Universidad de Concepción
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Matemática

An HDG method for a convection-diffusion equation with non-linear boundary conditions.

POR

Luciano Andrés Gajardo Chamblas

Tesis presentada a la Facultad de Ciencias Físicas y Matemáticas de la
Universidad de Concepción para optar al título profesional de
Ingeniero/a Civil Matemático/a

Profesor Guía: Dr. Manuel Solano Palma (Universidad de Conecpción).

Agosto de 2025,
Concepción, Chile.

An HDG method for a convection-diffusion equation with non-linear boundary conditions.

COMISIÓN EVALUADORA

Dr. Manuel Solano Palma [Profesor guía]

CI^{2}MA y Departamento de Ingeniería Matemática, Universidad de Concepción, Chile.

Dr. Rommel Bustinza

CI^{2}MA y Departamento de Ingeniería Matemática, Universidad de Concepción, Chile.

Dr. Patrick Vega Román

Departamento de Matemática y Ciencia de la Computación, Universidad de Santiago de Chile, Chile.

FECHA DE DEFENSA: 25 de Agosto de 2025.

Agradecimientos

En primer lugar, quiero expresar mi más profundo agradecimiento a mis padres, Juan Carlos y Elizabeth, por su amor y apoyo incondicional a lo largo de todo este proceso de estudios. Gracias a su cariño, consejos, regaloneos y a todas las veces que estuvieron a mi lado, tuve el privilegio de dedicarme por completo a aprender y crecer académicamente. Sin su respaldo constante no habría podido llegar hasta aquí. De corazón, muchas gracias por todo lo que han hecho por mí. También agradezco a mi madrina, Paola, por su constante presencia y apoyo en cada etapa de este camino.

Un agradecimiento muy especial a mi profesor guía, Manuel Solano, por brindarme la oportunidad de participar en este proyecto. Aprender y trabajar junto a él ha sido un verdadero privilegio. Le agradezco su dedicación, orientación, apoyo y sobre todo su paciencia durante este trabajo, así como en los cursos que tomé con él, los cuales me motivaron a seguir este camino académico. Espero que podamos mantenernos en contacto y colaborar nuevamente en el futuro.

Estoy también muy agradecido de los profesores que tuve a lo largo de la carrera. En particular, al profesor Freddy Paiva, por los gratos momentos compartidos y por la oportunidad de ser su alumno ayudante en varias ocasiones. Aprecio mucho esas experiencias, así como las conversaciones enriquecedoras, su sabiduría, sus enseñanzas, los regalos, los buenos consejos y, por supuesto, sus divertidos chistes.

Extiendo mis agradecimientos al profesor Gabriel Gatica, por todas las oportunidades que me brindó. Valoro enormemente sus clases interesantes, su sabiduría, su sentido del humor y también las veces en que me hizo aterrizar. Un reconocimiento especial al profesor Rommel Bustinza, quien formó parte de mi comisión evaluadora y de quien aprendí mucho tanto como estudiante como en mi rol de ayudante. También a la profesora Mónica Selva, con quien compartí más durante mi último semestre y con quien tuve enriquecedoras conversaciones.

Deseo destacar igualmente al profesor Leonardo Figueroa, con quien cursé numerosas asignaturas que fueron de gran interés y provecho; al profesor Rodolfo Rodríguez, por sus valiosas

clases, en particular el curso electivo que dictó; y a los profesores Dominique Spehner, Rodolfo Araya y Raimund Bürger, por las enseñanzas y el conocimiento compartido a lo largo de mi formación.

No puedo dejar de mencionar a los amigos que me acompañaron en este camino. En especial a Fernando Artaza y Estefanía Olivares, mis compañeros en el análisis numérico, por las risas, la amistad y las muchas anécdotas compartidas a lo largo de estos años. Espero que, a pesar de la distancia ahora que todos cerramos esta etapa, el contacto y el cariño se mantengan siempre. Gracias de corazón por todos los momentos vividos juntos, los llevo siempre conmigo.

También agradezco a los integrantes de un grupo cuyo nombre no puedo revelar acá: Camila Albornoz, Daniela Castañeda, Bruno Daveggio, Constanza Gacitúa, Annie González, Milene Gutiérrez, Yotan Hidalgo, Nikolas Jara y Ricardo Vega. Gracias por las jornadas de estudio, las risas y la buena onda que compartimos. Mención aparte para Brayan Sandoval, por su amistad y las conversaciones, aunque no siempre entendiera de qué me estaba hablando. Quiero destacar también al grupo de Runaterráneos, con Bastián Flores y Benjamín Mendieta, que junto con Fernando y Ricardo hicieron más divertidas varias noches de juego, a veces a costa de horas de sueño.

Agradezco igualmente al Centro de Investigación de Ingeniería Matemática (CI^2MA) por darme un espacio para trabajar, para tener clases y por el buen ambiente que siempre se vivió allí. Gracias en particular a la señora Lorena y a la señora Paola, cuyo trato amable y disponibilidad hicieron todo mucho más agradable.

Finalmente, gracias a los proyectos de ANID-Chile: FONDECYT N° 1240183, Anillo ACT210087 y Basal FB210005 por el apoyo para la realización de esta memoria.

Contents

Resumen

El principal objetivo de esta tesis es desarrollar un esquema de Galerkin Discontinuo Hibridizable (HDG) para una ecuación de convección-difusión con condiciones de contorno no lineales. La motivación proviene del proceso de ósmosis inversa aplicado a la desalinización de agua, que en su versión más completa involucra un sistema acoplado de ecuaciones de Navier-Stokes y convección-difusión, con incógnitas correspondientes a la presión, la velocidad del fluido y la concentración de sal. En este trabajo nos enfocamos exclusivamente en la ecuación de convección-difusión, considerando una condición de borde no lineal sobre una parte de la frontera, donde la única incógnita es la concentración de sal.

En primer lugar, se analiza la existencia y unicidad de solución del problema a nivel continuo mediante una formulación variacional mixta en el contexto de espacios de Banach, utilizando un enfoque basado en la perturbación de un punto de silla. Para garantizar el buen planteamiento del problema, se adopta una estrategia de punto fijo de Banach, resolviendo una versión linealizada del sistema en la que aparece una condición de borde del tipo Robin, y aplicando el teorema de Banach–Nečas–Babuška junto con la teoría de Babuška–Brezzi.

Posteriormente, se propone un esquema HDG para aproximar la solución de la formulación variacional continua, cuya estructura también es no lineal. Se emplea nuevamente un esquema de punto fijo, y para establecer el buen planteamiento del esquema linealizado se demuestra primero la dependencia continua respecto a los datos, utilizando argumentos de energía y dualidad. La existencia y unicidad del punto fijo en el esquema discreto se obtiene de manera similar al caso continuo, aunque bajo hipótesis más restrictivas.

Finalmente, se realiza un análisis de error a priori, estudiando las proyecciones de los errores y obteniendo resultados de convergencia óptimos bajo suposiciones similares a las consideradas en el análisis discreto. Por último, se presentan ensayos numéricos que corroboran las cotas teóricas obtenidas.

Abstract

The main objective of this thesis is to develop a Hybridizable Discontinuous Galerkin (HDG) scheme for a convection-diffusion equation with nonlinear boundary conditions. The motivation arises from the reverse osmosis process applied to water desalination, which, in its most complete form, involves a coupled system of Navier–Stokes and convection–diffusion equations, with unknowns corresponding to pressure, fluid velocity, and salt concentration. In this work, we focus exclusively on the convection–diffusion equation, considering a nonlinear boundary condition on part of the boundary, where the only unknown is the salt concentration.

First, we analyze the existence and uniqueness of the continuous problem using a mixed variational formulation in the context of Banach spaces, based on a saddle-point perturbation approach. To ensure the well-posedness of the problem, a Banach fixed-point strategy is adopted, solving a linearized version of the system involving a Robin-type boundary condition, and applying the Banach–Nečas–Babuška theorem along with the Babuška–Brezzi theory.

Then, an HDG scheme is proposed to approximate the solution of the continuous variational formulation, whose structure is also nonlinear. A fixed-point scheme is again employed, and to establish the well-posedness of the linearized scheme, we first prove continuous dependence on the data using energy and duality arguments. The existence and uniqueness of the fixed point in the discrete scheme are obtained similarly to the continuous case, but under more restrictive assumptions.

Finally, an *a priori* error analysis is carried out, studying the projections of the errors and obtaining optimal convergence results under assumptions similar to those considered in the discrete analysis. Lastly, numerical experiments are presented that confirm the theoretical bounds obtained.

CHAPTER 1

Introduction

The reverse osmosis process is nowadays one of the most commonly used techniques in water desalination plants [8]. Seawater flows into a channel at high pressure, and passes through the pores of a semi-permeable membrane that is able to retain colloidal matter and dissolved particles larger than 0.1-1.0 nm [12].

The mathematical model involves the Navier-Stokes equations for the fluid flow coupled to a convection-diffusion equation for the salt concentration. Solving this non-linear coupled system in the entire membrane module is expensive from the computational point of view. This is why, it is common to consider numerical simulations in a rectangular section, since desalination channels are typically represented by taking a rectangular cut of the channel. In this setting, inlet and outlet boundary conditions at the left (Γ_{in}) and right (Γ_{out}) boundaries of the channel, as depicted in Figure 1.1. The top and bottom boundaries, represent the semi-permeable membrane Σ where the gradient of the salt concentration depends in a nonlinear manner on the concentration [4].

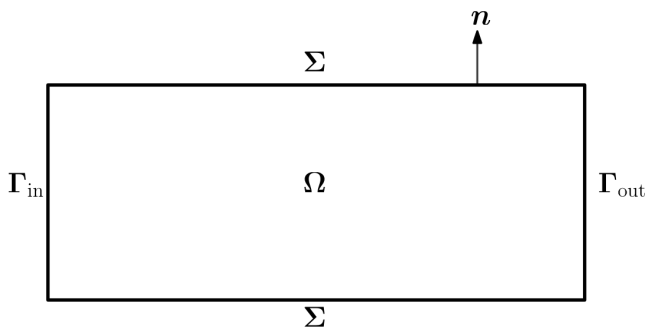

Figure 1.1: Sketch of the geometry.

The coupled system is usually solved by means of a fixed-point algorithm. That is, given a fluid velocity, the convection-diffusion equation is solved. Then, having the solution of the latter, the Navier-Stokes system is solved. The main goal of this work is to propose a hybridizable discontinuous Galerkin (HDG) scheme to approximate the solution of the convection-diffusion equation, as a stepping stone towards developing a HDG scheme for the coupled Navier-Stokes/convection-diffusion system.

1.1 Preliminaries

Let $\Omega \in \mathbb{R}^n, n \in \{2, 3\}$, be a bounded domain with polyhedral boundary $\partial\Omega$, and let $\boldsymbol{n}$ be the outward unit normal vector on $\partial\Omega$. We adopt standard notations for Lebesgue spaces $L^t(\Omega)$ and Sobolev spaces $W^{l,t}(\Omega)$, with $l \geq 0$ and $t \geq 1$, whose corresponding norms, either for the scalar and vectorial case, are denoted by $\|\cdot\|_{0,t;\Omega}$ and $\|\cdot\|_{l,t;\Omega}$, respectively. Note that $W^{0,t}(\Omega) = L^t(\Omega)$, and if $t = 2$ we write $H^l(\Omega)$ instead of $W^{t,2}(\Omega)$, with the corresponding norm and seminorm denoted by $\|\cdot\|_{l,\Omega}$ and $|\cdot|_{l,\Omega}$, respectively. In addition, $H^{1/2}(\partial\Omega)$ denotes the space of traces of $H^1(\Omega)$, and $H^{-1/2}(\partial\Omega)$ its dual space, provided with the duality pairing $\langle\cdot,\cdot\rangle_{\partial\Omega}$.

On the other hand, given any generic scalar functional space S, we let $[S]^d$ be the corresponding vector counterpart, whereas $\|\cdot\|$, with no subscripts, will be employed for the norm of any element or operator whenever there is no confusion about the space to which they belong. Vector-valued functions are boldfaced.

For any vector field $\boldsymbol{v} = (v_i)_{i=1,n}$, we set the gradient and divergence as

$$\nabla \boldsymbol{v} := \left(\frac{\partial v_i}{\partial v_j}\right)_{i,j=1,n} \quad \text{and} \quad \nabla \cdot \boldsymbol{v} := \sum_{i=1}^{n} \frac{\partial v_i}{\partial x_i}.$$

On the other hand, given $t \geq 1$, we introduce the Banach space

$$\boldsymbol{H}(\mathrm{div}_t;\Omega) := \{\boldsymbol{v} \in \boldsymbol{L}^2(\Omega) : \nabla \cdot \boldsymbol{v} \in L^t(\Omega)\}$$

equipped with the natural norm $\|\boldsymbol{v}\|_{\mathrm{div}_t;\Omega} := \|\boldsymbol{v}\|_{0,\Omega} + \|\nabla \cdot \boldsymbol{v}\|_{0,t;\Omega}$.

We recall that, proceeding as in [10, Theorem 1.7], one can prove that, in $\mathbb{R}^2$, for $t \in (1, +\infty]$ in there holds

$$\langle \boldsymbol{v} \cdot \boldsymbol{n}, \psi \rangle_{\partial\Omega} = \int_\Omega \psi \, \nabla \cdot \boldsymbol{v} + \int_\Omega \nabla \psi \cdot \boldsymbol{v}, \quad \forall (\boldsymbol{v}, \psi) \in \boldsymbol{H}(\mathrm{div}_t;\Omega) \times H^1(\Omega). \tag{1.1.1}$$

We recall some definitions and technical results concerning boundary conditions and extension operators from [10]. Let Γ_1 and Γ_2 be disjoint parts of $\partial\Omega$ such that $|\Gamma_1| \neq 0$ and $\partial\Omega = \overline{\Gamma}_1 \cup \overline{\Gamma}_2$. We define

$$H_{00}^{1/2}(\Gamma_2) := \{\mu|_{\Gamma_2} : \mu \in H^1(\Omega), \mu = 0 \quad \text{on} \quad \Gamma_1\} \tag{1.1.2}$$

Equivalently, if $E_{\Gamma_2,0} : H^{1/2}(\Gamma_2) \to L^2(\partial\Omega)$ is the extension operator

$$E_{\Gamma_2,0}(\mu) := \begin{cases} \mu & \text{on} \quad \Gamma_2 \\ 0 & \text{on} \quad \Gamma_1 \end{cases}, \quad \forall \mu \in H^{1/2}(\Gamma_2) \tag{1.1.3}$$

we have that

$$H_{00}^{1/2}(\Gamma_2) = \{\mu \in H^{1/2}(\Gamma_2) : E_{\Gamma_2,0}(\mu) \in H^{1/2}(\partial\Omega)\} \tag{1.1.4}$$

which is endowed with the norm $\|\mu\|_{1/2,00,\Gamma_2} := \|E_{\Gamma_2,0}(\mu)\|_{1/2,\partial\Omega}$. We recall that $H_{00}^{-1/2}(\Gamma_2)$ is the dual space of $H_{00}^{1/2}(\Gamma_2)$, then the duality between $H_{00}^{-1/2}(\Gamma_2)$ and $H_{00}^{1/2}(\Gamma_2)$, with respect to $L^2(\Gamma_2)$, is denoted by $\langle \cdot, \cdot \rangle_{\Gamma_2}$. In addition, given $\psi \in H^{-1/2}(\partial\Omega)$, its restriction to Γ_2, denoted

by $\psi|_{\Gamma_2}$ and defined by

$$\langle \psi|_{\Gamma_2}, \mu \rangle_{\Gamma_2} := \langle \psi, E_{\Gamma_2,0}(\mu) \rangle_{\partial\Omega}, \quad \forall \mu \in H_{00}^{1/2}(\Gamma_2), \tag{1.1.5}$$

clearly belongs to $H_{00}^{-1/2}(\Gamma_2)$. Moreover, we have that

$$\|\psi|_{\Gamma_2}\|_{-1/2,00,\Gamma_2} \leq \|\psi\|_{-1/2,\partial\Omega}, \quad \forall \psi \in H^{-1/2}(\partial\Omega) \tag{1.1.6}$$

1.2 The model problem

We consider a given fluid velocity $\boldsymbol{\beta} \in [L^4(\Omega)]^2$ (we can think this is a discretized velocity coming from a previous iteration of a fixed point algorithm) such that $\nabla \cdot \boldsymbol{\beta} = 0$ in Ω. We look for the salt concentration ϕ such that

$$\begin{cases} -\kappa\Delta\phi + \boldsymbol{\beta} \cdot \nabla\phi &= f & \text{in } \Omega, \\ \phi &= \phi_{\text{in}} & \text{in } \Gamma_{\text{in}}, \\ \kappa\nabla\phi \cdot \boldsymbol{n} + g(\phi) &= 0 & \text{in } \Gamma_N := \Sigma \cup \Gamma_{\text{out}}, \end{cases} \tag{1.2.1}$$

where κ is the solute diffusivity through the solvent, ϕ_{in} is the salt concentration at the inlet, and f is a source term. In addition, g is given by

$$g(\phi) := \begin{cases} 0 & \text{in } \Gamma_{\text{out}}, \\ a_3\phi + a_1\phi^2 & \text{in } \Sigma, \end{cases} \tag{1.2.2}$$

where $a_0 := A\Delta P$, $a_1 := AiRT$, and $a_2 := B$ are positive constants and can be calculated using the values in Table 6.2 of [1]; and $a_3 = a_2 - a_0$. In the reverse osmosis process f is zero, but we include it in the model to handle more general situations.

CHAPTER 2

Continuous solvability analysis

2.1 The mixed formulation

In this section, we follow [1] to derive a mixed formulation for (1.2.1) within a Banach spaces framework. In order to describe the geometry, we let Ω be an open bounded simply connected polygonal domain in $\mathbb{R}^2$ such that $\partial\Omega$ is divided in three parts: Γ_{in} (inlet), Γ_{out} (outlet) and Σ (wall) such that $\partial\Omega = \Gamma_{\text{in}} \cup \overline{\Sigma} \cup \Gamma_{\text{out}}$, as depicted in Figure 1.1.

In order to derive a mixed formulation, we set $\boldsymbol{q} := -\kappa\nabla\phi$. In this way, (1.2.1) can be rewritten equivalently as follows: Find $(\boldsymbol{q}, \phi)$ in suitable spaces to be indicated below, such that

$$\left\{\begin{array}{rcll} \kappa^{-1}\boldsymbol{q} + \nabla\phi &=& 0 & \text{in } \Omega, \\ \nabla\cdot\boldsymbol{q} - \kappa^{-1}\boldsymbol{\beta}\cdot\boldsymbol{q} &=& f & \text{in } \Omega, \\ \phi &=& \phi_{\text{in}}, & \text{in } \Gamma_{\text{in}}, \\ \boldsymbol{q}\cdot\boldsymbol{n} - g(\phi) &=& 0 & \text{in } \Gamma_N. \end{array}\right. \tag{2.1.1}$$

We first consider the change of variable $\phi_* = \phi - \phi_{\text{in}}$ in order to obtain a homogeneous Dirichlet

boundary condition. We notice that $g(\phi) = g(\phi_* + \phi_{\text{in}}) = \varphi(\phi_*)\phi_* + g(\phi_{\text{in}})$, where

$$\varphi(\zeta) := \begin{cases} 0 & \text{in } \Gamma_{\text{out}}, \\ c_3 + c_1\zeta & \text{in } \Sigma. \end{cases} \tag{2.1.2}$$

Here $c_3 := a_3 + 2a_1\phi_{\text{in}}$, $c_1 := a_1$. Then, since ϕ_{in} is constant, (2.1.1) becomes

$$\begin{cases} \kappa^{-1}\boldsymbol{q} + \nabla\phi_* &= 0 & \text{in } \Omega, \\ \nabla\cdot\boldsymbol{q} - \kappa^{-1}\boldsymbol{\beta}\cdot\boldsymbol{q} &= f & \text{in } \Omega, \\ \phi_* &= 0, & \text{in } \Gamma_{\text{in}}, \\ \boldsymbol{q}\cdot\boldsymbol{n} - \varphi(\phi_*)\phi_* &= g(\phi_{\text{in}}) & \text{in } \Gamma_N. \end{cases} \tag{2.1.3}$$

We begin by looking originally for $\phi_* \in H^1(\Omega)$. Then multiplying the first equation of (2.1.3) by $\boldsymbol{r} \in \boldsymbol{H}(\text{div}_t;\Omega)$ with $t \in]1,\infty[$, applying the integration by parts formula, we find that

$$\kappa^{-1}\int_\Omega \boldsymbol{q}\cdot\boldsymbol{r} - \int_\Omega \phi_*\,\nabla\cdot\boldsymbol{r} + \langle \boldsymbol{r}\cdot\boldsymbol{n}, \phi_*\rangle_{\partial\Omega} = 0, \quad \forall \boldsymbol{r} \in \boldsymbol{H}(\text{div}_t;\Omega). \tag{2.1.4}$$

It is clear that the first term on the left-hand side is well-defined for $\boldsymbol{q} \in \boldsymbol{L}^2(\Omega)$. In addition, knowing that $\nabla\cdot\boldsymbol{r} \in L^t(\Omega)$, we realize from Hölder's inequality that it is sufficient to look for $\phi_* \in L^{t'}(\Omega)$ such that $\frac{1}{t} + \frac{1}{t'} = 1$. Since traces of $L^{t'}(\Omega)$-functions are not defined, we introduce a Lagrange multiplier $\lambda := \phi_*\big|_{\Gamma_N} \in H^{1/2}_{00}(\Gamma_N)$, and realize that (2.1.4) becomes

$$\kappa^{-1}\int_\Omega \boldsymbol{q}\cdot\boldsymbol{r} - \int_\Omega \phi_*\,\nabla\cdot\boldsymbol{r} + \langle \boldsymbol{r}\cdot\boldsymbol{n}, \lambda\rangle_{\Gamma_N} = 0, \quad \forall \boldsymbol{r} \in \boldsymbol{H}(\text{div}_t;\Omega). \tag{2.1.5}$$

Here we used the fact that $\phi_*\big|_{\Gamma_{\text{in}}} = 0$, and $\boldsymbol{r}\cdot\boldsymbol{n}$ is well defined since $\boldsymbol{r} \in \boldsymbol{H}(\text{div}_t;\Omega)$. Next, from the second equation of (2.1.3), with $\psi \in L^s(\Omega)$ as test function, we obtain

$$\int_\Omega \nabla\cdot\boldsymbol{q}\;\psi - \kappa^{-1}\int_\Omega \boldsymbol{q}\cdot\boldsymbol{\beta}\;\psi = \int_\Omega f\psi, \quad \forall \psi \in L^s(\Omega). \tag{2.1.6}$$

For the second term on the left-hand side, since $\boldsymbol{q} \in \boldsymbol{L}^2(\Omega)$ by Hölder's inequality, there holds

$$\left|\int_\Omega \boldsymbol{q}\cdot\boldsymbol{\beta}\;\psi\right| \le \|\boldsymbol{q}\|_{0;\Omega}\;\|\boldsymbol{\beta}\|_{0,p;\Omega}\;\|\psi\|_{0,s;\Omega}\;, \tag{2.1.7}$$

where $\frac{1}{s}+\frac{1}{p}=\frac{1}{2}$ with $s,p\in(2,+\infty)$. The other terms are well-defined if we assume that $\nabla\cdot\boldsymbol{q}$ and f are in $L^{s'}(\Omega)$, where s' is the conjugate of s . Since we would like to seek $\boldsymbol{q}$ and $\boldsymbol{r}$ in the same space, it follows that $s'=t$, so ϕ_* and ψ belong to the same space. We can write t and t' in terms of p as

$$t=\frac{2p}{p+2}\quad\text{and}\quad t'=\frac{2p}{p-2}.$$

Since we are interested in the Navier-Stokes/convection-diffusion equation coupled problems, we know that $p=4$ (see [1, Section 2]), so $t'=4$ and $t=4/3$.

Now for the fourth equation of (2.1.3), after testing against $\mu\in H_{00}^{1/2}(\Gamma_N)$ and using the definition of the Lagrange multiplier λ, it follows that

$$\langle\boldsymbol{q}\cdot\boldsymbol{n},\mu\rangle_{\Gamma_N}-\langle\varphi(\lambda)\lambda,\mu\rangle_{\Gamma_N}=\langle g(\phi_{\text{in}}),\mu\rangle_{\Gamma_N},\quad\forall\mu\in H_{00}^{1/2}(\Gamma_N)\tag{2.1.8}$$

Consequently, we introduce the following spaces

$$\boldsymbol{H}:=\boldsymbol{H}(\text{div}_{4/3};\Omega),\quad Q_1:=L^4(\Omega),\quad Q_2:=H_{00}^{1/2}(\Gamma_N),\ \text{and}\ \boldsymbol{Q}:=Q_1\times Q_2.$$

and equipping these spaces with the norms

$$\begin{aligned}\|\boldsymbol{r}\|_{\boldsymbol{H}}&:=\|\boldsymbol{r}\|_{\text{div};4/3;\Omega}&&\forall\boldsymbol{r}\in\boldsymbol{H},\\ \|(\psi,\mu)\|_{\boldsymbol{Q}}&:=\|\psi\|_{0,4;\Omega}+\|\mu\|_{1/2,00;\Gamma_N}&&\forall(\psi,\mu)\in\boldsymbol{Q},\\ \|(\boldsymbol{r},(\psi,\mu))\|_{\boldsymbol{H}\times\boldsymbol{Q}}&:=\|\boldsymbol{r}\|_{\boldsymbol{H}}+\|(\psi,\mu)\|_{\boldsymbol{Q}}&&\forall(\boldsymbol{r},(\psi,\mu))\in\boldsymbol{H}\times\boldsymbol{Q}.\end{aligned}$$

We define the bilinear forms $a:\boldsymbol{H}\times\boldsymbol{H}\to\mathbb{R}$, and $b,c:\boldsymbol{H}\times\boldsymbol{Q}\to\mathbb{R}$, for all $\boldsymbol{q},\boldsymbol{r}\in\boldsymbol{H}$ and all $(\psi,\mu)\in\boldsymbol{Q}$, as

$$\begin{aligned}a(\boldsymbol{q},\boldsymbol{r})&:=(\kappa^{-1}\boldsymbol{q},\boldsymbol{r})_\Omega,\\ b(\boldsymbol{r},(\psi,\mu))&:=-(\psi,\nabla\cdot\boldsymbol{r})_\Omega+\langle\boldsymbol{r}\cdot\boldsymbol{n},\mu\rangle_{\Gamma_N},\\ c(\boldsymbol{r},(\psi,\mu))&:=(\kappa^{-1}\boldsymbol{r}\cdot\boldsymbol{\beta},\psi)_\Omega.\end{aligned}$$

We define the linear functional $\mathbf{G} \in \boldsymbol{Q}'$, for all $(\psi, \mu) \in \boldsymbol{Q}$, as

$$\mathbf{G}(\psi, \mu) := -(f, \psi)_\Omega + \langle g(\phi_{\text{in}}), \mu \rangle_\Sigma = -(f, \psi)_\Omega + \langle c_2, \mu \rangle_\Sigma, \tag{2.1.9}$$

where $c_2 := a_3\phi_{\text{in}} + a_1\phi_{\text{in}}^2$. Now, given $\zeta \in Q_2$, we define the bilienar from $d_\zeta : Q_2 \times Q_2 \to \mathbb{R}$, for all $(\lambda, \mu) \in Q_2$, as

$$d_\zeta(\lambda, \mu) := \langle \varphi(\zeta)\lambda, \mu \rangle_\Sigma.$$

Hence, we arrive at the next variational formulation: Find $(\boldsymbol{q}, (\phi_*, \lambda)) \in \boldsymbol{H} \times \boldsymbol{Q}$ such that

$$\begin{array}{rcrcl} a(\boldsymbol{q}, \boldsymbol{r}) & + & b(\boldsymbol{r}, (\phi_*, \lambda)) & = & 0, \\ b(\boldsymbol{q}, (\psi, \mu)) & + & c(\boldsymbol{q}, (\psi, \mu)) - d_\lambda(\lambda, \mu) & = & \mathbf{G}(\psi, \mu), \end{array} \tag{2.1.10}$$

for all $\Big(\boldsymbol{r}, (\psi, \mu)\Big) \in \boldsymbol{H} \times \boldsymbol{Q}$.

2.2 Fixed-point operator

We begin by rewriting (2.1.10) as an equivalent fixed-point equation. We define $T : Q_2 \to Q_2$ as the operator defined for each $\zeta \in Q_2$ as

$$T(\zeta) := \lambda^\zeta, \tag{2.2.1}$$

where $(\boldsymbol{q}^\zeta, (\phi_*^\zeta, \lambda^\zeta)) \in \boldsymbol{H} \times \boldsymbol{Q}$ is the unique solution (to be confirmed later) of the following linearized problem: Find $(\boldsymbol{q}^\zeta, (\phi_*^\zeta, \lambda^\zeta)) \in \boldsymbol{H} \times \boldsymbol{Q}$ such that

$$\begin{array}{rcrcl} a(\boldsymbol{q}^\zeta, \boldsymbol{r}) & + & b(\boldsymbol{r}, (\phi_*^\zeta, \lambda^\zeta)) & = & 0, \\ b(\boldsymbol{q}^\zeta, (\psi, \mu)) & + & c(\boldsymbol{q}^\zeta, (\psi, \mu)) - d_\zeta(\lambda^\zeta, \mu) & = & \mathbf{G}(\psi, \mu), \end{array} \tag{2.2.2}$$

for all $(\boldsymbol{r}, (\psi, \mu)) \in \boldsymbol{H} \times \boldsymbol{Q}$. We observe that solving (2.1.10) is equivalent to seeking a fixed point of T, that is: Find $\lambda \in Q_2$ such that $T(\lambda) = \lambda$.

Before stating the boundedness of all variational forms involved in (2.1.10) and (2.2.2), we

recall the following injections. First, $\boldsymbol{c_t} : H^{1/2}(\partial\Omega) \to L^t(\partial\Omega)$ denotes the compact injection from $H^{1/2}(\partial\Omega)$ into $L^t(\partial\Omega)$ for $t \geq 1$ (see [9, Theorem B.46] for the case $d = 1$, $s = 1/2$, and $p = 2$). Second, $\boldsymbol{i_t} : H^1(\Omega) \to L^t(\Omega)$ denotes the continuous injection from $H^1(\Omega)$ into $L^t(\Omega)$.

Now, in order to prove that the operator T is well defined, we can define the following quantities

$$\|a\| := \kappa^{-1}, \quad \|b\| := \|\boldsymbol{i_4}\| + 2, \quad \|c\| := \kappa^{-1}\|\boldsymbol{\beta}\|_{0,4;\Omega}, \quad \|\mathbf{G}\| := \|f\|_{0,4/3,\Omega} + \|\boldsymbol{c_2}\|\,|c_2|\,|\Sigma|^{1/2}, \tag{2.2.3}$$

such that the following holds

$$|a(\boldsymbol{q},\boldsymbol{r})| \leq \|a\|\ \|\boldsymbol{q}\|_{\boldsymbol{H}}\ \|\boldsymbol{r}\|_{\boldsymbol{H}} \qquad \forall \boldsymbol{q},\boldsymbol{r} \in \boldsymbol{H}, \tag{2.2.4a}$$

$$|b(\boldsymbol{r},(\psi,\mu))| \leq \|b\|\ \|\boldsymbol{r}\|_{\boldsymbol{H}}\ \|(\psi,\mu)\|_{\boldsymbol{Q}} \qquad \forall (\boldsymbol{r},(\psi,\mu)) \in \boldsymbol{H}\times\boldsymbol{Q}, \tag{2.2.4b}$$

$$|c(\boldsymbol{r},(\psi,\mu))| \leq \|c\|\ \|\boldsymbol{r}\|_{\boldsymbol{H}}\ \|(\psi,\mu)\|_{\boldsymbol{Q}} \qquad \forall (\boldsymbol{r},(\psi,\mu)) \in \boldsymbol{H}\times\boldsymbol{Q}, \tag{2.2.4c}$$

$$|\mathbf{G}(\psi,\mu)| \leq \|\mathbf{G}\|\ \|(\psi,\mu)\|_{\boldsymbol{Q}} \qquad \forall (\psi,\mu) \in \boldsymbol{Q}, \tag{2.2.4d}$$

where $\|a\|$ and $\|c\|$ are obtained by direct applications of Cauchy-Schwarz and Hölder inequalities. For the boundedness of b, we proceed similarly as in [1, Appendix A.1.]. Thus, by Hölder's inequality we have

$$|b(\boldsymbol{r},(\psi,\mu))| \leq \|\boldsymbol{r}\|_{\boldsymbol{H}}\ \|(\psi,\mu)\|_{\boldsymbol{Q}} + |\langle \boldsymbol{r}\cdot\boldsymbol{n},\mu\rangle_{\Gamma_N}|.$$

Then, by (1.1.5) and (1.1.1) we have

$$\langle \boldsymbol{r}\cdot\boldsymbol{n},\mu\rangle_{\Gamma_N} = \langle \boldsymbol{r}\cdot\boldsymbol{n}, E_{\Gamma_N,0}(\mu)\rangle_{\partial\Omega} = \int_\Omega \tilde{\gamma}_0^{-1}(E_{\Gamma_N,0}(\mu))\ \nabla\cdot\boldsymbol{r} + \int_\Omega \boldsymbol{r}\cdot\nabla\big(\tilde{\gamma}_0^{-1}(E_{\Gamma_N,0}(\mu))\big),$$

where $\tilde{\gamma}_0^{-1} : H^{1/2}(\partial\Omega) \to [H_0^1(\Omega)]^\perp$ is the right inverse of the trace operator $\gamma_0 : H^1(\Omega) \to H^{1/2}(\partial\Omega)$ (see [10, Section 1.3.4]). Thus, applying Cauchy-Schwarz and Hölder's inequalities we obtain

$$\begin{aligned}\int_\Omega \tilde{\gamma}_0^{-1}(E_{\Gamma_N,0}(\mu))\ \nabla\cdot\boldsymbol{r} &\leq \|\tilde{\gamma}_0^{-1}(E_{\Gamma_N,0}(\mu))\|_{0,4;\Omega}\ \|\nabla\cdot\boldsymbol{r}\|_{0,4/3;\Omega} \leq \|\boldsymbol{i_4}\|\ \|\tilde{\gamma}_0^{-1}(E_{\Gamma_N,0}(\mu))\|_{1,\Omega}\ \|\boldsymbol{r}\|_{\boldsymbol{H}} \\ &= \|\boldsymbol{i_4}\|\ \|E_{\Gamma_N,0}(\mu)\|_{1/2,\partial\Omega}\ \|\boldsymbol{r}\|_{\boldsymbol{H}} = \|\boldsymbol{i_4}\|\ \|\boldsymbol{r}\|_{\boldsymbol{H}}\ \|\mu\|_{1/2,00,\Gamma_N}\end{aligned}$$

$$\leq \|\boldsymbol{i_4}\| \ \|\boldsymbol{r}\|_{\boldsymbol{H}} \ \|(\psi,\mu)\|_{\boldsymbol{Q}},$$

and

$$\begin{aligned}\int_\Omega \boldsymbol{r}\cdot\nabla\Big(\tilde{\gamma}_0^{-1}(E_{\Gamma_N,0}(\mu))\Big) &\leq \|\boldsymbol{r}\|_{0,\Omega} \ \|\nabla\left(\tilde{\gamma}_0^{-1}(E_{\Gamma_N,0}(\mu))\right)\|_{0,\Omega} \leq \|\boldsymbol{r}\|_{\boldsymbol{H}} \ \|\tilde{\gamma}_0^{-1}(E_{\Gamma_N,0}(\mu))\|_{1,\Omega}\\ &= \|\boldsymbol{r}\|_{\boldsymbol{H}} \ \|E_{\Gamma_N,0}(\mu)\|_{1/2,\partial\Omega} = \|\boldsymbol{r}\|_{\boldsymbol{H}} \ \|\mu\|_{1/2,00,\Gamma_N}\\ &\leq \|\boldsymbol{r}\|_{\boldsymbol{H}} \ \|(\psi,\mu)\|_{\boldsymbol{Q}}.\end{aligned}$$

From this, we obtain $\|b\|$. Next, in order to get $\|\mathbf{G}\|$, since $g(\phi_{\text{in}}) \in L^2(\Gamma_N)$, by Cauchy-Scwarz and Hölder's inequalities we get that

$$\begin{aligned}\mathbf{G}(\psi,\mu) = -\int_\Omega f\psi - \int_\Sigma c_2\mu &\leq \|f\|_{0,4/3;\Omega} \ \|\psi\|_{0,4;\Omega} + |c_2| \ \|1\|_{0,\Sigma} \ \|\mu\|_{0,\Sigma}\\ &\leq \|f\|_{0,4/3,\Omega} \ \|(\psi,\mu)\|_{\boldsymbol{Q}} + |c_2| \ |\Sigma|^{1/2} \ \|E_{\Gamma_N,0}(\mu)\|_{0,\partial\Omega}\\ &\leq \|f\|_{0,4/3,\Omega} \ \|(\psi,\mu)\|_{\boldsymbol{Q}} + \|\boldsymbol{c_2}\| \ |c_2| \ |\Sigma|^{1/2} \ \|E_{\Gamma_N,0}(\mu)\|_{1/2,\partial\Omega}\\ &\leq \|f\|_{0,4/3,\Omega} \ \|(\psi,\mu)\|_{\boldsymbol{Q}} + \|\boldsymbol{c_2}\| \ |c_2| \ |\Sigma|^{1/2} \ \|(\psi,\mu)\|_{\boldsymbol{Q}}.\end{aligned}$$

From this, we can obtain directly $\|\mathbf{G}\|$. Now we need a bound for d_ζ, and for this purpose we notice that

$$\|\varphi(\zeta)\|_{0,\Sigma} \leq |c_3| \ |\Sigma|^{1/2} + |c_1| \ \|\zeta\|_{0,\Sigma} \leq |c_3| \ |\Sigma|^{1/2} + |c_1| \ \|\boldsymbol{c_2}\|\|\zeta\|_{1/2,00;\Gamma_N}.$$

Now, the bilinear form d_ζ will be bounded in a ball. More precisely, let us define

$$B := \{\zeta \in Q_2 : \|\zeta\|_{1/2,00;\Gamma_N} \leq R\}, \tag{2.2.5}$$

with $R > 0$ to be defined in Lemma 2.2.2. Now, for every $\zeta \in B$, we get that $\|\varphi(\zeta)\|_{0,\Sigma} \leq M_\varphi$, where $M_\varphi := |c_3| \ |\Sigma|^{1/2} + |c_1| \ \|\boldsymbol{c_2}\|R$. Under this assumption, we have that

$$|d_\zeta(\lambda,\mu)| = |\langle\varphi(\zeta)\lambda,\mu\rangle_\Sigma| \leq \|\varphi(\zeta)\|_{0,\Sigma} \ \|\lambda\|_{0,4;\Sigma} \ \|\mu\|_{0,4;\Sigma} \leq M_\varphi\|\boldsymbol{c_4}\|^2\|\lambda\|_{1/2,00,\Gamma_N} \ \|\mu\|_{1/2,00,\Gamma_N}$$

and we can define

$$\|d_\zeta\| := M_\varphi \|\boldsymbol{c_4}\|^2 \tag{2.2.6}$$

and get that $d_\zeta(\lambda,\mu) \leq \|d_\zeta\|\|\lambda\|_{1/2,00;\Gamma_N}\|\mu\|_{1/2,00;\Gamma_N}$. We showed that all variational forms involved in (2.1.10) and (2.2.2) are bounded.

2.2.1 Well-definedness of the operator T

We will show that (2.2.2) is well-posed, and therefore the operator T is well-defined. For this purpose, we will use the Banach-Nečas-Babuška [9, Theorem 2.6], along with the Banach version of Babuška-Brezzi theory. We begin by remarking that, being $L^t(\Omega)$ reflexive for each $t > 1$, all the spaces involved, namely $\boldsymbol{H}(\mathrm{div}_{4/3};\Omega)$, $L^4(\Omega)$ and $H^{1/2}_{00}(\Gamma_N)$ are easily shown to be reflexive as well. Also, we notice that $\mathbf{G}$ is linear and bounded.

First, in what follows we address the solvability of the next problem, which satisfies hypotheses of [9, Theorem 2.34]: Find $(\boldsymbol{q},(\phi_*,\lambda)) \in \boldsymbol{H}\times\boldsymbol{Q}$ such that

$$\begin{array}{rcl} a(\boldsymbol{q},\boldsymbol{r}) + b(\boldsymbol{r},(\phi_*,\lambda)) &=& 0, \\ b(\boldsymbol{q},(\psi,\mu)) &=& \mathbf{G}(\psi,\mu), \end{array} \tag{2.2.7}$$

for all $(\boldsymbol{r},(\psi,\mu)) \in \boldsymbol{H}\times\boldsymbol{Q}$.

Now, letting $\boldsymbol{V}$ be the null space of the linear and bounded operator induced by b, we readily see that

$$\boldsymbol{V} = \{\boldsymbol{t}\in\boldsymbol{H} : \nabla\cdot\boldsymbol{t} = 0,\ (\boldsymbol{t}\cdot\boldsymbol{n})|_{\Gamma_N} = 0\}. \tag{2.2.8}$$

It is straightforward to see from the definition of a that, for each $\boldsymbol{r}\in\boldsymbol{V}$, there holds

$$a(\boldsymbol{r},\boldsymbol{r}) \geq \alpha\|\boldsymbol{r}\|^2_{\boldsymbol{H}}.$$

with $\alpha = \kappa^{-1}/2$. Now, we provide the corresponding inf-sup condition for b. We recall that its proof is basically an adaptation of the work done in [3, Lemma 3.4] and [10, Section 2.4.2].

Lemma 2.2.1. *There exists a positive constant β such that*

$$\sup_{\substack{\boldsymbol{q}\in\boldsymbol{H}\\ \boldsymbol{q}\neq 0}} \frac{b(\boldsymbol{q},(\psi,\mu))}{\|\boldsymbol{q}\|_{\boldsymbol{H}}} \geq \beta\|(\psi,\mu)\|_{\boldsymbol{Q}}, \quad \forall(\psi,\mu)\in\boldsymbol{Q}. \tag{2.2.9}$$

Proof. Given $\psi \in Q_1$, we set $\psi_{4/3} := |\psi|^2\psi$ and observe that $|\psi_{4/3}|^{4/3} = |\psi|^4$, which implies that $\psi_{4/3} \in L^{4/3}(\Omega)$ and that

$$\int_\Omega \psi\psi_{4/3} = \|\psi\|_{0,4;\Omega}\ \|\psi_{4/3}\|_{0,4/3;\Omega}.$$

Then , given $\psi_{4/3}$, we let $z_1 \in \boldsymbol{H}^1_{\Gamma_{\rm in}}(\Omega)$ be the unique weak solution of the boundary problem:

$$\left\{\begin{array}{rcll} -\Delta z_1 &=& \psi_{4/3} & \text{in } \Omega,\\ z_1 &=& 0 & \text{in } \Gamma_{\rm in},\\ \nabla z_1\cdot\boldsymbol{n} &=& 0, & \text{in } \Gamma_N. \end{array}\right.$$

By the Lax-Milgram Lemma, we get that $\|z_1\|_{1,\Omega} \leq C_P^{-2}\ \|\boldsymbol{i_4}\|\ \|\psi_{4/3}\|_{0,4/3;\Omega}$, where C_P is the Poincaré constant, such that $\|z_1\|_{1,\Omega} \leq C_P|z|_{1,\Omega}$.

Thus, by defining $\widetilde{q_1} := \nabla z_1$ we notice that $\nabla\cdot\widetilde{q_1} = -\psi_{4/3} \in L^{4/3}(\Omega)$ and that $\widetilde{q_1}\cdot\boldsymbol{n}\big|_{\Gamma_N} = 0$, and then $\widetilde{q_1} \in \boldsymbol{H}$. A simple calculation leads to $\|\widetilde{q_1}\|_{\boldsymbol{H}} \leq (C_P^{-2}\|\boldsymbol{i_4}\| + 1)\ \|\psi_{4/3}\|_{0,4/3;\Omega}$.

Now, it is easy to get that

$$S \geq (C_P^{-2}\|\boldsymbol{i_4}\| + 1)^{-1}\|\psi\|_{0,4;\Omega}, \tag{2.2.10}$$

where $S := \sup\limits_{\substack{\boldsymbol{q}\in\boldsymbol{H}\\ \boldsymbol{q}\neq 0}} \dfrac{b(\boldsymbol{q},(\psi,\mu))}{\|\boldsymbol{q}\|_{\boldsymbol{H}}}$.

Now, given $\mu \in H^{1/2}_{00}(\Gamma_N)$, we let $z_2 \in \boldsymbol{H}^1_{\Gamma_{\rm in}}(\Omega)$ be the unique weak solution of the boundary problem:

$$\left\{\begin{array}{rcll} -\Delta z_2 &=& 0 & \text{in } \Omega,\\ z_2 &=& 0 & \text{in } \Gamma_{\rm in},\\ \nabla z_2\cdot\boldsymbol{n} &=& \mathcal{R}_{00}^{-1}(\mu), & \text{in } \Gamma_N. \end{array}\right.$$

where $\mathcal{R}_{00} : H_{00}^{-1/2}(\Gamma_N) \to H_{00}^{1/2}(\Gamma_N)$ is the corresponding Riesz mapping. By the Lax-Milgram Lemma, we get that $\|z_2\|_{1,\Omega} \leq C_P^{-2}\ C_{tr}\ \|\mu\|_{1/2,00,\Gamma_N}$, where C_{tr} comes from the trace inequality.

Thus, by defining $\widetilde{q}_2 := \nabla z_2$ we notice that $\nabla \cdot \widetilde{q}_2 = 0 \in L^{4/3}(\Omega)$ and that $\widetilde{q}_2 \cdot \boldsymbol{n}\Big|_{\Gamma_N} = \mathcal{R}_{00}^{-1}(\mu)$, and then $\widetilde{q}_2 \in \boldsymbol{H}$. A simple calculation leads to $\|\widetilde{q}_2\|_{\boldsymbol{H}} \leq C_P^{-2}\ C_{tr}\ \|\mu\|_{1/2,00,\Gamma_N}$.

Now, it is easy to get that

$$S \geq C_P^2\ C_{tr}^{-1}\ \|\mu\|_{1/2,00,\Gamma_N}. \tag{2.2.11}$$

From (2.2.10) and (2.2.11) we get the inf-sup condition for b, where

$$\beta := \frac{1}{2}\min\{(C_P^{-2}\|\boldsymbol{i_4}\| + 1)^{-1}, C_P^2\ C_{tr}^{-1}\}.$$

□

In other words, thanks to [9, Theorem 2.34], the non-perturbed problem (2.2.7) is well-posed. In addition, (2.2.2) is also well-posed according to the following result.

Lemma 2.2.2. *Let us assume that the radius R in (2.2.5), and the given data c_1, c_2, c_3, κ and $\boldsymbol{\beta}$ satisfy*

$$\kappa^{-1}\|\boldsymbol{\beta}\|_{0,4;\Omega} + \|\boldsymbol{c_4}\|^2\left(|c_3|\ |\Sigma|^{1/2} + c_1\ \|\boldsymbol{c_2}\|R\right) < \alpha_A/2.$$

Then problem (2.2.16) has a unique solution, equivalently T is well-posed. Moreover, the following a priori estimate holds

$$\|(\boldsymbol{q}^\zeta, (\phi_*^\zeta, \lambda^\zeta))\|_{\boldsymbol{H}\times\boldsymbol{Q}} \leq \frac{2}{\alpha_A}\left(\|f\|_{0,4/3,\Omega} + \|\boldsymbol{c_2}\|\ |c_2|\ |\Sigma|^{1/2}\right). \tag{2.2.12}$$

Proof. Consider the following problem: Find $(\boldsymbol{q}, (\phi_*, \lambda)) \in \boldsymbol{H} \times \boldsymbol{Q}$ such that

$$\boldsymbol{A}\Big((\boldsymbol{q}, (\phi_*, \lambda)), (\boldsymbol{r}, (\psi, \mu))\Big) \quad = \quad \tilde{\mathbf{G}}(\boldsymbol{r}, (\psi, \mu)), \tag{2.2.13}$$

for all $(\boldsymbol{r}, (\psi, \mu)) \in \boldsymbol{H} \times \boldsymbol{Q}$, where $\boldsymbol{A} : (\boldsymbol{H} \times \boldsymbol{Q}) \times (\boldsymbol{H} \times \boldsymbol{Q}) \to \mathbb{R}$ be the bilinear form given by

$$\boldsymbol{A}\Big((\boldsymbol{q}, (\phi_*, \lambda)), (\boldsymbol{r}, (\psi, \mu))\Big) := a(\boldsymbol{q}, \boldsymbol{r}) + b(\boldsymbol{r}, (\phi_*, \lambda)) + b(\boldsymbol{q}, (\psi, \mu)),$$

and $\tilde{\mathbf{G}} : \boldsymbol{H} \times \boldsymbol{Q} \to \mathbb{R}$ is the linear functional defined by $\tilde{\mathbf{G}}(\boldsymbol{r}, (\psi, \mu)) := \mathbf{G}(\psi, \mu)$. We notice that problems (2.2.7) and (2.2.13) are equivalent, and since (2.2.7) is well-posed, by [9, Theorem 2.6] there exists $\alpha_A > 0$ such that, for all $(\boldsymbol{q}, (\phi_*, \lambda)) \in \boldsymbol{H} \times \boldsymbol{Q}$

$$S_A := \sup_{\substack{(\boldsymbol{r},(\psi,\mu))\in \boldsymbol{H}\times\boldsymbol{Q} \\ (\boldsymbol{r},(\psi,\mu))\neq 0}} \frac{\boldsymbol{A}\Big((\boldsymbol{q}, (\phi_*, \lambda)), (\boldsymbol{r}, (\psi, \mu))\Big)}{\|(\boldsymbol{r}, (\psi, \mu))\|_{\boldsymbol{H}\times\boldsymbol{Q}}} \geq \alpha_A \, \|(\boldsymbol{q}, (\phi_*, \lambda))\|_{\boldsymbol{H}\times\boldsymbol{Q}}, \tag{2.2.14}$$

which implies that there exists $(\tilde{\boldsymbol{r}}, (\tilde{\psi}, \tilde{\mu})) \in \boldsymbol{H} \times \boldsymbol{Q}$, with $(\tilde{\boldsymbol{r}}, (\tilde{\psi}, \tilde{\mu})) \neq \mathbf{0}$ such that

$$\boldsymbol{A}\Big((\boldsymbol{q}, (\phi_*, \lambda)), (\tilde{\boldsymbol{r}}, (\tilde{\psi}, \tilde{\mu}))\Big) \geq \alpha_A \|(\boldsymbol{q}, (\phi_*, \lambda))\|_{\boldsymbol{H}\times\boldsymbol{Q}} \|(\tilde{\boldsymbol{r}}, (\tilde{\psi}, \tilde{\mu}))\|_{\boldsymbol{H}\times\boldsymbol{Q}}, \quad \forall (\boldsymbol{q}, (\phi_*, \lambda)) \in \boldsymbol{H} \times \boldsymbol{Q}. \tag{2.2.15}$$

Now, (2.2.2) can be stated, equivalently, as: Find $(\boldsymbol{q}^\zeta, (\phi_*^\zeta, \lambda^\zeta)) \in \boldsymbol{H} \times \boldsymbol{Q}$ such that

$$\boldsymbol{A}_\zeta\Big((\boldsymbol{q}^\zeta, (\phi_*^\zeta, \lambda^\zeta)), (\boldsymbol{r}, (\psi, \mu))\Big) \;=\; \tilde{\mathbf{G}}(\boldsymbol{r}, (\psi, \mu)), \tag{2.2.16}$$

for all $(\boldsymbol{r}, (\psi, \mu)) \in \boldsymbol{H} \times \boldsymbol{Q}$, where $\boldsymbol{A}_\zeta : (\boldsymbol{H} \times \boldsymbol{Q}) \times (\boldsymbol{H} \times \boldsymbol{Q}) \to \mathbb{R}$ is the bilinear form given by

$$\boldsymbol{A}_\zeta\Big((\boldsymbol{q}, (\phi_*, \lambda)), (\boldsymbol{r}, (\psi, \mu))\Big) := \boldsymbol{A}\Big((\boldsymbol{q}, (\phi_*, \lambda)), (\boldsymbol{r}, (\psi, \mu))\Big) + c(\boldsymbol{q}, (\psi, \mu)) + d_\zeta(\lambda, \mu).$$

We will now show that the assumptions of Theorem 2.6 in [9] are satisfied. Let us define

$$S_\zeta := \sup_{\substack{(\boldsymbol{r},(\psi,\mu))\in \boldsymbol{H}\times\boldsymbol{Q} \\ (\boldsymbol{r},(\psi,\mu))\neq 0}} \frac{\boldsymbol{A}_\zeta\Big((\boldsymbol{q}, (\phi_*, \lambda)), (\boldsymbol{r}, (\psi, \mu))\Big)}{\|(\boldsymbol{r}, (\psi, \mu)\|_{\boldsymbol{H}\times\boldsymbol{Q}}}.$$

Then, given $(\widehat{\boldsymbol{r}}, (\widehat{\psi}, \widehat{\mu})) \in \boldsymbol{H} \times \boldsymbol{Q}$, and thanks to (2.2.3) and (2.2.6) we get that

$$\begin{aligned}
\frac{\boldsymbol{A}_\zeta\Big((\boldsymbol{q}, (\phi_*, \lambda)), (\widehat{\boldsymbol{r}}, (\widehat{\psi}, \widehat{\mu}))\Big)}{\|(\widehat{\boldsymbol{r}}, (\widehat{\psi}, \widehat{\mu})\|_{\boldsymbol{H}\times\boldsymbol{Q}}} &= \frac{\boldsymbol{A}\Big((\boldsymbol{q}, (\phi_*, \lambda)), (\widehat{r}, (\widehat{\psi}, \widehat{\mu}))\Big)}{\|(\widehat{\boldsymbol{r}}, (\widehat{\psi}, \widehat{\mu}))\|_{\boldsymbol{H}\times\boldsymbol{Q}}} + \frac{c(\boldsymbol{q}, (\widehat{\psi}, \widehat{\mu}))}{\|(\widehat{\boldsymbol{r}}, (\widehat{\psi}, \widehat{\mu}))\|_{\boldsymbol{H}\times\boldsymbol{Q}}} + \frac{d_\zeta(\lambda, \widehat{\mu})}{\|(\widehat{\boldsymbol{r}}, (\widehat{\psi}, \widehat{\mu}))\|_{\boldsymbol{H}\times\boldsymbol{Q}}} \\
&\geq \frac{\boldsymbol{A}\Big((\boldsymbol{q}, (\phi_*, \lambda)), (\widehat{r}, (\widehat{\psi}, \widehat{\mu}))\Big)}{\|(\widehat{\boldsymbol{r}}, (\widehat{\psi}, \widehat{\mu}))\|_{\boldsymbol{H}\times\boldsymbol{Q}}} - \|c\|\|\boldsymbol{q}\|_{\boldsymbol{H}} - \|d_\zeta\|\|\lambda\|_{1/2,00;\Gamma_N}.
\end{aligned}$$

And now taking supremum over $\boldsymbol{H} \times \boldsymbol{Q}$ we get

$$S_\zeta \geq (\alpha_A - \|c\| - \|d_\zeta\|)\, \|(\boldsymbol{q}, (\phi_*, \lambda))\|_{\boldsymbol{H}\times\boldsymbol{Q}} = \left(\alpha_A - \kappa^{-1}\|\boldsymbol{\beta}\|_{0,4;\Omega} - M_\varphi \|\boldsymbol{c_4}\|^2\right) \|(\boldsymbol{q}, (\phi_*, \lambda))\|_{\boldsymbol{H}\times\boldsymbol{Q}},$$

where we have used (2.2.3) and (2.2.6). Moreover, since $M_\varphi = |c_3|\ |\Sigma|^{1/2} + c_1\ \|\boldsymbol{c_2}\| R$, then the assumption of this lemma implies that

$$S_\zeta \geq \frac{\alpha_A}{2} \|(\boldsymbol{q}, (\phi_*, \lambda))\|_{\boldsymbol{H}\times\boldsymbol{Q}},$$

proving that the first assumption of Theorem 2.6 of [9] is satisfied.

On the other hand, we consider $(\tilde{\boldsymbol{r}}, (\tilde{\psi}, \tilde{\mu})) \in \boldsymbol{H} \times \boldsymbol{Q}$ from (2.2.15). Then, since $\boldsymbol{A}$ and d_ζ are symmetric, thanks to (2.2.4) and (2.2.14) we have

$$\begin{aligned}
\boldsymbol{A}_\zeta\Big((\tilde{\boldsymbol{r}}, (\tilde{\psi}, \tilde{\mu})), (\boldsymbol{q}, (\phi_*, \lambda))\Big) &= \boldsymbol{A}\Big((\tilde{\boldsymbol{r}}, (\tilde{\psi}, \tilde{\mu})), (\boldsymbol{q}, (\phi_*, \lambda))\Big) + c(\tilde{\boldsymbol{r}}, (\phi_*, \lambda)) + d_\zeta(\tilde{\mu}, \lambda) \\
&= \boldsymbol{A}\Big((\boldsymbol{q}, (\phi_*, \lambda)), (\tilde{\boldsymbol{r}}, (\tilde{\psi}, \tilde{\mu}))\Big) + c(\tilde{\boldsymbol{r}}, (\phi_*, \lambda)) + d_\zeta(\lambda, \tilde{\mu}) \\
&\geq (\alpha_A - \|c\| - \|d_\zeta\|)\, \|(\boldsymbol{q}, (\phi_*, \lambda))\|_{\boldsymbol{H}\times\boldsymbol{Q}}\ \|(\tilde{\boldsymbol{r}}, (\tilde{\psi}, \tilde{\mu}))\|_{\boldsymbol{H}\times\boldsymbol{Q}} \\
&\geq \frac{\alpha_A}{2} \|(\boldsymbol{q}, (\phi_*, \lambda))\|_{\boldsymbol{H}\times\boldsymbol{Q}}\ \|(\tilde{\boldsymbol{r}}, (\tilde{\psi}, \tilde{\mu}))\|_{\boldsymbol{H}\times\boldsymbol{Q}}.
\end{aligned}$$

Therefore, if $\sup_{(\boldsymbol{r},(\psi,\mu))\in\boldsymbol{H}\times\boldsymbol{Q}} \boldsymbol{A}_\zeta\Big((\boldsymbol{r}, (\psi, \mu)), (\boldsymbol{q}, (\phi_*, \lambda))\Big) = 0$, then $(\boldsymbol{q}, (\phi_*, \lambda)) = 0$, proving the second assumption of Theorem 2.6 in [9]. Thus, (2.2.16) has a unique solution and we have the *a priori* estimate

$$\|(\boldsymbol{q}^\zeta, (\phi_*^\zeta, \lambda^\zeta))\|_{\boldsymbol{H}\times\boldsymbol{Q}} \leq \frac{2}{\alpha_A} \left(\|f\|_{0,4/3,\Omega} + \|\boldsymbol{c_2}\|\ |c_2|\ |\Sigma|^{1/2}\right).$$

□

2.2.2 Solvability analysis of the fixed-point scheme

Knowing that the operator T is well-defined, we now focus on the solvability of the fixed point. Let $\zeta \in B$. Then, we have

$$\|T(\zeta)\|_{1/2,00;\Gamma_N} = \|\lambda^\zeta\|_{1/2,00;\Gamma_N} \leq \|(\boldsymbol{q}^\zeta, (\phi_*^\zeta, \lambda^\zeta))\|_{\boldsymbol{H}\times\boldsymbol{Q}} \leq \frac{2}{\alpha_A}\left(\|f\|_{0,4/3,\Omega} + \|\boldsymbol{c_2}\|\ |c_2|\ |\Sigma|^{1/2}\right).$$

If we assume that the data f and c_2 satisfy

$$\frac{2}{\alpha_A}\left(\|f\|_{0,4/3,\Omega} + \|\boldsymbol{c_2}\|\ |c_2|\ |\Sigma|^{1/2}\right) < R, \tag{2.2.17}$$

then $T(\zeta) \in B$.

On the other hand, for $i \in \{1,2\}$, let $\zeta_i \in B$ and set $T(\zeta_i) := \lambda^{\zeta_i}$, where $(\boldsymbol{q}^{\zeta_i}, (\phi_*^{\zeta_i}, \lambda^{\zeta_i})) \in \boldsymbol{H}\times\boldsymbol{Q}$ is the only solution to (2.2.2) with ζ_i as data. We have that

$$\|\lambda^{\zeta_i}\|_{1/2,00;\Gamma_N} \leq \frac{2}{\alpha_A}\left(\|f\|_{0,4/3,\Omega} + \|\boldsymbol{c_2}\|\ |c_2|\ |\Sigma|^{1/2}\right). \tag{2.2.18}$$

If we define $\zeta := \zeta_1 - \zeta_2$ and $(\boldsymbol{q}^\zeta, (\phi_*^\zeta, \lambda^\zeta)) := (\boldsymbol{q}^{\zeta_1} - \boldsymbol{q}^{\zeta_2}, (\phi_*^{\zeta_1} - \phi_*^{\zeta_2}, \lambda^{\zeta_1} - \lambda^{\zeta_2}))$, then, according to (2.2.16), for all $(\boldsymbol{r}, (\psi, \mu)) \in \boldsymbol{H}\times\boldsymbol{Q}$

$$\boldsymbol{A}_{\zeta_1}\Big((\boldsymbol{q}^{\zeta_1}, (\phi_*^{\zeta_1}, \lambda^{\zeta_1})), (\boldsymbol{r}, (\psi, \mu))\Big) - \boldsymbol{A}_{\zeta_2}\Big((\boldsymbol{q}^{\zeta_2}, (\phi_*^{\zeta_2}, \lambda^{\zeta_2})), (\boldsymbol{r}, (\psi, \mu))\Big) = 0.$$

Rewriting this expression,

$$\boldsymbol{A}\Big((\boldsymbol{q}^\zeta, (\phi_*^\zeta, \lambda^\zeta)), (\boldsymbol{r}, (\psi, \mu))\Big) + c(\boldsymbol{q}^\zeta, (\psi, \mu)) = d_{\zeta_1}(\lambda^{\zeta_1}, \mu) - d_{\zeta_2}(\lambda^{\zeta_2}, \mu),$$

i.e.,

$$\boldsymbol{A}\Big((\boldsymbol{q}^\zeta, (\phi_*^\zeta, \lambda^\zeta)), (\boldsymbol{r}, (\psi, \mu))\Big) + c(\boldsymbol{q}^\zeta, (\psi, \mu)) = \langle\varphi(\zeta_1)\lambda^\zeta, \mu\rangle_\Sigma + \langle(\varphi(\zeta_1) - \varphi(\zeta_2))\lambda^{\zeta_2}, \mu\rangle_\Sigma.$$

This implies that

$$\begin{aligned}\boldsymbol{A}_{\zeta_1}\Big((\boldsymbol{q}^\zeta,(\phi_*^\zeta,\lambda^\zeta),(\boldsymbol{r},(\psi,\mu))\Big) &= \boldsymbol{A}\Big((\boldsymbol{q}^\zeta,(\phi_*^\zeta,\lambda^\zeta),(\boldsymbol{r},(\psi,\mu))\Big) + c(\boldsymbol{q}^\zeta,(\psi,\mu)) - d_{\zeta_1}(\lambda^\zeta,\mu)\\ &= \langle(\varphi(\zeta_1)-\varphi(\zeta_2))\lambda^{\zeta_2},\mu\rangle_\Sigma\\ &\le c_1\|\zeta\|_{0,\Sigma}\|\lambda^{\zeta_2}\|_{0,4;\Sigma}\|\mu\|_{0,4;\Sigma}\\ &\le c_1\|\boldsymbol{c_2}\|\|\boldsymbol{c}_4\|^2\|\zeta\|_{1/2,00;\Gamma_N}\|\lambda^{\zeta_2}\|_{1/2,00;\Gamma_N}\|\mu\|_{1/2,00;\Gamma_N},\end{aligned}$$

where we used the definition (2.1.2). Also, by (2.2.12) and (2.2.17),

$$\boldsymbol{A}_{\zeta_1}\Big((\boldsymbol{q}^\zeta,(\phi_*^\zeta,\lambda^\zeta),(\boldsymbol{r},(\psi,\mu))\Big) \le c_1\|\boldsymbol{c_2}\|\|\boldsymbol{c}_4\|^2 R\,\|\zeta\|_{1/2,00;\Gamma_N}\|\mu\|_{1/2,00;\Gamma_N}.$$

Thus, we have

$$\begin{aligned}\frac{\alpha_A}{2}\|(\boldsymbol{q}^\zeta,(\phi_h^z,\lambda^\zeta))\| &\le \sup_{\substack{(\boldsymbol{r},(\psi,\mu))\in\boldsymbol{H}\times\boldsymbol{Q}\\(\boldsymbol{r},(\psi,\mu))\neq 0}}\frac{\boldsymbol{A}_{\zeta_1}\Big((\boldsymbol{q}^\zeta,(\phi_*^\zeta,\lambda^\zeta)),(\boldsymbol{r},(\psi,\mu))\Big)}{\|(\boldsymbol{r},(\psi,\mu)\|_{\boldsymbol{H}\times\boldsymbol{Q}}}\\ &\le c_1\|\boldsymbol{c_2}\|\|\boldsymbol{c}_4\|^2\,R\|\zeta\|_{1/2,00;\Gamma_N}.\end{aligned}$$

Thus,

$$\|\lambda^\zeta\|_{1/2,00;\Gamma_N} \le L_T\|\zeta\|_{1/2,00;\Gamma_N}$$

with

$$L_T := \frac{2}{\alpha_A}c_1\|\boldsymbol{c_2}\|\|\boldsymbol{c}_4\|^2\,R. \tag{2.2.19}$$

In other words, we have proved the following result, thanks to Banach fixed-point theorem.

Theorem 2.2.3. *If the data f and c_2 satisfy*

$$\frac{2}{\alpha_A}\left(\|f\|_{0,4/3,\Omega} + \|\boldsymbol{c_2}\|\ |c_2|\ |\Sigma|^{1/2}\right) < R \tag{2.2.20}$$

and the data c_1 is small enough such that $L_T < 1$, then (2.2.16) has a unique fixed-point.

Remark 1. In reverse osmosis models, $f = 0$ and c_0, c_1 and c_2 are small parameters according to Table 6.2 of [1].

CHAPTER 3

An HDG Scheme and well-posedness

3.1 HDG formulation

Let $\mathcal{T}_h$ a simplicial shape-regular triangulation of Ω of meshsize h, with mesh-regularity parameter ϱ. For a simplex K, we denote its diameter h_K and outward unit normal by $\boldsymbol{n}_K$, writing $\boldsymbol{n}$ instead of $\boldsymbol{n}_K$ when there is no confusion. Similarly, for a facet e, we denote be h_e its diameter and write $\boldsymbol{n}$ instead of $\boldsymbol{n}_e$ to refer to its normal vector. We also consider, by simplicity, that the triangulation does not have hanging nodes. The set of facets and boundary facets of $\mathcal{T}_h$ are denoted by $\mathcal{E}_h$ and $\mathcal{E}_h^\partial$, respectively.

For each scalar-valued function η and ζ, we define

$$(\eta,\zeta)_{\mathcal{T}_h} := \sum_{K\in\mathcal{T}_h} (\eta,\zeta)_K \quad \text{and} \quad \langle\eta,\zeta\rangle_{\partial\mathcal{T}_h} := \sum_{K\in\mathcal{T}_h} \langle\eta,\zeta\rangle_{\partial K}\,.$$

Vector-valued functions are boldfaced and, for $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$, we write

$$(\boldsymbol{\eta},\boldsymbol{\zeta})_{\mathcal{T}_h} := \sum_{i=1}^{d}(\eta_i,\zeta_i)_{\mathcal{T}_h} \quad \text{and} \quad \langle\boldsymbol{\eta},\boldsymbol{\zeta}\rangle_{\partial\mathcal{T}_h} := \sum_{i=1}^{d}\langle\eta_i,\zeta_i\rangle_{\partial\mathcal{T}_h}\,.$$

These inner products defined on the mesh induce the norms

$$\|\cdot\|_{\mathcal{T}_h} := \left(\sum_{K\in\mathcal{T}_h}\|\cdot\|_K^2\right)^{1/2}, \quad \|\cdot\|_{\partial\mathcal{T}_h} := \left(\sum_{K\in\mathcal{T}_h}\|\cdot\|_{\partial K}^2\right)^{1/2}.$$

The problem (2.1.3), can be written as

$$\begin{cases} \kappa^{-1}\boldsymbol{q}+\nabla\phi_* &= 0 & \text{in } \Omega,\\ \nabla\cdot(\boldsymbol{q}+\boldsymbol{\beta}\phi_*) &= f & \text{in } \Omega,\\ \phi_* &= 0 & \text{in } \Gamma_{\text{in}},\\ (\boldsymbol{q}+\boldsymbol{\beta}\phi_*)\cdot\boldsymbol{n}-\varphi(\phi_*)\phi_* &= \boldsymbol{\beta}\cdot\boldsymbol{n}\phi_*+g(\phi_{\text{in}}) & \text{in } \Gamma_N. \end{cases} \tag{3.1.1}$$

The approximation of the solution will be looked in the following finite dimensional spaces:

$$\boldsymbol{H}_h := \left\{\boldsymbol{v}\in[L^2(\mathcal{T}_h)]^2 \,:\, v\Big|_K\in[\mathbb{P}_k(K)]^d\,,\ \forall\, K\in\mathcal{T}_h\right\}, \tag{3.1.2a}$$

$$Q_h := \left\{w\in L^2(\mathcal{T}_h) \,:\, w\Big|_K\in\mathbb{P}_k(K)\,,\ \forall\, K\in\mathcal{T}_h\right\}, \tag{3.1.2b}$$

$$\mathrm{M}_h := \left\{\mu\in L^2(\mathcal{E}_h) \,:\, \mu\Big|_e\in\mathbb{P}_k(e)\,,\ \forall\, e\in\mathcal{E}_h\right\}. \tag{3.1.2c}$$

The HDG scheme reads: Find $(\boldsymbol{q}_h,\phi_h,\widehat{\phi}_h)\in\boldsymbol{H}_h\times Q_h\times M_h$, an approximation of $(\boldsymbol{q},\phi_*,\phi_*|_{\mathcal{E}_h})$, such

$$(\kappa^{-1}\boldsymbol{q}_h,\boldsymbol{r}_h)_{\mathcal{T}_h}-(\phi_h,\nabla\cdot\boldsymbol{r}_h)_{\mathcal{T}_h}+\langle\boldsymbol{r}_h\cdot\boldsymbol{n},\widehat{\phi}_h\rangle_{\partial\mathcal{T}_h}=0, \tag{3.1.3a}$$

$$-(\boldsymbol{q}_h+\boldsymbol{\beta}\phi_h,\nabla w_h)_{\mathcal{T}_h}+\langle\widehat{\boldsymbol{q}_h+\boldsymbol{\beta}\phi_h}\cdot\boldsymbol{n},w_h\rangle_{\partial\mathcal{T}_h}=(f,w_h)_{\mathcal{T}_h}, \tag{3.1.3b}$$

$$\langle\widehat{\boldsymbol{q}_h+\boldsymbol{\beta}\phi_h}\cdot\boldsymbol{n},\mu_h\rangle_{\partial\mathcal{T}_h\setminus\Gamma_{\text{in}}}=\langle\varphi(\phi_*)\phi_*+\boldsymbol{\beta}\cdot\boldsymbol{n}\widehat{\phi}_h+g(\phi_{\text{in}}),\mu_h\rangle_{\Gamma_N}, \tag{3.1.3c}$$

$$\langle\widehat{\phi}_h,\mu_h\rangle_{\Gamma_{\text{in}}}=0, \tag{3.1.3d}$$

for all $(\boldsymbol{r}_h, w_h, \mu_h) \in \boldsymbol{H}_h \times Q_h \times \mathrm{M}_h$, where the numerical trace for the total flux is given by

$$\widehat{\boldsymbol{q}_h + \boldsymbol{\beta}\phi_h} := \boldsymbol{q}_h + \boldsymbol{\beta}\widehat{\phi}_h + \tau(\phi_h - \widehat{\phi}_h)\boldsymbol{n} \quad \text{on } \partial\mathcal{T}_h, \tag{3.1.3e}$$

and τ is a stabilization parameter such that it satisfies (S_1) and (S_2). We notice that, since ϕ_{in} is constant, if the HDG scheme has an unique solution, then $(\boldsymbol{q}_h, \phi_h + \phi_{\text{in}}, \widehat{\phi}_h + \phi_{\text{in}})$ will be the unique solution for the problem with non zero Dirichlet condition. In particular, we set $\tau_\beta := \tau - \frac{1}{2}\boldsymbol{\beta}\cdot\boldsymbol{n}$ and assume that

(S_1) there exists a constant $\gamma_0 > 0$ such that $|\tau_\beta|_{\partial\mathcal{T}_h} > \gamma_0$,

(S_2) τ is constant on each facet.

3.2 Fixed-point scheme

We begin by rewriting (3.1.3) as an equivalent fixed point equation. We define $T_h : \mathrm{M}_h \to \mathrm{M}_h$ as the operator defined for each $z \in \mathrm{M}_h$ as

$$T_h(z) := \widehat{\phi}_h^z \tag{3.2.1}$$

where $(\boldsymbol{q}_h^z, \phi_h^z, \widehat{\phi}_h^z) \in \boldsymbol{H}_h \times Q_h \times \mathrm{M}_h$ is the unique solution (to be confirmed later in Section 3.5) of the following linearized problem:

$$(\kappa^{-1}\boldsymbol{q}_h^z, \boldsymbol{r}_h)_{\mathcal{T}_h} - (\phi_h^z, \nabla\cdot\boldsymbol{r}_h)_{\mathcal{T}_h} + \langle \boldsymbol{r}_h\cdot\boldsymbol{n}, \widehat{\phi}_h^z\rangle_{\partial\mathcal{T}_h} = 0, \tag{3.2.2a}$$

$$-(\boldsymbol{q}_h^z + \boldsymbol{\beta}\phi_h^z, \nabla w_h)_{\mathcal{T}_h} + \langle \widehat{\boldsymbol{q}_h^z + \boldsymbol{\beta}\phi_h^z}\cdot\boldsymbol{n}, w_h\rangle_{\partial\mathcal{T}_h} = (f, w_h)_{\mathcal{T}_h}, \tag{3.2.2b}$$

$$\langle \widehat{\boldsymbol{q}_h^z + \boldsymbol{\beta}\phi_h^z}\cdot\boldsymbol{n}, \mu_h\rangle_{\partial\mathcal{T}_h\setminus\Gamma_{\text{in}}} = \langle \varphi(z)\widehat{\phi}_h^z + \boldsymbol{\beta}\widehat{\phi}_h^z\cdot\boldsymbol{n} + g(\phi_{\text{in}}), \mu_h\rangle_{\Gamma_N}, \tag{3.2.2c}$$

$$\langle \widehat{\phi}_h^z, \mu_h\rangle_{\Gamma_{\text{in}}} = 0, \tag{3.2.2d}$$

for all $(\boldsymbol{r}_h, w_h, \mu_h) \in \boldsymbol{H}_h \times Q_h \times \mathrm{M}_h$.

We observe that solving (3.1.3) is equivalent to seeking a fixed point of T_h, that is: Find $\lambda \in M_h^\Sigma$ such that $T_h(\lambda) = \lambda$, where M_h^Σ is the restriction of M_h to Σ.

3.3 Preliminaries

Norm equivalences and Sobolev inequality. During the stability analysis, we will be using the following norms and results: Let A be a subset of $\mathcal{E}_h$, for a scalar-valued function η, we define

$$\|\eta\|_{\infty,A} \coloneqq \max_{e\in A}\|\eta\|_{\infty,e}.$$

We will also use the following results: Given $v \in \mathrm{M}_h$ and $e \in \mathcal{E}_h$, there exists a constant $C_{eq} > 0$, independent of h, such that

$$h_e^{-1/2}\|v\|_{0,e} \leq \|v\|_{\infty,e} \leq C_{eq}\ h_e^{-\frac{1}{2}}\|v\|_{0,e}. \tag{3.3.1}$$

Let $v \in Q_h$, $K \in \mathcal{T}_h$, and e an edge of K. We know that [7, Lemma 1.46]

$$h_K^{1/2}\|v\|_{0,e} \leq C_{tr}^e\|v\|_{0,K}, \tag{3.3.2}$$

where C_{tr}^e depends only on ϱ and k.

We now recall the Sobolev's inequality [2, Lemma 4.34]. Let $\widehat{\Omega}$ a domain of diameter $\hat{d}$ and is star-shaped with respect to a ball B. If v is in $W^{m,p}(\Omega)$ where either (i) $1 < p < \infty$ and $m > 2/p$, or (ii) $p = 1$ and $m \geq 2$, then v can be identified by a continuous function in Ω and there exists a positive Sobolev constant C_{sob}, depending on the diameter $\hat{d}$ and chunkiness parameter (cf. [2, Definition 4.2.16])

$$\|v\|_{\infty,\widehat{\Omega}} \leq C_{sob}\|v\|_{m,p,\widehat{\Omega}}. \tag{3.3.3}$$

Projections. We denote by P_M the L^2-projection from $L^2(\mathcal{E}_h)$ into M_h. Let $e \in \mathcal{E}_h$ and $K \in \mathcal{T}$ an element with e as an edge. For all $s \in \{0, ..., k+1\}$ and all $v \in H^s(K)$, there exists a positive constant C_1, independent of the meshsize, such that

$$\|v - P_M v\|_e \leq C_1 h_K^{s-1/2}|v|_{s,K}. \tag{3.3.4}$$

Moreover, for all $\mu \in H^s(e)$, there exists $C_2 > 0$, independent of the meshsize, such that

$$\|\mu - P_M\mu\|_e \leq Ch_e^s|\mu|_{s,e}. \tag{3.3.5}$$

We refer to [7, Lemmas 1.58 and 1.59] for more details.

We now recall the HDG projections suitable for our problem (cf. [5]). On any simplex K, the projection $\Pi_h(\boldsymbol{q},\phi) := (\Pi_V\boldsymbol{q}, \Pi_W\phi)$ is the element of $\mathbf{P}_k(K) \times P_k(K)$ which solves the equations

$$\Big((\Pi_V\boldsymbol{q} - \boldsymbol{q}) - \boldsymbol{\beta}(\Pi_W\phi - \phi), \boldsymbol{r}\Big)_K = 0 \qquad \forall \boldsymbol{r} \in [\mathbb{P}_{k-1}(K)]^2, \tag{3.3.6a}$$

$$\Big(\Pi_W\phi - \phi, w\Big)_K = 0 \qquad \forall w \in \mathbb{P}_{k-1}(K), \tag{3.3.6b}$$

$$\Big\langle\Big((\Pi_V\boldsymbol{q} - \boldsymbol{q}) + \boldsymbol{\beta}(P_M\phi - \phi)\Big)\cdot\boldsymbol{n} + \tau(\Pi_W\phi - \phi), \mu\Big\rangle_e = 0 \qquad \forall \mu \in \mathbb{P}_k(e), \tag{3.3.6c}$$

for all faces e of the simplex K.

According to Theorem 2.1 in [5], if Assumption (S_1) holds true, the system (3.3.6) is uniquely solvable for $\Pi_V\boldsymbol{q}$ and $\Pi_W\phi$. Moreover, we have the following approximation properties:

$$\|\Pi_W\phi - \phi\|_{0,K} \leq C^1_{W,K}h_K^{\ell_\phi+1}|\phi|_{\ell_\phi+1,K} + C^2_{W,K}h_K^{\ell_{\boldsymbol{q}}+1}|\nabla\cdot\boldsymbol{q}|_{\ell_{\boldsymbol{q}},K}, \tag{3.3.7a}$$

$$\|\Pi_V\boldsymbol{q} - \boldsymbol{q}\|_{0,K} \leq C^1_{V,K}h_K^{\ell_{\boldsymbol{q}}+1}|\boldsymbol{q}|_{\ell_{\boldsymbol{q}}+1,K} + C^2_{V,K}h_K^{\ell_\phi+1}|\phi|_{\ell_\phi+1,K}. \tag{3.3.7b}$$

for $\ell_\phi, \ell_{\boldsymbol{q}} \in [0,k]$. Here, as shown in [5], if h is sufficiently small, the values $C^1_{W,K}$, $C^2_{W,K}$, $C^1_{V,K}$ and $C^2_{V,K}$ can be bounded independently if the meshsize. More precisely, there exists a constant $C > 0$ depending on the polynomial degree and the shape-regularity constant, and

$$\begin{aligned}
C^1_{W,K} &\leq C\frac{(\tau_K^{\max} + |\boldsymbol{\beta}|_{1,\infty,K})}{\gamma_0},\\
C^2_{W,K} &\leq \frac{C}{\gamma_0},\\
C^1_{V,K} &\leq C\left(1 + \frac{|\boldsymbol{\beta}|_{1,\infty,K}}{\gamma_0}\right),\\
C^2_{V,K} &\leq C\left(\tau_K^* + \frac{|\boldsymbol{\beta}|_{1,\infty,K}(\tau_K^{\max} + |\boldsymbol{\beta}|_{1,\infty,K})}{\gamma_0}\right),
\end{aligned}$$

where

$$\tau_K^{\max} := \max_{e\in\partial K} |\tau|_{\partial K}|,$$
$$\tau_K^{*} := \max_{e\in\partial K} |\tau|_{\partial K\setminus e^*}|$$

and e^* is any face of K at which $|\tau|_{\partial K}|$ attains its maximum.

In order to use a duality argument, we need to introduce an auxiliary projection, the HDG projection, associated to Π_h^* as defined in [5]. On any simplex $K \in \mathcal{T}_h$, $\Pi_h^*(\boldsymbol{Q}, \Psi) = (\Pi_V^*\boldsymbol{Q}, \Pi_W^*\Psi)$ is the element of $\mathbf{P}_k(K) \times P_k(K)$ determined by requiring that

$$\Big((\Pi_V^*\boldsymbol{Q} - \boldsymbol{Q}) - (\Pi_W^*\Psi - \Psi)\boldsymbol{\beta}, \boldsymbol{r}\Big)_K = 0 \qquad \forall \boldsymbol{r} \in [\mathbb{P}_{k-1}(K)]^2 \tag{3.3.8a}$$
$$\Big(\Pi_W^*\Psi - \Psi, w\Big)_K = 0 \qquad \forall w \in \mathbb{P}_{k-1}(K) \tag{3.3.8b}$$
$$\Big\langle(\Pi_V^*\boldsymbol{Q} - \boldsymbol{Q})\cdot\boldsymbol{n} + (\tau - \boldsymbol{\beta}\cdot\boldsymbol{n})(\Pi_W^*\Psi - \Psi), \mu\Big\rangle_e = 0 \qquad \forall \mu \in \mathbb{P}_k(e) \tag{3.3.8c}$$

for all faces e of K.

According to [5, Theorem A.1], if $k \geq 0$ and assumption (S_1) hold true, the projection Π_h^* is well-posed and satisfies the following approximation properties:

$$\|\Pi_W^*\Psi - \Psi\|_{0,K} \leq C_{W,K}^{1,*} h_K^{\ell_\Psi+1} |\Psi|_{\ell_\Psi+1,K} + C_{W,K}^{2,*} h_K^{\ell_Q+1} |\nabla\cdot\boldsymbol{Q}|_{\ell_Q,K}, \tag{3.3.9a}$$
$$\|\Pi_V^*\boldsymbol{Q} - \boldsymbol{Q}\|_{0,K} \leq C_{V,K}^{1,*} h_K^{\ell_Q+1} |\boldsymbol{Q}|_{\ell_Q+1,K} + C_{V,K}^{2,*} h_K^{\ell_\Psi+1} |\Psi|_{\ell_\Psi+1,K}, \tag{3.3.9b}$$

for $\ell_\Psi, \ell_Q \in [0, k]$. Here, as shown in [5], if h is sufficiently small, the values $C_{W,K}^{1,*}, C_{W,K}^{2,*}, C_{V,K}^{1,*}$ and $C_{V,K}^{2,*}$ can be bounded independently of the meshsize. More precisely, there exists a constant $C > 0$ depending on the polynomial degree and the shape-regularity constant, and

$$C_{W,K}^{1,*} \leq C\frac{((\tau\boldsymbol{\beta})_K^{\max} + |\boldsymbol{\beta}|_{1,\infty,K})}{\gamma_0},$$
$$C_{W,K}^{2,*} \leq \frac{C}{\gamma_0},$$
$$C_{V,K}^{1,*} \leq C\left(1 + \frac{|\boldsymbol{\beta}|_{1,\infty,K}}{\gamma_0}\right),$$

$$C_{V,K}^{2,*} \leq C\left((\tau\boldsymbol{\beta})_K^* + \frac{|\boldsymbol{\beta}|_{1,\infty,K}\left((\tau\boldsymbol{\beta})_K^{\max} + |\boldsymbol{\beta}|_{1,\infty,K}\right)}{\gamma_0}\right),$$

where

$$(\tau\boldsymbol{\beta})_K^{\max} := \max|(\tau - \boldsymbol{\beta}\cdot\boldsymbol{n})|_{\partial K}|,$$
$$(\tau\boldsymbol{\beta})_K^{*} := \max|(\tau - \boldsymbol{\beta}\cdot\boldsymbol{n})|_{\partial K\setminus e^*}|,$$

and e^* is the face of K at which $|(\tau - \boldsymbol{\beta}\cdot\boldsymbol{n})|_{\partial K}|$ attains its maximum.

3.4 Stability estimates

We are interested in showing the stability estimates associated with (3.4.1). Now, as we will see in the next sections, to show the contraction property of T_h and also to obtain the error bounds, the same stability estimates will be required, but they will be associated to HDG schemes as the one in (3.4.1) with different right-hand sides. This is the reason why we will analyze the following more general HDG scheme. Given $z \in \mathrm{M}_h$, $\boldsymbol{I} \in \boldsymbol{L}^2(\Omega)$ and $\Lambda \in L^2(\Gamma_N)$ such that $\Lambda|_{\Gamma_{\text{out}}} = 0$, find $(\boldsymbol{q}_h^z, \phi_h^z, \widehat{\phi}_h^z) \in \boldsymbol{H}_h \times Q_h \times \mathrm{M}_h$ such that

$$(\kappa^{-1}\boldsymbol{q}_h^z, \boldsymbol{r}_h)_{\mathcal{T}_h} - (\phi_h^z, \nabla\cdot\boldsymbol{r}_h)_{\mathcal{T}_h} + \langle \boldsymbol{r}_h\cdot\boldsymbol{n}, \widehat{\phi}_h^z\rangle_{\partial\mathcal{T}_h} = (\kappa^{-1}\boldsymbol{I}, \boldsymbol{r}_h)_{\mathcal{T}_h}, \tag{3.4.1a}$$
$$-(\boldsymbol{q}_h^z + \boldsymbol{\beta}\phi_h^z, \nabla w_h)_{\mathcal{T}_h} + \langle \widehat{\boldsymbol{q}_h^z + \boldsymbol{\beta}\phi_h^z}\cdot\boldsymbol{n}, w_h\rangle_{\partial\mathcal{T}_h} = (f, w_h)_{\mathcal{T}_h}, \tag{3.4.1b}$$
$$\langle \widehat{\boldsymbol{q}_h^z + \boldsymbol{\beta}\phi_h^z}\cdot\boldsymbol{n}, \mu_h\rangle_{\partial\mathcal{T}_h\setminus\Gamma_{\text{in}}} = \langle \varphi(z)\widehat{\phi}_h^z + \boldsymbol{\beta}\widehat{\phi}_h^z\cdot\boldsymbol{n} + \Lambda, \mu_h\rangle_{\Gamma_N}, \tag{3.4.1c}$$
$$\langle \widehat{\phi}_h^z, \mu_h\rangle_{\Gamma_{\text{in}}} = 0, \tag{3.4.1d}$$

for all $(\boldsymbol{r}_h, w_h, \mu_h) \in \boldsymbol{H}_h \times Q_h \times \mathrm{M}_h$. Here, addition, $\boldsymbol{I}$ will be zero when we prove well-posedness and will be orthogonal to polynomials of degree less than or equal to $k-1$ when we derive the error estimates.

Let us define

$$E := \left(\|\kappa^{-1/2}\boldsymbol{q}_h^z\|_{\mathcal{T}_h}^2 + \|\tau_\beta^{1/2}(\phi_h^z - \widehat{\phi}_h^z)\|_{\partial\mathcal{T}_h}^2\right)^{1/2}. \tag{3.4.2}$$

3.4.1 Energy estimate.

Lemma 3.4.1. *Suppose that assumption (S_2) is satisfied. Then we have*

$$E^2 + \frac{1}{2}\||\boldsymbol{\beta}\cdot\boldsymbol{n}|^{1/2}\ \widehat{\phi}_h^z\|_{\Gamma_N}^2 = -\langle\varphi(z)\widehat{\phi}_h^z,\widehat{\phi}_h^z\rangle_\Sigma - \langle\Lambda,\widehat{\phi}_h^z\rangle_\Sigma + (f,\phi_h^z)_{\mathcal{T}_h} + (\kappa^{-1}\boldsymbol{I},\boldsymbol{q}_h^z)_{\mathcal{T}_h}. \tag{3.4.3}$$

Proof. From the first equation of the HDG scheme (3.4.1a) with $\boldsymbol{r}_h = \boldsymbol{q}_h^z$ and integration by parts, we get that

$$\|\kappa^{-1/2}\boldsymbol{q}_h^z\|_{\mathcal{T}_h}^2 + (\boldsymbol{q}_h^z,\nabla\phi_h^z)_{\mathcal{T}_h} - \langle\boldsymbol{q}_h^z\cdot\boldsymbol{n},\phi_h^z\rangle_{\partial\mathcal{T}_h} + \langle\boldsymbol{q}_h^z\cdot\boldsymbol{n},\widehat{\phi}_h^z\rangle_{\partial\mathcal{T}_h} = (\kappa^{-1}\boldsymbol{I},\boldsymbol{q}_h^z)_{\mathcal{T}_h}.$$

From the second equation of the HDG scheme (3.4.1b) with $w_h := \phi_h^z$ we get that

$$-(\boldsymbol{q}_h^z,\nabla\phi_h^z)_{\mathcal{T}_h} - (\boldsymbol{\beta}\ \phi_h^z,\nabla\phi_h^z)_{\mathcal{T}_h} + \langle\boldsymbol{q}_h^z\cdot\boldsymbol{n},\phi_h^z\rangle_{\partial\mathcal{T}_h} + \langle\boldsymbol{\beta}\cdot\boldsymbol{n}\ \widehat{\phi}_h^z,\phi_h^z\rangle_{\partial\mathcal{T}_h} + \langle\tau(\phi_h^z-\widehat{\phi}_h^z),\phi_h^z\rangle_{\partial\mathcal{T}_h} = (f,\phi_h^z)_{\mathcal{T}_h}.$$

By adding the last two equalities we get that

$$\begin{aligned}&\|\kappa^{-1/2}\boldsymbol{q}_h^z\|_{\mathcal{T}_h}^2 - (\boldsymbol{\beta}\ \phi_h^z,\nabla\phi_h^z)_{\mathcal{T}_h} + \langle\boldsymbol{\beta}\cdot\boldsymbol{n}\ \widehat{\phi}_h^z,\phi_h^z-\widehat{\phi}_h^z\rangle_{\partial\mathcal{T}_h}\\ &\quad + \langle\tau(\phi_h^z-\widehat{\phi}_h^z),\phi_h^z-\widehat{\phi}_h^z\rangle_{\partial\mathcal{T}_h} + \langle\widehat{\boldsymbol{q}_h^z+\boldsymbol{\beta}\phi_h^z}\cdot\boldsymbol{n},\widehat{\phi}_h^z\rangle_{\partial\mathcal{T}_h} = (f,\phi_h^z)_{\mathcal{T}_h} + (\kappa^{-1}\boldsymbol{I},\boldsymbol{q}_h^z)_{\mathcal{T}_h}.\end{aligned}$$

From the third equation of the HDG scheme (3.4.1c) with $\mu_h := \widehat{\phi}_h^z$ we get that

$$\langle\widehat{\boldsymbol{q}_h^z+\boldsymbol{\beta}\phi_h^z}\cdot\boldsymbol{n},\widehat{\phi}_h^z\rangle_{\partial\mathcal{T}_h} = \langle\varphi(z)\widehat{\phi}_h^z + \boldsymbol{\beta}\cdot\boldsymbol{n}\ \widehat{\phi}_h^z + \Lambda,\widehat{\phi}_h^z\rangle_{\Gamma_N} + \langle\widehat{\boldsymbol{q}_h^z+\boldsymbol{\beta}\phi_h^z}\cdot\boldsymbol{n},\widehat{\phi}_h^z\rangle_{\Gamma_{in}}$$

and from the fourth equation of the HDG scheme (3.4.1d), we obtain that

$$\langle\widehat{\phi}_h^z,\widehat{\boldsymbol{q}_h^z+\boldsymbol{\beta}\phi_h^z}\cdot\boldsymbol{n}\rangle_{\Gamma_{in}} = 0.$$

Then,

$$\begin{aligned}&\|\kappa^{-1/2}\boldsymbol{q}_h^z\|_{\mathcal{T}_h}^2 - (\boldsymbol{\beta}\ \phi_h^z,\nabla\phi_h^z)_{\mathcal{T}_h} + \langle\boldsymbol{\beta}\cdot\boldsymbol{n}\ \widehat{\phi}_h^z,\phi_h^z-\widehat{\phi}_h^z\rangle_{\partial\mathcal{T}_h}\\ &\quad + \langle\tau(\phi_h^z-\widehat{\phi}_h^z),\phi_h^z-\widehat{\phi}_h^z\rangle_{\partial\mathcal{T}_h} + \langle\varphi(z)\widehat{\phi}_h^z,\widehat{\phi}_h^z\rangle_\Sigma + \langle\boldsymbol{\beta}\cdot\boldsymbol{n}\ \widehat{\phi}_h^z,\widehat{\phi}_h^z\rangle_{\Gamma_N}\end{aligned}$$

$$+ \langle \Lambda, \widehat{\phi}_h^z \rangle_\Sigma = (f, \phi_h^z)_{\mathcal{T}_h} + (\kappa^{-1}\boldsymbol{I}, \boldsymbol{q}_h^z)_{\mathcal{T}_h}.$$

By integration by parts and some simple algebraic manipulation, we get that

$$(\boldsymbol{\beta}\ \phi_h^z, \nabla\phi_h^z)_{\mathcal{T}_h} = \frac{1}{2}\langle \boldsymbol{\beta}\cdot\boldsymbol{n}\ (\phi_h^z - \widehat{\phi}_h^z), \phi_h^z - \widehat{\phi}_h^z\rangle_{\partial\mathcal{T}_h} + \frac{1}{2}\langle \boldsymbol{\beta}\cdot\boldsymbol{n}\ \widehat{\phi}_h^z, \phi_h^z\rangle_{\partial\mathcal{T}_h} + \frac{1}{2}\langle \boldsymbol{\beta}\cdot\boldsymbol{n}\ \widehat{\phi}_h^z, \phi_h^z - \widehat{\phi}_h^z\rangle_{\partial\mathcal{T}_h}$$

and we have that

$$\begin{aligned}\|\kappa^{-1/2}\boldsymbol{q}_h^z\|_{\mathcal{T}_h}^2 - \frac{1}{2}\langle \boldsymbol{\beta}\cdot\boldsymbol{n}\ \widehat{\phi}_h^z, \widehat{\phi}_h^z\rangle_{\partial\mathcal{T}_h} + \langle (\tau - \frac{1}{2}\boldsymbol{\beta}\cdot\boldsymbol{n})\ (\phi_h^z - \widehat{\phi}_h^z), \phi_h^z - \widehat{\phi}_h^z\rangle_{\partial\mathcal{T}_h} + \langle \Lambda, \widehat{\phi}_h^z\rangle_\Sigma \\ + \langle \varphi(z)\widehat{\phi}_h^z, \widehat{\phi}_h^z\rangle_\Sigma + \langle \boldsymbol{\beta}\cdot\boldsymbol{n}\ \widehat{\phi}_h^z, \widehat{\phi}_h^z\rangle_{\Gamma_N} = (f, \phi_h^z)_{\mathcal{T}_h} + (\kappa^{-1}\boldsymbol{I}, \boldsymbol{q}_h^z)_{\mathcal{T}_h}.\end{aligned}$$

Since $\nabla\cdot\boldsymbol{\beta} = 0, \boldsymbol{\beta} \in \boldsymbol{H}(\mathrm{div};\Omega)$. By the characterization of $\boldsymbol{H}(\mathrm{div};\Omega)$ on a mesh $\mathcal{T}_h$, the normal component of $\boldsymbol{\beta}$ is well defined on each internal face, $\partial\mathcal{T}_h \setminus \partial\Omega$, and in our case it vanishes. Hence, we obtain

$$\langle \boldsymbol{\beta}\cdot\boldsymbol{n}\ \widehat{\phi}_h^z, \widehat{\phi}_h^z\rangle_{\partial\mathcal{T}_h} = \langle \boldsymbol{\beta}\cdot\boldsymbol{n}\ \widehat{\phi}_h^z, \widehat{\phi}_h^z\rangle_{\Gamma_{in}} + \langle \boldsymbol{\beta}\cdot\boldsymbol{n}\ \widehat{\phi}_h^z, \widehat{\phi}_h^z\rangle_{\Gamma_N} = \langle \boldsymbol{\beta}\cdot\boldsymbol{n}\ \widehat{\phi}_h^z, \widehat{\phi}_h^z\rangle_{\Gamma_N},$$

which implies that

$$\begin{aligned}\|\kappa^{-1/2}\boldsymbol{q}_h^z\|_{\mathcal{T}_h}^2 + \|\tau_\beta^{1/2}(\phi_h^z - \widehat{\phi}_h^z)\|_{\partial\mathcal{T}_h}^2 + \frac{1}{2}\langle \boldsymbol{\beta}\cdot\boldsymbol{n}\ \widehat{\phi}_h^z, \widehat{\phi}_h^z\rangle_{\Gamma_N} + \langle \Lambda, \widehat{\phi}_h^z\rangle_\Sigma \\ + \langle \varphi(z)\widehat{\phi}_h^z, \widehat{\phi}_h^z\rangle_\Sigma = (f, \phi_h^z)_{\mathcal{T}_h} + (\kappa^{-1}\boldsymbol{I}, \boldsymbol{q}_h^z)_{\mathcal{T}_h}.\end{aligned}$$

□

As we observe, we will need to bound $\|\phi_h^z\|_{\mathcal{T}_h}$ and $\|\widehat{\phi}_h^z\|_\Sigma$. To that end, as usual, we will proceed by a duality argument.

3.4.2 Duality argument

Given $\Theta \in L^2(\Omega)$ and $z \in M_h$, we define $(\boldsymbol{Q}, \Psi)$ to be the solution of the auxliary problem

$$\begin{cases} \kappa^{-1}\boldsymbol{Q} + \nabla\Psi &= 0 & \text{in } \Omega, \\ \nabla\cdot(\boldsymbol{Q} - \boldsymbol{\beta}\Psi) &= \Theta & \text{in } \Omega, \\ \Psi &= 0, & \text{in } \Gamma_{\text{in}}, \\ (\boldsymbol{Q} - \boldsymbol{\beta}\Psi)\cdot\boldsymbol{n} &= \varphi(z)\Psi & \text{in } \Gamma_N. \end{cases} \tag{3.4.4}$$

We assume that the solution of this problem (3.4.4) satisfies the following regularity estimate: There exists a constant C_{reg} such that

$$\|\boldsymbol{Q}\|_{1,\Omega} + \|\Psi\|_{2,\Omega} \le C_{reg}\|\Theta\|_{0,\Omega}. \tag{3.4.5}$$

This holds [11], for example, when Ω is a convex polyhedral domain.

Lemma 3.4.2. *For any $\Theta \in L^2(\Omega)$ we have $(\phi_h^z, \Theta)_{\mathcal{T}_h} = \sum_{i=1}^{5} T_i$, where*

$$\begin{aligned} T_1 &:= (\kappa^{-1}\boldsymbol{q}_h^z, \Pi_V^*\boldsymbol{Q} - \boldsymbol{Q})_{\mathcal{T}_h}, & T_2 &:= (f, \Pi_W^*\Psi)_{\mathcal{T}_h}, \\ T_3 &:= -\langle\varphi(z)\widehat{\phi}_h^z, P_M\Psi - \Psi\rangle_\Sigma, & T_4 &:= -\langle\Lambda, P_M\Psi\rangle_\Sigma, \\ T_5 &:= -(\kappa^{-1}\boldsymbol{I}, \Pi_V^*\boldsymbol{Q} - \boldsymbol{Q})_{\mathcal{T}_h} + (\boldsymbol{I}, \nabla\Psi - \nabla\Psi_h)_{\mathcal{T}_h}, \end{aligned}$$

for all $\Psi_h \in Q_h$.

Proof. By the second equation of the dual problem (3.4.4), integration by parts, and the first equation defining Π_h^* (3.3.8a) with $r := \nabla\phi_h^z$, we have

$$\begin{aligned} (\phi_h^z, \Theta)_{\mathcal{T}_h} &= (\phi_h^z, \nabla\cdot(\boldsymbol{Q} - \boldsymbol{\beta}\Psi))_{\mathcal{T}_h} \\ &= -(\nabla\phi_h^z, \boldsymbol{Q} - \boldsymbol{\beta}\Psi)_{\mathcal{T}_h} + \langle\phi_h^z, (\boldsymbol{Q} - \boldsymbol{\beta}\Psi)\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} \\ &= -(\nabla\phi_h^z, \Pi_V^*\boldsymbol{Q} - \boldsymbol{\beta}\Pi_W^*\Psi)_{\mathcal{T}_h} + \langle\phi_h^z, (\boldsymbol{Q} - \boldsymbol{\beta}\Psi)\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h}. \end{aligned}$$

Integrating by parts the first term of the right-hand side and using that $\nabla \cdot \boldsymbol{\beta} = 0$ we obtain

$$(\phi_h^z, \Theta)_{\mathcal{T}_h} = (\phi_h^z, \nabla \cdot \Pi_V^* \boldsymbol{Q})_{\mathcal{T}_h} - (\phi_h^z, \boldsymbol{\beta} \cdot \nabla(\Pi_W^* \Psi))_{\mathcal{T}_h} - \langle \phi_h^z, (\Pi_V^* \boldsymbol{Q} - \boldsymbol{\beta} \Pi_W^* \Psi) \cdot \boldsymbol{n} \rangle_{\partial \mathcal{T}_h} + \langle \phi_h^z, (\boldsymbol{Q} - \boldsymbol{\beta} \Psi) \cdot \boldsymbol{n} \rangle_{\partial \mathcal{T}_h}.$$

Now by the first equation of the HDG scheme (3.4.1a) with $\boldsymbol{r}_h := \Pi_V^* \boldsymbol{Q}$, and the first equation of the dual problem (3.4.4), we get that

$$\begin{aligned}
(\phi_h^z, \Theta)_{\mathcal{T}_h} &= (\kappa^{-1} \boldsymbol{q}_h^z, \Pi_V^* \boldsymbol{Q})_{\mathcal{T}_h} + \langle \Pi_V^* \boldsymbol{Q} \cdot \boldsymbol{n}, \widehat{\phi}_h^z \rangle_{\partial \mathcal{T}_h} - (\phi_h^z, \boldsymbol{\beta} \cdot \nabla(\Pi_W^* \Psi))_{\mathcal{T}_h} \\
&\quad - \langle \phi_h^z, (\Pi_V^* \boldsymbol{Q} - \boldsymbol{Q}) \cdot \boldsymbol{n} - \boldsymbol{\beta} \cdot \boldsymbol{n} \, (\Pi_W^* \Psi - \Psi) \rangle_{\partial \mathcal{T}_h} - (\kappa^{-1} \boldsymbol{I}, \Pi_V^* \boldsymbol{Q})_{\mathcal{T}_h} \\
&= (\kappa^{-1} \boldsymbol{q}_h^z, \Pi_V^* \boldsymbol{Q} - \boldsymbol{Q})_{\mathcal{T}_h} + (\kappa^{-1} \boldsymbol{q}_h^z, \boldsymbol{Q})_{\mathcal{T}_h} + \langle \Pi_V^* \boldsymbol{Q} \cdot \boldsymbol{n}, \widehat{\phi}_h^z \rangle_{\partial \mathcal{T}_h} - (\phi_h^z, \boldsymbol{\beta} \cdot \nabla(\Pi_W^* \Psi))_{\mathcal{T}_h} \\
&\quad - \langle \phi_h^z, (\Pi_V^* \boldsymbol{Q} - \boldsymbol{Q}) \cdot \boldsymbol{n} - \boldsymbol{\beta} \cdot \boldsymbol{n} \, (\Pi_W^* \Psi - \Psi) \rangle_{\partial \mathcal{T}_h} - (\kappa^{-1} \boldsymbol{I}, \Pi_V^* \boldsymbol{Q})_{\mathcal{T}_h} \\
&= T_1 - (\boldsymbol{q}_h^z, \nabla \Psi)_{\mathcal{T}_h} + \langle \Pi_V^* \boldsymbol{Q} \cdot \boldsymbol{n}, \widehat{\phi}_h^z \rangle_{\partial \mathcal{T}_h} - (\phi_h^z, \boldsymbol{\beta} \cdot \nabla(\Pi_W^* \Psi))_{\mathcal{T}_h} \\
&\quad - \langle \phi_h^z, (\Pi_V^* \boldsymbol{Q} - \boldsymbol{Q}) \cdot \boldsymbol{n} - \boldsymbol{\beta} \cdot \boldsymbol{n} \, (\Pi_W^* \Psi - \Psi) \cdot \boldsymbol{n} \rangle_{\partial \mathcal{T}_h} - (\kappa^{-1} \boldsymbol{I}, \Pi_V^* \boldsymbol{Q})_{\mathcal{T}_h}.
\end{aligned}$$

Now, using the second equation of the HDG scheme (3.4.1b) on the fourth term of the right-hand side, with $w_h := \Pi_W^* \Psi$, we get that

$$\begin{aligned}
(\phi_h^z, \Theta)_{\mathcal{T}_h} = \; & T_1 - (\boldsymbol{q}_h^z, \nabla \Psi)_{\mathcal{T}_h} + \langle \Pi_V^* \boldsymbol{Q} \cdot \boldsymbol{n}, \widehat{\phi}_h^z \rangle_{\partial \mathcal{T}_h} + (f, \Pi_W^* \Psi)_{\mathcal{T}_h} + (\boldsymbol{q}_h^z, \nabla(\Pi_W^* \Psi))_{\mathcal{T}_h} \\
& - \langle \widehat{\boldsymbol{q}_h^z + \boldsymbol{\beta} \phi_h^z} \cdot \boldsymbol{n}, \Pi_W^* \Psi \rangle_{\partial \mathcal{T}_h} - \langle \phi_h^z, (\Pi_V^* \boldsymbol{Q} - \boldsymbol{Q}) \cdot \boldsymbol{n} - \boldsymbol{\beta} \cdot \boldsymbol{n} \, (\Pi_W^* \Psi - \Psi) \rangle_{\partial \mathcal{T}_h} \\
& - (\kappa^{-1} \boldsymbol{I}, \Pi_V^* \boldsymbol{Q})_{\mathcal{T}_h} \\
= \; & T_1 + (\boldsymbol{q}_h^z, \nabla(\Pi_W^* \Psi - \Psi_h))_{\mathcal{T}_h} + \langle \Pi_V^* \boldsymbol{Q} \cdot \boldsymbol{n}, \widehat{\phi}_h^z \rangle_{\partial \mathcal{T}_h} + (f, \Pi_W^* \Psi)_{\mathcal{T}_h} \\
& - \langle \widehat{\boldsymbol{q}_h^z + \boldsymbol{\beta} \phi_h^z} \cdot \boldsymbol{n}, \Pi_W^* \Psi \rangle_{\partial \mathcal{T}_h} - \langle \phi_h^z, (\Pi_V^* \boldsymbol{Q} - \boldsymbol{Q}) \cdot \boldsymbol{n} - \boldsymbol{\beta} \cdot \boldsymbol{n} \, (\Pi_W^* \Psi - \Psi) \rangle_{\partial \mathcal{T}_h} \\
& - (\kappa^{-1} \boldsymbol{I}, \Pi_V^* \boldsymbol{Q})_{\mathcal{T}_h}.
\end{aligned}$$

Now, using integration by parts and the second equation that defines Π_h^* (3.3.8b) with $w := \nabla \cdot \boldsymbol{q}_h^z$, the third term of the right-hand side becomes

$$(\boldsymbol{q}_h^z, \nabla(\Pi_W^* \Psi - \Psi))_{\mathcal{T}_h} = -(\nabla \cdot \boldsymbol{q}_h^z, \Pi_W^* \Psi - \Psi)_{\mathcal{T}_h} + \langle \boldsymbol{q}_h^z \cdot \boldsymbol{n}, \Pi_W^* \Psi - \Psi \rangle_{\partial \mathcal{T}_h}$$

$$= \langle \boldsymbol{q}_h^z \cdot \boldsymbol{n}, \Pi_W^* \Psi - \Psi \rangle_{\partial \mathcal{T}_h} + (\boldsymbol{q}_h^z, \nabla(\Psi - \Psi_h))_{\mathcal{T}_h}.$$

Using the last equality, we get that

$$\begin{aligned}(\phi_h^z, \Theta)_{\mathcal{T}_h} =& T_1 + \langle \boldsymbol{q}_h^z \cdot \boldsymbol{n}, \Pi_W^* \Psi - \Psi \rangle_{\partial \mathcal{T}_h} + \langle \Pi_V^* \boldsymbol{Q} \cdot \boldsymbol{n}, \widehat{\phi}_h^z \rangle_{\partial \mathcal{T}_h} + (f, \Pi_W^* \Psi)_{\mathcal{T}_h} \\ &- \langle \widehat{\boldsymbol{q}_h^z + \boldsymbol{\beta} \phi_h^z} \cdot \boldsymbol{n}, \Pi_W^* \Psi \rangle_{\partial \mathcal{T}_h} \\ &- \langle \phi_h^z, (\Pi_V^* \boldsymbol{Q} - \boldsymbol{Q}) \cdot \boldsymbol{n} - \boldsymbol{\beta} \cdot \boldsymbol{n} \ (\Pi_W^* \Psi - \Psi) \rangle_{\partial \mathcal{T}_h} - (\kappa^{-1} \boldsymbol{I}, \Pi_V^* \boldsymbol{Q})_{\mathcal{T}_h}.\end{aligned}$$

Now using the third equation of the HDG scheme (3.4.1c) with $\mu_h := P_M \Psi$, and the third equation of the dual problem (3.4.4), the fifth term on the right-hand side becomes

$$\begin{aligned}-\langle \widehat{\boldsymbol{q}_h^z + \boldsymbol{\beta} \phi_h^z} \cdot \boldsymbol{n}, \Pi_W^* \Psi \rangle_{\partial \mathcal{T}_h} &= -\langle \widehat{\boldsymbol{q}_h^z + \boldsymbol{\beta} \phi_h^z} \cdot \boldsymbol{n}, \Pi_W^* \Psi - P_M \Psi \rangle_{\partial \mathcal{T}_h} - \langle \widehat{\boldsymbol{q}_h^z + \boldsymbol{\beta} \phi_h^z} \cdot \boldsymbol{n}, P_M \Psi \rangle_{\partial \mathcal{T}_h} \\ &= -\langle \widehat{\boldsymbol{q}_h^z + \boldsymbol{\beta} \phi_h^z} \cdot \boldsymbol{n}, \Pi_W^* \Psi - P_M \Psi \rangle_{\partial \mathcal{T}_h} - \langle \widehat{\boldsymbol{q}_h^z + \boldsymbol{\beta} \phi_h^z} \cdot \boldsymbol{n}, P_M \Psi \rangle_{\partial \mathcal{T}_h \setminus \Gamma_{in}} \\ &\quad - \langle \widehat{\boldsymbol{q}_h^z + \boldsymbol{\beta} \phi_h^z} \cdot \boldsymbol{n}, P_M \Psi \rangle_{\Gamma_{in}} \\ &= -\langle \widehat{\boldsymbol{q}_h^z + \boldsymbol{\beta} \phi_h^z} \cdot \boldsymbol{n}, \Pi_W^* \Psi - P_M \Psi \rangle_{\partial \mathcal{T}_h} - \langle \varphi(z) \widehat{\phi}_h^z, P_M \Psi \rangle_{\Sigma} \\ &\quad - \langle \boldsymbol{\beta} \cdot \boldsymbol{n} \ \widehat{\phi}_h^z, P_M \Psi \rangle_{\Gamma_N} - \langle \Lambda, P_M \Psi \rangle_{\Sigma}.\end{aligned}$$

Expanding the numerical trace of the total flux, we get the next expression.

$$\begin{aligned}(\phi_h^z, \Theta)_{\mathcal{T}_h} = \ & T_1 + \langle \boldsymbol{q}_h^z \cdot \boldsymbol{n}, \Pi_W^* \Psi - \Psi \rangle_{\partial \mathcal{T}_h} + \langle \Pi_V^* \boldsymbol{Q} \cdot \boldsymbol{n}, \widehat{\phi}_h^z \rangle_{\partial \mathcal{T}_h} + (f, \Pi_W^* \Psi)_{\mathcal{T}_h} \\ &- \langle \boldsymbol{q}_h^z \cdot \boldsymbol{n}, \Pi_W^* \Psi - P_M \Psi \rangle_{\partial \mathcal{T}_h} - \langle \widehat{\phi}_h^z, \boldsymbol{\beta} \cdot \boldsymbol{n} (\Pi_W^* \Psi - P_M \Psi) \rangle_{\partial \mathcal{T}_h} \\ &- \langle \tau(\phi_h^z - \widehat{\phi}_h^z), \Pi_W^* \Psi - P_M \Psi \rangle_{\partial \mathcal{T}_h} - \langle \varphi(z) \widehat{\phi}_h^z, P_M \Psi \rangle_{\Sigma} - \langle \widehat{\phi}_h^z, \boldsymbol{\beta} \cdot \boldsymbol{n} \ P_M \Psi \rangle_{\Gamma_N} \\ &- \langle \phi_h^z, (\Pi_V^* \boldsymbol{Q} - \boldsymbol{Q}) \cdot \boldsymbol{n} \rangle_{\partial \mathcal{T}_h} + \langle \phi_h^z, \boldsymbol{\beta} \cdot \boldsymbol{n} \ (\Pi_W^* \Psi - \Psi) \rangle_{\partial \mathcal{T}_h} - (\kappa^{-1} \boldsymbol{I}, \Pi_V^* \boldsymbol{Q})_{\mathcal{T}_h} - \langle \Lambda, P_M \Psi \rangle_{\Sigma}.\end{aligned}$$

Re-arranging terms,

$$\begin{aligned}(\phi_h^z, \Theta)_{\mathcal{T}_h} =& T_1 + \langle \Pi_V^* \boldsymbol{Q} \cdot \boldsymbol{n}, \widehat{\phi}_h^z \rangle_{\partial \mathcal{T}_h} + (f, \Pi_W^* \Psi)_{\mathcal{T}_h} - \langle \widehat{\phi}_h^z, \boldsymbol{\beta} \cdot \boldsymbol{n} (\Pi_W^* \Psi - P_M \Psi) \rangle_{\partial \mathcal{T}_h} \\ &- \langle \tau(\phi_h^z - \widehat{\phi}_h^z), \Pi_W^* \Psi - P_M \Psi \rangle_{\partial \mathcal{T}_h} - (\kappa^{-1} \boldsymbol{I}, \Pi_V^* \boldsymbol{Q})_{\mathcal{T}_h} - \langle \Lambda, P_M \Psi \rangle_{\Sigma} \\ &- \langle \varphi(z) \widehat{\phi}_h^z, P_M \Psi \rangle_{\Sigma} - \langle \widehat{\phi}_h^z, \boldsymbol{\beta} \cdot \boldsymbol{n} \ P_M \Psi \rangle_{\Gamma_N}\end{aligned}$$

$$+ \langle \phi_h^z, (\boldsymbol{Q} - \Pi_V^* \boldsymbol{Q}) \cdot \boldsymbol{n} - \boldsymbol{\beta} \cdot \boldsymbol{n} \ (\Psi - \Pi_W^* \Psi) \rangle_{\partial \mathcal{T}_h}.$$

Adding and subtracting $\widehat{\phi}_h^z$ in the last term

$$\begin{aligned}
(\phi_h^z, \Theta)_{\mathcal{T}_h} &= T_1 + \langle \Pi_V^* \boldsymbol{Q} \cdot \boldsymbol{n}, \widehat{\phi}_h^z \rangle_{\partial \mathcal{T}_h} + (f, \Pi_W^* \Psi)_{\mathcal{T}_h} \\
&\quad - \langle \widehat{\phi}_h^z, \boldsymbol{\beta} \cdot \boldsymbol{n} (\Pi_W^* \Psi - P_M \Psi) \rangle_{\partial \mathcal{T}_h} - \langle \tau (\phi_h^z - \widehat{\phi}_h^z), \Pi_W^* \Psi - P_M \Psi \rangle_{\partial \mathcal{T}_h} \\
&\quad - (\kappa^{-1} \boldsymbol{I}, \Pi_V^* \boldsymbol{Q})_{\mathcal{T}_h} - \langle \Lambda, P_M \Psi \rangle_{\Sigma} \\
&\quad - \langle \varphi(z) \widehat{\phi}_h^z, P_M \Psi \rangle_{\Sigma} - \langle \widehat{\phi}_h^z, \boldsymbol{\beta} \cdot \boldsymbol{n} \ P_M \Psi \rangle_{\Gamma_N} \\
&\quad + \langle \phi_h^z - \widehat{\phi}_h^z, (\boldsymbol{Q} - \Pi_V^* \boldsymbol{Q}) \cdot \boldsymbol{n} - \boldsymbol{\beta} \cdot \boldsymbol{n} \ (\Psi - \Pi_W^* \Psi) \rangle_{\partial \mathcal{T}_h} \\
&\quad + \langle \widehat{\phi}_h^z, (\boldsymbol{Q} - \Pi_V^* \boldsymbol{Q}) \cdot \boldsymbol{n} - \boldsymbol{\beta} \cdot \boldsymbol{n} \ (\Psi - \Pi_W^* \Psi) \rangle_{\partial \mathcal{T}_h} \\
&= T_1 + (f, \Pi_W^* \Psi)_{\mathcal{T}_h} - (\kappa^{-1} \boldsymbol{I}, \Pi_V^* \boldsymbol{Q})_{\mathcal{T}_h} - \langle \Lambda, P_M \Psi \rangle_{\Sigma} \\
&\quad - \langle \varphi(z) \widehat{\phi}_h^z, P_M \Psi \rangle_{\Sigma} - \langle \widehat{\phi}_h^z, \boldsymbol{\beta} \cdot \boldsymbol{n} \ P_M \Psi \rangle_{\Gamma_N} \\
&\quad + \langle \phi_h^z - \widehat{\phi}_h^z, (\boldsymbol{Q} - \Pi_V^* \boldsymbol{Q}) \cdot \boldsymbol{n} + (\tau - \boldsymbol{\beta} \cdot \boldsymbol{n}) \ (\Psi - \Pi_W^* \Psi) \rangle_{\partial \mathcal{T}_h} \\
&\quad + \langle \widehat{\phi}_h^z, \boldsymbol{Q} \cdot \boldsymbol{n} - \boldsymbol{\beta} \cdot \boldsymbol{n} \ (\Psi - P_M \Psi) \rangle_{\partial \mathcal{T}_h}
\end{aligned}$$

The seventh term vanishes thanks to (3.3.8c). Moreover, since $\widehat{\phi}_h^z$ is single-valued, and that $\widehat{\phi}_h^z = 0$ on Γ_{in}, together with the facts that $\boldsymbol{\beta}$ and $\boldsymbol{Q}$ are in $\boldsymbol{H}(\mathrm{div}; \Omega)$, the contribution of the last term is only on the boundary $\partial\Omega$, the contribution of the last term is only on the boundary Γ_N. Thus,

$$\begin{aligned}
(\phi_h^z, \Theta)_{\mathcal{T}_h} &= T_1 + (f, \Pi_W^* \Psi)_{\mathcal{T}_h} - \langle \varphi(z) \widehat{\phi}_h^z, P_M \Psi \rangle_{\Sigma} - \langle \widehat{\phi}_h^z, \boldsymbol{\beta} \cdot \boldsymbol{n} \ P_M \Psi \rangle_{\Gamma_N} \\
&\quad + \langle \widehat{\phi}_h^z, \boldsymbol{Q} \cdot \boldsymbol{n} - \boldsymbol{\beta} \cdot \boldsymbol{n} \ (\Psi - P_M \Psi) \rangle_{\Gamma_N} - (\kappa^{-1} \boldsymbol{I}, \Pi_V^* \boldsymbol{Q})_{\mathcal{T}_h} - \langle \Lambda, P_M \Psi \rangle_{\Sigma} \\
&= T_1 + (f, \Pi_W^* \Psi)_{\mathcal{T}_h} - \langle \varphi(z) \widehat{\phi}_h^z - \boldsymbol{\beta} \cdot \boldsymbol{n} \ \widehat{\phi}_h^z, P_M \Psi \rangle_{\Sigma} - \langle \widehat{\phi}_h^z, \boldsymbol{\beta} \cdot \boldsymbol{n} \ P_M \Psi \rangle_{\Gamma_{\text{out}}} \\
&\quad + \langle \widehat{\phi}_h^z, \boldsymbol{Q} \cdot \boldsymbol{n} - \boldsymbol{\beta} \cdot \boldsymbol{n} \ (\Psi - P_M \Psi) \rangle_{\Gamma_{\text{out}}} + \langle \widehat{\phi}_h^z, \boldsymbol{Q} \cdot \boldsymbol{n} - \boldsymbol{\beta} \cdot \boldsymbol{n} \ (\Psi - P_M \Psi) \rangle_{\Sigma} \\
&\quad - (\kappa^{-1} \boldsymbol{I}, \Pi_V^* \boldsymbol{Q})_{\mathcal{T}_h} - \langle \Lambda, P_M \Psi \rangle_{\Sigma} \\
&= T_1 + (f, \Pi_W^* \Psi)_{\mathcal{T}_h} - \langle \varphi(z) \widehat{\phi}_h^z - \boldsymbol{\beta} \cdot \boldsymbol{n} \ \widehat{\phi}_h^z, P_M \Psi \rangle_{\Sigma} \\
&\quad + \langle \widehat{\phi}_h^z, \boldsymbol{Q} \cdot \boldsymbol{n} - \boldsymbol{\beta} \cdot \boldsymbol{n} \ (\Psi - P_M \Psi) \rangle_{\Sigma} - (\kappa^{-1} \boldsymbol{I}, \Pi_V^* \boldsymbol{Q})_{\mathcal{T}_h} - \langle \Lambda, P_M \Psi \rangle_{\Sigma} \\
&= T_1 + (f, \Pi_W^* \Psi)_{\mathcal{T}_h} - \langle \varphi(z) \widehat{\phi}_h^z, P_M \Psi \rangle_{\Sigma} - \kappa^{-1} \boldsymbol{I}, \Pi_V^* \boldsymbol{Q})_{\mathcal{T}_h} - \langle \Lambda, P_M \Psi \rangle_{\Sigma}
\end{aligned}$$

$$
\begin{aligned}
&+ \langle \widehat{\phi}_h^z, \boldsymbol{Q}\cdot\boldsymbol{n} - \boldsymbol{\beta}\cdot\boldsymbol{n}\,\Psi\rangle_\Sigma \\
&= T_1 + (f, \Pi_W^*\Psi)_{\mathcal{T}_h} - \langle \varphi(z)\widehat{\phi}_h^z, P_M\Psi\rangle_\Sigma - (\kappa^{-1}\boldsymbol{I}, \Pi_V^*\boldsymbol{Q})_{\mathcal{T}_h} - \langle \Lambda, P_M\Psi\rangle_\Sigma \\
&\quad + \langle \widehat{\phi}_h^z, -\varphi(z)\Psi\rangle_\Sigma \\
&= T_1 + (f, \Pi_W^*\Psi)_{\mathcal{T}_h} - \langle \varphi(z)\widehat{\phi}_h^z, P_M\Psi - \Psi\rangle_\Sigma - (\kappa^{-1}\boldsymbol{I}, \Pi_V^*\boldsymbol{Q})_{\mathcal{T}_h} - \langle \Lambda, P_M\Psi\rangle_\Sigma.
\end{aligned}
$$

Finally, we observe that

$$
\begin{aligned}
(\kappa^{-1}\boldsymbol{I}, \Pi_V^*\boldsymbol{Q})_{\mathcal{T}_h} &= (\kappa^{-1}\boldsymbol{I}, \Pi_V^*\boldsymbol{Q} - \boldsymbol{Q})_{\mathcal{T}_h} + (\kappa^{-1}\boldsymbol{I}, \boldsymbol{Q})_{\mathcal{T}_h} \\
&= (\kappa^{-1}\boldsymbol{I}, \Pi_V^*\boldsymbol{Q} - \boldsymbol{Q})_{\mathcal{T}_h} - (\boldsymbol{I}, \nabla\Psi)_{\mathcal{T}_h} \\
&= (\kappa^{-1}\boldsymbol{I}, \Pi_V^*\boldsymbol{Q} - \boldsymbol{Q})_{\mathcal{T}_h} - (\boldsymbol{I}, \nabla\Psi - \nabla\Psi_h)_{\mathcal{T}_h},
\end{aligned}
$$

for all $\Psi_h \in Q_h$, where we have used the fact that $\boldsymbol{I}$ is orthogonal to polynomials of degree $k-1$. The result follows.

□

Corollary 3.4.1. *Suppose that Assumption (3.4.5) is satisfied, $f \in L^2(\Omega)$ and $k \geq 1$. There exists a positive constant C_d, independent of h, such that*

$$
\|\phi_h^z\|_{\mathcal{T}_h} \leq C_d \Big(h\ \|\kappa^{-1/2}\boldsymbol{q}_h^z\|_{\mathcal{T}_h} + \|f\|_{0,\Omega} + h\|\boldsymbol{I}\|_{\mathcal{T}_h} + \|\Lambda\|_\Sigma + h^{3/2}\|\varphi(z)\|_{\infty,\Sigma}\ \|\widehat{\phi}_h^z\|_\Sigma \Big). \tag{3.4.6}
$$

Proof. We have to estimate the terms on the right-hand side of the identity of Lemma 3.4.2. For this purpose, we use (3.3.9) with $\ell_Q = 0$ and $\ell_\Psi = 1$ to get

$$
\begin{aligned}
\|\Pi_W^*\Psi - \Psi\|_{0,K} &\leq \max\{C_W^{1,*}, C_W^{2,*}\}\max\{1, \sqrt{d}\ \}\ h\Big(|Q|_{1,K} + |\Psi|_{2,K}\Big), \\
\|\Pi_V^*\boldsymbol{Q} - \boldsymbol{Q}\|_{0,K} &\leq \max\{C_V^{1,*}, C_V^{2,*}\}\ h\Big(|Q|_{1,K} + |\Psi|_{2,K}\Big).
\end{aligned}
$$

where $C_W^{1,*} := \max_K C_{W,K}^{1,*}$, $C_W^{2,*} := \max_K C_{W,K}^{2,*}$, $C_V^{1,*} := \max_K C_{V,K}^{1,*}$ and $C_V^{2,*} := \max_K C_{V,K}^{2,*}$. Now we get that

$$
\|\Pi_W^*\Psi - \Psi\|_{\mathcal{T}_h} \leq \sqrt{2}\max\{C_W^{1,*}, C_W^{2,*}\}\max\{1, \sqrt{d}\ \}\ C_{reg}\ h\|\Theta\|_{0,\Omega}, \tag{3.4.7}
$$

$$\|\Pi_V^* \boldsymbol{Q} - \boldsymbol{Q}\|_{\mathcal{T}_h} \leq \sqrt{2}\max\{C_V^{1,*}, C_V^{2,*}\} C_{reg}\ h \|\Theta\|_{0,\Omega}. \tag{3.4.8}$$

With this, by the Cauchy-Schwarz inequality and (3.4.8) we get that

$$|T_1| \leq \|\kappa^{-1}\boldsymbol{q}_h^z\|_{\mathcal{T}_h}\ \|\Pi_V^*\boldsymbol{Q} - \boldsymbol{Q}\|_{\mathcal{T}_h} \leq C_{T_1} h \|\kappa^{-1/2}\boldsymbol{q}_h^z\|_{\mathcal{T}_h} \|\Theta\|_{0,\Omega},$$

where $C_{T_1} := \kappa^{-1/2}\sqrt{2}\max\{C_V^{1,*}, C_V^{2,*}\} C_{reg}$.

For T_2, we have

$$|T_2| \leq C_{T_2} \|f\|_{0,\Omega} \|\Theta\|_{0,\Omega},$$

where $C_{T_2} := C_{reg} C_{\Pi_W^*}$ and $C_{\Pi_W^*} > 0$ is a constant that depends on the projection Π_W^*. For T_3,

$$\begin{aligned}
|T_3| &\leq \sum_{e\subset\Sigma} \|\widehat{\phi}_h^z\|_{0,e}\ \|\varphi(z)\|_{\infty,e}\ \|P_M\Psi - \Psi\|_{0,e} \\
&\leq \|\varphi(z)\|_{\infty,\Sigma} \sum_{e\subset\Sigma} \|\widehat{\phi}_h^z\|_{0,e}\ C''_{app}\ h_{K_e}^{3/2} |\Psi|_{2,K_e} \\
&\leq C''_{app}\ h^{3/2} \|\varphi(z)\|_{\infty,\Sigma} \Big(\sum_{e\subset\Sigma} \|\widehat{\phi}_h^z\|_{0,e}^2 \Big)^{1/2} \Big(\sum_{K\in\mathcal{T}_h} \|\Psi\|_{2,K}^2 \Big)^{1/2} \\
&\leq C_{T_3}\ h^{3/2} \|\varphi(z)\|_{\infty,\Sigma}\ \|\widehat{\phi}_h^z\|_{\Sigma}\ \|\Theta\|_{0,\Omega},
\end{aligned}$$

where K_e denotes the element that has the edge e, and $C_{T_3} := C''_{app} C_{reg}$.

Now, by the Cauchy-Schwarz inequality and assumption (3.4.5) we get that

$$|T_4| \leq C_{reg} \|\Lambda\|_{\Sigma} \|\Theta\|_{0,\Omega}.$$

For the last term, using (3.4.8), we have that

$$|T_5| \leq \|\kappa^{-1/2}\boldsymbol{I}\|_{\mathcal{T}_h} \sqrt{2}\max\{C_V^{1,*}, C_V^{2,*}\} C_{reg}\ h \|\Theta\|_{0,\Omega} + \|\kappa^{-1/2}\boldsymbol{I}\|_{\mathcal{T}_h} \|\kappa^{1/2}(\nabla\Psi - \nabla\Psi_h)\|_{\mathcal{T}_h}.$$

Moreover, if we take Ψ_h as the L^2-projection over Q_h, we know (cf. [7]) that there exists a

positive constant C_2, independent of the meshsize, such that

$$\|\kappa^{1/2}(\nabla\Psi - \nabla\Psi_h)\|_{\mathcal{T}_h} \leq C_2 h\|\Psi\|_{2,\Omega} \leq C_2 C_{reg} h\|\Theta\|_{0,\Omega}$$

We conclude that there exists $C_{T_5} > 0$, independent of h, such that

$$|T_5| \leq C_{T_5} h\|\kappa^{-1/2}\boldsymbol{I}\|_{\mathcal{T}_h}\|\Theta\|_{0,\Omega},$$

with $C_{T_5} := \sqrt{2}\max\{C_V^{1,*}, C_V^{2,*}\} + \kappa^{1/2}C_2$.

The result in (3.4.6) follows from the choice of $\Theta = \phi_h^z$ and $C_d := \max\{C_{T_1}, C_{T_2}, C_{T_3}, C_{reg}, C_{T_5}\}$.

□

On the other hand, following the proof of [6, Theorem 4.1], we see that given $K \in \mathcal{T}_h$ and $p \in \mathbb{P}_k(\partial K)$, there exist $\boldsymbol{r} \in [\mathbb{P}_k(K)]^d$ and a constant $\tilde{C}$ such that $\boldsymbol{r}\cdot\boldsymbol{n} = p$ in ∂K, and $\|\boldsymbol{r}\|_{0,K} \leq \tilde{C}h^{1/2}\|p\|_{0,\partial K}$. Then, by the first equation of the HDG scheme (3.4.1a), and considering $p = \widehat{\phi}_h^z$, we obtain that

$$\begin{aligned}
\|\widehat{\phi}_h^z\|_{0,\partial K}^2 &= -(\kappa^{-1}\boldsymbol{q}_h^z, \boldsymbol{r}_h)_K + (\phi_h^z, \nabla\cdot\boldsymbol{r}_h)_K + (\kappa^{-1}\boldsymbol{I}, \boldsymbol{r}_h)_K \\
&\leq \|\kappa^{-1}\boldsymbol{q}_h^z\|_{0,K}\ \|\boldsymbol{r}_h\|_{0,K} + C_{inv}\ h_K^{-1}\ \|\phi_h^z\|_{0,K}\ \|\boldsymbol{r}_h\|_{0,K} + \|\kappa^{-1}\boldsymbol{I}\|_{0,K}\|\boldsymbol{r}_h\|_{0,K} \\
&\leq (\|\kappa^{-1}\boldsymbol{q}_h^z\|_{0,K} + C_{inv}\ h_K^{-1}\ \|\phi_h^z\|_{0,K})\ \tilde{C}h^{1/2}\|\widehat{\phi}_h^z\|_{0,\partial K} + \tilde{C}h^{1/2}\|\kappa^{-1}\boldsymbol{I}\|_{0,K}\|\widehat{\phi}_h^z\|_{0,\partial K},
\end{aligned}$$

which, implies that

$$\|\widehat{\phi}_h^z\|_h \leq \widetilde{C_1} h\|\kappa^{-1/2}\boldsymbol{q}_h^z\|_{\mathcal{T}_h} + \widetilde{C_2}\|\phi_h^z\|_{\mathcal{T}_h} + \tilde{C}h\|\kappa^{-1}\boldsymbol{I}\|_{\mathcal{T}_h}, \tag{3.4.9}$$

with $\widetilde{C_1} := \kappa^{1/2}\sqrt{2}\ \tilde{C}$ and $\widetilde{C_2} := \sqrt{2}\ \tilde{C}C_{inv}$. Here

$$\|\cdot\|_h := \left(\sum_{K\in\mathcal{T}_h} h_K\|\cdot\|_{\partial K}^2\right)^{1/2}.$$

3.4.3 Energy bound

Lemma 3.4.3. *Let ϵ_Λ a positive parameter at our disposal such that $\epsilon_\Lambda = 0$ if $\Lambda = 0$ and $\epsilon_\Lambda > 0$ otherwise. Moreover, we assume that $\epsilon_\Lambda < \min_\Sigma |\boldsymbol{\beta}\cdot\boldsymbol{n}|$.*

If $f = 0$, then

$$\frac{1}{2}E^2 + \left(\frac{1}{2}(c_\beta - \epsilon_\Lambda) - \|\varphi(z)\|_{\infty,\Sigma}\right)\|\widehat{\phi}_h^z\|_\Sigma^2 \leq \frac{1}{2\epsilon_\Lambda}\|\Lambda\|_\Sigma^2 + \frac{1}{2}\|\kappa^{-1/2}\boldsymbol{I}\|_{\mathcal{T}_h}^2,$$

where $c_\beta := \min_\Sigma |\boldsymbol{\beta}\cdot\boldsymbol{n}|$. Moreover, if $f \neq 0$, then

$$\begin{aligned}\frac{1}{2}E^2 + \left(\frac{1}{4}(c_\beta - \epsilon_\Lambda) - \|\varphi(z)\|_{\infty,\Sigma}\right)\|\widehat{\phi}_h^z\|_\Sigma^2 &\\ \leq C_d\left(h^2C_d + 1 + \frac{hC_d}{2} + \frac{C_d}{2} + \frac{h^3C_d}{c_\beta - \epsilon_\Lambda}\|\varphi(z)\|_{\infty,\Sigma}^2\right)\|f\|_{0,\Omega}^2 &\\ + \frac{1}{\epsilon_\Lambda}\|\Lambda\|_\Sigma^2 + \frac{1}{2}(2+h)\|\kappa^{-1/2}\boldsymbol{I}\|_{\mathcal{T}_h}^2,&\end{aligned} \tag{3.4.10}$$

Proof. By Cauchy-Schwarz inequality applied to (3.4.1) we get that

$$E^2 + \frac{c_\beta}{2}\|\widehat{\phi}_h^z\|_\Sigma^2 \leq \|\varphi(z)\|_{\infty,\Sigma}\|\widehat{\phi}_h^z\|_\Sigma^2 + \|\Lambda\|_\Sigma\|\widehat{\phi}_h^z\|_\Sigma + \|f\|_{0,\Omega}\|\phi_h^z\|_{\mathcal{T}_h} + \|\kappa^{-1/2}\boldsymbol{I}\|_{\mathcal{T}_h}\|\kappa^{-1/2}\boldsymbol{q}_h^z\|_{\mathcal{T}_h}.$$

Now, by Young's inequality, there exist positive parameters ϵ_E and ϵ_Λ at our disposal such that

$$\begin{aligned}\|\Lambda\|_\Sigma\|\widehat{\phi}_h^z\|_\Sigma &\leq \frac{1}{2\epsilon_\Lambda}\|\Lambda\|_\Sigma^2 + \frac{\epsilon_\Lambda}{2}\|\widehat{\phi}_h^z\|_\Sigma^2,\\ \|\kappa^{-1/2}\boldsymbol{I}\|_{\mathcal{T}_h}\|\kappa^{-1/2}\boldsymbol{q}_h^z\|_{\mathcal{T}_h} &\leq \frac{1}{2\epsilon_E}\|\kappa^{-1/2}\boldsymbol{I}\|_{\mathcal{T}_h}^2 + \frac{\epsilon_E}{2}\|\kappa^{-1/2}\boldsymbol{q}_h^z\|_{\mathcal{T}_h}^2.\end{aligned}$$

Thus, the result for $f = 0$ follows by choosing $\epsilon_E = 1$.

On the other hand, if $f \neq 0$, then by Young's inequality, there exist positive parameters $\epsilon_1, \epsilon_2, \epsilon_3$ and ϵ_4 at our disposal such that, thanks to Corollary 3.4.1,

$$\begin{aligned}\|f\|_{0,\Omega}\|\phi_h^z\|_{\mathcal{T}_h} \leq C_d h\left(\frac{1}{2\epsilon_1}\|f\|_{0,\Omega}^2 + \frac{\epsilon_1}{2}E^2\right) + C_d\|f\|_{0,\Omega}^2 + C_d h\left(\frac{1}{2\epsilon_2}\|f\|_{0,\Omega}^2 + \frac{\epsilon_2}{2}\|\kappa^{-1/2}\boldsymbol{I}\|_{\mathcal{T}_h}^2\right)&\\ + C_d\left(\frac{1}{2\epsilon_3}\|f\|_{0,\Omega}^2 + \frac{\epsilon_3}{2}\|\Lambda\|_\Sigma^2\right) + C_d h^{3/2}\left(\frac{1}{2\epsilon_4}\|\varphi(z)\|_{\infty,\Sigma}^2\|f\|_{0,\Omega}^2 + \frac{\epsilon_4}{2}\|\widehat{\phi}_h^z\|_\Sigma^2\right).&\end{aligned}$$

So, now we have that

$$\left(1-\frac{\epsilon_E}{2}-C_d h\frac{\epsilon_1}{2}\right)E^2+\left(\frac{c_\beta}{2}-\frac{\epsilon_\Lambda}{2}-C_d h^{3/2}\frac{\epsilon_4}{2}-\|\varphi(z)\|_{\infty,\Sigma}\right)\|\widehat{\phi}_h^z\|_\Sigma^2$$
$$\leq C_d\left(\frac{h}{2\epsilon_1}+1+\frac{h}{2\epsilon_2}+\frac{1}{2\epsilon_3}+\frac{h^{3/2}\|\varphi(z)\|_{\infty,\Sigma}^2}{2\epsilon_4}\right)\|f\|_{0,\Omega}^2$$
$$+\left(\frac{1}{2\epsilon_\Lambda}+\frac{C_d\epsilon_3}{2}\right)\|\Lambda\|_\Sigma^2+\left(\frac{1}{2\epsilon_E}+\frac{C_d h\epsilon_2}{2}\right)\|\kappa^{-1/2}\boldsymbol{I}\|_{\mathcal{T}_h}^2.$$

By choosing $\epsilon_E=\frac{1}{2}$, $\epsilon_1=(C_d^{-1}h^{-1})/2$, $\epsilon_2=C_d^{-1}$, $\epsilon_3=C_d^{-1}/\epsilon_\Lambda$ and $\epsilon_4=\left(C_d^{-1}h^{-3/2}(c_\beta-\epsilon_\Lambda)\right)/2$, we obtain the result. □

Corollary 3.4.2. *Let us assume that there exists a positive constant M_∞, independent of h, such that for all $z\in M_h^\Sigma$, it holds that*

$$\|\varphi(z)\|_{\infty,\Sigma}\leq M_\infty. \tag{3.4.11}$$

If $M_\infty<(c_\beta-\epsilon_\Lambda)/4$, there exists a constant $C_\varphi>0$ independent of h such that, when $f=0$,

$$E^2+C_\varphi\|\widehat{\phi}_h^z\|_\Sigma^2\leq\frac{1}{\epsilon_\Lambda}\|\Lambda\|_\Sigma^2+\|\kappa^{-1/2}\boldsymbol{I}\|_{\mathcal{T}_h}^2.$$

In addition, if $f\neq 0$, then there exists a constant $C_E>0$ and $h_0>0$, independent of h, such that

$$E^2+C_\varphi\|\widehat{\phi}_h^z\|_\Sigma^2\leq C_E\left(\|f\|_{0,\Omega}^2+\frac{1}{\epsilon_\Lambda}\|\Lambda\|_\Sigma^2+(2+h)\|\kappa^{-1/2}\boldsymbol{I}\|_{\mathcal{T}_h}^2\right),$$

for all $h<h_0$.

Proof. Let us define $C_\varphi:=\frac{1}{2}(c_\beta-\epsilon_\Lambda)-2M_\infty>0$. The result when $f=0$ follows directly from the assumption over $\varphi(z)$ and Lemma 3.4.2. If $f\neq 0$, we have

$$\frac{1}{2}E^2+\frac{1}{2}C_\varphi\|\widehat{\phi}_h^z\|_\Sigma^2\leq C_d\left(h^2C_d+1+\frac{hC_d}{2}+\frac{C_d}{2}+\frac{h^3C_d}{c_\beta-\epsilon_\Lambda}\|\varphi(z)\|_{\infty,\Sigma}^2\right)\|f\|_{0,\Omega}^2$$

$$+\frac{1}{\epsilon_\Lambda}\|\Lambda\|_\Sigma^2+\frac{1}{2}\left(2+h\right)\|\kappa^{-1/2}\boldsymbol{I}\|_{\mathcal{T}_h}^2.$$

We notice that

$$\frac{h^3C_d}{c_\beta-\epsilon_\Lambda}\|\varphi(z)\|_{\infty,\Sigma}^2<\frac{h^3C_d}{16}(c_\beta-\epsilon_\Lambda).$$

The result follows thanks to Lemma 3.4.3, where

$$C_E:=\max\left\{1,C_d\left(2C_dh_0^2+C_dh_0+\frac{h_0^3C_d}{8}(c_\beta-\epsilon_\Lambda)+C_d+2\right)\right\}.$$

□

3.5 Solvability analysis of the fixed point scheme

First of all, the operator T_h is well-defined. In fact, given $z\in \mathrm{M}_h$, if $f=0$ and $\phi_{\text{in}}=0$, by the Fredholm alternative, it is enough to show that (3.2.2) has only the trivial solution. The latter follows directly from Corollaries 3.4.2 and 3.4.1 and the estimate in (3.4.9), noticing that (3.2.2) is a particular case of (3.4.1) with $\Lambda=0$ and $\boldsymbol{I}=\mathbf{0}$.

Now, by Corollary 3.4.2 with $\Lambda=g(\phi_{\text{in}})$ and $\boldsymbol{I}=\mathbf{0}$, if $f=0$ we have

$$C_\varphi\|T_h(z)\|_\Sigma^2=C_\varphi\|\widehat{\phi}_h^z\|_\Sigma^2\leq\frac{1}{\epsilon_\Lambda}\|g(\phi_{\text{in}})\|_\Sigma^2=\frac{c_2^2}{\epsilon_\Lambda}|\Sigma|,$$

and if $f\neq 0$ we have

$$C_\varphi\|T_h(z)\|_\Sigma^2=C_\varphi\|\widehat{\phi}_h^z\|_\Sigma^2\leq C_E\left(\|f\|_{0,\Omega}^2+\frac{1}{\epsilon_\Lambda}\|g(\phi_{\text{in}})\|_\Sigma^2\right)=C_E\left(\|f\|_{0,\Omega}^2+\frac{c_2^2}{\epsilon_\Lambda}|\Sigma|\right).$$

Let us define the ball

$$B_2:=\{z\in M_h^\Sigma:\|z\|_\Sigma\leq R_2\},\tag{3.5.1}$$

where $R_2>0$. Thus, we have the following result.

Lemma 3.5.1. *Suppose that the assumptions of Corollary 3.4.2 are satisfied. Moreover, let us*

assume that R_2 *and the data* c_2 *satisfy*

$$\frac{c_2^2}{C_\varphi \epsilon_\Lambda}|\Sigma| \leq R_2^2 \tag{3.5.2}$$

when $f = 0$*; and* R_2 *and the data* f *and* c_2 *satisfy*

$$\frac{C_E}{C_\varphi}\left(\|f\|_{0,\Omega}^2 + \frac{c_2^2}{\epsilon_\Lambda}|\Sigma|\right) \leq R_2^2, \tag{3.5.3}$$

when $f \neq 0$*. Then,* $T_h(B_2) \subset B_2$*.*

On the other hand, for $i \in \{1,2\}$, let $z_i \in B_2$ such that the hypothesis over the data of Lemma 3.5.1 and Corollary 3.4.2 are satisfied. We set $T_h(z_i) := \widehat{\phi}_h^{z_i}$, where $(\boldsymbol{q}_h^{z_i}, \phi_h^{z_i}, \widehat{\phi}_h^{z_i})$ is the only solution to (3.2.2) with z_i as data. In this scenario $\Lambda = g(\phi_{\text{in}})$ and $\boldsymbol{I} = 0$. Moreover, $g(\phi_{\text{in}})|_\Sigma = a_3\phi_{\text{in}} + a_1\phi_{\text{in}}^2 =: c_2$. By Lemma 3.5.1, we have,

$$C_\varphi \|\widehat{\phi}_h^{z_i}\|_\Sigma^2 \leq C_\varphi R_2^2. \tag{3.5.4}$$

If we define $z := z_1 - z_2$ and $(\boldsymbol{q}_h^z, \phi_h^z, \widehat{\phi}_h^z) := (\boldsymbol{q}_h^{z_1} - \boldsymbol{q}_h^{z_2}, \phi_h^{z_1} - \phi_h^{z_2}, \widehat{\phi}_h^{z_1} - \widehat{\phi}_h^{z_2})$, then $(\boldsymbol{q}_h^z, \phi_h^z, \widehat{\phi}_h^z) \in \boldsymbol{H}_h \times Q_h \times \mathrm{M}_h$ satisfy

$$(\kappa^{-1}\boldsymbol{q}_h^z, \boldsymbol{r}_h)_{\mathcal{T}_h} - (\phi_h^z, \nabla\cdot\boldsymbol{r}_h)_{\mathcal{T}_h} + \langle \boldsymbol{r}_h\cdot\boldsymbol{n}, \widehat{\phi}_h^z\rangle_{\partial\mathcal{T}_h} = 0, \tag{3.5.5a}$$

$$-(\boldsymbol{q}_h^z + \boldsymbol{\beta}\phi_h^z, \nabla w_h)_{\mathcal{T}_h} + \langle \widehat{\boldsymbol{q}_h^z + \boldsymbol{\beta}\phi_h^z}\cdot\boldsymbol{n}, w_h\rangle_{\partial\mathcal{T}_h} = 0, \tag{3.5.5b}$$

$$\langle \widehat{\boldsymbol{q}_h^z + \boldsymbol{\beta}\phi_h^z}\cdot\boldsymbol{n}, \mu_h\rangle_{\partial\mathcal{T}_h\setminus\Gamma_{\text{in}}} = \langle \varphi(z_1)\widehat{\phi}_h^z + \boldsymbol{\beta}\widehat{\phi}_h^z\cdot\boldsymbol{n} + \Lambda, \mu_h\rangle_{\Gamma_N}, \tag{3.5.5c}$$

$$\langle \widehat{\phi}_h^z, \mu_h\rangle_{\Gamma_{\text{in}}} = 0, \tag{3.5.5d}$$

for all $(\boldsymbol{r}_h, w_h, \mu_h) \in \boldsymbol{H}_h \times Q_h \times \mathrm{M}_h$, with

$$\Lambda := (\varphi(z_1) - \varphi(z_2))\widehat{\phi}_h^{z_2}.$$

Also thanks to Lemma 3.5.1, since in this case $f = 0$ and $\boldsymbol{I} = 0$ we have

$$C_\varphi \|\widehat{\phi}_h^z\|_\Sigma^2 \leq \frac{1}{\epsilon_\Lambda} \|\Lambda\|_\Sigma^2. \tag{3.5.6}$$

Now, we note that

$$\|\varphi(x) - \varphi(y)\|_{\infty,\Sigma} = c_1 \|x - y\|_{\infty,\Sigma} \quad \forall x, y \in L^\infty(\Sigma), \tag{3.5.7}$$

and if we define $e^* \subseteq \Sigma$ such that $\|z\|_{\infty,\Sigma} = \|z\|_{\infty,e^*}$, then by (3.3.1) we get the following.

$$\|\Lambda\|_\Sigma \leq c_1 \|z_1 - z_2\|_{\infty,\Sigma} \|\widehat{\phi}_h^{z_2}\|_\Sigma \leq c_1 C_{eq} h_{e^*}^{-1/2} \|z_1 - z_2\|_{0,e^*} \|\widehat{\phi}_h^{z_2}\|_\Sigma. \tag{3.5.8}$$

Then

$$C_\varphi \|\widehat{\phi}_h^z\|_\Sigma^2 \leq \frac{c_1^2}{\epsilon_\Lambda} C_{eq}^2 h_{e^*}^{-1} R_2^2 \|z_1 - z_2\|_{0,e^*}^2.$$

Therefore, we have that

$$\|T_h(z_1) - T_h(z_2)\|_\Sigma^2 = \|\widehat{\phi}_h^z\|_\Sigma^2 \leq L_{T_h}^2 \|z\|_{0,e^*}^2,$$

with

$$L_{T_h} := \frac{c_1 C_{eq} h_{e^*}^{-1/2} R_2}{\epsilon_\Lambda C_\varphi^{1/2}}.$$

In other words, we have the following result.

Theorem 3.5.2. *Let us assume that the data satisfies hypotheses of Corollary 3.4.2 and Lemma 3.5.1. Moreover, let us assume that the data c_2 satisfies*

$$\frac{c_2^2}{C_\varphi \epsilon_\Lambda} |\Sigma| \leq R_2^2 \tag{3.5.9}$$

when $f = 0$; and that the data f and c_2 satisfy

$$\frac{C_E}{C_\varphi} \left(\|f\|_{0,\Omega}^2 + \left(\frac{1}{\epsilon_\Lambda} + 1 \right) c_2^2 |\Sigma| \right) \leq R_2^2, \tag{3.5.10}$$

and the data c_1 is small enough such that $L_{T_h} < 1$, then (3.1.3) has a unique fixed-point.

Remark on the feasibility of the smallness assumption. In the reverse osmosis model [1], $f = 0$, and we expect that assumptions (3.4.11) and (3.5.3) hold since c_1, c_2, and c_3 are extremely small. On the other hand, in order to ensure $L_{T_h} < 1$, it is necessary that $c_1 h_{e^*}^{-1/2}$ is sufficiently small. In other words, for a given c_1, our result guarantees the existence and uniqueness of the solution for all mesh sizes h such that $c_1 h_{e^*}^{-1/2}$ is small enough.

CHAPTER 4

A priori error analysis

4.1 Analysis of the projection of the errors

In this section we will prove the error estimates for $\boldsymbol{q}$ and ϕ. For this purpose, let us define the projection of errors as follows.

$$e_{\boldsymbol{q}} := \Pi_V \boldsymbol{q} - \boldsymbol{q}_h, \tag{4.1.1a}$$

$$e_{\phi} := \Pi_W \phi - \phi_h, \tag{4.1.1b}$$

$$e_{\hat{\phi}} := P_M \phi - \widehat{\phi}_h, \tag{4.1.1c}$$

$$\widehat{e} \cdot \boldsymbol{n} := P_M \left((\boldsymbol{q} + \boldsymbol{\beta}\phi) \cdot \boldsymbol{n} - \widehat{\boldsymbol{q}_h + \boldsymbol{\beta}\phi_h} \cdot \boldsymbol{n} \right), \tag{4.1.1d}$$

and we define the error of the projection as

$$\boldsymbol{I}_{\boldsymbol{q}} := \boldsymbol{q} - \Pi_V \boldsymbol{q}, \tag{4.1.2a}$$

$$\boldsymbol{I}_{\phi} := \phi - \Pi_W \phi, \tag{4.1.2b}$$

$$\boldsymbol{I}_{\hat{\phi}} := \phi - P_M\phi, \tag{4.1.2c}$$

so we can write $\boldsymbol{q} - \boldsymbol{q}_h = \boldsymbol{I}_{\boldsymbol{q}} + e_{\boldsymbol{q}}$, $\phi - \phi_h = \boldsymbol{I}_\phi + e_\phi$ and $(\phi - \widehat{\phi}_h)|_e = \boldsymbol{I}_{\hat{\phi}} + e_{\hat{\phi}}$ for all $e \in \mathcal{E}_h$.

We begin our analysis by establishing the following equality

$$\widehat{e} \cdot \boldsymbol{n} = e_{\boldsymbol{q}} \cdot \boldsymbol{n} + \boldsymbol{\beta} \cdot \boldsymbol{n}\ e_{\hat{\phi}} + \tau(e_\phi - e_{\hat{\phi}}) \quad \text{on} \quad \partial\mathcal{T}_h. \tag{4.1.3}$$

To prove this, let $K \in \mathcal{T}_h$ and $e \subseteq \partial K$. Then, by the definition of $e_{\boldsymbol{q}}$, e_ϕ, and $e_{\hat{\phi}}$, we have for all $\mu \in M_h$ that

$$\begin{aligned}
\langle e_{\boldsymbol{q}} \cdot \boldsymbol{n} + \boldsymbol{\beta} \cdot \boldsymbol{n}\ e_{\hat{\phi}} + \tau(e_\phi - e_{\hat{\phi}}), \mu_h\rangle_e &= \langle \Pi_V \boldsymbol{q} \cdot \boldsymbol{n} + \boldsymbol{\beta} \cdot \boldsymbol{n}\ P_M\phi + \tau(\Pi_W\phi - P_M\phi)\rangle_e \\
&\quad - \langle \boldsymbol{q}_h \cdot \boldsymbol{n} + \boldsymbol{\beta} \cdot \boldsymbol{n}\ \widehat{\phi}_h + \tau(\phi_h - \widehat{\phi}_h)\rangle_e \\
&= \langle \Pi_V \boldsymbol{q} \cdot \boldsymbol{n} + \boldsymbol{\beta} \cdot \boldsymbol{n}\ P_M\phi + \tau(\Pi_W\phi - P_M\phi)\rangle_e \\
&\quad - \langle \widehat{\boldsymbol{q}_h + \boldsymbol{\beta}\phi} \cdot \boldsymbol{n}, \mu_h\rangle_e.
\end{aligned}$$

Finally, from the third equation defining Π_h (3.3.6c) we notice that

$$\langle \Pi_V \boldsymbol{q} \cdot \boldsymbol{n} + \boldsymbol{\beta} \cdot \boldsymbol{n}\ P_M\phi + \tau(\Pi_W\phi - P_M\phi)\rangle_e = \langle P_M((\boldsymbol{q} + \boldsymbol{\beta}\phi) \cdot \boldsymbol{n}), \mu_h\rangle_e.$$

Now it follows directly from the definition that the projection of errors (4.1.1) satisfies

$$(\kappa^{-1} e_{\boldsymbol{q}}, \boldsymbol{r}_h)_{\mathcal{T}_h} - (e_\phi, \nabla \cdot \boldsymbol{r}_h)_{\mathcal{T}_h} + \langle \boldsymbol{r}_h \cdot \boldsymbol{n}, e_{\hat{\phi}}\rangle_{\partial\mathcal{T}_h} = -(\kappa^{-1}\boldsymbol{I}_{\boldsymbol{q}}, \boldsymbol{r}_h)_{\mathcal{T}_h}, \tag{4.1.4a}$$

$$-(e_{\boldsymbol{q}} + \boldsymbol{\beta} e_\phi, \nabla w_h)_{\mathcal{T}_h} + \langle \widehat{e} \cdot \boldsymbol{n}, w_h\rangle_{\partial\mathcal{T}_h} = 0, \tag{4.1.4b}$$

$$\langle \widehat{e} \cdot \boldsymbol{n}, \mu_h\rangle_{\partial\mathcal{T}_h \setminus \Gamma_{\text{in}}} = \langle \varphi(\widehat{\phi}_h) e_{\hat{\phi}} + \boldsymbol{\beta} \cdot \boldsymbol{n} e_{\hat{\phi}} + \Lambda, \mu_h\rangle_{\Gamma_N}, \tag{4.1.4c}$$

$$\langle e_{\hat{\phi}}, \mu_h\rangle_{\Gamma_{\text{in}}} = 0, \tag{4.1.4d}$$

for all $(\boldsymbol{r}_h, w_h, \mu_h) \in \boldsymbol{H}_h \times Q_h \times \mathrm{M}_h$, where $\Lambda := \left(\varphi(\phi) - \varphi(\widehat{\phi}_h)\right) P_M\phi + \varphi(\phi)\boldsymbol{I}_{\widehat{\phi}} + \boldsymbol{\beta} \cdot \boldsymbol{n}\boldsymbol{I}_{\widehat{\phi}}$.

Now we proceed as in the general HDG scheme (3.4.1), with $f = 0$ and $\boldsymbol{I} = -\boldsymbol{I}_{\boldsymbol{q}}$.

Let us define

$$\mathcal{E}_h := \left(\|\kappa^{-1/2} e_{\boldsymbol{q}}\|^2_{\mathcal{T}_h} + \|\tau_\beta^{1/2}(e_\phi - e_{\hat{\phi}})\|^2_{\partial\mathcal{T}_h} \right)^{1/2}. \tag{4.1.5}$$

Theorem 4.1.1. *Let us assume that the hypotheses of Corollaries 3.4.2 and 3.4.1 hold true. If $\phi|_\Sigma \in L^\infty(\Sigma)$, then*

$$\begin{aligned}\|\Lambda\|_\Sigma \leq & c_1 \left(\|\boldsymbol{I}_{\widehat{\phi}}\|_\Sigma + C_{eq} \|h^{-1/2} \boldsymbol{I}_{\widehat{\phi}}\|_\Sigma + \|\phi\|_{\infty,\Sigma} \right) \|e_{\hat{\phi}}\|_\Sigma \\ & + \left(|c_3| + c_1 \|\boldsymbol{I}_{\widehat{\phi}}\|_{\infty,\Sigma} + c_1 \|\phi\|_{\infty,\Sigma} + \|\boldsymbol{\beta}\cdot\boldsymbol{n}\|_{\infty,\Sigma} \right) \|\boldsymbol{I}_{\widehat{\phi}}\|_\Sigma,\end{aligned}$$

$$\mathcal{E}_h + C_\varphi^{1/2} \|e_{\hat{\phi}}\|_\Sigma \leq \|\kappa^{-1/2} \boldsymbol{I}_{\boldsymbol{q}}\|_{\mathcal{T}_h} + \frac{1}{\epsilon_\Lambda^{1/2}} \|\Lambda\|_\Sigma,$$

and

$$\|e_\phi\|_{\mathcal{T}_h} \leq C_d \left(h \|\kappa^{-1/2} e_{\boldsymbol{q}}\|_{\mathcal{T}_h} + h \|\boldsymbol{I}_{\boldsymbol{q}}\|_{\mathcal{T}_h} + \|\Lambda\|_\Sigma + h^{3/2} \|\varphi(\widehat{\phi}_h)\|_{\infty,\Sigma} \, \|e_{\hat{\phi}}\|_\Sigma \right).$$

Proof. The second and third inequalities follow from the energy identity in Corollary 3.4.2 and the duality estimate in Corollary 3.4.1. It remains to bound $\|\Lambda\|_\Sigma$ and we proceed as follows.

$$\begin{aligned}\|\Lambda\|_\Sigma &= \| \left(\varphi(\phi) - \varphi(\widehat{\phi}_h) \right) P_M \phi + \varphi(\phi) \boldsymbol{I}_{\widehat{\phi}} + \boldsymbol{\beta}\cdot\boldsymbol{n} \boldsymbol{I}_{\widehat{\phi}} \|_\Sigma \\ &\leq \| \left(\varphi(\phi) - \varphi(\widehat{\phi}_h) \right) P_M \phi \|_\Sigma + \| (\varphi(\phi) + \boldsymbol{\beta}\cdot\boldsymbol{n}) \boldsymbol{I}_{\widehat{\phi}} \|_\Sigma \\ &\leq \| \left(\varphi(\phi) - \varphi(\widehat{\phi}_h) \right) \boldsymbol{I}_{\widehat{\phi}} \|_\Sigma + \| \left(\varphi(\phi) - \varphi(\widehat{\phi}_h) \right) \phi \|_\Sigma + \| (\varphi(\phi) + \boldsymbol{\beta}\cdot\boldsymbol{n}) \boldsymbol{I}_{\widehat{\phi}} \|_\Sigma.\end{aligned}$$

We bound each of these three terms separately. First, by (3.5.7),

$$\begin{aligned}\| \left(\varphi(\phi) - \varphi(\widehat{\phi}_h) \right) \boldsymbol{I}_{\widehat{\phi}} \|_\Sigma \leq & c_1 \|h^{1/2} (\phi - \widehat{\phi}_h)\|_{\infty,\Sigma} \, \|h^{-1/2} \boldsymbol{I}_{\widehat{\phi}}\|_\Sigma \\ \leq & c_1 \|h^{1/2} \boldsymbol{I}_{\widehat{\phi}}\|_{\infty,\Sigma} \, \|h^{-1/2} \boldsymbol{I}_{\widehat{\phi}}\|_\Sigma + c_1 \|h^{-1/2} \boldsymbol{I}_{\widehat{\phi}}\|_\Sigma \, \|h^{1/2} e_{\hat{\phi}}\|_{\infty,\Sigma} \\ \leq & c_1 \|h^{1/2} \boldsymbol{I}_{\widehat{\phi}}\|_{\infty,\Sigma} \, \|h^{-1/2} \boldsymbol{I}_{\widehat{\phi}}\|_\Sigma + c_1 C_{eq} \|h^{-1/2} \boldsymbol{I}_{\widehat{\phi}}\|_\Sigma \, \|e_{\hat{\phi}}\|_\Sigma,\end{aligned}$$

where we have used (3.3.1). Second, by (3.5.7) ,

$$\|\Big(\varphi(\phi) - \varphi(\widehat{\phi}_h)\Big)\,\phi\|_\Sigma \leq c_1\|\phi - \widehat{\phi}_h\|_\Sigma\ \|\phi\|_{\infty,\Sigma} \leq c_1\|\phi\|_{\infty,\Sigma}\ \|\boldsymbol{I}_{\widehat{\phi}}\|_\Sigma + c_1\|\phi\|_{\infty,\Sigma}\ \|e_{\hat{\phi}}\|_\Sigma.$$

Finally, also by (3.5.7)

$$\|(\varphi(\phi) + \boldsymbol{\beta}\cdot\boldsymbol{n})\boldsymbol{I}_{\widehat{\phi}}\|_\Sigma \leq (|c_3| + c_1\|\phi\|_{\infty,\Sigma} + \|\boldsymbol{\beta}\cdot\boldsymbol{n}\|_{\infty,\Sigma})\,\|\boldsymbol{I}_{\widehat{\phi}}\|_\Sigma.$$

Then,

$$\begin{aligned}\|\Lambda\|_\Sigma \leq & c_1\left(C_{eq}\|h^{-1/2}\boldsymbol{I}_{\widehat{\phi}}\|_\Sigma + \|\phi\|_{\infty,\Sigma}\right)\|e_{\hat{\phi}}\|_\Sigma \\ & + \left(c_1\|\boldsymbol{I}_{\widehat{\phi}}\|_{\infty,\Sigma} + |c_3| + 2c_1\|\phi\|_{\infty,\Sigma} + \|\boldsymbol{\beta}\cdot\boldsymbol{n}\|_{\infty,\Sigma}\right)\|\boldsymbol{I}_{\widehat{\phi}}\|_\Sigma.\end{aligned}$$

□

Corollary 4.1.1. *In addition to the hypotheses of Corollaries 3.4.2 and 3.4.1, let us suppose that $\phi \in H^{k+2}(\Omega)$ and $\boldsymbol{q} \in \boldsymbol{H}^{k+1}(\Omega)$. There hold*

$$\begin{aligned}\|\boldsymbol{q} - \boldsymbol{q}_h\|_{\mathcal{T}_h} &\lesssim h^{k+1}\left(|\boldsymbol{q}|_{k+1,\Omega} + \|\phi\|_{k+2,\Omega}\right), \\ \|\phi - \widehat{\phi}_h\|_h + \|e_\phi\|_{\mathcal{T}_h} &\lesssim (h^{k+2} + c_1h^{k+1})\left(|\boldsymbol{q}|_{k+1,\Omega} + \|\phi\|_{k+2,\Omega} + \|\boldsymbol{\beta}\cdot\boldsymbol{n}\|_{\infty,\Sigma}\right), \\ \|\phi - \phi_h\|_{\mathcal{T}_h} &\lesssim h^{k+1}\left(|\boldsymbol{q}|_{k+1,\Omega} + \|\phi\|_{k+2,\Omega}\right).\end{aligned}$$

Proof. From properties of the L^2-projection (cf. (3.3.5)), for each $e \in \mathcal{E}_h$, we have

$$\|\boldsymbol{I}_{\widehat{\phi}}\|_e \lesssim h_e^{k+1}|\phi|_{k+1,e}.$$

Moreover, by (3.5.7) and (3.3.3),

$$\|(\varphi(\phi) + \boldsymbol{\beta}\cdot\boldsymbol{n})\boldsymbol{I}_{\widehat{\phi}}\|_\Sigma \lesssim (|c_3| + c_1C_{sob}\|\phi\|_{k+1,\Sigma} + \|\boldsymbol{\beta}\cdot\boldsymbol{n}\|_{\infty,\Sigma})\,h^{k+1}|\phi|_{k+1,\Sigma}.$$

In addition, by a scaling argument, (3.3.3), and (3.3.5), we have

$$\|\boldsymbol{I}_{\widehat{\phi}}\|_{\infty,\Sigma} \lesssim h^{k+1/2}|\phi|_{k+1,\Sigma}.$$

Then, thanks to the previous theorem,

$$\begin{aligned}\|\Lambda\|_{\Sigma} \leq & c_1 \left(C_{eq}h^{k+1/2}|\phi|_{k+1,\Sigma} + C_{sob}\|\phi\|_{k+1,\Sigma}\right) \|e_{\hat{\phi}}\|_{\Sigma} \\ & + \left(c_1 h^{k+1/2}|\phi|_{k+1,\Sigma} + |c_3| + 2c_1 C_{sob}\|\phi\|_{k+1,\Sigma} + \|\boldsymbol{\beta}\cdot\boldsymbol{n}\|_{\infty,\Sigma}\right) h^{k+1}\|\phi\|_{k+1,\Sigma}, \end{aligned} \tag{4.1.6}$$

and, for c_1 sufficiently small and the second estimate of the previous Theorem, we have

$$\mathcal{E}_h + \|e_{\hat{\phi}}\|_{\Sigma} \lesssim \|\kappa^{-1/2}\boldsymbol{I}_{\boldsymbol{q}}\|_{\mathcal{T}_h} + h^{k+1}\|\phi\|_{k+1,\Sigma} \lesssim \|\kappa^{-1/2}\boldsymbol{I}_{\boldsymbol{q}}\|_{\mathcal{T}_h} + h^{k+1}\|\phi\|_{k+2,\Omega},$$

since, by the continuous trace inequality (cf. [2]),

$$\|\phi\|^2_{k+1,\Sigma} = \sum_{|\alpha|\leq k+1} \|D^{\alpha}\phi\|^2_{\Sigma} \leq \sum_{|\alpha|\leq k+1} \|D^{\alpha}\phi\|^2_{\partial\Omega} \lesssim \sum_{|\alpha|\leq k+1} \|D^{\alpha}\phi\|^2_{1,\Omega} \lesssim \|\phi\|^2_{k+2,\Omega}. \tag{4.1.7}$$

Moreover, from (3.3.7), we have

$$\|\kappa^{-1/2}\boldsymbol{I}_{\boldsymbol{q}}\|_{\mathcal{T}_h} \lesssim h^{k+1}\left(|\boldsymbol{q}|_{k+1,\Omega} + |\phi|_{k+1,\Omega}\right)$$

and then

$$\mathcal{E}_h + \|e_{\hat{\phi}}\|_{\Sigma} \lesssim h^{k+1}\left(|\boldsymbol{q}|_{k+1,\Omega} + |\phi|_{k+2,\Omega}\right), \tag{4.1.8}$$

which implies that

$$\|\boldsymbol{q} - \boldsymbol{q}_h\|_{\mathcal{T}_h} \leq h^{k+1}\left(|\boldsymbol{q}|_{k+1,\Omega} + |\phi|_{k+2,\Omega}\right).$$

In addition, replacing (4.1.8) in (4.1.6), and using the fact that $\|\phi\|_{k+1,\Sigma} \lesssim \|\phi\|_{k+2,\Omega}$, we obtain

$$\begin{aligned}\|\Lambda\|_{\Sigma} \lesssim\ & c_1 h^{k+1}\|\phi\|_{k+2,\Omega}\left(|\boldsymbol{q}|_{k+1,\Omega}+\|\phi\|_{k+2,\Omega}\right)\\ &+\left((c_1+|c_3|)\|\phi\|_{k+2,\Omega}+\|\boldsymbol{\beta}\cdot\boldsymbol{n}\|_{\infty,\Sigma}\right)h^{k+1}\|\phi\|_{k+2,\Omega}\\ \lesssim\ & c_1 h^{k+1}\|\phi\|_{k+2,\Omega}\left(|\boldsymbol{q}|_{k+1,\Omega}+\|\phi\|_{k+2,\Omega}+\|\boldsymbol{\beta}\cdot\boldsymbol{n}\|_{\infty,\Sigma}\right). \end{aligned} \tag{4.1.9}$$

On the other hand, by the last estimate of the previous theorem,

$$\|e_\phi\|_{\mathcal{T}_h} \lesssim h\|\kappa^{-1/2}e_{\boldsymbol{q}}\|_{\mathcal{T}_h}+h\|\boldsymbol{I}_{\boldsymbol{q}}\|_{\mathcal{T}_h}+\|\Lambda\|_{\Sigma}+h^{3/2}\|\varphi(\widehat{\phi}_h)\|_{\infty,\Sigma}\ \|e_{\hat{\phi}}\|_{\Sigma}.$$

Now, thanks to assumption (3.4.11) and (4.1.7),

$$\begin{aligned} h\|\kappa^{-1/2}e_{\boldsymbol{q}}\|_{\mathcal{T}_h} &\lesssim h^{k+2}\left(|\boldsymbol{q}|_{k+1,\Omega}+\|\phi\|_{k+2,\Omega}\right),\\ h\|\boldsymbol{I}_{\boldsymbol{q}}\|_{\mathcal{T}_h} &\lesssim h^{k+2}\left(|\boldsymbol{q}|_{k+1,\Omega}+|\phi|_{k+1,\Omega}\right),\\ h^{3/2}\|\varphi(\widehat{\phi}_h)\|_{\infty,\Sigma}\|e_{\hat{\phi}}\| &\lesssim M_\infty h^{k+5/2}\left(|\boldsymbol{q}|_{k+1,\Omega}+\|\phi\|_{k+2,\Omega}\right)\end{aligned}$$

and, from the estimate over Λ in (4.1.9), it follows that

$$\begin{aligned}\|e_\phi\|_{\mathcal{T}_h} \lesssim& h^{k+2}\left(|\boldsymbol{q}|_{k+1,\Omega}+\|\phi\|_{k+2,\Omega}\right)+c_1h^{k+1}\|\phi\|_{k+2,\Omega}\left(|\boldsymbol{q}|_{k+1,\Omega}+\|\phi\|_{k+2,\Omega}+\|\boldsymbol{\beta}\cdot\boldsymbol{n}\|_{\infty,\Sigma}\right)\\ \lesssim& (h^{k+2}+c_1h^{k+1})\left(|\boldsymbol{q}|_{k+1,\Omega}+\|\phi\|_{k+2,\Omega}+\|\boldsymbol{\beta}\cdot\boldsymbol{n}\|_{\infty,\Sigma}\right),\end{aligned}$$

and

$$\|\phi-\phi_h\|_{\mathcal{T}_h} \lesssim h^{k+1}\left(|\boldsymbol{q}|_{k+1,\Omega}+\|\phi\|_{k+2,\Omega}\right).$$

Finally, from (3.4.9) we have

$$\|e_{\hat{\phi}}\|_h \lesssim h\|\kappa^{-1/2}e_{\boldsymbol{q}}\|_{\mathcal{T}_h}+\|e_\phi\|_{\mathcal{T}_h}+h\|\kappa^{-1}\boldsymbol{I}_{\boldsymbol{q}}\|_{\mathcal{T}_h}$$

which also implies that

$$\|\phi - \widehat{\phi}_h\|_h \lesssim \left(h^{k+2} + c_1 h^{k+1}\right)\left(|\boldsymbol{q}|_{k+1,\Omega} + \|\phi\|_{k+2,\Omega} + \|\boldsymbol{\beta}\cdot\boldsymbol{n}\|_{\infty,\Sigma}\right).$$

□

CHAPTER 5

Computational results

In this chapter, we present two examples to illustrate the performance of the HDG scheme. We begin by introducing some additional notation. The individual errors are denoted by

$$e(\boldsymbol{q}) := \|\boldsymbol{q} - \boldsymbol{q}_h\|_{\mathcal{T}_h}, \quad e(\phi) := \|\phi - \phi_h\|_{\mathcal{T}_h}, \quad e(\widehat{\phi}) := \|\phi - \widehat{\phi}_h\|_h.$$

In turn, the experimental orders of convergence are defined as:

$$r(*) := \frac{\log(e(*)/e'(*))}{\log(h/h')} \quad \forall * \in \{\boldsymbol{q}.\phi, \widehat{\phi}\},$$

where e and e' denote the errors computed on two consecutive meshes of sizes h and h', respectively.

We implemented the fixed-point scheme specified in (3.2.2) in MATLAB. That is, given an initial value ϕ^0 for the concentration in the membrane Σ, we compute the corresponding solution to the HDG scheme (3.2.2) and obtain a concentration ϕ^1. The process is repeated

until the following stopping criterion is achieved:

$$\|\alpha_1 - \alpha_0\|_2 / \|\alpha_0\|_2 < tol,$$

where α_1 and α_0 are the vectors associated with the degree of freedom ϕ^1 and ϕ^0, resp.; tol is a given tolerance and $\|\cdot\|_2$ is the Euclidean norm. In what follows, we show two manufactured solutions.

Example 1

In the first numerical test, we consider a domain Ω with $L = M = 1$, that is $\Omega = (0,1)^2$, and set the parameters $a_1 = a_3 = 10^{-4}$ and $\kappa = 1$; and the inlet concentration $\phi_{\text{in}} = \phi|_{\Gamma_{\text{in}}}$. We recall that $c_1 = a_1$, $c_2 = a_3\phi_{\text{in}} + a_1\phi_{\text{in}}^2$ and $c_3 = a_3 + 2a_1\phi_{\text{in}}$. In addition, we set the velocity $\boldsymbol{\beta}$ and the manufactured solution ϕ as:

$$\boldsymbol{\beta}(x,y) = \begin{pmatrix} x \\ -y \end{pmatrix}, \quad \phi(x,y) = \sin(x)\sin(y)$$

and, on each edge e, we take $\tau|_e = \max_e \boldsymbol{\beta} \cdot \boldsymbol{n} + 1$ such that conditions (S_1) and (S_2) are satisfied.

In order to use the above manufactured solution, we introduce artificial volumetric and boundary sources that make this solution satisfy the equations. The following table shows the history of convergence, where we have used a tolerance of 10^{-8} and an initial concentration $\phi^0 = 1$.

Since $c_1 = 10^{-4} < h$ and c_3 is also of order 10^{-4}, the conditions of Theorem 3.5.2 are fulfilled, guaranteeing the existence of a unique fixed point. Second, according to Corollary 4.1.1, we expect an order of convergence of h^{k+1} for the errors in $\boldsymbol{q}$ and ϕ. This is exactly what we observe in our numerical experiments reported in Table 5.1. Moreover, we see in this table superconvergence of h^{k+2} for the error of the numerical trace, which agrees with the result in Corollary 4.1.1, since $c_1 = 10^{-4} < h$. We point out that for $k = 3$ and $N = 12764$, the results

are affected by round-off errors.

k	N	h	iter	$\|\boldsymbol{q}-\boldsymbol{q}_h\|_{\mathcal{T}_h}$	$r(\boldsymbol{q})$	$\|\phi-\phi_h\|_{\mathcal{T}_h}$	$r(\phi)$	$\|\phi-\widehat{\phi}\|_h$	$r(\widehat{\phi})$
1	215	0.068199	4	5.92e-04	-	1.39e-04	-	5.13e-05	-
	798	0.035400	4	1.53e-04	2.06	3.81e-05	1.97	6.74e-06	3.09
	3207	0.017658	4	3.96e-05	1.95	9.75e-06	1.96	9.09e-07	2.88
	12764	0.008851	4	9.55e-06	2.06	2.43e-06	2.01	1.10e-07	3.05
2	215	0.068199	4	6.39e-06	-	2.63e-06	-	2.70e-07	-
	798	0.035400	4	9.20e-07	2.96	4.01e-07	2.87	1.70e-08	4.22
	3207	0.017658	4	1.17e-07	2.96	4.90e-08	3.02	1.20e-09	3.82
	12764	0.008851	4	1.48e-08	3.00	6.31e-09	2.97	6.87e-11	4.14
3	215	0.068199	4	9.04e-08	-	5.79e-08	-	3.07e-09	-
	798	0.035400	4	5.79e-09	4.19	3.70e-09	4.20	7.95e-11	5.57
	3207	0.017658	4	4.08e-10	3.81	2.44e-10	3.91	3.18e-12	4.63
	12764	0.008851	4	2.34e-11	4.14	1.47e-11	4.07	4.27e-13	2.91

Table 5.1: History of convergence for Example 1.

Example 2

The second test also considers the unit square domain $\Omega = (0,1)^2$, with $L = M = 1$. We maintain the same parameters as before: $a_1 = a_3 = 10^{-4}$ and $\kappa = 1$, with the inlet concentration set as $\phi_{\text{in}} = \phi|_{\Gamma_{\text{in}}}$, but now the velocity field and manufactured solution are given by:

$$\boldsymbol{\beta}(x,y) = \begin{pmatrix} \beta_{\text{in}} \\ 0 \end{pmatrix}, \quad \phi(x,y) = x^2\left(y(1-y) + \exp(-y/\sqrt{\varepsilon}) + \exp(-(1-y)/\sqrt{\varepsilon})\right),$$

where β_{in} stands for the inlet mean feed fluid velocity, and $\varepsilon := 1/\beta_{\text{in}}$. The stabilization parameter τ is defined analogously to the first test, ensuring compliance with assumptions (S_1) and (S_2).

As before, to use the above manufactured solution, we introduce artificial volumetric and boundary sources that make this solution satisfy the equations. The following tables show the history of convergence, where we have used a tolerance of 10^{-8} and an initial concentration $\phi^0 = 1$.

Given that $c_1 = 10^{-4} < h$ and c_3 is of the same order, the hypotheses of Theorem 3.5.2

remain valid, ensuring a unique fixed-point solution. In line with Corollary 4.1.1, the numerical errors in $\boldsymbol{q}$ and ϕ converge with order h^{k+1}, which is corroborated by the results , as we see in Table . Furthermore, the numerical trace error displays superconvergent behavior of order h^{k+2}, in agreement with the theory, once again due to $c_1 < h$. As in the previous example, we observe that for $k = 3$ and $N = 12764$, the accuracy is limited by round-off errors.

k	N	h	iter	$\|\boldsymbol{q} - \boldsymbol{q}_h\|_{\mathcal{T}_h}$	$r(\boldsymbol{q})$	$\|\phi - \phi_h\|_{\mathcal{T}_h}$	$r(\phi)$	$\|\phi - \widehat{\phi}\|_h$	$r(\widehat{\phi})$
1	215	0.068199	4	1.79e-03	-	3.93e-04	-	2.07e-04	-
	798	0.035400	4	4.94e-04	1.97	1.10e-04	1.94	2.89e-05	3.00
	3207	0.017658	4	1.26e-04	1.96	2.77e-05	1.98	3.79e-06	2.92
	12764	0.008851	4	3.21e-05	1.99	7.05e-06	1.98	4.81e-07	2.99
2	215	0.068199	4	8.94e-06	-	2.28e-06	-	3.14e-07	-
	798	0.035400	4	1.28e-06	2.96	3.43e-07	2.89	2.15e-08	4.09
	3207	0.017658	4	1.68e-07	2.92	4.34e-08	2.97	1.66e-09	3.69
	12764	0.008851	4	2.06e-08	3.04	5.51e-09	2.99	8.79e-11	4.25
3	215	0.068199	4	1.31e-07	-	6.47e-08	-	3.61e-09	-
	798	0.035400	4	8.04e-09	4.25	4.11e-09	4.20	1.06e-10	5.37
	3207	0.017658	4	6.00e-10	3.73	2.61e-10	3.96	5.57e-12	4.24
	12764	0.008851	4	3.32e-11	4.19	1.58e-11	4.06	1.98e-12	1.50

Table 5.2: History of convergence for Example 2 for $\beta_{\text{in}} = 1$.

CHAPTER 6

Conclusions and future work

In this chapter, we summarize the main contributions of this work and give possible future research directions.

6.1 Conclusions

In this work, we studied a mathematical model for salt concentration in reverse osmosis desalination processes, involving a convection-diffusion equation with nonlinear boundary conditions on the membrane surface.

We analyzed a mixed formulation for the continuous problem and proved its well-posedness using a fixed-point argument. Then, we introduced a Hybridizable Discontinuous Galerkin scheme to approximate the salt concentration. Under several assumptions, we proved the existence and uniqueness of the discrete solution by applying a fixed-point strategy, where the nonlinear boundary condition was linearized, leading to a Robin-type boundary condition. In particular, we showed that the scheme is well-posed provided the data of the problem and the

meshsize satisfy certain assumptions.

We also carried out an *a priori* error analysis and obtained optimal convergence rates. These theoretical results were confirmed by numerical experiments implemented in MATLAB, which demonstrated the expected performance of the scheme.

In summary, we established and analyzed an HDG scheme for a convection-diffusion equation with nonlinear boundary conditions, demonstrating stability and convergence under suitable assumptions on the data and mesh size.

6.2 Future Work

Based on the results of this work, some interesting directions for future research can be considered:

- Relax some of the assumptions made in the analysis of the HDG scheme. In particular, the assumption $c_1 h^{-1/2}$ sufficiency small, we believe can be relaxed according to our numerical results.
- Extend the methodology to the coupled system involving the Navier–Stokes equations.
- Motivated by the application to reverse osmosis process, the inlet concentration ϕ_{in} was taken to be constant. It is possible to consider more general settings for the analysis, for example allowing ϕ_{in} to be a general function. In this case, ϕ_{in} must first be properly extended to $\partial\Omega$ and then construct a $H^1(\Omega)$-function whose trace is ϕ_{in}.
- Develop an *a posteriori* error analysis for the salt concentration.

Bibliography

[1] Isaac Bermúdez, Jessika Camaño, Ricardo Oyarzúa, and Manuel Solano, *A conforming mixed finite element method for a coupled navier–stokes/transport system modeling reverse osmosis processes*, Computer Methods in Applied Mechanics and Engineering **433** (2025), 117527.

[2] Susanne C. Brenner and L. Ridgway Scott, *The mathematical theory of finite element methods*, Springer New York, 2008.

[3] Jessika Camaño, Carlos García, and Ricardo Oyarzúa, *Analysis of a momentum conservative mixed-FEM for the stationary Navier–Stokes problem*, Numerical Methods for Partial Differential Equations **37** (2021), no. 5, 2895–2923.

[4] Nicolás Carro, David Mora, and Jesus Vellojin, *A finite element model for concentration polarization and osmotic effects in a membrane channel*, International Journal for Numerical Methods in Fluids **96** (2024), no. 5, 601–625.

[5] Y. Chen and B. Cockburn, *Analysis of variable-degree HDG methods for convection-diffusion equations. Part i: general nonconforming meshes*, IMA Journal of Numerical Analysis **32** (2012), no. 4, 1267–1293.

[6] Bernardo Cockburn, Jayadeep Gopalakrishnan, and Francisco-Javier Sayas, *A projection-based error analysis of HDG methods*, Mathematics of Computation **79** (2010), no. 271, 1351–1367.

[7] Daniele Antonio Di Pietro and Alexandre Ern, *Mathematical aspects of discontinuous galerkin methods*, Springer Berlin Heidelberg, 2012.

[8] Joyner Eke, Ahmed Yusuf, Adewale Giwa, and Ahmed Sodiq, *The global status of desalination: An assessment of current desalination technologies, plants and capacity*, Desalination **495** (2020), 114633.

[9] Alexandre Ern and Jean-Luc Guermond, *Theory and Practice of Finite Elements*, Springer New York, 2004.

[10] Gabriel N. Gatica, *A Simple Introduction to the Mixed Finite Element Method: Theory and Applications*, Springer International Publishing, 2014.

[11] P. Grisvard, *Elliptic problems in nonsmooth domains*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, 1985.

[12] Clara Skuse, Alejandro Gallego-Schmid, Adisa Azapagic, and Patricia Gorgojo, *Can emerging membrane-based desalination technologies replace reverse osmosis?*, Desalination **500** (2021), 114844.