universität
wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„Virtual element methods for problems in acoustics and fluid dynamics“

verfasst von / submitted by

Dipl.-Ing. Alexander Pichler, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doktor der Naturwissenschaften (Dr. rer. nat.)

Wien, 2019 / Vienna, 2019

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 796 605 405
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Doktoratsstudium Naturwissenschaften: Mathematik
Betreut von / Supervisor:	Univ.-Prof. Ilaria Perugia, PhD

At this point, I want to thank my supervisor, Prof. Ilaria Perugia, for her patience, her support and her belief in me. Especially in times when things did not work out as expected, her positive attitude, together with her enormous knowledge and experience, helped me a lot to keep going and never give up. In particular, I am really grateful that she gave me the opportunity to be part of her fantastic group (thank you to all of you!). Beside the variety of new knowledge that I was able to gain during my doctoral studies, I also had the chance to get to know so many wonderful people and build new, also transnational, friendships. So, I can only say *Grazie mille per tutto!*

Here, I want to thank especially Dr. Lorenzo Mascotto, one of my collaborators. I will never forget the many fruitful, but sometimes also crazy discussions we had. For me, it was always a pleasure to work with you!

Additionally, my thanks go to my colleagues in the office for the good and productive working atmosphere. I really loved coming to work every day.

Moreover, I want to thank Prof. Lourenço Beirão da Veiga from the University of Milano-Bicocca that he gave me the possibility to visit him in Milano for 3 months and work together with him and Dr. Giuseppe Vacca, which even resulted in a joint paper. Here, my thanks also go to Dr. Franco Dassi, Michele Visinoni, and Prof. Alessandro Russo for their help whenever I needed anything. I enjoyed every second with all of you and you will always stay in my heart!

Finally, I also want to thank the FWF (project P 29197-N32) and the Doctoral Program (DK) through the FWF Project W1245 for their financial support, and the Vienna Doctoral School of Mathematics (VDS) for being a member.

Last but not least, my deepest gratitude goes to my family for constantly supporting and backing me in my decisions.

Vienna, August 2019

Abstract

This thesis deals with the design and analysis of virtual element methods (VEM) for problems in acoustics and fluid dynamics. By an implicit definition of the basis functions and suitable projectors onto spaces of known functions, such as polynomials or plane waves, VEM are capable of coping with general polytopal meshes and thus provide more freedom in mesh generation, in comparison to standard finite element methods. In particular, VEM allow an easier handling of complex domain and geometry data, as in reservoir simulations (including the presence of cracks), an automatic inclusion of hanging nodes, more efficient and easier adaptivity, and a higher robustness to mesh deformation.

As acoustic and fluid dynamics model problems, the Helmholtz problem and the miscible displacement of incompressible fluids in porous media, respectively, are considered.

For the Helmholtz problem, a VEM is introduced that additionally fulfills the Trefftz property, i.e. the employed basis functions belong to the kernel of the Helmholtz operator. This feature allows to reach a given accuracy with significantly less degrees of freedom than with standard (non-Trefftz) methods. Unlike other Trefftz methods in the Helmholtz literature, which typically employ fully discontinuous trial and test functions, the interelement continuity constraints are imposed here in a nonconforming sense. This allows the construction of an edgewise orthogonalization-and-filtering process, which significantly mitigates the ill-conditioning due to the plane wave basis and, at the same time, reduces the number of degrees of freedom without deteriorating the accuracy. Such a numerical recipe is not directly applicable to other methods, such as the plane wave VEM (a virtual version of a partition of unity method) or the discontinuous Galerkin method (DG), and renders, as shown in a variety of numerical test cases (including a study of dispersion and dissipation properties), the presented method utterly competitive when compared to these technologies, especially in the high-order case and when approximating highly oscillatory problems. The theoretical analysis of this method is carried out in an elegant framework, where the best approximation error for Trefftz VE functions can be bounded in terms of the best approximation error for piecewise discontinuous Trefftz functions, such as plane waves. The nonconforming Trefftz VEM approach is introduced and analyzed for the Laplace problem, before focusing on the full Helmholtz problem.

Regarding the fluid dynamics part, the model problem mentioned above can be described by a nonlinear coupling of an elliptic equation for the pressure with a parabolic one for the fluid concentration. Since the pressure appears in the concentration equation only through the corresponding velocity field, a mixed method can be chosen to approximate both pressure and velocity in the pressure equation simultaneously. A semidiscrete and a fully discrete formulation of this problem in the VE context are presented. Moreover, *a priori* error estimates are derived for the latter case, where as time discretization scheme, a computationally cheap choice, namely a backward Euler method that is explicit in the nonlinear terms, is made. The theoretical results are demonstrated in a series of numerical tests.

Kurzfassung

Diese Arbeit befasst sich mit der Einführung und Analyse von Virtuellen Elemente Methoden (VEM) für Problemstellungen in Akustik und Fluiddynamik. Implizite Definitionen der Basisfunktionen und geeignete Projektoren auf Räume von bekannten Funktionen, wie Polynome oder ebene Wellen, ermöglichen den Einsatz von allgemeinen polytopalen Netzen und bieten daher im Vergleich zu Finiten Elemente Methoden mehr Freiheit bei der Konstruktion von Zerlegungen des Grundgebietes. Dies inkludiert eine natürliche Verwendung von hängenden Knoten, mehr Flexibilität bei adaptiven Verfeinerungen und eine erhöhte Robustheit gegenüber Deformationen.

Als Modellprobleme in Akustik und Fluiddynamik werden das Helmholtz Problem und das Problem der Vermischung von inkompressiblen Fluiden in porösen Medien betrachtet.

Das erstere Problem betreffend wird eine VEM eingeführt, die zusätzlich über die Trefftz Eigenschaft verfügt, d.h. die verwendeten Basisfunktionen liegen im Kern des Helmholtz Operators. Dies gestattet, eine vorgegebene Approximationsgenauigkeit mit deutlich weniger Freiheitsgraden als mit Standardmethoden (nicht-Trefftz Methoden) zu erreichen. Anders als bei in der Literatur existierenden Trefftz Methoden, welche typischerweise stückweise stetige, aber global unstetige Basisfunktionen verwenden, werden hier die Stetigkeitsbedingungen entlang der Kanten in einem nichtkonformen Sinn festgesetzt. Dadurch ist es möglich, einen kantenweisen Orthogonalisierungs- und Filterprozess zu konstruieren, welcher zum einen die schlechte Konditionierung der Basis aus ebenen Wellen mildert, und zum anderen gleichzeitig die Anzahl der Freiheitsgrade ohne Genauigkeitsverlust reduziert. Ein solches Verfahren ist nicht direkt in anderen Methoden, wie der *plane wave VEM* (eine virtuelle Version einer Methode der Zerlegung der Eins) oder der DG (*discontinuous Galerkin*) Methode anwendbar. Zahlreiche numerische Experimente (inklusive einer Studie der Dispersions- und Dissipationseigenschaften) demonstrieren die Vorteile der vorgestellten Methode speziell im Fall einer großen Anzahl an verwendeten Basisfunktionen und für stark oszillierende Probleme. Die theoretische Analyse wird dabei in einem eleganten Rahmen ausgeführt, in welchem der Bestapproximationsfehler für Trefftz VE Funktionen durch jenen für stückweise stetige und global unstetige Trefftz Funktionen abgeschätzt werden kann. Zum besseren Verständnis wird zuerst eine nichtkonforme Trefftz VEM für das Laplace Problem eingeführt und analysiert.

Das zweite Problem kann durch eine nichtlineare Kopplung einer elliptischen Gleichung für den Druck mit einer parabolischen für die Konzentration beschrieben werden. Dadurch dass der Druck in der Gleichung für die Konzentration nur durch das entsprechende Geschwindigkeitsfeld auftritt, kann eine gemischte Methode zur gleichzeitigen Approximation von Druck und Geschwindigkeit in der Gleichung für den Druck verwendet werden. Dabei werden semidiskrete und volldiskrete Formulierungen dieses Problems im Kontext der VE vorgestellt. Darüber hinaus werden *a priori* Fehlerabschätzungen für den letzteren Fall hergeleitet, wobei als Zeitdiskretisierungsschema eine rechengünstige Methode, nämlich ein Rückwärts-Euler-Verfahren, welches explizit in den nichtlinearen Termen ist, gewählt wird. Die hergeleiteten theoretischen Resultate werden mittels numerischer Experimente überprüft.

Contents

Chapter 1

Introduction

In this chapter, after a motivation for the use of virtual element methods (VEM) in general and for the model problems to be considered throughout this thesis in Section 1.1, a short overview of the state-of-the-art methodologies and techniques for efficiently computing numerical approximations to the model problems is provided in Section 1.2. Finally, in Section 1.3, the outline of this thesis, together with the main ideas contained in each chapter, is reported.

1.1 Motivation

Throughout the last years, Galerkin methods based on polytopal decompositions of the computational domain for the numerical approximation of boundary value problems arising from physical applications have been attracting a vast amount of attention. In order to name a few, we mention the discontinuous Galerkin methods (DG) [11], the hybridized discontinuous Galerkin methods (HDG) [81], the hybrid high-order methods (HHO) [88], the mimetic finite difference methods (MFD) [42, 140], the high-order boundary element method based finite element methods (BEM based FEM) [163], and the virtual element methods (VEM) [31,36]. Among their advantages are a large flexibility in dealing with complex geometry data, an automatic inclusion of hanging nodes, more efficient and easier adaptivity, and a higher robustness to mesh deformation.

In this thesis, we focus on VEM. Introduced in 2013 in [31, 36] and already applied to a variety of model problems, such as general second-order elliptic problems [38], quasilinear elliptic problems [64], eigenvalue problems [107,152], the Stokes problem [8,44], the elasticity problem [35], the Cahn-Hilliard equation [9], biharmonic [13] and polyharmonic [12] problems, discrete fracture network simulations [49], and topology optimization [10,105], VEM can be seen as a generalization of standard polynomial based FEM to the case of general polytopal meshes and as the ultimate evolution of MFD. However, the main difference in comparison to FEM is that, for VEM, the basis functions are not known explicitly in closed form, but rather are solutions to local differential problems that mimic the target problem. This idea justifies the adjective *virtual* in the name of the method. In order to deal with such implicitly defined functions, suitable projectors from VE approximation spaces onto spaces of known functions need to be employed, and proper discrete bilinear/sesquilinear forms mimicking the continuous counterparts have to be constructed. All these tools have to be *computable*, i.e. they need to have the feature that they can be computed only in terms of a suitably chosen set of degrees of freedom.

This framework goes hand in hand with a series of advantages. For example, VEM can be directly extended to highly-regular [45] and nonconforming [18,69] approximation spaces, combined with domain decomposition techniques [53] and adaptive mesh refinement [67], and adapted to curved domains [47]. Moreover, it also allows the construction of spaces incorporating certain physical properties, such as divergence-free velocity spaces for Stokes [44], or the Trefftz property, as discussed in the Section 1.2 below.

The two problems considered in this thesis are a Helmholtz-type boundary value problem which models acoustic wave propagation in the time-harmonic case, and the miscible displacement of incompressible fluids in porous media, which is relevant in oil industry and is also used in modeling the environmental pollution. Both problems are described in detail in Section 2.2; here, we point out the main difficulties arising in their numerical approximation.

The Helmholtz problem can be written in an abstract way as

$$\begin{cases} -\Delta u - k^2 u = 0 & \text{in } \Omega \\ + \text{ boundary conditions} & \text{on } \Gamma := \partial\Omega, \end{cases}$$

where the quantity u, the so-called *phasor* of the acoustic pressure or of the acoustic velocity potential, should be determined for given *wave number* $k > 0$ and computational domain $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$. For high wave numbers, due to the oscillatory behavior of the analytical solutions, the numerical approximation by standard Galerkin based methods represents intrinsic difficulties, see e.g. [24, 86, 93]. This can already be seen when considering the following 1D problem, see also [86]:

$$\begin{cases} -u'' - k^2 u = 0 & \text{in } \Omega := (0, 1) \\ u'(1) - \mathrm{i}ku(1) = 0, & u(0) = 1. \end{cases} \tag{1.1}$$

The exact solution here is given by $u(x) = e^{\mathrm{i}kx}$. Let Ω be discretized into intervals all with the same length $h > 0$. We use standard lowest-order FEM and plot the real parts of the exact solution, its interpolant, and the numerical solution for $h = 0.01$ and different values of k in Figure 1.1.

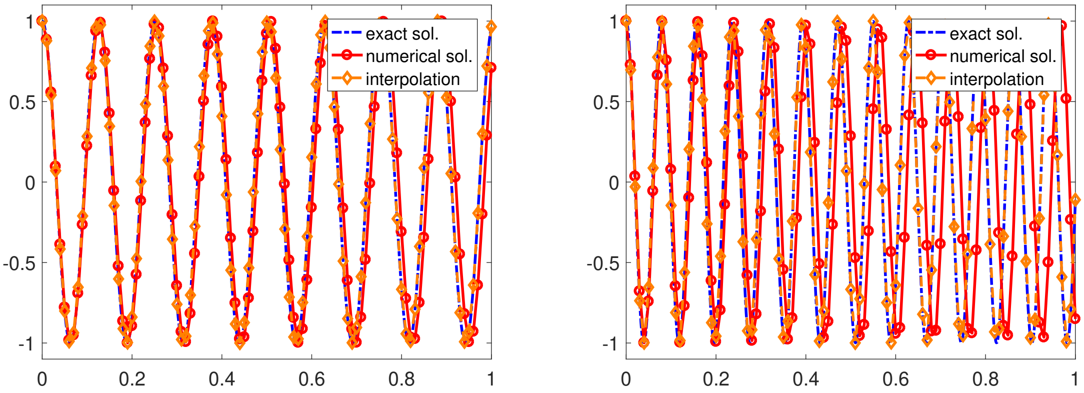


Figure 1.1: Real parts of the exact solution (*blue*), interpolated solution (*orange*), and numerical solution (*red*) to the 1D Helmholtz problem (1.1) for $h = 0.01$ with $k = 50$ (*left*) and 80 (*right*).

Moreover, in Figure 1.2, we display the relative discretization errors in the H^1 seminorm for the h-version and different values of k against the mesh size h, and the products kh and k^3h^2.

From both figures, we first observe that, for high wave numbers, the mesh size needs to be sufficiently small in order to obtain accurate numerical approximations. Furthermore, the numerical method fails to reproduce the correct oscillating behavior of the analytical solution; this phenomenon is called *numerical dispersion.* Next, there is a pre-asymptotic regime before convergence is achieved, which is wider for larger values of k. Finally, the condition that kh is small is not enough to guarantee to be in the convergence regime; rather the product k^3h^2 needs to be small. In fact, one can show for problem (1.1), when using lowest-order FEM, the error estimate [126]

$$\frac{|u - u_h|_{H^1}}{|u|_{H^1}} \leq C_1 kh + C_2 k^3 h^2, \quad kh < 1, \tag{1.2}$$

where C_1, C_2 are positive constants, and u_h is the numerical approximation of u. The first term on the right-hand side of (1.2) represents the best approximation error, and the second the pollution error. This widening discrepancy between the best approximation error and the discretization error for large values of k is also known as *pollution effect* in literature and cannot be avoided in higher dimensions, see [24].

Regarding the miscible displacement of one incompressible fluid by another in a porous medium, this problem can be expressed in abstract form by

$$\begin{cases} \phi \dfrac{\partial c}{\partial t} + \boldsymbol{u} \cdot \nabla c - \operatorname{div}(D(\boldsymbol{u})\nabla c) = f(c), \quad \operatorname{div} \boldsymbol{u} = G, \quad \boldsymbol{u} = h(c, p), & \text{in } \Omega \times [0, T], \\ \quad + \text{ boundary conditions} \quad + \text{ initial condition}, \end{cases} \tag{1.3}$$

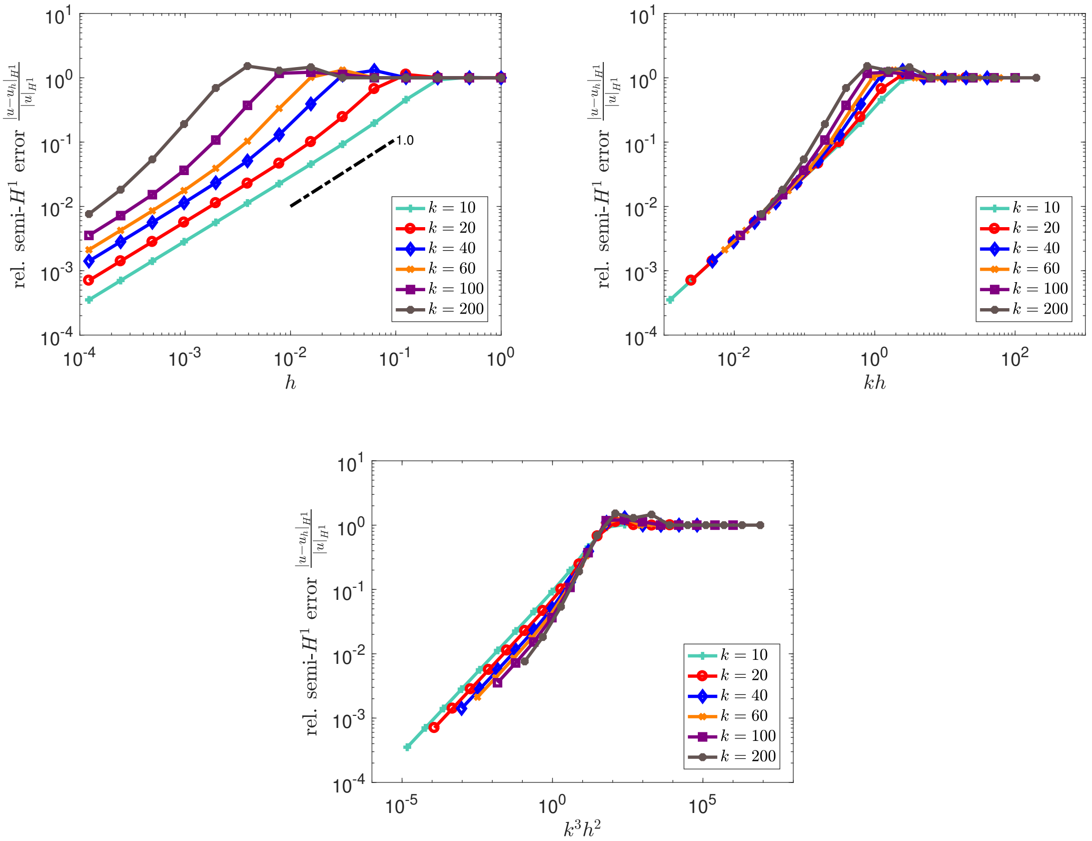


Figure 1.2: h-version of lowest-order FEM for the 1D Helmholtz problem (1.1) with different values of k. The relative discretization error in the H^1-seminorm is plotted against the mesh size h (*top-left*), against the product kh (*top-right*), and against k^3h^2 (*bottom*).

where $T > 0$, and one is interested in finding the velocity $\boldsymbol{u}$, the pressure p, and the concentration c, which are all coupled in a nonlinear fashion. We refer to Section 2.2.2 for further details.

Problem (1.3) is time-dependent, and thus numerical time integration techniques, such as the explicit or implicit Euler method, or the Crank-Nicolson method, have to be applied. In the case of explicit Euler, a *Courant-Friedrichs-Lewy (CFL) condition* linking the time stepping size to the spatial mesh size is needed to guarantee stability of the overall numerical scheme.

Moreover, the first equation in (1.3) is of diffusion-convection-reaction type and thus numerical instabilities are expected to occur in strongly convection-dominated situations, for which the equation nearly has hyperbolic character. More precisely, when the so-called *Péclet number*, which is defined as the ratio between convective strength and diffusive conductance, exceeds a critical value, spurious oscillations in the form of *over-* and *undershoots* in the numerical solution result in space, see Figure 1.3. To this purpose, several strategies have been proposed to mitigate these oscillations. Among them are, for FEM, the stabilizations by local projection [106] or by suitably constructed bubble functions [61, 103], and the Streamline Upwind Petrov-Galerkin method (SUPG) [62, 63, 108, 164]. In the framework of VEM, SUPG stabilization techniques have been discussed in e.g. [48, 52]. Additionally to all of those, flux-correcting transport schemes working at the algebraic level have been introduced, see e.g. [135, 136].

1.2 State-of-the-art

Concerning the Helmholtz problem, the aim of this thesis lies in the construction and analysis of a VEM incorporating the Trefftz property, i.e., given a discretization of the computational domain, the local approximation spaces consist of functions that belong elementwise to the kernel of the

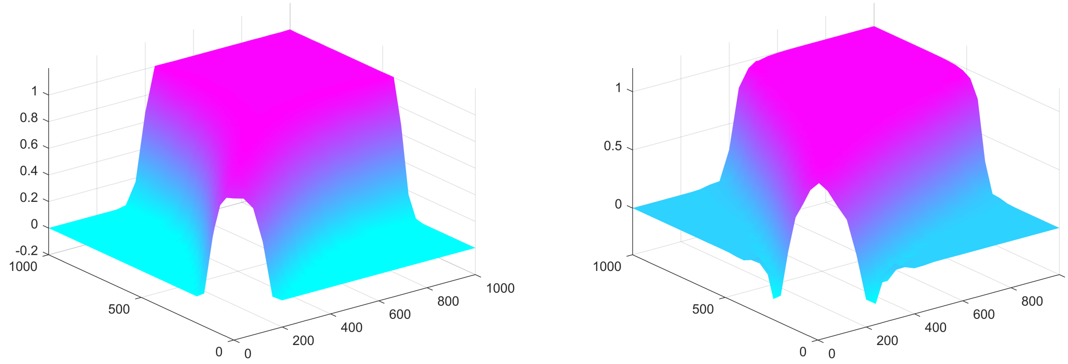


Figure 1.3: Numerical solutions for a strongly convection-dominated miscible displacement problem with an injection source at (1000,1000) and a production well at (0,0) (Test B in Example 2; Section 7.3) when using the VEM introduced in Section 7.1 endowed with a flux-correcting transport scheme (*left*; see also Figure 7.10) and without (*right*). In the right plot, one can clearly see strong oscillations and undershoots of the numerical solution around the origin.

Helmholtz operator. Representatives of Trefftz functions for Helmholtz are, for example, plane waves, evanescent waves, Fourier Bessel functions, fundamental solutions, and multipoles.

Named after Erich Trefftz (1888-1937) who first had the idea to incorporate *a priori* knowledge about the differential problem to be discretized into the approximation spaces [172], Trefftz methods for the Helmholtz problem look back on a long history. Here, we mention the ultra weak variational formulation (UWVF) [70–72,87,124], discontinuous methods based on Lagrange multipliers [101] and on least square formulations [151], the plane wave discontinuous Galerkin method (PWDG) [112,118,120], which can in fact be seen as a generalization of the UWVF, the wave based method [85], and the variational theory of complex rays [161]; we refer to [121] for an overview. The big advantage of such methods in comparison with non-Trefftz methods is that, when solving homogeneous problems, significantly less degrees of freedom are needed in order to achieve a given accuracy and thus also less computational effort is required. This comes however at the price of a possibly severe ill-conditioning.

For applications of Trefftz methods to the time-harmonic Maxwell equations, we refer to [70, 119,150]. We mention here that the Trefftz approach has also been applied to other problems than time-harmonic wave propagation, for instance to advection-diffusion problems [131,132], to wave problems in time-domain [25,94,133,134] and to Friedrichs systems [153].

In the context of VEM, the Helmholtz problem has already been tackled in a first approach in [159], giving rise to the plane wave VEM (PWVEM). This method is characterized by the use of plane waves that are modulated with lowest-order harmonic VE functions, and can be seen as a virtual version of the classical partition of unity method (PUM) [22]. Importantly, it is not a full Trefftz method.

The Trefftz VEM presented in this thesis for the 2D case is a full Trefftz method making use of plane waves. It further belongs to the class of *nonconforming* methods (à la Crouzeix-Raviart) as the interelement continuity constraints are imposed edgewise by requiring that the jumps across edges of the numerical solution against traces of plane waves are zero. In the context of VEM, nonconforming methods were first introduced in [18] for the Poisson problem, and then extended to the approximation of general elliptic problems in [69] and of the Stokes problem in [68].

In the numerical study of our new method, we also focus on the dispersion and dissipation errors. For the Helmholtz problem, in the framework of standard conforming FEM, a full dispersion analysis was carried out in [86] for dimensions one to three. In particular, in [24], it was shown that the pollution effect can be avoided in 1D, but not in higher dimensions, and a generalized pollution-free FEM in 1D was constructed. Moreover, in [125], a link between the results of the dispersion analysis and the numerical analysis was established, and, in [4], quantitative, fully explicit estimates for the behavior and decay rates of the dispersion error were derived in dependence on the order of the method relative to the mesh size and the wave number. In the context of non-conforming methods, dispersion analyses have been performed for DG-FEM in [5, 6], for the discontinuous Petrov-Galerkin method (DPG) in [113], and for UWVF/PWDG in [111]. Recently, a dispersion analysis for HDG was carried out in [114].

For the considered miscible displacement problem, which is sometimes referred to as *Peaceman model* after it was introduced by Peaceman and Rachford in 1962 in [157] (see also e.g. [29,76,156]), existence of solutions to that problem was studied under several assumptions; we mention the works [7, 77, 92, 102].

Moreover, a variety of numerical schemes and methodologies for the computation of approximate solutions to that problem has been proposed. In the framework of finite differences, we highlight the works [89,155,156]. For FEM, a mixed method was introduced in [90] to approximate both the velocity and the pressure, thus avoiding the differentiation of the latter quantity. In [123], a two-grid mixed FEM approach was employed, and, in [99,100], mixed FEM were combined with a modified method of characteristics. Additionally, Eulerian–Lagrangian localised adjoint methods were used in [176]. In the setting of mixed finite volume methods (MFV), the problem has been tackled in e.g. [73]. Discrete duality finite volume schemes for the model problem were introduced and studied in [74, 75]. Finally, in the setting of DG, we mention the works [28,162,168,177], and we note that a unified convergence analysis of numerical schemes was carried out in [91].

The VEM literature that has been relevant for our study includes the works on mixed VEM [30, 60], and on VEM for parabolic problems [174] and for general second-order elliptic problems [38].

1.3 Structure of the thesis

Here, we report the structure of this thesis and outline the main ideas of each chapter.

- Chapter 2: In this chapter, preliminaries related to the basic notation, the functional spaces, and the mesh assumptions are given. Moreover, the model problems considered throughout this thesis are introduced and described in detail.

- Chapter 3: As already mentioned above, one aim of this thesis is the design and analysis of a nonconforming Trefftz VEM for the Helmholtz problem. In this chapter, the focus lies on the simpler case of the Laplace problem, for which a nonconforming Trefftz VEM is presented. This method can be seen as the intermediate conformity level between the continuous harmonic VEM [79], which is a conforming Trefftz VEM for the Laplace problem, and the fully discontinuous harmonic DG-FEM [122,138,139].

  In addition to the construction of the method, its complete h- and p-version analysis is carried out, namely the study of its convergence behavior, when fixing the dimension of the local spaces and refining the mesh, and when fixing a mesh and increasing the dimension of the local spaces, respectively. Corresponding quasi-optimal error bounds that are explicit in terms of the mesh size h and of the degree of accuracy of the method p are derived.

  Moreover, the hp-version is studied numerically. In this case, suitable combinations of h- and p-refinements are taken. Similarly as for the harmonic VEM of [79] and the harmonic DG-FEM of [122], faster exponential convergence in terms of the number of degrees of freedom N than for standard FEM [19, 166] and VEM [39, 40] is observed, namely, $\exp(-b\sqrt[2]{N})$ instead of $\exp(-b\sqrt[3]{N})$, where b is a positive constant.

  This research has been published in [143].

- Chapter 4: This chapter, based on [144], deals with the construction and analysis of a nonconforming Trefftz VEM for the Helmholtz problem with impedance boundary conditions. By firstly defining the local Trefftz VE spaces as the spaces of Trefftz functions whose impedance traces are edgewise traces of plane waves, and then taking the degrees of freedom as Dirichlet moments with respect to plane waves, the global nonconforming Trefftz VE spaces can be built by gluing the local spaces together in a nonconforming way. In this sense, the presented method differs from most of the Trefftz methods in the Helmholtz literature, which typically employ fully discontinuous trial and test functions.

  Furthermore, the nonconforming approach provides, analogously as in the case of the Laplace problem, an elegant theoretical framework for the derivation of h-version error estimates.

More precisely, it allows to determine an upper bound of the best approximation error for Trefftz VE functions in terms of the best approximation error for piecewise discontinuous plane waves.

- Chapter 5: Here, numerical aspects of nonconforming Trefftz VEM for Helmholtz boundary value problems are in the spotlight. Firstly, the extension of the Trefftz VEM introduced in Chapter 4 to the case of mixed boundary conditions is discussed. This will be reflected in the definition of the nonconforming Trefftz VE spaces and the related discrete sesquilinear forms. Afterwards, implementation aspects are presented. Due to the severe ill-conditioning of the plane wave traces, a numerical recipe based on an edgewise orthogonalization-and-filtering procedure is introduced at the practical level to mitigate this ill-conditioning. At the same time, the presented strategy also allows to significantly reduce the number of degrees of freedom without deteriorating the convergence of the method.

  After testing the method endowed with this procedure in several numerical experiments, including an acoustic scattering problem, h-, p-, and hp-versions, a comparison of its performance with that of PWVEM [159] and UWVF/PWDG [71, 112] is carried out. Moreover, following the dispersion analysis for UWVF/PWDG in [111], a numerical study of the dispersion and dissipation properties is performed.

  This chapter is based on [145, 158].

- Chapter 6: In this chapter, following [146], the nonconforming Trefftz VEM introduced in Chapters 4 and 5 is extended to the case of the fluid-fluid interface problem, that is, a Helmholtz problem with piecewise constant wave number. To this purpose, local Trefftz VE spaces with possibly different wave number have to be coupled. Moreover, in order to capture the physical behavior of the analytical solutions, such local spaces are also enriched with additional special functions, like evanescent waves. For both issues, the nonconforming setting provides an elegant framework. Since the basis functions are related to edges, it suffices to simply consider, on each edge of the underlying mesh, the union of the restrictions of all plane waves and evanescent waves related to adjacent elements to the given edge, and then to apply the orthogonalization-and-filtering process mentioned above to get rid of almost linearly dependent basis functions and decrease the number of degrees of freedom.

  Numerical experiments with the h-version of the proposed method, the p-version with quasi-uniform meshes, and the hp-version with isotropic and anisotropic mesh refinements are presented.

- Chapter 7: This chapter deals with the fluid dynamics model problem. Following the related work in [46], after introducing suitable VE spaces and projectors, VEM for the semidiscrete (continuous in time and discrete in space) and fully discrete (in time and space) formulations of the miscible displacement problem, respectively, are discussed. Due to the fact that the quantities of interest, namely the velocity, pressure, and concentration, are coupled in a nonlinear way, a backward Euler method that is explicit in the nonlinear terms is chosen for the latter formulation. This approach is advantageous for two reasons: firstly, it leads to some sort of decoupling of the total system, in such a way that the resulting system is computationally cheap to be solved; secondly, no CFL condition is required to guarantee stability of the method.

  After deriving L^2 error estimates for the velocity, pressure, and concentration discretization errors, the focus lies on the investigation of the numerical performance of the method in a series of experiments. Among them are academic ones, as well as a more realistic, strongly convection-dominated one. For the latter test case, the presented VEM is further endowed with a flux-corrected transport scheme at the algebraic level, leading to a quite robust performance.

- Chapter 8: In this chapter, an outlook to future research topics is given.

Chapter 2

Preliminaries

In this chapter, we firstly fix the basic notation and introduce the functional spaces needed in the rest of this thesis in Section 2.1. Then, in Section 2.2, the considered model problems are described in detail. Next, in Section 2.3, the focus lies on the definition of regular polygonal decompositions of the computational domain. Finally, in Section 2.4, broken Sobolev spaces are introduced.

2.1 Basic notation and functional spaces

Here, we fix the notation and introduce the functional spaces employed throughout this thesis. For a more detailed discussion, we refer to [2, 59, 97, 147, 167].

First of all, we write $\mathbb{N}$ and $\mathbb{N}_0$ for the sets of natural numbers without and including 0, respectively. Given $r \in \mathbb{R}$, we denote by $\mathbb{R}_{>r}$ and $\mathbb{R}_{\geq r}$ the sets of all real numbers that are greater than, and greater than or equal to r, respectively. In addition, given $m \in \mathbb{N}$, we define $\mathbb{N}_{\geq m}$ as the set of all natural numbers that are greater than or equal to m. The set of complex numbers is denoted by $\mathbb{C}$, and the imaginary unit by i. Given a complex number $z \in \mathbb{C}$, $\mathrm{Re}z$ and $\mathrm{Im}z$ are the real and imaginary part of z, respectively.

Furthermore, given any domain $D \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$, i.e. an open, non-empty and connected subset of $\mathbb{R}^d$, and $\ell \in \mathbb{N}_0$, we denote by $\mathbb{P}_\ell(D)$ and $\mathbb{H}_\ell(D)$ the spaces of polynomials and harmonic polynomials up to order ℓ over D, respectively; moreover, we set $\mathbb{P}_{-1}(D) := \mathbb{H}_{-1}(D) := \emptyset$.

Next, we introduce *Lipschitz domains*. Let $\Omega \subset \mathbb{R}^d$ be a domain. Then, Ω is a *Lipschitz domain* if its boundary $\Gamma := \partial\Omega$ is compact and can be locally represented (after a possible rotation and translation) as the graph of a Lipschitz continuous function. If, additionally, Ω is bounded, then Ω is called *bounded Lipschitz domain*. By $\mathbf{n}_\Omega$, we denote the unit normal vector on $\partial\Omega$ pointing outside Ω.

Finally, $B_r(\mathbf{x}_0)$ is the ball centered at $\mathbf{x}_0 \in \mathbb{R}^d$ and with radius r.

Lebesgue spaces. The space of Lebesgue integrable functions on Ω to the power $p \in [1, \infty)$ is defined as

$$L^p(\Omega) := \{f : \|f\|_{L^p(\Omega)} < \infty\},$$

with the norm

$$\|f\|_{L^p(\Omega)} := \left(\int_\Omega |f|^p \,\mathrm{d}x\right)^{\frac{1}{p}}.$$

For $p = \infty$, $L^\infty(\Omega)$ is defined as the space of essentially bounded measurable functions with norm

$$\|f\|_{L^\infty(\Omega)} := \text{ess sup}\{|f(\mathbf{x})| : \mathbf{x} \in \Omega\} := \inf_{D\subset\Omega,\, \mu(D)=0} \sup_{\mathbf{x}\in\Omega\setminus D} |f(\mathbf{x})|,$$

where $\mu(D)$ here denotes the measure of D. The spaces $L^p(\Omega)$ are Banach spaces. For $p = 2$, the corresponding space, i.e. $L^2(\Omega)$, is a Hilbert space with inner product (for real-valued f, g), sesquilinear form (for complex-valued f, g), and norm given by

$$(f, g)_{0,\Omega} := \int_\Omega fg \,\mathrm{d}x, \qquad (f, g)_{0,\Omega} := \int_\Omega f\overline{g} \,\mathrm{d}x, \qquad \|f\|_{0,\Omega} := \left(\int_\Omega |f|^2 \,\mathrm{d}x\right)^{\frac{1}{2}},$$

respectively. From now on, we will focus on the real-valued case. We underline that the following concepts can be easily generalized to the complex-valued one. For $p = 2$, the following inequality holds.

Lemma 2.1.1 (Cauchy-Schwarz). *Let $f, g \in L^2(\Omega)$. Then, $fg \in L^1(\Omega)$ and*

$$\int_\Omega fg \, dx \leq \|f\|_{0,\Omega} \|g\|_{0,\Omega}.$$

Weak derivatives. We introduce the multi-index notation. For $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$, we write

$$|\boldsymbol{\alpha}| := \sum_{i=1}^d \alpha_i, \qquad \boldsymbol{\alpha}! := \alpha_1! \cdots \alpha_d!, \qquad \mathbf{x}^{\boldsymbol{\alpha}} := x_1^{\alpha_1} \cdots x_d^{\alpha_d},$$

for all $\mathbf{x} \in \Omega$, and, for $u : \Omega \to \mathbb{R}$,

$$D^{\boldsymbol{\alpha}} u(\mathbf{x}) := \partial_1^{\alpha_1} \cdots \partial_d^{\alpha_d} u(\mathbf{x}),$$

where $\partial_j := \frac{\partial}{\partial x_j}$ are the partial derivatives. The space $C^m(\Omega)$, $m \in \mathbb{N}_0 \cup \{\infty\}$, is the space of m-times continuously differentiable functions, and $C_0^m(\Omega)$ is the set of functions in $C^m(\Omega)$ with compact support. Denoting by $L^1_{\text{loc}}(\Omega)$ the space of locally integrable functions, and given $u \in L^1_{\text{loc}}(\Omega)$ and $\boldsymbol{\alpha} \in \mathbb{N}_0^d$, a function $v := D^{\boldsymbol{\alpha}} u \in L^1_{\text{loc}}(\Omega)$ is called *$\boldsymbol{\alpha}$-th order weak partial derivative* of u if

$$\int_\Omega v \, \varphi \, \mathrm{d}x = (-1)^{|\boldsymbol{\alpha}|} \int_\Omega u \, D^{\boldsymbol{\alpha}} \varphi \, \mathrm{d}x \quad \forall \varphi \in C_0^\infty(\Omega).$$

Sobolev spaces of non-negative integer order. Having this, the *Sobolev space* $W^{m,p}(\Omega)$, with $m \in \mathbb{N}_0$ and $p \in \mathbb{N} \cup \{\infty\}$, is defined as the set of all functions $u \in L^p(\Omega)$, such that the weak derivatives $D^{\boldsymbol{\alpha}} u \in L^p(\Omega)$ exist for all $|\boldsymbol{\alpha}| \leq m$:

$$W^{m,p}(\Omega) := \{u \in L^p(\Omega) : \, D^{\boldsymbol{\alpha}} u \in L^p(\Omega), \, |\boldsymbol{\alpha}| \leq m\}. \tag{2.1}$$

These spaces are equipped with the norms

$$\|u\|_{W^{m,p}(\Omega)} := \begin{cases} \left(\sum_{|\boldsymbol{\alpha}| \leq m} \|D^{\boldsymbol{\alpha}} u\|^p_{L^p(\Omega)}\right)^{\frac{1}{p}}, & 1 \leq p < \infty \\ \max_{|\boldsymbol{\alpha}| \leq m} \|D^{\boldsymbol{\alpha}} u\|_{L^\infty(\Omega)}, & p = \infty. \end{cases}$$

Alternatively, $W^{m,p}(\Omega)$ can be introduced by

$$W^{m,p}(\Omega) := \overline{C^\infty(\Omega)}^{\|\cdot\|_{W^{m,p}(\Omega)}},$$

i.e., for every $u \in W^{m,p}(\Omega)$, there exists a sequence $\{\varphi_j\}_{j \in \mathbb{N}} \subset C^\infty(\Omega)$ with

$$\lim_{j \to \infty} \|u - \varphi_j\|_{W^{m,p}(\Omega)} = 0.$$

For $p = 2$, we write

$$H^m(\Omega) := W^{m,2}(\Omega), \qquad \|u\|_{m,\Omega} := \|u\|_{W^{m,2}(\Omega)},$$

and we define the inner product and the seminorm

$$(u, v)_{m,\Omega} := \sum_{|\boldsymbol{\alpha}| \leq m} (D^{\boldsymbol{\alpha}} u, D^{\boldsymbol{\alpha}} v)_{0,\Omega}, \qquad |u|_{m,\Omega} := \left(\sum_{|\boldsymbol{\alpha}| = m} \|D^{\boldsymbol{\alpha}} u\|^2_{0,\Omega}\right)^{\frac{1}{2}}.$$

Moreover, given $k > 0$, the k-weighted Sobolev norm is

$$\|u\|^2_{m,k,\Omega} := \sum_{j=1}^m k^{2(m-j)} |u|^2_{j,\Omega}. \tag{2.2}$$

Sobolev spaces of positive fractional order. Going back to (2.1), one can also introduce Sobolev spaces of fractional order. More precisely, the *Sobolev-Slobodeckij* space $W^{\sigma,p}(\Omega)$, with $0 < \sigma < 1$ and $1 \leq p < \infty$, is given by

$$W^{\sigma,p}(\Omega) := \left\{ u \in L^p(\Omega) : \, |u|_{W^{\sigma,p}(\Omega)} < \infty \right\},$$

where the seminorm and norm are

$$|u|_{W^{\sigma,p}(\Omega)} := \left(\int_\Omega \int_\Omega \frac{|u(\mathbf{x}) - u(\mathbf{y})|^p}{|\mathbf{x} - \mathbf{y}|^{d+\sigma p}} \, \mathrm{d}x \, \mathrm{d}y \right)^{\frac{1}{p}}, \qquad \|u\|_{W^{\sigma,p}(\Omega)} := \left(\|u\|^p_{L^p(\Omega)} + |u|^p_{W^{\sigma,p}(\Omega)} \right)^{\frac{1}{p}}.$$

Given $s = m + \sigma$ with $m \in \mathbb{N}_0$ and $\sigma \in (0,1)$, we define

$$W^{s,p}(\Omega) := \{ u \in W^{m,p}(\Omega) : \, D^{\boldsymbol{\alpha}} u \in W^{\sigma,p}(\Omega), \, |\boldsymbol{\alpha}| \leq m \}.$$

The corresponding seminorm and norm are

$$|u|_{W^{s,p}(\Omega)} := \left(\sum_{|\boldsymbol{\alpha}|=m} |D^{\boldsymbol{\alpha}} u|^p_{W^{\sigma,p}(\Omega)} \right)^{\frac{1}{p}}, \qquad \|u\|_{W^{s,p}(\Omega)} := \left(\sum_{|\boldsymbol{\alpha}|\leq m} \|D^{\boldsymbol{\alpha}} u\|^p_{W^{\sigma,p}(\Omega)} \right)^{\frac{1}{p}}.$$

Similarly as above, for $p = 2$, we write

$$\begin{aligned} H^\sigma(\Omega) &:= W^{\sigma,2}(\Omega), & |u|_{\sigma,\Omega} &:= |u|_{W^{\sigma,2}(\Omega)}, & \|u\|_{\sigma,\Omega} &:= \|u\|_{W^{\sigma,2}(\Omega)} \\ H^s(\Omega) &:= W^{s,2}(\Omega), & |u|_{s,\Omega} &:= |u|_{W^{s,2}(\Omega)}, & \|u\|_{s,\Omega} &:= \|u\|_{W^{s,2}(\Omega)}. \end{aligned}$$

In this case, one can define again inner products.

Note that, alternatively, Sobolev spaces of fractional order can also be introduced via interpolation theory, see e.g. [173].

Sobolev spaces of negative order. Sobolev spaces with negative order can be defined by duality. More precisely, for $s < 0$ and $1 < p < \infty$, we have

$$W^{s,p}(\Omega) := [W_0^{-s,q}(\Omega)]',$$

where $'$ denotes the dual space, q is such that $\frac{1}{p} + \frac{1}{q} = 1$, and

$$W_0^{-s,q}(\Omega) := \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{W^{-s,q}(\Omega)}}.$$

The corresponding norm is

$$\|u\|_{W^{s,p}(\Omega)} := \sup_{\varphi \in W_0^{-s,q}(\Omega) \setminus \{0\}} \frac{\langle u, \varphi \rangle_\Omega}{\|\varphi\|_{W^{-s,q}(\Omega)}},$$

where $\langle \cdot, \cdot \rangle_\Omega$ denotes the *duality product.*

Sobolev spaces on the boundary. We focus on the case $p = 2$. Sobolev spaces $H^s(\Gamma)$ on the boundary Γ can be defined by using the following lemma, see e.g. [147, Thm. 3.37].

Lemma 2.1.2 (Trace theorem)**.** *Let Ω be a bounded Lipschitz domain. Then, there exists a bounded linear operator $\gamma : \, H^s(\Omega) \to H^{s-\frac{1}{2}}(\Gamma)$, for $s \in (\frac{1}{2}, \frac{3}{2})$, with*

$$\|u\|_{s-\frac{1}{2},\Gamma} \leq c_T \|u\|_{s,\Omega} \quad \forall u \in H^s(\Omega),$$

which is the unique extension of $\gamma u = u_{|_\Gamma}$. Moreover, there exists a continuous right inverse $\varepsilon : \, H^{s-\frac{1}{2}}(\Gamma) \to H^s(\Omega)$ with

$$\gamma \varepsilon w = w, \qquad \|\varepsilon w\|_{s,\Omega} \leq c_{IT} \|w\|_{s-\frac{1}{2},\Gamma},$$

for all $w \in H^{s-\frac{1}{2}}(\Gamma)$.

Thus, the Sobolev space $H^s(\Gamma)$, $0 < s < 1$, can be defined as the image of $H^{s+\frac{1}{2}}(\Omega)$ under γ with corresponding norm

$$\|u\|_{s,\Gamma} := \inf_{U \in H^{s+\frac{1}{2}}(\Omega):\, \gamma U = u} \|U\|_{s+\frac{1}{2},\Omega}.$$

If Ω is a $C^{k-1,1}$ domain, i.e., roughly speaking, a domain whose boundary can be described locally (after a possible rotation of the coordinate system) with functions whose k-th order derivatives are Lipschitz continuous, then γ is well-defined for $\frac{1}{2} < s \leq k$.

Alternatively, one can introduce $H^s(\Gamma)$ by requiring that the Sobolev-Slobodeckij norm

$$\|u\|_{s,\Gamma} := \left(\|u\|^2_{L^2(\Gamma)} + \int_\Gamma \int_\Gamma \frac{|u(\mathbf{x}) - u(\mathbf{y})|^2}{|\mathbf{x} - \mathbf{y}|^{d-1+2s}} \, \mathrm{d}s_x \, \mathrm{d}s_y \right)^{\frac{1}{2}}$$

is finite.

Furthermore, for $s < 0$, we define

$$H^s(\Gamma) := (H^{-s}(\Gamma))'$$

with the corresponding norm

$$\|u\|_{s,\Gamma} := \sup_{\varphi \in H^{-s}(\Gamma) \setminus \{0\}} \frac{\langle u, \varphi \rangle_\Gamma}{\|\varphi\|_{s,\Gamma}},$$

where $\langle \cdot, \cdot \rangle_\Gamma$ is again the duality pairing.

Since needed later on, here, we also highlight the definition of the $H^{\frac{1}{2}}(\Gamma)$ inner product for complex-valued functions in 2D:

$$(u, v)_{\frac{1}{2},\Gamma} = (u, v)_{0,\Gamma} + \int_\Gamma \int_\Gamma \frac{(u(\boldsymbol{\xi}) - u(\boldsymbol{\eta}))\overline{(v(\boldsymbol{\xi}) - v(\boldsymbol{\eta}))}}{|\boldsymbol{\xi} - \boldsymbol{\eta}|^2} d\boldsymbol{\xi} \, d\boldsymbol{\eta},$$

where $(\cdot, \cdot)_{0,\Gamma}$ is the $L^2(\Gamma)$ inner product and $|\cdot|$ denotes the Euclidean distance.

In addition, by $H^1_0(\Omega)$ and $H^1_g(\Omega)$, we denote the Sobolev spaces of H^1 functions with traces equal to zero and equal to a given function $g \in H^{\frac{1}{2}}(\Gamma)$, respectively.

Now, given $\Sigma \subset \Gamma$, we introduce the Sobolev space $H^s(\Sigma)$, for $0 \leq s < 1$, via

$$H^s(\Sigma) := \{v = \widetilde{v}_{|\Sigma} : \widetilde{v} \in H^s(\Gamma)\},$$

with the associated norm

$$\|v\|_{s,\Sigma} := \inf_{\widetilde{v} \in H^s(\Gamma):\, \widetilde{v}_{|\Sigma} = v} \|\widetilde{v}\|_{s,\Gamma}.$$

For $-1 < s < 0$, we define the corresponding Sobolev space again by duality:

$$H^s(\Sigma) := [H^{-s}_0(\Sigma)]',$$

where, denoting by supp the support,

$$H^{-s}_0(\Sigma) := \{v = \widetilde{v}_{|\Sigma} : \widetilde{v} \in H^{-s}(\Gamma), \operatorname{supp} \widetilde{v} \subset \Sigma\}.$$

Finally, Sobolev spaces over edges e can be defined by

$$H^m(e) := \{u \in L^2(e) : \|u\|_{m,e} < \infty\},$$

for $m \in \mathbb{N}$, where the norm is given by

$$\|u\|_{m,e} := \left(\sum_{j=1}^{m} \|\partial^j_t u\|^2_{0,e} \right)^{\frac{1}{2}},$$

with ∂^j_t denoting the j-th order tangential derivative along e. Sobolev spaces on e with fractional order are defined via interpolation theory, and such with negative orders via duality.

The $H^1(\Gamma)$ norm is defined as the square root of the sum of the squares of the $H^1(e)$ norms for all edges e belonging to Γ.

2.2 Model problems

In this section, we describe the model problems considered in this thesis.

2.2.1 Acoustic model problem

As acoustic model problems, we consider Helmholtz boundary value problems of the following type:

Given a (polygonal) bounded domain $\Omega \subset \mathbb{R}^2$ with boundary Γ split into

$$\Gamma = \overline{\Gamma_D} \cup \overline{\Gamma_N} \cup \overline{\Gamma_R}, \quad \Gamma_D \cap \Gamma_N = \emptyset, \quad \Gamma_D \cap \Gamma_R = \emptyset, \quad \Gamma_N \cap \Gamma_R = \emptyset, \quad |\Gamma_R| > 0, \tag{2.3}$$

and wave number $k > 0$ (with corresponding wave length $\lambda = \frac{2\pi}{k}$), the problem reads as

$$\begin{cases} \text{find } u \in H^1(\Omega) \text{ such that} & \\ \qquad -\Delta u - k^2 u = 0 & \text{in } \Omega \\ \qquad u = g_D & \text{on } \Gamma_D \\ \qquad \nabla u \cdot \mathbf{n}_\Omega = g_N & \text{on } \Gamma_N \\ \qquad \nabla u \cdot \mathbf{n}_\Omega + \mathrm{i}k\theta u = g_R & \text{on } \Gamma_R, \end{cases} \tag{2.4}$$

where $\theta \in \{-1, 1\}$, $g_D \in H^{\frac{1}{2}}(\Gamma_D)$, $g_N \in H^{-\frac{1}{2}}(\Gamma_N)$, and $g_R \in H^{-\frac{1}{2}}(\Gamma_R)$. The unknown u represents the *phasor* of the sound pressure or the acoustic velocity potential, see below.

Physical motivation. The Helmholtz equation can be seen as a special case of the general second-order, time-dependent, hyperbolic wave equation

$$-\Delta U(x,t) + \frac{1}{c^2}\frac{\partial^2}{\partial t^2}U(x,t) = 0, \tag{2.5}$$

where c is the speed of sound, and U is either the sound pressure or the acoustic velocity potential. More precisely, by plugging the time-harmonic ansatz

$$U(x,t) = \mathrm{Re}\{u(x)e^{-\mathrm{i}\omega t}\}$$

with angular frequency ω into (2.5) and defining $k := \frac{\omega}{c}$, the Helmholtz equation is obtained.

The model problem (2.4) is related to *acoustic scattering*: Given an obstacle D and a time-harmonic *incident field* u_{inc}, one is interested in computing the *total field* $u = u_{\mathrm{inc}} + u_{\mathrm{sca}}$, where u_{sca} is the *scattered field*. Typically, homogeneous Dirichlet (*sound-soft*) or homogeneous Neumann (*sound-hard*) boundary conditions are imposed on the boundary of the scatterer D. Thus, the associated problems to be solved for the total field are of the form

$$\begin{cases} -\Delta u - k^2 u = 0 & \text{in } \mathbb{R}^2\backslash\overline{D} \\ \qquad u = 0 & \text{on } \partial D \end{cases} \qquad \begin{cases} -\Delta u - k^2 u = 0 & \text{in } \mathbb{R}^2\backslash\overline{D} \\ \quad \nabla u \cdot \mathbf{n}_D = 0 & \text{on } \partial D. \end{cases}$$

The behavior of u_{sca} at infinity is described by the *Sommerfeld radiation condition*

$$\lim_{r\to\infty} r^{\frac{d-1}{2}} \left(\partial_r u_{\mathrm{sca}} - \mathrm{i}k u_{\mathrm{sca}}\right) = 0, \quad r = |\mathbf{x}|, \tag{2.6}$$

where ∂_r is the radial derivative; in our case, $d = 2$. This condition excludes that waves are reflected back from ∞ and guarantees unique solvability of the scattering problems. For $\theta = -1$, the so-called *impedance boundary condition*

$$\nabla u \cdot \mathbf{n}_\Omega + \mathrm{i}k\theta u = g_R$$

in (2.4) plays the role of a first-order approximation of (2.6).

In Figure 2.1, the real parts of the incident field, the scattered field, and the total field, respectively, are portrayed in the sound-soft case, where $D \subset \mathbb{R}^2$ is a circular scatterer and u_{inc} is a plane wave with wave number $k = 40$ traveling along $\mathbf{d} = [\cos(\varphi^{\mathrm{inc}}), \sin(\varphi^{\mathrm{inc}})]$ with $\varphi^{\mathrm{inc}} = \frac{\pi}{6}$.

Note that, for easier understanding, a Treffz VEM will be introduced and analyzed for the special case $\Gamma_D = \Gamma_N = \emptyset$ in Chapter 4 first, before tackling the generalization to the full problem (2.4) in Chapter 5.

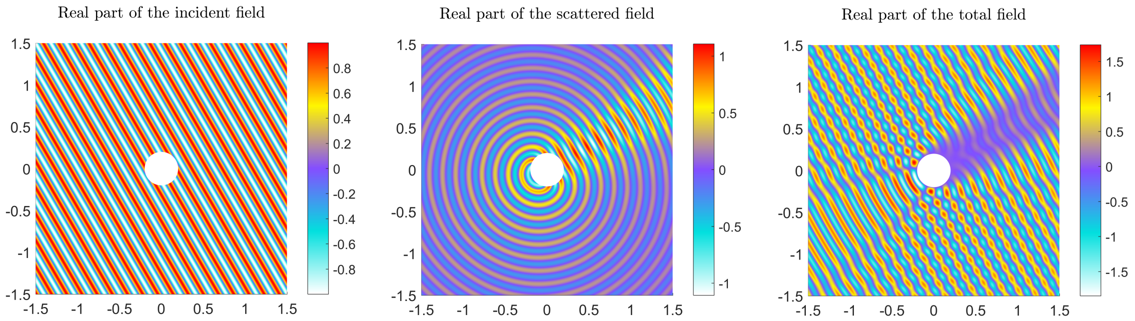


Figure 2.1: Acoustic scattering at a circular obstacle D. Real parts of the incident plane wave field with $k = 40$ (*left*), the corresponding scattered field (*center*), and the total field (*right*).

Weak formulation. The weak formulation of (2.4) reads

$$\begin{cases} \text{find } u \in V_{g_D} \text{ such that} \\ b_k(u,v) = F(v) \quad \forall v \in V_0, \end{cases} \tag{2.7}$$

where

$$\begin{aligned} V_{g_D} &:= H^1_{g_D,\Gamma_D}(\Omega) := \left\{ v \in H^1(\Omega) \,:\, v_{|\Gamma_D} = g_D \right\} \\ V_0 &:= H^1_{0,\Gamma_D}(\Omega) := \left\{ v \in H^1(\Omega) \,:\, v_{|\Gamma_D} = 0 \right\}, \end{aligned} \tag{2.8}$$

and where, for all $u, v \in H^1(\Omega)$,

$$b_k(u,v) := a_k(u,v) + \mathrm{i}k\theta \int_{\Gamma_R} u\overline{v}\,\mathrm{d}s, \quad F(v) := \int_{\Gamma_N} g_N \overline{v}\,\mathrm{d}s + \int_{\Gamma_R} g_R \overline{v}\,\mathrm{d}s, \tag{2.9}$$

with

$$a_k(u,v) := \int_\Omega \nabla u \cdot \overline{\nabla v}\,\mathrm{d}x - k^2 \int_\Omega u\overline{v}\,\mathrm{d}x.$$

Since we are assuming that $|\Gamma_R| > 0$, existence and uniqueness of solutions to the problem (2.4) follow from the Fredholm alternative and a continuation argument.

Theorem 2.2.1. *Under the assumption* (2.3) *on the polygonal domain* Ω*, problem* (2.4) *is uniquely solvable.*

Proof. We firstly note that the sesquilinear form $b_k(\cdot,\cdot)$ in (2.7) is continuous and satisfies a Gårding inequality [147, p.118], namely

$$\operatorname{Re}\{b_k(u,u) + 2k^2\|u\|^2_{0,\Omega}\} \geq \|u\|^2_{1,k,\Omega} \quad \forall u \in H^1(\Omega).$$

Owing to the Fredholm alternative [147, Thm. 4.11, 4.12], the problem (2.4) admits a unique solution if and only if the homogeneous adjoint problem to (2.4) with homogeneous boundary conditions, which is obtained by switching the sign in front of the term $\mathrm{i}k\theta u$ in the Robin boundary condition of the boundary value problem (2.4), admits only the trivial solution 0.
In order to show this, we consider the variational formulation of the homogeneous adjoint problem with homogeneous boundary conditions. After testing with $v = u$, and taking the imaginary part, we deduce $u = 0$ on Γ_R. Consequently, also $\nabla u \cdot \mathbf{n}_\Omega = 0$ on Γ_R, due to the definition of the impedance trace.
Let now $U \subset \mathbb{R}^2$ be an open, connected set such that $U \cap \partial\Omega = \Gamma_R$ and $\operatorname{meas}(U \backslash \overline{\Omega}) > 0$. We define $\widetilde{\Omega} := \Omega \cup U$ and $\widetilde{u} : \widetilde{\Omega} \to \mathbb{C}$ as the extension of u by zero in $\widetilde{\Omega} \setminus \Omega$. Then, $\widetilde{u}$ solves a homogeneous Helmholtz equation in $\widetilde{\Omega}$. Application of the unique continuation principle, see e.g. [15], leads to $\widetilde{u} = 0$ in $\widetilde{\Omega}$, and therefore $u = 0$ in Ω. □

To the best of our knowledge, elliptic regularity results and k-explicit stability estimates are not available so far for the solution to the general problem (2.4). Results have been derived only in specific cases, such as in the case of an interior impedance problem [96, Thm. 2.4], or a sound-soft acoustic scattering problem with a star-shaped scatterer [120, Prop. 2.1]. We highlight that the existence and the uniqueness of solutions can also be shown for Helmholtz-type boundary value problems with nonconstant wave number, see e.g. [115].

2.2.2 Fluid dynamics model problem

The model problem we consider here is the miscible displacement of one incompressible fluid by another in a porous medium. This problem can be formulated in terms of a system of partial differential equations, where a parabolic diffusion-convection-reaction type equation is nonlinearly coupled with an elliptic system, see [73, 90, 123, 156], and is of particular relevance in oil industry, but is also encountered in modeling the environmental pollution.

Given a (convex, polygonal) bounded domain $\Omega \subset \mathbb{R}^2$, playing the role of a reservoir of unit thickness, and given a time interval $J := [0, T]$, for $T > 0$, we are interested in finding $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x}, t)$, representing the Darcy velocity (volume of fluid flowing cross a unit across-section per unit time), the pressure $p = p(\boldsymbol{x}, t)$ in the fluid mixture, and the concentration $c = c(\boldsymbol{x}, t)$ of one of the two fluids (amount of the fluid per unit volume in the fluid mixture), with $(\boldsymbol{x}, t) \in \Omega_T := \Omega \times J$, such that

$$\begin{cases} \phi\,\dfrac{\partial c}{\partial t} + \boldsymbol{u} \cdot \nabla c - \operatorname{div}(D(\boldsymbol{u})\nabla c) = q^+(\widehat{c} - c) \\ \operatorname{div}\, \boldsymbol{u} = G \\ \boldsymbol{u} = -a(c)(\nabla p - \boldsymbol{\gamma}(c)), \end{cases} \tag{2.10}$$

where $\phi = \phi(\boldsymbol{x})$ is the porosity of the medium, $q^+ = q^+(\boldsymbol{x}, t)$ and $q^- = q^-(\boldsymbol{x}, t)$ are the (non-negative) injection and production source terms, respectively, $\widehat{c} = \widehat{c}(\boldsymbol{x}, t)$ is the concentration of the injected fluid, and

$$G := q^+ - q^-. \tag{2.11}$$

Moreover, $D(\boldsymbol{u}) \in \mathbb{R}^{2\times 2}$ is the diffusion tensor given by

$$D(\boldsymbol{u}) := \phi\left[d_m I + |\boldsymbol{u}|(d_\ell E(\boldsymbol{u}) + d_t E^\perp(\boldsymbol{u}))\right], \tag{2.12}$$

with matrices

$$E(\boldsymbol{u}) := \left(\frac{\boldsymbol{u}_i \boldsymbol{u}_j}{|\boldsymbol{u}|^2}\right)_{i,j=1,2} = \frac{\boldsymbol{u}\boldsymbol{u}^T}{|\boldsymbol{u}|^2}, \quad E^\perp(\boldsymbol{u}) := I - E(\boldsymbol{u}),$$

and molecular diffusion coefficient d_m, longitudinal dispersion coefficient d_ℓ, and transversal dispersion coefficient d_t. Furthermore, $\boldsymbol{\gamma}(c)$ in (2.10) describes the force density due to gravity (typically written as $\boldsymbol{\gamma}(c) = \gamma_0(c)\boldsymbol{\rho}$ with $\gamma_0(c)$ being the density of the fluid and $\boldsymbol{\rho}$ the gravitational acceleration vector), and $a(c) = a(c, \boldsymbol{x})$ is the scalar-valued function given by

$$a(c) := \frac{k}{\mu(c)},$$

where $k = k(\boldsymbol{x})$ represents the permeability of the porous rock, and $\mu(c)$ is the viscosity of the fluid mixture, which can be modeled by

$$\mu(c) = \mu(0)\left(1 + \left(M^{\frac{1}{4}} - 1\right)c\right)^{-4}, \quad \text{in } [0, 1],$$

with mobility ratio $M := \frac{\mu(0)}{\mu(1)}$. Note that μ can be set to $\mu(0)$ for $c < 0$, and to $\mu(1)$ for $c > 1$. We also highlight that, in the literature, k is sometimes assumed to be a tensor. The analysis carried out in Chapter 7 can be straightforwardly generalized to that case.

Assuming impermeability of $\partial\Omega$, the system (2.10) is then closed by requiring *no-flow boundary conditions* of the form

$$\begin{cases} \boldsymbol{u} \cdot \boldsymbol{n} = 0 & \text{on } \partial\Omega \times J \\ D(\boldsymbol{u})\nabla c \cdot \boldsymbol{n} = 0 & \text{on } \partial\Omega \times J, \end{cases} \tag{2.13}$$

and initial condition

$$c(\boldsymbol{x}, 0) = c_0(\boldsymbol{x}) \quad \text{in } \Omega, \tag{2.14}$$

where $0 \le c_0(\boldsymbol{x}) \le 1$ is an initial concentration. By use of the divergence theorem, the boundary conditions (2.13) directly imply the following compatibility condition for q^+ and q^-:

$$\int_\Omega q^+(\boldsymbol{x}, t)\,\mathrm{d}x = \int_\Omega q^-(\boldsymbol{x}, t)\,\mathrm{d}x,$$

for every $t \in J$.

We highlight that, in the theoretical analysis in Section 7.2, we will always assume sufficient regularity of the exact solution and the involved functions, such as q^+, q^-, $\widehat{c}$, etc., as motivated there. Moreover, we will make use of the following assumptions.

First of all, we suppose that the functions a and ϕ are positive and uniformly bounded from below and above, i.e. there exist positive constants a_*, a^*, ϕ_*, and ϕ^*, such that

$$a_* \le a(z, \boldsymbol{x}) \le a^*, \qquad \phi_* \le \phi(\boldsymbol{x}) \le \phi^*, \tag{2.15}$$

for all $\boldsymbol{x} \in \Omega$ and $z = z(t)$. For the sake of readability, we define

$$A(z)(\boldsymbol{x}) := a^{-1}(z, \boldsymbol{x}).$$

Additionally, we will make use of the following relation of the diffusion and dispersion coefficients, which was observed in laboratory experiments:

$$0 < d_m \le d_t \le d_\ell. \tag{2.16}$$

Finally, we recall that the source terms q^+ and q^- are, as usual, assumed to be non-negative functions.

Existence of weak solutions to this model problem was shown in [102] for $\boldsymbol{\gamma}(c) = 0$. An extension of this result to 3D spatial domains, including the presence of $\boldsymbol{\gamma}(c)$ and various boundary conditions, was discussed in [77].

Weak formulation. Here, we derive a weak formulation for the model problem described above. To this purpose, we firstly introduce the Sobolev space

$$H(\mathrm{div}; \Omega) := \{\boldsymbol{v} \in [L^2(\Omega)]^2 : \, \mathrm{div}\, \boldsymbol{v} \in L^2(\Omega)\}.$$

Then, we define the velocity space $\boldsymbol{V}$, the pressure space Q, and the concentration space Z by

$$\begin{aligned} \boldsymbol{V} &:= \{\boldsymbol{v} \in H(\mathrm{div}; \Omega) : \, \boldsymbol{v} \cdot \boldsymbol{n} = 0 \;\text{ on } \partial\Omega\} \\ Q &:= L^2_0(\Omega) := \{\varphi \in L^2(\Omega) : \, (\varphi, 1)_{0,\Omega} = 0\} \\ Z &:= H^1(\Omega), \end{aligned} \tag{2.17}$$

respectively. These spaces are endowed, respectively, with the following norms:

$$\|\boldsymbol{u}\|^2_{\boldsymbol{V}} := \|\boldsymbol{u}\|^2_{0,\Omega} + \|\mathrm{div}\, \boldsymbol{u}\|^2_{0,\Omega}, \qquad \|q\|^2_Q := \|q\|^2_{0,\Omega}, \qquad \|z\|^2_Z := \|z\|^2_{1,\Omega} := \|\nabla z\|^2_{0,\Omega} + \|z\|^2_{0,\Omega}.$$

Note that $\mathrm{div}\, \boldsymbol{V} = Q$.

As usual in the framework of parabolic problems, we use the notation

$$\boldsymbol{u}(t)(x) := \boldsymbol{u}(x, t), \qquad p(t)(x) := p(x, t), \qquad c(t)(x) := c(x, t). \tag{2.18}$$

For $0 \le a \le b$, we further introduce

$$\|\boldsymbol{v}\|_{L^2(a,b;\boldsymbol{V})} := \left(\int_a^b \|\boldsymbol{v}(t)\|^2_{\boldsymbol{V}}\,\mathrm{d}x \right)^{\frac{1}{2}}, \quad \|\boldsymbol{v}\|_{L^\infty(a,b;\boldsymbol{V})} := \operatorname*{ess\,sup}_{t\in[a,b]} \|\boldsymbol{v}(t)\|_{\boldsymbol{V}};$$

analogously for p and c.

Having this, the continuous problem reads as follows: find $c \in L^2(0,T;Z) \cap C^0([0,T];L^2(\Omega))$, $\boldsymbol{u} \in L^2(0,T;\boldsymbol{V})$, and $p \in L^2(0,T;Q)$, such that

$$\begin{cases} \mathcal{M}\left(\dfrac{\partial c(t)}{\partial t}, z\right) + (\boldsymbol{u}(t) \cdot \nabla c(t), z)_{0,\Omega} + \mathcal{D}(\boldsymbol{u}(t); c(t), z) = \left(q^+(\widehat{c} - c)(t), z\right)_{0,\Omega} \\ \qquad\qquad\qquad\qquad \mathcal{A}(c(t); \boldsymbol{u}(t), \boldsymbol{v}) + B(\boldsymbol{v}, p(t)) = (\boldsymbol{\gamma}(c(t)), \boldsymbol{v})_{0,\Omega} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad B(\boldsymbol{u}(t), q) = -\left(G(t), q\right)_{0,\Omega}, \end{cases} \tag{2.19}$$

for all $\boldsymbol{v} \in \boldsymbol{V}$, $q \in Q$, $z \in Z$, for almost all $t \in J$, and with initial condition $c(0) = c_0$, where

$$\begin{aligned} \mathcal{M}(c,z) &:= (\phi\, c, z)_{0,\Omega}, & \mathcal{D}(\boldsymbol{u}; c, z) &:= (D(\boldsymbol{u})\nabla c, \nabla z)_{0,\Omega}, \\ \mathcal{A}(c; \boldsymbol{u}, \boldsymbol{v}) &:= (A(c)\boldsymbol{u}, \boldsymbol{v})_{0,\Omega}, & B(\boldsymbol{v}, q) &:= -\left(\operatorname{div} \boldsymbol{v}, q\right)_{0,\Omega}. \end{aligned} \tag{2.20}$$

Note that $c \in L^2(0,T;Z) \cap C^0([0,T];L^2(\Omega))$ implies $\frac{\partial c}{\partial t} \in L^2(0,T;Z')$, see e.g. [160, Thm. 11.1.1].

For the sake of readability, we suppressed (t) in (2.20). From now on, we will use the convention that by writing $\boldsymbol{u}$, we mean in fact $\boldsymbol{u}(t)$; similarly for the other functions depending on space and time. In general it will be clear from the context whether $\boldsymbol{u}$ represents $\boldsymbol{u}(t)$ for a fixed $t \in J$, i.e. as a function of space only, or for varying $\boldsymbol{x}$ and t, as a function of both space and time.

Moreover, we will use the following alternative form for the concentration equation:

$$\mathcal{M}\left(\frac{\partial c}{\partial t}, z\right) + \Theta(\boldsymbol{u}, c; z) + \mathcal{D}(\boldsymbol{u}; c, z) = \left(q^+\, \widehat{c}, z\right)_{0,\Omega}, \tag{2.21}$$

where

$$\Theta(\boldsymbol{u}, c; z) := \frac{1}{2}\Big[(\boldsymbol{u} \cdot \nabla c, z)_{0,\Omega} + ((q^+ + q^-)\, c, z)_{0,\Omega} - (\boldsymbol{u}, c\,\nabla z)_{0,\Omega} \Big].$$

This version is obtained from the original one in (2.19) by rewriting the convective term as

$$(\boldsymbol{u} \cdot \nabla c, z)_{0,\Omega} = \frac{1}{2}\big[\, (\boldsymbol{u} \cdot \nabla c, z)_{0,\Omega} - (G, c\, z)_{0,\Omega} - (\boldsymbol{u}, c\,\nabla z)_{0,\Omega} \,\big],$$

where we firstly integrated by parts, then employed the fact that $\nabla \cdot \boldsymbol{u} = G$, together with the definition of G in (2.11), and afterwards combined this term with $(q^+\, c, z)_{0,\Omega}$ from the right-hand side of (2.19). This representation was inspired by the theory of VEM for general elliptic problems [69] and helps to ensure that properties of the continuous bilinear will be preserved after discretization.

In the rest of this section, we summarize some properties of the forms $\mathcal{M}(\cdot,\cdot)$, $\mathcal{A}(\cdot,\cdot,\cdot)$, and $\mathcal{D}(\cdot;\cdot,\cdot)$, all defined in (2.20), which will be needed later on.

To start with, for $\mathcal{M}(\cdot,\cdot)$, it directly holds with the Cauchy-Schwarz inequality and (2.15)

$$\mathcal{M}(c,z) \leq \phi^* \|c\|_{0,\Omega} \|z\|_{0,\Omega}, \qquad \mathcal{M}(z,z) \geq \phi_* \|z\|_{0,\Omega}^2,$$

for all $c, z \in Z$.

Concerning $\mathcal{A}(\cdot;\cdot,\cdot)$, again employing (2.15), for all $c \in L^\infty(\Omega)$ and $\boldsymbol{u}, \boldsymbol{v} \in [L^2(\Omega)]^2$, we have

$$\mathcal{A}(c; \boldsymbol{u}, \boldsymbol{v}) \leq \frac{1}{a_*} \|\boldsymbol{u}\|_{0,\Omega} \|\boldsymbol{v}\|_{0,\Omega}.$$

Further, if $c \in L^2(\Omega)$, $\boldsymbol{u} \in [L^\infty(\Omega)]^2$ and $\boldsymbol{v} \in [L^2(\Omega)]^2$, it holds true that

$$\mathcal{A}(c; \boldsymbol{u}, \boldsymbol{v}) \leq \|A(c)\|_{0,\Omega} \|\boldsymbol{u}\|_{\infty,\Omega} \|\boldsymbol{v}\|_{0,\Omega}.$$

We also have the coercivity bound

$$\mathcal{A}(c; \boldsymbol{v}, \boldsymbol{v}) \geq \frac{1}{a^*} \|\boldsymbol{v}\|_{0,\Omega}^2$$

for all $c \in L^\infty(\Omega)$ and $\boldsymbol{v} \in [L^2(\Omega)]^2$, from which, after defining the kernel

$$\mathcal{K} := \{\boldsymbol{v} \in \boldsymbol{V} : B(\boldsymbol{v}, q) = 0 \quad \forall q \in Q\}, \tag{2.22}$$

coercivity of $\mathcal{A}(c;\cdot,\cdot)$ on $\mathcal{K}$ in the norm $\|\cdot\|_V$ follows.

Regarding $\mathcal{D}(\cdot;\cdot,\cdot)$, the following continuity properties can be shown. Firstly, for all $\boldsymbol{u} \in [L^\infty(\Omega)]^2$ and $c, z \in H^1(\Omega)$, we have

$$\mathcal{D}(\boldsymbol{u}; c, z) \leq \phi^* \left[d_m + \|\boldsymbol{u}\|_{\infty,\Omega}(d_\ell + d_t)\right] \|\nabla c\|_{0,\Omega} \|\nabla z\|_{0,\Omega}, \tag{2.23}$$

which follows directly from the Cauchy-Schwarz inequality, the definition of $D(\boldsymbol{u})$ in (2.12), and the fact that $|E(\boldsymbol{u})\boldsymbol{v}| \leq |\boldsymbol{v}|$ and $|E^\perp(\boldsymbol{u})\boldsymbol{v}| \leq |\boldsymbol{v}|$ for all $\boldsymbol{v} \in \mathbb{R}^2$. Moreover, for all $\boldsymbol{u} \in [L^2(\Omega)]^2$ and $c, z \in H^1(\Omega)$ with $\nabla c \in L^\infty(\Omega)$, we have the bound

$$\mathcal{D}(\boldsymbol{u}; c, z) \leq \|D(\boldsymbol{u})\|_{0,\Omega} \|\nabla c\|_{\infty,\Omega} \|\nabla z\|_{0,\Omega} \leq \eta_{\mathcal{D}}(1 + \|\boldsymbol{u}\|_{0,\Omega}) \|\nabla c\|_{\infty,\Omega} \|\nabla z\|_{0,\Omega}, \tag{2.24}$$

with matrix norm $\|D(\boldsymbol{u})\|_{0,\Omega} := \left(\sum_{i,j=1}^2 \|D_{i,j}(\boldsymbol{u})\|_{0,\Omega}^2\right)^{\frac{1}{2}}$, and some positive constant $\eta_{\mathcal{D}}$ depending only on d_m, d_ℓ, and d_t. In addition, coercivity of $\mathcal{D}(\boldsymbol{u};\cdot,\cdot)$ for all $\boldsymbol{u} \in [L^\infty(\Omega)]^2$, with respect to $\|\cdot\|_{0,\Omega}$, follows from

$$\begin{aligned}
(D(\boldsymbol{u})\,\boldsymbol{\mu}, \boldsymbol{\mu})_{0,\Omega} &= (\phi\, d_m\, \boldsymbol{\mu}, \boldsymbol{\mu})_{0,\Omega} + (\phi\, |\boldsymbol{u}|\, (d_\ell E(\boldsymbol{u}) + d_t E^\perp(\boldsymbol{u}))\, \boldsymbol{\mu}, \boldsymbol{\mu})_{0,\Omega} \\
&\geq \phi_*\, d_m\, \|\boldsymbol{\mu}\|_{0,\Omega}^2 + (\phi\, |\boldsymbol{u}|(d_\ell - d_t) E(\boldsymbol{u})\boldsymbol{\mu}, \boldsymbol{\mu})_{0,\Omega} + (\phi\, |\boldsymbol{u}| d_t\, \boldsymbol{\mu}, \boldsymbol{\mu})_{0,\Omega} \\
&\geq \phi_* \left(d_m\, \|\boldsymbol{\mu}\|_{0,\Omega}^2 + d_t\, \||\boldsymbol{u}|^{\frac{1}{2}} \boldsymbol{\mu}\|_{0,\Omega}^2\right)
\end{aligned} \tag{2.25}$$

for all $\boldsymbol{\mu} \in [L^2(\Omega)]^2$, where we also employed (2.15) and (2.16).

2.3 Regular polygonal decompositions

We restrict here to the 2D case. Given a polygonal bounded domain $\Omega \subset \mathbb{R}^2$, we introduce here the concept of regular sequences of polygonal decompositions of Ω.

To this purpose, let $\{\mathcal{T}_n\}_{n\in\mathbb{N}}$ be a sequence of *conforming* polygonal decompositions of Ω, i.e., for each $n \in \mathbb{N}$, every internal edge e of $\mathcal{T}_n$ is contained in the boundary of precisely two elements in the decomposition. Note that this definition also allows for elements that have adjacent edges lying on the same line.

For all $n \in \mathbb{N}$, with each $\mathcal{T}_n$, we associate $\mathcal{E}_n$, $\mathcal{E}_n^I$ and $\mathcal{E}_n^B$, denoting the set of its edges, internal edges, and boundary edges, respectively.

Moreover, with each element K of $\mathcal{T}_n$, we associate $\mathcal{E}^K$, the set of its edges. For all $K \in \mathcal{T}_n$ and for all $n \in \mathbb{N}$, we set

$$h_K := \operatorname{diam}(K), \quad n_K := \operatorname{card}(\mathcal{E}^K),$$

and we denote by $\mathbf{x}_K$ the centroid of K. The normal vector pointing outward of K is denoted by $\mathbf{n}_K$. Finally, the mesh size of $\mathcal{T}_n$ is defined by

$$h := \max_{K\in\mathcal{T}_n} h_K.$$

Given an edge $e \in \mathcal{E}_n$, h_e denotes its length.

The sequence $\{\mathcal{T}_n\}_{n\in\mathbb{N}}$ is called a *regular sequence of polygonal decompositions* if the following assumptions are satisfied:

(**G1**) (*uniform star-shapedness*) there exist $\rho \in (0, \frac{1}{2}]$, $0 < \rho_0 \leq \rho$, such that, for all $n \in \mathbb{N}$ and for all $K \in \mathcal{T}_n$, there exist points $\mathbf{x}_{0,K} \in K$ for which the ball $B_{\rho h_K}(\mathbf{x}_{0,K})$ is contained in K, and K is star-shaped with respect to $B_{\rho_0 h_K}(\mathbf{x}_{0,K})$;

(**G2**) (*uniformly non-degenerating edges*) for all $n \in \mathbb{N}$ and for all $K \in \mathcal{T}_n$, it holds $h_e \geq \rho_0 h_K$ for all edges e of K, where ρ_0 is the same constant as in (**G1**).

The assumptions (**G1**) and (**G2**) imply the following property:

(**G3**) (*uniform boundedness of the number of edges*) there exists a constant $\Lambda \in \mathbb{N}$ such that, for all $n \in \mathbb{N}$ and for all $K \in \mathcal{T}_n$, card($\mathcal{E}^K$)$\leq \Lambda$, that is, the number of edges of each element is uniformly bounded.

We point out that, in this definition, we are not requiring any quasi-uniformity on the size of the elements. The above assumptions are standard in the VE literature and are needed for the theoretical error analysis. A discussion of VEM under more general mesh assumptions is the topic of [43, 58]. Some examples of meshes fulfilling (**G1**)-(**G3**) are shown in Figure 2.2 (left, center). The mesh in Figure 2.2 (right) does not satisfy (**G1**).

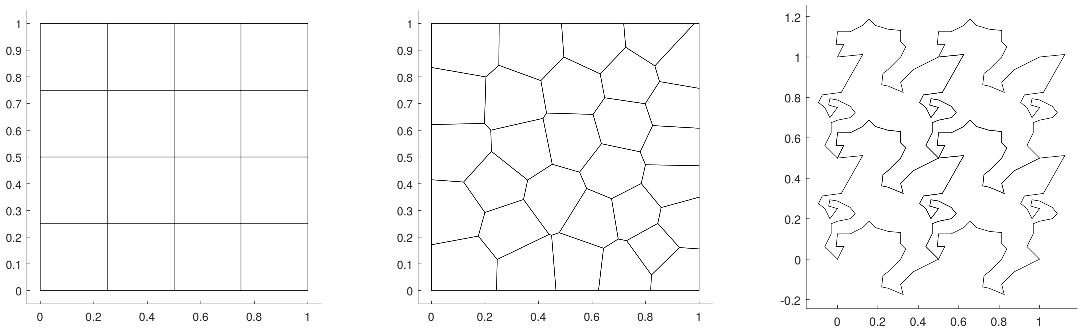


Figure 2.2: Examples of polygonal decompositions: regular Cartesian mesh (*left*), Voronoi-Lloyd mesh [169] (*center*), and mesh made of Escher horses (*right*). Whereas the former two meshes satisfy (**G1**)-(**G3**), the Escher mesh does not fulfil (**G1**).

Remark 1. For the theoretical analyses of the different methods presented in this thesis, a number of standard functional inequalities (such as the Poincaré inequality and trace inequalities) will be employed. It can be proven that the constants appearing in such inequalities depend solely on the parameters ρ_0 and Λ introduced in (**G1**)-(**G3**). For ease of notation, such a dependence will be omitted.

Important inequalities. Here, we recall some useful inequalities. Let $K \in \mathcal{T}_n$ be a fixed mesh element. The first inequality is the *multiplicative trace inequality*, see e.g. [59, Thm. 1.6.6],

$$\|v\|_{0,\partial K}^2 \leq C_M \|v\|_{0,K} \|v\|_{1,K} \quad \forall v \in H^1(K),$$

which also implies

$$\|v\|_{0,\partial K}^2 \leq C_M \|v\|_{0,K} \left(h_K^{-1} \|v\|_{0,K} + |v|_{1,K} \right) \quad \forall v \in H^1(K), \tag{2.26}$$

where C_M may differ in the two equations, but depends solely on the shape of K. From this, the following *trace inequality* follows, see also e.g. [104]:

$$\|v\|_{0,\partial K}^2 \leq C_T \left(h_K^{-1} \|v\|_{0,K}^2 + h_K |v|_{1,K}^2 \right) \quad \forall v \in H^1(K), \tag{2.27}$$

where C_T again depends solely on the shape of K. Furthermore, we recall the following Poincaré-Friedrichs inequalities, see e.g. [57]:

$$\|\xi\|_{0,K} \leq C_P h_K \left(|\xi|_{1,K} + \mathrm{meas}(\Upsilon)^{-1} \left| \int_\Upsilon \xi \, \mathrm{d}s \right| \right) \quad \forall \xi \in H^1(K), \tag{2.28}$$

$$\|\xi\|_{0,K} \leq C_P h_K \left(|\xi|_{1,K} + h_K^{-2} \left| \int_K \xi \, \mathrm{d}x \right| \right) \quad \forall \xi \in H^1(K), \tag{2.29}$$

where Υ is a measurable subset of ∂K with 1D positive measure, and $C_P > 0$ depends only on the shape of K.

In the sequel, when writing $a \lesssim b$, we mean that there exists a constant $c > 0$, independent of the discretization parameters and of the problem data, such that $a \leqslant c\,b$. For $a \lesssim b$ and $b \lesssim a$ simultaneously, we use $a \approx b$.

2.4 Broken Sobolev spaces

We introduce the broken Sobolev spaces of order $s > 0$, subordinated to a decomposition $\mathcal{T}_n$, by

$$H^s(\mathcal{T}_n) := \prod_{K \in \mathcal{T}_n} H^s(K) = \{v \in L^2(\Omega) : v_{|_K} \in H^s(K) \quad \forall K \in \mathcal{T}_n\}, \tag{2.30}$$

which are equipped with the corresponding broken seminorms and norms

$$|v|^2_{s,\mathcal{T}_n} := \sum_{K \in \mathcal{T}_n} |v|^2_{s,K}, \qquad \|v\|^2_{s,\mathcal{T}_n} := \sum_{K \in \mathcal{T}_n} \|v\|^2_{s,K}. \tag{2.31}$$

Particular emphasis is given to the broken H^1 bilinear form

$$(u, v)_{1,\mathcal{T}_n} := \sum_{K \in \mathcal{T}_n} (\nabla u, \nabla v)_{0,K}.$$

For the Helmholtz problem with wave number $k > 0$, it is more suitable to endow (2.30) with the weighted broken Sobolev norm

$$\|v\|^2_{s,k,\mathcal{T}_n} := \sum_{K \in \mathcal{T}_n} \|v\|^2_{s,k,K},$$

instead of $\|v\|^2_{s,\mathcal{T}_n}$, where, for every $K \in \mathcal{T}_n$, $\|v\|^2_{s,k,K}$ was defined in (2.2).

Chapter 3

Trefftz virtual element method for the Laplace problem

As mentioned in the introduction, our aim is to design and analyze a Trefftz VEM for the acoustic model problem (2.4). In contrast to "standard" methods, the interelement continuity constraints will be imposed in a nonconforming fashion. This leads to a series of advantages, as we will see in Chapters 4 and 5. Although not being an acoustic problem, it is useful to firstly consider a Laplace problem of the type

$$\begin{cases} -\Delta u = 0 & \text{in } \Omega \\ \quad u = g & \text{on } \Gamma, \end{cases} \tag{3.1}$$

where $\Omega \subset \mathbb{R}^2$ is a polygonal bounded domain with boundary $\Gamma := \partial\Omega$, and $g \in H^{\frac{1}{2}}(\Gamma)$ is a given boundary datum, since the construction and analysis there already give valuable insight into how such a method can be designed for the full Helmholtz problem.

The outline of this chapter is as follows. After designing a nonconforming Trefftz VEM for the Laplace problem (3.1) in Section 3.1, an abstract error analysis, including the derivation of h- and p-error estimates, is performed in Section 3.2. Finally, in Section 3.3, details on the implementation of the method are given and numerical results validating the theoretical estimates are presented.

The contents of this chapter have been published in [143], and can be seen as the counterpart of the conforming harmonic VEM [79] in the nonconforming framework.

3.1 Nonconforming Trefftz virtual element methods

In this section, we design a nonconforming Trefftz VEM for the approximation of the model problem (3.1).

To this purpose, we firstly state the weak formulation of (3.1), which reads

$$\begin{cases} \text{find } u \in V_g \text{ such that} \\ a_0(u,v) = 0 \quad \forall v \in V_0, \end{cases} \tag{3.2}$$

where

$$a_0(u,v) := (\nabla u, \nabla v)_{0,\Omega}, \quad V_g := H^1_g(\Omega), \quad V_0 := H^1_0(\Omega). \tag{3.3}$$

Well-posedness of problem (3.2) follows from a lifting argument and the Lax-Milgram lemma.

Given a mesh $\mathcal{T}_n$ as described in Section 2.3, our aim is to approximate problem (3.2) with a method of the following type:

$$\begin{cases} \text{find } u_n \in V^{\Delta,p}_{n,g} \text{ such that} \\ a_{0,n}(u_n, v_n) = 0 \quad \forall v_n \in V^{\Delta,p}_{n,0}, \end{cases} \tag{3.4}$$

where the space of trial functions $V_{n,g}^{\Delta,p}$ and the space of test functions $V_{n,0}^{\Delta,p}$ are finite dimensional (nonconforming) spaces on $\mathcal{T}_n$, "mimicking" the infinite dimensional spaces V_g and V_0, defined in (3.3), respectively. Moreover, $a_{0,n}(\cdot,\cdot) : V_{n,g}^{\Delta,p} \times V_{n,0}^{\Delta,p} \to \mathbb{R}$ is a *computable* discrete bilinear form mimicking its continuous counterpart defined again in (3.3). Such approximation spaces and discrete bilinear forms have to be tailored so that method (3.4) is well-posed and provides "good" h- and p-approximation estimates.

The outline of this section is as follows. After introducing suitable global approximation spaces $V_{n,g}^{\Delta,p}$ and $V_{n,0}^{\Delta,p}$ in Section 3.1.1, a set of local projectors is defined in 3.1.2, which are then needed for the construction of $a_{0,n}(\cdot,\cdot)$ in Section 3.1.3.

3.1.1 Nonconforming Trefftz virtual element spaces

Here, we specify the global approximation spaces $V_{n,g}^{\Delta,p}$ and $V_{n,0}^{\Delta,p}$ in the formulation (3.4). To this purpose, we firstly introduce local Trefftz VE spaces and then define the global spaces with respect to the local ones.

Local Trefftz VE spaces. In order to describe the local Trefftz VE spaces, we fix the *degree of accuracy* $p \in \mathbb{N}$. Then, we define, for all $n \in \mathbb{N}$ and for all $K \in \mathcal{T}_n$, the local Trefftz VE space

$$V^{\Delta}(K) := \{v_n \in H^1(K) \mid \Delta v_n = 0 \text{ in } K,\ (\nabla v_n \cdot \mathbf{n}_K)_{|_e} \in \mathbb{P}_{p-1}(e)\ \forall e \in \mathcal{E}^K\}. \tag{3.5}$$

In words, $V^{\Delta}(K)$ consists of *harmonic* functions with piecewise (discontinuous) polynomial normal traces on the boundary of K. The dimension of the space $V^{\Delta}(K)$ is $n_K p$, where we recall that n_K is the number of edges of K. Importantly, the space of harmonic polynomials of degree p is included as a subspace, i.e. $\mathbb{H}_p(K) \subset V^{\Delta}(K)$. This inclusion is essential for the derivation of best approximation estimates, see Section 3.2.2 below. Note that $V^{\Delta}(K)$ is in fact a modification of the corresponding Trefftz VE space introduced in [79] for the construction of a conforming Trefftz VEM for problem (3.2).

A set of $n_K p$ degrees of freedom for $V^{\Delta}(K)$ is the following. Given $v_n \in V^{\Delta}(K)$,

$$\frac{1}{h_e}\int_e v_n m_r^e \,\mathrm{d}s \quad \forall r = 0,\ldots,p-1,\ \forall e \in \mathcal{E}^K, \tag{3.6}$$

where $\{m_r^e\}_{r=0,\ldots,p-1}$ is *any* basis of $\mathbb{P}_{p-1}(e)$. These functionals are indeed *unisolvent* for $V^{\Delta}(K)$, i.e. by fixing their values for $v_n \in V^{\Delta}(K)$, the function v_n is uniquely determined in the space $V^{\Delta}(K)$. This can be seen as follows. If $v_n \in V^{\Delta}(K)$ has all the degrees of freedom set to 0, we need to show that this implies $v_n = 0$. Thus, let v_n be a function with this property. Then,

$$|\nabla v_n|_{1,K}^2 = \int_K \underbrace{(-\Delta v_n)}_{=0} v_n \,\mathrm{d}x + \int_{\partial K} (\nabla v_n \cdot \mathbf{n}_K)\, v_n \,\mathrm{d}s = \sum_{e \in \mathcal{E}^K} \int_e \underbrace{(\nabla v_n \cdot \mathbf{n}_K)}_{\in \mathbb{P}_{p-1}(e)} v_n \,\mathrm{d}s = 0,$$

which means that v_n has to be constant. This, in addition to

$$h_e v_n = \int_e v_n \,\mathrm{d}s = \int_e 1\, v_n \,\mathrm{d}s = 0,$$

for some edge $e \in \mathcal{E}^K$, implies $v_n = 0$, providing unisolvency.

We denote by $\{\varphi_{j,r}\}_{\substack{j=1\ldots,n_K \\ r=0,\ldots,p-1}}$ the local canonical basis associated with the set of degrees of freedom (3.6), namely

$$\mathrm{dof}_{i,s}(\varphi_{j,r}) = \begin{cases} 1 & \text{if } i = j \text{ and } s = r \\ 0 & \text{otherwise} \end{cases} \qquad \forall\, i,j = 1,\ldots,n_K,\ \forall\, s,r = 0,\ldots,p-1. \tag{3.7}$$

The indices i and j refer to the edge, whereas the indices s and r refer to the polynomials employed in the definition of the local degrees of freedom (3.6). It is worth to note that the local canonical basis consists of functions that are not explicitly known inside the element and whose polynomial normal traces over the boundary are also unknown. In this sense, the functions are *virtual.*

Global Trefftz VE spaces. We now focus on the global level. Before defining the global nonconforming Trefftz VE spaces, though, we need to fix some additional notation and to introduce the global nonconforming Sobolev space of order $r \in \mathbb{N}$ with respect to the decomposition $\mathcal{T}_n$ incorporating boundary conditions in a *nonconforming sense.*

Given any internal edge $e \in \mathcal{E}_n^I$ shared by the polygons K^- and K^+ in $\mathcal{T}_n$, we denote by $\mathbf{n}_{K^\pm}^e$ the two outer normal unit vectors with respect to $K^\pm$. For simplicity, we will later only write $\mathbf{n}_{K^\pm}$ instead of $\mathbf{n}_{K^\pm}^e$. Moreover, for boundary edges $e \in \mathcal{E}_n^B$, we recall that $\mathbf{n}_\Omega$ is the normal unit vector pointing outside Ω. Having this, for any $v \in H^1(\mathcal{T}_n)$, we set the jump operator across an edge $e \in \mathcal{E}_n$ to

$$[\![v]\!] := \begin{cases} v_{|_{K^+}} \mathbf{n}_{K^+} + v_{|_{K^-}} \mathbf{n}_{K^-} & \text{if } e \in \mathcal{E}_n^I \\ v\mathbf{n}_\Omega & \text{if } e \in \mathcal{E}_n^B. \end{cases} \tag{3.8}$$

Notice that $[\![v]\!]$ is vector-valued. Then, given $g \in H^{\frac{1}{2}}(\Gamma)$ and $r \in \mathbb{N}$, we define

$$\begin{aligned} H_g^{1,nc}(\mathcal{T}_n, r) := \{v \in H^1(\mathcal{T}_n) \mid & \int_e [\![v]\!] \cdot \mathbf{n}\, q_{r-1} \,\mathrm{d}s = 0 \quad \forall q_{r-1} \in \mathbb{P}_{r-1}(e),\ \forall e \in \mathcal{E}_n^I \\ & \int_e [\![v]\!] \cdot \mathbf{n}\, q_{r-1} \,\mathrm{d}s = \int_e g q_{r-1} \,\mathrm{d}s \quad \forall q_{r-1} \in \mathbb{P}_{r-1}(e),\ \forall e \in \mathcal{E}_n^B\}, \end{aligned} \tag{3.9}$$

where $\mathbf{n}$ is either of the two normal unit vectors to e, but fixed, if $e \in \mathcal{E}_n^I$, and $\mathbf{n} = \mathbf{n}_\Omega$, if $e \in \mathcal{E}_n^B$. In the homogeneous case, definition (3.9) becomes

$$H_0^{1,nc}(\mathcal{T}_n, r) := \{v \in H^1(\mathcal{T}_n) \mid \int_e [\![v]\!] \cdot \mathbf{n}\, q_{r-1} \,\mathrm{d}s = 0 \quad \forall q_{r-1} \in \mathbb{P}_{r-1}(e),\ \forall e \in \mathcal{E}_n\}. \tag{3.10}$$

Importantly, the seminorm $|\cdot|_{1,\mathcal{T}_n}$ is actually a norm for functions in $H_0^{1,nc}(\mathcal{T}_n, r)$. In [57], the validity of the following Poincaré inequality was proven: there exists a positive constant c_P only depending on Ω such that, for all $r \in \mathbb{N}$,

$$\|v\|_{0,\Omega} \le c_P |v|_{1,\mathcal{T}_n} \quad \forall v \in H_0^{1,nc}(\mathcal{T}_n, r). \tag{3.11}$$

We are ready to define global nonconforming Trefftz VE spaces which incorporate Dirichlet boundary conditions in a "nonconforming sense". Let $\widetilde{p} \in \mathbb{N}$ be a given parameter, representing the *order of nonconformity.* For any $g \in H^{\frac{1}{2}}(\Gamma)$, we set

$$V_{n,g}^{\Delta,\widetilde{p}} := \{v_n \in H_g^{1,nc}(\mathcal{T}_n, \widetilde{p}) \mid v_{n|_K} \in V^\Delta(K)\ \forall K \in \mathcal{T}_n\}. \tag{3.12}$$

We observe the following facts:

- Definition (3.12) includes the space of test functions $V_{n,0}^{\Delta,\widetilde{p}}$, by selecting $g = 0$.
- By taking $\widetilde{p} = p$, where p is the degree of accuracy entering in (3.5), the local degrees of freedom can be easily coupled into a global set. The resulting global set of degrees of freedom is of dimension card$(\mathcal{E}_n)p$. From now on, $\widetilde{p} = p$.
- $V_{n,g}^{\Delta,p} \not\subseteq V_g = H_g^1(\Omega)$, thus the method is not conforming.
- Dirichlet boundary conditions on Γ are imposed weakly via the definition of the nonconforming spaces (3.9) and (3.10). For instance, given a Dirichlet datum g, on all boundary edges $e \in \mathcal{E}_n^B$, we set

$$\int_e [\![v_n]\!] \cdot \mathbf{n}_\Omega\, q_{p-1}^e \,\mathrm{d}s = \int_e v_n q_{p-1}^e \,\mathrm{d}s = \int_e g q_{p-1}^e \,\mathrm{d}s \quad \forall v_n \in V_{n,g}^{\Delta,p},\ \forall q_{p-1}^e \in \mathbb{P}_{p-1}(e).$$

Remark 2. We highlight that, at the discrete level, one should also take into account the approximation of the Dirichlet boundary condition g. In practice, assuming $g \in H^{\frac{1}{2}+\varepsilon}(\Gamma)$, for any $\varepsilon > 0$ arbitrarily small, and denoting by g_p the approximation of g obtained by interpolating g at the $p+1$ Gauß-Lobatto nodes on each edge in $\mathcal{E}_n^B$, one should define the trial space as

$$V_{n,g}^{\Delta,p} := \{v_n \in H_{g_p}^{1,nc}(\mathcal{T}_n, p) \mid v_{n|_K} \in V^\Delta(K)\ \forall K \in \mathcal{T}_n\}.$$

With this definition, in the forthcoming analysis (see Proposition 3.2.1, Theorems 3.2.2 and 3.2.5, Proposition 3.2.7, and Theorem 3.2.8 below), an additional term related to the approximation of the Dirichlet datum via Gauß-Lobatto interpolants should be taken into account. However, following [51, Theorem 4.2, Theorem 4.5], it is possible to show that the h- and p-rates of convergence of the method are not spoilt by this term. For this reason and for the sake of simplicity, we will neglect in the following the presence of this term and assume that the approximation space is the one defined in (3.12).

3.1.2 Local projectors

Before specifying, in the next section, the choice for the bilinear form $a_{0,n}(\cdot,\cdot)$ in (3.2), we introduce two projectors that are computable by employing only the degrees of freedom defined in (3.6). The first one is the edge L^2 projector onto the space of polynomials of degree $p-1$

$$\begin{aligned}&\Pi^{0,e}_{p-1} : V^{\Delta}(K)|_e \to \mathbb{P}_{p-1}(e),\\ &\qquad\int_e (v_n - \Pi^{0,e}_{p-1} v_n) q^e_{p-1}\,\mathrm{d}s = 0 \quad \forall v_n \in V^{\Delta}(K),\, \forall q^e_{p-1} \in \mathbb{P}_{p-1}(e).\end{aligned} \tag{3.13}$$

The second one is the bulk H^1 projector onto the space of harmonic polynomials of degree p

$$\begin{aligned}&\Pi^{\nabla,\Delta,K}_p = \Pi^{\nabla,K}_p : V^{\Delta}(K) \to \mathbb{H}_p(K),\\ &\qquad\int_K \nabla(v_n - \Pi^{\nabla,K}_p v_n)\cdot\nabla q^{\Delta}_p\,\mathrm{d}x = 0 \quad \forall v_n \in V^{\Delta}(K),\, \forall q^{\Delta}_p \in \mathbb{H}_p(K),\\ &\qquad\int_{\partial K} (v_n - \Pi^{\nabla,K}_p v_n)\,\mathrm{d}s = 0 \quad \forall v_n \in V^{\Delta}(K),\end{aligned} \tag{3.14}$$

where the last condition is imposed in order to define the projector in a unique way. Whereas the former projector is clearly computable from the degrees of freedom (3.6), computability of the second can be seen after integration by parts

$$\int_K \nabla v_n \cdot \nabla q^{\Delta}_p\,\mathrm{d}x = -\int_K v_n \underbrace{\Delta q^{\Delta}_p}_{=0}\,\mathrm{d}x + \int_{\partial K} v_n (\nabla q^{\Delta}_p \cdot \mathbf{n}_K)\,\mathrm{d}s,$$

and employing that the Neumann trace of a harmonic polynomial of degree p is a polynomial of degree $p-1$, together with the degrees of freedom (3.6).

3.1.3 Discrete bilinear forms

In this section, we complete the definition of the method (3.4) by introducing a suitable bilinear form $a_{0,n}(\cdot,\cdot)$, which is explicitly computable. We follow here the typical VEM gospel [31, 40, 79]. Defining the local bilinear forms on polygons $K \in \mathcal{T}_n$ as

$$a^K_0(u,v) := (\nabla u, \nabla v)_{0,K} \quad \forall u,\, v \in H^1(K),$$

we highlight that $a^K_0(\cdot,\cdot)$ are not explicitly computable on the whole discrete spaces since an explicit representation of functions in the Trefftz VE spaces is not available in closed form. Thus,

$$a_{0,n}(u,v) = \sum_{K\in\mathcal{T}_n} a^K_0(u,v)$$

cannot be taken. Therefore, we aim at introducing explicit computable local discrete bilinear forms $a^K_{0,n}(\cdot,\cdot)$ which mimic their continuous counterparts $a^K_0(\cdot,\cdot)$, and then we define $a_{0,n}(\cdot,\cdot)$ by means of these local approximations. To this purpose, we observe that the Pythagorean theorem yields

$$a^K_0(u_n,v_n) = a^K_0(\Pi^{\nabla,K}_p u_n, \Pi^{\nabla,K}_p v_n) + a^K_0((I-\Pi^{\nabla,K}_p)u_n, (I-\Pi^{\nabla,K}_p)v_n) \quad \forall u_n,\, v_n \in V^{\Delta}(K), \tag{3.15}$$

where we recall that $\Pi^{\nabla,K}_p$ is defined in (3.14). The first term on the right-hand side of (3.15) is computable, whereas the second is not. Thus, following [79] and the references therein, we replace this term by a *computable* symmetric bilinear form $S^K_0 : \ker(\Pi^{\nabla,K}_p) \times \ker(\Pi^{\nabla,K}_p) \to \mathbb{R}$, such that

$$c_*(p)|v_n|^2_{1,K} \le S^K_0(v_n,v_n) \le c^*(p)|v_n|^2_{1,K} \quad \forall v_n \in \ker(\Pi^{\nabla,K}_p), \tag{3.16}$$

where $c_*(p)$ and $c^*(p)$ are two positive constants which may depend on p, but are independent of K and, in particular, of h_K. An explicit choice of $S_0^K(\cdot,\cdot)$ is given in (3.22) below.

Hence, depending on the choice of the stabilization, a class of candidates for the local discrete symmetric bilinear forms is

$$a_{0,n}^K(u_n, v_n) := a_0^K(\Pi_p^{\nabla,K} u_n, \Pi_p^{\nabla,K} v_n) + S_0^K((I - \Pi_p^{\nabla,K})u_n, (I - \Pi_p^{\nabla,K})v_n). \tag{3.17}$$

The forms $a_{0,n}^K(\cdot,\cdot)$ satisfy the two following properties:

(**P1**) **p-harmonic consistency**: for all $K \in \mathcal{T}_n$ and for all $p \in \mathbb{N}$,

$$a_0^K(q_p^\Delta, v_n) = a_{0,n}^K(q_p^\Delta, v_n) \quad \forall q_p^\Delta \in \mathbb{H}_p(K),\, \forall v_n \in V^\Delta(K); \tag{3.18}$$

(**P2**) **stability**: for all $K \in \mathcal{T}_n$ and for all $p \in \mathbb{N}$,

$$\alpha_*(p)|v_n|_{1,K}^2 \le a_{0,n}^K(v_n, v_n) \le \alpha^*(p)|v_n|_{1,K}^2 \quad \forall v_n \in V^\Delta(K), \tag{3.19}$$

where $\alpha_*(p) = \min(1, c_*(p))$ and $\alpha^*(p) = \max(1, c^*(p))$.

Property (**P1**) justifies to refer to p as degree of accuracy of the method, since whenever either of its two entries is a harmonic polynomial of degree p, the local discrete bilinear form can be computed exactly, up to machine precision. Moreover, (**P2**) implies continuity since $a_{0,n}^K(\cdot,\cdot)$ is assumed to be symmetric:

$$\begin{aligned} a_{0,n}^K(u_n, v_n) &\le \left(a_{0,n}^K(u_n, u_n)\right)^{\frac{1}{2}} \left(a_{0,n}^K(v_n, v_n)\right)^{\frac{1}{2}} \\ &\le \alpha^*(p)|u_n|_{1,K}|v_n|_{1,K} \quad \forall u_n, v_n \in V^\Delta(K). \end{aligned} \tag{3.20}$$

As indicated above, the global discrete bilinear form is then defined as

$$a_{0,n}(u_n, v_n) := \sum_{K \in \mathcal{T}_n} a_{0,n}^K(u_n, v_n) \quad \forall u_n \in V_{n,g_1}^{\Delta,p},\, \forall v_n \in V_{n,g_2}^{\Delta,p} \tag{3.21}$$

for all $g_1, g_2 \in H^{\frac{1}{2}}(\Gamma)$.

Choice of the stabilization. Here, we introduce an explicit stabilization $S_0^K(\cdot,\cdot)$ with explicit bounds of the constants $c_*(p)$ and $c^*(p)$ in (3.16).

For all $K \in \mathcal{T}_n$, define

$$S_0^K(u_n, v_n) := \sum_{e \in \mathcal{E}^K} \frac{p}{h_e} (\Pi_{p-1}^{0,e} u_n, \Pi_{p-1}^{0,e} v_n)_{0,e} \quad \forall u_n,\, v_n \in \ker(\Pi_p^{\nabla,K}). \tag{3.22}$$

Then, the following result holds true.

Theorem 3.1.1. *Assume that the mesh assumptions (**G1**) and (**G2**), introduced in Section 2.3, hold true. Then, for any $K \in \mathcal{T}_n$, the stabilization $S_0^K(\cdot,\cdot)$ defined in* (3.22) *satisfies* (3.16) *with the bounds*

$$c_*(p) \gtrsim p^{-2}, \qquad c^*(p) \lesssim \begin{cases} p\left(\frac{\log(p)}{p}\right)^{\frac{\lambda_K}{2}} & \text{if } K \text{ is convex} \\ p\left(\frac{\log(p)}{p}\right)^{\frac{\lambda_K}{2\omega_K} - \varepsilon} & \text{otherwise} \end{cases} \tag{3.23}$$

for all $\varepsilon > 0$ arbitrarily small, where the hidden constants in (3.23) *are independent of h and p, and where $\omega_K \pi$ and $\lambda_K \pi$, with ω_K and $\lambda_K \in (0, 2)$, denote the largest interior and the smallest exterior angles of K, respectively.*

Proof. We assume, without loss of generality, that $h_K = 1$; the general result follows from a scaling argument.

For any function v_n in $V^{\Delta}(K)$, we have

$$\begin{aligned}|v_n|_{1,K}^2 &= -\int_K \underbrace{(\Delta v_n)}_{=0} v_n \,\mathrm{d}x + \int_{\partial K} \nabla v_n \cdot \mathbf{n}_K \, v_n \,\mathrm{d}s \\ &= \sum_{e\in\mathcal{E}^K} \int_e \nabla v_n \cdot \mathbf{n}_K (\Pi_{p-1}^{0,e} v_n)\,\mathrm{d}s \le \|\nabla v_n \cdot \mathbf{n}_K\|_{0,\partial K} \|\Pi_{p-1}^{0,\partial K} v_n\|_{0,\partial K}\end{aligned} \tag{3.24}$$

where we have set, with an abuse of notation, $(\Pi_{p-1}^{0,\partial K} v_n)_{|e} = \Pi_{p-1}^{0,e}(v_{n_{|e}})$. We prove that

$$\|\nabla v_n \cdot \mathbf{n}_K\|_{0,\partial K} \lesssim p^{\frac{3}{2}} \|\nabla v_n \cdot \mathbf{n}_K\|_{-\frac{1}{2},\partial K}. \tag{3.25}$$

To this end, we set, for the sake of simplicity, $r_p := \nabla v_n \cdot \mathbf{n}_K$, and consider the case $r_p \neq 0$. One has $r_p \in L^2(\partial K)$ with $r_{p|_e} \in \mathbb{P}_p(e)$ for all $e \in \mathcal{E}^K$. In general, $r_p \notin H^{\frac{1}{2}}(\partial K)$. Further, we introduce the piecewise bubble function $b_{\partial K} \in H^{\frac{1}{2}}(\partial K)$ defined edgewise as

$$(b_{\partial K})_{|e}(\mathbf{x}) := (\beta \circ \phi_e^{-1})(\mathbf{x}) \quad \forall e \in \mathcal{E}^K,$$

where $\phi_e : [-1,1] \to e$ is the linear transformation mapping the interval $[-1,1]$ to the edge e, and $\beta : [-1,1] \to [0,1]$ is the 1D quadratic bubble function $\beta(x) := 4(1-x^2)$.

Using the definition of the $H^{-\frac{1}{2}}(\partial K)$ norm, $r_p \in L^2(\partial K)$, and $r_p b_{\partial K} \in H^{\frac{1}{2}}(\partial K)\backslash\{0\}$, we have

$$\|r_p\|_{-\frac{1}{2},\partial K} = \sup_{\psi \in H^{\frac{1}{2}}(\partial K)\backslash\{0\}} \frac{(r_p,\psi)_{0,\partial K}}{\|\psi\|_{\frac{1}{2},\partial K}} \ge \frac{(r_p, r_p b_{\partial K})_{0,\partial K}}{\|r_p b_{\partial K}\|_{\frac{1}{2},\partial K}} = \frac{\|r_p b_{\partial K}^{\frac{1}{2}}\|_{0,\partial K}^2}{\|r_p b_{\partial K}\|_{\frac{1}{2},\partial K}}. \tag{3.26}$$

Then, we note that the two following polynomial p-inverse inequalities hold true:

$$\|r_p b_{\partial K}\|_{0,e} \le \|r_p b_{\partial K}^{\frac{1}{2}}\|_{0,e}, \quad |r_p b_{\partial K}|_{1,e} \lesssim p \|r_p b_{\partial K}^{\frac{1}{2}}\|_{0,e} \quad \forall e \in \mathcal{E}^K. \tag{3.27}$$

The first one is a direct consequence of the fact that the range of $b_{\partial K}$ is $[0,1]$, and the second one follows from [26, Lemma 2]. Using (3.27), summing over all edges $e \in \mathcal{E}^K$, and applying interpolation theory, lead to

$$\|r_p b_{\partial K}\|_{\frac{1}{2},\partial K} \lesssim p^{\frac{1}{2}} \|r_p b_{\partial K}^{\frac{1}{2}}\|_{0,\partial K},$$

which, together with (3.26), gives

$$\|r_p\|_{-\frac{1}{2},\partial K} \gtrsim p^{-\frac{1}{2}} \|r_p b_{\partial K}^{\frac{1}{2}}\|_{0,\partial K} \gtrsim p^{-\frac{3}{2}} \|r_p\|_{0,\partial K},$$

where, in the last inequality, [50, Lemma 4] was applied. The bound (3.25) follows immediately.

From (3.24) and (3.25), taking also into account that $\Delta v_n = 0$ in K, we get

$$|v_n|_{1,K}^2 \lesssim p^{\frac{3}{2}} \|\nabla v_n \cdot \mathbf{n}_K\|_{-\frac{1}{2},\partial K} \|\Pi_{p-1}^{0,\partial K} v_n\|_{0,\partial K} \lesssim p^{\frac{3}{2}} |v_n|_{1,K} \|\Pi_{p-1}^{0,\partial K} v_n\|_{0,\partial K},$$

where in the last step we have used a Neumann trace inequality, see e.g. [166, Theorem A.33]. This proves the first inequality of (3.16) with $c_*(p) \gtrsim p^{-2}$.

In order to prove the second one, we can write

$$\|\Pi_{p-1}^{0,\partial K} v_n\|_{0,\partial K} \le \|v_n\|_{0,\partial K} \lesssim \|v_n\|_{0,K}^{\frac{1}{2}} |v_n|_{1,K}^{\frac{1}{2}}, \tag{3.28}$$

where we have used the stability of the L^2 projection, the multiplicative trace inequality (2.26), and the Poincaré inequality (2.28), which is valid since $v_n \in \ker(\Pi_p^{\nabla,K})$ and thus has zero mean value on ∂K, see (3.14).

Let us estimate the first factor on the right-hand side of (3.28). To this end, we define $\overline{v}_n$ as the average of v_n over the polygon K. A triangle inequality yields

$$\|v_n\|_{0,K} \le \|v_n - \overline{v}_n\|_{0,K} + \|\overline{v}_n\|_{0,K}. \tag{3.29}$$

Recalling that v_n has zero average over ∂K, we have

$$\|\overline{v}_n\|_{0,K} = |K|^{\frac{1}{2}}|\overline{v}_n| = \frac{|K|^{\frac{1}{2}}}{|\partial K|}\left|\int_{\partial K}\overline{v}_n - v_n\,\mathrm{d}s\right|.$$

A Cauchy-Schwarz inequality, together with the multiplicative trace inequality (2.26), yields

$$\|\overline{v}_n\|_{0,K} \lesssim \|v_n - \overline{v}_n\|_{0,K}^{\frac{1}{2}}|v_n|_{1,K}^{\frac{1}{2}}.$$

Inserting this inequality in (3.29) gives

$$\|v_n\|_{0,K} \lesssim \|v_n - \overline{v}_n\|_{0,K} + \|v_n - \overline{v}_n\|_{0,K}^{\frac{1}{2}}|v_n|_{1,K}^{\frac{1}{2}}. \tag{3.30}$$

From [79, Lemma 3.2], we have

$$\|v_n - \overline{v}_n\|_{0,K}^{\frac{1}{2}} \lesssim \begin{cases} \left(\frac{\log(p)}{p}\right)^{\lambda_K}|v_n|_{1,K} & \text{if } K \text{ is convex}\\ \left(\frac{\log(p)}{p}\right)^{\frac{\lambda_K}{\omega_K}-\varepsilon}|v_n|_{1,K} & \text{otherwise}\end{cases}$$

for all $\varepsilon > 0$ arbitrarily small. Inserting this into (3.30) gives

$$\|v_n\|_{0,K} \lesssim \begin{cases} \left(\frac{\log(p)}{p}\right)^{\frac{\lambda_K}{2}}|v_n|_{1,K} & \text{if } K \text{ is convex}\\ \left(\frac{\log(p)}{p}\right)^{\frac{\lambda_K}{2\omega_K}-\varepsilon}|v_n|_{1,K} & \text{otherwise,}\end{cases}$$

which, together with (3.28), gives (3.16) with $c^*(p)$ as in (3.23). □

Owing to (3.19) and (3.23) one deduces

$$\alpha_*(p) \gtrsim p^{-2}, \qquad \alpha^*(p) \lesssim \begin{cases} p\left(\frac{\log(p)}{p}\right)^{\frac{\lambda_K}{2}} & \text{if } K \text{ is convex}\\ p\left(\frac{\log(p)}{p}\right)^{\frac{\lambda_K}{2\omega_K}-\varepsilon} & \text{otherwise}\end{cases}$$

for all $\varepsilon > 0$ arbitrarily small.

Remark 3. In the conforming Trefftz VEM setting [79], the following local stabilization forms were introduced:

$$S_0^K(u_n, v_n) = (u_n, v_n)_{\frac{1}{2},\partial K} \quad \forall K \in \mathcal{T}_n.$$

It was proven that employing such stabilization forms leads to have stability constants $\alpha_*(p)$ and $\alpha^*(p)$ that are independent of the degree of accuracy p. However, in the present nonconforming setting, such a stabilization is not computable, as the Dirichlet traces of functions in the local VE spaces are not available in closed form.

Finally, we investigate numerically the behavior of the conditioning of the global VE matrix in terms of the degree of accuracy p, when employing the local stabilization forms in (3.22). In Figure 3.1, we plot the condition number for different values of p, when computing the global stiffness matrix on a Cartesian mesh, a Voronoi-Lloyd mesh, and an Escher horses mesh, see Figure 2.2, and note that it grows algebraically with p. We remark that the condition number of standard (non-Trefftz) VEM can grow exponentially or algebraically with p, depending on the choice of the internal degrees of freedom. This was investigated in [142].

3.2 *A priori* error analysis

This section is dedicated to perform an *a priori* analysis for the method defined in (3.4).

To this purpose, in Section 3.2.1, we first derive approximation estimates for Trefftz VE functions in the global approximation spaces $V_{n,g}^{\Delta,p}$ and $V_{n,0}^{\Delta,p}$, introduced in (3.12). Then, in Section 3.2.2, an abstract error analysis is carried out. Such analysis is instrumental for the derivation of h- and p-error estimates in the H^1 seminorm, which is the topic of Section 3.2.3. Afterwards, in Section 3.2.4, corresponding L^2 error bounds for the discretization error are provided. Finally, in Section 3.2.5, some hints on the extension of the method to the 3D case are given and the main differences between 2D and 3D are pointed out.

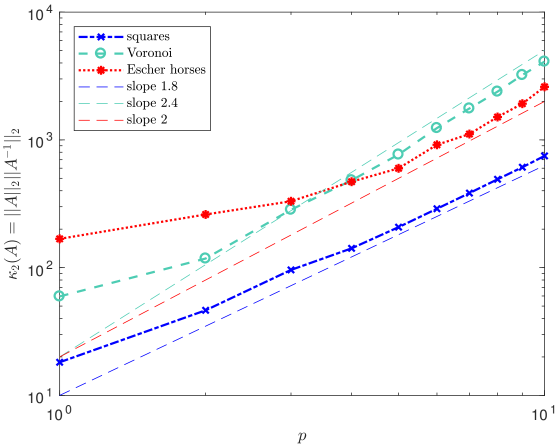


Figure 3.1: Condition number for different values of p of the global stiffness matrix obtained with the local stabilization forms in (3.22). A Cartesian mesh, a Voronoi-Lloyd mesh, and an Escher horses mesh have been considered. We observe algebraic growth of the condition number with p for all the tested meshes.

3.2.1 Approximation properties of functions in Trefftz virtual element spaces

This section deals with the approximation properties of functions in the nonconforming Trefftz VE spaces $V_{n,g}^{\Delta,p}$ and $V_{n,0}^{\Delta,p}$, introduced in (3.12).

Since h- and p-approximation properties of harmonic functions via harmonic polynomials are known, see e.g. [23, 122], we want to relate best approximation estimates in the nonconforming Trefftz VE spaces to the corresponding ones in *discontinuous* harmonic polynomial spaces. In particular, we prove the following result, which can be seen as a generalization of those in the case of nonconforming finite element methods, see [175].

Proposition 3.2.1. *Given $g \in H^{\frac{1}{2}}(\Gamma)$, let $u \in V_g$, where V_g is defined in* (3.3)*. Then, for any polygonal partition $\mathcal{T}_n$ of Ω, there exists $u_I \in V_{n,g}^{\Delta,p}$, such that*

$$|u - u_I|_{1,\mathcal{T}_n} \leq |u - q_p^\Delta|_{1,\mathcal{T}_n} \quad \forall q_p^\Delta \in \mathcal{S}^{p,\Delta,-1}(\mathcal{T}_n),$$

where $\mathcal{S}^{p,\Delta,-1}(\mathcal{T}_n)$ is the space of discontinuous piecewise harmonic polynomials of degree p, i.e.

$$\mathcal{S}^{p,\Delta,-1}(\mathcal{T}_n) := \{q \in L^2(\Omega) : q_{|_K} \in \mathbb{H}_p(K) \ \forall K \in \mathcal{T}_n\}. \tag{3.31}$$

Proof. Define $u_I \in V_{n,g}^{\Delta,p}$ by

$$\int_e (u - u_I) q_{p-1}^e \, \mathrm{d}s = 0 \quad \forall q_{p-1}^e \in \mathbb{P}_{p-1}(e), \forall e \in \mathcal{E}_n, \tag{3.32}$$

that is, we fix the degrees of freedom (3.6) of u_I to be equal to the values of the same functionals applied to the solution u. Then, for any $K \in \mathcal{T}_n$, we have

$$|u - u_I|_{1,K}^2 = \int_K \nabla(u - u_I) \cdot \nabla(u - q_p^\Delta) \, \mathrm{d}x + \int_K \nabla(u - u_I) \cdot \nabla(q_p^\Delta - u_I) \, \mathrm{d}x. \tag{3.33}$$

For the second term on the right-hand side of (3.33), we use integration by parts to estimate

$$\begin{aligned}\int_K \nabla(u - u_I) \cdot \nabla(q_p^\Delta - u_I) \, \mathrm{d}x = &- \int_K (u - u_I) \Delta(q_p^\Delta - u_I) \, \mathrm{d}x + \int_{\partial K} (u - u_I) \nabla(q_p^\Delta - u_I) \cdot \mathbf{n}_K \, \mathrm{d}s \\ &= 0,\end{aligned}$$

where the first integral on the right-hand side is 0 since both q_p^Δ and u_I lie in the kernel of the Laplace operator, and the second is 0 owing to the definition of u_I in (3.32). Hence, (3.33) can be estimated with the Cauchy-Schwarz inequality by

$$|u - u_I|_{1,K}^2 \leq |u - u_I|_{1,K} |u - q_p^\Delta|_{1,K}.$$

Dividing by $|u - u_I|_{1,K}$ and summing over all elements $K \in \mathcal{T}_n$ leads to the result. □

As compared to [143, Prop. 3.1], with a slight modification of the proof, we have eliminated the constant 2 in the upper bound. Moreover, we note that, with a similar proof of that of Proposition 3.2.1, one can show an equivalent result for the nonconforming (non-Trefftz) VE spaces of [18]; see Proposition 3.2.7 below.

3.2.2 Abstract error analysis

Along the lines of [31,39,79], we provide here an abstract error analysis of the method (3.4), taking the nonconformity of the approximation into account. To this purpose, we introduce the auxiliary bilinear form

$$\mathcal{N}_n : H^1(\Omega) \times H_0^{1,nc}(\mathcal{T}_n, p) \to \mathbb{R}, \quad \mathcal{N}_n(u, v) := \sum_{e\in\mathcal{E}_n} \int_e \nabla u \cdot [\![v]\!] \, \mathrm{d}s. \tag{3.34}$$

The following convergence result holds true, mimicking those of [80, Thm. 15] for nonconforming finite element methods.

Theorem 3.2.2. *Assume that the mesh assumptions **(G1)** and **(G2)**, introduced in Section 2.3, hold true. Then, the nonconforming Trefftz VEM* (3.4) *with Trefftz VE spaces as in* (3.12) *employing $\widetilde{p} = p$, discrete bilinear form $a_{0,n}(\cdot,\cdot)$ as in* (3.21)*, and local stabilization forms $S_0^K(\cdot,\cdot)$ satisfying* (3.16)*, is well-posed and the following bound holds true:*

$$|u - u_n|_{1,\mathcal{T}_n} \le \frac{\alpha^*(p)}{\alpha_*(p)} \left\{ 4 \inf_{q_p^\Delta \in \mathcal{S}^{p,\Delta,-1}(\mathcal{T}_n)} |u - q_p^\Delta|_{1,\mathcal{T}_n} + \sup_{v_n \in V_{n,0}^{\Delta,p}} \frac{\mathcal{N}_n(u, v_n)}{|v_n|_{1,\mathcal{T}_n}} \right\}, \tag{3.35}$$

where we recall that $\mathcal{S}^{p,\Delta,-1}(\mathcal{T}_n)$ is defined in (3.31)*, $\mathcal{N}_n(\cdot,\cdot)$ is given in* (3.34)*, and the stability constants $\alpha_*(p)$ and $\alpha^*(p)$ are introduced in* (3.19).

Proof. Well-posedness of the method follows from (3.11), (3.19) and the Lax-Milgram lemma.

For the bound (3.35), we observe that

$$|u - u_n|_{1,\mathcal{T}_n} \le |u - u_I|_{1,\mathcal{T}_n} + |u_n - u_I|_{1,\mathcal{T}_n} \quad \forall u_I \in V_{n,g}^{\Delta,p}.$$

We estimate the second term on the right-hand side. Set $\delta_n := u_n - u_I$. Since $u_n, u_I \in V_{n,g}^{\Delta,p}$, then $\delta_n \in V_{n,0}^{\Delta,p}$. Therefore, for all $q_p^\Delta \in \mathcal{S}^{p,\Delta,-1}(\mathcal{T}_n)$, using (3.19), (3.4) and (3.18), we have

$$\begin{aligned} |\delta_n|_{1,\mathcal{T}_n}^2 &= \sum_{K\in\mathcal{T}_n} |\delta_n|_{1,K}^2 \le \frac{1}{\alpha_*(p)} \sum_{K\in\mathcal{T}_n} a_{0,n}^K(\delta_n, \delta_n) = -\frac{1}{\alpha_*(p)} \sum_{K\in\mathcal{T}_n} a_{0,n}^K(u_I, \delta_n) \\ &= -\frac{1}{\alpha_*(p)} \left\{ \sum_{K\in\mathcal{T}_n} \left[a_{0,n}^K(u_I - q_p^\Delta, \delta_n) + a_0^K(q_p^\Delta - u, \delta_n) \right] + \sum_{K\in\mathcal{T}_n} a_0^K(u, \delta_n) \right\}. \end{aligned}$$

The last term on the right-hand side can be rewritten in the spirit of nonconforming methods. More precisely, we observe that an integration by parts, the fact that $\Delta u = 0$ in every $K \in \mathcal{T}_n$, and the definition (3.34), yield

$$\sum_{K\in\mathcal{T}_n} a_0^K(u, \delta_n) = \sum_{K\in\mathcal{T}_n} \int_{\partial K} \nabla u \cdot \mathbf{n}_K \, \delta_n \, \mathrm{d}s = \sum_{e\in\mathcal{E}_n} \int_e \nabla u \cdot [\![\delta_n]\!] \, \mathrm{d}s = \mathcal{N}_n(u, \delta_n).$$

This, together with the stability property (3.19), and the triangle and the Cauchy-Schwarz inequalities, gives

$$|\delta_n|_{1,\mathcal{T}_n}^2 \le \frac{1}{\alpha_*(p)} \Big[\Big(\alpha^*(p)(|u_I - u|_{1,\mathcal{T}_n} + |u - q_p^\Delta|_{1,\mathcal{T}_n}) + |q_p^\Delta - u|_{1,\mathcal{T}_n} \Big) |\delta_n|_{1,\mathcal{T}_n} + \mathcal{N}_n(u, \delta_n) \Big].$$

Therefore, using Proposition 3.2.1 and $\alpha^*(p) \ge 1$, we obtain

$$\begin{aligned} |\delta_n|_{1,\mathcal{T}_n} &\le \frac{1}{\alpha_*(p)} \left[\alpha^*(p) 2 |u - q_p^\Delta|_{1,\mathcal{T}_n} + |q_p^\Delta - u|_{1,\mathcal{T}_n} + \frac{\mathcal{N}_n(u, \delta_n)}{|\delta_n|_{1,\mathcal{T}_n}} \right] \\ &\le \frac{\alpha^*(p)}{\alpha_*(p)} \left[3 |u - q_p^\Delta|_{1,\mathcal{T}_n} + \frac{\mathcal{N}_n(u, \delta_n)}{|\delta_n|_{1,\mathcal{T}_n}} \right], \end{aligned}$$

and bound (3.35) readily follows. □

We refer to the term $\frac{\alpha^*(p)}{\alpha_*(p)}$ appearing in (3.35) as *pollution factor.*

Remark 4. It is interesting to note that the counterpart of Theorem 3.2.2 in the conforming version of the Trefftz VEM in [79] states that the error of the method is bounded, up to a constant times the pollution factor $\frac{\alpha^*(p)}{\alpha_*(p)}$, by a best approximation error with respect to piecewise discontinuous harmonic polynomials, plus the best approximation error with respect to functions in the global approximation space. In the nonconforming setting here, however, the latter term is not present, thanks to Proposition 3.2.1. The additional term here is related to the nonconformity.

3.2.3 h- and p-error analysis

In this section, we derive error estimates for the h- and p-versions of the method (3.4). For the sake of clarity, by h-version, we here mean the strategy of keeping the degree of accuracy p fixed and only decreasing the mesh size h, whereas, for the p-version, h is fixed and the convergence of the method is studied for increasing p. To this purpose, we discuss how to estimate the two terms on the right-hand side of (3.35) in terms of h and p.

The first term, i.e. the best approximation error with respect to discontinuous harmonic polynomials, can be dealt with following [148,149]. In particular, we recall the following result from [148, Theorem 2.9] (see also [149, Chapter II]).

Lemma 3.2.3. *Under the star-shapedness assumption (**G2**) in Section 2.3, for a given $K \in \mathcal{T}_n$, we denote by $\lambda_K \pi$, $0 < \lambda_K < 2$, its smallest exterior angle. Then, for every harmonic function u in $H^{s+1}(K)$, $s \geq 0$, there exists a sequence $\{q_p^\Delta\}_{p\in\mathbb{N}}$, with $q_p^\Delta \in \mathbb{H}_p(K)$ for all $p \in \mathbb{N}$ with $p \geq s-1$, such that*

$$|u - q_p^\Delta|_{1,K} \leq c h_K^s \left(\frac{\log(p)}{p}\right)^{\lambda_K s} \|u\|_{s+1,K}, \tag{3.36}$$

for some positive constant c depending only on ρ_0.

Remark 5. We underline that the p-version approximation of harmonic functions by means of harmonic polynomials has different rates of convergence than that of generic (non-harmonic) functions by means of full polynomials. In particular, from (3.36), one deduces that, on convex elements, a better convergence rate is achieved (i.e. harmonic functions can be better approximated by polynomials than generic functions, even by considering harmonic polynomials only), while on non-convex elements, the rate of approximation gets worse (i.e. the best approximation of harmonic functions by full polynomials fails to be achieved with harmonic polynomials).

Next, we prove an upper bound for the nonconformity term $\mathcal{N}_n(u, v_n)$ introduced in (3.34). To this purpose, we use tools of nonconforming methods and hp-analysis. We need to require on the sequence of meshes $\{\mathcal{T}_n\}_{n\in\mathbb{N}}$, in addition to (**G1**)-(**G3**) in Section 2.3, the following quasi-uniformity assumption:

(**G4**) there exists a constant $\rho_1 \geq 1$ such that, for all $n \in \mathbb{N}$ and for all K_1 and K_2 in $\mathcal{T}_n$, it holds $h_{K_2} \leq \rho_1 h_{K_1}$.

Before formulating the estimate, we define Ω_{ext} as an extension of the domain Ω, subordinated to polygonal decompositions. More precisely, let $\widetilde{\mathcal{T}}_n$ be a triangulation of Ω which is given by the union of local triangulations $\widetilde{\mathcal{T}}_n(K)$ over each polygon $K \in \mathcal{T}_n$ ($\widetilde{\mathcal{T}}_n$ is nested in $\mathcal{T}_n$); such local triangulations are obtained by connecting the vertices of K to the center of the ball with respect to which K is star-shaped, see assumption (**G2**). With each triangle $T \in \widetilde{\mathcal{T}}_n$, we associate $Q(T)$, a parallelogram obtained by reflecting T with respect to the midpoint of one of its edges, which is arbitrarily fixed. Then, we set

$$\Omega_{\text{ext}} := \bigcup_{T\in\widetilde{\mathcal{T}}_n} Q(T). \tag{3.37}$$

Notice that Ω_{ext} could coincide with Ω.

With this, we have all the ingredients to prove the following lemma, which provides an upper bound for the nonconformity term $\mathcal{N}_n(u, v_n)$ in (3.35).

Lemma 3.2.4. *Assume that the mesh assumptions (**G1**)-(**G4**) are satisfied. Then, for all $s \geq 1$ and for all $u \in H^{s+1}(\Omega_{\text{ext}})$, the following bound holds true:*

$$|\mathcal{N}_n(u, v_n)| \leq c\, d^s \frac{h^{\min(s,p)}}{p^s} \|u\|_{s+1,\Omega_{\text{ext}}} |v_n|_{1,\mathcal{T}_n} \quad \forall v_n \in V^{\Delta,p}_{n,0},$$

where c is a positive constant depending only on ρ_0, ρ_1, and Λ, and d is a positive constant.

Proof. Without loss of generality, let us assume that $h = 1$, so that $\rho_1^{-1} \leq h_K \leq 1$ for all $K \in \mathcal{T}_n$, due to the assumption (**G4**); the general assertion follows from a scaling argument.

First, we observe that, for all $v_n \in V^{\Delta,p}_{n,0}$, the definition of nonconforming spaces and basic properties of orthogonal projectors yield

$$\begin{aligned} |\mathcal{N}_n(u, v_n)| &= \left| \sum_{e\in\mathcal{E}_n} \int_e \nabla u \cdot [\![v_n]\!] \,\mathrm{d}s \right| = \left| \sum_{e\in\mathcal{E}_n} \int_e (\nabla u - \Pi^{0,e}_{p-1}(\nabla u)) \cdot ([\![v_n]\!] - \Pi^{0,e}_{p-1}[\![v_n]\!]) \,\mathrm{d}s \right| \\ &\leq \sum_{e\in\mathcal{E}_n} \left\| \nabla u - \Pi^{0,e}_{p-1}(\nabla u) \right\|_{0,e} \left\| [\![v_n]\!] - \Pi^{0,e}_{p-1}[\![v_n]\!] \right\|_{0,e}, \end{aligned} \tag{3.38}$$

where we have denoted by $\Pi^{0,e}_{p-1}$, with an abuse of notation, the L^2 projector onto the vectorial polynomial spaces of degree $p - 1$ on e.

In order to estimate the first term on the right-hand side of (3.38), we proceed as follows. Let us consider $\widetilde{\mathcal{T}}_n$, the union of the local triangulations $\widetilde{\mathcal{T}}_n(K)$ of each $K \in \mathcal{T}_n$ defined as above. The triangulation $\widetilde{\mathcal{T}}_n$ has the property that each $T \in \widetilde{\mathcal{T}}_n$ is star-shaped with respect to a ball of radius greater than or equal to $\rho_2 h_T$, where ρ_2 is a positive constant and h_T is the diameter of the triangle T, see [152]. Let now $e \in \mathcal{E}_n$ be fixed and $K \in \mathcal{T}_n$ be a polygon with $e \in \mathcal{E}^K$. Then,

$$\|\nabla u - \Pi^{0,e}_{p-1}(\nabla u)\|_{0,e} \leq \|\nabla u - \Pi^{0,T}_{p-1}(\nabla u)\|_{0,e},$$

where $\Pi^{0,T}_{p-1}$ is the L^2 projector onto the space of vectorial polynomials of degree at most $p - 1$ over T, and T is the triangle in $\widetilde{\mathcal{T}}_n(K)$ with $e \subset \partial T$ (this inequality holds true because the restriction of $\Pi^{0,T}_{p-1}(\nabla u)$ to e is a vectorial polynomial of degree $p - 1$).

For any $v \in H^2(T)$, due to [78, Theorem 3.1], we have

$$\|\nabla v - \Pi^{0,T}_{p-1}(\nabla v)\|_{0,e} \leq \frac{\sqrt{5}+1}{\sqrt{2}} p^{-\frac{1}{2}} |\nabla v|_{1,T}. \tag{3.39}$$

Using that $\Pi^{0,T}_{p-1} \nabla q_p = \nabla q_p$ for all $q_p \in \mathbb{P}_p(T)$, owing to (3.39), we get

$$\|\nabla u - \Pi^{0,T}_{p-1}(\nabla u)\|_{0,e} = \|(\nabla(u - q_p)) - \Pi^{0,T}_{p-1}(\nabla(u - q_p))\|_{0,e} \lesssim p^{-\frac{1}{2}} |\nabla(u - q_p)|_{1,T}.$$

Applying now standard hp-polynomial approximation results, see e.g. [39, Lemma 5.1], we obtain for every $q_p \in \mathbb{P}_p(T)$,

$$|\nabla(u - q_{p-1})|_{1,T} \lesssim d^s p^{-s+1} |\nabla u|_{s,Q(T)}, \tag{3.40}$$

where d is a positive constant and $Q(T)$ is the parallelogram given by the union of T and its reflection defined above.

Moving to the second term in (3.38), assuming that $e = \partial T^- \cap \partial T^+$, where $T^\pm \in \widetilde{\mathcal{T}}_n$ and $T^\pm \subset K^\pm$, we have

$$\left\| [\![v_n]\!] - \Pi^{0,e}_{p-1}[\![v_n]\!] \right\|_{0,e} \leq \|v_{n|_{T^+}} - \Pi^{0,T^+}_{p-1} v_{n|_{T^+}}\|_{0,e} + \|v_{n|_{T^-}} - \Pi^{0,T^-}_{p-1} v_{n|_{T^-}}\|_{0,e}.$$

Then, applying once again [78, Theorem 3.1], we deduce

$$\left\| [\![v_n]\!] - \Pi^{0,e}_{p-1}[\![v_n]\!] \right\|_{0,e} \lesssim p^{-\frac{1}{2}} \left(|v_{n|_{T^+}}|_{1,T^+} + |v_{n|_{T^-}}|_{1,T^-} \right).$$

By combining the bounds of the two terms on the right-hand side of (3.38) and the definition of the extended domain Ω_{ext} in (3.37), we get the assertion. □

We are now ready to state the main h- and p-error estimate result.

Theorem 3.2.5. *Let $\{\mathcal{T}_n\}_{n\in\mathbb{N}}$ be a sequence of polygonal decompositions satisfying **(G1)**-**(G4)**. Further, let u and u_n be the solutions to* (3.2) *and* (3.4)*, respectively; we assume that, with a slight abuse of notation, u is the restriction to Ω of a function $u \in H^{s+1}(\Omega_{\text{ext}})$, $s \geq 1$, where Ω_{ext} is defined in* (3.37)*. Then, the following a priori h- and p-error estimate holds true:*

$$|u - u_n|_{1,\mathcal{T}_n} \leq c\, d^s \, \frac{\alpha^*(p)}{\alpha_*(p)} \, h^{\min(s,p)} \left\{ \left(\frac{\log(p)}{p} \right)^{\min_{K\in\mathcal{T}_n}(\lambda_K)\, s} + p^{-s} \right\} \|u\|_{s+1,\Omega_{\text{ext}}},$$

where c is a positive constant depending only on ρ_0, ρ_1, and Λ, d is a positive constant, $\lambda_K \pi$ denotes the smallest exterior angle of K for each $K \in \mathcal{T}_n$, and $\frac{\alpha^*(p)}{\alpha_*(p)}$ is the pollution factor appearing in (3.35)*, which is related to the choice of the stabilization.*

Proof. It is enough to combine Theorem 3.2.2 with Lemmata 3.2.3 and 3.2.4. □

Assuming, moreover, that u, the solution to the problem (3.2), is the restriction to Ω of an analytic function defined over Ω_{ext}, where Ω_{ext} was introduced in (3.37), it is possible to prove the following result.

Theorem 3.2.6. *Let **(G1)**-**(G4)** be valid and assume that u, the solution to the problem* (3.2)*, is the restriction to Ω of an analytic function defined over Ω_{ext}, given in* (3.37)*. Then, the following a priori p-error estimate holds true:*

$$|u - u_n|_{1,\mathcal{T}_n} \leq c \exp(-b\, p),$$

for some positive constants b and c, depending again only on ρ_0, ρ_1, and Λ.

Proof. The assertion follows by combining Theorem 3.2.5 with the tools employed in [39, Theorem 5.2]. □

Remark 6. We highlight that the construction involving the collection of parallelograms in (3.37) is instrumental for proving Theorem 3.2.6. In order to derive the bound of Theorem 3.2.6 from that of Theorem 3.2.5, one needs to know the explicit dependence on s of the constant in the bound of Theorem 3.2.5. This comes at the price of involving the extended domain Ω_{ext}. If one were interested in approximating solutions with finite Sobolev regularity, then there would be no need of employing the construction with the parellelograms $Q(T)$. In particular, equation (3.40) would be valid also with the norm over the triangle T, instead of over $Q(T)$, on the right-hand side. As a consequence, the bounds in Lemma 3.2.4 and in Theorem 3.2.5 would be valid also with the norm of u over Ω, instead of over Ω_{ext}, on the right-hand sides. See [39] for additional details on the hp-version in the case of the standard VEM setting.

3.2.4 Error estimates in the L^2 norm

This section is devoted to prove an upper bound for the L^2 error of method (3.4) in terms of the energy error and the best approximation error with respect to piecewise discontinuous harmonic polynomials.

To this purpose, we firstly recall the definition of nonconforming (non-Trefftz) VE spaces introduced in [18] for the approximation of the Poisson problem, and then we prove hp-best approximation estimates by functions in those spaces. The obtained results will be instrumental for proving L^2 error estimates for method (3.4). As above, we assume that p, the degree of accuracy, is equal to the nonconformity parameter appearing in (3.9).

Let $K \in \mathcal{T}_n$. We define, for $p \in \mathbb{N}$ arbitrary,

$$V(K) := \left\{ v_n \in H^1(K) \mid \Delta v_n \in \mathbb{P}_{p-2}(K),\, (\nabla v_n \cdot \mathbf{n}_K)_{|_e} \in \mathbb{P}_{p-1}(e) \;\forall e \in \mathcal{E}_n \right\}.$$

It is proven in [18, Lemma 3.1] that the following set of functionals is a set of degrees of freedom for the space $V(K)$. Given $v_n \in V(K)$, we associate the edge moments defined in (3.6)

$$\frac{1}{h_e} \int_e v_n m_\alpha^e \, \mathrm{d}s, \quad \forall \alpha = 0, \ldots, p-1, \, \forall e \in \mathcal{E}^K, \tag{3.41}$$

plus the bulk moments of the form

$$\frac{1}{|K|}\int_K v_n m_{\boldsymbol{\alpha}}\,\mathrm{d}x, \quad \forall|\boldsymbol{\alpha}| = 0,\dots,p-2, \tag{3.42}$$

where $\{m_{\boldsymbol{\alpha}}\}_{|\boldsymbol{\alpha}|=0}^{p-2}$ is *any* basis of $\mathbb{P}_{p-2}(K)$.

For all $g \in H^{\frac{1}{2}}(\Gamma)$, the global nonconforming spaces in (3.9) are defined as in the harmonic case:

$$V_{n,g}^r := \left\{ v_n \in H_g^{1,nc}(\mathcal{T}_n, r) \mid v_{n|_K} \in V(K)\ \forall K \in \mathcal{T}_n \right\}. \tag{3.43}$$

The set of global degrees of freedom is obtained by a standard nonconforming coupling of the local counterparts. The precise treatment of Dirichlet boundary conditions should be dealt with as in Remark 2.

We show that, in the H^1 seminorm, the error between a regular target function and its interpolant in the space $V_{n,g}^p$ defined in (3.43) can be estimated by the best approximation error in the space of piecewise discontinuous polynomials of degree at most p. As in Proposition 3.2.1, we have eliminated the constant 2 of the corresponding bound in [143, Prop. 3.8].

Proposition 3.2.7. *Given $g \in H^{\frac{1}{2}}(\Gamma)$, let $\psi \in V_g$, where V_g is defined in* (3.3)*. For every polygonal partition $\mathcal{T}_n$ of Ω, there exists $\psi_I \in V_{n,g}^p$, with $V_{n,g}^p$ given in* (3.43)*, such that*

$$|\psi - \psi_I|_{1,\mathcal{T}_n} \le |\psi - q_p|_{1,\mathcal{T}_n} \quad \forall q_p \in \mathcal{S}^{p,-1}(\mathcal{T}_n),$$

where $\mathcal{S}^{p,-1}(\mathcal{T}_n)$ is the space of piecewise discontinuous polynomials, that is,

$$\mathcal{S}^{p,-1}(\mathcal{T}_n) := \{q \in L^2(\Omega) : q_{|_K} \in \mathbb{P}_p(K)\,\forall K \in \mathcal{T}_n\}. \tag{3.44}$$

Proof. The proof follows the lines of that of Proposition 3.2.1. Given $\psi \in V_g$, we define $\psi_I \in V_{n,g}^p$ by imposing its degrees of freedom as follows:

$$\begin{aligned} &\frac{1}{h_e}\int_e (\psi_I - \psi) q_{p-1}^e\,\mathrm{d}s = 0 \quad \forall q_{p-1}^e \in \mathbb{P}_{p-1}(e),\, \forall e \in \mathcal{E}^K,\, \forall K \in \mathcal{T}_n \\ &\frac{1}{|K|}\int_K (\psi_I - \psi) q_{p-2}\,\mathrm{d}x = 0 \quad \forall q_{p-2} \in \mathbb{P}_{p-2}(K),\, \forall K \in \mathcal{T}_n. \end{aligned} \tag{3.45}$$

It is important to note that, since the degrees of freedom (3.41) and (3.42) are unisolvent for the space $V_{n,g}^p$, the interpolant ψ_I is defined in a unique way. Having this, we write, for all $K \in \mathcal{T}_n$,

$$|\psi - \psi_I|_{1,K}^2 = \int_K \nabla(\psi - \psi_I)\cdot\nabla(\psi - q_p)\,\mathrm{d}x + \int_K \nabla(\psi - \psi_I)\cdot\nabla(q_p - \psi_I)\,\mathrm{d}x \quad \forall q_p \in \mathcal{S}^{p,-1}(\mathcal{T}_n).$$

The second integral on the right-hand side is again zero, which can be seen as follows:

$$\begin{aligned} \int_K \nabla(\psi - \psi_I)\cdot\nabla(q_p - \psi_I)\,\mathrm{d}x &= -\int_K (\psi - \psi_I)\,\Delta(q_p - \psi_I)\,\mathrm{d}x + \int_{\partial K} (\psi - \psi_I)\,\nabla(q_p - \psi_I)\cdot\mathbf{n}_K\,\mathrm{d}s \\ &= 0, \end{aligned}$$

where we firstly integrated by parts, and then used the definition of ψ_I in (3.45). Thus, with the Cauchy-Schwarz inequality, it holds

$$|\psi - \psi_I|_{1,K}^2 \le |\psi - \psi_I|_{1,K}|\psi - q_p|_{1,K},$$

from which after dividing by $|\psi - \psi_I|_{1,K}$ and summation over all elements, the statement follows. □

We are now ready to prove a bound of the L^2 error of the method. For simplicity, we restrict ourselves here to the case that Ω is convex and is split into a collection of convex polygons. In the non-convex case, slightly worse error estimates can be proven, as discussed in Remark 7 below.

Theorem 3.2.8. *Let Ω be a polygonal convex domain and let $\{\mathcal{T}_n\}_{n\in\mathbb{N}}$ be a sequence of decompositions into convex polygons satisfying the mesh assumptions **(G1)**-**(G4)**. Let u and u_n be the solutions to* (3.2) *and* (3.4)*, respectively; we assume, with a slight abuse of notation, that u is the restriction to Ω of a function $u \in H^{s+1}(\Omega_{\text{ext}})$, $s \geq 1$, where Ω_{ext} is defined in* (3.37)*. Then,*

$$\|u-u_n\|_{0,\Omega} \leq c\left\{\frac{h^{\min(s,p)+1}}{p^{s+1}}\|u\|_{s+1,\Omega_{\text{ext}}}\right.$$
$$\left.+\max\left(\frac{h}{p}, h\,\alpha^*(p)\left(\frac{\log(p)}{p}\right)^{\max_{K\in\mathcal{T}_n}\lambda_K}\right)\left(|u-u_n|_{1,\mathcal{T}_n}+\inf_{q_p^\Delta\in\mathcal{S}^{p,\Delta,-1}(\mathcal{T}_n)}|u-q_p^\Delta|_{1,\mathcal{T}_n}\right)\right\},$$

where c is a positive constant depending only on ρ_0, ρ_1, ρ_2, and Λ, $\alpha^*(p)$ is the "upper" stability constant appearing in (3.19)*, $\mathcal{S}^{p,\Delta,-1}(\mathcal{T}_n)$ is defined in* (3.31)*, and $\lambda_K\pi$ denotes the smallest exterior angle of K for each $K \in \mathcal{T}_n$.*

Proof. We consider the following dual problem: Find $\psi \in H^1(\Omega)$ such that

$$\begin{cases}-\Delta\psi = u-u_n & \text{in } \Omega\\ \psi = 0 & \text{on } \Gamma.\end{cases} \tag{3.46}$$

Standard stability and a priori regularity theory implies that $\psi \in H^2(\Omega)$ and

$$\|\psi\|_{2,\Omega} \lesssim \|u-u_n\|_{0,\Omega}, \tag{3.47}$$

where the hidden constant depends only on the domain Ω, see e.g. [116, Theorem 3.2.1.2].

Using (3.46) and (3.34), and taking into account that $u-u_n \in H_0^{1,nc}(\mathcal{T}_n,p)$, we obtain the following equivalent expression for the L^2 error:

$$\begin{aligned}\|u-u_n\|_{0,\Omega}^2 &= \sum_{K\in\mathcal{T}_n}\int_K(-\Delta\psi)(u-u_n)\,\mathrm{d}x\\ &= \sum_{K\in\mathcal{T}_n}\left\{\int_K\nabla\psi\cdot\nabla(u-u_n)\,\mathrm{d}x-\int_{\partial K}\nabla\psi\cdot\mathbf{n}_K\,(u-u_n)\,\mathrm{d}s\right\}\\ &= \sum_{K\in\mathcal{T}_n}a_0^K(\psi-\psi_I,u-u_n)+\sum_{K\in\mathcal{T}_n}a_0^K(\psi_I,u-u_n)-\mathcal{N}_n(\psi,u-u_n)\\ &=: T_1+T_2+T_3,\end{aligned} \tag{3.48}$$

where ψ_I is the (unique) function in $V_{n,0}^p$, the enlarged space of functions with zero Dirichlet traces introduced in (3.43), defined from ψ via (3.45); in particular, ψ_I is not piecewise harmonic, in general.

We begin by estimating term T_1. Owing to the Cauchy-Schwarz inequality and Proposition 3.2.7, we have

$$|T_1| \leq |\psi-\psi_I|_{1,\mathcal{T}_n}|u-u_n|_{1,\mathcal{T}_n} \leq |\psi-q_p|_{1,\mathcal{T}_n}|u-u_n|_{1,\mathcal{T}_n} \quad \forall q_p \in \mathcal{S}^{p,-1}(\mathcal{T}_n),$$

where $\mathcal{S}^{p,-1}(\mathcal{T}_n)$ is the space of piecewise discontinuous polynomials introduced in (3.44). By taking q_p equal to the best approximation of ψ in $\mathcal{S}^{p,-1}(\mathcal{T}_n)$ and using [39, Lemma 4.2], together with (3.47), we have

$$|T_1| \lesssim \frac{h}{p}\|\psi\|_{2,\Omega}|u-u_n|_{1,\mathcal{T}_n} \lesssim \frac{h}{p}\|u-u_n\|_{0,\Omega}|u-u_n|_{1,\mathcal{T}_n}.$$

Next, we focus on term T_3 on the right-hand side of (3.48). Following the same steps as in the proof of Lemma 3.2.4, we obtain

$$|T_3| = |\mathcal{N}_n(\psi,u-u_n)| \leq \sum_{e\in\mathcal{E}_n}\left\|\nabla\psi-\Pi_{p-1}^{0,e}(\nabla\psi)\right\|_{0,e}\left\|[\![u-u_n]\!]-\Pi_{p-1}^{0,e}[\![u-u_n]\!]\right\|_{0,e},$$

where $\Pi^{0,e}_{p-1}$ denotes here again, with an abuse of notation, the L^2 projector onto vectorial polynomial spaces. Applying [78, Theorem 3.1] and [39, Lemma 4.1] similarly as in the proof of Lemma 3.2.4, together with (3.47) ($|\nabla\psi|_{1,K} \le \|\psi\|_{2,K}$), we get

$$|T_3| \lesssim \frac{h}{p}\|u-u_n\|_{0,\Omega}|u-u_n|_{1,\mathcal{T}_n}.$$

Finally, we study term T_2 on the right-hand side of (3.48), which can be split as

$$T_2 = \sum_{K\in\mathcal{T}_n} a_0^K(\psi_I, u-u_n) = \sum_{K\in\mathcal{T}_n} a_0^K(u,\psi_I) - \sum_{K\in\mathcal{T}_n} a_0^K(u_n,\psi_I) =: T_4 + T_5. \tag{3.49}$$

The first term T_4 is related to the nonconformity of the discretization spaces, whereas the second term T_5 reflects the fact that method (3.4) does not employ the original bilinear form.

We start to estimate term T_4. Using computations analogous to those in the proof of Lemma 3.2.4, it is possible to deduce

$$\begin{aligned}|T_4| &= \left|\sum_{K\in\mathcal{T}_n} a_0^K(u,\psi_I)\right| = \left|\sum_{K\in\mathcal{T}_n}\int_{\partial K}\nabla u\cdot\mathbf{n}_K\,\psi_I\,\mathrm{d}s\right| = |\mathcal{N}_n(u,\psi_I)| = |\mathcal{N}_n(u,\psi_I-\psi)|\\ &\le \sum_{e\in\mathcal{E}_n}\left\|\nabla u - \Pi^{0,e}_{p-1}(\nabla u)\right\|_{0,e}\left\|[\![\psi_I-\psi]\!] - \Pi^{0,e}_{p-1}[\![\psi_I-\psi]\!]\right\|_{0,e},\end{aligned}$$

where in the fourth identity we used the fact that $\mathcal{N}_n(u,\psi)=0$, which holds since u and ψ are sufficiently regular, and in the last step we used (3.38). Again, $\Pi^{0,e}_{p-1}$ has to be understood as the L^2 projection onto the vectorial polynomial spaces of degree at most $p-1$ on e. Applying [78, Theorem 3.1], Proposition 3.2.7, [39, Lemma 4.2], and finally (3.47), leads to

$$|T_4| \lesssim p^{-1}|\nabla(u-\Pi^\nabla_p u)|_{1,\mathcal{T}_n}|\psi-\psi_I|_{1,\mathcal{T}_n} \lesssim \frac{h^{\min(s,p)}}{p^s}\|u\|_{s+1,\Omega_{\mathrm{ext}}}\frac{h}{p}\|u-u_n\|_{0,\Omega},$$

where we recall that Ω_{ext} is defined in (3.37) and where Π^∇_p is any piecewise energy projector from $H^1(K)$ into $\mathbb{P}_p(K)$, for all $K\in\mathcal{T}_n$.

Finally, it remains to treat term T_5 on the right-hand side of (3.49). To this purpose, we consider the following splittings of ψ and ψ_I. Firstly, we split ψ into $\psi=\psi^1+\psi^2$, where ψ^1 and ψ^2 are, element by element, solutions to the local problems

$$\begin{cases}-\Delta\psi^1 = -\Delta\psi & \text{in } K\\ \quad\psi^1 = 0 & \text{on } \partial K,\end{cases}\qquad \begin{cases}-\Delta\psi^2 = 0 & \text{in } K\\ \quad\psi^2 = \psi & \text{on } \partial K,\end{cases} \tag{3.50}$$

for all $K\in\mathcal{T}_n$. Using (3.46), we can also observe that $\psi^2-\psi$ solves the local problems

$$\begin{cases}-\Delta(\psi-\psi^2) = u-u_n & \text{in } K\\ \quad\psi-\psi^2 = 0 & \text{on } \partial K.\end{cases}$$

Then, (local) standard *a priori* regularity theory and, afterwards, summation over all elements $K\in\mathcal{T}_n$ imply the global bound

$$\|\psi^2-\psi\|_{2,\mathcal{T}_n} \lesssim \|u-u_n\|_{0,\Omega}, \tag{3.51}$$

where the broken norm $\|\cdot\|_{2,\mathcal{T}_n}$ is defined in (2.31). With the triangle inequality, (3.47), and (3.51), we get

$$\|\psi^2\|_{2,\mathcal{T}_n} \le \|\psi-\psi^2\|_{2,\mathcal{T}_n} + \|\psi\|_{2,\Omega} \lesssim \|u-u_n\|_{0,\Omega}. \tag{3.52}$$

Secondly, we split $\psi_I\in V^p_{n,0}$ into $\psi_I=\psi_I^1+\psi_I^2$. We define ψ_I^2 as the unique element in $V^{\Delta,p}_{n,0}$ introduced in (3.12), which satisfies

$$\frac{1}{h_e}\int_e \psi_I^2 q^e_{p-1}\,\mathrm{d}s = \frac{1}{h_e}\int_e \psi_I q^e_{p-1}\,\mathrm{d}s \quad \forall q^e_{p-1}\in\mathbb{P}_{p-1}(e), \forall e\in\mathcal{E}_n. \tag{3.53}$$

Existence and uniqueness of ψ_I^2 follow from the fact that ψ_I^2 is defined via unisolvent degrees of freedom for the space $V_{n,0}^{\Delta,p}$. Owing to (3.53), the definition of ψ_I in (3.45), and (3.50), we deduce

$$\frac{1}{h_e}\int_e \psi_I^2 q_{p-1}^e \,\mathrm{d}s = \frac{1}{h_e}\int_e \psi_I q_{p-1}^e \,\mathrm{d}s = \frac{1}{h_e}\int_e \psi q_{p-1}^e \,\mathrm{d}s = \frac{1}{h_e}\int_e \psi^2 q_{p-1}^e \,\mathrm{d}s \quad \forall q_{p-1}^e \in \mathbb{P}_{p-1}(e),$$

on every edge $e \in \mathcal{E}_n$. This entails that ψ_I^2 approximates ψ^2 in the sense of Proposition 3.2.1. Having this, the function $\psi_I^1 = \psi_I - \psi_I^2 \in V_{n,0}^p$ satisfies

$$\begin{cases} \dfrac{1}{|e|}\displaystyle\int_e \psi_I^1 q_{p-1}^e \,\mathrm{d}s = 0 & \forall q_{p-1}^e \in \mathbb{P}_{p-1}(e),\, \forall e \in \mathcal{E}^K,\, \forall K \in \mathcal{T}_n, \\ \dfrac{1}{|K|}\displaystyle\int_K \psi_I^1 q_{p-2} \,\mathrm{d}x = \dfrac{1}{|K|}\displaystyle\int_K (\psi_I - \psi_I^2) q_{p-2} \,\mathrm{d}x & \forall q_{p-2} \in \mathbb{P}_{p-2}(K),\, \forall K \in \mathcal{T}_n. \end{cases}$$

Moreover, since $u_n \in V_{n,g}^{\Delta,p}$, ψ_I^1 has the essential feature that it satisfies

$$a_0^K(u_n, \psi_I^1) = \int_K \underbrace{(-\Delta u_n)}_{=0} \psi_I^1 \,\mathrm{d}x + \underbrace{\int_{\partial K} (\nabla u_n \cdot \mathbf{n}_K) \psi_I^1 \,\mathrm{d}s}_{=0} = 0. \tag{3.54}$$

We have now all the tools for estimating term T_5. Using (3.54), (3.4), and (3.18), we get

$$\begin{aligned} |T_5| &= \left| \sum_{K\in\mathcal{T}_n} a_0^K(u_n, \psi_I^2) \right| = \left| \sum_{K\in\mathcal{T}_n} \left\{ a_{0,n}^K(u_n, \psi_I^2) - a_0^K(u_n, \psi_I^2) \right\} \right| \\ &= \left| \sum_{K\in\mathcal{T}_n} \left\{ a_{0,n}^K(u_n - q_p^\Delta, \psi_I^2 - \widetilde{q}_p^\Delta) - a_0^K(u_n - q_p^\Delta, \psi_I^2 - \widetilde{q}_p^\Delta) \right\} \right| \quad \forall q_p^\Delta,\, \widetilde{q}_p^\Delta \in \mathcal{S}^{p,\Delta,-1}(\mathcal{T}_n), \end{aligned}$$

where we recall that $\mathcal{S}^{p,\Delta,-1}(\mathcal{T}_n)$ is defined in (3.31). It is important to highlight that it is in fact a key point of the error analysis to have piecewise harmonic functions in both entries of the discrete bilinear form. By applying the continuity property (3.20) and the Cauchy-Schwarz inequality, then the triangle inequality and Proposition 3.2.1, we deduce

$$\begin{aligned} |T_5| &\lesssim \alpha^*(p) |u_n - q_p^\Delta|_{1,\mathcal{T}_n} |\psi_I^2 - \widetilde{q}_p^\Delta|_{1,\mathcal{T}_n} \\ &\leq \alpha^*(p)(|u - u_n|_{1,\mathcal{T}_n} + |u - q_p^\Delta|_{1,\mathcal{T}_n})(|\psi^2 - \psi_I^2|_{1,\mathcal{T}_n} + |\psi^2 - \widetilde{q}_p^\Delta|_{1,\mathcal{T}_n}) \\ &\lesssim \alpha^*(p)(|u - u_n|_{1,\mathcal{T}_n} + |u - q_p^\Delta|_{1,\mathcal{T}_n})|\psi^2 - \widetilde{q}_p^\Delta|_{1,\mathcal{T}_n}. \end{aligned}$$

Thanks to Lemma 3.2.3 (here, $s = 1$) and the bound (3.52), we have

$$\begin{aligned} |T_5| &\lesssim \alpha^*(p)(|u - u_n|_{1,\mathcal{T}_n} + |u - q_p^\Delta|_{1,\mathcal{T}_n})\, h \left(\frac{\log(p)}{p} \right)^{\min_{K\in\mathcal{T}_n} \lambda_K} \left(\sum_{K\in\mathcal{T}_n} \|\psi^2\|_{2,K}^2 \right)^{\frac{1}{2}} \\ &\lesssim \alpha^*(p)(|u - u_n|_{1,\mathcal{T}_n} + |u - q_p^\Delta|_{1,\mathcal{T}_n}) h \left(\frac{\log(p)}{p} \right)^{\min_{K\in\mathcal{T}_n} \lambda_K} \|u - u_n\|_{0,\Omega}, \end{aligned}$$

where we recall that, for any $K \in \mathcal{T}_n$, $\lambda_K \pi$ denotes the smallest exterior angle of K.

By combining the estimates on all the terms T_1 to T_5, we get the assertion. □

Remark 7. As already highlighted, the case of non-convex Ω can be treated analogously. More precisely, given ω the largest reentrant angle of Ω, the solution to (3.2) belongs to $H^{1+t}(\Omega)$, with $t = \frac{\pi}{\omega} - \varepsilon$ for all $\varepsilon > 0$ arbitrarily small. Standard stability and *a priori* regularity theory, see [21, Theorem 2.1], gives

$$\|\psi\|_{1+t,\Omega} \leq c\|u - u_n\|_{0,\Omega}$$

for some positive constant c depending only on the domain Ω. An analogous bound is valid for the counterpart of (3.51) in the non-convex case. Having this, a straightforward modification of the proof of Theorem 3.2.8 leads to the h- and p-error bounds

$$\begin{aligned}\|u-u_n\|_{0,\Omega} \le \Bigg\{ & c\,\frac{h^{\min(s,p)+t}}{p^{s+t}}\|u\|_{s+1,\Omega_{\mathrm{ext}}} \\ & + \max\left(\left(\frac{h}{p}\right)^t,\, h^t\,\alpha^*(p)\left(\frac{\log(p)}{p}\right)^{\max_{K\in\mathcal{T}_n}(\lambda_K)\,t}\right) \\ & \cdot\left(|u-u_n|_{1,\mathcal{T}_n} + \inf_{q_p^\Delta\in\mathcal{S}^{p,\Delta,-1}(\mathcal{T}_n)}|u-q_p^\Delta|_{1,\mathcal{T}_n}\right)\Bigg\},\end{aligned}$$

where c is a positive constant depending only on the constants ρ_0, ρ_1, ρ_2, and Λ appearing in (**G1**)-(**G4**) and in the proof of Lemma 3.2.4.

The presence of non-convex polygons in the decomposition $\mathcal{T}_n$ leads to a possible additional loss in the convergence rate in p of the L^2 error, which will depend on the largest interior and exterior angles of the polygons.

3.2.5 Hints for the extension to the 3D case

Here, we give a hint concerning the extension of the method and its corresponding analysis to 3D.

Regarding the definition of local Trefftz VE spaces, one mimics the strategy suggested in [18] and defines, for every polyhedron K in $\mathbb{R}^3$ and any fixed $p\in\mathbb{N}$,

$$V^\Delta(K) := \left\{v_n\in H^1(K)\mid \Delta v_n = 0 \text{ in } K,\ (\nabla v_n\cdot\mathbf{n}_K)_{|_F}\in\mathbb{P}_{p-1}(F)\ \forall F \text{ faces of } K\right\}.$$

We observe that the definition of the local 3D space is a straightforward extension of its 2D counterpart. This is however not the case when using *conforming* VEM. In that situation, typically, one also requires to have a modified version of the local VE spaces on each face, see [3]. On the one hand, this allows the construction of continuous functions over the boundary of a polyhedron, as well as the construction of projectors onto proper polynomial spaces; on the other, it complicates the p-analysis of the method. In the nonconforming framework, however, one does not need to fix any sort of continuity across the interface between faces of a polyhedron and thus it suffices to impose that normal derivatives are polynomials.

The global 3D nonconforming space is built as in the 2D case. Also, the degrees of freedom are given by scaled face moments with respect to polynomials up to order $p-1$.

Next, the abstract definition of the 2D local discrete bilinear form in (3.17) can also be employed in the 3D case. The (properly scaled) 3D counterpart of the 2D explicit stabilization defined in (3.22) would be

$$S_0^K(u_n,v_n) = \sum_{F \text{ faces of } K}\frac{p}{h_F}(\Pi_{p-1}^{0,F}u_n,\Pi_{p-1}^{0,F}v_n)_{0,F},$$

where, for any face F, $\Pi_{p-1}^{0,F}$ denotes the L^2 projector onto $\mathbb{P}_{p-1}(F)$ of the traces on F of functions in the 3D VE space. Nonetheless, it is not clear whether explicit bounds in terms of p of the stability constants appearing in (3.16) can be proven for this form. In fact, in the 2D case, hp-polynomial inverse estimates in 1D were the key tool for proving Theorem 3.1.1. In the 3D framework, one needs to employ hp-polynomial inverse estimates on general polygons based on weighted norms. We highlight that the approach of [66, Chapter 3], see also [65], could be followed in order to prove such hp-weighted inverse inequalities. However, as this extension is quite technical, we do not investigate it here.

Independently of the specific choice of the stabilization, provided that it is symmetric and satisfies (3.19), the abstract error analysis is dealt with similarly to the 2D case, see Theorem 3.2.2. The only modification is in the definition of the nonconformity term, which in 3D is defined as

$$\mathcal{N}_n(u,v) = \sum_{F\in\mathcal{E}_n^3}\int_F \nabla u\cdot[\![v]\!]_F\,\mathrm{d}s$$

for all conforming functions u and all nonconforming functions v, where $\mathcal{E}_n^3$ denotes the set of faces in the polyhedral decomposition, and $[\![\cdot]\!]_F$ is defined as in (3.8) in terms of normal derivatives over faces.

The proof of h- and p-error bounds for this nonconforming term follows the same lines as in the 2D case since [78, Theorem 3.1] holds true on simplices in arbitrary space dimension. For the best approximation error, one should use the 3D version of Lemma 3.2.3, which can be found e.g. in [150, Theorem 3.12].

3.3 Numerical results

We firstly discuss some details on the implementation in Section 3.3.1. Then, in Section 3.3.2, we present numerical tests for the h-version and the p-version of the method, validating the theoretical results obtained in Section 3.1. We conclude with a discussion and some tests on the hp-version in Section 3.3.3.

3.3.1 Details on the implementation

In this section, we discuss some practical aspects concerning the implementation of the nonconforming Trefftz VEM in 2D. We employ henceforth the notation of [36]. It is worth to underline that we present herein only the case with uniform degree of accuracy. As a first step, we begin by fixing the notation for the various bases instrumental for the construction of the method.

Basis of $\mathbb{P}_{p-1}(e)$ for a given $e \in \mathcal{E}^K$. Using the same notation as in (3.6), we denote by $\{m_r^e\}_{r=0,\dots,p-1}$ the basis of $\mathbb{P}_{p-1}(e)$, $e \in \mathcal{E}^K$. The choice we make is

$$m_r^e(\mathbf{x}) := \mathbb{L}_r\left(\phi_e^{-1}(\mathbf{x})\right) \quad \forall r = 0, \dots, p-1, \tag{3.55}$$

where $\phi_e : [-1, 1] \to e$ is the linear transformation mapping the interval $[-1, 1]$ to the edge e, and $\mathbb{L}_r$ is the Legendre polynomial of degree r over $[-1, 1]$. We recall, see e.g. [166], for future use the orthogonality property

$$(m_r^e, m_s^e)_{0,e} = \frac{h_e}{2}\int_{-1}^{1} \mathbb{L}_r(t)\mathbb{L}_s(t)\,\mathrm{d}t = \frac{h_e}{2r+1}\delta_{rs} \quad \forall r, s = 0, \dots, p-1, \tag{3.56}$$

where δ_{rs} is the Kronecker delta (1 if $r = s$, 0 otherwise).

Basis of $\mathbb{H}_p(K)$ for a given $K \in \mathcal{T}_n$. We denote by $\{q_\alpha^\Delta\}_{\alpha=1,\dots,n_p^\Delta}$ the basis of the space of harmonic polynomials $\mathbb{H}_p(K)$, where $n_p^\Delta := \dim \mathbb{H}_p(K) = 2p + 1$. The choice we make is

$$\begin{aligned}
q_1^\Delta(\mathbf{x}) &= 1;\\
q_{2\ell}^\Delta(\mathbf{x}) &= \sum_{k=1,\, k \text{ odd}}^{\ell} (-1)^{\frac{k-1}{2}} \binom{\ell}{k} \left(\frac{x - x_K}{h_K}\right)^{\ell-k} \left(\frac{y - y_K}{h_K}\right)^{k} \quad \forall \ell = 1, \dots, p;\\
q_{2\ell+1}^\Delta(\mathbf{x}) &= \sum_{k=0,\, k \text{ even}}^{\ell} (-1)^{\frac{k}{2}} \binom{\ell}{k} \left(\frac{x - x_K}{h_K}\right)^{\ell-k} \left(\frac{y - y_K}{h_K}\right)^{k} \quad \forall \ell = 1, \dots, p.
\end{aligned}$$

The fact that this is actually a basis for $\mathbb{H}_p(K)$ is proven in e.g. [17, Theorem 5.24].

Basis for $V^\Delta(K)$ for a given $K \in \mathcal{T}_n$. For this local VE space introduced in (3.5), we employ the canonical basis $\{\varphi_{j,r}\}_{\substack{j=1,\dots,n_K\\ r=0,\dots,p-1}}$ defined though (3.7), where we recall that n_K denotes the number of edges of K.

In the following, we derive the matrix representation of the local discrete bilinear form introduced in (3.17). We begin with the computation of the matrix representation of the projector $\Pi_p^{\nabla,K}$ acting from $V(K)$ to $\mathbb{H}_p(K)$ and defined in (3.14). To this purpose, given any basis function $\varphi_{j,r} \in V^\Delta(K)$, $j = 1, \dots, n_K$, $r = 0, \dots, p-1$, we expand $\Pi_p^{\nabla,K}\varphi_{j,r}$ in terms of basis $\{q_\alpha^\Delta\}_{\alpha=1,\dots,n_p^\Delta}$ of $\mathbb{H}_p(K)$, i.e.

$$\Pi_p^{\nabla,K}\varphi_{j,r} = \sum_{\alpha=1}^{n_p^\Delta} s_\alpha^{(j,r)} q_\alpha^\Delta. \tag{3.57}$$

Using (3.14) and testing (3.57) with functions q_β^Δ, $\beta = 1, \dots, n_p^\Delta$, we get that the coefficients $s_\alpha^{(j,r)}$ can be computed by solving for $\boldsymbol{s}^{(j,r)} := [s_1^{(j,r)}, \dots, s_{n_p^\Delta}^{(j,r)}]^T$ the $n_p^\Delta \times n_p^\Delta$ algebraic linear system

$$\boldsymbol{G}\boldsymbol{s}^{(j,r)} = \boldsymbol{b}^{(j,r)},$$

where

$$\boldsymbol{G} = \begin{bmatrix} (q_1^\Delta, 1)_{0,\partial K} & (q_2^\Delta, 1)_{0,\partial K} & \cdots & (q_{n_p^\Delta}^\Delta, 1)_{0,\partial K} \\ 0 & (\nabla q_2^\Delta, \nabla q_2^\Delta)_{0,K} & \cdots & (\nabla q_{n_p^\Delta}^\Delta, \nabla q_2^\Delta)_{0,K} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & (\nabla q_{n_p^\Delta}^\Delta, \nabla q_2^\Delta)_{0,K} & \cdots & (\nabla q_{n_p^\Delta}^\Delta, \nabla q_{n_p^\Delta}^\Delta)_{0,K} \end{bmatrix}, \quad \boldsymbol{b}^{(j,r)} = \begin{bmatrix} (\varphi_{j,r}, 1)_{0,\partial K} \\ (\nabla\varphi_{j,r}, \nabla q_2^\Delta)_{0,K} \\ \vdots \\ (\nabla\varphi_{j,r}, \nabla q_{n_p^\Delta}^\Delta)_{0,K} \end{bmatrix}.$$

Collecting all the $n_K p$ (column) vectors $\boldsymbol{b}^{(j,r)}$ in a matrix $\boldsymbol{B} := [\boldsymbol{b}^{(1,1)}, \dots, \boldsymbol{b}^{(n_K,p)}] \in \mathbb{R}^{n_p^\Delta \times n_K p}$, the matrix representation $\boldsymbol{\Pi}_*$ of the projector $\Pi_p^{\nabla,K}$ acting from $V^\Delta(K)$ to $\mathbb{H}_p(K)$ is given by

$$\boldsymbol{\Pi}_* = \boldsymbol{G}^{-1}\boldsymbol{B} \in \mathbb{R}^{n_p^\Delta \times n_K p}.$$

Subsequently, we define

$$\boldsymbol{D} := \begin{bmatrix} \mathrm{dof}_{1,1}(q_1^\Delta) & \cdots & \mathrm{dof}_{1,1}(q_{n_p^\Delta}^\Delta) \\ \vdots & \ddots & \vdots \\ \mathrm{dof}_{n_K,p}(q_1^\Delta) & \cdots & \mathrm{dof}_{n_K,p}(q_{n_p^\Delta}^\Delta) \end{bmatrix} \in \mathbb{R}^{n_K p \times n_p^\Delta}.$$

Let $\boldsymbol{\Pi}$ be the matrix representation of the operator $\Pi_p^{\nabla,K}$ seen now as a map from $V^\Delta(K)$ into $V^\Delta(K) \supseteq \mathbb{H}_p(K)$. Then, following [36], it is possible to show that

$$\boldsymbol{\Pi} = \boldsymbol{D}\boldsymbol{G}^{-1}\boldsymbol{B} \in \mathbb{R}^{n_K p \times n_K p}.$$

Next, denoting by $\widetilde{\boldsymbol{G}} \in \mathbb{R}^{n_p^\Delta \times n_p^\Delta}$ the matrix coinciding with $\boldsymbol{G}$ apart from the first row which is set to zero, the matrix representation of the bilinear form in (3.17) is

$$(\boldsymbol{\Pi}_*)^T \, \widetilde{\boldsymbol{G}} \, (\boldsymbol{\Pi}_*) + (\boldsymbol{I} - \boldsymbol{\Pi})^T \, \boldsymbol{S} \, (\boldsymbol{I} - \boldsymbol{\Pi}).$$

Here, $\boldsymbol{S}$ denotes the matrix representation of an explicit stabilization $S^K(\cdot,\cdot)$. For the stabilization defined in (3.22), we have, for all $k, \ell = 1, \dots, n_K$ and $r, s = 0, \dots, p-1$,

$$\boldsymbol{S}((k-1)n_K + r, (\ell-1)n_K + s) = \sum_{i=1}^{n_K} \frac{p}{h_{e_i}} (\Pi_{p-1}^{0,e_i}\varphi_{\ell,s}, \Pi_{p-1}^{0,e_i}\varphi_{k,r})_{0,e_i}.$$

By expanding $\Pi_{p-1}^{0,e_i}\varphi_{\ell,s}$ and $\Pi_{p-1}^{0,e_i}\varphi_{k,r}$ in the basis $\{m_\gamma^{e_i}\}_{\gamma=0,\dots,p-1}$ of $\mathbb{P}_{p-1}(e_i)$, i.e.

$$\Pi_{p-1}^{0,e_i}\varphi_{\ell,s} = \sum_{\gamma=0}^{p-1} t_\gamma^{(\ell,s),e_i} m_\gamma^{e_i}, \quad \Pi_{p-1}^{0,e_i}\varphi_{k,r} = \sum_{\zeta=0}^{p-1} t_\zeta^{(k,r),e_i} m_\zeta^{e_i}, \tag{3.58}$$

we can write

$$\boldsymbol{S}((k-1)n_K + r, (\ell-1)n_K + s) = \sum_{i=1}^{n_K} \sum_{\gamma=0}^{p-1} \sum_{\zeta=0}^{p-1} t_\gamma^{(\ell,s),e_i} t_\zeta^{(k,r),e_i} \frac{p}{h_{e_i}} (m_\gamma^{e_i}, m_\zeta^{e_i})_{0,e_i}.$$

For the basis defined in (3.55), using the orthogonality of the Legendre polynomials (3.56), this expression can be simplified leading to a diagonal stability matrix $\mathbf{S}$:

$$\boldsymbol{S}((k-1)n_K + r, (k-1)n_K + r) = \sum_{i=1}^{n_K} \sum_{\zeta=0}^{p-1} \frac{p}{2r+1} (t_\zeta^{(k,r),e_i})^2.$$

For fixed $i, k \in \{1, \dots, n_K\}$ and $r \in \{0, \dots, p-1\}$, the coefficients $t_\zeta^{(k,r),e_i}$ are obtained by testing $\Pi_{p-1}^{0,e_i}\varphi_{k,r}$, defined in (3.58), with $m_\zeta^{e_i}$, $\zeta = 0, \dots, p-1$, and by taking into account the definition of Π_{p-1}^{0,e_i} in (3.13), the orthogonality relation (3.56) and the definition of $\varphi_{k,r}$ in (3.7). This gives

$$t_\zeta^{(k,r),e_i} = \frac{2\zeta+1}{h_{e_i}}(\varphi_{k,r}, m_\zeta^{e_i})_{0,e_i} = (2\zeta+1)\delta_{ik}\delta_{r\zeta} \quad \forall \zeta = 0, \dots, p-1.$$

The global system of linear equations corresponding to method (3.4) is assembled as in the standard nonconforming FEM. Finally, one imposes the Dirichlet boundary datum g in a nonconforming fashion by

$$\int_e u_n q_{p-1}^e \,\mathrm{d}s = \int_e g q_{p-1}^e \,\mathrm{d}s \quad \forall q_{p-1}^e \in \mathbb{P}_{p-1}(e),$$

where, in practice, g is replaced by g_p, see Remark 2.

3.3.2 Numerical results: h- and p-version

In this section, we present numerical experiments validating the theoretical error estimates in the L^2 and $H^1(\mathcal{T}_n)$ (H^1, for short) norms discussed in Theorems 3.2.5, 3.2.6, and 3.2.8.

For the following numerical experiments, we consider boundary value problems of the form (3.1), on $\Omega := (0,1)^2$, with known exact solutions given by

- $u_1(x,y) = e^x \sin(y)$,
- $u_2(x,y) = u_2(r,\theta) = r^2\,(\log(r)\sin(2\theta) + \theta\cos(2\theta))$.

We underline that u_1 is an analytic function in Ω, whereas $u_2 \in H^{3-\epsilon}(\Omega)$ for every $\epsilon > 0$ arbitrarily small; moreover, u_2 represents the natural singular solution at $\mathbf{0} = (0,0)$ of the Poisson problem on a square domain, see e.g. [21].

We discretize these problems on sequences of quasi-uniform Cartesian meshes and Voronoi-Lloyd meshes of the type shown in Figure 2.2, left and center, respectively. Moreover, we test on a problem with exact solution u_1 on the domain Ω given by the union of four Escher horses as in Figure 2.2, right.

It is important to note that, since an explicit representation of the numerical approximation u_n inside each element is not available, due to the "virtuality" of the basis functions, we cannot compute the L^2 and H^1 errors of the method directly. Instead, we compute the following relative errors between u and $\Pi_p^\nabla u_n$, where Π_p^∇ is defined in (3.14):

$$\frac{\|u - \Pi_p^\nabla u_n\|_{0,\Omega}}{\|u\|_{0,\Omega}}, \qquad \frac{\|u - \Pi_p^\nabla u_n\|_{1,\mathcal{T}_n}}{\|u\|_{1,\Omega}}. \tag{3.59}$$

We observe that the "computable" H^1 error in (3.59) is related to the exact H^1 error. In fact, thanks to Theorem 3.2.2, we have

$$\begin{aligned} |u - u_n|_{1,\mathcal{T}_n} &\lesssim \inf_{q_p^\Delta \in \mathcal{S}^{p,\Delta,-1}(\mathcal{T}_n)} |u - q_p^\Delta|_{1,\mathcal{T}_n} + \sup_{v_n \in V_{n,0}^{\Delta,p}} \frac{\mathcal{N}_n(u, v_n)}{|v_n|_{1,\mathcal{T}_n}} \\ &\le |u - \Pi_p^\nabla u_n|_{1,\mathcal{T}_n} + \sup_{v_n \in V_{n,0}^{\Delta,p}} \frac{\mathcal{N}_n(u, v_n)}{|v_n|_{1,\mathcal{T}_n}}; \end{aligned}$$

the convergence of the second term on the right-hand side is provided in Lemma 3.2.4. Moreover, by the triangle inequality and the stability of the H^1-projection, one also has

$$\begin{aligned} |u - \Pi_p^\nabla u_n|_{1,\mathcal{T}_n} &\le |u - \Pi_p^\nabla u|_{1,\mathcal{T}_n} + |\Pi_p^\nabla(u - u_n)|_{1,\mathcal{T}_n} \\ &\le |u - \Pi_p^\nabla u|_{1,\mathcal{T}_n} + |u - u_n|_{1,\mathcal{T}_n}; \end{aligned}$$

the convergence of the second term on the right-hand side is provided in Lemma 3.2.3.

Numerical results: h-version

In this section, we verify the algebraic rate of convergence of the h-version of the method, validating thus Theorems 3.2.5 and 3.2.8 for different degrees of accuracy $p = 1, 2, 3, 4, 5$.

The numerical results for the problems in $\Omega = (0, 1)^2$ with exact solutions u_1 and u_2, obtained on sequences of Cartesian and Voronoi-Lloyd meshes, are depicted in Figure 3.2 and Figure 3.3.

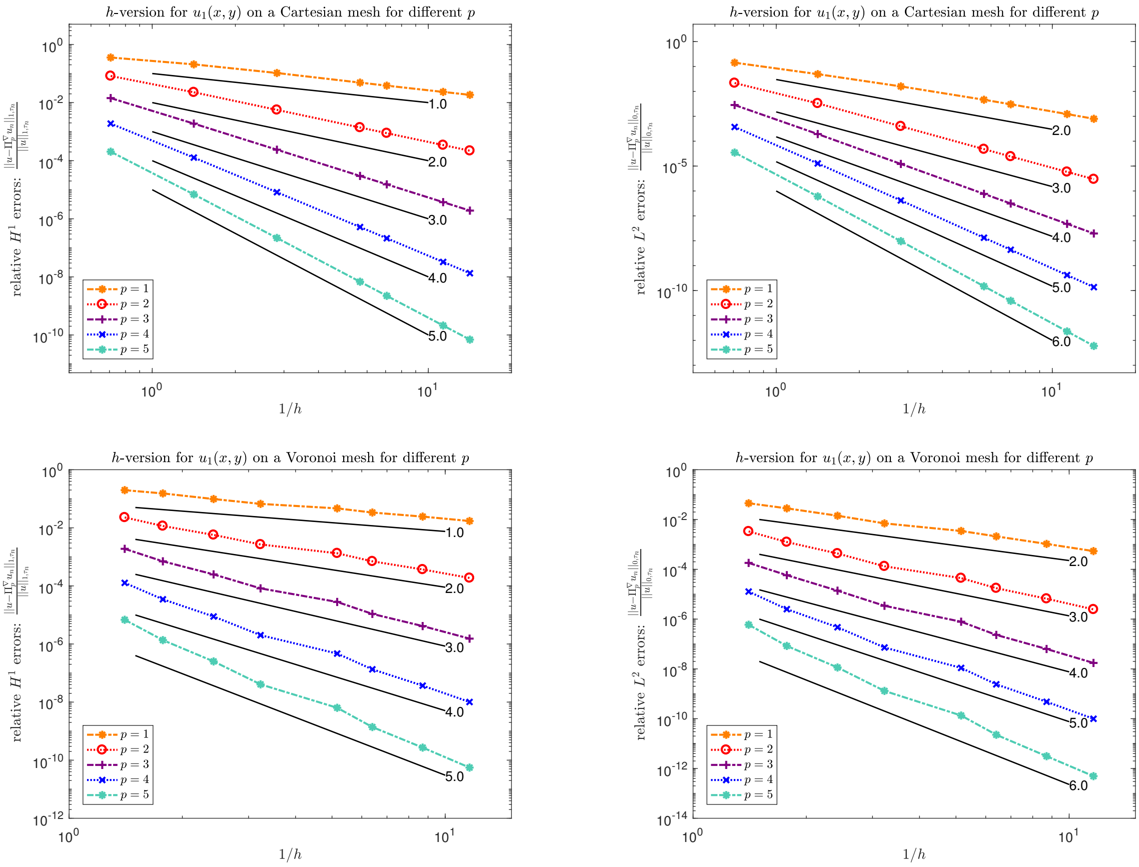


Figure 3.2: Convergence of the h-version of the method for the analytic solution u_1 on quasi-uniform Cartesian (*first row*) and Voronoi-Lloyd (*second row*) meshes; relative H^1 errors (*left*) and relative L^2 errors (*right*) defined in (3.59).

From Theorems 3.2.5 and 3.2.8, we expect the H^1 and L^2 errors to behave like $\mathcal{O}(h^{\min(t,p)})$ and $\mathcal{O}(h^{\min(t,p)+1})$, respectively, where $t+1$ is the regularity of the exact solution u, and p is the degree of accuracy. The numerical results in Figures 3.2 and 3.3 are in agreement with these theoretical estimates. In fact, for u_1, which belongs to $H^s(\Omega)$ for all $s \geqslant 0$, we see that the H^1 error actually converges with order $\mathcal{O}(h^p)$, and the L^2 error with order $\mathcal{O}(h^{p+1})$ for all degrees of accuracy. On the other hand, we observe convergence rates 1 and 2, respectively, for $p = 1$, and convergence rates 2 and 3, respectively, for $p = 2, 3, 4, 5$. This is due to the fact that the expected convergence is of order $\mathcal{O}(h^{\min\{2-\epsilon,p\}})$ in the H^1 norm and $\mathcal{O}(h^{\min\{2-\epsilon,p\}+1})$ in the L^2 norm.

Numerical results: p-version

Here, we validate the exponential convergence of the p-version of the method for the model problem (3.1) with exact solution u_1 on $\Omega = (0, 1)^2$ on a Cartesian mesh and a Voronoi mesh made of four elements, respectively, as well as on the domain Ω given by the union of four Escher horses (see Figure 2.2, right). The obtained results are depicted in Figure 3.4, where the logarithm of the relative errors defined in (3.59) is plotted against the polynomial degree p.

One can clearly observe that the exponential convergence predicted in Theorem 3.2.6 is attained, even when employing a very coarse mesh with (non-convex) non-star-shaped elements, as the one

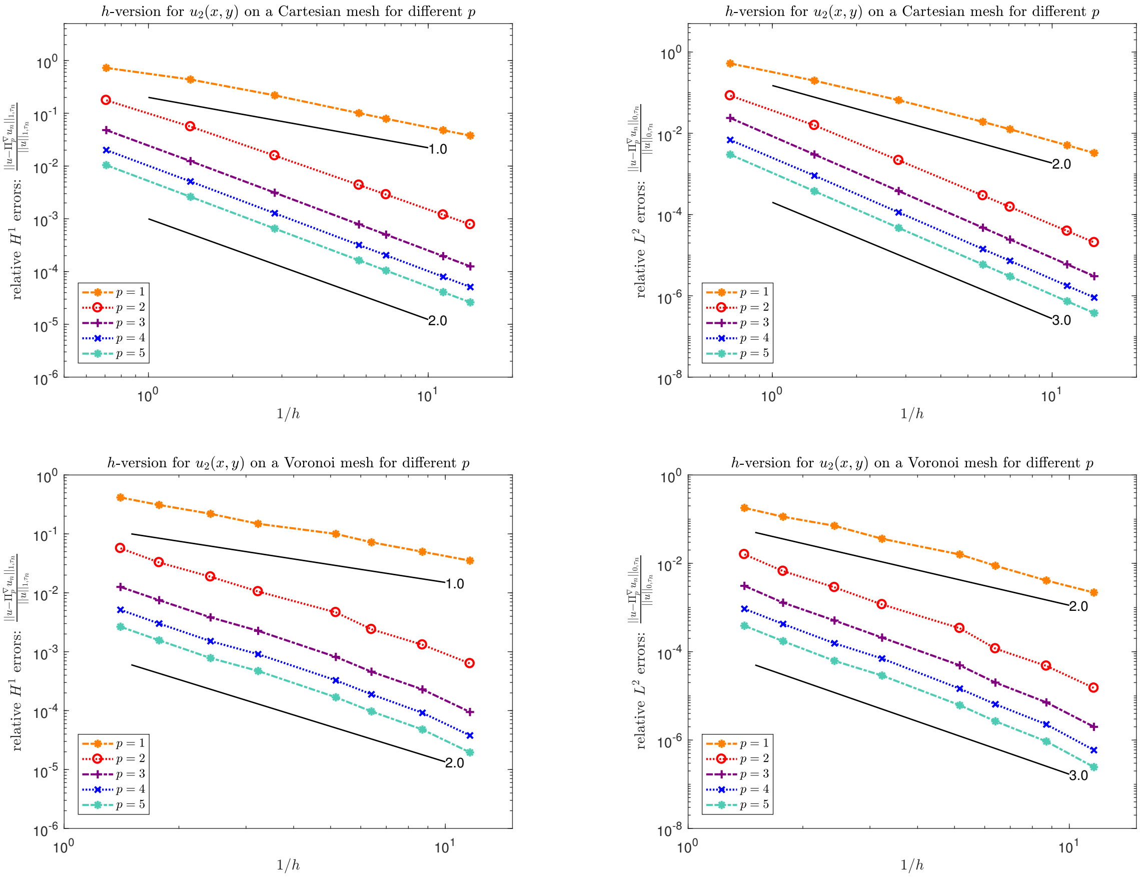


Figure 3.3: Convergence of the h-version of the method for the solution u_2 with finite Sobolev regularity on quasi-uniform Cartesian (*first row*) and Voronoi-Lloyd (*second row*) meshes; relative H^1 errors (*left*) and relative L^2 errors (*right*) defined in (3.59).

in Figure 2.2, right.

3.3.3 The hp-version and approximation of corner singularities

So far, both the theoretical analysis and the numerical tests were performed considering approximation spaces with uniform degree of accuracy p and with quasi-uniform meshes.

In general, however, the solutions to elliptic problems over polygonal domains have natural singularities arising in neighbourhoods of the corners of the domain. In particular, for problem (3.2) in a domain Ω with reentrant corners, the solution might have a regularity lower than H^2, even if the Dirichlet boundary datum g is smooth; for a precise functional setting regarding regularity of solutions to elliptic PDEs, we refer to [21,116,166] and the references therein. This implies that both the h- and the p-versions of standard Galerkin methods, in general, have limited approximation properties. In particular, employing quasi-uniform meshes and uniform degree of accuracy does not entail any sort of exponential convergence.

A possible way to recover exponential convergence, even in presence of corner singularities, is to use the so-called *hp-strategy* firstly designed by Babuška and Guo [19–21] in the FEM framework, and then generalized to the VEM in [40]. This strategy consists in combining mesh refinements towards the corners of the domain with p-refinements in the elements where the solution is sufficiently smooth. In this section, we discuss and numerically test an hp-version of the presented nonconforming Trefftz VEM.

To this purpose, we recall the concept of sequences of geometrically graded polygonal meshes $\{\mathcal{T}_n\}_{n\in\mathbb{N}}$. For a given $n \in \mathbb{N}$, $\mathcal{T}_n$ is a polygonal mesh consisting of $n+1$ layers, where we define a *layer* as follows. The 0-th layer is the set of all polygons in $\mathcal{T}_n$ abutting the vertices of Ω; the other

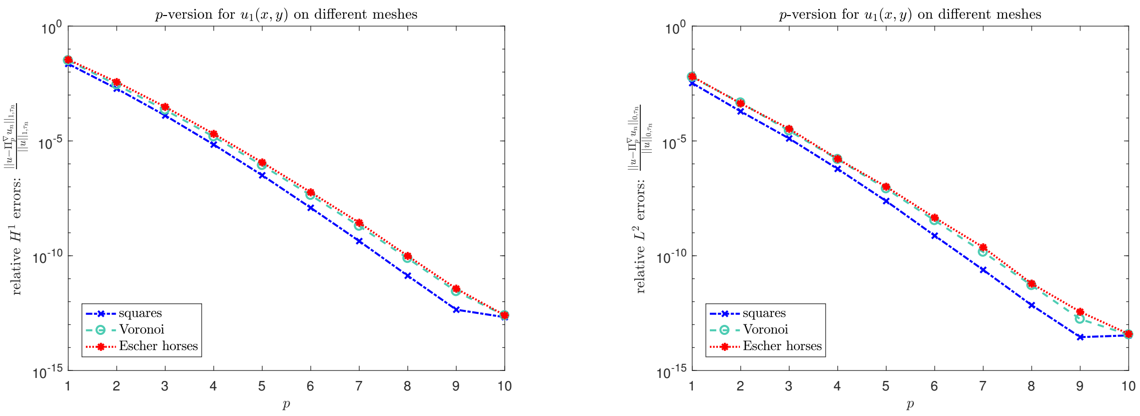


Figure 3.4: Convergence of the p-version of the method for the analytic solution u_1 on a quasi-uniform Cartesian mesh, a Voronoi-Lloyd mesh, and a Escher horses mesh; relative H^1 errors (*left*) and relative L^2 errors (*right*) defined in (3.59).

layers are defined inductively by requiring that the ℓ-th layer consists of those polygons, which abut the polygons in the $(\ell-1)$-th layer. More precisely, for all $\ell = 1, \dots, n$, we set

$$Ł_{n,\ell} := Ł_\ell := \left\{ K \in \mathcal{T}_n \mid \overline{K} \cap \overline{K_{\ell-1}} \neq \emptyset \text{ for some } K_{\ell-1} \in Ł_{\ell-1},\ K \not\subseteq \cup_{j=0}^{\ell-1} L_j \right\}.$$

The *hp*-gospel states that, in order to achieve exponential convergence of the error, one has to employ geometrically graded sequences of meshes. For this reason, we consider sequences $\{\mathcal{T}_n\}_{n\in\mathbb{N}}$ satisfying (**G1**)-(**G3**), but not (**G4**); we require instead

(**G5**) for all $n \in \mathbb{N}$, there exists $\sigma \in (0,1)$, called *grading parameter*, such that

$$h_K \approx \begin{cases} \sigma^n & \text{if } K \in Ł_0 \\ \frac{1-\sigma}{\sigma} \operatorname{dist}(K, \mathcal{V}^\Omega) & \text{if } K \in Ł_\ell,\ \ell = 1, \dots, n, \end{cases} \tag{3.60}$$

where $\mathcal{V}^\Omega$ denotes the set of vertices of the polygonal domain Ω.

Sequences $\{\mathcal{T}_n\}_{n\in\mathbb{N}}$ satisfying (**G5**) have the property that the layers "near" the corners of the domain consist of elements with measure converging to zero, whereas the other layers consist of polygons with fixed size. In Figure 3.5, we depict three meshes that represent the third elements $\mathcal{T}_3$ in certain sequences of meshes of the L-shaped domain

$$\Omega := (-1,1)^2 \setminus (-1,0)^2, \tag{3.61}$$

which are graded, for simplicity, only towards the vertex **0**.

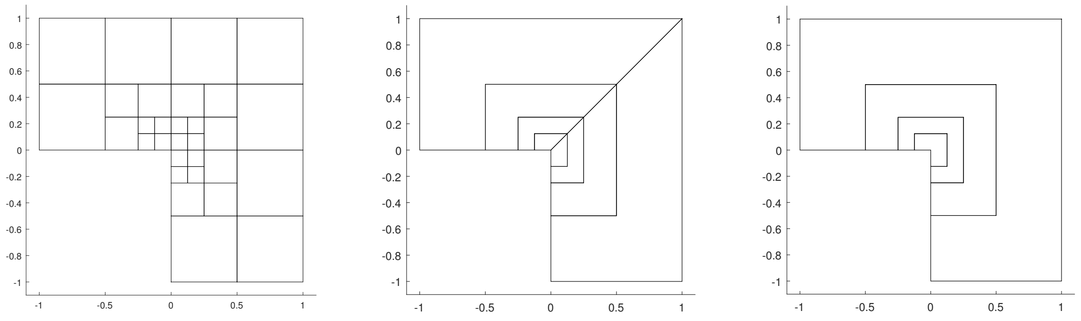


Figure 3.5: Third element $\mathcal{T}_3$ in three different sequences of geometrically graded meshes (type (a)-(c) from *left* to *right*) with $\sigma = 0.5$.

In order to completely describe the *hp*-strategy, we need to introduce Trefftz VE spaces with non-uniform degrees of accuracy. This can be done as follows. Firstly, for all $n \in \mathbb{N}$, we order the elements in $\mathcal{T}_n$ as K_1, K_2, $\dots$, $K_{\operatorname{card}(\mathcal{T}_n)}$. Then, we consider a vector $\mathbf{p}_n \in \mathbb{N}^{\operatorname{card}(\mathcal{T}_n)}$ whose entries are defined by

$$(\mathbf{p}_n)_j := \begin{cases} 1 & \text{if } K_j \in L_0 \\ \max(1, \lceil \mu(\ell+1) \rceil) & \text{if } K_j \in \mathrm{L}_\ell,\ \ell = 1, \dots, n, \end{cases} \tag{3.62}$$

where μ is a positive parameter to be assigned, and where $\lceil \cdot \rceil$ is the ceiling function. Having $\mathbf{p}_n$ for all $n \in \mathbb{N}$, we consider the elements e_1, e_2, $\dots$, $e_{\operatorname{card}(\mathcal{E}_n)}$ in $\mathcal{E}_n$ and define a vector $\mathbf{p}_n^{\mathcal{E}} \in \mathbb{N}^{\operatorname{card}(\mathcal{E}_n)}$, whose entries are built using the following rule (*maximum rule*):

$$(\mathbf{p}_n^{\mathcal{E}})_j := \begin{cases} (\mathbf{p}_n)_i & \text{if } e_j \in \mathcal{E}_n^B \text{ and } e_j \subset \partial K_i \\ \max((\mathbf{p}_n)_{i_1}, (\mathbf{p}_n)_{i_2}) & \text{if } e_j \in \mathcal{E}_n^I \text{ and } e_j \subset \partial K_{i_1} \cap \partial K_{i_2}. \end{cases}$$

Finally, for all $K \in \mathcal{T}_n$, we pinpoint the local Trefftz VE spaces with non-uniform degrees of accuracy

$$V^{\Delta}(K) := \left\{ v_n \in H^1(K) \mid \Delta v_n = 0 \text{ in } K,\ (\nabla v_n \cdot \mathbf{n}_K)_{|e_j} \in \mathbb{P}_{(\mathbf{p}_n^{\mathcal{E}})_j}(e_j)\, \forall e_j \text{ edge of } K \right\}.$$

The global nonconforming space and the set of global degrees of freedom are defined similarly to those for the case of uniform degree, see Section 3.1. The difference is that now the degrees of freedom and the corresponding “level of nonconformity” of the method vary from edge to edge. This approach is similar to that discussed in [40] for the *hp*-version of the conforming standard VEM.

Under this construction, one should be able to prove the following convergence result in terms of the number of degrees of freedom: there exists $\mu > 0$ such that the choice (3.62) guarantees

$$|u - u_n|_{1,\mathcal{T}_n} \leq c \exp\left(-b \sqrt[2]{\#\text{dofs}}\right), \tag{3.63}$$

for some positive constants b and c, depending on u, ρ_0, Λ, and σ, where #dofs denotes the number of degrees of freedom of the discretization space. This exponential convergence in terms of the dimension of the approximation space was proven for conforming Trefftz VEM in [79] and for Trefftz DG-FEM in [122]. In the present nonconforming Trefftz VEM, the setting of the proof of such exponential convergence would follow the same lines as that of the two methods mentioned above. Thus, we omit a detailed analysis and present here only some numerical results.

We underline that the exponential convergence in (3.63) is faster (in terms of the dimension of the space) than that of standard *hp*-FEM [166] and *hp*-VEM [40], whose decay rate is $\mathcal{O}(\exp\left(-b\sqrt[3]{\#\text{dofs}}\right))$, due to the use of harmonic subspaces instead of complete FE or VE spaces.

For our numerical tests, we consider the boundary value problems (3.2) on the L-shaped domain Ω defined in (3.61), with exact solution

$$u_3(x,y) = u_3(r,\theta) = r^{\frac{2}{3}} \sin\left(\frac{2}{3}\theta + \frac{\pi}{3}\right).$$

We note that $u_3 \in H^{\frac{5}{3}-\epsilon}(\Omega)$ for every $\epsilon > 0$ arbitrarily small, and also $u_3 \in H^{\frac{5}{3}-\epsilon}(\Omega_{\text{ext}})$, where Ω_{ext} is defined in (3.37); we stress that u_3 is the natural solution, singular at $\mathbf{0} = (0,0)$, which arises when solving a Poisson problem in the L-shaped domain Ω.

In Figure 3.6, we show the convergence of the *hp*-version of the method for different values of the grading parameter σ used in (3.60) and with degrees of accuracy graded according to (3.62), having set $\mu = 1$. We plot the logarithm of the relative H^1 error (3.59) against the square root of the number of degrees of freedom.

Note that, due to the different number of degrees of freedom for each type of mesh, the range of the coordinates varies from plot to plot. The straight lines for $\sigma = 0.5$ and $\sigma = \sqrt{2} - 1$ indicate agreement with (3.63) for meshes of type (a) and (b). However, when employing the mesh of type (c) with all grading parameters, and when employing grading parameter $\sigma = (\sqrt{2} - 1)^2$ for meshes of all types, we do not observe exponential convergence (3.63). In the former case, we deem that this is due to the shape of the elements, whereas, in the latter, this could be due to the fact

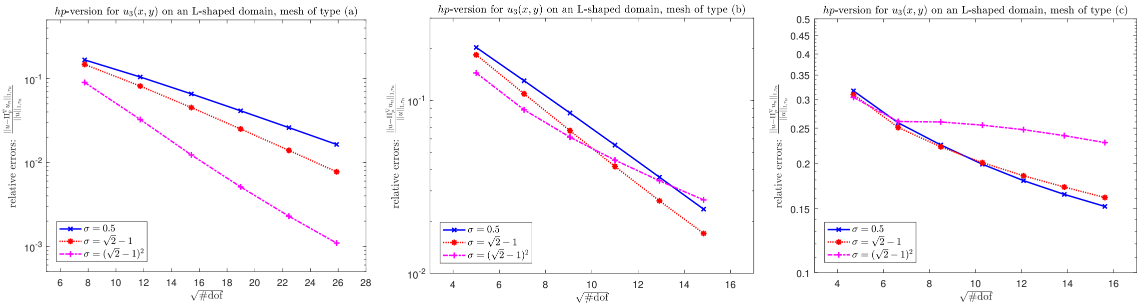


Figure 3.6: Convergence of the hp-version of the method for the solution u_3 on an L-shaped domain Ω, for the three sequences of graded meshes represented in Figure 3.5; relative H^1 errors defined in (3.59). The grading parameter σ is set to $1/2$, $\sqrt{2}-1$ and $(\sqrt{2}-1)^2$.

that the size of the elements in the outer layers is too large if picking the parameter μ in (3.62) equal to 1.

We point out that, in the framework of the conforming Trefftz VEM [79], a similar behaviour for the mesh of type (c) was observed. Instead, when employing the hp-version of the standard (non-Trefftz) VEM [40], the performance is more robust and the decay of the error is always straight exponential. This suboptimal behaviour might be intrinsic in the use of harmonic polynomials, or might be due to the choice of the harmonic polynomial basis employed in the construction of the method, see Section 3.3.1.

With this preliminary section in mind, we now turn to the Helmholtz problem.

Chapter 4

Trefftz virtual element method for the Helmholtz problem

In this chapter, we focus on the Helmholtz problem given in (2.4) with $\Gamma_D = \Gamma_N = \emptyset$ and $\theta = 1$, which, given a bounded convex polygon $\Omega \subset \mathbb{R}^2$ with boundary $\Gamma = \partial\Omega$ and $g \in H^{-\frac{1}{2}}(\Gamma)$, reads

$$\begin{cases} -\Delta u - k^2 u = 0 & \text{in } \Omega, \\ \nabla u \cdot \mathbf{n}_\Omega + \mathrm{i}ku = g & \text{on } \Gamma, \end{cases} \tag{4.1}$$

where $k > 0$ is the wave number, and where we recall that i is the imaginary unit and $\mathbf{n}_\Omega$ is the unit normal vector on Γ pointing outside Ω. The general case will be considered in Chapter 5. In the spirit of the previous chapter, we are interested in the construction and analysis of a nonconforming Trefftz VEM for (4.1).

The outline of this chapter is as follows. Section 4.1 deals with the design of a Trefftz VEM for (4.1) where the interelement continuity constraints are again imposed in a nonconforming fashion. In contrast to the previous chapter, the method will be based on plane waves instead of polynomials. In Section 4.2, an abstract error analysis is carried out and h-version error estimates are derived. Due to the fact that the presented methods suffers of strong ill-conditioning at the practical level, no numerical results are shown in this chapter, but are rather postponed to Chapter 5. There, a full discussion on this topic is carried out and a numerical recipe to mitigate the ill-conditioning and to render the method competitive is presented.

The material of this chapter has been published in [144].

4.1 Nonconforming Trefftz virtual element methods

In this section, we introduce a nonconforming Trefftz VE formulation for the problem (4.1).

To this purpose, we firstly consider the variational formulation corresponding to (4.1), which reads

$$\begin{cases} \text{find } u \in V \text{ such that} \\ b_k(u,v) = \int_\Gamma g\overline{v}\,\mathrm{d}s \quad \forall v \in V, \end{cases} \tag{4.2}$$

where $V := H^1(\Omega)$ and where

$$b_k(u,v) := a_k(u,v) + \mathrm{i}k \int_\Gamma u\overline{v}\,\mathrm{d}s \quad \forall u,\, v \in V, \tag{4.3}$$

with

$$a_k(u,v) := \int_\Omega \nabla u \cdot \overline{\nabla v}\,\mathrm{d}x - k^2 \int_\Omega u\overline{v}\,\mathrm{d}x \quad \forall u,\, v \in V.$$

Problem (4.2) is well-posed for all wave numbers k and, due to the convexity assumption on Ω, $u \in H^2(\Omega)$, if we assume in addition $g \in H^{\frac{1}{2}}(\Gamma)$, see e.g. [149, Proposition 8.1.4].

Given a mesh $\mathcal{T}_n$ as described in Section 2.3, our aim is to design a numerical method having the following structure:

$$\begin{cases} \text{find } u_h \in V_h^{\Delta+k^2} \text{ such that} \\ b_{k,h}(u_h, v_h) = F_h(v_h) \quad \forall v_h \in V_h^{\Delta+k^2}, \end{cases} \tag{4.4}$$

where, for all $n \in \mathbb{N}$, $V_h^{\Delta+k^2}$ is a finite dimensional space subordinated to $\mathcal{T}_n$, $b_{k,h}(\cdot,\cdot) \ : V_h^{\Delta+k^2} \times V_h^{\Delta+k^2} \to \mathbb{C}$ is a *computable* sesquilinear form mimicking its continuous counterpart $b(\cdot,\cdot)$ defined in (4.3), and the functional $F_h(\cdot) : V_h^{\Delta+k^2} \to \mathbb{C}$ is a *computable* counterpart of $\int_\Gamma g\overline{v}\,\mathrm{d}s$.

The reason why we do not employ the continuous sesquilinear forms and right-hand side is that the functions in the nonconforming Trefftz VE spaces are not known in closed form; therefore, the continuous sesquilinear forms and right-hand side are not computable.

The outline of this section is the following. After fixing the basic notation on plane waves in Section 4.1.1, local Trefftz VE spaces, as well as the global space $V_h^{\Delta+k^2}$ in (4.4), are introduced in Section 4.1.2. Then, in Section 4.1.3, local projectors mapping from the local Trefftz VE spaces into spaces of plane waves are defined; such projectors will allow to define suitable $b_{k,h}(\cdot,\cdot)$ and $F_h(\cdot)$ in (4.4), see Section 4.1.4.

4.1.1 Plane wave spaces

Here, we introduce the plane wave spaces. To this purpose, given $p = 2q+1$ for some $q \in \mathbb{N}$, we introduce the set of indices $\mathcal{J} := \{1,\dots,p\}$ and the set of pairwise different normalized directions $\{\mathbf{d}_\ell\}_{\ell\in\mathcal{J}}$. As q plays the same role as the polynomial degree in the approximation properties of plane wave spaces, we refer to q as *effective plane wave degree.* For every $\ell \in \mathcal{J}$, we define the plane wave traveling along the direction $\mathbf{d}_\ell$ as

$$w_\ell(\mathbf{x}) := e^{\mathrm{i}k\mathbf{d}_\ell\cdot\mathbf{x}}. \tag{4.5}$$

Moreover, for every $K \in \mathcal{T}_n$, recalling that $\mathbf{x}_K$ is the centroid of K, we pinpoint the bulk plane wave related to K by

$$w_\ell^K(\mathbf{x}) := e^{\mathrm{i}k\mathbf{d}_\ell\cdot(\mathbf{x}-\mathbf{x}_K)}{}_{|K}. \tag{4.6}$$

Examples of plane waves with $k = 20$ traveling along different directions are given in Figure 4.1.

Then, we introduce the local plane wave space on the element $K \in \mathcal{T}_n$ by

$$\mathbb{PW}_p(K) := \operatorname{span}\left\{w_\ell^K,\ \ell \in \mathcal{J}\right\}. \tag{4.7}$$

We make the following assumption on the plane wave directions:

(D1) (*minimum angle*) there exists a constant $0 < \delta \leq 1$ with the property that the directions $\{\mathbf{d}_\ell\}_{\ell\in\mathcal{J}}$ are such that the minimum angle between two directions is larger than or equal to $\frac{2\pi}{p}\delta$, and the angle between two neighbouring directions is strictly smaller than π.

The global discontinuous plane wave space with uniform p is given by

$$\mathbb{PW}_p(\mathcal{T}_n) := \prod_{K\in\mathcal{T}_n} \mathbb{PW}_p(K) = \{v \in L^2(\Omega) : v_{|K} \in \mathbb{PW}_p(K) \quad \forall K \in \mathcal{T}_n\}.$$

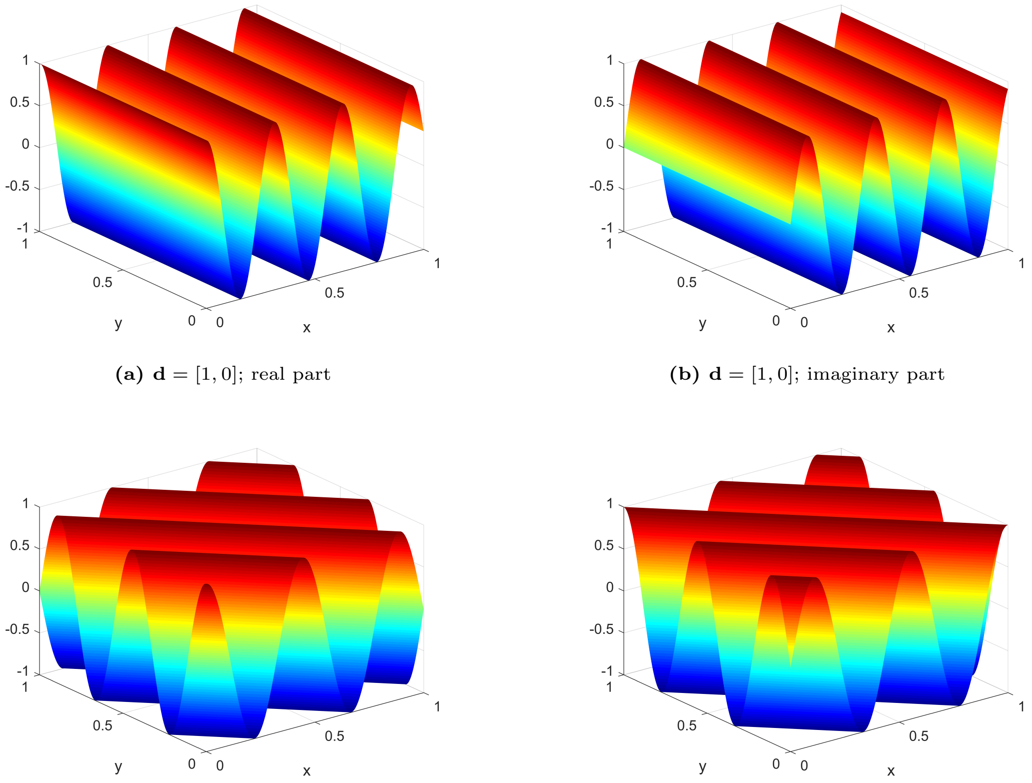
For the same p, we also introduce the spaces of traces of plane waves on the mesh edges. Given $e \in \mathcal{E}_n$ and $\mathbf{x}_e$ its midpoint, we define, for any $\ell \in \mathcal{J}$,

$$w_\ell^e(\mathbf{x}) := e^{\mathrm{i}k\mathbf{d}_\ell\cdot(\mathbf{x}-\mathbf{x}_e)}{}_{|e}. \tag{4.8}$$

We denote by $\mathbb{PW}_p(e)$ the space spanned by w_ℓ^e, $\ell = 1,\dots,p$.

We observe that, while the dimension of $\mathbb{PW}_p(K)$ is equal to p for all $K \in \mathcal{T}_n$, the dimension of $\mathbb{PW}_p(e)$ could in principle be smaller. In fact, if

$$\mathbf{d}_j \cdot (\mathbf{x}-\mathbf{x}_e) = \mathbf{d}_\ell \cdot (\mathbf{x}-\mathbf{x}_e) \quad \forall \mathbf{x} \in e \tag{4.9}$$



(a) $\mathbf{d} = [1, 0]$; real part

(b) $\mathbf{d} = [1, 0]$; imaginary part

(c) $\mathbf{d} = [\frac{\pi}{4}, \frac{\pi}{4}]$; real part

(d) $\mathbf{d} = [\frac{\pi}{4}, \frac{\pi}{4}]$; imaginary part

Figure 4.1: Plane waves with $k = 20$ traveling along the direction $\mathbf{d}$.

for some $j, \ell \in \{1, \ldots, p\}$, $j > \ell$, then $w_j^e(\mathbf{x}) = w_\ell^e(\mathbf{x})$. Thus, in order to have linearly independent functions, we have to check for all the indices in $\mathcal{J}$ whether (4.9) is satisfied. Whenever this is the case, we remove, without loss of generality, the index j from $\mathcal{J}$. The resulting set of indices is denoted by $\mathcal{J}_e'$. Clearly, it holds $\mathrm{card}(\mathcal{J}_e') \leq p$.

In the forthcoming analysis (see Section 4.2) we also need to employ constant functions on the edges. To this purpose, if there exists a direction $\mathbf{d}_* \in \{\mathbf{d}_\ell\}_{\ell \in \mathcal{J}}$ such that

$$\mathbf{d}_* \cdot (\mathbf{x} - \mathbf{x}_e) = 0 \quad \forall \mathbf{x} \in e, \tag{4.10}$$

that is, $\mathbf{d}_*$ is orthogonal to the edge e, then $\mathrm{span}\{w_\ell^e,\ \ell \in \mathcal{J}_e'\}$ already contains the constant functions. In this case, we set $\mathcal{J}_e := \mathcal{J}_e'$; otherwise, we define $\mathcal{J}_e := \mathcal{J}_e' \cup \{p+1\}$ and set $w_{p+1}^e(\mathbf{x}) := 1$. Finally, we introduce $p_e := \mathrm{card}(\mathcal{J}_e)$.

This whole procedure goes under the name of *filtering process*. It is summarized again in the form of a pseudocode in Algorithm 1.

In Figure 4.2, we depict the *filtering process* applied to all the possible configurations along a given edge $e \in \mathcal{E}_n$.

Remark 8. We note that, in the definitions (4.6) and (4.8), we also consider a shift by the barycenters of the elements and the midpoints of the edges, respectively. This actually does not change the nature of the basis since it simply results in a multiplication between a nonshifted plane wave with a constant. However, this additional notation may be useful when implementing the method, as it helps to remember when dealing with bulk and/or edge plane waves, see Section 5.2.

Remark 9. Here, we highlight that the filtering process guarantees that the plane wave traces are linearly independent on each edge. However, at the practical level, one would expect that, if two different directions $\mathbf{d}_j$ and $\mathbf{d}_\ell$ lead to almost the same values when computing $\mathbf{d}_j \cdot (\mathbf{x} - \mathbf{x}_e)$ and $\mathbf{d}_\ell \cdot (\mathbf{x} - \mathbf{x}_e)$, the method becomes instable due to ill-conditioning. This is in fact the case. To this purpose, in order to deal with such situations, at the practical level, Algorithm 1 will be

Algorithm 1 *Filtering process*

```
For all edges $e \in \mathcal{E}_n$:
    1. Remove redundant plane waves
        • Initialize $\mathcal{J}_e' := \mathcal{J} := \{1, \dots, p\}$;
        • For all indices in $\mathcal{J}_e'$, check whether (4.9) is satisfied;
        • Whenever this is the case for some pair $j, \ell \in \mathcal{J}_e'$ with $j > \ell$, remove index $j$ from $\mathcal{J}_e'$;
    2. Add the constants
        • Check whether there exists a direction $\mathbf{d}_* \in \{\mathbf{d}_\ell\}_{\ell \in \mathcal{J}}$ such that (4.10) is fulfilled.
        • If this is the case, set $\mathcal{J}_e := \mathcal{J}_e'$; otherwise, set $\mathcal{J}_e := \mathcal{J}_e' \cup \{p+1\}$ and $w_{p+1}^e(\mathbf{x}) := 1$.
    3. Define $p_e := \mathrm{card}(\mathcal{J}_e)$.
```

replaced by a modified filtering process based on an eigendecomposition of the plane wave edge mass matrices in Chapter 5.

With these definitions of the set of indices $\mathcal{J}_e$ and of the corresponding functions w_ℓ^e on each edge $e \in \mathcal{E}_n$, we define the plane wave trace space of dimension p_e as

$$\mathbb{PW}_p^c(e) := \mathrm{span}\left\{w_\ell^e\,,\, \ell \in \mathcal{J}_e\right\}, \tag{4.11}$$

where the superscript c indicates that the space includes the constants.

We denote the space of piecewise discontinuous traces over ∂K as

$$\mathbb{PW}_p^c(\partial K) = \{w^{\partial K} \in L^2(\partial K) : {w^{\partial K}}_{|_e} \in \mathbb{PW}_p^c(e) \quad \forall e \in \mathcal{E}^K\}. \tag{4.12}$$

4.1.2 Nonconforming Trefftz virtual element spaces

In this section, we specify the nonconforming Trefftz VE space $V_h^{\Delta+k^2}$ in (4.4). To this purpose, we firstly introduce local Trefftz VE spaces, and then define corresponding ones at the global level.

Local Trefftz VE spaces. Given $K \in \mathcal{T}_n$, we denote the impedance trace of a function $v \in H^1(K)$ on ∂K by

$$\gamma_I^K(v) := \nabla v \cdot \mathbf{n}_K + \mathrm{i}kv, \tag{4.13}$$

where we recall that $\mathbf{n}_K$ is the unit normal vector on ∂K pointing outside K.

Then, given $p \in \mathbb{N}$, for every $n \in \mathbb{N}$ and $K \in \mathcal{T}_n$, we introduce the *local* Trefftz VE space

$$\begin{aligned} V^{\Delta+k^2}(K) := \big\{v_h \in H^1(K) \mid \Delta v_h + k^2 v_h = 0 \text{ in } K, \\ \gamma_I^K(v_h)_{|_e} \in \mathbb{PW}_p^c(e)\ \forall e \in \mathcal{E}^K\big\}, \end{aligned} \tag{4.14}$$

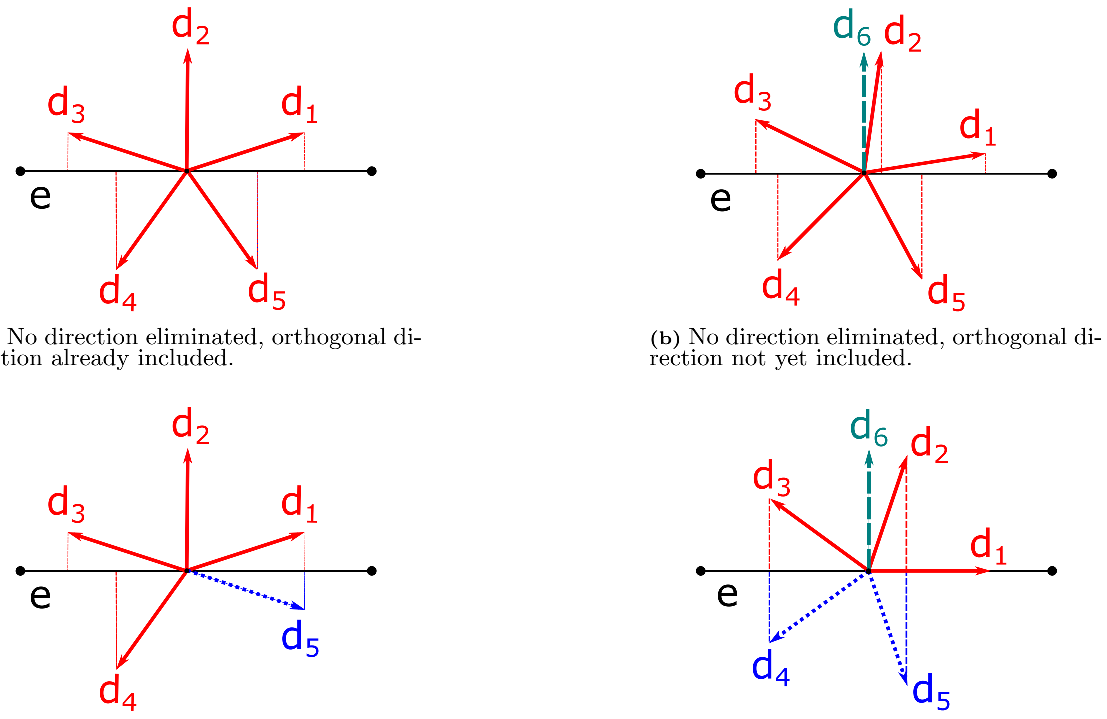
where we recall that $\mathbb{PW}_p^c(e)$ is given in (4.11). In words, this space consists of all functions in $H^1(K)$ which lie in the kernel of the Helmholtz operator and whose impedance traces are edgewise equal to traces of plane waves including constants.

It can be easily seen that $\mathbb{PW}_p(K) \subset V^{\Delta+k^2}(K)$, which will be essential for deriving *a priori* error estimates for the discretization error. However, the space $V^{\Delta+k^2}(K)$ also contains other functions that are not available in closed form, whence the term *virtual* in the name of the method.

For each $K \in \mathcal{T}_n$, the dimension p_K of the discrete space $V^{\Delta+k^2}(K)$ in (4.14) coincides with the sum over all $e \in \mathcal{E}^K$ of the dimension p_e of the edge plane wave spaces $\mathbb{PW}_p^c(e)$ in (4.11):

$$p_K := \dim V^{\Delta+k^2}(K) = \sum_{e \in \mathcal{E}^K} p_e,$$

where we recall that $p_e \leq p + 1$.



(a) No direction eliminated, orthogonal direction already included.

(b) No direction eliminated, orthogonal direction not yet included.

(c) One direction eliminated, orthogonal direction already included.

(d) Two directions eliminated, orthogonal direction not yet included.

Figure 4.2: *Filtering process.* We depict all the possible configurations. In solid lines, the directions that are kept; in dotted lines, the directions that are eliminated accordingly with (4.9); in dashed lines, the orthogonal direction that has to be possibly added in order to include constants.

Having this, we define the following set of functionals. Given $K \in \mathcal{T}_n$, we consider the moments on each edge $e \in \mathcal{E}^K$ with respect to functions in the space $\mathbb{PW}_p^c(e)$ defined in (4.11):

$$\mathrm{dof}_{e,\ell}(v_h) := \frac{1}{h_e}\int_e v_h \overline{w_\ell^e}\,\mathrm{d}s \quad \forall e \in \mathcal{E}^K, \forall \ell \in \mathcal{J}_e. \tag{4.15}$$

We prove that this set provides a set of unisolvent degrees of freedom for all $K \in \mathcal{T}_n$, provided that the following assumption on the wave number k is satisfied:

(**A1**) the wave number k is such that k^2 is not a Dirichlet-Laplace eigenvalue on K for all $K \in \mathcal{T}_n$.

For a given $K \in \mathcal{T}_n$, the assumption (**A1**) results in a condition on the product $h_K k$. More precisely, for any simply connected element K, the smallest Dirichlet-Laplace eigenvalue on K satisfies

$$\lambda_1 \geq \frac{a}{\rho_K^2},$$

where ρ_K denotes the radius of the largest ball contained in K and where $a \geq 0.6197$, see e.g. [27]. As a consequence, assuming that

$$h_K k \leq \sqrt{c_0 a} \tag{4.16}$$

for some $c_0 \in (0,1]$, we deduce

$$k^2 = \frac{h_K^2 k^2}{h_K^2} \leq \frac{c_0 a}{h_K^2} \leq \frac{c_0 a}{\rho_K^2} \leq \lambda_1,$$

which means that (4.16) guarantees that k^2 is not a Dirichlet-Laplace eigenvalue on K.

Lemma 4.1.1. *Suppose that the assumption (**A1**) holds true. Then, for every $K \in \mathcal{T}_n$, the set of functionals* (4.15) *is a unisolvent set of degrees of freedom for $V^{\Delta+k^2}(K)$.*

Proof. Given $K \in \mathcal{T}_n$, we firstly observe that the dimension of the local space $V^{\Delta+k^2}(K)$ is equal to the number of functionals in (4.15). Thus, we only need to prove that, given any $v_h \in V^{\Delta+k^2}(K)$ such that all degrees of freedom (4.15) are zero, then $v_h = 0$.

To this end, we observe that an integration by parts, together with the fact that v_h belongs to the kernel of the Helmholtz operator, yields

$$\begin{aligned}
&|v_h|^2_{1,K} - k^2\|v_h\|^2_{0,K} - \mathrm{i}k\|v_h\|^2_{0,\partial K}\\
&= \int_K \nabla v_h \cdot \overline{\nabla v_h}\,\mathrm{d}x - k^2 \int_K v_h\overline{v_h}\,\mathrm{d}x - \mathrm{i}k\int_{\partial K} v_h\overline{v_h}\,\mathrm{d}s\\
&= \int_K v_h \underbrace{\overline{(-\Delta v_h - k^2 v_h)}}_{=0}\,\mathrm{d}x + \int_{\partial K} v_h\overline{(\nabla v_h \cdot \mathbf{n}_K + \mathrm{i}kv_h)}\,\mathrm{d}s \qquad (4.17)\\
&= \sum_{e\in\mathcal{E}^K} \int_e v_h\overline{\gamma_I^K(v_h)_{|e}}\,\mathrm{d}s = 0,
\end{aligned}$$

where in the last identity we also used the facts that, owing to the definition of the space $V^{\Delta+k^2}(K)$ in (4.14), the impedance trace of v_h is an element of the space (4.12), and that the degrees of freedom (4.15) of v_h are zero. Thus, the imaginary part on the left-hand side of (4.17) is zero and one deduces that $v_h = 0$ on ∂K. Since v_h is also solution to a homogeneous Helmholtz equation, the assertion follows thanks to the assumption (**A1**). □

Having this, we denote by $\{\varphi_{e,\ell}\}_{e\in\mathcal{E}^K,\,\ell\in\mathcal{J}_e}$ the local canonical basis, where

$$\mathrm{dof}_{\widetilde{e},\widetilde{\ell}}(\varphi_{e,\ell}) = \delta_{(e,\ell),(\widetilde{e},\widetilde{\ell})} = \begin{cases} 1 & \text{if } (e,\ell) = (\widetilde{e},\widetilde{\ell})\\ 0 & \text{otherwise.}\end{cases} \qquad (4.18)$$

Global Trefftz VE spaces. We now focus on the global level. The global nonconforming Sobolev space with respect to $\mathcal{T}_n$ and the underlying plane wave spaces with $p \in \mathbb{N}$ reads

$$H^{1,nc}(\mathcal{T}_n) := \left\{v \in H^1(\mathcal{T}_n) : \int_e [\![v]\!] \cdot \mathbf{n}^e\,\overline{w^e}\,\mathrm{d}s = 0 \quad \forall w^e \in \mathbb{PW}^c_p(e),\, \forall e \in \mathcal{E}^I_n\right\}, \qquad (4.19)$$

where $\mathbf{n}^e$ is either $\mathbf{n}^e_{K^+}$ or $\mathbf{n}^e_{K^-}$, but fixed.

Then, the *global* Trefftz VE space is given by

$$V_h^{\Delta+k^2} := \{v_h \in H^{1,nc}(\mathcal{T}_n) : v_{h|K} \in V^{\Delta+k^2}(K) \quad \forall K \in \mathcal{T}_n\}. \qquad (4.20)$$

As above, the set of global degrees of freedom is obtained by coupling the local degrees of freedom on the interfaces between elements. Clearly, $V_h^{\Delta+k^2} \not\subseteq H^1(\Omega)$.

We underline that the definition of the degrees of freedom (4.15) is actually tailored for building discrete trial and test spaces that are nonconforming in the sense of (4.19). Besides, they will be used in the construction of projectors mapping onto spaces of plane waves. This is the topic of the next Section 4.1.3.

Remark 10. Under the choice of the degrees of freedom in (4.15), the dimension of the global space is larger than that of plane wave discontinuous Galerkin methods [150]. However, due to a modified filtering process, see Remark 9, at the practical level, the dimension of the nonconforming Trefftz VE space can be reduced without losing in terms of accuracy. A numerical comparison carried out in Chapter 5 shows that the nonconforming Trefftz VEM and the plane wave discontinuous Galerkin method have a comparable behavior in terms of accuracy versus number of degrees of freedom and, in some occasions, the former performs even better.

4.1.3 Local projectors

In this section, we introduce local projectors mapping functions in local Trefftz VE spaces (4.14) onto plane waves. Such projectors will play a central role in the construction of the computable sesquilinear form $b_{k,h}(\cdot,\cdot)$ and functional $F_h(\cdot)$ for the method (4.4).

To start with, given $K \in \mathcal{T}_n$, we define the local sesquilinear form

$$a_k^K(u,v) := \int_K \nabla u \cdot \overline{\nabla v} \,\mathrm{d}x - k^2 \int_K u\overline{v} \,\mathrm{d}x \quad \forall u, v \in H^1(K). \tag{4.21}$$

Note that

$$a_k(u,v) = \sum_{K\in\mathcal{T}_n} a_k^K(u,v) \quad \forall u,\, v \in V.$$

Then, we introduce the local projector

$$\begin{aligned} \Pi_p^K &: V^{\Delta+k^2}(K) \to \mathbb{PW}_p(K) \\ a_k^K(\Pi_p^K u_h, w^K) &= a_k^K(u_h, w^K) \quad \forall u_h \in V^{\Delta+k^2}(K), \forall w^K \in \mathbb{PW}_p(K). \end{aligned} \tag{4.22}$$

Note that this projector is *computable* by means of the degrees of freedom (4.15) without the need of explicit knowledge of the functions of $V^{\Delta+k^2}(K)$. Indeed, an integration by parts and the fact that any plane wave $w^K \in \mathbb{PW}_p(K)$ belongs to the kernel of the Helmholtz operator lead to

$$\begin{aligned} a_k^K(u_h, w^K) &= \int_K \nabla u_h \cdot \overline{\nabla w^K} \,\mathrm{d}x - k^2 \int_K u_h \overline{w^K} \,\mathrm{d}x \\ &= \sum_{e\in\mathcal{E}^K} \int_e u_h \overline{(\nabla w^K \cdot \mathbf{n}_K)} \,\mathrm{d}s \quad \forall u_h \in V^{\Delta+k^2}(K), \forall w^K \in \mathbb{PW}_p(K). \end{aligned}$$

Since $(\nabla w^K \cdot \mathbf{n})_{K|_e} \in \mathbb{PW}_p^c(e)$ for all $e \in \mathcal{E}^K$, computability is guaranteed by the choice of the degrees of freedom in (4.15).

In the following proposition we prove that Π_p^K is well-defined.

Proposition 4.1.2. *Assume that K is an element of a mesh that satisfies the mesh assumption **(G1)** introduced in Section 2.3. Then, the following two statements hold true:*

1. *Denoting by μ_2 the smallest positive Neumann-Laplace eigenvalue in K, it holds*

$$\mu_2 \geqslant \frac{C_\Delta \pi^2}{h_K^2},$$

*where $C_\Delta \in (0,1]$ only depends on the shape of K, i.e. on ρ_0 and ρ in the assumption **(G1)**.*

2. *Assume that the assumption **(D1)** on the plane wave directions holds true. If $h_K k$ is such that there exists a constant $C_1 > 0$ with*

$$0 < h_K k < C_1 \leq \min\left\{\frac{\sqrt{C_\Delta}\pi}{\sqrt{2}}, 0.5538\right\},$$

then $k^2 < \mu_2$, and in particular it follows that Π_p^K is well-defined and continuous. More precisely, there exists a constant $\beta(h_K k) > 0$, uniformly bounded away from zero as $h_K k \to 0$, such that

$$\|\Pi_p^K u_h\|_{1,k,K} \leq \frac{1}{\beta(h_K k)} \|u_h\|_{1,k,K} \quad \forall u_h \in V^{\Delta+k^2}(K).$$

Note that, whenever K is convex, $C_\Delta = 1$, see e.g. [154], and hence $\min\left\{\frac{\sqrt{C_\Delta}\pi}{\sqrt{2}}, 0.5538\right\} = 0.5538$.

Proof. For the proof of the first part, we refer to [55, 137], and for the second part, to [159, Propositions 2.1 and 2.3]. □

Remark 11. In order to numerically investigate the condition for well-posedness of Π_p^K, we plot the minimal (absolute) eigenvalues of the matrix $\boldsymbol{A}^{\widehat{K}} := \{a^{\widehat{K}}(w_\ell^{\widehat{K}}, w_j^{\widehat{K}})\}_{\ell,j=1,\dots,p}$ in terms of the wave number k on the reference element $\widehat{K} = (0,1)^2$, see Figure 4.3. On this domain, the Neumann-Laplace eigenvalues $\nu_{m,n}$ are known explicitly:

$$\nu_{m,n} = \pi^2(m^2 + n^2), \quad m, n \in \mathbb{N}_0.$$

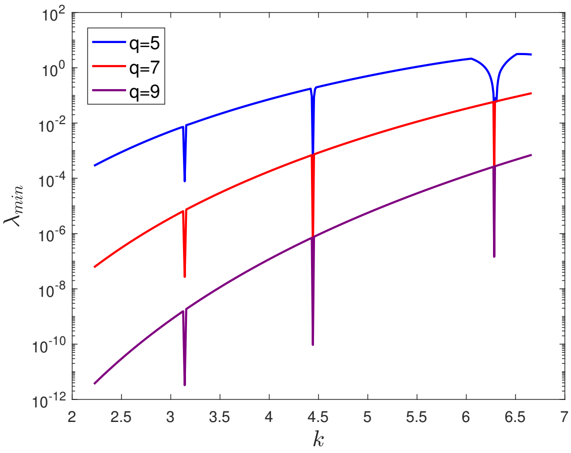


Figure 4.3: Minimal (absolute) eigenvalues of the matrix $\boldsymbol{A}^{\widehat{K}}$, see Remark 11.

We observe that, for wave numbers k close to the square roots of the eigenvalues $\nu_{m,n}$, the minimal (absolute) eigenvalue of $\boldsymbol{A}^{\widehat{K}}$ is actually some orders of magnitude lower than outside the neighborhoods of $\sqrt{\nu_{m,n}}$. Therefore, when k^2 is close to a Neumann-Laplace eigenvalue, the continuity constant of Π_p^K may deteriorate.

In addition, given a function $v_h \in V^{\Delta+k^2}(K)$, we define, on every edge $e \in \mathcal{E}^K$, the projector

$$\begin{aligned} &\Pi_p^{0,e} : V^{\Delta+k^2}(K)_{|_e} \to \mathbb{PW}_p^c(e) \\ &\int_e \Pi_p^{0,e}(v_{h|e})\overline{w^e}\,\mathrm{d}s = \int_e v_{h|_e}\overline{w^e}\,\mathrm{d}s \quad \forall v_h \in V^{\Delta+k^2}(K),\ \forall w^e \in \mathbb{PW}_p^c(e). \end{aligned} \tag{4.23}$$

The computability of this projector for functions in $V^{\Delta+k^2}(K)$ is again provided by the choice of the degrees of freedom in (4.15). Clearly, $\Pi_p^{0,e}(u_{h|e})$ coincides with the $L^2(e)$ projection of $u_{h|_e}$ onto $\mathbb{PW}_p^c(e)$.

Remark 12. The projector $\Pi_p^{0,e}$ is not defined for functions in the nonconforming space $V_h^{\Delta+k^2}$ in (4.20), but rather for the restrictions of such functions to the elements of the mesh. However, in order to avoid a cumbersome notation in the following, we will not highlight such restrictions whenever it is clear from the context.

The following approximation result holds true.

Proposition 4.1.3. *Let $K \in \mathcal{T}_n$ and $e \in \mathcal{E}^K$. For all $u \in H^1(K)$, it holds*

$$\|u - \Pi_p^{0,e}u\|_{0,e} \le C_0 h_K^{\frac{1}{2}} |u - w^K|_{1,K} \quad \forall w^K \in \mathbb{PW}_p(K), \tag{4.24}$$

where the constant $C_0 > 0$ only depends on the shape of K.

Proof. We firstly note that, for each $K \in \mathcal{T}_n$ and $e \in \mathcal{E}^K$, the definition of $\Pi_p^{0,e}$ in (4.23) yields

$$\|u - \Pi_p^{0,e}u\|_{0,e} \le \|u - c - w^K\|_{0,e} \le \|u - c - w^K\|_{0,\partial K} \quad \forall w^K \in \mathbb{PW}_p(K),\ \forall c \in \mathbb{C}.$$

By selecting

$$c = \frac{1}{|K|}\int_K (u - w^K)\,\mathrm{d}x,$$

and using the trace inequality (2.27) together with the Poincaré-Friedrichs inequality (2.29), we get

$$\begin{aligned} \|u - c - w^K\|_{0,\partial K}^2 &\le C_T\left(h_K^{-1}\|u - c - w^K\|_{0,K}^2 + h_K|u - w^K|_{1,K}^2\right) \\ &\le C_T(C_P^2 + 1)h_K|u - w^K|_{1,K}^2, \end{aligned}$$

from which we have (4.24) with $C_0^2 := C_T(C_P^2 + 1)$. □

For future use, we denote by $\Pi_p^{0,\Gamma}$ the L^2 projector

$$\Pi_p^{0,\Gamma} : L^2(\Gamma) \to \prod_{e\in\mathcal{E}_n^B} \mathbb{PW}_p^c(e). \tag{4.25}$$

We highlight that, for any $v_h \in V_h$, the identity $\Pi_p^{0,\Gamma}(v_{h|\Gamma})_{|e} = \Pi_p^{0,e}(v_{h|e})$ holds for all boundary edges $e \in \mathcal{E}_n^B$.

4.1.4 Discrete sesquilinear forms and right-hand side

We specify the sesquilinear form $b_{k,h}(\cdot,\cdot)$ and the functional $F_h(\cdot)$ characterizing the method (4.4).

Construction of $b_{k,h}(\cdot,\cdot)$.

Following [31], and analogously as in Section 3.1.3, the definition of Π_p^K in (4.22) yields

$$a_k^K(u_h,v_h) = a_k^K(\Pi_p^K u_h, \Pi_p^K v_h) + a_k^K((I-\Pi_p^K)u_h,(I-\Pi_p^K)v_h) \quad \forall u_h,\, v_h \in V^{\Delta+k^2}(K). \tag{4.26}$$

As above, the first term on the right-hand side of (4.26) is computable, whereas the second one is not, and thus has to be replaced by a proper *computable* sesquilinear form $S_k^K(\cdot,\cdot)$, the *stabilization*; see Section 5.3 for an explicit choice. Having this, we set, for all u_h, $v_h \in V^{\Delta+k^2}(K)$,

$$a_{k,h}^K(u_h,v_h) := a_k^K(\Pi_p^K u_h, \Pi_p^K v_h) + S_k^K\left((I-\Pi_p^K)u_h,(I-\Pi_p^K)v_h\right). \tag{4.27}$$

We point out that $a_{k,h}^K(\cdot,\cdot)$ satisfies the following *plane wave consistency property*:

$$\begin{aligned} a_{k,h}^K(w^K,v_h) = a_k^K(w^K,v_h), \quad & a_{k,h}^K(v_h,w^K) = a_k^K(v_h,w^K) \\ & \forall w^K \in \mathbb{PW}_p(K),\, \forall v_h \in V^{\Delta+k^2}(K). \end{aligned} \tag{4.28}$$

Moreover, we replace the boundary integral term in $b_k(\cdot,\cdot)$ in (4.3) with

$$\mathrm{i}k\int_\Gamma u_h\overline{v_h}\,\mathrm{d}s \quad \rightsquigarrow \quad \mathrm{i}k\int_\Gamma (\Pi_p^{0,\Gamma}u_h)\overline{(\Pi_p^{0,\Gamma}v_h)}\,\mathrm{d}s \quad \forall u_h, v_h \in V_h^{\Delta+k^2},$$

where $\Pi_p^{0,\Gamma}$ is defined in (4.25). Hence, the global sesquilinear form $b_{k,h}(\cdot,\cdot)$ in (4.4) is given by

$$b_{k,h}(u_h,v_h) := a_{k,h}(u_h,v_h) + \mathrm{i}k\int_\Gamma (\Pi_p^{0,\Gamma}u_h)\overline{(\Pi_p^{0,\Gamma}v_h)}\,\mathrm{d}s \quad \forall u_h, v_h \in V_h^{\Delta+k^2}, \tag{4.29}$$

where

$$a_{k,h}(u_h,v_h) := \sum_{K\in\mathcal{T}_n} a_{k,h}^K(u_h,v_h). \tag{4.30}$$

In the subsequent error analysis of the method (4.4), see Theorem 4.2.4 below, we will require continuity of the local sesquilinear forms $a_{k,h}^K(\cdot,\cdot)$ given in (4.27), as well as a discrete Gårding inequality for $b_{k,h}(\cdot,\cdot)$ defined in (4.29). If the stabilization forms $S_k^K(\cdot,\cdot)$ satisfy

$$\alpha_h\|v_h\|_{1,k,K}^2 - 2k^2\|v_h\|_{0,K}^2 \le S_k^K(v_h,v_h) \le \gamma_h\|v_h\|_{1,k,K}^2 \quad \forall v_h \in \ker(\Pi_p^K),\, \forall K\in\mathcal{T}_n, \tag{4.31}$$

for some positive constants α_h and γ_h, then, by proceeding as in Theorem [159, Proposition 4.1], one can prove that the local continuity assumptions and the local Gårding inequalities of Theorem 4.2.4 are satisfied.

Construction of $F_h(\cdot)$.

We set $g_h := \Pi_p^{0,\Gamma}g$, where $\Pi_p^{0,\Gamma}$ is defined in (4.25). Using the approximation g_h instead of g allows us to define the computable functional

$$F_h(v_h) := \int_\Gamma g_h\overline{v_h}\,\mathrm{d}s = \int_\Gamma g\overline{(\Pi_p^{0,\Gamma}v_h)}\,\mathrm{d}s. \tag{4.32}$$

In order to avoid additional complications in the forthcoming analysis, we will assume that the integral in (4.32) can be computed exactly. In practice, such integrals are approximated with high-order quadrature formulas.

4.2 *A priori* error analysis

In this section, we firstly prove approximation properties of functions in Trefftz VE spaces in Section 4.2.1. Then, in Section 4.2.2, we deduce an abstract error result which is instrumental for the derivation of *a priori* error estimates in Section 4.2.3.

4.2.1 Approximation properties of functions in Trefftz virtual element spaces

In order to discuss the approximation properties for the nonconforming Trefftz VE spaces, we recall the following local h-version best approximation result from [117, Theorem 5.2] for plane wave spaces in 2D.

Theorem 4.2.1. *Assume that K is an element of a mesh $\mathcal{T}_n$ satisfying the mesh assumption **(G1)** introduced in Section 2.3. In addition, let $u \in H^{s+1}(K)$, $s \in \mathbb{R}_{\geq 1}$, be such that $\Delta u + k^2 u = 0$, and let $\mathbb{PW}_p(K)$ be the plane wave space with directions $\{\mathbf{d}_\ell\}_{\ell=1,\dots,p}$, $p = 2q+1$, $q \in \mathbb{N}_{\geq 2}$, satisfying the assumption **(D1)** in Section 4.1.1. Then, for every $L \in \mathbb{R}$ with $1 \leq L \leq \min(q,s)$, there exists $w^K \in \mathbb{PW}_p(K)$ such that, for every $0 \leq j \leq L$, it holds*

$$\|u - w^K\|_{j,k,K} \leq c_{PW}(kh_K) h_K^{L+1-j} \|u\|_{L+1,k,K},$$

where

$$c_{PW}(t) := Ce^{\left(\frac{7}{4}-\frac{3}{4}\rho\right)t}\left(1 + t^{j+q+8}\right), \tag{4.33}$$

and the constant $C > 0$ depends on q, j, L, ρ, ρ_0, and the directions $\{\mathbf{d}_\ell\}$, but is independent of k, h_K, and u. Note that the constant $c_{PW}(kh_K)$ in (4.33) *is uniformly bounded as $h_K \to 0$.*

In the ensuing result, we prove that the best approximation error of functions in the nonconforming Trefftz VE space $V_h^{\Delta+k^2}$ can be estimated by the best error in (discontinuous) plane wave spaces. This can be seen as a generalization of Proposition 3.2.1 for the Laplace problem to the Helmholtz problem.

Theorem 4.2.2. *Consider a family of meshes $\{\mathcal{T}_n\}_{n\in\mathbb{N}}$ satisfying the assumptions **(G1)**-**(G3)** in Section 2.3, and **(A1)** in Section 4.1.2, and let $V_h^{\Delta+k^2}$ be the nonconforming Trefftz VE space defined in* (4.20) *with directions $\{\mathbf{d}_\ell\}_{\ell=1,\dots,p}$, $p = 2q+1$, $q \in \mathbb{N}_{\geq 2}$, satisfying the assumption **(D1)** in Section 4.1.1. Further, assume that, on every element $K \in \mathcal{T}_n$, k and h_K are such that kh_K is sufficiently small, see condition* (4.49) *below. Then, for any $u \in H^1(\Omega)$, there exists a function $u_I \in V_h^{\Delta+k^2}$ such that*

$$\|u - u_I\|_{1,k,\mathcal{T}_n} \leq c_{BA}(kh)\|u - w^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} \quad \forall w^{\mathcal{T}_n} \in \mathbb{PW}_p(\mathcal{T}_n), \tag{4.34}$$

where

$$c_{BA}(t) := 2\frac{\delta}{\delta-1}(1 + C_P^2 t^2)\left(C_T(C_P^2+1)t + 2\right), \tag{4.35}$$

with C_T from (2.27)*, C_P from* (2.28)*, and $\delta > 1$ from condition* (4.49) *below, remains uniformly bounded as $t \to 0$.*

Proof. Given $u \in H^1(\Omega)$, we define its "interpolant" u_I in $V_h^{\Delta+k^2}$ in terms of its degrees of freedom as follows:

$$\int_e (u_I - u)\overline{w_\ell^e}\,\mathrm{d}s = 0 \quad \forall \ell \in \mathcal{J}_e,\, \forall e \in \mathcal{E}^K,\, \forall K \in \mathcal{T}_n, \tag{4.36}$$

where the functions w_ℓ^e are defined in (4.8).

We stress that, with this definition, u_I is automatically an element of $H^{1,nc}(\mathcal{T}_n)$ introduced in (4.19). Moreover, the definition (4.36) implies that the average of $u - u_I$ on every edge $e \in \mathcal{E}^K$, $K \in \mathcal{T}_n$, is zero, thanks to the fact that the space $\mathbb{PW}_p^c(e)$ contains the constants for all edges e. This, together with the Poincaré-Friedrichs inequality (2.28), gives, for each element $K \in \mathcal{T}_n$,

$$\|u - u_I\|_{0,K} \leq C_P h_K |u - u_I|_{1,K}. \tag{4.37}$$

In order to obtain (4.34), we start by proving local approximation estimates. To this end, let $K \in \mathcal{T}_n$ be fixed. By using the triangle inequality, we obtain

$$|u - u_I|_{1,K} \leq |u - w^K|_{1,K} + |u_I - w^K|_{1,K} \quad \forall w^K \in \mathbb{PW}_p(K). \tag{4.38}$$

Concerning the second term, by using an integration by parts, taking into account that both u_I and w^K belong to the kernel of the Helmholtz operator, and by employing the definition of the impedance trace γ_I^K, we get, for every constant $c_K \in \mathbb{C}$,

$$\begin{aligned}
|u_I - w^K|_{1,K}^2 &= \int_K \nabla(u_I - w^K) \cdot \overline{\nabla(u_I - w^K)}\,\mathrm{d}x \\
&= -\int_K \Delta(u_I - w^K)\overline{(u_I - w^K)}\,\mathrm{d}x \\
&\quad + \int_{\partial K} \nabla(u_I - w^K - c_K) \cdot \mathbf{n}_K\,\overline{(u_I - w^K)}\,\mathrm{d}s \\
&= k^2 \int_K (u_I - w^K)\overline{(u_I - w^K)}\,\mathrm{d}x + \int_{\partial K} \gamma_I^K(u_I - w^K - c_K)\overline{(u_I - w^K)}\,\mathrm{d}s \\
&\quad - \mathrm{i}k \int_{\partial K} (u_I - w^K - c_K)\overline{(u_I - w^K)}\,\mathrm{d}s.
\end{aligned} \tag{4.39}$$

Taking now into account that $\gamma_I^K(u_I - w^K - c_K)_{|_e}$ belongs to the space $\mathbb{PW}_p^c(e)$ introduced in (4.11), for each edge $e \in \mathcal{E}^K$, the definition of u_I in (4.36) implies

$$\int_{\partial K} \gamma_I^K(u_I - w^K - c_K)\overline{(u_I - w^K)}\,\mathrm{d}s = \int_{\partial K} \gamma_I^K(u_I - w^K - c_K)\overline{(u - w^K)}\,\mathrm{d}s. \tag{4.40}$$

Using the definition of impedance traces, inserting (4.40) in (4.39), integrating by parts back, and using that both u_I and w^K belong to the kernel of the Helmholtz operator lead to

$$\begin{aligned}
|u_I - w^K|_{1,K}^2 &= k^2 \int_K (u_I - w^K)\overline{(u_I - w^K)}\,\mathrm{d}x + \int_K \nabla(u_I - w^K) \cdot \overline{\nabla(u - w^K)}\,\mathrm{d}x \\
&\quad + \int_K \Delta(u_I - w^K)\overline{(u - w^K)}\,\mathrm{d}x + \mathrm{i}k \int_{\partial K} (u_I - w^K - c_K)\overline{(u - u_I)}\,\mathrm{d}s \\
&= k^2 \int_K (u_I - w^K)\overline{(u_I - u)}\,\mathrm{d}x + \int_K \nabla(u_I - w^K) \cdot \overline{\nabla(u - w^K)}\,\mathrm{d}x \\
&\quad + \mathrm{i}k \int_{\partial K} (u_I - w^K - c_K)\overline{(u - u_I)}\,\mathrm{d}s =: Z_1 + Z_2 + Z_3.
\end{aligned} \tag{4.41}$$

We derive bounds for the three terms Z_1-Z_3 separately. For Z_1, we use the Cauchy-Schwarz and the triangle inequalities, the inequality (4.37), and the bound $a^2 + ab \leqslant \frac{1}{2}(3a^2 + b^2)$, for all $a, b \geq 0$, to get

$$\begin{aligned}
|Z_1| &= \left| k^2 \int_K (u_I - w^K)\overline{(u_I - u)}\,\mathrm{d}x \right| \leq k^2 \left(\|u - u_I\|_{0,K}^2 + \|u - w^K\|_{0,K} \|u - u_I\|_{0,K} \right) \\
&\leq k^2 \left\{ C_P^2 h_K^2 |u - u_I|_{1,K}^2 + C_P h_K |u - u_I|_{1,K} \|u - w^K\|_{0,K} \right\} \\
&\leq \frac{k^2}{2} \left\{ 3 C_P^2 h_K^2 |u - u_I|_{1,K}^2 + \|u - w^K\|_{0,K}^2 \right\}.
\end{aligned} \tag{4.42}$$

The term Z_2 can be estimated by applying the Cauchy-Schwarz inequality and $ab \leq \frac{1}{2}(a^2 + b^2)$:

$$|Z_2| = \left| \int_K \nabla(u_I - w^K) \cdot \overline{\nabla(u - w^K)}\,\mathrm{d}x \right| \leq \frac{1}{2} \left(|u_I - w^K|_{1,K}^2 + |u - w^K|_{1,K}^2 \right). \tag{4.43}$$

Finally, for the term Z_3, by employing the Cauchy-Schwarz and the triangle inequalities, and again the bound $a^2 + ab \leqslant \frac{1}{2}(3a^2 + b^2)$, we obtain

$$\begin{aligned}
|Z_3| &= \left| \mathrm{i}k \int_{\partial K} (u_I - w^K - c_K)\overline{(u - u_I)}\,\mathrm{d}s \right| \\
&\leq k \left(\|u - u_I\|_{0,\partial K}^2 + \|u - w^K - c_K\|_{0,\partial K} \|u - u_I\|_{0,\partial K} \right) \\
&\leq \frac{k}{2} \left(3 \|u - u_I\|_{0,\partial K}^2 + \|u - w^K - c_K\|_{0,\partial K}^2 \right).
\end{aligned} \tag{4.44}$$

Combining the trace inequality (2.27) with (4.37) yields

$$\begin{aligned}\|u-u_I\|^2_{0,\partial K} &\le C_T\left(h_K^{-1}\|u-u_I\|^2_{0,K}+h_K|u-u_I|^2_{1,K}\right)\\ &\le C_T(C_P^2+1)h_K|u-u_I|^2_{1,K}.\end{aligned} \tag{4.45}$$

Similarly, making use of the trace inequality (2.27) and the Poincaré-Friedrichs inequality (2.29), after selecting $c_K=\frac{1}{|K|}\int_K(u-w^K)\,\mathrm{d}x$, leads to

$$\begin{aligned}\|u-w^K-c_K\|^2_{0,\partial K} &\le C_T\left(h_K^{-1}\|u-w^K-c_K\|^2_{0,K}+h_K|u-w^K|^2_{1,K}\right)\\ &\le C_T(C_P^2+1)h_K|u-w^K|^2_{1,K}.\end{aligned} \tag{4.46}$$

By plugging (4.45) and (4.46) into (4.44), we obtain

$$|Z_3|\le\frac{1}{2}C_T(C_P^2+1)kh_K\left(3|u-u_I|^2_{1,K}+|u-w^K|^2_{1,K}\right). \tag{4.47}$$

Inserting the three bounds (4.42), (4.43), and (4.47) into (4.41), and moving the contribution $\frac{1}{2}|u_I-w^K|^2_{1,K}$ to the left-hand side, yield

$$\begin{aligned}\frac{1}{2}|u_I-w^K|^2_{1,K}\le{}&\frac{3}{2}kh_K\left(C_P^2kh_K+C_T(C_P^2+1)\right)|u-u_I|^2_{1,K}\\ &+\frac{1}{2}k^2\|u-w^K\|^2_{0,K}+\frac{1}{2}\left(1+C_T(C_P^2+1)kh_K\right)|u-w^K|^2_{1,K}.\end{aligned} \tag{4.48}$$

From (4.38), the bound $(a+b)^2\leqslant 2(a^2+b^2)$, and (4.48), we get, further taking the definition of the norm $\|\cdot\|_{1,k,K}$ into account,

$$\begin{aligned}|u-u_I|^2_{1,K}&\le 2|u-w^K|^2_{1,K}+2|u_I-w^K|^2_{1,K}\\ &\le 6kh_K(C_P^2kh_K+C_T(C_P^2+1))|u-u_I|^2_{1,K}+2k^2\|u-w^K\|^2_{0,K}\\ &\quad+2\left(C_T(C_P^2+1)kh_K+2\right)|u-w^K|^2_{1,K}\\ &\le 6kh_K(C_P^2kh_K+C_T(C_P^2+1))|u-u_I|^2_{1,K}+2\left(C_T(C_P^2+1)kh_K+2\right)\|u-w^K\|^2_{1,k,K}.\end{aligned}$$

Under the assumption that k and h_K are such that

$$6kh_K(C_P^2kh_K+C_T(C_P^2+1))\le\frac{1}{\delta} \tag{4.49}$$

for some $\delta>1$, we obtain

$$|u-u_I|^2_{1,K}\le 2\frac{\delta}{\delta-1}\left(C_T(C_P^2+1)kh_K+2\right)\|u-w^K\|^2_{1,k,K}\quad\forall w^K\in\mathbb{PW}_p(K). \tag{4.50}$$

From the definition of $\|\cdot\|_{1,k,K}$ in (2.2), inequality (4.37), and the estimate (4.50), we get

$$\begin{aligned}\|u-u_I\|^2_{1,k,K}&=|u-u_I|^2_{1,K}+k^2\|u-u_I\|^2_{0,K}\le(1+C_P^2(kh_K)^2)|u-u_I|^2_{1,K}\\ &\le 2\frac{\delta}{\delta-1}(1+C_P^2(kh_K)^2)\left(C_T(C_P^2+1)kh_K+2\right)\|u-w^K\|^2_{1,k,K}.\end{aligned}$$

The assertion follows by summing over all elements $K\in\mathcal{T}_n$ and taking the square root. □

By combining Theorem 4.2.1 with Theorem 4.2.2, we have the following best approximation error bound.

Corollary 4.2.3. *Under the assumptions of Theorems 4.2.1 and 4.2.2, for $u\in H^{s+1}(\Omega)$, $s\in\mathbb{R}_{\ge1}$, satisfying $\Delta u+k^2u=0$, the following bound holds true:*

$$\|u-u_I\|_{1,k,\mathcal{T}_n}\le C^*(kh)h^\zeta\|u\|_{\zeta+1,k,\mathcal{T}_n},$$

where $\zeta:=\min(q,s)$ and

$$C^*(t):=Ce^{\left(\frac{7}{4}-\frac{3}{4}\rho\right)t}\left(1+t^{q+9}\right)2\frac{\delta}{\delta-1}(1+C_P^2t^2)\left(C_T(C_P^2+1)t+2\right),$$

with $C>0$ depending on q, ρ, ρ_0, and $\{\mathbf{d}_\ell\}_{\ell=1,\dots,p}$, but independent of k, h, and u, and with δ as in Theorem 4.2.2.

4.2.2 Abstract error analysis

In this section, we prove existence and uniqueness of the discrete solution to the method (4.4), and we derive *a priori* error bounds, provided that the mesh size is sufficiently small.

To this purpose, we consider a variational formulation of (4.1) obtained by testing with functions in $V_h^{\Delta+k^2}$. Given u the exact solution to problem (4.1), we have, for all functions $v_h \in V_h^{\Delta+k^2}$,

$$0 = \sum_{K\in\mathcal{T}_n} \int_K (-\Delta u - k^2 u)\overline{v_h}\,\mathrm{d}x = \sum_{K\in\mathcal{T}_n} \left[\int_K \left(\nabla u \cdot \overline{\nabla v_h} - k^2 u\overline{v_h}\right)\mathrm{d}x - \int_{\partial K} (\nabla u \cdot \mathbf{n}_K)\overline{v_h}\,\mathrm{d}s\right],$$

and therefore

$$\sum_{K\in\mathcal{T}_n} a^K(u,v_h) - \mathcal{N}_h(u,v_h) + \mathrm{i}k\int_\Gamma u\overline{v_h}\,\mathrm{d}s = \int_\Gamma g\overline{v_h}\,\mathrm{d}s, \tag{4.51}$$

where the nonconformity term $\mathcal{N}_h(\cdot,\cdot)$ is defined as

$$\mathcal{N}_h(u,v_h) := \sum_{K\in\mathcal{T}_n} \int_{\partial K\setminus\Gamma} (\nabla u\cdot\mathbf{n}_K)\,\overline{v_h}\,\mathrm{d}s = \sum_{e\in\mathcal{E}_n^I} \int_e \nabla u \cdot \overline{[\![v_h]\!]}\,\mathrm{d}s. \tag{4.52}$$

Now, we have all the ingredients to prove the following abstract error result.

Theorem 4.2.4. *Let the assumptions **(G1)**-**(G3)** in Section 2.3, **(D1)** in Section 4.1.1, and **(A1)** in Section 4.1.2 hold true. Moreover, assume that $u \in H^2(\Omega)$, where u is the solution to* (4.2)*. Further, let the number of plane waves be $p = 2q+1$, $q \in \mathbb{N}_{\geq 2}$, and let the local stabilization forms $S_k^K(\cdot,\cdot)$ be such that the following properties are valid:*

- (local discrete continuity) *there exists a constant $\gamma_h > 0$ such that*

$$|a_{k,h}^K(v_h,z_h)| \leq \gamma_h \|v_h\|_{1,k,K}\|z_h\|_{1,k,K} \quad \forall v_h, z_h \in V^{\Delta+k^2}(K),\ \forall K\in\mathcal{T}_n; \tag{4.53}$$

- (discrete Gårding inequality) *there exists a constant $\alpha_h > 0$ such that*

$$Re[b_{k,h}(v_h,v_h)] + 2k^2\|v_h\|_{0,\Omega}^2 \geq \alpha_h \|v_h\|_{1,k,\mathcal{T}_n}^2 \quad \forall v_h \in V_h^{\Delta+k^2}. \tag{4.54}$$

Then, provided that k and h are chosen such that k^2h is sufficiently small, see condition (4.93) *below, the method* (4.4) *admits a unique solution $u_h \in V_h^{\Delta+k^2}$ which satisfies*

$$\begin{aligned}\|u-u_h\|_{1,k,\mathcal{T}_n} \lesssim\ & \aleph_1(k,h)\|u-w^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} + h\,\aleph_2(k,h)|u-w^{\mathcal{T}_n}|_{2,\mathcal{T}_n} \\ & + h^{\frac{1}{2}}\,\aleph_2(k,h)\|g-\Pi_p^{0,\Gamma}g\|_{0,\Gamma} \quad \forall w^{\mathcal{T}_n}\in\mathbb{PW}_p(\mathcal{T}_n),\end{aligned} \tag{4.55}$$

with

$$\begin{aligned}\aleph_1(k,h) &:= \frac{(kh+\gamma_h+1)(k(\varsigma(k,h)+\vartheta(k,h))+c_{BA}(kh)+1)}{\alpha_h} + c_{BA}(kh),\\ \aleph_2(k,h) &:= \frac{k(1+\vartheta(k,h))+1}{\alpha_h},\end{aligned} \tag{4.56}$$

where $\Pi_p^{0,\Gamma}$ is defined in (4.25)*, the hidden constants in* (4.55) *are independent of h and k, and*

$$\varsigma(k,h) := (1+kh)(1+d_\Omega k)h, \quad \vartheta(k,h) := c_{BA}(kh)\varsigma(k,h), \tag{4.57}$$

$c_{BA}(kh)$ being given in (4.35) *and d_Ω a positive constant depending only on Ω.*

Proof. We prove the error bound (4.55) under a condition on k^2h in five steps. Existence and uniqueness of discrete solutions, under the same assumption on k^2h, will follow as in [165].

Step 1: Triangle inequality: Let u_h satisfy (4.4). By the triangle inequality, we get

$$\|u-u_h\|_{1,k,\mathcal{T}_n} \leq \|u-u_I\|_{1,k,\mathcal{T}_n} + \|u_h-u_I\|_{1,k,\mathcal{T}_n}, \tag{4.58}$$

where $u_I \in V_h^{\Delta+k^2}$ is defined as in (4.36). The first term on the right-hand side of (4.58) can be estimated by using Theorem 4.2.2. We focus on the second one. By setting $\delta_h := u_h - u_I \in V_h^{\Delta+k^2}$ and using the discrete Gårding inequality (4.54), we obtain

$$\alpha_h \|\delta_h\|^2_{1,k,\mathcal{T}_n} \leq \mathrm{Re}\,[b_{k,h}(\delta_h,\delta_h)] + 2k^2\|\delta_h\|^2_{0,\Omega} =: I + II. \tag{4.59}$$

Step 2: Estimate of the term I in (4.59): The identity in (4.4), the definitions of $b_{k,h}(\cdot,\cdot)$ in (4.29), of $F_h(\cdot)$ in (4.32), and of the projector $\Pi_p^{0,\Gamma}$ in (4.25), together with the plane wave consistency property (4.28) yield

$$\begin{aligned} b_{k,h}(\delta_h,\delta_h) &= b_{k,h}(u_h - u_I,\delta_h) = b_{k,h}(u_h,\delta_h) - b_{k,h}(u_I,\delta_h)\\ &= \int_\Gamma g\overline{(\Pi_p^{0,\Gamma}\delta_h)}\,\mathrm{d}s - a_{k,h}(u_I,\delta_h) - \mathrm{i}k\int_\Gamma (\Pi_p^{0,\Gamma}u_I)\overline{(\Pi_p^{0,\Gamma}\delta_h)}\,\mathrm{d}s\\ &= \int_\Gamma (\Pi_p^{0,\Gamma}g)\overline{\delta_h}\,\mathrm{d}s - \sum_{K\in\mathcal{T}_n} a^K_{k,h}(u_I,\delta_h) - \mathrm{i}k\int_\Gamma (\Pi_p^{0,\Gamma}u_I)\overline{(\Pi_p^{0,\Gamma}\delta_h)}\,\mathrm{d}s\\ &= \int_\Gamma (\Pi_p^{0,\Gamma}g - g)\overline{\delta_h}\,\mathrm{d}s + \int_\Gamma g\overline{\delta_h}\,\mathrm{d}s - \sum_{K\in\mathcal{T}_n} a^K_{k,h}(u_I - w^{\mathcal{T}_n},\delta_h)\\ &\qquad - \sum_{K\in\mathcal{T}_n} a^K(w^{\mathcal{T}_n},\delta_h) - \mathrm{i}k\int_\Gamma (\Pi_p^{0,\Gamma}u_I)\overline{\delta_h}\,\mathrm{d}s, \end{aligned}$$

where $w^{\mathcal{T}_n} \in \mathbb{PW}_p(\mathcal{T}_n)$. Consequently, by applying the identity (4.51), we get

$$\begin{aligned} b_{k,h}(\delta_h,\delta_h) &= \int_\Gamma (\Pi_p^{0,\Gamma}g - g)\overline{\delta_h}\,\mathrm{d}s + \sum_{K\in\mathcal{T}_n} a^K(u,\delta_h) - \mathcal{N}_h(u,\delta_h) + \mathrm{i}k\int_\Gamma u\overline{\delta_h}\,\mathrm{d}s\\ &\qquad - \sum_{K\in\mathcal{T}_n} a^K_{k,h}(u_I - w^{\mathcal{T}_n},\delta_h) - \sum_{K\in\mathcal{T}_n} a^K(w^{\mathcal{T}_n},\delta_h) - \mathrm{i}k\int_\Gamma (\Pi_p^{0,\Gamma}u_I)\overline{\delta_h}\,\mathrm{d}s\\ &= \sum_{K\in\mathcal{T}_n} a^K_k(u - w^{\mathcal{T}_n},\delta_h) - \sum_{K\in\mathcal{T}_n} a^K_{k,h}(u_I - w^{\mathcal{T}_n},\delta_h) + \mathrm{i}k\int_\Gamma (u - \Pi_p^{0,\Gamma}u_I)\overline{\delta_h}\,\mathrm{d}s\\ &\qquad + \int_\Gamma (\Pi_p^{0,\Gamma}g - g)\overline{\delta_h}\,\mathrm{d}s - \mathcal{N}_h(u,\delta_h) =: R_1 + R_2 + R_3 + R_4 + R_5. \end{aligned}$$

We note that

$$I = \mathrm{Re}\,[b_{k,h}(\delta_h,\delta_h)] \leq |b_{k,h}(\delta_h,\delta_h)| \leq |R_1| + |R_2| + |R_3| + |R_4| + |R_5|, \tag{4.60}$$

and we proceed by deriving bounds for each of the five terms appearing on the right-hand side of (4.60). The term R_1 can be estimated by using the continuity of the local continuous sesquilinear forms:

$$|R_1| = \left|\sum_{K\in\mathcal{T}_n} a^K_k(u - w^{\mathcal{T}_n},\delta_h)\right| \leq \|u - w^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n}\|\delta_h\|_{1,k,\mathcal{T}_n}. \tag{4.61}$$

For R_2, we make use of the local discrete continuity assumption (4.53):

$$\begin{aligned} |R_2| &\leq \left|\sum_{K\in\mathcal{T}_n} a^K_{k,h}(u_I - w^{\mathcal{T}_n},\delta_h)\right| \leq \gamma_h\|u_I - w^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n}\|\delta_h\|_{1,k,\mathcal{T}_n}\\ &\leq \gamma_h\left\{\|u - u_I\|_{1,k,\mathcal{T}_n} + \|u - w^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n}\right\}\|\delta_h\|_{1,k,\mathcal{T}_n}. \end{aligned} \tag{4.62}$$

Regarding R_3, it holds with the properties of $\Pi_p^{0,\Gamma}$ and the definition of u_I in (4.36)

$$R_3 = \mathrm{i}k\int_\Gamma (u - \Pi_p^{0,\Gamma}u)\overline{\delta_h}\,\mathrm{d}s + \mathrm{i}k\int_\Gamma \Pi_p^{0,\Gamma}(u - u_I)\overline{\delta_h}\,\mathrm{d}s = \mathrm{i}k\int_\Gamma (u - \Pi_p^{0,\Gamma}u)\overline{\delta_h}\,\mathrm{d}s. \tag{4.63}$$

By using the definition and the properties of the L^2 projector $\Pi_p^{0,\Gamma}$ in (4.25), and by applying the $L^2(e)$, for all $e \in \mathcal{E}_n^B$, and the ℓ^2 Cauchy-Schwarz inequalities, we derive

$$\begin{aligned} |R_3| = \left| \mathrm{i}k \int_\Gamma (u - \Pi_p^{0,\Gamma} u)\overline{\delta_h}\,\mathrm{d}s \right| &= \left| \mathrm{i}k \sum_{e \in \mathcal{E}_n^B} \int_e (u - \Pi_p^{0,e} u)\overline{(\delta_h - c)}\,\mathrm{d}s \right| \\ \leq k \sum_{e \in \mathcal{E}_n^B} \|u - \Pi_p^{0,e} u\|_{0,e} \|\delta_h - c\|_{0,e} &\leq k \left(\sum_{e \in \mathcal{E}_n^B} \|u - \Pi_p^{0,e} u\|_{0,e}^2 \right)^{\frac{1}{2}} \left(\sum_{e \in \mathcal{E}_n^B} \|\delta_h - c\|_{0,e}^2 \right)^{\frac{1}{2}}, \end{aligned} \tag{4.64}$$

for any edgewise complex constant function c. We estimate the two terms on the right-hand side of (4.64) as follows. Given $e \in \mathcal{E}_n^B$, by using (4.24) and the definition of the norm $\|\cdot\|_{1,k,K}$, we have

$$\|u - \Pi_p^{0,e} u\|_{0,e} \lesssim h_K^{\frac{1}{2}} |u - w^K|_{1,K} \leq h_K^{\frac{1}{2}} \|u - w^K\|_{1,k,K} \quad \forall w^K \in \mathbb{PW}_p(K),$$

where K is the unique polygon in $\mathcal{T}_n$ such that $e \in \mathcal{E}^K \cap \mathcal{E}_n^B$.

Concerning the second term on the right-hand side of (4.64), we make use of the trace inequality (2.27) and the Poincaré-Friedrichs inequality (2.29), choosing $c = \frac{1}{|K|}\int_K \delta_h\,\mathrm{d}x$, to obtain

$$\begin{aligned} \|\delta_h - c\|_{0,e} \leq & \|\delta_h - c\|_{0,\partial K} \lesssim h_K^{-\frac{1}{2}} \|\delta_h - c\|_{0,K} + h_K^{\frac{1}{2}} |\delta_h|_{1,K} \\ & \lesssim h_K^{\frac{1}{2}} |\delta_h|_{1,K} \leq h_K^{\frac{1}{2}} \|\delta_h\|_{1,k,K}. \end{aligned} \tag{4.65}$$

Thus,

$$|R_3| \lesssim kh \|u - w^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} \|\delta_h\|_{1,k,\mathcal{T}_n} \quad \forall w^{\mathcal{T}_n} \in \mathbb{PW}_p(\mathcal{T}_n). \tag{4.66}$$

For the term R_4, by mimicking what was done in (4.64) and (4.65), i.e. making appear an edgewise constant on Γ and using the Poincaré inequality, we get

$$|R_4| = \left| \int_\Gamma (g - \Pi_p^{0,\Gamma} g)\overline{\delta_h}\,\mathrm{d}s \right| \lesssim h^{\frac{1}{2}} \|g - \Pi_p^{0,\Gamma} g\|_{0,\Gamma} \|\delta_h\|_{1,k,\mathcal{T}_n}. \tag{4.67}$$

Finally, we study the nonconformity term R_5 on the right-hand side of (4.60). Using the definitions of the nonconforming space $V_h^{\Delta+k^2}$ in (4.20) and of the projector $\Pi_p^{0,e}$ in (4.23), together with the Cauchy-Schwarz inequality, yield

$$\begin{aligned} |R_5| = |\mathcal{N}_h(u, \delta_h)| = \left| \sum_{e \in \mathcal{E}_n^I} \int_e \nabla u \cdot [\![\delta_h]\!]\,\mathrm{d}s \right| &= \left| \sum_{e \in \mathcal{E}_n^I} \int_e (\nabla u - \Pi_p^{0,e}(\nabla u)) \cdot \mathbf{n}_e \overline{(\delta_h^+ - \delta_h^-)}\,\mathrm{d}s \right| \\ \leq \left| \sum_{e \in \mathcal{E}_n^I} \int_e (\nabla u - \Pi_p^{0,e}(\nabla u)) \cdot \mathbf{n}_e \overline{(\delta_h^+ - c^+)}\,\mathrm{d}s \right| & \\ + \left| \sum_{e \in \mathcal{E}_n^I} \int_e (\nabla u - \Pi_p^{0,e}(\nabla u)) \cdot \mathbf{n}_e \overline{(\delta_h^- - c^-)}\,\mathrm{d}s \right| & \\ \leq \left(\sum_{e \in \mathcal{E}_n^I} \|\nabla u \cdot \mathbf{n}_e - \Pi_p^{0,e}(\nabla u \cdot \mathbf{n}_e)\|_{0,e}^2 \right)^{\frac{1}{2}} \left(\sum_{e \in \mathcal{E}_n^I} \|\delta_h^+ - c^+\|_{0,e}^2 \right)^{\frac{1}{2}} & \\ + \left(\sum_{e \in \mathcal{E}_n^I} \|\nabla u \cdot \mathbf{n}_e - \Pi_p^{0,e}(\nabla u \cdot \mathbf{n}_e)\|_{0,e}^2 \right)^{\frac{1}{2}} \left(\sum_{e \in \mathcal{E}_n^I} \|\delta_h^- - c^-\|_{0,e}^2 \right)^{\frac{1}{2}}, & \end{aligned} \tag{4.68}$$

for any edgewise complex constant functions c^+ and $c^- \in \mathbb{C}$. After applying the Cauchy-Schwarz inequality to both terms on the right-hand side of (4.68), we estimate the resulting terms as follows.

We begin with the bounds on the terms involving ∇u. Denoting by K^+ and K^- the two polygons in $\mathcal{T}_n$ with $e \in \mathcal{E}^{K^+} \cap \mathcal{E}^{K^-}$, owing to the trace inequality (2.27) and the inequality (2.29) for any $w^{K^\pm} \in \mathbb{PW}(K^\pm)$ and $(\mathbf{c}^\pm)_i = \frac{1}{|K^\pm|}\int_{K^\pm} \nabla(u - w^{K^\pm})\,\mathrm{d}x$, $i = 1, 2$, it holds

$$\begin{aligned}\sum_{e\in\mathcal{E}_n^I} \|\nabla u\cdot\mathbf{n}_e - \Pi_p^{0,e}(\nabla u\cdot\mathbf{n}_e)\|_{0,e}^2 &\le \sum_{e\in\mathcal{E}_n^I} \|(\nabla u - \nabla w^{K^\pm} - \mathbf{c}^\pm)\cdot\mathbf{n}_e\|_{0,e}^2\\ &\le \sum_{K\in\mathcal{T}_n} \|(\nabla u - \nabla w^{K^\pm} - \mathbf{c}^\pm)\cdot\mathbf{n}_e\|_{0,\partial K}^2 \lesssim \sum_{K\in\mathcal{T}_n} h_K|u-w|_{2,K}^2.\end{aligned}$$

For the terms with δ_h, we take $c^\pm = \frac{1}{|K^\pm|}\int_{K^\pm} \delta_h^\pm\,\mathrm{d}x$ and follow the computations in (4.65), to get

$$\|\delta_h^\pm - c^\pm\|_{0,e}^2 \lesssim h\|\delta_h^\pm\|_{1,k,K^\pm}^2,$$

Thus, a bound on the nonconformity term R_5 is given by

$$|R_5| = |\mathcal{N}_h(u,\delta_h)| \lesssim h|u - w^{\mathcal{T}_n}|_{2,\mathcal{T}_n}\|\delta_h\|_{1,k,\mathcal{T}_n} \quad \forall w^{\mathcal{T}_n} \in \mathbb{PW}_p(\mathcal{T}_n). \tag{4.69}$$

Collecting (4.61), (4.62), (4.66), (4.67), and (4.69), from (4.60), we get

$$\begin{aligned}I = \mathrm{Re}\,[b_{k,h}(\delta_h,\delta_h)] \lesssim \{&(kh+\gamma_h+1)\|u-w^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} + \gamma_h\|u-u_I\|_{1,k,\mathcal{T}_n}\\ &+ h^{\frac12}\|g-\Pi_p^{0,\Gamma}g\|_{0,\Gamma} + h|u-w^{\mathcal{T}_n}|_{2,\mathcal{T}_n}\}\|\delta_h\|_{1,k,\mathcal{T}_n}.\end{aligned} \tag{4.70}$$

Step 3: Estimate of the term II in (4.59): By using simple algebra and the definitions of δ_h and the norm $\|\cdot\|_{1,k,\mathcal{T}_n}$, we obtain

$$\begin{aligned}II = 2k^2\|\delta_h\|_{0,\Omega}^2 = 2k^2\|u_h-u_I\|_{0,\Omega}\|\delta_h\|_{0,\Omega} &\le 2k^2\left\{\|u-u_h\|_{0,\Omega} + \|u-u_I\|_{0,\Omega}\right\}\|\delta_h\|_{0,\Omega}\\ &\le 2\left\{k\|u-u_h\|_{0,\Omega} + \|u-u_I\|_{1,k,\mathcal{T}_n}\right\}\|\delta_h\|_{1,k,\mathcal{T}_n}.\end{aligned} \tag{4.71}$$

We plug (4.70) and (4.71) into (4.59) and divide by $\|\delta_h\|_{1,k,\mathcal{T}_n}$, deducing

$$\begin{aligned}\alpha_h\|u_h-u_I\|_{1,k,\mathcal{T}_n} \lesssim{}& (kh+\gamma_h+1)\|u-w^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} + (\gamma_h+1)\|u-u_I\|_{1,k,\mathcal{T}_n}\\ &+ h^{\frac12}\|g-\Pi_p^{0,\Gamma}g\|_{0,\Gamma} + h|u-w^{\mathcal{T}_n}|_{2,\mathcal{T}_n} + k\|u-u_h\|_{0,\Omega}.\end{aligned} \tag{4.72}$$

Step 4: Estimate of $\|u-u_h\|_{0,\Omega}$: We consider the auxiliary dual problem: find ψ such that

$$\begin{cases} -\Delta\psi - k^2\psi = u - u_h & \text{in } \Omega\\ \nabla\psi\cdot\mathbf{n}_\Omega - \mathrm{i}k\psi = 0 & \text{on } \Gamma.\end{cases} \tag{4.73}$$

The convexity of Ω and [149, Proposition 8.1.4] imply that the solution ψ to the weak formulation of (4.73) belongs to $H^2(\Omega)$ and that the stability bounds

$$\|\psi\|_{1,k,\mathcal{T}_n} \le d_\Omega\|u-u_h\|_{0,\Omega}, \qquad |\psi|_{2,\Omega} \lesssim (1+d_\Omega k)\|u-u_h\|_{0,\Omega} \tag{4.74}$$

are valid, with d_Ω being a positive universal constant depending only on Ω.

In addition, for all $K \in \mathcal{T}_n$, there exists $\psi^K \in \mathbb{PW}_p(K)$ such that, see [112, Propositions 3.12 and 3.13],

$$\begin{aligned}\|\psi-\psi^K\|_{0,K} &\lesssim h_K^2(|\psi|_{2,K} + k^2\|\psi\|_{0,K}),\\ |\psi-\psi^K|_{1,K} &\lesssim h_K(kh_K+1)(|\psi|_{2,K} + k^2\|\psi\|_{0,K}),\end{aligned} \tag{4.75}$$

where the hidden constants depend only on the shape of the element K and on p. Hence, by combining (4.75) with (4.74), there exists $\psi^{\mathcal{T}_n} \in \mathbb{PW}_p(\mathcal{T}_n)$ such that

$$\|\psi-\psi^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} \lesssim h(1+hk)(1+d_\Omega k)\|u-u_h\|_{0,\Omega} =: \varsigma(k,h)\|u-u_h\|_{0,\Omega}, \tag{4.76}$$

where the hidden constant is independent of h, k, and ψ.

Besides, thanks to Theorem 4.2.2, together with (4.74) and (4.75), defining the "interpolant" ψ_I of ψ as in (4.36),

$$\|\psi - \psi_I\|_{1,k,\mathcal{T}_n} \leq c_{BA}(kh)\|\psi - \psi^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} \lesssim \vartheta(k,h)\|u - u_h\|_{0,\Omega}, \tag{4.77}$$

where $c_{BA}(kh_K)$ and $\vartheta(k,h)$ are defined in (4.35) and (4.57), respectively.

From the definition of the dual problem (4.73), by integrating by parts and using the definition of $\mathcal{N}_h(\cdot,\cdot)$ in (4.52), we get

$$\begin{aligned}
\|u - u_h\|^2_{0,\Omega} &= \sum_{K\in\mathcal{T}_n} \int_K (-\Delta\psi - k^2\psi)\overline{(u-u_h)}\,\mathrm{d}x \\
&= \sum_{K\in\mathcal{T}_n} \left[\int_K \left(\nabla\psi \cdot \overline{\nabla(u-u_h)} - k^2\psi\overline{(u-u_h)}\right)\mathrm{d}x - \int_{\partial K}(\nabla\psi\cdot\mathbf{n}_K)\overline{(u-u_h)}\,\mathrm{d}s\right] \\
&= \sum_{K\in\mathcal{T}_n} a_k^K(\psi, u-u_h) - \mathcal{N}_h(\psi, u-u_h) - \int_\Gamma (\nabla\psi\cdot\mathbf{n}_\Omega)\overline{(u-u_h)}\,\mathrm{d}s \\
&= \sum_{K\in\mathcal{T}_n} a_k^K(\psi-\psi_I, u-u_h) + \sum_{K\in\mathcal{T}_n} a_k^K(\psi_I, u-u_h) - \mathcal{N}_h(\psi, u-u_h) - \mathrm{i}k\int_\Gamma \psi\overline{(u-u_h)}\,\mathrm{d}s \\
&=: S_1 + S_2 + S_3 + S_4.
\end{aligned} \tag{4.78}$$

Hence, we need to estimate the four terms on the right-hand side of (4.78). We begin with S_1. By using the continuity of the continuous local sesquilinear forms, together with (4.77), we have

$$\begin{aligned}
|S_1| = \left|\sum_{K\in\mathcal{T}_n} a_k^K(\psi-\psi_I, u-u_h)\right| &\leq \|\psi-\psi_I\|_{1,k,\mathcal{T}_n}\|u-u_h\|_{1,k,\mathcal{T}_n} \\
&\lesssim \vartheta(k,h)\|u-u_h\|_{0,\Omega}\|u-u_h\|_{1,k,\mathcal{T}_n}.
\end{aligned} \tag{4.79}$$

The nonconformity term S_3 can be estimated analogously as the term R_5 in (4.69). By taking the special choice $w^{\mathcal{T}_n} = 0$ and using (4.74), we arrive at

$$|S_3| = |\mathcal{N}_h(\psi, u-u_h)| \lesssim h(1 + d_\Omega k)\|u-u_h\|_{0,\Omega}\|u-u_h\|_{1,k,\mathcal{T}_n}. \tag{4.80}$$

It remains to control the terms S_2 and S_4. For S_2, we observe that using the identity (4.51), taking the complex conjugated of (4.4), and employing the definitions (4.29) and (4.32), give

$$\begin{aligned}
S_2 &= \sum_{K\in\mathcal{T}_n} a_k^K(\psi_I, u-u_h) = \sum_{K\in\mathcal{T}_n}\overline{a_k^K(u,\psi_I)} - \sum_{K\in\mathcal{T}_n} a_k^K(\psi_I, u_h) \\
&= \overline{\mathcal{N}_h(u,\psi_I)} + \int_\Gamma \overline{g}\psi_I\,\mathrm{d}s + \mathrm{i}k\int_\Gamma \overline{u}\psi_I\,\mathrm{d}s + \sum_{K\in\mathcal{T}_n}\left\{-a_{k,h}^K(\psi_I,u_h) + a_{k,h}^K(\psi_I,u_h) - a_k^K(\psi_I,u_h)\right\} \\
&= \overline{\mathcal{N}_h(u,\psi_I)} + \int_\Gamma \overline{g}(\psi_I - \Pi_p^{0,\Gamma}\psi_I)\,\mathrm{d}s + \mathrm{i}k\int_\Gamma \overline{(u - \Pi_p^{0,\Gamma}u_h)}\psi_I\,\mathrm{d}s \\
&\quad + \sum_{K\in\mathcal{T}_n}\left\{a_{k,h}^K(\psi_I,u_h) - a_k^K(\psi_I,u_h)\right\}.
\end{aligned}$$

We deduce

$$\begin{aligned}
S_2 + S_4 = &\overline{\mathcal{N}_h(u,\psi_I)} + \int_\Gamma \overline{g}(\psi_I - \Pi_p^{0,\Gamma}\psi_I)\,\mathrm{d}s \\
&+ \mathrm{i}k\left(\int_\Gamma \psi_I\overline{(u - \Pi_p^{0,\Gamma}u_h)}\,\mathrm{d}s - \int_\Gamma \psi\overline{(u-u_h)}\,\mathrm{d}s\right) \\
&+ \sum_{K\in\mathcal{T}_n}\left\{a_{k,h}^K(\psi_I,u_h) - a_k^K(\psi_I,u_h)\right\} =: T_1 + T_2 + T_3 + T_4.
\end{aligned} \tag{4.81}$$

The term T_1 can be estimated by using (4.69), (4.74), and (4.77):

$$\begin{aligned}
|T_1| = |\overline{\mathcal{N}_h(u,\psi_I)}| &\leq h|u - w^{\mathcal{T}_n}|_{2,\mathcal{T}_n}\|\psi_I\|_{1,k,\mathcal{T}_n} \\
&\leq h|u - w^{\mathcal{T}_n}|_{2,\mathcal{T}_n}\left(\|\psi\|_{1,k,\mathcal{T}_n} + \|\psi-\psi_I\|_{1,k,\mathcal{T}_n}\right) \\
&\lesssim h(1+\vartheta(k,h))|u - w^{\mathcal{T}_n}|_{2,\mathcal{T}_n}\|u-u_h\|_{0,\Omega}
\end{aligned} \tag{4.82}$$

for any $w^{\mathcal{T}_n} \in \mathbb{PW}_p(\mathcal{T}_n)$.

For T_2, we observe that, with the definition of the projector $\Pi_p^{0,\Gamma}$ given in (4.23), it follows

$$|T_2| = \left| \int_\Gamma g\overline{(\psi_I - \Pi_p^{0,\Gamma}\psi_I)}\,\mathrm{d}s \right| = \left| \int_\Gamma (g - \Pi_p^{0,\Gamma} g)\overline{(\psi_I - c)}\,\mathrm{d}s \right|,$$

where c is any edgewise complex constant. By doing similar computations as in (4.64) and (4.65), and employing also (4.74) and (4.77), we get

$$\begin{aligned}|T_2| &\lesssim h^{\frac{1}{2}} \|g - \Pi_p^{0,\Gamma} g\|_{0,\Gamma} \|\psi_I\|_{1,k,\mathcal{T}_n} \le h^{\frac{1}{2}} \|g - \Pi_p^{0,\Gamma} g\|_{0,\Gamma} \left(\|\psi\|_{1,k,\mathcal{T}_n} + \|\psi - \psi_I\|_{1,k,\mathcal{T}_n}\right) \\ &\le h^{\frac{1}{2}}(1 + \vartheta(k,h))\|g - \Pi_p^{0,\Gamma} g\|_{0,\Gamma} \|u - u_h\|_{0,\Omega}.\end{aligned} \tag{4.83}$$

The term T_4 can be estimated using the plane wave consistency property (4.28), the continuity of the sesquilinear forms $a_{k,h}(\cdot,\cdot)$ and $a_k^K(\cdot,\cdot)$, and the approximation estimates (4.76) and (4.77):

$$\begin{aligned}|T_4| &= \left| \sum_{K \in \mathcal{T}_n} \left\{ a_{k,h}^K(u_h, \psi_I) - a_k^K(u_h, \psi_I) \right\} \right| \\ &\le \sum_{K \in \mathcal{T}_n} \left| a_{k,h}^K(u_h - w^{\mathcal{T}_n}, \psi_I - \psi^{\mathcal{T}_n}) - a_k^K(u_h - w^{\mathcal{T}_n}, \psi_I - \psi^{\mathcal{T}_n}) \right| \\ &\le (\gamma_h + 1)\|u_h - w^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} \|\psi_I - \psi^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} \\ &\le (\gamma_h + 1)\left\{\|u - w^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} + \|u - u_h\|_{1,k,\mathcal{T}_n}\right\}\left\{\|\psi - \psi_I\|_{1,k,\mathcal{T}_n} + \|\psi - \psi^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n}\right\} \\ &\lesssim (\gamma_h + 1)\left\{\|u - w^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} + \|u - u_h\|_{1,k,\mathcal{T}_n}\right\}\{\vartheta(k,h) + \varsigma(k,h)\}\|u - u_h\|_{0,\Omega},\end{aligned} \tag{4.84}$$

for all $w^{\mathcal{T}_n}, \psi^{\mathcal{T}_n} \in \mathbb{PW}_p(\mathcal{T}_n)$.

Finally, we derive bounds for T_3. We compute

$$\begin{aligned}|T_3| &= k\left| \int_\Gamma \psi\overline{(u - u_h)}\,\mathrm{d}s - \int_\Gamma \psi_I\overline{(u - \Pi_p^{0,\Gamma} u_h)}\,\mathrm{d}s \right| \\ &= k\left| \int_\Gamma (\psi - \psi_I)\overline{(u - u_h)}\,\mathrm{d}s - \int_\Gamma \psi_I\overline{(u_h - \Pi_p^{0,\Gamma} u_h)}\,\mathrm{d}s \right|.\end{aligned}$$

Using the definitions of ψ_I as in (4.36) and of $\Pi_p^{0,\Gamma}$ in (4.25), we obtain

$$\begin{aligned}|T_3| &= k\left| \int_\Gamma (\psi - \psi_I)\overline{(u - u_h - \Pi_p^{0,\Gamma}(u - u_h))}\,\mathrm{d}s - \int_\Gamma (\psi_I - \Pi_p^{0,\Gamma}\psi_I)\overline{(u_h - \Pi_p^{0,\Gamma} u_h)}\,\mathrm{d}s \right| \\ &= k\Bigg| \int_\Gamma (\psi - \psi_I)\overline{(u - u_h - \Pi_p^{0,\Gamma}(u - u_h))}\,\mathrm{d}s - \int_\Gamma (\psi_I - \Pi_p^{0,\Gamma}\psi_I)\overline{(u_h - u)}\,\mathrm{d}s \\ &\qquad - \int_\Gamma (\psi_I - \Pi_p^{0,\Gamma}\psi_I)\overline{(u - \Pi_p^{0,\Gamma} u)}\,\mathrm{d}s \Bigg| \\ &=: k|T_3^A - T_3^B - T_3^C| \le k\left(|T_3^A| + |T_3^B| + |T_3^C|\right).\end{aligned} \tag{4.85}$$

We estimate the three terms on the right-hand side of (4.85) with tools analogous to those employed so far. The term T_3^A can be estimated using the Cauchy-Schwarz inequality, the trace inequality (2.27), the definition of ψ_I, the Poincaré-Friedrichs inequality (4.37), and the identity (4.24) with $w^K = 0$:

$$\begin{aligned}|T_3^A| &= \left| \int_\Gamma (\psi - \psi_I)\overline{(u - u_h - \Pi_p^{0,\Gamma}(u - u_h))}\,\mathrm{d}s \right| \le \|\psi - \psi_I\|_{0,\Gamma} \|u - u_h - \Pi_p^{0,\Gamma}(u - u_h)\|_{0,\Gamma} \\ &\lesssim h\|\psi - \psi_I\|_{1,k,\mathcal{T}_n} \|u - u_h\|_{1,k,\mathcal{T}_n}.\end{aligned} \tag{4.86}$$

For T_3^B, we can do analogous computations as in (4.64) and (4.65), getting

$$\begin{aligned}|T_3^B| &= \left| \int_\Gamma (\psi_I - \Pi_p^{0,\Gamma}\psi_I)\overline{(u - u_h)}\,\mathrm{d}s \right| \lesssim h\|\psi_I - \psi^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} \|u - u_h\|_{1,k,\mathcal{T}_n} \\ &\le h(\|\psi - \psi_I\|_{1,k,\mathcal{T}_n} + \|\psi - \psi^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n})\|u - u_h\|_{1,k,\mathcal{T}_n} \quad \forall \psi^{\mathcal{T}_n} \in \mathbb{PW}_p(\mathcal{T}_n).\end{aligned} \tag{4.87}$$

The term T_3^C is estimated by using (4.24):

$$\begin{aligned}|T_3^C| = \left|\int_\Gamma (\psi_I - \Pi_p^{0,\Gamma}\psi_I)\overline{(u - \Pi_p^{0,\Gamma}u)}\,\mathrm{d}s\right| &\lesssim h\|\psi_I - \psi^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n}\|u - w^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} \\ &\leq h(\|\psi - \psi_I\|_{1,k,\mathcal{T}_n} + \|\psi - \psi^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n})\|u - w^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} \quad \forall w^{\mathcal{T}_n},\ \psi^{\mathcal{T}_n} \in \mathbb{PW}_p(\mathcal{T}_n).\end{aligned} \tag{4.88}$$

Plugging (4.86), (4.87), and (4.88) in (4.85), and using (4.76) and (4.77), yield

$$\begin{aligned}|T_3| &\lesssim kh(\|\psi - \psi^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} + \|\psi - \psi_I\|_{1,k,\mathcal{T}_n})\left(\|u - w^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} + \|u - u_h\|_{1,k,\mathcal{T}_n}\right) \\ &\lesssim kh\left(\|u - w^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} + \|u - u_h\|_{1,k,\mathcal{T}_n}\right)(\varsigma(k,h) + \vartheta(k,h))\|u - u_h\|_{0,\Omega}.\end{aligned} \tag{4.89}$$

Collecting and inserting (4.82), (4.83), (4.84), and (4.89) into (4.81), we obtain the following bound:

$$\begin{aligned}|S_2 + S_4| \lesssim \big\{&(1 + \vartheta(k,h))(h|u - w^{\mathcal{T}_n}|_{2,\mathcal{T}_n} + h^{\frac{1}{2}}\|g - \Pi_p^{0,e}g\|_{0,\Gamma}) \\ &+ (\gamma_h + 1 + kh)(\varsigma(k,h) + \vartheta(k,h)) \\ &\quad\big[\|u - w^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} + \|u - u_h\|_{1,k,\mathcal{T}_n}\big]\big\}\|u - u_h\|_{0,\Omega} \quad \forall w^{\mathcal{T}_n} \in \mathbb{PW}_p(\mathcal{T}_n).\end{aligned} \tag{4.90}$$

After inserting next (4.79), (4.80), and (4.90) into (4.78), and dividing by $\|u - u_h\|_{0,\Omega}$, we have

$$\begin{aligned}\|u - u_h\|_{0,\Omega} \lesssim\ &(1 + \vartheta(k,h))(h|u - w^{\mathcal{T}_n}|_{2,\mathcal{T}_n} + h^{\frac{1}{2}}\|g - \Pi_p^{0,e}g\|_{0,\Gamma}) \\ &+ (\gamma_h + 1 + kh)(\varsigma(k,h) + \vartheta(k,h))\|u - w^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} \\ &+ \{(\gamma_h + 1 + kh)(\varsigma(k,h) + \vartheta(k,h)) + h(1 + d_\Omega k)\}\|u - u_h\|_{1,k,\mathcal{T}_n}.\end{aligned} \tag{4.91}$$

Step 5: Conclusion: We plug (4.91) in (4.72) and (4.72) in (4.58), obtaining

$$\begin{aligned}\|u - u_h\|_{1,k,\mathcal{T}_n} \lesssim\ &\frac{(kh + \gamma_h + 1)(1 + k(\varsigma(k,h) + \vartheta(k,h)))}{\alpha_h}\|u - w^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n} \\ &+ \left(\frac{\gamma_h + 1}{\alpha_h} + 1\right)\|u - u_I\|_{1,k,\mathcal{T}_n} + \frac{(k(1 + \vartheta(k,h)) + 1)h}{\alpha_h}|u - w^{\mathcal{T}_n}|_{2,\mathcal{T}_n} \\ &+ \frac{(k(1 + \vartheta(k,h)) + 1)h^{\frac{1}{2}}}{\alpha_h}\|g - \Pi_p^{0,e}g\|_{0,\Gamma} \\ &+ \frac{k\{(\gamma_h + 1 + kh)(\varsigma(k,h) + \vartheta(k,h)) + h(1 + d_\Omega k)\}}{\alpha_h}\|u - u_h\|_{1,k,\mathcal{T}_n},\end{aligned} \tag{4.92}$$

for all plane waves $w^{\mathcal{T}_n} \in \mathbb{PW}_p(\mathcal{T}_n)$, where $\varsigma(k,h)$ and $\vartheta(k,h)$ are given in (4.57).

Assuming that k^2h is sufficiently small, for instance, having set $\widetilde{c}$ the hidden constant in (4.92),

$$\widetilde{c}\,\frac{k\{(\gamma_h + 1 + kh)(\varsigma(k,h) + \vartheta(k,h)) + h(1 + d_\Omega k)\}}{\alpha_h} \leq \frac{1}{\nu}, \tag{4.93}$$

for some $\nu > 1$, we can bring the last term on the right-hand side of (4.92) to the left-hand side and obtain, further using (4.34), the desired bound (4.55). □

4.2.3 *A priori* error bounds

From Theorem 4.2.4, we deduce *a priori* error bounds in terms of h. The best approximation terms with respect to plane waves on the right-hand side of (4.55), namely $\|u - w^{\mathcal{T}_n}\|_{1,k,\mathcal{T}_n}$ and $|u - w^{\mathcal{T}_n}|_{2,\mathcal{T}_n}$, can be estimated using Theorem 4.2.1. A bound for the third term, namely $\|g - \Pi_p^{0,\Gamma}g\|_{0,\Gamma}$, is given in the following proposition.

Proposition 4.2.5. *Let $\mathcal{T}_n$ satisfy the assumptions **(G1)**-**(G3)** in Section 2.3, and let $\{\mathbf{d}_\ell\}_{\ell=1,\ldots,p}$, $p = 2q + 1$, $q \in \mathbb{N}_{\geq 2}$, be a given set of plane wave directions fulfilling the assumption **(D1)** in Section 4.1.1. Assuming that h is sufficiently small, see* (4.94) *below, and given g defined on Γ with $g_e := g_{|_e} \in H^{s-\frac{1}{2}}(e)$ for all $e \in \mathcal{E}_n^B$ and for some $s \in \mathbb{R}_{\geq 1}$, we have*

$$\|g - \Pi_p^{0,\Gamma}g\|_{0,\Gamma} \lesssim e^{\left(\frac{7}{4} - \frac{3}{4}\rho_{\max}\right)\sigma(kh)}\left(1 + [\sigma(kh)]^{q+9}\right)h^{\zeta + \frac{1}{2}}\sum_{e \in \mathcal{E}_n^B}\|G\|_{\zeta+1,k,D_e},$$

where $\zeta := \min(q,s)$, $\Pi_p^{0,\Gamma}$ is defined in (4.25)*, the constant $\sigma > 1$ with $\sigma \approx 1$, and G and $\rho_{\max}$ are set in* (4.95) *and* (4.96) *below, respectively.*

Proof. Associated with every boundary edge $e \in \mathcal{E}_n^B$, we consider a domain D_e with C^∞-boundary and diameter $h_{D_e} = \sigma h$, where $\sigma > 1$ is such that $h_{D_e} \approx h$, and D_e satisfies

- $e \in \partial D_e$;
- there exist $\rho_{D_e} \in (0, \frac{1}{2}]$ and $0 < \rho_{0,D_e} \leq \rho_{D_e}$, such that the ball $B_{\rho_{D_e} h_{D_e}}$ is contained in D_e, and D_e is star-shaped with respect to $B_{\rho_{0,D_e} h_{D_e}}$;
- it holds that
$$k^2 \text{ is not a Dirichlet-Laplace eigenvalue in } D_e, \tag{4.94}$$
which means that k^2 fulfils the counterpart of the condition (4.16) on D_e.

A graphical example of D_e with smooth boundary is provided in Figure 4.4. The construction of such domains is based on convolution techniques, as done in [95].

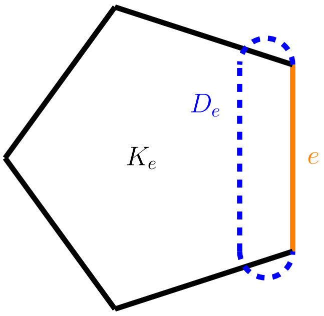


Figure 4.4: A possible construction for the domain D_e with smooth boundary, given a boundary edge $e \in \mathcal{E}_n^B \cap K_e$, for some polygon K_e belonging to a mesh $\mathcal{T}_n$.

Note that the requirement on σ guarantees a uniformly bounded overlapping of the collection of extended domains D_e associated with all the boundary edges $e \in \mathcal{E}_n^B$. More precisely, there exists $N \in \mathbb{N}$ such that, for all $\mathbf{x} \in \mathbb{R}^2$, $\mathbf{x}$ belongs to the intersection of at most N domains D_e, $e \in \mathcal{E}_n^B$. Owing to the smoothness of ∂D_e, $e \in \mathcal{E}_n^B$, it is possible to extend g_e to an $H^{s-\frac{1}{2}}(\partial D_e)$ function, following e.g. [97, Sect. 5.4], which we denote by $\widetilde{g_e}$. Note that $\widetilde{g_e}_{|_e} = g_e$.

Next, we consider the Helmholtz problem

$$\begin{cases} -\Delta G - k^2 G = 0 & \text{in } D_e \\ \qquad\quad G = \widetilde{g_e} & \text{on } \partial D_e. \end{cases} \tag{4.95}$$

Well-posedness follows from the fact that k^2 is not a Dirichlet-Laplace eigenvalue in D_e, see (4.94). Denoting by γ^{-1} the continuous right-inverse trace operator, see Lemma 2.1.2, and introducing $G_0 := G - \gamma^{-1}\widetilde{g_e}$, we can rewrite (4.95) as a Helmholtz problem with zero Dirichlet boundary conditions:

$$\begin{cases} -\Delta G_0 - k^2 G_0 = f_0 & \text{in } D_e \\ \qquad\quad G_0 = 0 & \text{on } \partial D_e, \end{cases}$$

with right-hand side $f_0 := (\Delta + k^2)(\gamma^{-1}\widetilde{g_e}) \in H^{s-2}(D_e)$.

Standard regularity theory [97, Sect. 6.3] implies $G_0 \in H^s(D_e)$ and therefore $G \in H^s(D_e)$. Then, by using the definition of the projector $\Pi_p^{0,\Gamma}$ in (4.25) on every edge $e \in \mathcal{E}_n^B$, we obtain

$$\|g - \Pi_p^{0,\Gamma} g\|_{0,e} \leq \|g_e - w^{D_e} - c_e\|_{0,e} \quad \forall w^{D_e} \in \mathbb{PW}_p(D_e), \forall c_e \in \mathbb{C}.$$

By applying the trace inequality (2.27), selecting $c_e = \frac{1}{|D_e|} \int_{D_e} (G - w^{D_e})\, \mathrm{d}x$, and using the Poincaré inequality (2.29), we get

$$\|g - \Pi_p^{0,\Gamma} g\|_{0,e} \leq C_T^{\frac{1}{2}} (C_P^2 + 1)^{\frac{1}{2}} h^{\frac{1}{2}} |G - w^{D_e}|_{1,D_e} \quad \forall w^{D_e} \in \mathbb{PW}_p(D_e).$$

For $s = 1$, this can be estimated by simply taking $w^{D_e} = 0$. Provided that $s \in \mathbb{R}_{>1}$, we can use Theorem 4.2.1 to get

$$\|g - \Pi_p^{0,\Gamma} g\|_{0,e} \lesssim e^{\left(\frac{7}{4} - \frac{3}{4}\rho_{D_e}\right)kh_{D_e}} \left(1 + (kh_{D_e})^{q+9}\right) h_{D_e}^{\zeta+\frac{1}{2}} \|G\|_{\zeta+1,k,D_e},$$

where $\zeta := \min(q, s-1)$, and the hidden constant is independent of k, h_{D_e}, and G. Defining

$$\rho_{\max} := \max_{D_e} \rho_{D_e}, \tag{4.96}$$

and summing over all edges $e \in \mathcal{E}_n^B$ give the desired result. □

The following theorem states the *a priori* error estimate associated with the method (4.4).

Theorem 4.2.6. *Let $u \in H^{s+1}(\Omega)$, $s \in \mathbb{R}_{\geq 1}$, be the exact solution to* (4.2). *Under the same assumptions as in Theorem 4.2.4 and Proposition 4.2.5, the following a priori error bound is valid:*

$$\begin{aligned}\|u - u_h\|_{1,k,\mathcal{T}_n} \lesssim{}& c_{PW}(kh) h^{\zeta_{1,2}} \left(\aleph_1(k,h) + \aleph_2(k,h)\right) \|u\|_{\zeta_{1,2}+1,k,\mathcal{T}_n} \\ &+ h^{\zeta_3+1}\, \aleph_2(k,h) e^{\left(\frac{7}{4} - \frac{3}{4}\rho_{\max}\right)\sigma(kh)} \left(1 + [\sigma(kh)]^{q+9}\right) \sum_{e \in \mathcal{E}_n^B} \|G\|_{\zeta_3+1,k,D_e},\end{aligned}$$

where $\zeta_{1,2} := \min(q, s)$, $\zeta_3 := \min(q, s-1)$, the constants $c_{PW}(kh)$, $\aleph_1(k,h)$, and $\aleph_2(k,h)$ are defined in (4.33) *and* (4.56)*, respectively, and where $\rho_{\max}$, σ, G, and D_e are constructed in Proposition 4.2.5.*

Proof. The assertion follows directly by combining the abstract error estimate (4.55) in Theorem 4.2.4 with best approximation estimates. More precisely, the first and second terms on the right-hand side of (4.55) can be estimated by means of Theorem 4.2.1. For the third term, Theorem 4.2.5 can be applied. □

So far, we have studied the method (4.4) at the theoretical level. In particular, we have proven that it is well-posed and we have derived convergence rates for the h-version of the method. However, as already mentioned in Remark 9, in its present version, the method is affected by strong ill-conditioning at the practical level. To this purpose, we do not show convergence plots here, but rather refer to the next Chapter 5, where the issue of ill-conditioning is discussed in full detail, a numerical strategy to mitigate the ill-conditioning and to render the method competitive is introduced, and a variety of numerical experiments is presented.

Chapter 5

Trefftz virtual element method for the Helmholtz problem: numerical aspects

In this chapter, we focus on numerical aspects of a nonconforming Trefftz VEM (ncTVEM) for the full Helmholtz problem (2.4) with $k > 0$:

$$\begin{cases} \text{find } u \in H^1(\Omega) \text{ such that} & \\ \qquad -\Delta u - k^2 u = 0 & \text{in } \Omega \\ \qquad\qquad\qquad u = g_D & \text{on } \Gamma_D \\ \qquad\quad \nabla u \cdot \mathbf{n}_\Omega = g_N & \text{on } \Gamma_N \\ \nabla u \cdot \mathbf{n}_\Omega + \mathrm{i}k\theta u = g_R & \text{on } \Gamma_R, \end{cases} \tag{5.1}$$

where $\theta \in \{-1, 1\}$, $g_D \in H^{\frac{1}{2}}(\Gamma_D)$, $g_N \in H^{-\frac{1}{2}}(\Gamma_N)$, $g_R \in H^{-\frac{1}{2}}(\Gamma_R)$, and $\Gamma = \partial\Omega$ with

$$\Gamma = \overline{\Gamma_D} \cup \overline{\Gamma_N} \cup \overline{\Gamma_R}, \quad \Gamma_D \cap \Gamma_N = \emptyset, \quad \Gamma_D \cap \Gamma_R = \emptyset, \quad \Gamma_N \cap \Gamma_R = \emptyset, \quad |\Gamma_R| > 0.$$

To this purpose, we firstly discuss how the method (4.4) for the Helmholtz problem (4.1) can be extended to the general case (5.1). This is the topic of Section 5.1. Then, in Section 5.2, details on the implementation of the method are given, and, in Section 5.3, a first set of numerical experiments is presented, underlining the problematic related to strong ill-conditioning. In Section 5.4, a numerical recipe to mitigate this ill-conditioning is introduced. Based on a series of numerical experiments, convergence of the new modified nonconforming Trefftz VEM is validated at the practical level. Furthermore, a comparison with UWVF/PWDG [71, 112, 118] and PWVEM [159] portrays the advantages of the new method. Finally, in Section 5.5, dispersion and dissipation properties are investigated for ncTVEM and PWVEM, and are compared to UWVF/PWDG and standard polynomial based finite element methods.

The material of this chapter, apart from Section 5.5, has been published in [145]. We highlight that the content of Section 5.5 is based on the work [158], which has been submitted for publication.

5.1 Extension of the nonconforming Trefftz VEM to the full Helmholtz problem

We now consider the general Helmholtz problem (5.1). In this section, we discuss the design of a corresponding nonconforming Trefftz VEM. Although not proven in this section, we expect that the analysis of Section 4.2 for the case $\Gamma_R = \Gamma$ can be carried over to the full problem (5.1), provided that k-explicit stability estimates are available there.

Similarly as in (4.4), given the variational formulation (2.7) corresponding to (5.1), namely

$$\begin{cases}\text{find } u \in V_{g_D} \text{ such that} \\ b_k(u,v) = F(v) \quad \forall v \in V_0,\end{cases}$$

where V_{g_D} and V_0 are defined in (2.8), and $b_k(\cdot,\cdot)$ and $F(\cdot)$ were introduced in (2.9), we are interested in constructing a nonconforming Trefftz VEM of the form

$$\begin{cases}\text{find } u_h \in V_{h,g_D}^{\Delta+k^2} \text{ such that} \\ b_{k,h}(u_h,v_h) = F_h(v_h) \quad \forall v_h \in V_{h,0}^{\Delta+k^2},\end{cases} \tag{5.2}$$

where the spaces $V_{h,g_D}^{\Delta+k^2}$ and $V_{h,0}^{\Delta+k^2}$ will be specified in Section 5.1.1, and the sesquilinear form $b_{k,h}(\cdot,\cdot)$ and the functional $F_h(\cdot)$ will be defined in Section 5.1.3. To this end, suitable projectors need to be introduced. This is done in Section 5.1.2.

5.1.1 Nonconforming Trefftz virtual element spaces

Here, we introduce the spaces $V_{h,g_D}^{\Delta+k^2}$ and $V_{h,0}^{\Delta+k^2}$ defining (5.2).

Let $\mathcal{T}_n$ be a mesh as described in Section 2.3 that is conforming with respect to the decomposition (2.3), i.e., for all boundary edges $e \in \mathcal{E}_n^B$, e is contained in only one amidst Γ_D, Γ_N, and Γ_R. In the sequel, we will use the following notation for the set of "Dirichlet, Neumann, and impedance (Robin)" edges:

$$\mathcal{E}_n^D = \{e \in \mathcal{E}_n^B \,:\, e \subseteq \Gamma_D\}, \quad \mathcal{E}_n^N = \{e \in \mathcal{E}_n^B \,:\, e \subseteq \Gamma_N\}, \quad \mathcal{E}_n^R = \{e \in \mathcal{E}_n^B \,:\, e \subseteq \Gamma_R\}.$$

Further, given the effective degree $q \in \mathbb{N}$, let $\{\mathbf{d}_\ell\}_{\ell\in\mathcal{J}}$, $\mathcal{J} := \{1,\dots,p\}$, $p = 2q+1$, be a set of pairwise different directions satisfying (**D1**) in Section 4.1.1. We use the same notation for the plane waves and plane wave spaces as in that section. In particular, we recall that w_ℓ^K denotes the bulk plane wave traveling along the direction $\mathbf{d}_\ell$, w_ℓ^e is the corresponding trace on a given edge e, $\mathbb{PW}_p(K)$ and $\mathbb{PW}_p(e)$ are the spaces spanned by the sets of bulk and edge plane waves, respectively, $\mathbb{PW}_p^c(e)$ is the set of plane wave traces after the filtering process (see Algorithm 1) with index set $\mathcal{J}_e$, and p_e is the dimension of $\mathbb{PW}_p^c(e)$.

Local Trefftz VE spaces. Now, we introduce, for any $K \in \mathcal{T}_n$, the *local* Trefftz VE space

$$\begin{aligned} V_{\Gamma_R}^{\Delta+k^2}(K) := \big\{ v_h \in H^1(K) \mid \Delta v_h + k^2 v_h = 0 \text{ in } K, \\ v_{h|_e} \in \mathbb{PW}_p^c(e) \quad \forall e \in \mathcal{E}^K \cap (\mathcal{E}_n^D \cup \mathcal{E}_n^N), \\ \gamma_I^K(v_h)_{|_e} \in \mathbb{PW}_p^c(e) \quad \forall e \in \mathcal{E}^K \setminus (\mathcal{E}_n^D \cup \mathcal{E}_n^N) \big\}, \end{aligned} \tag{5.3}$$

where we recall that $\gamma_I^K(v_h) = \nabla v_h \cdot \mathbf{n}_K + \mathrm{i}k\theta v_h$ is the impedance trace of v_h. In words, this space consists of all functions in $H^1(K)$ in the kernel of the Helmholtz operator, whose Dirichlet traces on boundary Dirichlet and Neumann edges, and impedance traces on interior and Robin edges are equal to traces of plane waves including constants. As above, $\mathbb{PW}_p(K) \subset V_{\Gamma_R}^{\Delta+k^2}(K)$, but $V_{\Gamma_R}^{\Delta+k^2}(K)$ also contains other functions whose explicit representation is not available in closed form, and which are henceforth *virtual*. We denote $p_K := \dim(V_{\Gamma_R}^{\Delta+k^2}(K)) = \sum_{e\in\mathcal{E}^K} p_e$. Moreover, we set $\mathcal{M}_K := \{1,\dots,n_K\}$.

For each edge $e_r \in \mathcal{E}^K$, $r \in \mathcal{M}_K$, we consider the same set of functionals as in (4.15), i.e.

$$\mathrm{dof}_{r,j}(v_h) := \frac{1}{h_{e_r}} \int_{e_r} v_h \overline{w_j^{e_r}} \,\mathrm{d}s \quad \forall r \in \mathcal{M}_K, \forall j \in \mathcal{J}_{e_r}. \tag{5.4}$$

Similarly as shown in Lemma 4.1.1 for the space $V^{\Delta+k^2}(K)$ in (4.14), this set constitutes a set of degrees of freedom also for functions in $V_{\Gamma_R}^{\Delta+k^2}(K)$, as proven in the forthcoming result.

Lemma 5.1.1. *Suppose that the assumption (**A1**) in Section 4.1.2 is satisfied. Then, the set of functionals in* (5.4) *defines a set of unisolvent degrees of freedom for the local space* $V_{\Gamma_R}^{\Delta+k^2}(K)$ *introduced in* (5.3).

Proof. If $\mathcal{E}^K \cap (\mathcal{E}_n^D \cup \mathcal{E}_n^N) = \emptyset$, the proof is identical to that of Lemma 4.1.1. Otherwise, we observe that, if $v_h \in V_{\Gamma_R}^{\Delta+k^2}(K)$ is such that all the associated functionals in (5.4) are zero, then $v_{h|_e} = 0$ on each edge $e \in \mathcal{E}^K \cap (\mathcal{E}_n^D \cup \mathcal{E}_n^N)$. This follows from the fact that $v_{h|_e} \in \mathbb{PW}_p^c(e)$ for all $e \in \mathcal{E}^K \cap (\mathcal{E}_n^D \cup \mathcal{E}_n^N)$, and from the definition of the functionals in (5.4). This, combined with an integration by parts, leads to

$$|v_h|_{1,K}^2 - k^2\|v_h\|_{0,K}^2 - \mathrm{i}k\theta\|v_h\|_{0,\partial K\setminus(\Gamma_D\cup\Gamma_N)}^2 = \int_{\partial K\setminus(\Gamma_D\cup\Gamma_N)} v_h\overline{\gamma_I^K(v_h)}\,\mathrm{d}s = 0, \tag{5.5}$$

since $\gamma_I^K(v_h)_{|_e} \in \mathbb{PW}_p^c(e)$ for all $e \in \mathcal{E}^K \setminus (\mathcal{E}_n^D \cup \mathcal{E}_n^N)$. Taking the imaginary part on both sides in (5.5) gives $v_h = 0$ on $\partial K\setminus(\Gamma_D \cup \Gamma_N)$, and therefore $v_h = 0$ on ∂K. Next, recalling that v_h belongs to the kernel of the Helmholtz operator and k^2 is not a Dirichlet-Laplace eigenvalue due to (**A1**), we deduce $v_h = 0$ in K, which is the assertion. □

Having this, analogously to (4.18), the set of local canonical basis functions $\{\varphi_{s,\ell}\}_{s\in\mathcal{M}_K,\ell\in\mathcal{J}_{e_s}}$ associated with the set of degrees of freedom (5.4) is defined by duality, i.e.

$$\mathrm{dof}_{r,j}(\varphi_{s,\ell}) = \delta_{r,s}\delta_{j,\ell} \quad \forall r,s \in \mathcal{M}_K,\ \forall j \in \mathcal{J}_{e_r},\ \forall \ell \in \mathcal{J}_{e_s}, \tag{5.6}$$

where δ is the Kronecker delta.

Global Trefftz VE spaces. Next, we construct the global Trefftz VE space, assuming uniform p; the case when p may vary from element to element is discussed in Section 5.4.3 below. We pinpoint the global nonconforming Sobolev space associated with $\mathcal{T}_n$ incorporating a Dirichlet boundary datum $\widetilde{g} \in H^{\frac{1}{2}}(\Gamma_D)$ in a nonconforming fashion:

$$\begin{aligned} H_{\widetilde{g}}^{1,nc}(\mathcal{T}_n) := \{v \in H^1(\mathcal{T}_n) \,:\, &\int_e (v^+ - v^-)\,\overline{w^e}\,\mathrm{d}s = 0 \quad \forall w^e \in \mathbb{PW}_p^c(e),\ \forall e \in \mathcal{E}_n^I,\\ &\int_e (v - \widetilde{g})\overline{w^e}\,\mathrm{d}s = 0 \quad \forall w^e \in \mathbb{PW}_p^c(e),\ \forall e \in \mathcal{E}_n^D\}, \end{aligned} \tag{5.7}$$

where, on each internal edge $e \in \mathcal{E}_n^I$ with $e \subseteq \partial K^- \cap \partial K^+$ for some K^-, $K^+ \in \mathcal{T}_n$, the functions v^- and v^+ are the Dirichlet traces of v from K^- and K^+, respectively.

The *global* nonconforming Trefftz VE trial and test spaces are given by

$$V_{h,g_D}^{\Delta+k^2} = \{v_h \in H_{g_D}^{1,nc}(\mathcal{T}_n) : v_{h|_K} \in V_{\Gamma_R}^{\Delta+k^2}(K) \quad \forall K \in \mathcal{T}_n\} \tag{5.8}$$

and

$$V_{h,0}^{\Delta+k^2} = \{v_h \in H_0^{1,nc}(\mathcal{T}_n) : v_{h|_K} \in V_{\Gamma_R}^{\Delta+k^2}(K) \quad \forall K \in \mathcal{T}_n\}, \tag{5.9}$$

respectively. In both cases, the set of global degrees of freedom is given by

$$\frac{1}{h_e}\int_e v_h\overline{w_j^e}\,\mathrm{d}s \quad \forall e \in \mathcal{E}_n,\ \forall j \in \mathcal{J}_e.$$

Remark 13. Owing to the definition (5.7), the Dirichlet boundary conditions are imposed weakly via the definition of moments with respect to plane waves. At the computational level, one can approximate g_D by taking a sufficiently high-order Gauß-Lobatto interpolant.

5.1.2 Local projectors

We employ the same projectors as in Section 4.1.3, where we replace $V^{\Delta+k^2}(K)$ by $V_{\Gamma_R}^{\Delta+k^2}(K)$. More precisely, for every $K \in \mathcal{T}_n$, we have the projector

$$\begin{aligned} \Pi_p^K &: V_{\Gamma_R}^{\Delta+k^2}(K) \to \mathbb{PW}_p(K)\\ &a_k^K(\Pi_p^K u_h, w^K) = a_k^K(u_h, w^K) \quad \forall u_h \in V_{\Gamma_R}^{\Delta+k^2}(K),\ \forall w^K \in \mathbb{PW}_p(K), \end{aligned} \tag{5.10}$$

where $a_k^K(\cdot,\cdot)$ is defined in (4.21).

Furthermore, on any boundary edge $e \in \mathcal{E}_n^B$, denoting by $K_e \in \mathcal{T}_n$ the adjacent element of e, we have the $L^2(e)$ projector

$$\begin{aligned} &\Pi_p^{0,e} : V_{\Gamma_R}^{\Delta+k^2}(K_e)_{|e} \to \mathbb{PW}_p^c(e) \\ &\qquad \int_e (\Pi_p^{0,e} u_h)\overline{w^e}\,\mathrm{d}s = \int_e u_h \overline{w^e}\,\mathrm{d}s \quad \forall u_h \in V_{\Gamma_R}^{\Delta+k^2}(K_e),\, \forall w^e \in \mathbb{PW}_p^c(e). \end{aligned} \tag{5.11}$$

In the sequel, we will use the notation $\Pi_p^{0,\omega}$ to denote the L^2 projector onto the space $\prod_{e\in\omega}\mathbb{PW}_p^c(e)$ defined edgewise by (5.11), where ω is either Γ_R or Γ_N.

Both types of projectors are computable by using the degrees of freedom (5.4), see Section 4.1.3.

5.1.3 Discrete sesquilinear forms and right-hand side

Here, we specify a computable sesquilinear form $b_{k,h}(\cdot,\cdot)$ and a computable functional $F_h(\cdot)$ occurring in the method (5.2). Those forms are in fact modifications of those introduced in Section 4.1.4.

Construction of $b_{k,h}(\cdot,\cdot)$.

For the discrete sesquilinear form $b_{k,h}(\cdot,\cdot)$, we employ

$$b_{k,h}(u_h, v_h) := a_{k,h}(u_h, v_h) + \mathrm{i}k\theta \int_{\Gamma_R} (\Pi_p^{0,\Gamma_R} u_h)\overline{(\Pi_p^{0,\Gamma_R} v_h)}\,\mathrm{d}s \quad \forall u_h \in V_{h,g_D}^{\Delta+k^2},\, \forall v_h \in V_{h,0}^{\Delta+k^2}, \tag{5.12}$$

where $a_{k,h}(\cdot,\cdot)$ is given as in (4.27), i.e.

$$\begin{aligned} a_{k,h}(u_h, v_h) &= \sum_{K\in\mathcal{T}_n} a_{k,h}^K(u_h, v_h) \\ &= \sum_{K\in\mathcal{T}_n} a_k^K(\Pi_p^K u_h, \Pi_p^K v_h) + S_k^K\left((I-\Pi_p^K)u_h, (I-\Pi_p^K)v_h\right), \end{aligned} \tag{5.13}$$

with a proper computable sesquilinear form $S_k^K(\cdot,\cdot)$ mimicking $a_k^K(\cdot,\cdot)$. In order to guarantee well-posedness of the method, some conditions on the choice of $S_k^K(\cdot,\cdot)$ are needed, see Theorem 4.2.4. In Section 5.4.3, we discuss effects of the choice of the stabilization on the numerical performance of the method.

Construction of $F_h(\cdot)$.

Concerning the right-hand side $F_h(\cdot)$, the choice we make is

$$F_h(v_h) := \int_{\Gamma_N} g_N \overline{(\Pi_p^{0,\Gamma_N} v_h)}\,\mathrm{d}s + \int_{\Gamma_R} g_R \overline{(\Pi_p^{0,\Gamma_R} v_h)}\,\mathrm{d}s \quad v_h \in V_{h,0}^{\Delta+k^2}. \tag{5.14}$$

5.2 Details on the implementation

In this section, we give some details concerning the implementation of the method (5.2), involving in particular the computation of the two projectors Π_p^K and $\Pi_p^{0,e}$ in (5.10) and (5.11), respectively. Although the setting of the method (5.2) is rather different from the one of standard VEM, the same ideas and notation as in [36] can be employed. This section is the counterpart to Section 3.3.1 for the Laplace problem.

5.2.1 Assembly of the global system of linear equations

The global system of linear equations corresponding to the method (5.2) is assembled as in the standard nonconforming VEM [18, 143] and FEM [83]. For the sake of clarity, we firstly consider the case that $\Gamma_D = \emptyset$. The general case will be addressed in Section 5.2.5 below.

Given N_e the total number of edges of the mesh $\mathcal{T}_n$, let $\{\varphi_{\tilde{s},\tilde{\ell}}\}_{\tilde{s}=1,\ldots,N_e,\,\tilde{\ell}\in\mathcal{J}_{e_{\tilde{s}}}}$ be the set of canonical basis functions (5.6). In this section, we use the convention that the indices hooded by a tilde denote global indices, whereas those without stand for local ones.

Expanding u_h as $\sum_{\tilde{s}=1}^{N_e}\sum_{\tilde{\ell}=1}^{p_{e_{\tilde{s}}}} u_{\tilde{s},\tilde{\ell}}\varphi_{\tilde{s},\tilde{\ell}}$ and plugging this ansatz into (5.2) lead to

$$\begin{aligned}\sum_{\tilde{s}=1}^{N_e}\sum_{\tilde{\ell}=1}^{p_{e_{\tilde{s}}}} u_{\tilde{s},\tilde{\ell}}\left[a_{k,h}(\varphi_{\tilde{s},\tilde{\ell}},\varphi_{\tilde{r},\tilde{j}}) + \mathrm{i}k\theta\int_{\Gamma_R}(\Pi_p^{0,\Gamma_R}\varphi_{\tilde{s},\tilde{\ell}})\overline{(\Pi_p^{0,\Gamma_R}\varphi_{\tilde{r},\tilde{j}})}\,\mathrm{d}s\right]\\ = \int_{\Gamma_N} g_N\overline{(\Pi_p^{0,\Gamma_N}\varphi_{\tilde{r},\tilde{j}})}\,\mathrm{d}s + \int_{\Gamma_R} g_R\overline{(\Pi_p^{0,\Gamma_R}\varphi_{\tilde{r},\tilde{j}})}\,\mathrm{d}s \quad \forall\tilde{r}=1,\dots,N_e,\ \forall\tilde{j}=1,\dots,p_{e_{\tilde{r}}},\end{aligned} \tag{5.15}$$

where, with a slight abuse of notation, we relabeled by $1,\dots,p_{e_{\tilde{s}}}$ the indices in $\mathcal{J}_{e_{\tilde{s}}}$ that remain after the filtering process (see Algorithm 1); similarly for the ones in $\mathcal{J}_{e_{\tilde{r}}}$. Then, (5.15) can be represented as the linear system

$$(\boldsymbol{A}+\boldsymbol{R})\boldsymbol{u} = \boldsymbol{f}, \tag{5.16}$$

where $\boldsymbol{A},\boldsymbol{R}\in\mathbb{C}^{N_{\mathrm{dof}}\times N_{\mathrm{dof}}}$, $\boldsymbol{u}\in\mathbb{C}^{N_{\mathrm{dof}}}$, and $\boldsymbol{f}\in\mathbb{C}^{N_{\mathrm{dof}}}$, N_{dof} being the total number of global degrees of freedom, are matrices and vectors with entries defined by

$$\begin{aligned}\boldsymbol{A}_{(\tilde{r},\tilde{j}),(\tilde{s},\tilde{\ell})} &:= a_{k,h}(\varphi_{\tilde{s},\tilde{\ell}},\varphi_{\tilde{r},\tilde{j}}), & \boldsymbol{R}_{(\tilde{r},\tilde{j}),(\tilde{s},\tilde{\ell})} &:= \mathrm{i}k\theta\int_{\Gamma_R}(\Pi_p^{0,\Gamma_R}\varphi_{\tilde{s},\tilde{\ell}})\overline{(\Pi_p^{0,\Gamma_R}\varphi_{\tilde{r},\tilde{j}})}\,\mathrm{d}s,\\ \boldsymbol{u}_{(\tilde{s},\tilde{\ell})} &:= u_{\tilde{s},\tilde{\ell}}, & \boldsymbol{f}_{(\tilde{r},\tilde{j})} &:= \int_{\Gamma_N} g_N\overline{(\Pi_p^{0,\Gamma_N}\varphi_{\tilde{r},\tilde{j}})}\,\mathrm{d}s + \int_{\Gamma_R} g_R\overline{(\Pi_p^{0,\Gamma_R}\varphi_{\tilde{r},\tilde{j}})}\,\mathrm{d}s.\end{aligned}$$

Note that here the subindex $(\tilde{r},\tilde{j})$ is associated with the index $\sum_{\tilde{t}=1}^{\tilde{r}-1} p_{e_{\tilde{t}}} + \tilde{j}$. The computation of $\boldsymbol{A}$, $\boldsymbol{R}$, and $\boldsymbol{f}$ are described in the forthcoming Sections 5.2.2, 5.2.3, and 5.2.4, respectively.

5.2.2 Computation of the matrix $\boldsymbol{A}$

To start with, the global matrix $\boldsymbol{A}\in\mathbb{C}^{N_{\mathrm{dof}}\times N_{\mathrm{dof}}}$ with entries

$$\boldsymbol{A}_{(\tilde{r},\tilde{j}),(\tilde{s},\tilde{\ell})} = a_{k,h}(\varphi_{\tilde{s},\tilde{\ell}},\varphi_{\tilde{r},\tilde{j}}) \tag{5.17}$$

is assembled by means of the local matrices $\boldsymbol{A}^K\in\mathbb{C}^{p_K\times p_K}$ that are given via

$$\boldsymbol{A}^K_{(r,j),(s,\ell)} := a_k^K(\Pi_p^K\varphi_{s,\ell},\Pi_p^K\varphi_{r,j}) + S_k^K\left((I-\Pi_p^K)\varphi_{s,\ell},(I-\Pi_p^K)\varphi_{r,j}\right),$$

where $\{\varphi_{s,\ell}\}_{s\in\mathcal{M}_K,\,\ell\in\mathcal{J}_{e_s}}$ denotes the local basis of $V_{\Gamma_R}^{\Delta+k^2}(K)$. The computation of such local matrices is performed in various steps.

Computation of the bulk projector Π_p^K in (5.10)**.** Let $\varphi_{s,\ell}\in V_{\Gamma_R}^{\Delta+k^2}(K)$, $s\in\mathcal{M}_K$, $\ell\in\mathcal{J}_{e_s}$, be the canonical basis function. We write $\Pi_p^K\varphi_{s,\ell}\in\mathbb{PW}_p(K)$ as linear combination of plane waves

$$\Pi_p^K\varphi_{s,\ell} = \sum_{\zeta=1}^{p}\gamma_\zeta^{K(s,\ell)}w_\zeta^K.$$

Plugging this ansatz into (5.10) and testing with plane waves lead to the system of linear equations

$$\boldsymbol{G}^K\boldsymbol{\gamma}^{K(s,\ell)} = \boldsymbol{b}^{K(s,\ell)},$$

where $\boldsymbol{G}^K\in\mathbb{C}^{p\times p}$, $\boldsymbol{\gamma}^{K(s,\ell)}\in\mathbb{C}^p$, $\boldsymbol{b}^{K(s,\ell)}\in\mathbb{C}^p$, for all $s\in\mathcal{M}_K$ and $\ell\in\mathcal{J}_{e_s}$, are defined as

$$\boldsymbol{G}^K := \begin{bmatrix} a_k^K(w_1^K,w_1^K) & \cdots & a_k^K(w_p^K,w_1^K)\\ \vdots & \ddots & \vdots\\ a_k^K(w_1^K,w_p^K) & \cdots & a_k^K(w_p^K,w_p^K)\end{bmatrix},\ \boldsymbol{\gamma}^{K(s,\ell)} := \begin{bmatrix}\gamma_1^{K(s,\ell)}\\ \vdots\\ \gamma_p^{K(s,\ell)}\end{bmatrix},\ \boldsymbol{b}^{K(s,\ell)} := \begin{bmatrix} a_k^K(\varphi_{s,\ell},w_1^K)\\ \vdots\\ a_k^K(\varphi_{s,\ell},w_p^K)\end{bmatrix}.$$

Collecting columnwise the $\boldsymbol{b}^{K(s,\ell)}$ leads to a matrix $\boldsymbol{B}^K := \left[\boldsymbol{b}^{K(1,1)},\dots,\boldsymbol{b}^{K(n_K,p_{e_{n_K}})}\right]\in\mathbb{C}^{p\times p_K}$.

The matrix $\boldsymbol{\Pi}_\star^K$ representing the action of Π_p^K from $V_{\Gamma_R}^{\Delta+k^2}(K)$ into $\mathbb{PW}_p(K)$ is then given by

$$\boldsymbol{\Pi}_\star^K = (\boldsymbol{G}^K)^{-1}\boldsymbol{B}^K\in\mathbb{C}^{p\times p_K}. \tag{5.18}$$

We introduce next the matrix

$$\boldsymbol{D}^K := \begin{bmatrix} \mathrm{dof}_{1,1}(w_1^K) & \cdots & \mathrm{dof}_{1,1}(w_p^K) \\ \vdots & \ddots & \vdots \\ \mathrm{dof}_{n_K, p_{e_{n_K}}}(w_1^K) & \cdots & \mathrm{dof}_{n_K, p_{e_{n_K}}}(w_p^K) \end{bmatrix} \in \mathbb{C}^{p_K \times p}.$$

Then, as in [36], the matrix $\boldsymbol{\Pi}^K$ representing the composition of the embedding of $\mathbb{PW}_p(K)$ into $V_{\Gamma_R}^{\Delta+k^2}(K)$ after Π_p^K can be expressed as

$$\boldsymbol{\Pi}^K = \boldsymbol{D}^K (\boldsymbol{G}^K)^{-1} \boldsymbol{B}^K \in \mathbb{C}^{p_K \times p_K}. \tag{5.19}$$

Matrix representation of $\boldsymbol{A}^K \in \mathbb{C}^{p_K \times p_K}$. The local VE stiffness matrix $\boldsymbol{A}^K$ is given by

$$\boldsymbol{A}^K = \overline{(\boldsymbol{\Pi}_\star^K)}^T \boldsymbol{G}^K \boldsymbol{\Pi}_\star^K + \overline{(\boldsymbol{I}^K - \boldsymbol{\Pi}^K)}^T \boldsymbol{S}^K (\boldsymbol{I}^K - \boldsymbol{\Pi}^K), \tag{5.20}$$

where $\boldsymbol{I}^K$ denotes the identity matrix of size $p_K \times p_K$, and $\boldsymbol{S}^K$ is the matrix representation of the local stabilization forms $S_k^K(\cdot,\cdot)$; for a specific choice of the stabilization, we refer to Section 5.4.3 below. Further, note that by using (5.18), it holds

$$\overline{(\boldsymbol{\Pi}_\star^K)}^T \boldsymbol{G}^K \boldsymbol{\Pi}_\star^K = \overline{(\boldsymbol{B}^K)}^T \overline{(\boldsymbol{G}^K)}^{-T} \boldsymbol{B}^K.$$

Computation of the local matrices $\boldsymbol{G}^K$, $\boldsymbol{B}^K$, and $\boldsymbol{D}^K$

The matrices $\boldsymbol{G}^K$, $\boldsymbol{B}^K$, and $\boldsymbol{D}^K$ can actually be computed exactly without numerical integration, but rather by using the definition of the degrees of freedom in (5.4) and the formula, see also [110, 159],

$$\Phi(z) := \int_0^1 e^{zt} \mathrm{d}t = \begin{cases} \frac{e^z - 1}{z} & \text{if } z \neq 0 \\ 1 & \text{if } z = 0 \end{cases} \qquad \forall z \in \mathbb{C}. \tag{5.21}$$

Computation of $\boldsymbol{G}^K \in \mathbb{C}^{p \times p}$. Given $j, \ell \in \mathcal{J}$, we compute, by using an integration by parts and taking into account the definition of the bulk plane waves w_j^K and w_ℓ^K, respectively,

$$\begin{aligned} \boldsymbol{G}_{j,\ell}^K = a_k^K(w_\ell^K, w_j^K) &= \sum_{r=1}^{n_K} \int_{e_r} (\nabla w_\ell^K \cdot \mathbf{n}_{K|_{e_r}}) \overline{w_j^K} \,\mathrm{d}s \\ &= \mathrm{i}k \sum_{r=1}^{n_K} e^{\mathrm{i}k(\mathbf{d}_j - \mathbf{d}_\ell)\cdot \mathbf{x}_K} (\mathbf{d}_\ell \cdot \mathbf{n}_{K|_{e_r}}) \int_{e_r} e^{\mathrm{i}k(\mathbf{d}_\ell - \mathbf{d}_j)\cdot \mathbf{x}} \,\mathrm{d}s. \end{aligned}$$

The integral over the edges e_r, $r \in \mathcal{M}_K$, on the right-hand side can be computed by application of the transformation rule. In fact, denoting by $\boldsymbol{a}_r$ and $\boldsymbol{b}_r$ the endpoints of the edge e_r, we obtain

$$\begin{aligned} \int_{e_r} e^{\mathrm{i}k(\mathbf{d}_\ell - \mathbf{d}_j)\cdot \mathbf{x}} \,\mathrm{d}s &= h_{e_r} e^{\mathrm{i}k(\mathbf{d}_\ell - \mathbf{d}_j)\cdot \boldsymbol{a}_r} \int_0^1 e^{\mathrm{i}k(\mathbf{d}_\ell - \mathbf{d}_j)\cdot(\boldsymbol{b}_r - \boldsymbol{a}_r)t} \,\mathrm{d}t \\ &= h_{e_r} e^{\mathrm{i}k(\mathbf{d}_\ell - \mathbf{d}_j)\cdot \boldsymbol{a}_r} \Phi\left(\mathrm{i}k(\mathbf{d}_\ell - \mathbf{d}_j)\cdot(\boldsymbol{b}_r - \boldsymbol{a}_r)\right), \end{aligned} \tag{5.22}$$

where Φ is defined in (5.21).

Computation of $\boldsymbol{B}^K \in \mathbb{C}^{p \times p_K}$. Given $s \in \mathcal{M}_K$, $\ell \in \mathcal{J}_{e_s}$, $j \in \mathcal{J}$, an integration by parts, the definitions of the local canonical basis functions in (5.6), and the definition of the degrees of freedom in (5.4) yield

$$\begin{aligned} \boldsymbol{B}_{j,(s,\ell)}^K = a_k^K(\varphi_{s,\ell}, w_j^K) = \sum_{r=1}^{n_K} \int_{e_r} \varphi_{s,\ell} \overline{(\nabla w_j^K \cdot \mathbf{n}_{K|_{e_r}})} \,\mathrm{d}s &= -\mathrm{i}k \sum_{r=1}^{n_K} (\mathbf{d}_j \cdot \mathbf{n}_{K|_{e_r}}) \int_{e_r} \varphi_{s,\ell} \overline{w_j^K} \,\mathrm{d}s \\ &= -\mathrm{i}k(\mathbf{d}_j \cdot \mathbf{n}_{K|_{e_s}}) e^{-\mathrm{i}k\mathbf{d}_j \cdot (\mathbf{x}_{e_s} - \mathbf{x}_K)} \int_{e_s} \varphi_{s,\ell} \underbrace{\overline{e^{\mathrm{i}k\mathbf{d}_j \cdot (\mathbf{x} - \mathbf{x}_{e_s})}}}_{= w_t^{e_s}} \,\mathrm{d}s \\ &= -\mathrm{i}k(\mathbf{d}_j \cdot \mathbf{n}_{K|_{e_s}}) e^{-\mathrm{i}k\mathbf{d}_j \cdot (\mathbf{x}_{e_s} - \mathbf{x}_K)} h_{e_s} \delta_{t,\ell}, \end{aligned}$$

where $t \in \mathcal{J}_{e_s}$ is the local index such that $w_t^{e_s} = e^{\mathrm{i}k\mathbf{d}_j \cdot (\mathbf{x} - \mathbf{x}_{e_s})}$ on e_s.

Computation of $\boldsymbol{D}^K \in \mathbb{C}^{p_K \times p}$. Given $r \in \mathcal{M}_K$, $j \in \mathcal{J}_{e_r}$, $\ell \in \mathcal{J}$, a direct computation gives

$$\boldsymbol{D}^K_{(r,j),\ell} = \mathrm{dof}_{r,j}(w^K_\ell) = \frac{1}{h_{e_r}} \int_{e_r} w^K_\ell \overline{w^{e_r}_j} \, \mathrm{d}s = \frac{1}{h_{e_r}} e^{\mathrm{i}k(\mathbf{d}_j \cdot \mathbf{x}_{e_r} - \mathbf{d}_\ell \cdot \mathbf{x}_K)} \int_{e_r} e^{\mathrm{i}k(\mathbf{d}_\ell - \mathbf{d}_j)\cdot \mathbf{x}} \, \mathrm{d}s.$$

The last term on the right-hand side can be computed as in (5.22).

5.2.3 Computation of the Robin boundary matrix $\boldsymbol{R}$

Recall that the Robin boundary matrix $\boldsymbol{R}$ is defined by

$$\boldsymbol{R}_{(\tilde{r},\tilde{j}),(\tilde{s},\tilde{\ell})} = \mathrm{i}k\theta \sum_{e \in \mathcal{E}^R_n} \int_e (\Pi^{0,e}_p \varphi_{\tilde{s},\tilde{\ell}}) \overline{(\Pi^{0,e}_p \varphi_{\tilde{r},\tilde{j}})} \, \mathrm{d}s. \tag{5.23}$$

The global matrix $\boldsymbol{R}$ is again assembled by means of the local matrices $\boldsymbol{R}^e \in \mathbb{C}^{p_e \times p_e}$ given by

$$\boldsymbol{R}^e_{(r,j),(s,\ell)} = \mathrm{i}k\theta \int_e (\Pi^{0,e}_p \varphi_{s,\ell}) \overline{(\Pi^{0,e}_p \varphi_{r,j})} \, \mathrm{d}s,$$

where $\{\varphi_{s,\ell}\}_{s \in \mathcal{M}_K, \ell \in \mathcal{J}_{e_s}}$ denotes the local basis of $V^{\Delta+k^2}_{\Gamma_R}(K)$, with K such that $e \subset \partial K \cap \Gamma_R$.

Let $e \in \mathcal{E}^R_n$ be a fixed boundary edge in $\mathcal{E}^R_n$ with local index $z \in \mathcal{M}_K$, where $K \in \mathcal{T}_n$ is the unique polygon with $e = \partial K \cap \Gamma_R$.

Computation of the edge projector $\Pi^{0,e}_p$ in (5.11). Let $\varphi_{z,\ell} \in V^{\Delta+k^2}_{\Gamma_R}(K)$, $\ell \in \mathcal{J}_e$, be a local canonical basis function. We firstly expand $\Pi^{0,e}_p \varphi_{z,\ell} \in \mathbb{PW}^c_p(e)$ in terms of the edge plane wave traces

$$\Pi^{0,e}_p \varphi_{z,\ell} = \sum_{\eta=1}^{p_e} \beta^{e(\ell)}_\eta w^e_\eta.$$

Inserting this ansatz into (5.11) and testing with edge plane waves lead to the linear system

$$\boldsymbol{G}^e_0 \boldsymbol{\beta}^{e(\ell)} = \boldsymbol{b}^{e(\ell)}_0.$$

Here, $\boldsymbol{G}^e_0 \in \mathbb{C}^{p_e \times p_e}$, $\boldsymbol{\beta}^{e(\ell)} \in \mathbb{C}^{p_e}$, $\boldsymbol{b}^{e(\ell)}_0 \in \mathbb{C}^{p_e}$, for all $\ell \in \mathcal{J}_e$, are defined as

$$\boldsymbol{G}^{\boldsymbol{e}}_{\boldsymbol{0}} := \begin{bmatrix} (w^e_1, w^e_1)_{0,e} & \cdots & (w^e_{p_e}, w^e_1)_{0,e} \\ \vdots & \ddots & \vdots \\ (w^e_1, w^e_{p_e})_{0,e} & \cdots & (w^e_{p_e}, w^e_{p_e})_{0,e} \end{bmatrix}, \quad \boldsymbol{\beta}^{e(\ell)} := \begin{bmatrix} \beta^{e(\ell)}_1 \\ \vdots \\ \beta^{e(\ell)}_{p_e} \end{bmatrix}, \quad \boldsymbol{b}^{e(\ell)}_0 := \begin{bmatrix} (\varphi_{z,\ell}, w^e_1)_{0,e} \\ \vdots \\ (\varphi_{z,\ell}, w^e_{p_e})_{0,e} \end{bmatrix}, \tag{5.24}$$

where we recall that $(\cdot,\cdot)_e$ is the complex L^2 inner product over e. Note that in fact $\boldsymbol{G}^{\boldsymbol{e}}_{\boldsymbol{0}} \in \mathbb{R}^{p_e \times p_e}$, see (5.26) below. Moreover, such matrix is positive definite for all $K \in \mathcal{T}_n$, and thus also invertible. Nevertheless, it is worth to underline that in presence of small elements and of a large number of plane waves, such matrix may become singular in machine precision. This problem will be analyzed in Section 5.3 and addressed in Section 5.4.

Consequently, collecting the $\boldsymbol{b}^{e(\ell)}_0$ columnwise into a matrix $\boldsymbol{B}^e_0 \in \mathbb{C}^{p_e \times p_K}$, the matrix representation of $\Pi^{0,e}_p$ is given by

$$\boldsymbol{\Pi}^{0,e}_\star = (\boldsymbol{G}^e_0)^{-1} \boldsymbol{B}^e_0.$$

Matrix representation of $\boldsymbol{R}^e$. The local edge VE boundary mass matrix $\boldsymbol{R}^e$ has the form

$$\boldsymbol{R}^e = \overline{\boldsymbol{\Pi}^{0,e}_\star}^T \boldsymbol{G}^e_0 \boldsymbol{\Pi}^{0,e}_\star = \overline{\boldsymbol{B}^e_0}^T (\overline{\boldsymbol{G}^e_0})^{-T} \boldsymbol{B}^e_0. \tag{5.25}$$

Computation of the local matrices $\boldsymbol{G}^e_0$ and $\boldsymbol{B}^e_0$

The matrices $\boldsymbol{G}^e_0$ and $\boldsymbol{B}^e_0$ can be computed exactly using the formula (5.21).

Computation of $\boldsymbol{G}_0^e \in \mathbb{R}^{p_e \times p_e}$. Given $j, \ell \in \mathcal{J}_e$ and denoting by $\boldsymbol{a}$ and $\boldsymbol{b}$ the endpoints of the edge e, it holds $(\boldsymbol{G}_0^e)_{j,j} = h_e$ and, if $j \neq \ell$,

$$(\boldsymbol{G}_0^e)_{j,\ell} = (w_\ell^e, w_j^e)_{0,e} = e^{\mathrm{i}k(\mathbf{d}_j - \mathbf{d}_\ell)\cdot \mathbf{x}_e} \int_e e^{\mathrm{i}k(\mathbf{d}_\ell - \mathbf{d}_j)\cdot \mathbf{x}} \,\mathrm{d}s = 2h_e \frac{\sin\left(k(\mathbf{d}_\ell - \mathbf{d}_j)\cdot \frac{\boldsymbol{b}-\boldsymbol{a}}{2}\right)}{k(\mathbf{d}_\ell - \mathbf{d}_j)\cdot(\boldsymbol{b}-\boldsymbol{a})} \in \mathbb{R}, \tag{5.26}$$

where we used (5.22) and the property $\sin(z) = \frac{1}{2\mathrm{i}}(e^{\mathrm{i}z} - e^{-\mathrm{i}z})$, $z \in \mathbb{C}$, in the last equality.

Computation of $\boldsymbol{B}_0^e \in \mathbb{C}^{p_e \times p_K}$. For all $j, \ell \in \mathcal{J}_e$, we have with the degrees of freedom in (5.4)

$$(\boldsymbol{B}_0^e)_{j,\ell} = (\varphi_{z,\ell}, w_j^e)_{0,e} = \int_e \varphi_{z,\ell} \overline{w_j^e} \,\mathrm{d}s = h_e \delta_{j,\ell}.$$

5.2.4 Computation of the right-hand side vector $\boldsymbol{f}$

We firstly recall that $\boldsymbol{f}$ is given by

$$\boldsymbol{f}_{(\tilde{r},\tilde{\jmath})} = \sum_{e \in \mathcal{E}_n^N} \int_e g_N \overline{(\Pi_p^{0,e} \varphi_{\tilde{r},\tilde{\jmath}})} \,\mathrm{d}s + \sum_{e \in \mathcal{E}_n^R} \int_e g_R \overline{(\Pi_p^{0,e} \varphi_{\tilde{r},\tilde{\jmath}})} \,\mathrm{d}s := \boldsymbol{f}^N_{(\tilde{r},\tilde{\jmath})} + \boldsymbol{f}^R_{(\tilde{r},\tilde{\jmath})}.$$

Once again, the global right-hand side $\boldsymbol{f}$ is assembled by means of the local vectors $\boldsymbol{f}^{N,e} \in \mathbb{C}^{p_e}$ and $\boldsymbol{f}^{R,e} \in \mathbb{C}^{p_e}$ that are defined as

$$\boldsymbol{f}^{N,e}_{(r,j)} = \int_e g_N \overline{(\Pi_p^{0,e} \varphi_{r,j})} \,\mathrm{d}s, \quad \boldsymbol{f}^{R,e}_{(r,j)} = \int_e g_R \overline{(\Pi_p^{0,e} \varphi_{r,j})} \,\mathrm{d}s,$$

where $\{\varphi_{s,\ell}\}_{s \in \mathcal{M}_K,\, \ell \in \mathcal{J}_{e_s}}$ denotes the local basis of $V_{\Gamma_R}^{\Delta+k^2}(K)$, with K such that either $e \subset \partial K \cap \Gamma_N$ or $e \subset \partial K \cap \Gamma_R$.

We only show the details concerning the computation of $\boldsymbol{f}^{N,e}$. The assembly of $\boldsymbol{f}^{R,e}$ is analogous. To this purpose, let $e \in \mathcal{E}_n^N$ be a fixed Neumann boundary edge with local index $z \in \mathcal{M}_K$, where $K \in \mathcal{T}_n$ is the unique polygon with $e = \partial K \cap \Gamma_N$. Then, for every $\ell \in \mathcal{J}_e$, denoting by $\mathbf{a}_z$ and $\mathbf{b}_z$ the endpoints of edge e, we have

$$\begin{aligned} \boldsymbol{f}_j^{N,e} &= \int_e g_N \overline{(\Pi_p^{0,e} \varphi_{z,j})} \,\mathrm{d}s = \sum_{\eta=1}^{p_e} \overline{\beta_\eta^{e(j)}} \int_e g_N \overline{w_\eta^e} \,\mathrm{d}s \\ &= \sum_{\eta=1}^{p_e} \overline{\beta_\eta^{e(j)}} h_e \int_0^1 g_N(\mathbf{a}_z + t(\mathbf{b}_z - \mathbf{a}_z)) e^{-\mathrm{i}k\mathbf{d}_j \cdot (\mathbf{a}_z + t(\mathbf{b} - \mathbf{a}_z) - \mathbf{x}_e)} \,\mathrm{d}t. \end{aligned} \tag{5.27}$$

The last integral can be approximated employing a Gauß-Lobatto quadrature formula. We remark that the computation of $\boldsymbol{f}$ is the only place where numerical quadrature may be required.

5.2.5 General case ($\Gamma_D \neq \emptyset$)

The general case with $\Gamma_D \neq \emptyset$ can be dealt with in a similar fashion. First of all, we implement the global matrices $\boldsymbol{A}$ and $\boldsymbol{R}$, and the right-hand side vector $\boldsymbol{f}$ as above. Then, in order to incorporate the Dirichlet boundary conditions, we additionally impose that the numerical solution u_h satisfies

$$\int_{e_\zeta} (u_h - g_D) \overline{w_j^e} \,\mathrm{d}s = 0 \quad \forall j = 1, \dots, p_{e_\zeta}, \forall e_\zeta \in \mathcal{E}_n^D,$$

which, using the expansion of u_h in terms of the canonical basis functions, leads to

$$\sum_{\tilde{s}=1}^{N_e} \sum_{\tilde{\ell}=1}^{p_{e_{\tilde{s}}}} u_{\tilde{s},\tilde{\ell}} \int_{e_\zeta} \varphi_{\tilde{s},\tilde{\ell}} \overline{w_j^{e_\zeta}} \,\mathrm{d}s = \int_{e_\zeta} g_D \overline{w_j^{e_\zeta}} \,\mathrm{d}s \quad \forall j = 1, \dots, p_{e_\zeta}, \forall e_\zeta \in \mathcal{E}_n^D.$$

With the canonical basis functions in (5.6) and the degrees of freedom in (5.4), it holds

$$u_{\zeta,j} = \frac{1}{h_{e_\zeta}} \int_{e_\zeta} g_D \overline{w_j^{e_\zeta}} \,\mathrm{d}s \quad \forall j = 1, \dots, p_{e_\zeta}, \forall e_\zeta \in \mathcal{E}_n^D. \tag{5.28}$$

This information is inserted in the linear system (5.16) by setting to zero all the entries in the rows of $\boldsymbol{A}$ corresponding to test functions associated with Dirichlet boundary edges, apart from the diagonal entry, which is set to one, and replacing the corresponding values of the vector $\boldsymbol{f}$ with the right-hand sides of (5.28).

5.3 The curse of ill-conditioning

In this section, we investigate the numerical performance of the method (5.2) for the problem (4.1), that is, we consider (5.1) with $\theta = 1$ and $\Gamma_R = \Gamma$. For these choices, (5.2) and (4.4) clearly coincide. As computational domain, we take $\Omega := (0,1)^2$. We will see that, as already mentioned in Remark 9, the present version of the method with the filtering process in Algorithm 1 does not deliver accurate results due to the strong ill-conditioning related to the plane wave bases. Therefore, in Section 5.4.1 below, we will propose a numerical recipe to mitigate such instabilities. All the tests were performed with `Matlab R2016b`.

For the numerical experiments in this section, the boundary datum g in (4.4) is cooked up in accordance with the analytical solutions

$$\begin{aligned} u_0(x,y) &:= \exp\left(\mathrm{i}kx\right), \\ u_1(x,y) &:= \exp\left(\mathrm{i}k\left(\cos\left(\frac{\pi}{4}\right)x + \sin\left(\frac{\pi}{4}\right)y\right)\right). \end{aligned} \tag{5.29}$$

The functions u_0 and u_1 are plane waves travelling in the directions $(1,0)$ and $(\frac{\pi}{4}, \frac{\pi}{4})$, respectively, see also Figure 4.1.

Since an exact representation of the numerical solution u_h is not available in closed form inside each element, it is not possible to compute the exact H^1 and L^2 discretization errors directly. Instead, analogously to what was done in Section 3.3, we compute the approximate relative errors

$$\frac{\|u - \Pi_p u_h\|_{1,k,\mathcal{T}_n}}{\|u\|_{1,k,\Omega}}, \qquad \frac{\|u - \Pi_p u_h\|_{0,\mathcal{T}_n}}{\|u\|_{0,\Omega}}, \tag{5.30}$$

where $\Pi_{p|K} = \Pi_p^K$, $K \in \mathcal{T}_n$, is the local projector given in (5.10). It is again possible to show that the errors (5.30) converge with the same rate as the exact relative H^1 and L^2 discretization errors.

We employ two different local stabilizations, which in matrix form read as follows:

- the identity stabilization

$$\boldsymbol{S}^K = \boldsymbol{I}^K, \tag{5.31}$$

where $\boldsymbol{I}^K \in \mathbb{C}^{p_K \times p_K}$ denotes the identity matrix;

- the *modified D-recipe* stabilization

$$\boldsymbol{S}^K_{(s,\ell),(r,j)} = a_k^K(\Pi_p^K \varphi_{r,j}, \Pi_p^K \varphi_{s,\ell})\, \delta_{r,s}\delta_{\ell,j}, \tag{5.32}$$

where δ denotes the Kronecker delta.

The former choice is the original VEM stabilization proposed in [31, 36], whereas the latter is a modification of the *diagonal recipe* (D-recipe), which was introduced in [41], and whose performance was investigated for high-order VEM and in presence of badly-shaped elements in [84, 142].

For the bulk plane wave space $\mathbb{PW}_p(K)$ in (4.7), we take $p = 2q+1$, $q \in \mathbb{N}$, plane waves with equidistributed directions $\{\mathbf{d}_\ell^{(0)}\}_{\ell=1}^p$ given by

$$\mathbf{d}_\ell^{(0)} = \left(\cos\left(\tfrac{2\pi}{p}(\ell-1)\right), \sin\left(\tfrac{2\pi}{p}(\ell-1)\right)\right). \tag{5.33}$$

We discretize the boundary value problem on sequences of regular Cartesian meshes and Voronoi-Lloyd meshes, see Figure 2.2, and investigate the h-version of the method for a fixed wave number $k = 10$ and different values of $q = 2$, 3, and 4. Note that in the case of u_0, since $u_0 \in \operatorname{span}\{w_\ell^K\}_{\ell=1}^p$ and owing to the consistency property (4.28) of the discrete bilinear

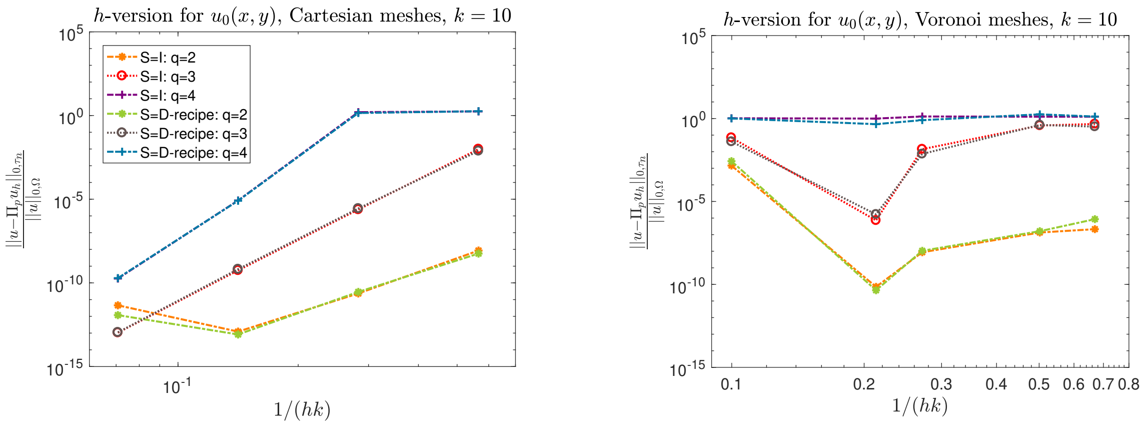


Figure 5.1: Approximate relative L^2 bulk errors for the h-version of the method for u_0 in (5.29) with $k=10$, $q=2$, 3, and 4, on Cartesian meshes (*left*) and Voronoi meshes (*right*) with directions $\{\mathbf{d}_\ell^{(0)}\}_{\ell=1}^{p}$ as in (5.33), and the identity and modified D-recipe stabilizations in (5.31) and (5.32), respectively.

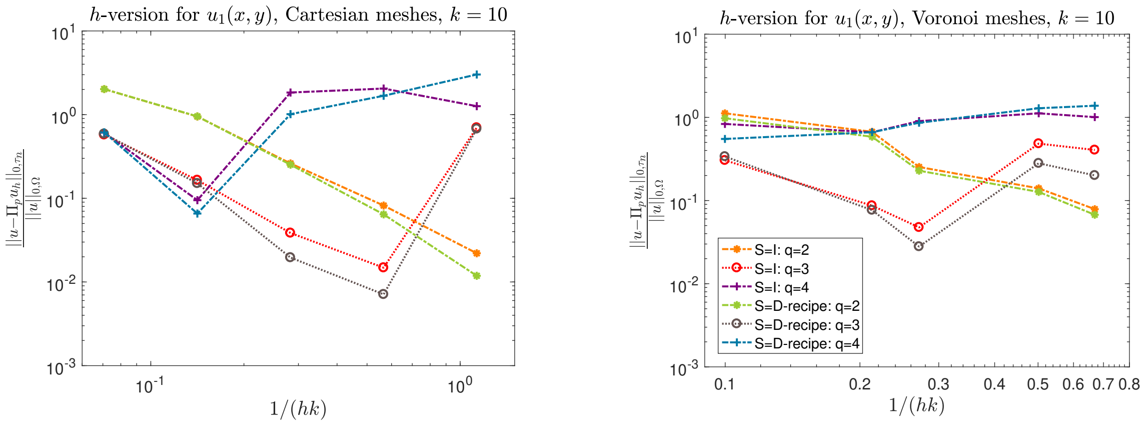


Figure 5.2: Approximate relative L^2 bulk errors for the h-version of the method for u_1 in (5.29) with $k=10$, $q=2$, 3, and 4, on Cartesian meshes (*left*) and Voronoi meshes (*right*) with directions $\{\mathbf{d}_\ell^{(0)}\}_{\ell=1}^{p}$ as in (5.33), and the identity and modified D-recipe stabilizations in (5.31) and (5.32), respectively.

form $b_{k,h}(\cdot,\cdot)$ in (5.12), the method should reproduce, up to machine precision, the exact solution. The approximate relative L^2 bulk errors defined in (5.30) are plotted in Figures 5.1 and 5.2.

In all the cases, we notice that the method becomes unstable after very few mesh refinements. This fact can be traced back to the computation of the Robin matrix $\boldsymbol{R}$ in (5.25) and of the right-hand side vector $\boldsymbol{f}$ in (5.27). Indeed, in both cases, we locally invert the edge plane wave mass matrices $\boldsymbol{G}_0^e$ in (5.26) on all boundary edges $e \in \mathcal{E}_n^B$. Such matrices are highly ill-conditioned; see Figure 5.3, where the condition number of the matrix $\boldsymbol{G}_0^e$ for the edge e with endpoints in $\boldsymbol{a}=[0,0]$ and $\boldsymbol{b}=[0,h]$ is depicted in dependence of h for the set of directions $\{\mathbf{d}_\ell^{(0)}\}_{\ell=1}^{p}$ in (5.33) and for different values of $q=2$, 3, and 4. In particular, one can also observe that the ill-conditioning grows together with the increase of the effective plane wave degree q and of the quantity $1/(hk)$.

This behavior is also reasonable from a heuristic point of view. More precisely, we firstly highlight that, in contrast to polynomials, plane waves cannot be directly scaled with characteristic lengths, such as the area of the element or the size of an edge. Indeed, replacing $\mathbf{x}$ in (4.6) or (4.8) by either $\frac{\mathbf{x}}{h_K}$ in the former or $\frac{\mathbf{x}}{h_e}$ in the latter case, results in bulk or edge plane waves with a different wave number. Thus, plane waves with fixed wave number and traveling in pairwise different directions become more and more linearly dependent for decreasing size of the elements and edges, respectively, and when increasing their numbers. Figure 5.4 visualizes this property.

Furthermore, it is imporant to note that the failure of the methods is really related to ill-conditiong, rather than to a poor approximation of the boundary integrals, since for exact solu-

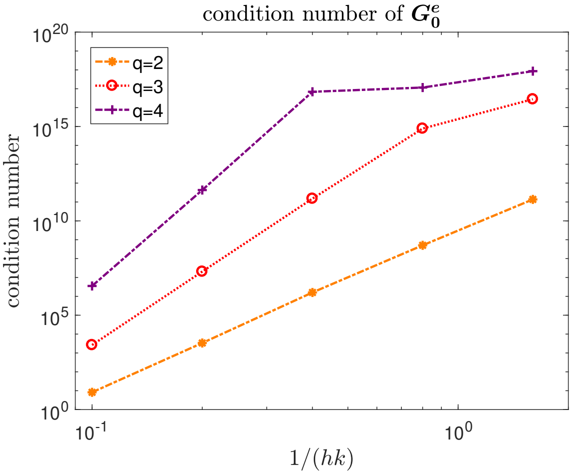


Figure 5.3: Condition number of G_0^e defined in (5.26) for the edge e with endpoints in $\boldsymbol{a} = [0, 0]$ and $\boldsymbol{b} = [0, h]$ in terms of hk for the set of directions $\{\mathbf{d}_\ell^{(0)}\}_{\ell=1}^p$ in (5.33) and different values of $q = 2$, 3, and 4.

tions u_0 and u_1 in (5.29), they are computed exactly. Out of curiosity, we also employed an overkill quadrature to handle the boundary conditions, and the results were practically the same.

Rebus sic stantibus, the present version of the method is not reliable. For this reason, we propose in Section 5.4 a numerical recipe to mitigate this ill-conditioning.

5.4 The modified nonconforming Trefftz VEM

In order to damp the strong ill-conditioning and to make the method reliable, we present a modified nonconforming Trefftz VEM in Section 5.4.1. Its implementation aspects are then described in Section 5.4.2. Finally, in Section 5.4.3, various numerical experiments and comparisons with other methods are shown.

5.4.1 A cure for the ill-conditioning

The main idea of the modification of the method lies in the substitution of the filtering process in Algorithm 1 by a modified version, which is explained in the forthcoming lines and summarized in the form of a pseudocode in Algorithm 2.

Starting from a set of $p = 2q + 1$, $q \in N$, plane waves with pairwise different propagation directions, we firstly compute, on each edge $e \in \mathcal{E}_n$, an eigendecomposition of the corresponding edge plane wave mass matrix $\boldsymbol{G}_\mathbf{0}^\mathbf{e}$:

$$\boldsymbol{G}_0^e \boldsymbol{Q}^e = \boldsymbol{Q}^e \boldsymbol{\Lambda}^e, \tag{5.34}$$

where $\boldsymbol{G}_0^e \in \mathbb{R}^{p\times p}$ is defined similarly as in (5.24), but with the difference that here the traces of *all* bulk plane waves in $\mathbb{PW}_p(K)$ are used, rather than only those after the filtering process in Algorithm 1, see also the definition of $\mathbb{PW}_p^c(e)$ in (4.11). Therefore, $\boldsymbol{G}_0^e$ can be singular (e.g. when two bulk plane waves have the same trace on e). Moreover, we do no longer require that the constants belong to the plane wave trace space. Note that this requirement was instrumental in the proof of the abstract error estimate in Section 4.2, but is not necessary in practice.

In the decomposition (5.34), the matrices $\boldsymbol{Q}^e \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\Lambda}^e \in \mathbb{R}^{p\times p}$ denote the eigenvector and eigenvalue matrices, respectively. Equivalently, the j-th column of $\boldsymbol{Q}^e$ contains the coefficients of the expansion of the new basis function $\widehat{w_j^e}$ with respect to the traces of the bulk plane waves w_ℓ^K, $\ell = 1, \ldots, p$, on e.

Next, we determine the positions of the eigenvalues on the diagonal of the matrix $\boldsymbol{\Lambda}^e$ which are zero or "close" to zero (up to a given tolerance σ), and we remove the corresponding columns of $\boldsymbol{Q}^e$. Doing so, we end up with a set of filtered orthogonalized plane waves. Having this, all the

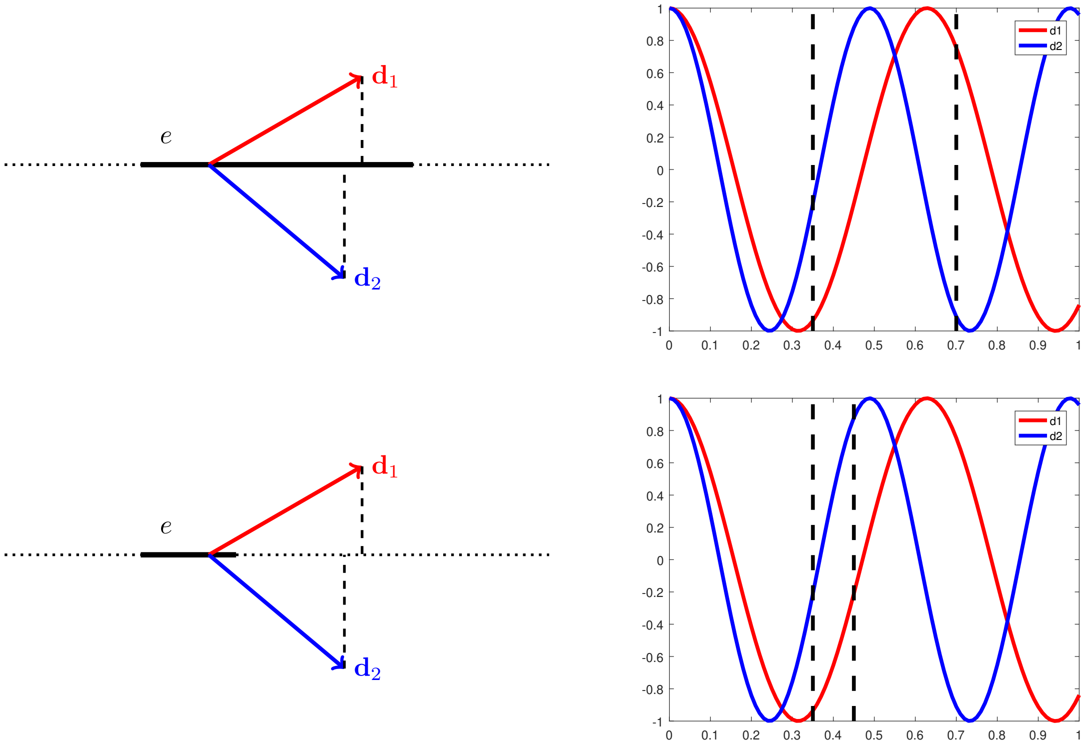


Figure 5.4: Real parts of the restrictions of two plane waves with fixed wave number and traveling along the directions $\mathbf{d}_1$ and $\mathbf{d}_2$, respectively, to given edges e. One can notice that, for smaller edges, the obtained plane wave traces are closer to being linearly dependent.

VE matrices discussed in Section 5.2 are computed employing the new filtered basis. Some details on this issue are given in Section 5.4.2.

As already expected from the discussion in Section 5.3, this modified filtering process is highly significant in presence of small edges and when employing a large number of initial plane wave basis functions. Moreover, it does not affect the rate of convergence of the method, as we will see in the numerical experiments.

Remark 14. We highlight that the influence of the choice of the parameter σ in Algorithm 2 on the convergence of the method will be discussed in Remark 16. Furthermore, we note that, from a practical point of view, due to the presence of eigenvalues/singular values close to zero, the computation of an orthogonal basis in `Matlab` via the eigendecomposition in step 1(b) in Algorithm 2 seems to be more robust than other procedures, such as SVD.

Remark 15. The strategy presented in Algorithm 2 seems to be natural in the nonconforming setting, also when employing other Trefftz functions, such as Fourier-Bessel functions, fundamental solutions, evanescent waves, etc., instead of plane waves. In fact, the basis functions are defined implicitly inside each elements by prescribing explicit conditions on the traces on each edge, and thus they can be modified edgewise without affecting their behavior on the other edges. This is not the case, for instance, in DG methods, where a modification of the basis functions implies a change in the behavior of such functions over *all* the edges. We deem that Algorithm 2 can be applied to other methods, whenever it is possible to split the basis functions into element bubbles and functions attached to single edges; this is for instance the case in all nonconforming virtual element methods.

5.4.2 Details on the implementation of the modified method

Here, we discuss some aspects of the implementation of the modified nonconforming Trefftz VEM.

Definition of the new degrees of freedom and canonical basis functions. Given $K \in \mathcal{T}_n$, let $\widehat{V}_{\Gamma_R}^{\Delta+k^2}(K)$ be defined similarly as $V_{\Gamma_R}^{\Delta+k^2}(K)$ in (5.3), where the only difference is that the

Algorithm 2 *Modified filtering process*

Let $\sigma > 0$ be a given tolerance.

1. For all the edges $e \in \mathcal{E}_n$:

   (a) Assemble the real-valued, symmetric, and possibly singular matrix $\boldsymbol{G}_0^e \in \mathbb{R}^{p\times p}$ given as in (5.26) by
   $$(\boldsymbol{G}_0^e)_{j,\ell} = (w_\ell^e, w_j^e)_{0,e} \quad \forall j,\ell = 1,\dots,p. \tag{5.35}$$

   (b) Starting from $\boldsymbol{G}_0^e$, compute the eigendecomposition (5.34):
   $$\boldsymbol{G}_0^e \boldsymbol{Q}^e = \boldsymbol{Q}^e \boldsymbol{\Lambda}^e,$$
   where $\boldsymbol{Q}^e \in \mathbb{R}^{p\times p}$ is a matrix whose columns are right-eigenvectors, and $\boldsymbol{\Lambda}^e \in \mathbb{R}^{p\times p}$ is a diagonal matrix containing the corresponding eigenvalues.

   (c) Determine the eigenvalues with (absolute) value smaller than the tolerance σ and remove the columns of $\boldsymbol{Q}^e$ corresponding to these eigenvalues. Denote the number of remaining columns of $\boldsymbol{Q}^e$ by $\widehat{p}_e \le p$. The remaining columns of $\boldsymbol{Q}^e$ are relabelled by $1,\dots,\widehat{p}_e$.

   (d) Define the new $L^2(e)$ orthogonal edge functions $\widehat{w}_\ell^e$, $\ell = 1,\dots,\widehat{p}_e$, in terms of the old ones w_r^e, $r = 1,\dots,p$, as
   $$\widehat{w}_\ell^e := \sum_{r=1}^{p} \boldsymbol{Q}_{r,\ell}^e\, w_r^e. \tag{5.36}$$

2. By using (5.36), build up the new local matrices $\widehat{\boldsymbol{G}}^K$, $\widehat{\boldsymbol{B}}^K$, and $\widehat{\boldsymbol{D}}^K$ for every element $K \in \mathcal{T}_n$, and assemble the global matrices $\widehat{\boldsymbol{A}}$, $\widehat{\boldsymbol{R}}$, and the global right-hand side vector $\widehat{\boldsymbol{f}}$.

space $\mathbb{PW}_p^c(e)$ in (5.3) is replaced by $\mathbb{PW}_p(K)_{|_e}$. In addition, given $e \in \mathcal{E}_n$, let $\{\widehat{w}_\ell^e\}_{\ell=1}^{\widehat{p}_e}$ be the set of the new (L^2 orthogonal) edge functions determined with Algorithm 2. The definitions of the global nonconforming Trefftz VE spaces in (5.8) and (5.9), and of the L^2 projector in (5.11) are changed accordingly.

Using (5.36), we modify the degrees of freedom and the definition of the canonical basis functions as follows. The new local degrees of freedom $\{\widehat{\mathrm{dof}}_{r,j}\}_{r=1,\dots,n_K,\, j=1,\dots,\widehat{p}_{e_r}}$ related to an element $K \in \mathcal{T}_n$ are given, for any $v_h \in \widehat{V}_{\Gamma_R}^{\Delta+k^2}(K)$, as

$$\widehat{\mathrm{dof}}_{r,j}(v_h) := \frac{1}{h_{e_r}} \int_{e_r} v_h \overline{\widehat{w}_j^{e_r}}\,\mathrm{d}s \quad \forall j = 1,\dots,\widehat{p}_{e_r}. \tag{5.37}$$

Further, the set of the new local canonical basis functions $\{\widehat{\varphi}_{s,\ell}\}_{s=1,\dots,n_K,\,\ell=1,\dots,\widehat{p}_{e_s}}$ associated with the local set of degrees of freedom (5.37) is the set of functions in the space $\widehat{V}_{\Gamma_R}^{\Delta+k^2}(K)$ with the property that

$$\widehat{\mathrm{dof}}_{r,j}(\widehat{\varphi}_{s,\ell}) = \delta_{r,s}\delta_{j,\ell}, \quad \forall r,s = 1,\dots,n_K,\, \forall j = 1,\dots,\widehat{p}_{e_r},\, \forall \ell = 1,\dots,\widehat{p}_{e_s}.$$

As usual, the sets of global degrees of freedom and of the canonical basis functions are obtained by coupling the local counterparts in a nonconforming fashion.

Next, we show how the new matrices $\widehat{\boldsymbol{G}}^K$, $\widehat{\boldsymbol{B}}^K$, $\widehat{\boldsymbol{D}}^K$, $\widehat{\boldsymbol{A}}$, and $\widehat{\boldsymbol{R}}$, and the new discrete right-hand side $\widehat{\boldsymbol{f}}$, counterparts of those described in Section 5.2, can be built starting from the original ones.

Computation of new local matrices.

- $\widehat{\boldsymbol{G}}^K$: This matrix coincides with $\boldsymbol{G}^K$ since it is computed via plane waves in the bulk.
- $\widehat{\boldsymbol{B}}^K$: For all $j = 1,\dots,p$, $s = 1,\dots,n_K$, $\ell = 1,\dots,\widehat{p}_{e_s}$, it holds
$$(\widehat{\boldsymbol{B}}^K)_{j,(s,\ell)} := a_k^K(\widehat{\varphi}_{s,\ell}, w_j^K) = -\mathrm{i}k(\mathbf{d}_j \cdot \mathbf{n}_{K|_{e_s}})e^{-\mathrm{i}k\mathbf{d}_j\cdot(\mathbf{x}_{e_s}-\mathbf{x}_K)} \int_{e_s} \widehat{\varphi}_{s,\ell}\, \overline{e^{\mathrm{i}k\mathbf{d}_j\cdot(\mathbf{x}-\mathbf{x}_{e_s})}}\,\mathrm{d}s,$$

which, after expressing the old edge function $w_j^{e_s}$ in terms of the novel ones

$$w_j^{e_s} = \sum_{\zeta=1}^{\widehat{p}_{e_s}} \overline{(\boldsymbol{Q}^e)^T_{\zeta,j}}\, \widehat{w}_\zeta^{e_s}, \tag{5.38}$$

and recalling that $\boldsymbol{Q}^e$ is the eigenvector matrix in (5.34), can be expressed as

$$(\widehat{\boldsymbol{B}}^K)_{j,(s,\ell)} = -\mathrm{i}k(\boldsymbol{Q}^e)^T_{\ell,j}(\mathbf{d}_j \cdot \mathbf{n}_{K|_{e_s}})e^{-\mathrm{i}k\mathbf{d}_j\cdot(\mathbf{x}_{e_s}-\mathbf{x}_K)}h_{e_s}.$$

- $\widehat{\boldsymbol{D}}^K$: Given $r \in \mathcal{M}_K$, $j = 1, \dots, \widehat{p}_{e_r}$, $\ell = 1, \dots, p$, a direct computation based on (5.38) gives

$$(\widehat{\boldsymbol{D}}^K)_{(r,j),\ell} := \widehat{\mathrm{dof}}_{r,j}(w_\ell^K) = \frac{1}{h_{e_r}}\int_{e_r} w_\ell^K \overline{\widehat{w}_j^{e_r}}\,\mathrm{d}s = \sum_{\zeta=1}^{p} \overline{\boldsymbol{Q}^e_{\zeta,j}}\frac{1}{h_{e_r}}\int_{e_r} w_\ell^K \overline{w_\zeta^{e_r}}\,\mathrm{d}s.$$

- $\widehat{\boldsymbol{A}}$: Starting from the local matrices

$$\widehat{\boldsymbol{A}}^K := \overline{\widehat{\boldsymbol{B}}^K}^T \overline{\widehat{\boldsymbol{G}}^K}^{-T} \widehat{\boldsymbol{B}}^K + \overline{(\widehat{\boldsymbol{I}}^K - \widehat{\boldsymbol{\Pi}}^K)}^T \widehat{\boldsymbol{S}}^K (\widehat{\boldsymbol{I}}^K - \widehat{\boldsymbol{\Pi}}^K),$$

see (5.20), $\widehat{\boldsymbol{A}}$ is assembled as in (5.17), where $\widehat{\boldsymbol{\Pi}}^K$ is defined similarly as in (5.19).

- $\widehat{\boldsymbol{R}}$: We need to compute, for all $\tilde{r}, \tilde{s} = 1, \dots, N_e$, $\tilde{j} = 1, \dots, \widehat{p}_{e_{\tilde{r}}}$, $\tilde{\ell} = 1, \dots, \widehat{p}_{e_{\tilde{s}}}$,

$$\widehat{\boldsymbol{R}}_{(\tilde{r},\tilde{j}),(\tilde{s},\tilde{\ell})} := \mathrm{i}k\theta \sum_{e\in\mathcal{E}_n^R} \int_e (\Pi_p^{0,e}\widehat{\varphi}_{\tilde{s},\tilde{\ell}})\overline{(\Pi_p^{0,e}\widehat{\varphi}_{\tilde{r},\tilde{j}})}\,\mathrm{d}s.$$

Given $e \in \mathcal{E}_n^R$, we only describe here the assembly of the matrix $\widehat{\boldsymbol{R}}^e \in \mathbb{C}^{p_e \times p_e}$, which takes into account the local contributions of the basis functions associated with e. Then, $\widehat{\boldsymbol{R}}$ is assembled as in (5.23). Given z the local index of e, for every $j, \ell = 1, \dots, \widehat{p}_e$, it holds

$$(\widehat{\boldsymbol{R}}^e)_{\ell,j} = \int_e (\Pi_p^{0,e}\widehat{\varphi}_{z,j})\overline{(\Pi_p^{0,e}\widehat{\varphi}_{z,\ell})}\,\mathrm{d}s. \tag{5.39}$$

By writing each $\Pi_p^{0,e}\widehat{\varphi}_{z,j}$, $j = 1, \dots, \widehat{p}_e$, as a linear combination of the $L^2(e)$ orthogonal plane waves $\widehat{w}_\theta^e$, $\theta = 1, \dots, \widehat{p}_e$, and inserting this into (5.39), one obtains

$$\widehat{\boldsymbol{R}}^e = \overline{(\widehat{\boldsymbol{B}}_0^e)}^T\, \overline{(\widehat{\boldsymbol{G}}_0^e)}^{-T}\, \widehat{\boldsymbol{B}}_0^e,$$

where

$$(\widehat{\boldsymbol{B}}_0^e)_{j,\ell} = (\widehat{\varphi}_{z,\ell}, \widehat{w}_j^e)_{0,e} = h_e\delta_{j,\ell} \quad \forall j, \ell = 1, \dots, \widehat{p}_e,$$

and

$$(\widehat{\boldsymbol{G}}_0^e)_{j,\ell} = (\widehat{w}_\ell^e, \widehat{w}_j^e)_{0,e} = \sum_{\zeta,\eta=1}^{\widehat{p}_e} \boldsymbol{Q}^e_{\eta,\ell}\overline{\boldsymbol{Q}^e_{\zeta,j}}\int_e w_\eta^e \overline{w_\zeta^e}\,\mathrm{d}s,$$

which can be represented as

$$\widehat{\boldsymbol{G}_0^e} = \overline{(\boldsymbol{Q}^e)}^T \boldsymbol{G}_0^e \boldsymbol{Q}^e$$

with $\boldsymbol{G}_0^e$ given in (5.35).

- $\widehat{\boldsymbol{f}} := \widehat{\boldsymbol{f}}^N + \widehat{\boldsymbol{f}}^R$: We restrict here ourselves to the computation of $\widehat{\boldsymbol{f}}^N$, which is given by

$$(\widehat{\boldsymbol{f}}^N)_{(\tilde{r},\tilde{j})} := \sum_{e\in\mathcal{E}_n^N}\int_e g_N\overline{(\Pi_p^{0,e}\widehat{\varphi}_{\tilde{r},\tilde{j}})}\,\mathrm{d}s \quad \forall \tilde{r} = 1, \dots, N_e,\ \forall \tilde{j} = 1, \dots, \widehat{p}_{e_{\tilde{r}}}.$$

The local vector $\widehat{\boldsymbol{f}}^{N,e} \in \mathbb{C}^{p_e}$ for a given $e \in \mathcal{E}_n^N$, with z denoting again the local index associated with e, has the form

$$(\widehat{\boldsymbol{f}}^{N,e})_\ell = \int_e g_N \overline{(\Pi_p^{0,e}\widehat{\varphi}_{z,\ell})}\,\mathrm{d}s = \sum_{\eta=1}^{\widehat{p}_e} \widehat{\beta}_\eta^{e(\ell)} \int_e g_N \overline{\widehat{w}_\eta^e}\,\mathrm{d}s = \sum_{\eta=1}^{\widehat{p}_e}\sum_{\zeta=1}^{p} \widehat{\beta}_\eta^{e(\ell)}\overline{(\boldsymbol{Q}^e)_{\zeta,\eta}} \int_e g_N \overline{w_\zeta^e}\,\mathrm{d}s,$$

where we recall that the matrix $\boldsymbol{Q}^e$ is the eigenvector matrix in (5.34).

- The Dirichlet boundary conditions are incorporated in the global system of linear equations as already shown in Section 5.2.5, by requiring that

$$\int_{e_\zeta} (u_h - g_D)\overline{\widehat{w}_j^{e_\zeta}}\,\mathrm{d}s = 0 \quad \forall j = 1,\ldots,\widehat{p}_{e_\zeta},\, \forall e_\zeta \in \mathcal{E}_n^D,$$

which leads to

$$u_{\zeta,j} = \frac{1}{h_{e_\zeta}} \int_{e_\zeta} g_D \overline{\widehat{w}_j^{e_\zeta}}\,\mathrm{d}s = \frac{1}{h_{e_\zeta}} \sum_{r=1}^{p} \overline{(\boldsymbol{Q}^{e_\zeta})_{r,j}} \int_{e_\zeta} g_D \overline{w_r^{e_\zeta}}\,\mathrm{d}s \quad \forall j = 1,\ldots,\widehat{p}_{e_\zeta},\, \forall e_\zeta \in \mathcal{E}_n^D.$$

Now, we investigate the numerical performance of this modified method.

5.4.3 Numerical results with the modified method

In this section, we discuss the h-, p-, and hp-versions of the modified method and assess the improvements in the numerical performance. We will see that the modified method is not only better conditioned, but also the number of degrees of freedom needed to achieve a given accuracy of the numerical approximation is significantly lower than in the original version in Section 5.2. Moreover, we compare the modified nonconforming Trefftz VEM with the PWVEM of [159] and with the more established UWVF/PWDG of [71, 112].

In all the numerical tests throughout this chapter, the tolerance σ in Algorithm 2 is set to 10^{-13}. Other choices and their influence on the method are discussed in Remark 16.

Additionally to the boundary value problems (5.1) with $\theta = 1$ and $\Gamma_R = \Gamma$ on $\Omega := (0,1)^2$ with known solutions u_0 and u_1 in (5.29), we consider problems of that form with exact solutions

$$\begin{aligned} u_2(x,y) &:= H_0^{(1)}(k|\mathbf{x}-\mathbf{x}_0|), \quad \mathbf{x}_0 = (-0.25, 0), \\ u_3(x,y) &:= J_\xi(kr)\cos(\xi\Theta), \quad \xi = \frac{2}{3}, \end{aligned} \tag{5.40}$$

where $H_0^{(1)}$ is the 0-th order Hankel functions of the first kind, J_ξ denotes the Bessel function of the first kind, and r and Θ are the polar coordinates of $(x, y-0.5)$, see [1, Chapters 9 and 10]. Note that the function u_2 is analytic over Ω, but u_3 has a singularity at $(0, 0.5)$; more precisely, $u_3 \in H^{\xi+1-\epsilon}(\Omega)$ for all $\epsilon > 0$ arbitrarily small, but $u_3 \notin H^{\xi+1}(\Omega)$. The real parts for the two test cases in (5.40) with $k = 20$ and the associated contour plots are depicted in Figure 5.5.

Furthermore, we consider experiments for a scattering problem in Section 5.4.3.

h-version

We firstly investigate the performance of the modified method for the *patch test* u_0 defined in (5.29) to check the consistency (4.28) and to validate the gain in robustness with respect to the original version, cf. Section 5.3. Let $\{\mathbf{d}_\ell^{(0)}\}_{\ell=1}^p$ be the set of directions given in (5.33). The numerical experiments are again performed on sequences of regular Cartesian meshes and Voronoi-Lloyd meshes, see Figure 2.2, for $k = 10$ and 20, and effective plane wave degrees $q = 4$ and 7. Recall that the number of used bulk plane waves is $p = 2q + 1$. Furthermore, we employ the *modified D-recipe stabilization* in (5.32). In Figure 5.6, the approximate relative H^1 bulk errors in (5.30) are plotted. We observe that the patch test is fulfilled for meshes with a moderately small mesh size. The plots indicate that the modified version is much more stable than the original one, see

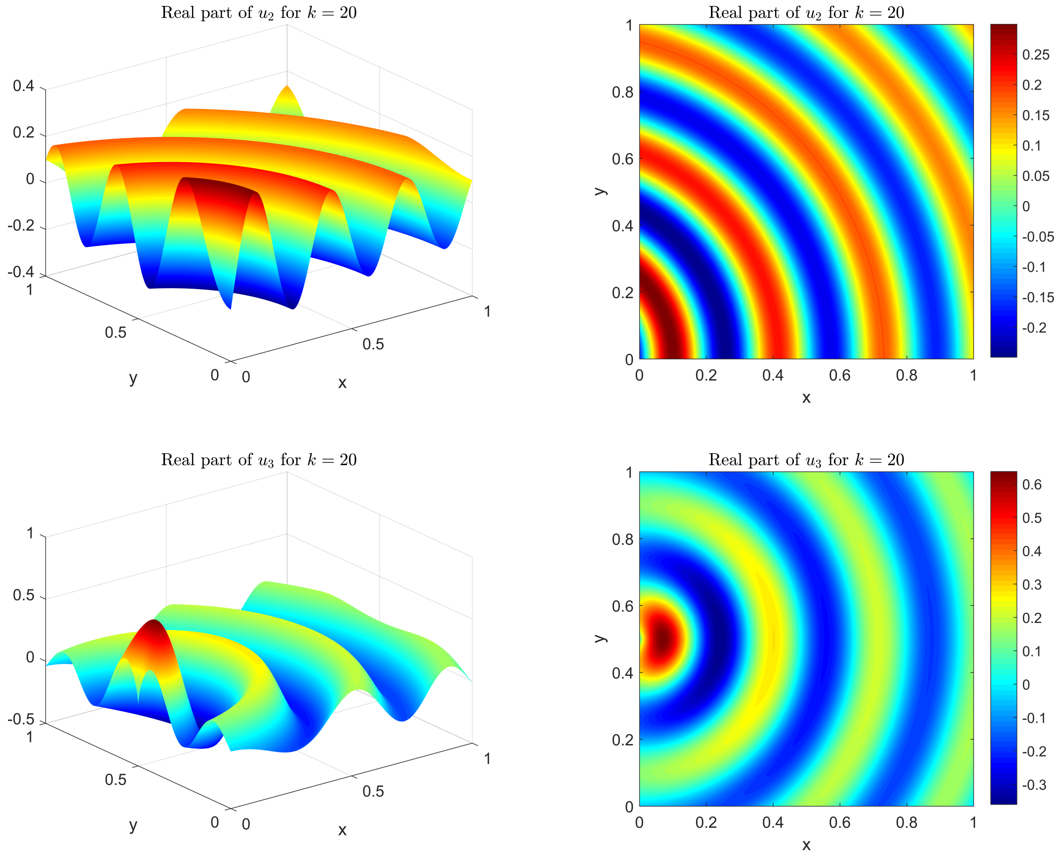


Figure 5.5: Real parts of the functions u_2 (*left*) and u_3 (*right*) defined in (5.40) for $k = 20$ and the corresponding contour plots.

Figure 5.1. Nevertheless, also this modified version is affected by ill-conditioning, which results in the increase of the errors for decreasing mesh size h, as typical of plane wave based methods.

As a second test, we investigate the h-version for the exact solution u_1 in (5.29) with $k = 10$, 20, and 40, and $q = 4$ and 7, employing the same choice of directions, meshes, and stabilizations as before. The numerical results are depicted for the Cartesian meshes in Figure 5.7 and Table 5.1 ($k = 20$, $q = 7$), and for the Voronoi meshes in Figure 5.8 and Table 5.2 ($k = 20$, $q = 7$). In all cases, the errors were computed accordingly with (5.30). In Tables 5.1 and 5.2, we further compare the number of degrees of freedom using the modified version of the method with the original one. The reduction of degrees of freedom in % is presented in the last column.

Here, we mention that the tests with exact solution u_2 give similar results to those for the smooth solution u_1 and are postponed to Section 5.4.4, where the modified nonconforming Trefftz VEM will be compared with PWVEM [159] and UWVF/PWDG [71, 112], respectively.

h	N_{dof}	rel. H^1 error	rate	rel. L^2 error	rate	N_{dof} orig.	red. (%)
1.414e+00	46	4.6885e-01	—	4.7153e-01	—	48	4.17
7.071e-01	120	1.3527e-01	1.793	1.3185e-01	1.838	144	16.67
3.535e-01	340	1.0540e-03	7.004	5.4861e-04	7.909	480	29.17
1.767e-01	1008	6.1594e-06	7.419	1.4439e-06	8.570	1728	41.67
8.838e-02	3264	4.2394e-08	7.183	4.4716e-09	8.335	6528	50.00
4.419e-02	10560	1.6544e-07	-1.964	7.3453e-08	-4.038	25344	58.33

Table 5.1: Relative errors for u_1 in (5.29) with $k = 20$, $q = 7$, and the directions $\{\mathbf{d}_\ell^{(0)}\}_{\ell=1}^{p}$ as in (5.33) on Cartesian meshes employing the modified method with the modified D-recipe stabilization (5.32).

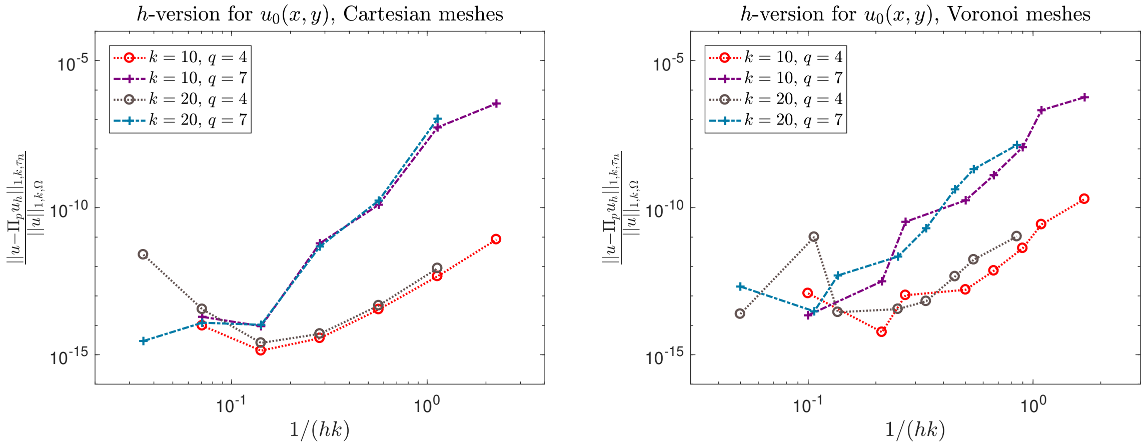


Figure 5.6: h-version of the method for u_0 in (5.29) with $k = 10$ and 20, and $q = 4$ and 7, with the sets of directions $\{\mathbf{d}_\ell^{(0)}\}_{\ell=1}^p$ as in (5.33) and the modified D-recipe stabilization (5.32) on Cartesian (*left*) and Voronoi meshes (*right*).

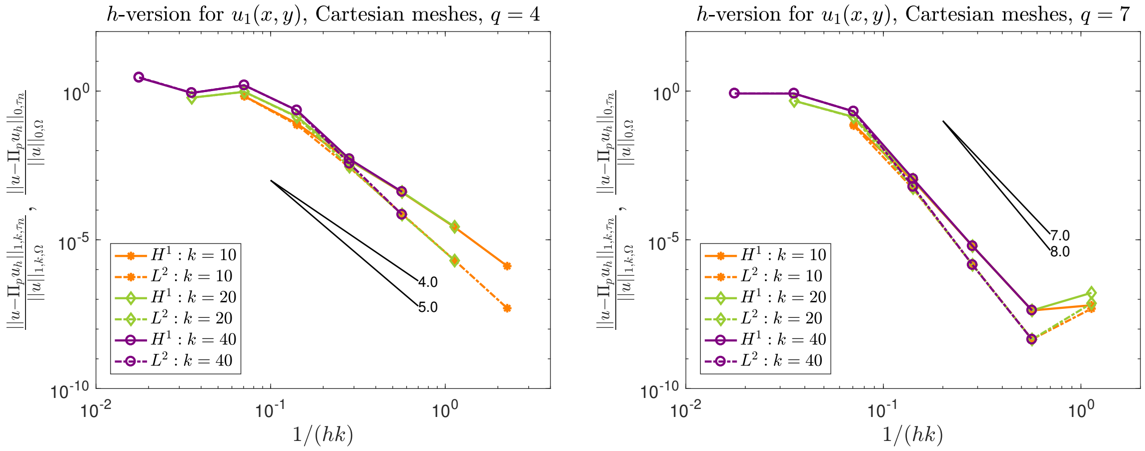


Figure 5.7: h-version of the modified method for u_1 in (5.29) with $k = 10$, 20, and 40, and $q = 4$ (*left*) and 7 (*right*), with the sets of directions $\{\mathbf{d}_\ell^{(0)}\}_{\ell=1}^p$ as in (5.33) and the modified D-recipe stabilization (5.32) on Cartesian meshes.

h	N_{dof}	rel. H^1 error	rel. L^2 error	N_{dof} orig.	red. (%)
1.001e+00	131	2.1704e-01	2.1440e-01	182	28.02
4.697e-01	224	7.5289e-02	7.4015e-02	359	37.60
3.688e-01	394	2.7605e-03	1.9061e-03	713	44.74
1.993e-01	695	2.4147e-04	1.0970e-04	1477	52.95
1.493e-01	1243	1.3955e-05	4.1303e-06	2960	58.01
1.111e-01	2206	1.7662e-06	3.9013e-07	5998	63.22
9.171e-02	4002	1.5165e-07	2.3002e-08	12092	66.90
5.896e-02	7282	2.1462e-08	3.0271e-09	24304	70.04

Table 5.2: Relative errors for u_1 in (5.29) with $k = 20$, $q = 7$, and the directions $\{\mathbf{d}_\ell^{(0)}\}_{\ell=1}^p$ as in (5.33) on Voronoi meshes employing the modified method with the modified D-recipe stabilization (5.32).

We observe from Figures 5.7 and 5.8, and Tables 5.1 and 5.2, that the approximate relative H^1 and L^2 discretization errors in (5.30) of the method approximately converge with rate 4 and 5 for $q = 4$, and 7 and 8 for $q = 7$, respectively. This is in agreement with the error estimate derived in Section 4.2.3, which established, for $h \to 0$ and analytic solutions, convergence rates of order q for the relative H^1 errors. Note that due to the fact that the Voronoi meshes are not nested, the slopes indicating the convergence order are not as straight as in the Cartesian case.

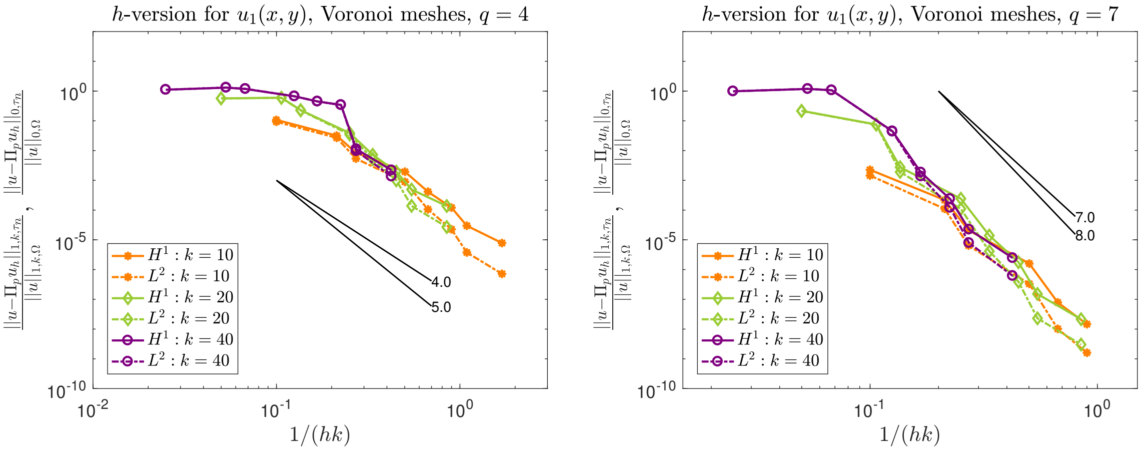


Figure 5.8: h-version of the modified method for u_1 in (5.29) with $k = 10$, 20, and 40, and $q = 4$ (*left*) and 7 (*right*), with the sets of directions $\{\mathbf{d}_\ell^{(0)}\}_{\ell=1}^p$ as in (5.33) and the modified D-recipe stabilization (5.32) on Voronoi meshes.

In addition, we notice that the number of degrees of freedom was reduced significantly by making use of the orthogonalization-and-filtering process described in Algorithm 2 in comparison to the original version of the method, which employs Algorithm 1.

Next, we employ the *identity stabilization* (5.31) and compare the performance with the modified D-recipe stabilization for u_1 using the same meshes and parameters as above. The results for the relative H^1 errors in (5.30) are shown in Table 5.3.

Cartesian				Voronoi			
h	N_{dof}	D-recipe	identity	h	N_{dof}	D-recipe	identity
1.414e+00	46	4.6885e-01	4.8651e-01	1.001e+00	131	2.1704e-01	2.3510e-01
7.071e-01	120	1.3527e-01	2.0525e-01	4.697e-01	224	7.5289e-02	9.3167e-02
3.535e-01	340	1.0540e-03	2.4615e-02	3.688e-01	394	2.7605e-03	2.4375e-02
1.767e-01	1008	6.1594e-06	1.7224e-03	1.993e-01	695	2.4147e-04	8.5729e-03
8.838e-02	3264	4.2394e-08	1.2786e-05	1.493e-01	1243	1.3955e-05	2.4687e-03
4.419e-02	10560	1.6544e-07	6.4752e-07	1.111e-01	2206	1.7662e-06	6.0640e-04

Table 5.3: Relative H^1 errors for u_1 in (5.29) with $k = 20$, $q = 7$, and the directions $\{\mathbf{d}_\ell^{(0)}\}_{\ell=1}^p$ as in (5.33) on Cartesian (*left*) and Voronoi (*right*) meshes employing the modified method with the D-recipe stabilization (5.32) and the identity stabilization (5.31).

Compared to the modified D-recipe stabilization, the method based on the identity stabilization behaves worse. Similar results are obtained for the relative L^2 errors in (5.30). This fact highlights that picking a "good" stabilization is an important issue in the design of VEM [41, 84, 142].

Thus, in the sequel, we will always consider the modified nonconforming Trefftz VEM endowed with the modified D-recipe stabilization (5.32).

As a last test in this section, we study the h-version of the method for the non-analytic solution u_3 in (5.40). Once again we perform the tests on the Cartesian meshes with $k = 10$, 20, and 40, and $q = 4$ and 7, in Figure 5.9. Similar results were obtained for the Voronoi meshes.

The observed convergence rate for the approximate H^1 bulk error in (5.30) is $\frac{2}{3}$ and that for the approximate L^2 bulk error is $\frac{5}{3}$. This corresponds to the expected convergence rates $\min\{s,q\}$ and $\min\{s,q\}+1$ for the H^1 and L^2 errors, respectively, where s is the regularity of the solution and q is the effective plane wave degree.

Remark 16. Here, we discuss and motivate the choice for the parameter σ in Algorithm 2, which so far has been set to 10^{-13}. In principle, it would have been more natural to take $\sigma = 10\,\text{eps}$, where eps denotes the machine epsilon. With this choice, it would be basically guaranteed that the span of the filtered orthogonalized edge plane wave functions coincides with the non-orthogonalized

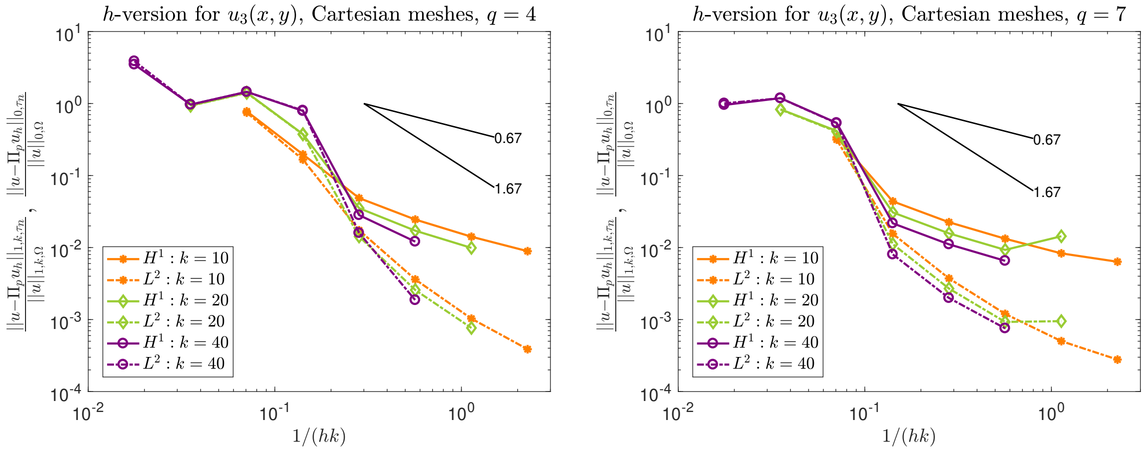


Figure 5.9: h-version of the method for u_3 in (5.29) with $k=10$, 20, and 40, and $q=4$ (*left*) and 7 (*right*), with the sets of directions $\{\mathbf{d}_\ell^{(0)}\}_{\ell=1}^p$ as in (5.33) and the modified D-recipe stabilization (5.32) on Cartesian meshes.

edge plane wave space, up to a negligible difference. However, we could observe from numerical experiments that with smaller choices of σ, such as 10^{-13}, it is possible to achieve the same accuracy as when employing $\sigma = 10\,\text{eps}$, but with less degrees of freedom, see Table 5.4, where we tested the h-version of the modified nonconforming Trefftz VEM with analytical solution u_2 in (5.40) on a sequence of Voronoi-Llyod meshes of the type in Figure 2.2 for the two above-mentioned choices of σ with $k=10$ and $q=7$.

	$\sigma = 10\,\text{eps}$		$\sigma = 10^{-13}$	
h	N_{dof}	rel. L^2 error	N_{dof}	rel. L^2 error
1.001346e+00	113	6.174135e-03	106	6.147714e-03
4.697545e-01	201	4.285982e-04	189	4.337061e-04
3.688297e-01	353	6.529610e-05	327	6.250524e-05
1.993180e-01	631	6.754430e-06	578	6.625276e-06
1.493758e-01	1139	1.572124e-07	1037	1.512503e-07
1.111597e-01	2053	6.369678e-08	1886	6.294611e-08
9.171171e-02	3745	2.514794e-08	3445	2.441118e-08

Table 5.4: h-version of the modified method for the analytical solution u_2 in (5.40), $k=10$, $q=7$, on Voronoi-Lloyd meshes for different choices of σ in Algorithm 2. The relative L^2 errors are computed accordingly with (5.30).

Application to an acoustic scattering problem. In this section, we consider the scattering of acoustic waves at sound-soft and sound-hard scatterers $\Omega_{Sc} \subset \mathbb{R}^2$, respectively, with polygonal boundary Γ_{Sc}, where the Sommerfeld radiation condition is approximated by an impedance boundary condition, see Section 2.2.1. The problems for the total field $u = u^I + u^S$, with incident field u^I and scattered field u^S, are

$$\text{(soft)} \begin{cases} -\Delta u - k^2 u = 0 & \text{in } \Omega \\ u = 0 & \text{on } \Gamma_{Sc} \\ \nabla u \cdot \mathbf{n}_\Omega - \mathrm{i}ku = g_R & \text{on } \Gamma_R, \end{cases} \qquad \text{(hard)} \begin{cases} -\Delta u - k^2 u = 0 & \text{in } \Omega \\ \nabla u \cdot \mathbf{n}_\Omega = 0 & \text{on } \Gamma_{Sc} \\ \nabla u \cdot \mathbf{n}_\Omega - \mathrm{i}ku = g_R & \text{on } \Gamma_R, \end{cases} \tag{5.41}$$

where $\Omega := \Omega_R \backslash \overline{\Omega_{Sc}}$, with Ω_R denoting the truncated domain with boundary Γ_R, and $g_R = \nabla u^I \cdot \mathbf{n}_\Omega - \mathrm{i}ku^I$ is the impedance trace of the incoming wave. For the numerical tests, we fix $\Omega = (-1,2) \times (0,3) \setminus [0,1] \times [1,2]$ and employ uniform Cartesian meshes, see Figure 5.10.

As incident fields, we consider u_0 and u_1 in (5.29), as well as the plane wave given by

$$u_4(x,y) := \exp\left(\mathrm{i}k\left(\cos\left(\frac{2\pi}{17}\right)x + \sin\left(\frac{2\pi}{17}\right)y\right)\right). \tag{5.42}$$

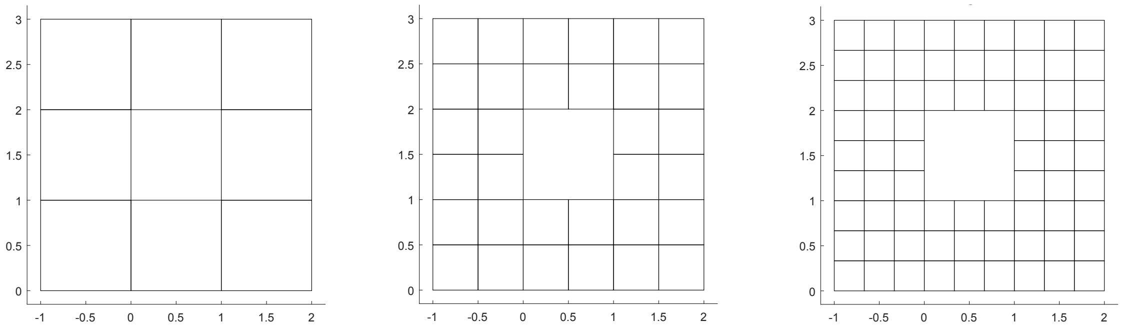


Figure 5.10: First three Cartesian meshes in the decomposition over the domain $\Omega = (-1, 2) \times (0, 3) \setminus [0, 1] \times [1, 2]$.

In Figures 5.11 and 5.12, the real parts of the computed total fields for the sound-hard and sound-soft cases, respectively, are plotted for the different incident fields with $k = 15$. As effective plane wave degree we choose $q = 10$ (namely $p = 21$ bulk plane waves).

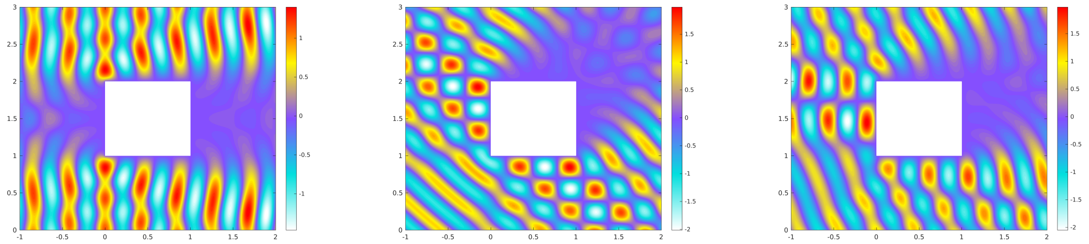


Figure 5.11: Real parts of the total fields for the sound-soft scattering employing as incident field the plane waves given by u_0 (*left*) and u_1 (*center*) in (5.29), and u_4 (*right*) in (5.42), with $k = 15$.

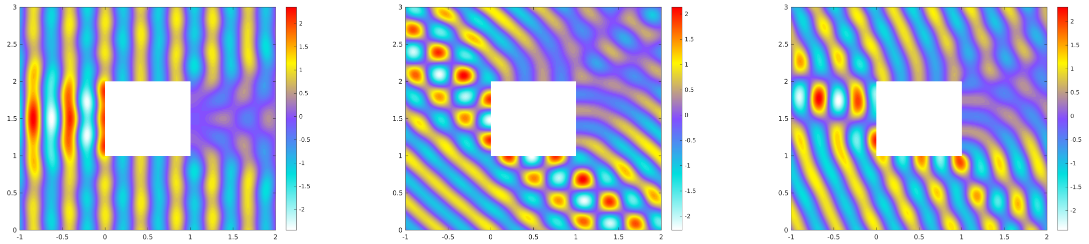


Figure 5.12: Real parts of the total fields for the sound-hard scattering employing as incident field the plane waves given by u_0 (*left*) and u_1 (*center*) in (5.29), and u_4 (*right*) in (5.42), with $k = 15$.

The relative errors are computed accordingly with (5.30), where, since an exact solution u is not known in closed form, u was chosen to be the discrete solution on a very fine mesh. In Figure 5.13, the obtained results are plotted. In both cases, the convergence rates are approximately 1.5 and 2.1 for the relative H^1 and L^2 errors, respectively.

p-version

We test numerically the p-version of the modified nonconforming Trefftz VEM, that is, we fix a mesh and increase the local effective degree q to achieve convergence. To this end, we consider the two meshes shown in Figure 5.14. Each of them consists of eight elements. The first one is a Voronoi-Lloyd mesh, and the second is a mesh whose elements are not star-shaped with respect to any ball. In the sequel, we will refer to these meshes as mesh (a) and mesh (b), respectively.

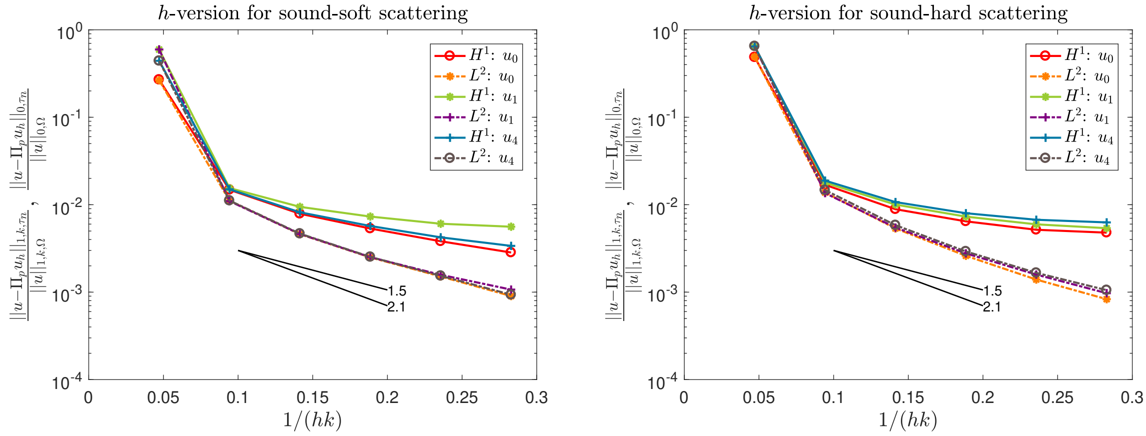


Figure 5.13: h-version of the modified method for the scattering problems (5.41) with $k = 15$ and $q = 10$. *Left:* sound-soft scattering; *right:* sound-hard scattering. The relative errors are computed accordingly with (5.30).

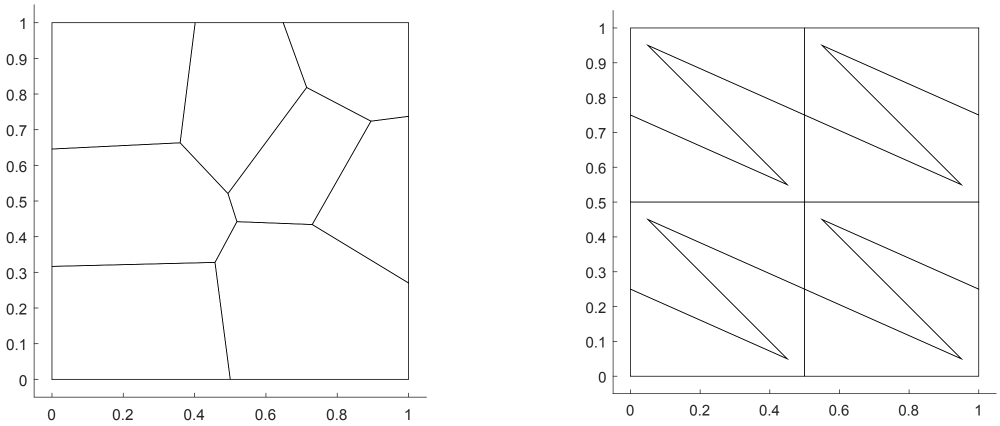


Figure 5.14: Different types of meshes made of eight elements; *left*: mesh (a), *right*: mesh (b).

As first test case, we take the one with analytical solution u_1 in (5.29), and we employ different values of $k = 10$, 20, and 40. The obtained numerical results are shown in Figure 5.15.

Analogously as for the h-version, the tests with analytical solution u_2 in (5.40) lead to similar results and are postponed to the forthcoming Section 5.4.4.

For both meshes, we observe that after a pre-asymptotic regime, the modified method is able to reach exponential convergence in terms of the effective degree q, before instability takes place, caused by the haunting ill-conditioning of the plane wave basis. The pre-asymptotic regime is much wider for higher wave numbers, which is typical of plane wave based methods. We underline that, despite the p-version of the nonconforming Trefftz VEM has not been investigated theoretically yet, the exponential decay of the error for analytic solutions is not surprising, cf. [39, 118, 166].

Next, we perform the same experiments on the non-analytic exact solution u_3 in (5.40). The corresponding plots are depicted in Figure 5.16. We notice that the convergence rate is not exponential any more, but rather algebraic. This is also an expected behavior of the p-version [39, 118, 166].

hp-version

Next, we numerically investigate the hp-version of the modified nonconforming Trefftz VEM, cf. Section 3.3.3 for the Laplace problem. In the framework of Trefftz methods for the Helmholtz equation, a full hp-analysis was carried out for UWVF/PWDG in [120], where exponential convergence in terms of the square root of the number of degrees of freedom was proven.

Here, we build approximation spaces with elementwise variable number of plane wave directions following the hp-approach for the Poisson problem introduced in [40]. To this end, we also need to take into account that interelement continuity has to be imposed in the nonconforming sense (5.7).

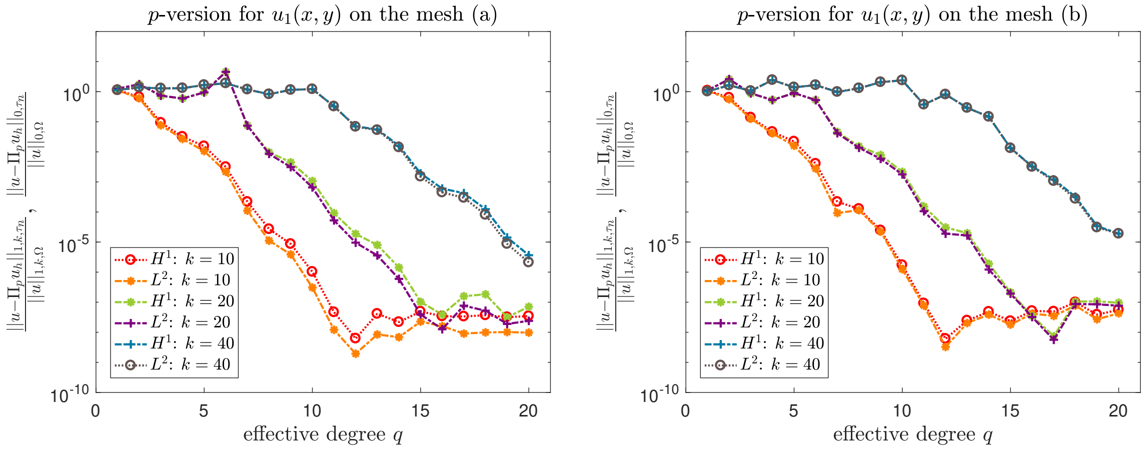


Figure 5.15: p-version of the modified method for u_1 in (5.29) on mesh (a) and (b) in Figure 5.14, *left* to *right*.

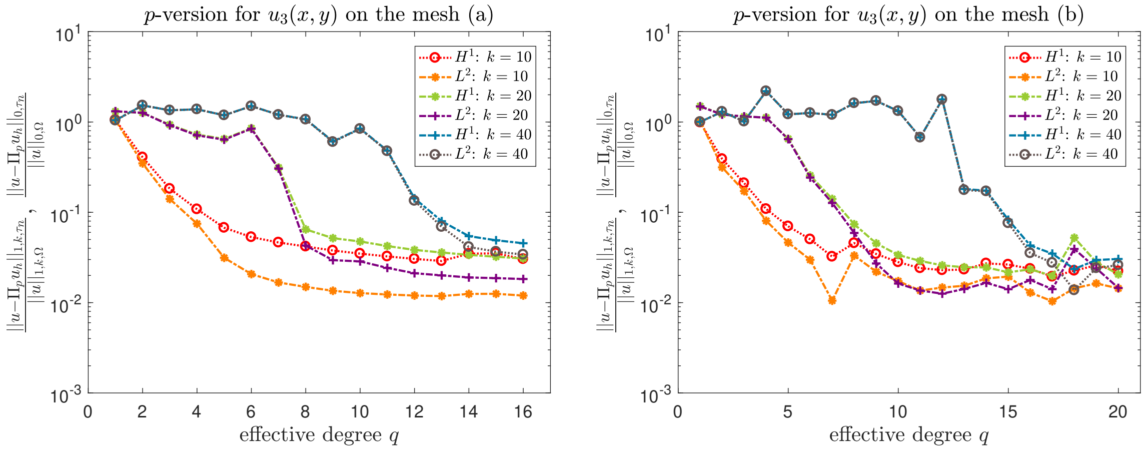


Figure 5.16: p-version of the modified method for u_3 in (5.40) on mesh (a) and (b) in Figure 5.14, *left* to *right*.

Let us assume that we aim at approximating the solution u_3 defined in (5.40) on $\Omega=(0,1)^2$; such function has a singularity at $\boldsymbol{\nu}=(0,0.5)$. We build a sequence of nested meshes that are refined towards $\boldsymbol{\nu}$ by proceeding as in Section 3.3.3. More precisely, we firstly set $\mathcal{T}_0=\{\Omega\}$. Next, for $n\in\mathbb{N}$, the mesh $\mathcal{T}_n$ is a polygonal mesh consisting of $n+1$ layers, where the 0-th layer $L_0:=L_{n,0}$ is the set of polygons abutting the singularity $\boldsymbol{\nu}$, whereas the ℓ-th layer is defined by induction as

$$L_\ell := L_{n,\ell} = \{K\in\mathcal{T}_n \,:\, \overline{K}\cap\overline{K_{\ell-1}}\neq\emptyset \text{ for some } K_{\ell-1}\in L_{\ell-1},\, K\not\subset\cup_{j=0}^{\ell-1}L_j\}.$$

Given the grading parameter $\mu\in(0,1)$, we require that

$$h_K \approx \begin{cases}\mu^n & \text{if } K\in L_0,\\ \frac{1-\mu}{\mu}\operatorname{dist}(K,\boldsymbol{\nu}) & \text{otherwise,}\end{cases} \tag{5.43}$$

for all $K\in\mathcal{T}_n$. Moreover, we increase the dimension of the local spaces as follows. We associate with each $K\in\mathcal{T}_n$ a number q_K, defined as

$$q_K=\ell+1 \quad \text{if } K\in L_{n,\ell},\ \ell=0,\dots,n-1, \tag{5.44}$$

and we build the local spaces $V_{\Gamma_R}^{\Delta+k^2}(K)$ in (5.3) by using Dirichlet/impedance traces that are edgewise in $\widetilde{\mathbb{PW}}_p^c(e)$, where the space $\widetilde{\mathbb{PW}}_p^c(e)$ is defined as follows.

Given $q_{max,n}=\max_{K\in\mathcal{T}_n} q_K$, we firstly consider the set of $p_{max,n}:=2q_{max,n}+1$ equidistributed directions $\{\widetilde{\mathbf{d}}_{\ell,n}\}_{\ell=1}^{p_{max,n}}$. On each element K, we pick a set of $2q_K+1$ directions obtained by removing $2(q_{max,n} - q_K)$ selected directions from the original set. Thus, elements abutting the singularity will have a small number of directions, which then increases linearly with the index of the layer. In order to select such directions to be removed, we order the set $\{\widetilde{\mathbf{d}}_{\ell,n}\}_{\ell=1}^{p_{max,n}}$ by picking first the directions with increasing odd indices and next those with even ones, see Figure 5.17. At this point, given the reordered set of directions $\{\widetilde{\widetilde{\mathbf{d}}}_{\ell,n}\}_{\ell=1}^{p_{max,n}}$, we remove the $2(q_{max,n} - q_K)$ directions having the largest indices. This procedure allows to build elementwise nested sets of directions with different cardinality. Then, nested spaces over each edge e of the mesh skeleton can be defined by fixing spaces of plane waves whose number of basis elements is given by the maximum of the local numbers q_K in (5.44) of the neighbouring elements:

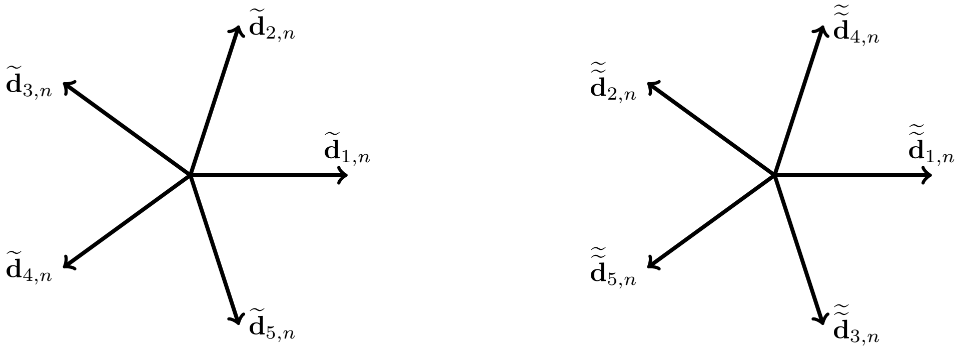

$\widetilde{\mathbf{d}}_{2,n}$
$\widetilde{\mathbf{d}}_{3,n}$
$\widetilde{\mathbf{d}}_{1,n}$
$\widetilde{\mathbf{d}}_{4,n}$
$\widetilde{\mathbf{d}}_{5,n}$
$\widetilde{\widetilde{\mathbf{d}}}_{4,n}$
$\widetilde{\widetilde{\mathbf{d}}}_{2,n}$
$\widetilde{\widetilde{\mathbf{d}}}_{1,n}$
$\widetilde{\widetilde{\mathbf{d}}}_{5,n}$
$\widetilde{\widetilde{\mathbf{d}}}_{3,n}$

Figure 5.17: *Left*: equidistributed set of directions $\{\widetilde{\mathbf{d}}_{\ell,n}\}_{\ell=1}^{p_{max,n}}$. *Right*: reordered set of directions $\{\widetilde{\widetilde{\mathbf{d}}}_{\ell,n}\}_{\ell=1}^{p_{max,n}}$. Firstly, one considers the directions with odd index and next those with even index.

$$\widetilde{\mathbb{PW}}_p^c(e) := \begin{cases} \operatorname{span}\left\{ e^{\mathrm{i}k\widetilde{\widetilde{\mathbf{d}}}_\ell(\mathbf{x}-\mathbf{x}_e)}{}_{|_e} \,:\, \ell = 1,\dots,2\max(q_{K_1},q_{K_2})+1 \right\} & \text{if } e \subset \mathcal{E}_n^I,\ e \subseteq \partial K_1 \cap \partial K_2 \\ \operatorname{span}\left\{ e^{\mathrm{i}k\widetilde{\widetilde{\mathbf{d}}}_\ell(\mathbf{x}-\mathbf{x}_e)}{}_{|_e} \,:\, \ell = 1,\dots,2q_K+1 \right\} & \text{if } e \subset \mathcal{E}_n^B,\ e \subseteq \partial K, \end{cases}$$

where K_1 and K_2, and K, denote the elements abutting edge e, if e is an interior edge and a boundary edge, respectively. This resembles the so-called *maximum rule* employed in *hp*-VEM [40].

A sequence of meshes satisfying the geometric refinement condition (5.43) towards $\boldsymbol{\nu}$, along with the distribution of effective degrees accordingly with (5.44), is depicted in Figure 5.18.

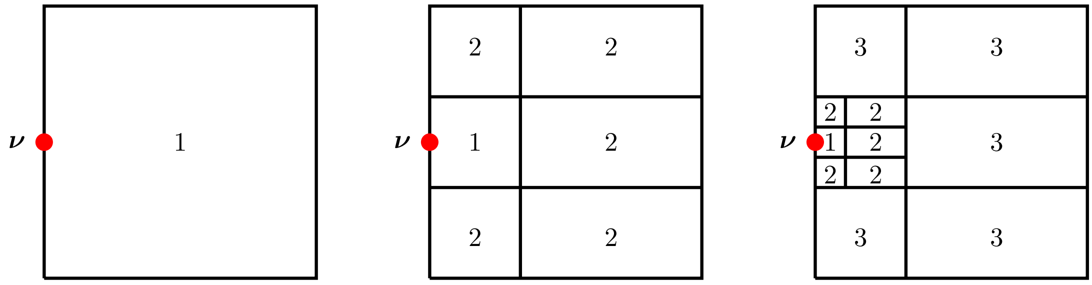


Figure 5.18: τ_0 (*left*), τ_1 (*center*), and τ_2 (*right*) of a sequence $\{\mathcal{T}_n\}_n$ of meshes graded toward $\boldsymbol{\nu}$ with grading parameter $\mu = 1/3$. The numbers inside the elements denote the effective degrees accordingly with (5.44).

We investigate the behavior of the modified version of the method presented in Section 5.4.1 for such a test case, where we select as wave numbers $k = 10$, 20, and 40, and as grading parameters $\mu = 1/2$ and $\mu = 1/3$, see (5.43). For the resulting error plot, we refer to Figure 5.23, where we also make a comparison with the plane wave discontinuous Galerkin method and which is discussed in more detail below. The approximate L^2 errors in (5.30) are plotted in terms of the quadratic root of the number of degrees of freedom.

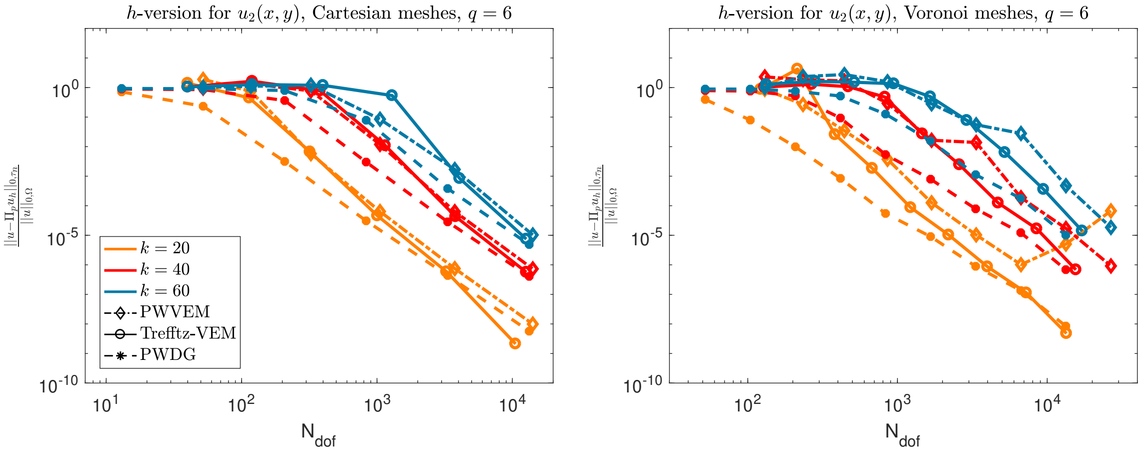
Focusing for the moment only on the nonconforming Trefftz VEM, we observe a decay of the error which is exponential in terms of the square root of the degrees of freedom instead of the cubic root as for standard hp-FEM [166] and hp-VEM [40]. This is typical of the Trefftz setting, see [120, 122] in the DG framework and [79, 143] in the VEM framework. Moreover, we want to highlight that after the pre-asymptotic regime, the relative errors decay extremely rapidly in terms of the number of degrees of freedom. This can be explained by the fact that, for smaller mesh sizes, more and more redundant plane wave directions are removed by the filtering process, compensating the increase in the number of edges, see also Section 5.4.1. The "paradox" here is that the errors of the method decrease exponentially, while the number of degrees of freedom seems to increase extremely slowly, especially in presence of high wave number. This behavior is really a peculiarity of Algorithm 2 and is denoted as *cliff effect.*

Remark 17. The strategy for designing approximation spaces discussed here is quite constructive. At the practical level, one can proceed alternatively as follows. Given an interior edge $e \in \mathcal{E}_n^I$ with the two adjacent elements $K_1, K_2 \in \mathcal{T}_n$, and associated plane wave bulk spaces $\mathbb{PW}_{p_1}(K)$ and $\mathbb{PW}_{p_2}(K)$, a proper edge plane wave space can be obtained by considering the restrictions of the $p_1 + p_2$ bulk plane waves to the edge and then applying the modified filtering process in Algorithm 2 to kick out redundancies. This approach also allows to handle the use of elementwise different basis functions, such as plane waves, Fourier-Bessel functions, evanescent waves, etc., in a very natural and simple fashion. For more details, we refer to Chapter 6.

5.4.4 Comparison of ncTVEM with PWVEM and UWVF/PWDG

Here, we compare the approximate relative L^2 errors in (5.30) of the modified nonconforming Trefftz VEM (ncTVEM) with those of PWVEM [159] and UWVF/PWDG [71, 112], whose structure will be shortly recalled in Section 5.5 below. Note that the definition of Π_p^K is the same for ncTVEM and PWVEM. For PWVEM, we took the stabilization proposed in [159]. Moreover, for PWDG, we chose the penalty parameters of the ultra weak variational formulation (UWVF) in [71].

h-version: To start with, we compare the h-versions of the methods on regular Cartesian meshes and on Voronoi meshes, both as shown in Figure 2.2, in terms of the number of degrees of freedom. Given a boundary value problem of the form (2.7) with $\Omega := (0,1)^2$, $\Gamma_R = \Gamma$, and exact solution u_2 in (5.40), we firstly choose $q = 6$ (which corresponds to $p = 13$) and $k = 20$, 40, and 60. Then, as a second test, we fix instead $k = 20$ and employ $q = 5$, 7, and 9. The results are shown in Figures 5.19 and 5.20, respectively.


h-version for $u_2(x,y)$, Cartesian meshes, $q = 6$

h-version for $u_2(x,y)$, Voronoi meshes, $q = 6$

Figure 5.19: Comparison of the h-version of ncTVEM with PWVEM and UWVF/PWDG for u_2 in (5.40), fixed $q = 6$, and $k = 20$, 40, and 60, on regular Cartesian (*left*) and Voronoi meshes (*right*). The legend is the same for both plots.

It can be noticed that, when comparing ncTVEM with UWVF/PWDG, we can approximately reach the same accuracy. For ncTVEM, the pre-asymptotic regime is broader, followed however by a "steeper" slope of the convergence rate. This broader pre-asymptotic area can be explained by the fact that, on coarse meshes, the removing procedure of Algorithm 2 is almost not performed,

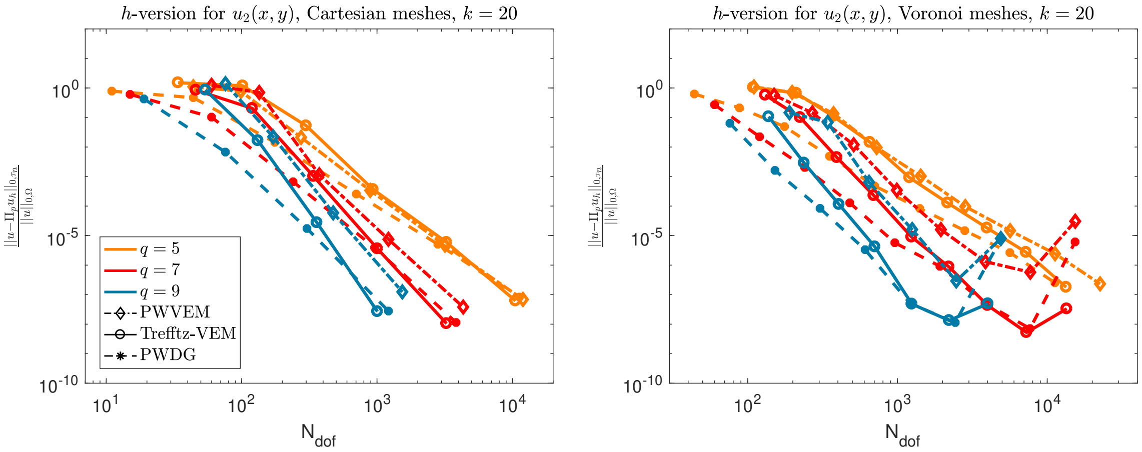


Figure 5.20: Comparison of the h-version of ncTVEM with PWVEM and UWVF/PWDG for u_2 in (5.40), fixed $k = 20$, and $q = 5$, 7, and 9, on regular Cartesian (*left*) and Voronoi meshes (*right*). The legend is the same for both plots.

and thus more degrees of freedom than in UWVF/PWDG are employed, whereas for fine meshes, the removing procedure has a huge impact. Furthermore, in the convergence regime, both methods lead to slightly better results than PWVEM.

***p*-version:** Here, we compare the three methods with $k = 20$, 40, and 60 for the exact solution u_2 in (5.40) on a regular Cartesian mesh and a Voronoi mesh with 16 elements each. The corresponding plot is Figure 5.21. Moreover, we test the methods with $k = 10$ and 20 for the exact solution u_3 in (5.40) on the same meshes, giving rise to Figure 5.22.

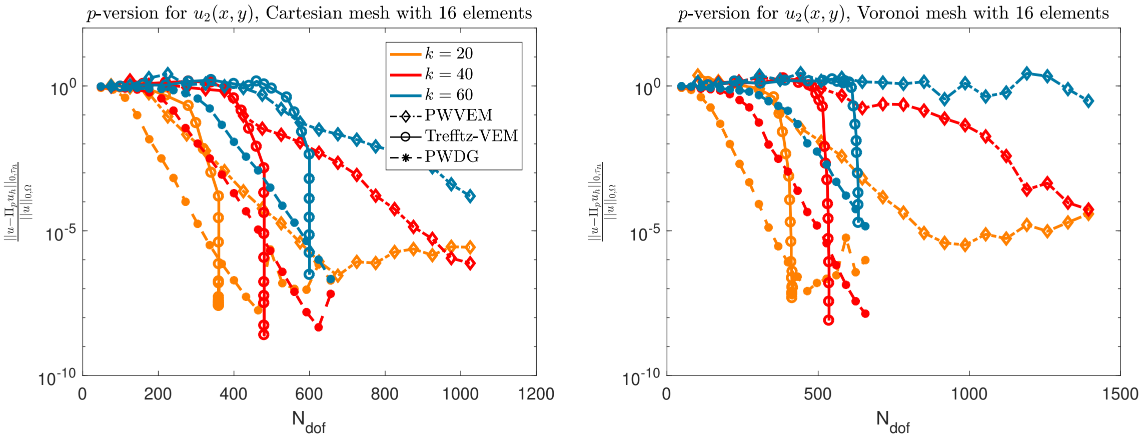


Figure 5.21: Comparison of the p-version of ncTVEM with PWVEM and UWVF/PWDG for u_2 in (5.40) on a regular Cartesian (*left*) and a Voronoi mesh (*right*) with 16 elements each. The legend is the same for both plots.

Here, for ncTVEM, one can observe the above-mentioned cliff effect, i.e. at some points the accuracy increases without increase of the number of degrees of freedom. This is a side effect of the orthogonalization-and-filtering process in Algorithm 2. With this strategy, one can even obtain in some situations a better accuracy in terms of the number of degrees of freedom than with UWVF/PWDG. Especially for u_2, ncTVEM and UWVF/PWDG outperform PWVEM.

***hp*-version:** Finally, we also compare the hp-version of ncTVEM with UWVF/PWDG for the experiment described above, namely, exact solution u_3 in (5.40), wave numbers $k = 10$, 20, and 40, and grading parameters $\mu = 1/2$ and $\mu = 1/3$. The associated error plot is Figure 5.23.

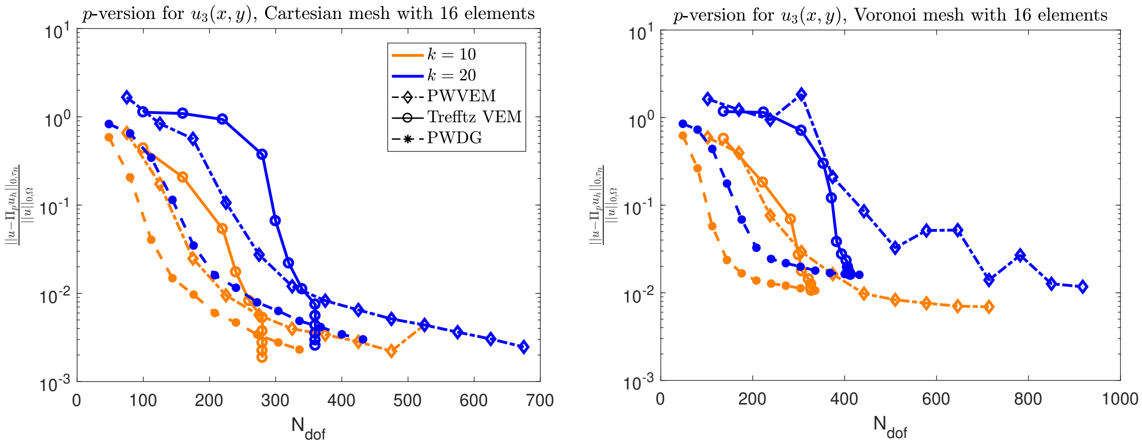


Figure 5.22: Comparison of the p-version of ncTVEM with PWVEM and UWVF/PWDG for u_2 in (5.40) on a regular Cartesian (*left*) and a Voronoi mesh (*right*) with 16 elements each. The legend is the same for both plots.

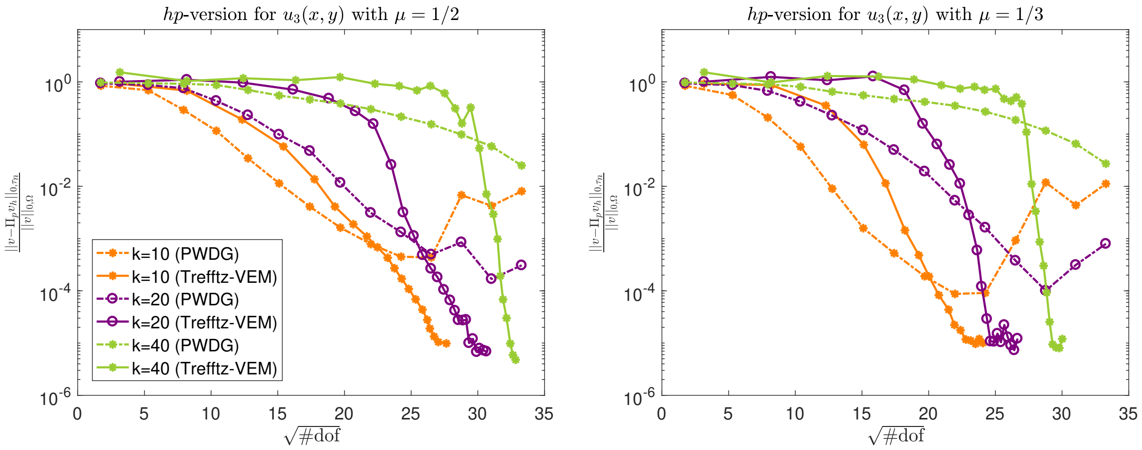


Figure 5.23: hp-version of ncTVEM and UWVF/PWDG on the test case u_3 in (5.40), by employing graded meshes as those in Figure 5.18 with wave numbers $k = 10$, 20, and 40, and grading parameters $\mu = 1/2$ (*left*) and $\mu = 1/3$ (*right*). The distribution of the effective plane wave degree indices is as in (5.44). In both plots, the approximate L^2 error (5.30) is plotted against the quadratic root of the number of degrees of freedom. The legend is the same for both plots.

We highlight that, for this test case, a much higher accuracy can be achieved for ncTVEM than for UWVF/PWDG. Moreover, in particular for high wave numbers, the filtering process helps to significantly reduce the number of degrees of freedom.

To conclude, we have seen that the orthogonalization-and-filtering procedure renders the method highly competitive.

Remark 18. At this point, we underline that we also tested an elementwise orthogonalization-and-filtering process in the spirit of Algorithm 2 for UWVF/PWDG. More precisely, for every $K \in \mathcal{T}_n$, starting from the $L^2(K)$ bulk plane wave mass matrix, we computed an eigendecomposition similarly to (5.34), and then filtered out those orthogonalized basis functions that were associated with eigenvalues close to zero, i.e. eigenvalues smaller than a parameter σ. By using this strategy, however, we could not observe any gain. In fact, for smaller σ ($\sim 1e-13$), basis functions started getting removed at a level where UWVF/PWDG was already outperformed by ncTVEM. On the other hand, for larger σ ($\sim 1e-10$), the accuracy of the method got affected.

5.5 Dispersion and dissipation properties

In this section, we investigate the dispersion and dissipation properties of ncTVEM and compare them to PWVEM [159] and UWVF/PWDG [71, 112]. As already mentioned in Chapter 1, numerical dispersion describes the failure of a numerical method to reproduce the correct oscillating behavior of the analytical solution. Thus, it represents, besides the discretization error, a possibility to measure deviations of a discrete solution from the corresponding continuous one in a qualitative and quantitative way.

The general strategy for a dispersion analysis can be summarized in the following two steps:

1. Consider the discretization scheme of the numerical method applied to $-\Delta u - k^2 u = 0$ in $\mathbb{R}^2$ using infinite meshes which are invariant under a discrete group of translations. Due to translation invariance, it is then possible to reduce the infinite mesh to a finite one.

2. Given a plane wave with wave number k traveling in a fixed direction, seek a so-called *discrete Bloch wave solution*, which can be regarded as a generalization of the given continuous plane wave based on the underlying approximating spaces, and determine for which (discrete) wave number k_n this Bloch wave solution actually solves the discrete variational formulation. This procedure leads to small nonlinear eigenvalue problems, which need to be solved.

Based on this strategy, in Section 5.5.1, we introduce the abstract framework for the computation of dispersion and dissipation, apt for a later comparison of ncTVEM with PWVEM and UWVF/PWDG. Then, in Section 5.5.2, we specify how ncTVEM, PWVEM, and UWVF/PWDG are realized in this general setting. Finally, in Section 5.5.3, a series of numerical experiments gives insight into the behavior of dispersion and dissipation for the different methods.

This section is the analogue of the study carried out in [111] for UWVF/PWDG.

5.5.1 Abstract dispersion analysis

Firstly, we fix the abstract setting for the dispersion analysis employing the notation of [110].

To this purpose, in order to remove possible dependencies of the dispersion on the boundary conditions of the problem, we consider the homogeneous Helmholtz equation $-\Delta u - k^2 u = 0$ on the unbounded domain $\Omega = \mathbb{R}^2$. Let $\mathcal{T}_n := \{K\}$ be a translation-invariant partition of Ω into polygons with mesh size $h := \max_{K \in \mathcal{T}_n} h_K$, where $h_K := \operatorname{diam}(K)$, i.e. there exists a set of elements $\widehat{K}_1, \ldots, \widehat{K}_r$, $r \in \mathbb{N}$, such that the whole infinite mesh can be covered in a non-overlapping way by shifts of the "reference" patch $\widehat{K} := \bigcup_{j=1}^{r} \widehat{K}_j$. In other words, this assumption implies the existence of translation vectors $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in \mathbb{R}^2$, such that every element $K \in \mathcal{T}_n$ can be written as a linear combination with coefficients in $\mathbb{N}_0$ of one of the reference polygons $\widehat{K}_\ell$, $\ell = 1, \ldots, r$. Some examples for translation-invariant meshes are shown in Figure 7.3. Moreover, as above, we denote by $\mathcal{E}^K$ the set of edges belonging to K.

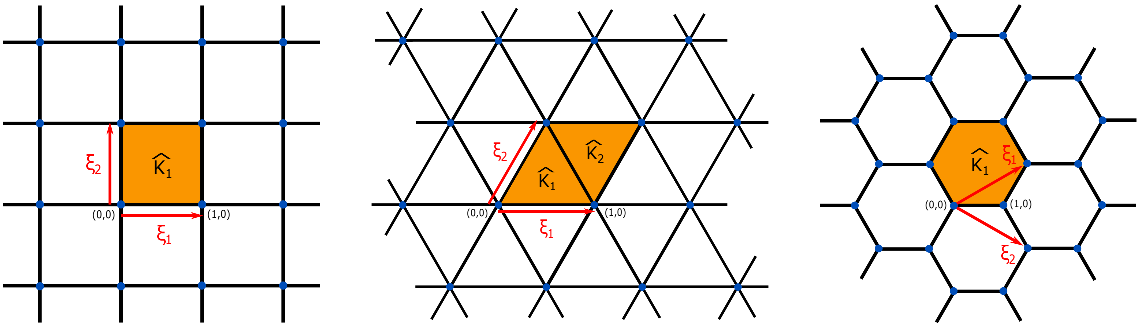


Figure 5.24: Examples of translation-invariant meshes with the corresponding translation vectors $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$: regular Cartesian mesh, triangular mesh, and hexagonal mesh, from *left* to *right*.

Let now $u(\mathbf{x}) = e^{\mathrm{i}k\mathbf{d}\cdot\mathbf{x}}$, $\mathbf{d} \in \mathbb{R}^2$ with $|\mathbf{d}| = 1$, be a plane wave with wave number k and traveling in direction $\mathbf{d}$. We denote by $\mathcal{V}_n$ the global approximation space resulting from the discretization of the homogeneous Helmholtz equation using a Galerkin based numerical method, and by $\widehat{\mathcal{V}}_n \subset \mathcal{V}_n$ a minimal subspace generating $\mathcal{V}_n$ by translations with

$$\boldsymbol{\xi}_{\mathbf{n}} := n_1\boldsymbol{\xi}_1 + n_2\boldsymbol{\xi}_2, \quad \mathbf{n} = (n_1, n_2) \in \mathbb{Z}^2. \tag{5.45}$$

More precisely, depending on the structure of the method, $\widehat{\mathcal{V}}_n$ is determined as follows.

1. *Edge-related* basis functions: In this case, the space $\widehat{\mathcal{V}}_n$ is the span of all basis functions related to a minimal set of edges $\{\eta_i\}_{i=1}^{\lambda^{(1)}}$, $\lambda^{(1)} \in \mathbb{N}$, such that all the other edges of the mesh are obtained by translations with $\boldsymbol{\xi}_n$ of the form (5.45). This is exactly the case of ncTVEM.
2. *Vertex-related* basis functions: Similarly as above, $\widehat{\mathcal{V}}_n$ is the span of all basis functions related to a minimal set of vertices $\{\nu_i\}_{i=1}^{\lambda^{(0)}}$, $\lambda^{(0)} \in \mathbb{N}$, such that all the other mesh vertices are obtained by translations with $\boldsymbol{\xi}_n$ of the form (5.45). Examples are FEM and PWVEM [159].
3. *Element-related* basis functions: Here, the space $\widehat{\mathcal{V}}_n$ is simply given as the span of all basis functions related to a minimal set of elements $\{\sigma_i\}_{i=1}^{\lambda^{(2)}}$, $\lambda^{(2)} \in \mathbb{N}$, such that all other elements of the mesh are obtained by a translation with a vector $\boldsymbol{\xi}_n$ of the form (5.45). One representative of this category is UWVF/PWDG [71,112].

In the following, we will refer to these minimal sets of edges $\{\eta_i\}_{i=1}^{\lambda^{(1)}}$, vertices $\{\nu_i\}_{i=1}^{\lambda^{(0)}}$, and elements $\{\sigma_i\}_{i=1}^{\lambda^{(2)}}$ as *fundamental sets* of edges, vertices, and elements, respectively.

As a direct consequence, every $v_n \in \mathcal{V}_n$ can be written as

$$v_n(\mathbf{x}) = \sum_{\mathbf{n}\in\mathbb{Z}^2} \widehat{v}_n(\mathbf{x} - \boldsymbol{\xi}_{\mathbf{n}}), \quad \widehat{v}_n \in \widehat{\mathcal{V}}_n.$$

Next, we define the discrete *Bloch wave* with wave number k_n and traveling in direction $\mathbf{d}$ by

$$u_n(\mathbf{x}) = \sum_{\mathbf{n}\in\mathbb{Z}^2} e^{\mathrm{i}k_n\mathbf{d}\cdot\boldsymbol{\xi}_{\mathbf{n}}} \widehat{u}_n(\mathbf{x} - \boldsymbol{\xi}_{\mathbf{n}}), \tag{5.46}$$

where $\widehat{u}_n \in \widehat{\mathcal{V}}_n$, and $k_n \in \mathbb{C}$ with $\mathrm{Re}(k_n) > 0$. Note that, since $\widehat{u}_n \in \widehat{\mathcal{V}}_n$, the infinite sum in (5.46) is in fact finite. Furthermore, given $\mathbf{d} \in \mathbb{R}^2$ with $|\mathbf{d}| = 1$, the Bloch wave u_n in (5.46) satisfies

$$u_n(\mathbf{x} + \boldsymbol{\xi}_{\boldsymbol{\ell}}) = e^{\mathrm{i}k_n\mathbf{d}\cdot\boldsymbol{\xi}_{\boldsymbol{\ell}}} u_n(\mathbf{x}),$$

for all $\boldsymbol{\ell} \in \mathbb{Z}^2$. This property follows directly by using the definition of the Bloch wave:

$$\begin{aligned} u_n(\mathbf{x} + \boldsymbol{\xi}_{\boldsymbol{\ell}}) &= \sum_{\mathbf{n}\in\mathbb{Z}^2} e^{\mathrm{i}k_n\mathbf{d}\cdot\boldsymbol{\xi}_{\mathbf{n}}} \widehat{u}_n(\mathbf{x} + \boldsymbol{\xi}_{\boldsymbol{\ell}} - \boldsymbol{\xi}_{\mathbf{n}}) = \sum_{n\in\mathbb{Z}^2} e^{\mathrm{i}k_n\mathbf{d}\cdot\boldsymbol{\xi}_{\mathbf{n}}} \widehat{u}_n(\mathbf{x} - \boldsymbol{\xi}_{\mathbf{n}-\boldsymbol{\ell}}) \\ &= e^{\mathrm{i}k_n\mathbf{d}\cdot\boldsymbol{\xi}_{\boldsymbol{\ell}}} \sum_{\mathbf{m}\in\mathbb{Z}^2} e^{\mathrm{i}k_n\mathbf{d}\cdot\boldsymbol{\xi}_{\mathbf{m}}} \widehat{u}_n(\mathbf{x} - \boldsymbol{\xi}_{\mathbf{m}}) = e^{\mathrm{i}k_n\mathbf{d}\cdot\boldsymbol{\xi}_{\boldsymbol{\ell}}} u_n(\mathbf{x}). \end{aligned}$$

Therefore, Bloch waves can be regarded as discrete counterparts, based on the approximation spaces, of continuous plane waves.

We recall the definition of the global (continuous) sesquilinear form

$$a_k(u, v) := \sum_{K\in\mathcal{T}_n} a_k^K(u, v) := \sum_{K\in\mathcal{T}_n} \left[\int_K \nabla u \cdot \overline{\nabla v} \,\mathrm{d}x - k^2 \int_K u\overline{v} \,\mathrm{d}x \right] \quad \forall u, v \in H^1(\mathbb{R}^2), \tag{5.47}$$

and we denote by $a_{k,n}(\cdot,\cdot)$ the global discrete sesquilinear form defining the numerical method under consideration. In Section 5.5.2, $\widehat{\mathcal{V}}_n$ and $a_{k,n}(\cdot,\cdot)$ will be specified for ncTVEM, PWVEM, and UWVF/PWDG, respectively.

Next, we define the discrete wave number $k_n \in \mathbb{C}$ as follows.

Definition 5.5.1. Given $k > 0$ and $\mathbf{d} \in \mathbb{R}^2$ with $|\mathbf{d}| = 1$, the *discrete wave number* $k_n \in \mathbb{C}$ is the number with minimal $|k - k_n|$, for which a discrete Bloch wave u_n of the form (5.46) is a solution to the discrete problem

$$a_{k,n}(u_n, \widehat{v}_n) = 0 \quad \forall \widehat{v}_n \in \widehat{\mathcal{V}}_n. \tag{5.48}$$

Due to the scaling invariance of the mesh, we can assume that $h = 1$. Notice that the wave number k on a mesh with $h = 1$ corresponds to the wave number $k_0 = \frac{k}{h_0}$ on a mesh with mesh size h_0.

Having this, the general procedure in the dispersion analysis now consists in finding those discrete wave numbers $k_n \in \mathbb{C}$ and coefficients $\widehat{u}_n \in \widehat{\mathcal{V}}_n$, for which a Bloch wave solution of the form (5.46) satisfies (5.48), and to measure the deviation of k_n from k afterwards. This strategy results in solving small nonlinear eigenvalue problems. In fact, by plugging the Bloch wave ansatz (5.46) into (5.48) and using the sesquilinearity of $a_{k,n}(\cdot,\cdot)$, we obtain

$$\sum_{\mathbf{n}\in\mathbb{Z}^2} e^{\mathrm{i}k_n\mathbf{d}\cdot\boldsymbol{\xi}_\mathbf{n}} a_{k,n}(\widehat{u}_n(\cdot-\boldsymbol{\xi}_\mathbf{n}),\widehat{v}_n) = 0 \quad \forall \widehat{v}_n \in \widehat{\mathcal{V}}_n. \tag{5.49}$$

Let $\{\widehat{\chi}_s\}_{s=1}^{\Xi} \subset \widehat{\mathcal{V}}_n$ be a set of basis functions for the space $\widehat{\mathcal{V}}_n$ that are related to fundamental elements, vertices, or edges, depending on the method. Then, we can expand $\widehat{u}_n$ in terms of this basis as

$$\widehat{u}_n = \sum_{t=1}^{\Xi} u_t\widehat{\chi}_t.$$

Plugging this ansatz into (5.49), testing with $\widehat{\chi}_s$, $s = 1,\dots,\Xi$, and interchanging the sums (this can be done since the infinite sum over $\mathbf{n}$ is in fact finite) yields

$$\sum_{t=1}^{\Xi} u_t \left(\sum_{\mathbf{n}\in\mathbb{Z}^2} e^{\mathrm{i}k_n\mathbf{d}\cdot\boldsymbol{\xi}_\mathbf{n}} a_{k,n}(\widehat{\chi}_t(\cdot-\boldsymbol{\xi}_\mathbf{n}),\widehat{\chi}_s) \right) = 0 \quad \forall s = 1,\dots,\Xi, \tag{5.50}$$

which can be represented as

$$\sum_{t=1}^{\Xi} \boldsymbol{T}_{s,t}(k_n)u_t = 0 \quad \forall s = 1,\dots,\Xi, \tag{5.51}$$

with

$$\boldsymbol{T}_{s,t}(k_n) := \sum_{\mathbf{n}\in\mathbb{Z}^2} e^{\mathrm{i}k_n\mathbf{d}\cdot\boldsymbol{\xi}_\mathbf{n}} a_{k,n}(\widehat{\chi}_t(\cdot-\boldsymbol{\xi}_\mathbf{n}),\widehat{\chi}_s). \tag{5.52}$$

The matrix problem corresponding to (5.51) has the form

$$\boldsymbol{T}(k_n)\mathbf{u} = \mathbf{0}, \tag{5.53}$$

where $\boldsymbol{T} : \mathbb{C} \to \mathbb{C}^{\Xi\times\Xi}$ is defined via (5.52), and $\boldsymbol{u} = (u_1,\dots,u_\Xi)^T \in \mathbb{C}^\Xi$. We highlight that $\boldsymbol{T}$ is a holomorphic map, and that (5.53) is a small nonlinear eigenvalue problem, which can be solved using e.g. an iterative method as done in [110], or a direct method based on a rational interpolation procedure [178] or on a contour integral approach [16,54]. For the numerical experiments presented in Section 5.5.3, we will make use of the latter, which we will denote by *contour integral method* (CIM) in the sequel. We stress that, due to the use of plane wave related basis functions, deriving an exact analytical solution to (5.53) is not even be possible for the lowest-order case.

5.5.2 Minimal generating subspaces and sesquilinear forms

In this section, we specify the minimal generating subspaces $\widehat{\mathcal{V}}_n$, the corresponding sets of basis functions $\{\widehat{\chi}_s\}_{s=1}^{\Xi}$, and the sequilinear forms $a_{k,n}(\cdot,\cdot)$ for ncTVEM and PWVEM, and we recall them from [110] for UWVF/PWDG. The basis functions for these three methods are edge-related, vertex-related, and element-related, respectively. In Figures 5.25-5.27 in Section 5.5.2, the stencils related to the fundamental sets of edges, vertices, and elements are depicted for these three methods and the meshes in Figure 5.24.

Before doing that, we recall that, given $\{\mathbf{d}_\ell\}_{\ell=1}^{p}$, $p = 2q + 1$, $q \in \mathbb{N}$, a set of equidistributed plane wave directions, w_ℓ is the plane wave traveling along $\mathbf{d}_\ell$ and $\mathbb{PW}_p(K)$ is the bulk plane wave space related to an element $K \in \mathcal{T}_n$, see (4.5) and (4.7), respectively.

Nonconforming Trefftz virtual element method (ncTVEM)

We firstly specify the minimal generating subspace $\widehat{\mathcal{V}}_n^{(1)}$. To this purpose, we shortly recall the construction of the method from Sections 4.1 and 5.4. Starting, on each edge e, from the set of plane wave traces $\{w_\ell^e\}_{\ell=1}^p$, see (4.8), we compute a set of $L^2(e)$ orthogonal functions $\{\widetilde{w}_\ell^e\}_{\ell=1}^{\widetilde{p}_e}$ using Algorithm 2. Defining then the space of filtered $L^2(e)$ orthogonalized plane wave traces by

$$\widetilde{\mathbb{PW}}(e) := \operatorname{span}\{\widetilde{w}_\ell^e,\, \ell \in \mathcal{J}_e\},$$

where $\mathcal{J}_e := \{1, \dots, \widetilde{p}_e\}$, the local Trefftz VE space related to an element $K \in \mathcal{T}_n$ is introduced by

$$\mathcal{V}_n^{(1)}(K) := \left\{ v_n \in H^1(K) \mid \Delta v_n + k^2 v_n = 0 \text{ in } K, \quad \gamma_I^K(v_n)_{|_e} \in \widetilde{\mathbb{PW}}(e)\, \forall e \in \mathcal{E}^K \right\}, \tag{5.54}$$

with impedance trace γ_I^K, see (4.13). The corresponding set of local degrees of freedom is given, for any $v_n \in \mathcal{V}_n^{(1)}(K)$, by

$$\operatorname{dof}_{e,\ell}(v_n) = \frac{1}{h_e} \int_e v_n \overline{\widetilde{w}_\ell^e}\, \mathrm{d}s \quad \forall e \in \mathcal{E}^K,\, \forall \ell \in \mathcal{J}_e,$$

and the canonical basis functions $\{\psi_{(e,\ell)}^K\}_{e \in \mathcal{E}^K,\, \ell \in \mathcal{J}_e} \subset \mathcal{V}_n^{(1)}(K)$ are defined via

$$\operatorname{dof}_{\widetilde{e},\widetilde{\ell}}\left(\psi_{(e,\ell)}^K\right) = \delta_{(e,\ell),(\widetilde{e},\widetilde{\ell})} = \begin{cases} 1 & \text{if } (e,\ell) = (\widetilde{e},\widetilde{\ell}) \\ 0 & \text{otherwise} \end{cases} \quad \forall e, \widetilde{e} \in \mathcal{E}^K,\, \forall \ell \in \mathcal{J}_e,\, \forall \widetilde{\ell} \in \mathcal{J}_{\widetilde{e}}.$$

Further, the global nonconforming Trefftz VE space $\mathcal{V}_n^{(1)}$ is

$$\mathcal{V}_n^{(1)} := \{v_n \in H^{1,nc}(\mathcal{T}_n) : v_{n|_K} \in \mathcal{V}_n^{(1)}(K) \quad \forall K \in \mathcal{T}_n\},$$

where $H^{1,nc}(\mathcal{T}_n)$ is the global nonconforming Sobolev space

$$H^{1,nc}(\mathcal{T}_n) := \left\{ v_n \in H^1(\mathcal{T}_n) : \int_e [\![v_n]\!]_N \cdot \mathbf{n}^e\, \overline{w^e}\, \mathrm{d}s = 0 \quad \forall w^e \in \widetilde{\mathbb{PW}}(e),\, \forall e \in \mathcal{E}_n \right\},$$

with $\mathbf{n}^e$ being a fixed unit normal vector to the edge e.

Let now $\{\eta_i\}_{i=1}^{\lambda^{(1)}}$ be a fundamental set of edges. Then, the set of basis functions $\{\widehat{\chi}_s^{(1)}\}_{s=1}^{\Xi}$ spanning the minimal generating subspace $\widehat{\mathcal{V}}_n^{(1)}$ is given by the union of the canonical basis functions related to $\{\eta_i\}_{i=1}^{\lambda^{(1)}}$. More precisely, for $s \leftrightarrow (i,j)$, $i \in \{1, \dots, \lambda^{(1)}\}$ and $j \in \mathcal{J}_{\eta_i}$, i.e. we identify s with the edge index i and the index j associated with the j-th orthogonalized plane wave basis function on this edge as above,

$$\widehat{\chi}_s^{(1)} = \widehat{\chi}_{(i,j)}^{(1)} := \Psi_{(\eta_i,j)},$$

where $\Psi_{(\eta_i,j)}$ is defined elementwise as follows. If $K \in \mathcal{T}_n$ is an element abutting the edge η_i, then $\Psi_{(\eta_i,j)|_{\eta_i}}$ coincides with the local canonical basis function associated with the (global) edge η_i and the j-th orthogonalized edge plane wave basis function; otherwise $\Psi_{(\eta_i,j)}$ is zero. Clearly, $\Xi = \sum_{i=1}^{\lambda^{(1)}} \widetilde{p}_{\eta_i}$.

Regarding the sesquilinear form $a_{k,n}^{(1)}(\cdot,\cdot)$, it is given by $a_{k,h}(\cdot,\cdot)$ in (4.30).

Plane wave virtual element method (PWVEM)

We shortly recall the structure of PWVEM introduced in [159], using the notation employed there.

To this purpose, given $K \in \mathcal{T}_n$, the lowest-order local VE space is defined as

$$\widetilde{\mathcal{V}}_n^{(0)}(K) := \{v \in H^1(K) : v|_{\partial K} \in C^0(\partial K),\, v|_e \in \mathbb{P}_1(e)\, \forall e \in \mathcal{E}^K,\, \Delta v = 0 \text{ in } K\}, \tag{5.55}$$

where $\mathcal{E}^K$ is the set of edges of K. We underline that $\widetilde{\mathcal{V}}_n^{(0)}(K)$ includes $\mathbb{P}_1(K)$, the space of linear polynomials over K, as a subspace. Moreover, it contains functions which cannot be written down explicitly in closed form, justifying the term *virtual* in the name of the method.

The space (5.55) is endowed with the local set of degrees of freedom given by the point values at the vertices V_s^K, $s=1,\ldots,n_K$, of K, where n_K denotes their number. Due to the unisolvency of the degrees of freedom, a set of canonical basis functions $\{\phi_r^K\}_{r=1}^{n_K}$ can be defined by duality, i.e. $\phi_r^K(V_s^K)=\delta_{rs}$, $r,s=1,\ldots,n_K$, with δ denoting the standard Kronecker delta. It can be easily shown that these basis functions actually form a partition of unity.

The local VE space is given by the modulation of canonical basis functions with plane waves:

$$\mathcal{V}_n^{(0)}(K) := \left\{ v\in H^1(K) : \, v=\sum_{r=1}^{n_K}\sum_{j=1}^{p} \alpha_{rj}^K \varphi_{(r,j)}^K, \, \alpha_{rj}^K\in\mathbb{C} \right\} \supset \mathbb{PW}_p(K), \tag{5.56}$$

where $\varphi_{(r,j)}^K := \phi_r^K w_j^K$. Note that the inclusion in (5.56) is a direct consequence of the properties of the canonical basis functions and is in fact essential for deriving best approximation estimates needed in the error analysis of the method.

The global plane wave VE space is defined in terms of the local ones:

$$\mathcal{V}_n^{(0)} := \left\{ v_n\in C^0(\mathbb{R}^2) : \, v_{n|_K}\in\mathcal{V}_n^{(0)}(K) \quad \forall K\in\mathcal{T}_n \right\}.$$

Moreover, the global sesquilinear form is given by

$$a_{k,n}^{(0)}(u_n,v_n) := a_{k,h}(u_n,v_n) \quad \forall u_n,v_n\in\mathcal{V}_n^{(0)}, \tag{5.57}$$

where $a_{k,h}(\cdot,\cdot)$ is defined in (4.30) with the only modification that the computable projector Π_p^K in (4.23) maps this time from $\mathcal{V}_n^{(0)}(K)$ into $\mathbb{PW}_p(K)$. Similarly as for ncTVEM, conditions on the stabilization terms are required. Additionally, in Section 5.5.3 below, the dispersion and dissipation properties of the method will be studied numerically for different choice of stabilizations that work fine in practice.

Given a fundamental set of vertices $\{\nu_i\}_{i=1}^{\lambda^{(0)}}$, the set of basis functions $\{\widehat{\chi}_s^{(0)}\}_{s=1}^{\Xi}\subset C^0(\mathbb{R}^2)$ is defined as follows. Let $s\leftrightarrow(i,j)$, with $i\in\{1,\ldots,\lambda^{(0)}\}$ and $j\in\{1,\ldots,p\}$, i.e. we identify s with the vertex index i and the direction index j. Then,

$$\widehat{\chi}_s^{(0)} = \widehat{\chi}_{(i,j)}^{(0)} := \Phi_{\nu_i} w_j \in C^0(\mathbb{R}^2), \tag{5.58}$$

where we recall that w_j is the plane wave traveling along the direction $\mathbf{d}_j$, and where Φ_{ν_i} is defined elementwise as follows. If $K\in\mathcal{T}_n$ is an element abutting the fundamental vertex ν_i, then $\Phi_{\nu_i|_K}$ coincides with the local canonical basis function in K which is associated with the (global) vertex ν_i; otherwise $\Phi_{\nu_i|_K}$ is set to zero. Taking into account the definitions of the degrees of freedom and of $\mathcal{V}_n^{(0)}(K)$ in (5.55), it can be easily seen that $\widehat{\chi}_s^{(0)}$ is in fact globally continuous with compact support. Clearly, $\Xi=\lambda^{(0)}p$.

To conclude, for PWVEM, the minimal generating subspace $\widehat{\mathcal{V}}_n^{(0)}$ of $\mathcal{V}_n^{(0)}$ is given as the span of the basis functions (5.58), $\Xi=\lambda^{(0)}p$, and the employed sesquilinear form is $a_{k,n}^{(0)}(\cdot,\cdot)$ in (5.57).

Ultra weak variational formulation (UWVF) / Plane wave discontinuous Galerkin method (PWDG)

For UWVF/PWDG, we refer to [110], where a complete dispersion analysis was carried out. Nevertheless, for the sake of completeness, we shortly recall here the definitions of the minimal generating subspace and the sesquilinear form adapted to our setting.

The global approximation space $\mathcal{V}_n^{(2)}$ is given by

$$\mathcal{V}_n^{(2)} := \{v_n\in L^2(\mathbb{R}^2) : \, v_{n|_K}\in\mathbb{PW}_p(K) \quad \forall K\in\mathcal{T}_n\}.$$

Moreover, the global sesquilinear form $a_{k,n}^{(2)}(\cdot,\cdot)$ is defined by

$$\begin{aligned} a_{k,n}^{(2)}(u_n,v_n) := &\sum_{K\in\mathcal{T}_n} a_k^K(u_n,v_n) - \int_{\mathcal{F}_n} [\![u]\!]_N\cdot\{\!\{\overline{\nabla_n v}\}\!\}\,\mathrm{d}s - \frac{\beta}{\mathrm{i}k}\int_{\mathcal{F}_n}[\![\nabla_n u_n]\!]_N\cdot[\![\overline{\nabla_n v_n}]\!]_N\,\mathrm{d}s \\ &- \int_{\mathcal{F}_n}\{\!\{\nabla_n u_n\}\!\}\cdot[\![\overline{v_n}]\!]_N\,\mathrm{d}s + \mathrm{i}k\alpha\int_{\mathcal{F}_n}[\![u_n]\!]_N\cdot[\![\overline{v_n}]\!]_N\,\mathrm{d}s, \quad \forall u_n,v_n\in\mathcal{V}_n^{(2)}, \end{aligned} \tag{5.59}$$

where $a_k^K(\cdot,\cdot)$ is given in (5.47), $\mathcal{F}_n$ is the mesh skeleton, $\alpha, \beta > 0$ are the flux parameters, ∇_n is the broken gradient, $[\![\cdot]\!]_N$ is the standard trace jump as defined in (3.8), and, for a given edge e, denoting by K^- and K^+ its adjacent elements,

$$\{\!\{\nabla_n u\}\!\} := \frac{1}{2}\left(\nabla u_{n|_{K^+}} + \nabla u_{n|_{K^-}}\right)$$

is the trace average.

Let now $\{\sigma_i\}_{i=1}^{\lambda^{(2)}}$ be a fundamental set of elements. Then, the basis functions $\{\widehat{\chi}_s^{(2)}\}_{s=1}^{\Xi}$ are given by $\{w_j^{\sigma_i}\}_{i=1,\ldots,\lambda^{(2)},j=1\ldots,p}$, where $s \leftrightarrow (i,j)$, i.e. s is identified with the element index i and the plane wave direction index j, and $\Xi = \lambda^{(2)}p$. As mentioned above, the minimal generating subspace $\widehat{\mathcal{V}}_n^{(2)} \subset \mathcal{V}_n^{(2)}$ is simply the span of the basis functions $\{\widehat{\chi}_s^{(2)}\}_{s=1}^{\Xi}$, and the sesquilinear form $a_{k,n}^{(2)}(\cdot,\cdot)$ is given in (5.59).

Overview of the stencils generating the minimal subspaces

In Figures 5.25-5.27, we illustrate the stencils of the basis functions for ncTVEM, PWVEM, and UWVF/PWDG, employing the meshes made of squares, triangles, and hexagons, respectively, depicted in Figure 5.24. The fundamental sets of edges, vertices, and elements are displayed in dark-blue, and the translation vectors $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ in red. Furthermore, the supports of the basis functions spanning the minimal generating subspaces are colored in light-blue for ncTVEM and PWVEM. Due to the locality of the basis functions, only those associated with the edges, vertices, and elements displayed in dark-blue and dark-yellow contribute to the sum (5.50). Integration only has to be performed over the elements K^ζ and the adjacent edges.

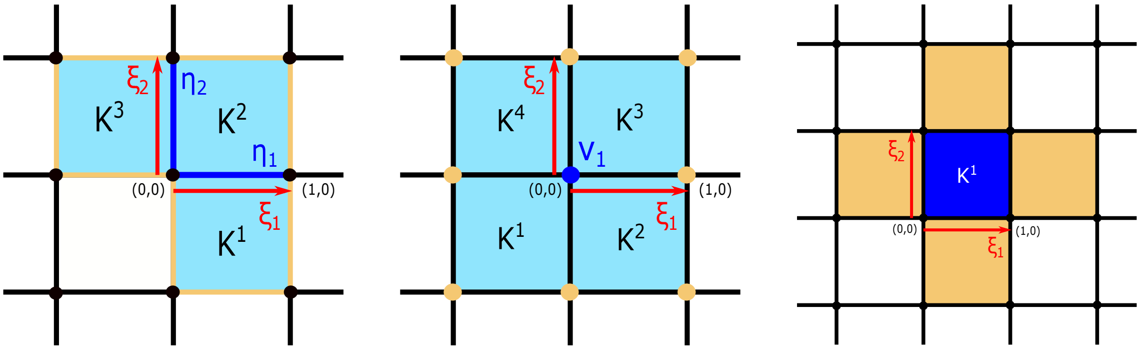


Figure 5.25: Stencils of the basis functions related to the fundamental sets of edges (ncTVEM), vertices (PWVEM), and elements (UWVF/PWDG), respectively, from *left* to *right*, when employing the meshes made of squares in Figure 5.24.

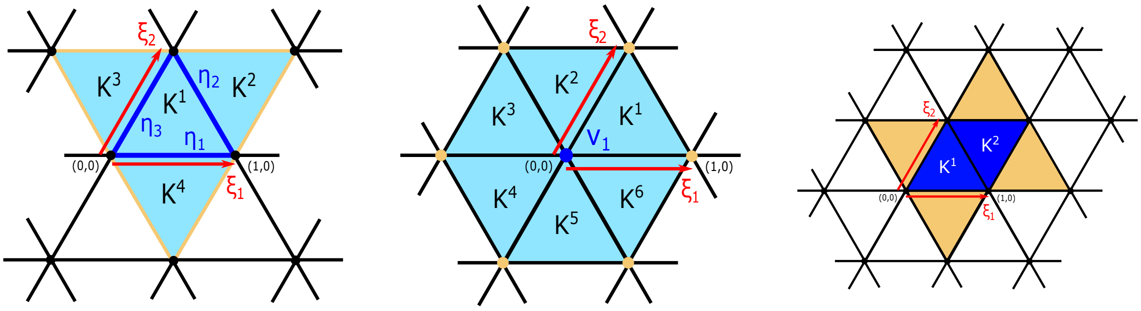


Figure 5.26: Stencils of the basis functions related to the fundamental sets of edges (ncTVEM), vertices (PWVEM), and elements (UWVF/PWDG), respectively, from *left* to *right*, when employing the meshes made of triangles in Figure 5.24.

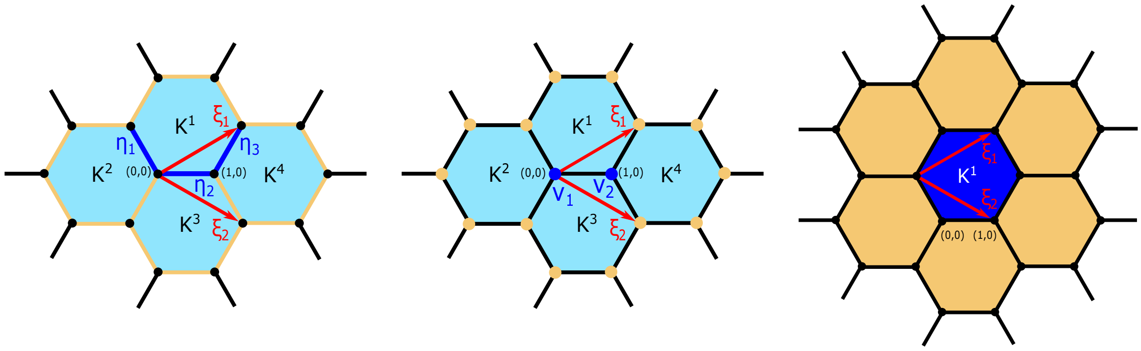


Figure 5.27: Stencils of the basis functions related to the fundamental sets of edges (ncTVEM), vertices (PWVEM), and elements (UWVF/PWDG), respectively, from *left* to *right*, when employing the meshes made of hexagons in Figure 5.24.

5.5.3 Numerical results

In this section, after fixing some parameters for the different methods and specifying the quantities to be compared, we present a series of numerical experiments using the meshes portrayed in Figure 5.24. Firstly, we investigate the qualitative behavior of dispersion and dissipation depending on the Bloch wave angle θ in Definition 5.46. Then, we compare the dispersion and dissipation errors against the effective plane wave degree q and against the dimensions of the minimal generating subspaces. Finally, the dependence of the errors on the wave number is studied.

Choice of the stabilizations in ncTVEM and PWVEM, and the parameters in PWDG. For ncTVEM, we take the modified D-recipe stabilization in (5.32). Furthermore, for PWVEM, we employ the stabilization suggested in [159]. More precisely, analogously as for ncTVEM, see (5.20), for PWVEM, the stabilization can be written locally, i.e. on each $K \in \mathcal{T}_n$, in matrix form as

$$\overline{(\mathbf{I}^K - \mathbf{\Pi}^K)}^T \mathbf{S}^K (\mathbf{I}^K - \mathbf{\Pi}^K),$$

where $\mathbf{\Pi}^K$ is the matrix representation of the composition of the embedding of $\mathbb{PW}_p(K)$ into $\mathcal{V}_n^{(0)}(K)$, after Π_p^K. The matrix $\mathbf{S}^K$ is a suitable approximation of the matrix with entries given by

$$[(r,j),(s,\ell)] \mapsto \int_K \left(\nabla\varphi^K_{(r,j)} \cdot \overline{\nabla\varphi^K_{(s,\ell)}} - k^2 \varphi^K_{(r,j)} \overline{\varphi^K_{(s,\ell)}}\right) \mathrm{d}x.$$

By using the notation of (5.56), we compute

$$\begin{aligned}\nabla\varphi^K_{(r,j)} \cdot \overline{\nabla\varphi^K_{(s,\ell)}} - k^2 \varphi^K_{(r,j)} \overline{\varphi^K_{(s,\ell)}} &= (\nabla\phi^K_r \cdot \overline{\nabla\phi^K_s}) w^K_j \overline{w^K_\ell} + \mathrm{i}k(\mathbf{d}_j \cdot \overline{\nabla\phi^K_s}) \phi^K_r w^K_j \overline{w^K_\ell} \\ &\quad - \mathrm{i}k(\mathbf{d}_\ell \cdot \nabla\phi^K_r) \overline{\phi^K_s} w^K_j \overline{w^K_\ell} + k^2(\mathbf{d}_j \cdot \mathbf{d}_\ell - 1) \phi^K_r \overline{\phi^K_s} w^K_j \overline{w^K_\ell}.\end{aligned} \tag{5.60}$$

Then, due to scaling considerations, the last three terms on the right-hand side are neglected, and the first one is simplified obtaining

$$\mathbf{S}^K_{(s,\ell),(r,j)} = \frac{\delta_{r,s}}{h_K^2} \int_K w^K_j \overline{w^K_\ell}\, \mathrm{d}x, \tag{5.61}$$

where δ is the usual Kronecker delta.

We highlight that, by taking the analogue of (5.32) for PWVEM, one does not recover numerically the expected theoretical rate of convergence of the method. On the other hand, (5.61) cannot be directly used in ncTVEM due to the fact that plane wave directions are filtered out on each edge, but are not removed in the bulk, which would lead to dimensional inconsistencies.

Finally, for PWDG, we use the choice of the flux parameters of the ultra weak variational formulation (UWVF), i.e. $\alpha = \beta = 1/2$.

Numerical quantities. Given a wave number $k > 0$ and letting k_n be the discrete wave number in Definition 5.5.1, we will study the following quantities:

- *dispersion error* $|\mathrm{Re}(k - k_n)|$, which describes the difference of the propagation velocities of the continuous and discrete plane wave solutions;
- *dissipation error* $|\mathrm{Im}(k_n)| = |\mathrm{Im}(k - k_n)|$, which represents the difference of the amplitudes (damping) of the continuous and discrete plane wave solutions;
- *total error* $|k - k_n|$, which measures the total deviation of the continuous and discrete wave numbers.

Dependence of dispersion and dissipation on the Bloch wave angle θ

Here, we study dispersion and dissipation of the different methods in dependence on the angle θ of the direction $\mathbf{d}$ in the definition of the Bloch wave in (5.46). Importantly, we are here interested in a qualitative comparison of the methods, rather than a quantitative one, which should be performed in terms of the dimensions of the minimal generating subspaces instead of the effective degrees, and which is discussed below.

To this purpose, in Figures 5.28-5.30, the numerical quantities $|\mathrm{Re}(k - k_n)|$ and $|\mathrm{Im}(k_n)|$ are plotted against θ for the meshes made of squares, triangles, and hexagons, respectively, shown in Figure 5.24. We took $k = 3$ and $q = 7$ for all those types of meshes (Figures 5.28-5.30, left). Moreover, for $k = 10$, we chose $q = 10$ for the squares (Figure 5.28, right) and the triangles (Figure 5.29, right), and $q = 13$ for the hexagons (Figure 5.30, right). We remark that the latter choice for q on the meshes made of hexagons is purely for demonstration purposes, in order to obtain a reasonable range for the errors, where one can see the behavior more clearly. Moreover, we recall that the wave number k here (mesh size $h = 1$) corresponds to the wave number $k_0 = \frac{k}{h_0}$ on a mesh with mesh size h_0.

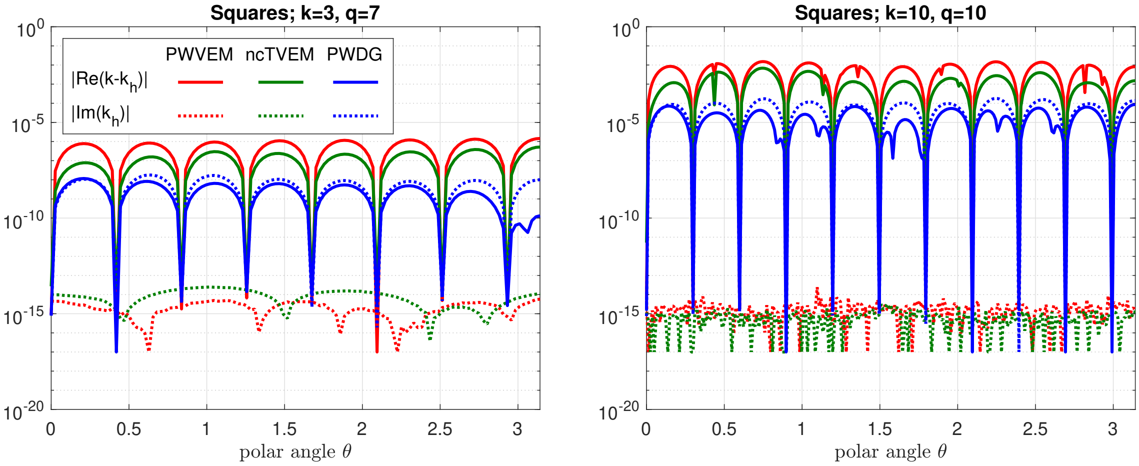


Figure 5.28: Dispersive and dissipative behavior of ncTVEM, PWVEM, and UWVF/PWDG in dependence on the polar angle θ of the Bloch wave direction $\mathbf{d}$ in (5.46) on the meshes made of squares in Figure 5.24, with $k = 3$ and $q = 7$ (*left*), and $k = 10$ and $q = 10$ (*right*).

We notice that dispersion and dissipation are zero, up to machine precision, for choices of the Bloch wave direction $\mathbf{d}$ in (5.46) coinciding with one of the plane wave directions $\{\mathbf{d}_j\}_{j=1}^{p}$ (here we always took equidistributed directions $\mathbf{d}_j$, where $\mathbf{d}_1 = (1, 0)$). This follows directly from the fact that, in this case, the Bloch wave satisfying (5.48) coincides with the corresponding plane wave traveling along the direction $\mathbf{d}$. Moreover, we observe that, for ncTVEM and PWVEM, the dispersion error dominates the dissipation error, whereas, for UWVF/PWDG, dissipation dominates dispersion. Furthermore, the dissipation $|\mathrm{Im}(k_n)|$ is basically zero for PWVEM.

Remark 19. We highlight that, in the case of VEM, the dissipation and dispersion behavior also hinges upon the choice of stabilization. To this purpose, for PWVEM, we compare the results

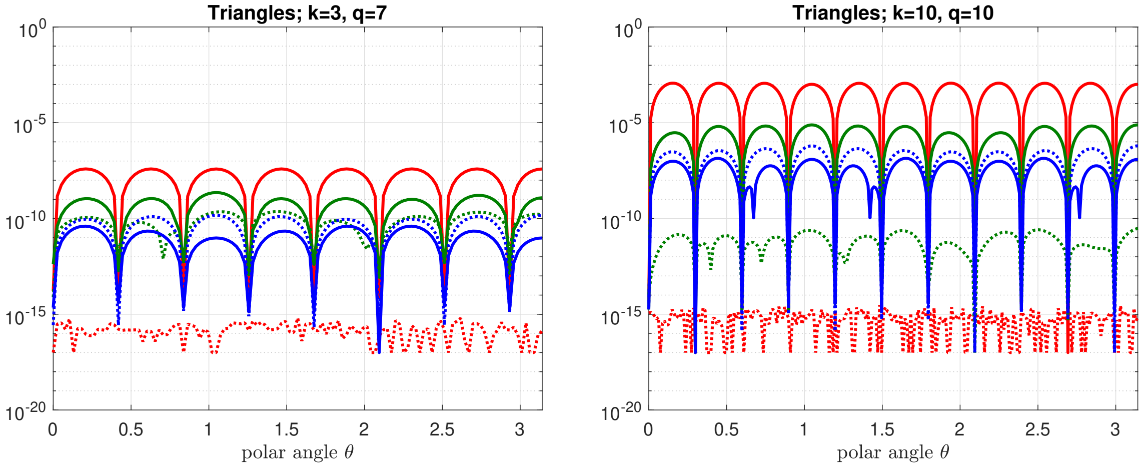


Figure 5.29: Dispersive and dissipative behavior of ncTVEM, PWVEM, and UWVF/PWDG in dependence on the polar angle θ of the Bloch wave direction $\mathbf{d}$ in (5.46) on the meshes made of triangles in Figure 5.24, with $k = 3$ and $q = 7$ (*left*), and $k = 10$ and $q = 1$ (*right*). The color legend is the same as in Figure 5.28.

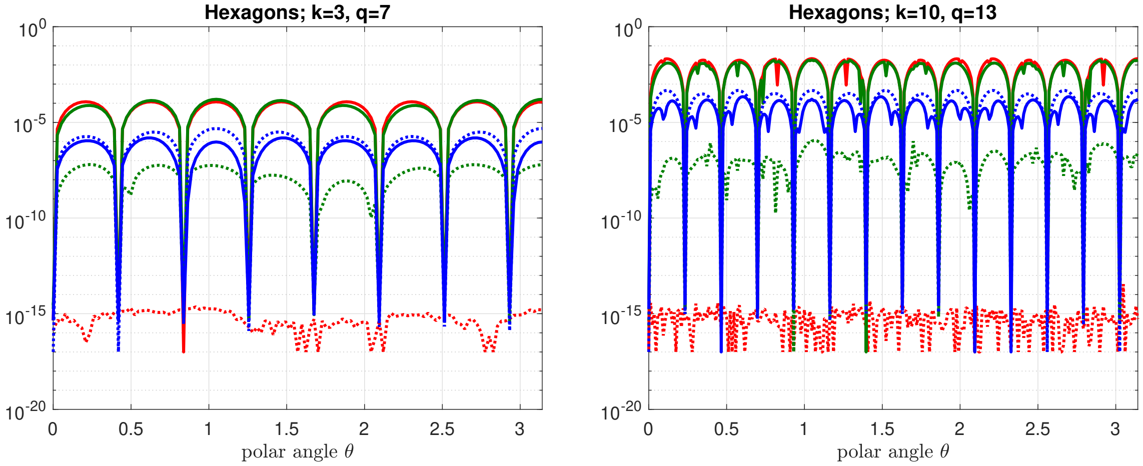


Figure 5.30: Dispersive and dissipative behavior of ncTVEM, PWVEM, and UWVF/PWDG in dependence on the polar angle θ of the Bloch wave direction $\mathbf{d}$ in (5.46) on the meshes made of hexagons in Figure 5.24, with $k = 3$ and $q = 7$ (*left*), and $k = 10$ and $q = 13$ (*right*). The color legend is the same as in Figure 5.28.

obtained when employing the *standard* stabilization in (5.61) with two alternative stabilizations that also lead to the correct convergence behavior for the discretization error in practice. More precisely, let $\Pi^{\nabla,K}_{p,1} : \widetilde{\mathcal{V}}^{(0)}_n(K) \to \mathbb{P}_1(K) \subset \widetilde{\mathcal{V}}^{(0)}_n(K)$ be the projector onto polynomials of degree at most one, defined by

$$\begin{cases} \int_K \nabla(\Pi^{\nabla,K}_{p,1} v_n) \cdot \nabla p_1 \,\mathrm{d}x & = \int_K \nabla v_n \cdot \nabla p_1 \,\mathrm{d}x \quad \forall p_1 \in \mathbb{P}_1(K) \\ \frac{1}{n_K} \sum_{i=1}^{n_K} (\Pi^{\nabla,K}_{p,1} v_n)(V_i^K) & = \frac{1}{n_K} \sum_{i=1}^{n_K} v_n(V_i^K), \end{cases}$$

for all $v_n \in \widetilde{\mathcal{V}}^{(0)}_n(K)$, where V_i^K, $i = 1, \dots, n_K$, are the vertices of K; see [31, 36]. We consider

- *standard*, which is the stabilization defined in (5.61);
- *stab 1*, which is the stabilization one gets by replacing ϕ_r^K and ϕ_s^K on the right-hand side of (5.60) with $\Pi^{\nabla,K}_{p,1}\phi_r^K$ and $\Pi^{\nabla,K}_{p,1}\phi_s^K$, respectively;
- *stab 2*, the resulting stabilization after substituting the right-hand side of (5.60) by

$$\delta_{r,s} \left[(\nabla(\Pi^{\nabla,K}_{p,1}\phi_r^K) \cdot \overline{\nabla(\Pi^{\nabla,K}_{p,1}\phi_s^K)} w_j^K \overline{w_\ell^K} + k^2(\mathbf{d}_j \cdot \mathbf{d}_\ell - 1)(\Pi^{\nabla,K}_{p,1}\phi_r^K)\overline{(\Pi^{\nabla,K}_{p,1}\phi_s^K)} w_j^K \overline{w_\ell^K} \right].$$

In Figure 5.31, we plot the dispersion error $|\mathrm{Re}(k - k_n)|$ for the three stabilizations, $k = 3$ and $q = 6$, on the meshes made of squares and triangles. The dissipation is zero, up to machine precision, in all cases and is thus not shown. One can observe a different behavior between *stab 1* and the other stabilizations.

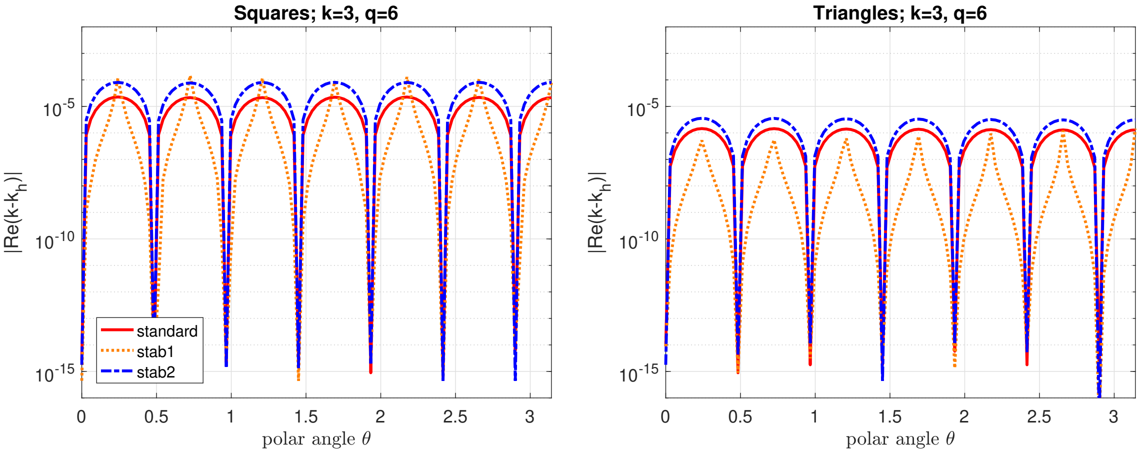


Figure 5.31: Dispersion error for PWVEM with different stabilizations in dependence on the polar angle θ of the Bloch wave direction $\mathbf{d}$ in (5.46) for fixed $k = 3$ and $q = 6$, on the meshes made of squares (*left*) and triangles (*right*) in Figure 5.24.

Exponential convergence of the dispersion error against the effective degree q

Here, we investigate the dependence of dispersion and dissipation on the effective plane wave degree q (namely, $p = 2q+1$ bulk plane waves). For fixed wave number k, we will observe exponential convergence of the total error for increasing q, as already seen in [110] for UWVF/PWDG. This result is not unexpected since also the p-versions for the discretization errors have exponential convergence, provided that the exact analytical solution is smooth; see [118] for UWVF/PWDG, and the numerical experiments in [159] and in Section 5.4 for PWVEM and ncTVEM, respectively. Moreover, we will make a comparison of these methods in terms of the total error versus the dimensions of the minimal generating subspaces.

To this purpose, we consider the following range for the wave number: $k \in \{2, 3, 4, 5\}$. We recall again that k here corresponds in fact to $k_0 = \frac{k}{h_0}$ on a mesh with mesh size h_0.

Dispersion and dissipation vs. effective degree q. In Figures 5.32-5.34, the relative dispersion error $|\mathrm{Re}(k - k_n)|/k$ and the relative damping error $|\mathrm{Im}(k_n)|/k$ are displayed against q, for the meshes made of squares, triangles, and hexagons, respectively. The maxima of the relative dispersion and the relative dissipation, respectively, are taken over a large set of Bloch wave directions $\mathbf{d}$. One can observe, after some preasymptotic regime, exponential convergence of the dispersion error for all methods, and of the dissipation error for UWVF/PWDG. Apart from some instabilities, the dissipation is close to machine precision for PWVEM. Furthermore, the dispersion error is consistently smaller for UWVF/PWDG than for PWVEM and ncTVEM.

Dispersion and dissipation vs. dimensions of minimal generating subspaces. From a computational point of view, it is also important to consider a comparison of the dispersion errors in terms of the dimensions of the minimal generating subspaces (density of the degrees of freedom). We directly compare the relative total errors $|k_n - k|/k$, thus measuring the total deviation of the discrete wave number from the continuous one. As above, the maxima over a large set of Bloch wave directions are taken. In Figure 5.35, those errors are displayed for the meshes in Figure 5.24. For ncTVEM, we can recognize the *cliff effect*, meaning that, at some point, the dispersion error decreases without increase of the dimension of the minimal generating subspace. Moreover, one can observe a direct correlation between the density of the degrees of freedom, which depends on

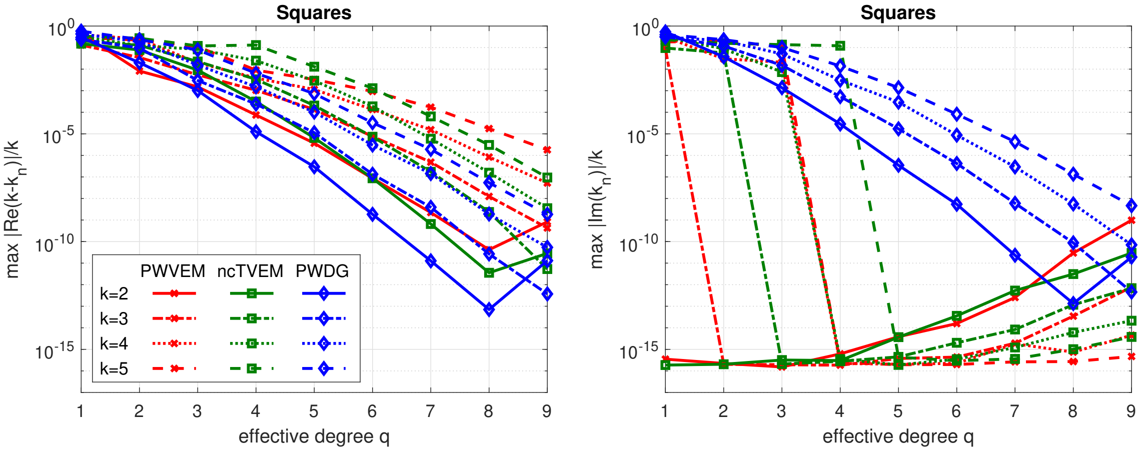


Figure 5.32: Relative dispersion (*left*) and relative dissipation (*right*) for the different methods in dependence on the effective degree q and the wave numbers $k = 2, \dots, 5$ on the meshes made of squares in Figure 5.24. The maxima over a large set of Bloch wave directions $\mathbf{d}$ are taken.

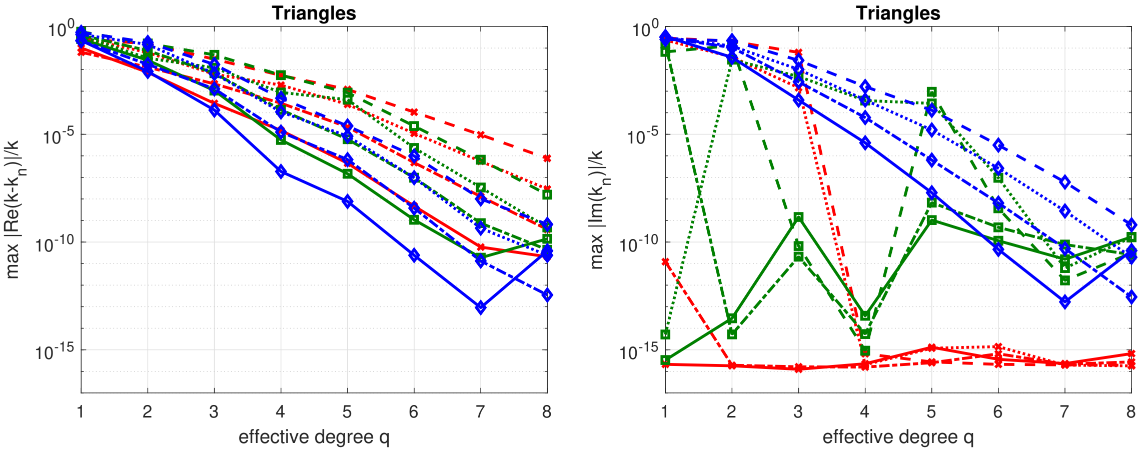


Figure 5.33: Relative dispersion (*left*) and relative dissipation (*right*) for the different methods in dependence on the effective degree q and the wave numbers $k = 2, \dots, 5$ on the meshes made of triangles in Figure 5.24. The maxima over a large set of Bloch wave directions $\mathbf{d}$ are taken. The color legend is the same as in Figure 5.32.

the shape of the meshes, see Figures 5.25-5.27, and the dispersion error plots (larger cardinalities of the fundamental sets lead to larger dispersion errors; as mentioned above, for ncTVEM, the filtering process leads to dimensionality reductions).

Comparison with standard FEM. Here, we highlight the advantages of using full Trefftz methods (ncTVEM, UWVF/PWDG) or methods that make use of Trefftz functions (PWVEM) in comparison to standard polynomial based methods, such as FEM, whose dispersion properties were studied in e.g. [4, 24, 86, 125]. We focus for simplicity on the meshes made of squares in Figure 5.24, since, in this case, the basis functions in FEM have a tensor product structure and an explicit dispersion relation can be derived [4, Theorem 3.1]:

$$\cos(k_n) = R_q(k), \tag{5.62}$$

where, denoting by $[\cdot/\cdot]_{z\cot z}$ and $[\cdot/\cdot]_{z\tan z}$ the Padé approximants to the functions $z\cot z$ and $z\tan z$, respectively,

$$R_q(2z) := \frac{[2N_0/2N_0-2]_{z\cot z} - [2N_e+2/2N_e]_{z\tan z}}{[2N_0/2N_0-2]_{z\cot z} + [2N_e+2/2N_e]_{z\tan z}},$$

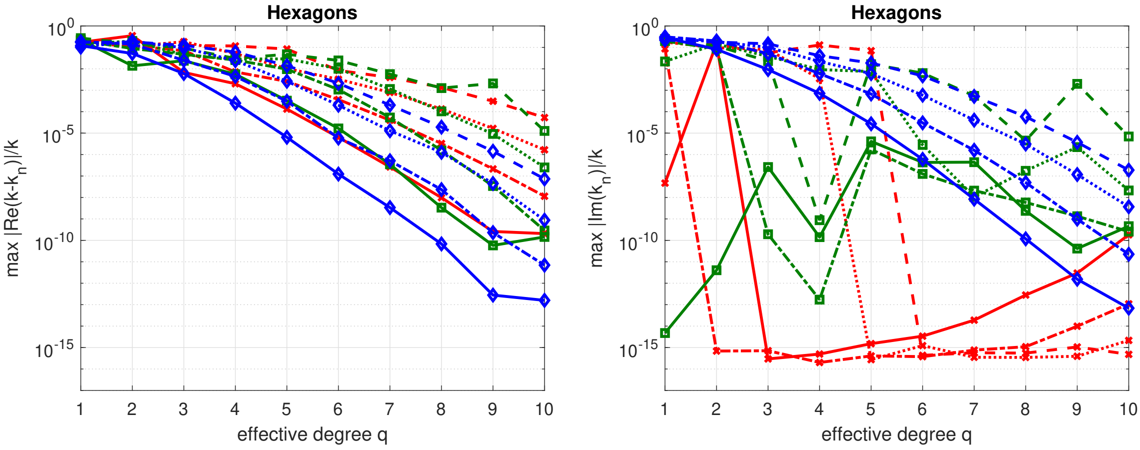


Figure 5.34: Relative dispersion (*left*) and relative dissipation (*right*) for the different methods in dependence on the effective degree q and the wave numbers $k = 2, \ldots, 5$ on the meshes made of hexagons in Figure 5.24. The maxima over a large set of Bloch wave directions $\mathbf{d}$ are taken. The color legend is the same as in Figure 5.32.

with $N_0 := \lfloor (q+1)/2 \rfloor$ and $N_e := \lfloor q/2 \rfloor$. From (5.62), one can see that only dispersion plays a role in FEM. In Figure 5.36, we display the relative total dispersion errors against the effective degree q (left) and against the dimensions of the minimal generating subspaces (right) for fixed $k = 3$. Similar results are obtained for other values of k and are not shown. One can clearly notice that the dispersion error for FEM is lower than for the other methods, when comparing it in terms of q, but higher, when comparing it in terms of the dimensions of the minimal generating subspaces.

Algebraic convergence of the dispersion error against the wave number k

We study the dispersion and dissipation properties of the three methods with respect to the wave number k. Due to the fact that $h = 1$, and k is related to the wave number k_0 on a mesh with mesh size h_0 by $k = kh = k_0 h_0$, the limit $k \to 0$ corresponds in fact to an h-version with $h_0 \to 0$ for fixed k_0. We will observe algebraic convergence of the total dispersion error in terms of k. This mimics the algebraic convergence of the discretization error in the h-version, proven in [159], in Section 4.2.2, and in [112], for PWVEM, ncTVEM, and UWVF/PWDG, respectively.

For the numerical experiments, we fix the effective degrees $q = 3, 5, 7$. We employ once again the meshes made of squares and triangles in Figure 5.24. Similar results have been obtained on the mesh made of hexagons. In Figure 5.37, the relative total errors $|k - k_n|/k$ determined over a large set of Bloch wave directions $\mathbf{d}$ are depicted against k. Algebraic convergence can be observed. Furthermore, larger values of q lead to smaller errors. The peaks occurring in the convergence regions of PWVEM and ncTVEM could be related to the presence of Neumann eigenvalues, and Dirichlet and Neumann eigenvalues, that have to be excluded in the construction of ncTVEM and PWVEM, respectively, in order to have a well-posed variational formulation, see Sections 4.1.2 and 4.1.3 for ncTVEM, and [159] for PWVEM. Moreover, the oscillations for larger and smaller values of k are related to the pre-asymptotic regime and the instability regime, which are typical of wave based methods.

In Table 5.5, we list some relative total errors for different values of k. They indicate a convergence behavior of

$$\max \frac{|k - k_n|}{|k|} \approx \mathcal{O}(k^\eta), \quad k \to 0, \tag{5.63}$$

where $\eta \in [2q-1, 2q]$. This was already observed in [110] for UWVF/PWDG .

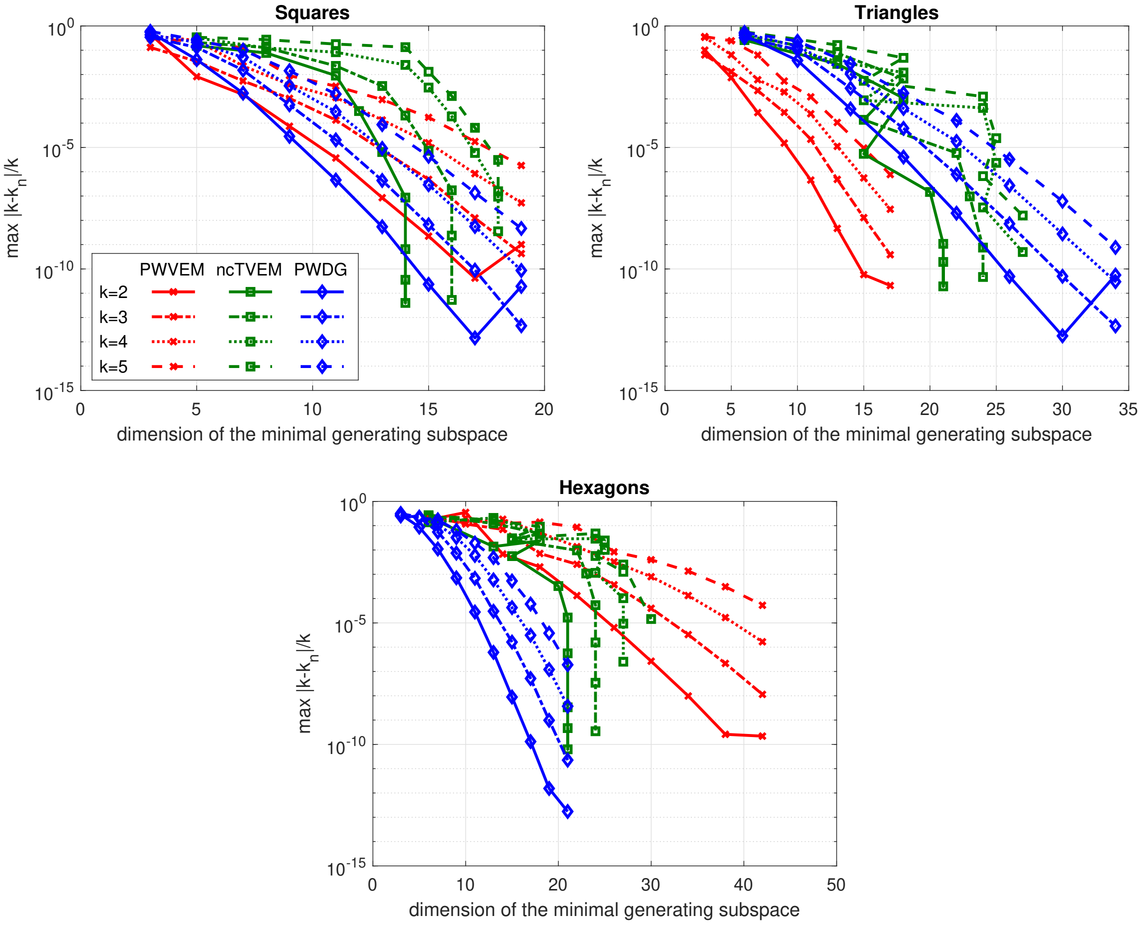


Figure 5.35: Relative total dispersion error in dependence on the dimensions of the minimal generating subspaces for different values of k on the meshes in Figure 5.24.

	method	squares					triangles				
		k	$\frac{\lvert k-k_n\rvert}{k}$	k	$\frac{\lvert k-k_n\rvert}{k}$	rate	k	$\frac{\lvert k-k_n\rvert}{k}$	k	$\frac{\lvert k-k_n\rvert}{k}$	rate
$q=3$	PWVEM	2	1.50e-03	0.3	4.59e-08	5.48	2	2.71e-04	0.3	3.42e-09	5.95
	ncTVEM	2	9.04e-03	0.3	3.69e-07	5.33	2	1.07e-03	0.3	4.09e-08	5.36
	PWDG	2	1.71e-03	0.3	1.04e-07	5.11	2	3.87e-04	0.3	3.04e-08	4.98
$q=5$	PWVEM	2	3.68e-06	0.8	5.09e-10	9.70	3	2.17e-05	2	4.54e-07	9.53
	ncTVEM	2	6.48e-06	0.8	1.21e-09	9.37	3	5.91e-06	2	1.47e-07	9.11
	PWDG	2	4.56e-07	0.8	1.47e-10	8.77	3	7.75e-07	2	1.97e-08	9.06
$q=7$	PWVEM	4	1.55e-05	2	2.23e-09	12.76	6	7.79e-05	4	5.57e-07	12.19
	ncTVEM	4	5.93e-06	2	6.54e-10	13.15	6	6.01e-06	4	3.39e-08	12.77
	PWDG	4	2.92e-07	2	2.33e-11	13.62	6	7.10e-07	4	2.76e-09	13.69

Table 5.5: Rates of the relative total error for $k \to 0$.

Remark 20. Clearly, similarly as above, dispersion and dissipation can be investigated again separately from each other. Here, we only show the results, depicted in Figure 5.38, for fixed $q = 5$ and varying k on the meshes made of squares. As already observed, one can deduce that ncTVEM and PWVEM are dispersion-dominated, whereas dissipation plays a major role for UWVF/PWDG.

The results of this section can be summarized as follows:

- Dispersion and dissipation hinge upon the choice of the Bloch wave direction.
- There is a link between dispersion and dissipation, and the level of conformity. Whereas the dissipation error is zero (up to machine precision) in the convergence regime for conforming

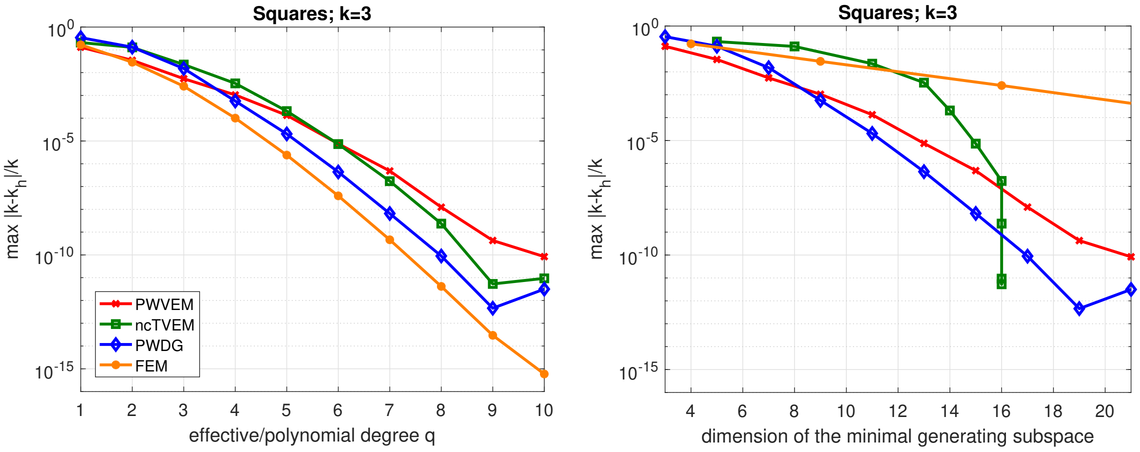


Figure 5.36: Comparison of the relative total errors for ncTVEM, PWVEM, UWVF/PWDG, and the standard polynomial based FEM on a mesh made of squares as in Figure 5.24 for fixed wave number $k = 3$, in dependence on the effective/polynomial degree q (*left*) and the dimension of the minimal generating subspaces (*right*). The maxima over a large set of Bloch wave directions $\mathbf{d}$ are taken.

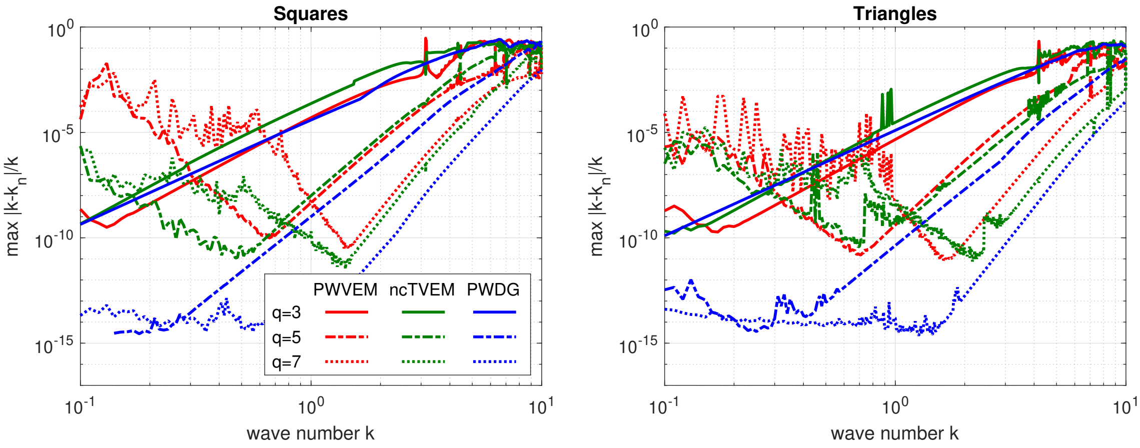


Figure 5.37: Relative total dispersion in dependence on the wave number k for fixed effective degrees $q = 3, 5, 7$. The maxima over a large set of Bloch wave directions $\mathbf{d}$ are taken. As meshes, those made of squares (*left*) and triangles (*right*) in Figure 5.24 are employed.

methods, such as PWVEM and FEM, it is much larger and even dominates the dispersion error for the fully discontinuous UWVF/PWDG. For ncTVEM, dispersion dominates dissipation, and the dissipation error is in general not zero, but is in most cases lower than for UWVF/PWDG.

- The dispersion error depends on the choice of stabilization.
- We observed for all methods exponential convergence of the relative total error with respect to the effective plane wave degree q, for $q \to \infty$.
- Moreover, the dispersion error is consistently smaller for UWVF/PWDG than for PWVEM and ncTVEM, when measured in terms of q, however, when compared to the dimensions of the minimal generating subspaces, the results depend on the element geometry, and thus on the density of the degrees of freedom.
- Concerning the comparison of the total error with respect to the wave number k, as $k \to 0$, algebraic convergence was observed. There, larger values of q lead to smaller errors.

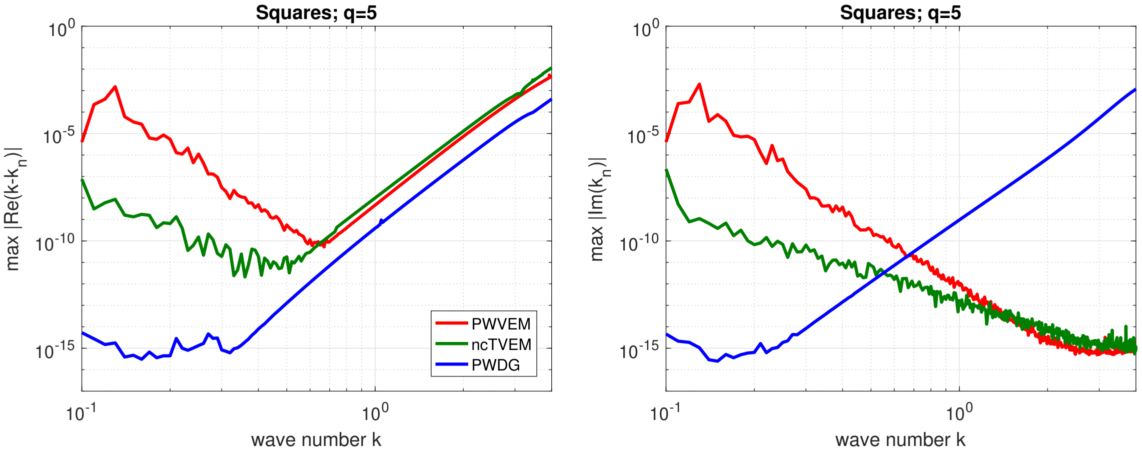


Figure 5.38: Relative dispersion (*left*) and relative dissipation (*right*) in dependence on the wave number k for fixed $q = 5$ on the meshes made of squares in Figure 5.24. The maxima over a large set of Bloch wave directions $\mathbf{d}$ are taken.

- Finally, the comparison with the standard polynomial based FEM highlighted the advantages of employing Trefftz based methods, such as ncTVEM and UWVF/PWDG, or methods that make use of Trefftz functions, like PWVEM, over standard polynomial based methods.

Chapter 6

Trefftz virtual element method for the fluid-fluid interface problem

In this chapter, we extend the nonconforming Trefftz VEM introduced in Chapters 4 and 5 to the case of the fluid-fluid interface problem, that is, a Helmholtz problem with piecewise constant wave number. More precisely, given a polygonal domain $\Omega \subset \mathbb{R}^2$, a piecewise (real-valued) constant wave number $\mathfrak{k} \in L^\infty(\Omega)$, and $g \in H^{-\frac{1}{2}}(\partial\Omega)$, we aim to approximate the solution to the problem

$$\begin{cases} -\Delta u - \mathfrak{k}^2 u = 0 & \text{in } \Omega \\ \nabla u \cdot \mathbf{n}_\Omega + \mathrm{i}\mathfrak{k} u = g & \text{on } \partial\Omega, \end{cases} \tag{6.1}$$

where we recall that $\mathbf{n}_\Omega$ denotes the unit normal vector on $\partial\Omega$ pointing outside Ω. With respect to the original approach, we address two additional issues: firstly, we define the coupling of local approximation spaces with piecewise constant wave numbers; secondly, we enrich such local spaces with special functions capturing the physical behavior of the solution to the target problem in the spirit of [141] and [170] for PWDG and the discontinuous enrichment method, respectively. We will see that this can be done in a natural fashion by simply supplementing the edge spaces with the corresponding traces of the functions and then applying Algorithm 2 to eliminate redundant basis functions and mitigate the strong ill-conditioning.

The outline of this chapter is as follows. After giving a more detailed description of the model problem in Section 6.1, we design a corresponding nonconforming Trefftz VEM in Section 6.2. Finally, in Section 6.3, numerical aspects are discussed and numerical results are presented.

The material of this chapter has been published in [146].

6.1 The fluid-fluid interface problem

In this section, we give a closer look at the model problem to be considered. Starting from (6.1), the corresponding weak formulation reads

$$\begin{cases} \text{find } u \in V := H^1(\Omega) \,\text{such that} \\ b_\mathfrak{k}(u,v) = F(v) \quad \forall v \in V, \end{cases} \tag{6.2}$$

where the sesquilinear form $b_\mathfrak{k}(\cdot,\cdot)$ is given by

$$b_\mathfrak{k}(u,v) := a_\mathfrak{k}(u,v) + \mathrm{i}(\mathfrak{k}u, v)_{0,\partial\Omega} \tag{6.3}$$

with

$$a_\mathfrak{k}(u,v) := \int_\Omega \nabla u \cdot \overline{\nabla v} \,\mathrm{d}x - \int_\Omega \mathfrak{k}^2 u \overline{v} \,\mathrm{d}x,$$

and the right-hand side is defined as

$$F(v) := \int_{\partial\Omega} g\overline{v} \,\mathrm{d}s. \tag{6.4}$$

Well-posedness of the problem (6.2) can be proven as in e.g. [115, Theorem 2.4].

For the sake of simplicity, we will assume in the following that the domain $\Omega = (-1,1)^2$ is split into two parts $\Omega_1 := (-1,1) \times (-1,0)$ and $\Omega_2 := (-1,1) \times (0,1)$, and that the wave number $\mathfrak{k}$ is piecewise constant over Ω_1 and Ω_2; more precisely, we set $k_i := \mathfrak{k}_{|\Omega_i} = n_i k$, $i = 1,2$, where $k \in \mathbb{R}$, and n_1, $n_2 \in \mathbb{R}$ with $n_1 > n_2$ are the so-called *refraction indices*. The more general situation with multiple refraction indices and subdomains is a straightforward modification of this case.

Denoting by $\Gamma := \partial\Omega_1 \cap \partial\Omega_2$ this time the interface between the two subdomains with fixed unit normal vector $\mathbf{n}_\Gamma$, problem (6.1) can be reformulated as the transmission problem

$$\begin{cases} \text{find } u_i \in H^1(\Omega_i),\ i = 1,2, \text{ such that} & \\ \qquad -\Delta u_i - k_i^2 u_i = 0 & \text{in } \Omega_i, \quad i = 1,2 \\ \nabla u_i \cdot \mathbf{n}_{\Omega_i} + \mathrm{i} k_i u_i = g & \text{on } \partial\Omega_i \setminus \Gamma, \quad i = 1,2 \\ \qquad u_1 = u_2 & \text{on } \Gamma \\ \qquad \nabla u_1 \cdot \mathbf{n}_\Gamma = \nabla u_2 \cdot \mathbf{n}_\Gamma & \text{on } \Gamma. \end{cases} \tag{6.5}$$

This model goes under the name of *fluid-fluid interface problem.* From a physical standpoint, it describes the propagation of waves through a domain split into two subdomains containing different fluids (e.g. water-air). Typically, some reflection/transmission phenomenon occurs at the interface Γ. For instance, assuming that there is an incoming traveling plane wave in Ω_1 with incident angle θ_{inc} formed by the direction of the incoming wave with the interface Γ, the model describes the propagation of such wave from Ω_1 to Ω_2. Depending on the angle θ_{inc}, a different behaviour may occur in Ω_1 and Ω_2.

In order to describe the two possible outcomes, we introduce the so-called *critical angle*

$$\theta_{\text{crit}} := \cos^{-1}\left(\frac{n_2}{n_1}\right). \tag{6.6}$$

If $\theta_{\text{inc}} \geq \theta_{\text{crit}}$, the incoming wave is partially refracted at Γ with angle θ_R (same measure as θ_{inc}) and transmitted in Ω_2 with transmission angle θ_{T}, computed by means of Snell's law

$$n_1 \cos(\theta_{\text{inc}}) = n_2 \cos(\theta_{\text{T}}).$$

Otherwise, if $\theta_{\text{inc}} < \theta_{\text{crit}}$, the incoming wave is totally refracted (with angle θ_{R}, having again the same measure as θ_{inc}); however, in the subdomain Ω_2 some evanescent modes, decaying exponentially with increasing distance from the interface Γ, appear. This phenomenon is known in the literature as *total internal reflection.* In Figure 6.1, the two different situations depending on the choice of θ_{inc} are depicted. A couple of explicit solutions to the problem (6.2) in the transmission and the total internal reflection cases, respectively, are described in Section 6.3.2.

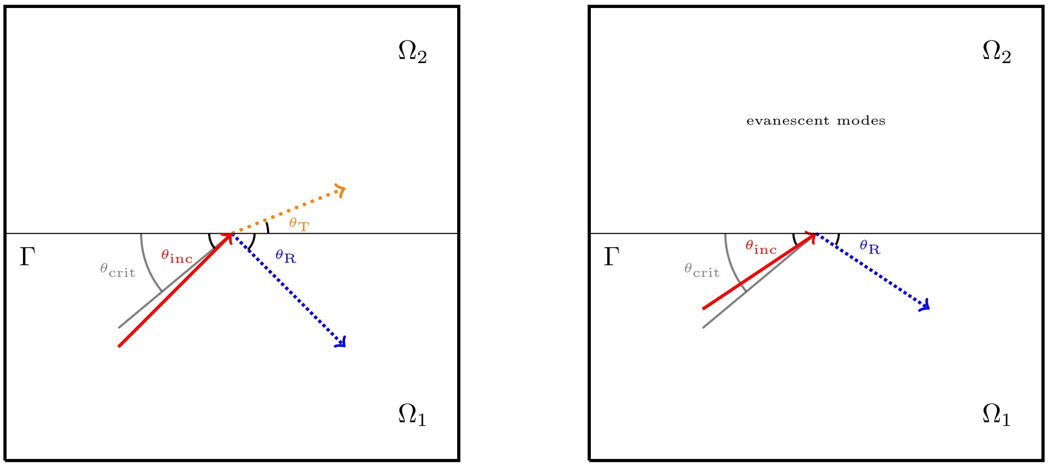


Figure 6.1: *Left:* $\theta_{\text{inc}} \geq \theta_{\text{crit}}$. The incoming wave is partially refracted at Γ and partially transmitted in form of a plane wave with direction given by the angle θ_{T} in Ω_2. *Right:* $\theta_{\text{inc}} < \theta_{\text{crit}}$. The incoming wave is totally refracted; only evanescent modes appear in Ω_2. *Legend:* the directions of the incident, the reflected, and the transmitted plane waves are straight red, dashed blue, and dotted orange, respectively. The critical angle θ_{crit} is depicted in grey.

6.2 Nonconforming Trefftz virtual element methods

Let $\mathcal{T}_n$ be a conforming mesh with respect to the interface Γ, i.e. $\mathcal{T}_n = \mathcal{T}_n^1 \cup \mathcal{T}_n^2$, where $\mathcal{T}_n^1$ and $\mathcal{T}_n^2$ are regular polygonal decompositions of Ω_1 and Ω_2 in the sense of Section 2.3. In particular, we do not consider here the case of elements cut by Γ and of elements on which $\mathfrak{k}$ may vary. However, hanging nodes are allowed. The case of meshes that are not conforming with respect to the interface is discussed in Section 6.3.2.

Having this, our goal is to design a Trefftz VEM of the following structure:

$$\begin{cases} \text{find } u_h \in V_h^{\Delta+\mathfrak{k}^2} \text{ such that} \\ b_{\mathfrak{k},h}(u_h, v_h) = F_h(v_h) \quad \forall v_h \in V_h^{\Delta+\mathfrak{k}^2}, \end{cases} \tag{6.7}$$

where $V_h^{\Delta+\mathfrak{k}^2}$ is a finite dimensional space, $b_{\mathfrak{k},h}(\cdot,\cdot) : [V_h^{\Delta+\mathfrak{k}^2}]^2 \to \mathbb{C}$ is a computable sesquilinear form mimicking its continuous counterpart $b_{\mathfrak{k}}(\cdot,\cdot)$ defined in (6.3), and $F_h(\cdot) : V_h^{\Delta+\mathfrak{k}^2} \to \mathbb{C}$ is a computable functional mimicking its continuous counterpart $F(\cdot)$ in (6.4).

The remainder of the section is organized as follows. After defining spaces of planes waves and evanescent waves in Section 6.2.1, we pinpoint the local and global nonconforming Trefftz VE spaces, together with a set of unisolvent degrees of freedom, in Section 6.2.2. Next, in Section 6.2.3, we introduce a couple of local (bulk and edge) projectors from local VE spaces into proper (plane/evanescent) wave spaces. Such operators, in addition to proper suitable stabilizations, are instrumental for the construction of the discrete sesquilinear form $b_{\mathfrak{k},h}(\cdot,\cdot)$ and right-hand side $F_h(\cdot)$ in (6.7), which is the topic of Section 6.2.4.

6.2.1 Plane waves and evanescent waves

In this section, we introduce spaces of plane waves and evanescent waves over elements and edges.

We firstly focus on the bulk spaces, namely plane wave based spaces over the elements in Ω_1, and spaces based on both plane waves and evanescent waves over the elements contained in Ω_2. The choice for the latter spaces is inspired by [141,170], where evanescent waves were added as special functions to the standard plane wave and Bessel spaces, respectively, to capture the evanescent modes occurring in specific situations described in Section 6.1. We anticipate that variants of such spaces are possible and will be discussed in Section 6.3.

To start with, given $K \in \mathcal{T}_n^1$ and a bunch of equidistributed normalized directions $\{\mathbf{d}_\ell^K\}_{\ell=1}^{p^K}$, $p^K = 2q^K + 1$, $q^K \in \mathbb{N}$, we denote, analogously to (4.6), by

$$w_\ell^{(1),K}(\mathbf{x}) := e^{\mathrm{i}k_1 \mathbf{d}_\ell^K \cdot (\mathbf{x} - \mathbf{x}_K)}{}_{|K} \tag{6.8}$$

the bulk plane wave with wave number k_1 and traveling along the directions $\mathbf{d}_\ell^K$. The bulk plane wave space over K is

$$\mathbb{PW}_{p^K}^{(1)}(K) := \operatorname{span}\left\{ w_\ell^{(1),K} \mid \ell = 1, \dots, p^K \right\}. \tag{6.9}$$

Note that we allow here for elementwise different numbers of plane waves; this notation is particularly suitable for developing an hp-version of the method in the spirit of Remark 17. For more details, we refer to Section 6.3.2.

Next, for all $K \in \mathcal{T}_n^2$, we define the bulk plane wave space $\mathbb{PW}_{p^K}^{(2)}(K)$ as the span of the plane waves $w_\ell^{(2),K}$, defined analogously to $w_\ell^{(1),K}$ in (6.8), but with wave number k_2 instead of k_1.

Following [141,170], we introduce a set of $\widetilde{p}^K = 2\widetilde{q}^K$, $\widetilde{q}^K \in \mathbb{N}_0$, evanescent waves, for all $K \in \mathcal{T}_n^2$. To this purpose, we firstly consider the set of equidistributed angles

$$\theta_{\widetilde{\ell}}^{\mathrm{EW}} = \frac{\widetilde{\ell}}{\widetilde{q}^K + 1} \theta_{\mathrm{crit}} \quad \forall \widetilde{\ell} = 1, \dots, \widetilde{q}^K,$$

where we recall that the critical angle θ_{crit} is computed as in (6.6). Then, the evanescent waves over K are defined, for all $j = 1, \dots, \widetilde{q}^K$, as

$$w_j^{\mathrm{EV},K}(\mathbf{x}) := e^{\mathrm{i}k \widehat{\mathbf{d}}_{\frac{j}{2}}^K \cdot (\mathbf{x} - \mathbf{x}_K)}{}_{|K}, \tag{6.10}$$

where k is the real number with $k_1 = n_1 k$ and $k_2 = n_2 k$, and $\widehat{\mathbf{d}}_{\frac{j}{2}}^K \in \mathbb{R} \times \mathbb{C}$ is given by

$$\widehat{\mathbf{d}}_{\frac{j}{2}}^K := \begin{cases} \left(-n_1 \cos\left(\theta^{\mathrm{EW}}_{\lceil \frac{j}{2} \rceil}\right), \mathrm{i}\sqrt{n_1^2 \cos\left(\theta^{\mathrm{EW}}_{\lceil \frac{j}{2} \rceil}\right)^2 - n_2^2}\right) & \text{if } j \text{ odd} \\ \left(n_1 \cos\left(\theta^{\mathrm{EW}}_{\frac{j}{2}}\right), \mathrm{i}\sqrt{n_1^2 \cos\left(\theta^{\mathrm{EW}}_{\frac{j}{2}}\right)^2 - n_2^2}\right) & \text{if } j \text{ even.} \end{cases} \tag{6.11}$$

Remark 21. Note that, similarly as above, the assumption of having sets of equidistributed directions and angles in the construction of the plane and evanescent wave spaces, respectively, is made for the sake of simplicity and could be relaxed in principle, without jeopardizing the approximation properties of the space of interest.

As one can notice from (6.10) and (6.11), the structure of an evanescent wave is similar to that of a plane wave; the difference is that the direction vector is complex-valued in the former case, whereas it is real-valued in the latter. As discussed and numerically proven in [141, 170], the evanescent waves are better suited than plane waves to capture the exponential decay of the evanescent modes appearing in the fluid-fluid interface problem for specific incident angles θ_{inc}, and therefore they could be added to the approximation space associated with the domain Ω_2 to improve the performance of the method.

We point out that the evanescent waves given by (6.10) satisfy the homogeneous Helmholtz problem in Ω_2. In Figure 6.2, we plot the real and imaginary part of the evanescent wave with parameters $k = 5$, $n_1 = 2$ and $n_2 = 1$ (critical angle $\theta_{\mathrm{crit}} = 60°$), and $\mathbf{x}_K = (0, 0)$.

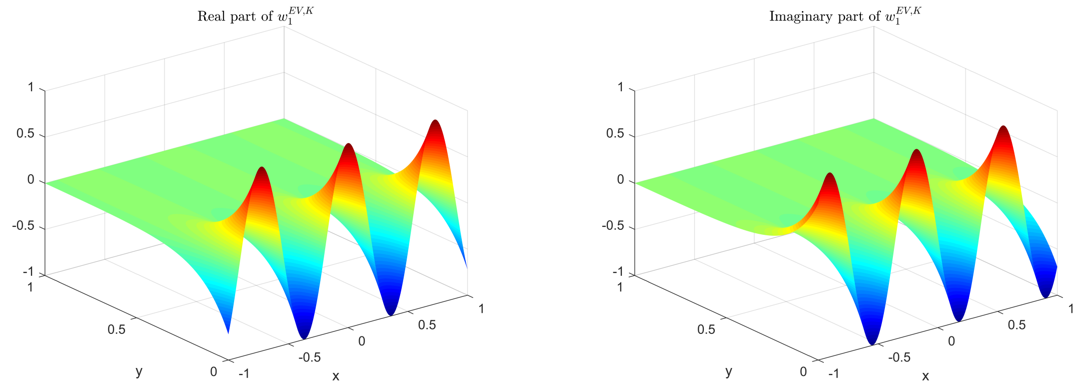


Figure 6.2: Real and imaginary parts of the first evanescent wave for $k = 5$, $n_1 = 2$, $n_2 = 1$, and $\mathbf{x}_K = (0, 0)$.

Finally, we define the space of *evanescent waves* over $K \in \mathcal{T}_n^2$ by

$$\mathbb{EW}_{\widetilde{p}^K}(K) := \operatorname{span}\left\{ w_j^{\mathrm{EV},K} \mid j = 1, \ldots, \widetilde{p}^K \right\},$$

and the space of plane waves *and* evanescent waves by

$$\widetilde{\mathbb{PW}}^{(2)}_{p^K, \widetilde{p}^K}(K) := \mathbb{PW}^{(2)}_{p^K}(K) \cup \mathbb{EW}_{\widetilde{p}^K}(K). \tag{6.12}$$

In the following, we shall also need spaces of traces of plane waves and evanescent waves over edges. To this purpose, we firstly need to fix some additional notation. We write $\mathcal{E}_n^{1,I}$ and $\mathcal{E}_n^{1,B}$ for the sets of interior edges in $\mathcal{T}_n^1$, and boundary edges in $\mathcal{T}_n^1$ not belonging to Γ, respectively. Similarly, we introduce the sets $\mathcal{E}_n^{2,I}$ and $\mathcal{E}_n^{2,B}$ for $\mathcal{T}_n^2$. The symbol $\mathcal{E}_n^\Gamma$ denotes the set of edges of $\mathcal{T}_n$ on Γ. Finally, we define $\mathcal{E}_n^I := \mathcal{E}_n^{1,I} \cup \mathcal{E}_n^{2,I}$, $\mathcal{E}_n^B := \mathcal{E}_n^{1,B} \cup \mathcal{E}_n^{2,B}$, and $\mathcal{E}_n := \mathcal{E}_n^I \cup \mathcal{E}_n^B \cup \mathcal{E}_n^\Gamma$.

For all edges $e \in \mathcal{E}_n$, we set

$$\widetilde{\mathbb{PW}}_{p_e}(e) := \begin{cases} \mathbb{PW}^{(1)}_{p^K}(K)_{|e}, & \text{if } e \in \mathcal{E}_n^{1,B} \cap \mathcal{E}^K \\ \mathbb{PW}^{(1)}_{p^{K^-}}(K^-)_{|e} \cup \mathbb{PW}^{(1)}_{p^{K^+}}(K^+)_{|e}, & \text{if } e \in \mathcal{E}_n^{1,I} \cap \mathcal{E}^{K^-} \cap \mathcal{E}^{K^+} \text{ with } \\ & \quad K^+,\, K^- \in \mathcal{T}_n^1,\, K^- \neq K^+ \\ \widetilde{\mathbb{PW}}^{(2)}_{p^K,\widetilde{p}^K}(K)_{|e} & \text{if } e \in \mathcal{E}_n^{2,B} \cap \mathcal{E}^K \\ \widetilde{\mathbb{PW}}^{(2)}_{p^{K^-},\widetilde{p}^{K^-}}(K^-)_{|e} \cup \widetilde{\mathbb{PW}}^{(2)}_{p^{K^+},\widetilde{p}^{K^+}}(K^+)_{|e} & \text{if } e \in \mathcal{E}_n^{2,I} \cap \mathcal{E}^{K^-} \cap \mathcal{E}^{K^+} \text{ with } \\ & \quad K^+,\, K^- \in \mathcal{T}_n^2,\, K^- \neq K^+ \\ \mathbb{PW}^{(1)}_{p^{K^-}}(K^-)_{|e} \cup \widetilde{\mathbb{PW}}^{(2)}_{p^{K^+},\widetilde{p}^{K^+}}(K^+)_{|e}, & \text{if } e \in \mathcal{E}_n^{\Gamma} \cap \mathcal{E}^{K^-} \cap \mathcal{E}^{K^+} \text{ with } \\ & \quad K^- \in \mathcal{T}_n^1,\, K^+ \in \mathcal{T}_n^2, \end{cases} \tag{6.13}$$

denoting by p_e the dimension of the space $\widetilde{\mathbb{PW}}_{p_e}(e)$.

In words, we consider spaces of traces of plane waves with wave number k_1 on all edges in $\mathcal{E}_n^{1,I} \cup \mathcal{E}_n^{1,B}$, spaces of traces of plane waves with wave number k_2 and evanescent waves on all edges in $\mathcal{E}_n^{2,I} \cup \mathcal{E}_n^{2,B}$, and, at the interface Γ, we consider traces of plane waves with the two different wave numbers k_1 and k_2 and evanescent waves. The definition (6.13) will be instrumental to build suitable nonconforming Sobolev spaces.

Remark 22. Whilst the dimensions of the bulk plane wave spaces $\mathbb{PW}^{(1)}_{p^K}(K)$ and $\widetilde{\mathbb{PW}}^{(2)}_{p^K,\widetilde{p}^K}(K)$ are given by p^K and $p^K + \widetilde{p}^K$, respectively, those of the spaces $\widetilde{\mathbb{PW}}_{p_e}(e)$ are in general smaller than or equal to the sum of the dimensions of the bulk spaces of the adjacent polygons. In fact, the restriction of two different plane waves onto a given edge could generate a 1D space only, see Figure 4.2. On the other hand, whenever the restrictions of two plane waves with different directions and wave numbers on a given edge are "close", numerical instabilities may occur, see also Section 5.4.1. In order to avoid this situation, we will employ the edgewise orthogonalization-and-filtering process introduced in Algorithm 2.

6.2.2 Nonconforming Trefftz virtual element spaces

Our aim here is to introduce local Trefftz VE spaces tailored for the fluid-fluid interface problem (6.2), and subsequently to patch them into a global space in a nonconforming fashion. To this end, we specify $V_h^{\Delta+\mathfrak{k}^2}$ in (6.7).

For every $K \in \mathcal{T}_n^1$, let $q_K \in \mathbb{N}$ be a fixed effective plane wave degree (namely, $2q_K + 1$ plane waves). Similarly, for all $K \in \mathcal{T}_n^2$, let $q_K + \widetilde{q}_K$ be the total effective degree, where $q_K \in \mathbb{N}$ and $\widetilde{q}_K \in \mathbb{N}_0$ are the effective plane wave (namely, $2q_K + 1$ plane waves) and effective evanescent wave (namely, $2\widetilde{q}_K$ evanescent waves) degrees, respectively. Moreover, to the set of edges $\mathcal{E}_n$, we associate a vector $\mathbf{p}_{\mathcal{E}_n} \in \mathbb{N}^{\mathrm{card}(\mathcal{E}_n)}$, whose j-th entry represents the dimension of the space $\widetilde{\mathbb{PW}}_{p_e}(e)$ defined in (6.13) on the j-th global edge e.

Local Trefftz VE spaces. Given $K \in \mathcal{T}_n$, we define the local Trefftz VE space

$$V^{\Delta+\mathfrak{k}^2}(K) := \{v_h \in H^1(K) \mid \Delta v_h + \mathfrak{k}^2 v_h = 0 \text{ in } K,\ (\nabla v_h \cdot \mathbf{n}_K + \mathrm{i}\mathfrak{k} v_h)_{|e} \in \widetilde{\mathbb{PW}}_{p_e}(e)\, \forall e \in \mathcal{E}^K\}, \tag{6.14}$$

where we recall that the edge spaces $\widetilde{\mathbb{PW}}_{p_e}(e)$ are defined in (6.13).

We point out that, for every element $K \in \mathcal{T}_n^1$, the space $V^{\Delta+\mathfrak{k}^2}(K)$ contains $\mathbb{PW}^{(1)}_{p^K}(K)$, the space of $p_K = 2q_K + 1$ plane waves with wave number k_1 defined in (6.9); besides, it contains additional functions that are not known in closed form (whence *virtual* functions) and that are locally Trefftz with impedance traces in the space $\widetilde{\mathbb{PW}}_{p_e}(e)$, for all edges $e \in \mathcal{E}^K$.

On the other hand, the local spaces over the elements $K \in \mathcal{T}_n^2$ are designed in such a way that they contain $\widetilde{\mathbb{PW}}^{(2)}_{p^K,\widetilde{p}^K}(K)$, the space of $p_K = 2q_K + 1$ plane waves with wave number k_2 and $\widetilde{p}^K = 2\widetilde{q}_K$ evanescent waves defined in (6.12); again, there are additional functions unknown in closed form inside (which however have impedance traces in the space of traces of plane and evanescent waves). Such additional functions will be instrumental for building nonconforming global spaces, as described below.

Associated to $V^{\Delta+\mathfrak{k}^2}(K)$, we consider the following set of linear functionals. For all $e \in \mathcal{E}^K$,

$$\mathrm{dof}_{e,\alpha}(v_h) := \frac{1}{h_e} \int_e v_h \overline{w_\alpha^e} \,\mathrm{d}s \quad \forall \alpha = 1, \dots, p_e, \tag{6.15}$$

where $\{w_\alpha^e\}_{\alpha=1}^{p_e}$ is *any* basis for the space $\widetilde{\mathbb{PW}}_{p_e}(e)$. By requiring that $\mathfrak{k}_{|K}$ is not a Dirichlet-Laplace eigenvalue on K and following the lines in the proof of Lemma 4.1.1, one can show that this set of functionals constitutes in fact a set of unisolvent degrees of freedom. We recall that the assumption on $\mathfrak{k}_{|K}$ actually results in a condition on the size of the product $h_K \mathfrak{k}_{|K}$, see (4.16). Hence, for h_K sufficiently small, $\mathfrak{k}_{|K}$ is not a Dirichlet-Laplace eigenvalue on K.

Having this, we introduce the set of local canonical basis functions $\{\varphi_{\hat{e},\hat{\alpha}}\}_{\hat{e},\hat{\alpha}}$ by duality:

$$\mathrm{dof}_{e,\alpha}(\varphi_{\hat{e},\hat{\alpha}}) = \delta_{e,\hat{e}} \delta_{\alpha,\hat{\alpha}}, \quad \forall e, \hat{e} \in \mathcal{E}_n, \, \forall \alpha = 1, \dots, p_e, \, \forall \hat{\alpha} = 1, \dots, p_{\hat{e}},$$

where δ here denotes the standard Kronecker delta.

Global Trefftz VE spaces. First, similarly as in the preceding chapters, we define the global nonconforming Sobolev space associated with the vector $\mathbf{p}_{\mathcal{E}_n}$:

$$H^{1,nc}_{\mathbf{p}_{\mathcal{E}_n}}(\mathcal{T}_n) := \left\{ v \in H^1(\mathcal{T}_n) \mid \int_e [\![v]\!] \cdot \mathbf{n}^e \overline{w^e} \,\mathrm{d}s = 0 \quad \forall w^e \in \widetilde{\mathbb{PW}}_{p_e}(e), \, \forall e \in \mathcal{E}_n^I \right\}. \tag{6.16}$$

We highlight that by using this construction, nonconforming Sobolev spaces can be straightforwardly generalized to the case of piecewise constant $\mathfrak{k}$ on more than two subdomains.

Then, the global nonconforming Trefftz virtual element space is introduced by

$$V_h^{\Delta+\mathfrak{k}^2} := \{ v_h \in H^{1,nc}_{\mathbf{p}_{\mathcal{E}_n}}(\mathcal{T}_n) \mid v_{h|K} \in V^{\Delta+\mathfrak{k}^2}(K) \, \forall K \in \mathcal{T}_n \}, \tag{6.17}$$

and the global set of the degrees of freedom is built via a nonconforming coupling (*à la* Crouzeix-Raviart) of the local counterparts (6.15).

6.2.3 Local projectors

In this section, we introduce local projectors which will be instrumental for the design of the method (6.7). These are in fact adaptations of those in Sections 4.1.3 and 5.1.2.

First of all, for all $K \in \mathcal{T}_n^1$, we define the local operator $\Pi^{(1),K}_{p^K} : V^{\Delta+\mathfrak{k}^2}(K) \to \mathbb{PW}^{(1)}_{p^K}(K)$ by

$$a^K_{\mathfrak{k}}(\Pi^{(1),K}_{p^K} v_h, w^{(1),K}) = a^K_{\mathfrak{k}}(v_h, w^{(1),K}), \tag{6.18}$$

for all $v_h \in V^{\Delta+\mathfrak{k}^2}(K)$ and $w^{(1),K} \in \mathbb{PW}^{(1)}_{p^K}(K)$. Such operator is computable by means of the degrees of freedom (6.15). In fact, an integration by parts yields

$$a^K_{\mathfrak{k}}(v_h, w^{(1),K}) = \int_{\partial K} v_h \overline{\nabla w^{(1),K} \cdot \mathbf{n}_K} \,\mathrm{d}s,$$

which is computable since $(\nabla w^{(1),K} \cdot \mathbf{n}_K)_{|e} \in \widetilde{\mathbb{PW}}_{p_e}(e)$ for all $e \in \mathcal{E}^K$.

Besides, $\Pi^{(1),K}_{p^K}$ is well-defined under the assumption that the size of the element K is sufficiently small, see Proposition 4.1.2 for more details.

For all $K \in \mathcal{T}_n^2$, we also introduce the local projector $\Pi^{(2),K}_{p^K,\widetilde{p}^K} : V^{\Delta+\mathfrak{k}^2}(K) \to \widetilde{\mathbb{PW}}^{(2)}_{p^K,\widetilde{p}^K}(K)$ which is defined analogously to $\Pi^{(1),K}_{p^K}$ in (6.18) with the only difference that the space $\mathbb{PW}^{(1)}_{p^K}(K)$ is replaced by $\widetilde{\mathbb{PW}}^{(2)}_{p^K,\widetilde{p}^K}(K)$. Well-posedness of $\Pi^{(2),K}_{p^K,\widetilde{p}^K}$ is provided by the invertibility of the matrix

$\mathbf{G}^{(2),K} \in \mathbb{C}^{(p^K+\widetilde{p}^K)\times(p^K+\widetilde{p}^K)}$ defined by

$$\mathbf{G}^{(2),K}_{j,\ell} := \begin{cases} a^K_{\mathfrak{k}}(w^{(2),K}_\ell, w^{(2),K}_j), & \text{if } j,\ell \leqslant p^K \\ a^K_{\mathfrak{k}}(w^{EV,K}_{\ell-p^K}, w^{(2),K}_j), & \text{if } j \leqslant p^K, \ell > p^K \\ a^K_{\mathfrak{k}}(w^{(2),K}_\ell, w^{EV,K}_{j-p^K}), & \text{if } j > p^K, \ell \leqslant p^K \\ a^K_{\mathfrak{k}}(w^{EV,K}_{\ell-p^K}, w^{EV,K}_{j-p^K}), & \text{if } j,\ell > p^K \end{cases} \tag{6.19}$$

for all $j, \ell = 1, \dots, p^K + \widetilde{p}^K$. By investigating the behaviour of the minimal (absolute) eigenvalue of $\mathbf{G}^{(2),K}$ in terms of the wave number k_2 on the reference element $K := (0,1)^2$, one can observe that such a minimal eigenvalue becomes very small when k_2^2 is close to a Neumann-Laplace eigenvalue $\nu_{m,n} := \pi^2(m^2+n^2)$, $m, n \in \mathbb{N}_0$, on K, see Figure 6.3. This indicates that, assuming k_2^2 to be separated from the Neumann-Laplace eigenvalues, the local projector $\Pi^{(2),K}_{p^K,\widetilde{p}^K}$ is well-defined.

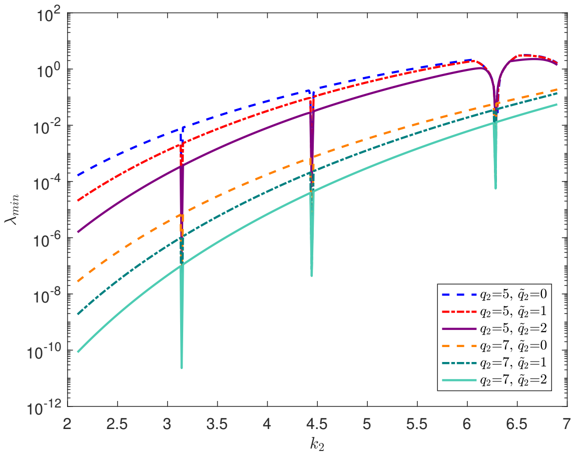


Figure 6.3: Minimal (absolute) eigenvalues of the matrix $\mathbf{G}^{(2),K}$ in (6.19) in terms of the wave number k_2 with $n_1 = 2$ and $n_2 = 1$. The effective plane and evanescent wave degrees are denoted by q_2 and $\widetilde{q}_2$, respectively.

The third operator we introduce is the boundary edge L^2 projector $\Pi^{0,e}_{p_e} : V^{\Delta+\mathfrak{k}^2}(K)_{|e} \to \widetilde{\mathbb{PW}}_{p_e}(e)$, which is defined for all edges $e \in \mathcal{E}^B_n$ by

$$(\Pi^{0,e}_{p_e} v_h, w^e)_{0,e} = (v_h, w^e)_{0,e},$$

for all $v_h \in V^{\Delta+\mathfrak{k}^2}(K)_{|e}$ and $w^e \in \widetilde{\mathbb{PW}}_{p_e}(e)$. Such a projector is directly computable from the local degrees of freedom in (6.15), and well-defined owing to the coercivity of the edge L^2 norm.

6.2.4 Discrete sesquilinear forms and right-hand side

Here, we specify the discrete sesquilinear form $b_{\mathfrak{k},h}(\cdot,\cdot)$ and the discrete right-hand side $F_h(\cdot)$ characterizing the method (6.7).

To begin with, we note again that the continuous counterparts $b_k(\cdot,\cdot)$ and $F(\cdot)$ in (6.3) and (6.4), respectively, are in general not explicitly computable when applied to functions in $V_h^{\Delta+\mathfrak{k}^2}$. Therefore, mimicking what was done in the previous chapters, we introduce, for all $K \in \mathcal{T}_n$, local computable stabilizing sesquilinear forms $S^K_{\mathfrak{k}}(\cdot,\cdot) : [\ker(\Pi^K)]^2 \to \mathbb{C}$, where Π^K is either $\Pi^{(1),K}_{p^K}$ or $\Pi^{(2),K}_{p^K,\widetilde{p}^K}$, depending on whether $K \in \mathcal{T}^1_n$ or $K \in \mathcal{T}^2_n$. Then, we propose a family of discrete sesquilinear forms $b_{\mathfrak{k},h}(\cdot,\cdot)$ defining method (6.7). More precisely, for all u_h, $v_h \in V_h^{\Delta+\mathfrak{k}^2}$, we set

$$b_{\mathfrak{k},h}(u_h, v_h) := \sum_{K\in\mathcal{T}_n} a^K_{\mathfrak{k},h}(u_{h|K}, v_{h|K}) + \mathrm{i}c^{\partial\Omega}_{\mathfrak{k},h}(\mathfrak{k}u_h, v_h),$$

where, for all $K \in \mathcal{T}_n$,

$$a^K_{\mathfrak{k},h}(u_h, v_h) := a^K_{\mathfrak{k}}(\Pi^K u_h, \Pi^K v_h) + S^K_{\mathfrak{k}}((I-\Pi^K)u_h, (I-\Pi^K)v_h) \tag{6.20}$$

with $\Pi^K = \Pi^{(1),K}_{p^K}$ for all $K \in \mathcal{T}_n^1$, and $\Pi^K = \Pi^{(2),K}_{p^K,\widetilde{p}^K}$ for all $K \in \mathcal{T}_n^2$, and where

$$c^{\partial\Omega}_{\mathfrak{k},h}(\mathfrak{k}u_h, v_h) := \sum_{e \in \mathcal{E}_n^B} (\mathfrak{k}\Pi^{0,e}_{p_e}(u_{h|e}), \Pi^{0,e}_{p_e}(v_{h|e}))_{0,e}.$$

Requirements on the stabilizations $S^K_{\mathfrak{k}}(\cdot,\cdot)$ in order to entail well-posedness and error estimates of the method (6.7) were discussed in Theorem 4.2.4. An explicit choice for $S^K_{\mathfrak{k}}(\cdot,\cdot)$ was provided in (5.32) and is recalled in (6.24) below.

The discrete right-hand side is defined, for all $v_h \in V_h^{\Delta+\mathfrak{k}^2}$, as

$$F_h(v_h) := \sum_{e \in \mathcal{E}_n^B} (g, \Pi^{0,e}_{p_e}(v_{h|e}))_{0,e}.$$

Note that the right-hand side is approximated by employing 1D quadrature formulas. In fact, this is the only occurrence where quadrature formulas are needed.

6.3 Numerical results

In this section, we firstly discuss some details of the implementation of method (6.7) in Section 6.3.1, and then, we present numerical experiments for a series of different test cases in Section 6.3.2.

6.3.1 Implementation aspects

The implementation of the method is performed analogously to the case of constant $\mathfrak{k}$, see Section 5.4.2. In particular, local matrices are computed and eventually patched into a global one.

Orthogonalization-and-filtering process. As discussed in detail in Section 5.3, due to ill-conditioning, we do not directly use, for all edges $e \in \mathcal{E}_n$, the traces of plane waves and evanescent waves, respectively, as basis functions for the spaces $\widetilde{\mathbb{PW}}_{p_e}(e)$. Instead, we apply Algorithm 2 to (i) automatically filter out redundancies in the edge basis functions, depending on the choice of the filtering parameter σ (which, for the numerical experiments, is set to 10^{-13}), and (ii) reduce the number of degrees of freedom. The resulting orthogonalized basis functions $\widehat{w}^e_\ell$ are hooded by a hat. Importantly, this strategy naturally dovetails with the supplement of special functions to the standard plane wave spaces and the use of plane wave spaces with varying degree from element to element. The traces of the corresponding functions are simply added edgewise first, as they are needed for the construction of the method (this leads to an increase of the number of degrees of freedom); afterwards, the relevant information is extracted using Algorithm 2 and the number of degrees of freedom is reduced significantly.

Local and global matrices. As in standard nonconforming FEM and VEM, the global system of linear equations is assembled in terms of the local contributions. Setting $\widehat{p}_K := \sum_{e \in \mathcal{E}^K} \widehat{p}_e$ and recalling that n_K denotes the number of edges of K, we define the following matrices:

- for all $K \in \mathcal{T}_n^1$:
  * $\mathbf{G}^{(1),K} \in \mathbb{C}^{p^K \times p^K}$ with $\mathbf{G}^{(1),K}_{j,\ell} := a^K_{\mathfrak{k}}(w^{(1),K}_\ell, w^{(1),K}_j)$, for all $j, \ell = 1, \dots, p^K$;
  * $\mathbf{D}^{(1),K} \in \mathbb{C}^{\widehat{p}_K \times p^K}$ with $\mathbf{D}^{(1),K}_{(r,j),\ell} := \widehat{\mathrm{dof}}_{r,j}(w^{(1),K}_\ell)$, for all $r = 1, \dots, n_K$, $j = 1, \dots, \widehat{p}_{e_r}$, and $\ell = 1, \dots, p^K$;
  * $\mathbf{B}^{(1),K} \in \mathbb{C}^{p^K \times \widehat{p}_K}$ with $\mathbf{B}^{(1),K}_{j,(s,\ell)} := a^K_{\mathfrak{k}}(\widehat{\varphi}_{s,\ell}, w^{(1),K}_j)$, for all $j = 1, \dots, p^K$, $s = 1, \dots, n_K$, and $\ell = 1, \dots, \widehat{p}_{e_s}$;
- for all $K \in \mathcal{T}_n^2$:
  * $\mathbf{G}^{(2),K} \in \mathbb{C}^{(p^K+\widetilde{p}^K)\times(p^K+\widetilde{p}^K)}$ as in (6.19);

- * $\mathbf{D}^{(2),K} \in \mathbb{C}^{\widehat{p}_K \times (p^K + \widetilde{p}^K)}$ with

$$\mathbf{D}^{(2),K}_{(r,j),\ell} := \begin{cases} \widehat{\mathrm{dof}}_{r,j}(w_\ell^{(2),K}), & \text{if } \ell \leq p^K \\ \widehat{\mathrm{dof}}_{r,j}(w_{\ell-p^K}^{EV,K}), & \text{if } \ell > p^K, \end{cases}$$

  for all $r = 1, \dots, n_K$ and $j = 1, \dots, \widehat{p}_{e_r}$;

- * $\mathbf{B}^{(2),K} \in \mathbb{C}^{(p^K + \widetilde{p}^K) \times \widehat{p}_K}$ with

$$\mathbf{B}^{(2),K}_{j,(s,\ell)} := \begin{cases} a_{\mathfrak{k}}^K(\widehat{\varphi}_{s,\ell}, w_j^{(2),K}), & \text{if } j \leq p^K \\ a_{\mathfrak{k}}^K(\widehat{\varphi}_{s,\ell}, w_{j-p^K}^{EV,K}), & \text{if } j > p^K, \end{cases}$$

  for all $s = 1, \dots, n_K$, and $\ell = 1, \dots, \widehat{p}_{e_s}$.

Having this, the matrix representation $\mathbf{A}^{(1),K}$ of $a_{\mathfrak{k},h}^K(\cdot,\cdot)$ is given, for all $K \in \mathcal{T}_n^1$, by

$$\mathbf{A}^{(1),K} := \overline{\mathbf{\Pi}_*^{(1),K}}^T \mathbf{G}^{(1),K} \mathbf{\Pi}_*^{(1),K} + \overline{(\mathbf{I}^{(1),K} - \mathbf{\Pi}^{(1),K})}^T \mathbf{S}^{(1),K} (\mathbf{I}^{(1),K} - \mathbf{\Pi}^{(1),K}),$$

where $\mathbf{I}^{(1),K} \in \mathbb{C}^{\widehat{p}_K \times \widehat{p}_K}$ is the identity matrix, $\mathbf{S}^{(1),K}$ is the matrix representation of the stabilizing sesquilinear form $S_{\mathfrak{k}}^K(\cdot,\cdot)$, and

$$\mathbf{\Pi}_*^{(1),K} := (\mathbf{G}^{(1),K})^{-1} \mathbf{B}^{(1),K} \in \mathbb{C}^{p^K \times \widehat{p}_K}, \quad \mathbf{\Pi}^{(1),K} := \mathbf{D}^{(1),K} (\mathbf{G}^{(1),K})^{-1} \mathbf{B}^{(1),K} \in \mathbb{C}^{\widehat{p}_K \times \widehat{p}_K}.$$

The matrix $\mathbf{A}^{(2),K}$ related to $a_{\mathfrak{k},h}^K(\cdot,\cdot)$ for $K \in \mathcal{T}_n^2$ is computed analogously.

Regarding the Robin part, given $e \in \mathcal{E}_n^B$, the matrix representation $\mathbf{R}^e$ of $(\mathfrak{k}\Pi_{p_e}^{0,e}\cdot, \Pi_{p_e}^{0,e}\cdot)_{0,e}$ is

$$\mathbf{R}^e := \overline{\mathbf{B}_0^e}^T \overline{\mathbf{G}_0^e}^{-T} \mathbf{B}_0^e,$$

where $\mathbf{G}_0^e$ and $\mathbf{B}_0^e \in \mathbb{C}^{\widehat{p}_e \times \widehat{p}_e}$ are given by $(\mathbf{G}_0^e)_{j,\ell} := (\widehat{w}_\ell^e, \widehat{w}_j^e)_{0,e}$ and $(\mathbf{B}_0^e)_{j,\ell} := (\widehat{\varphi}_{e,\ell}, \widehat{w}_j^e)_{0,e} = h_e \delta_{j,\ell}$, for all $j, \ell = 1, \dots, \widehat{p}_e$, respectively.

The right-hand side $F_h(v_h)$ is computed by expressing $\Pi_{p_e}^{0,e}(v_{h|e})$ in terms of the orthogonalized basis functions $\widehat{w}_\ell^e$ and using numerical integration.

6.3.2 Numerical experiments

In this section, we employ the method (6.7) to approximate the solution to (6.2) in three different test cases, using the notation of Section 6.1:

- `test case 1`: given an incoming traveling plane wave with $\theta_{\text{inc}} \geq \theta_{\text{crit}}$, this wave is partially reflected at the interface Γ and a plane wave is transmitted in the subdomain Ω_2;
- `test case 2`: given an incoming traveling plane wave with $\theta_{\text{inc}} < \theta_{\text{crit}}$, the wave is completely reflected and evanescent modes appear in Ω_2;
- `test case 3`: we consider the same situation as in `test case 1`, but employ here meshes with elements that are cut by the interface Γ.

Note that for all the test cases, the exact solution is known in closed form. In fact, assuming that u_{inc} is an incoming traveling plane wave with angle θ_{inc} and wave number k_1, i.e.

$$u_{\text{inc}}(\mathbf{x}) := \exp(ik_1 \mathbf{d} \cdot \mathbf{x}), \quad \mathbf{d} := (\cos(\theta_{\text{inc}}), \sin(\theta_{\text{inc}})),$$

the solution to the global problem (6.2) is given by

$$u := \begin{cases} u_{\text{inc}} + u_R & \text{in } \Omega_1 \\ u_T & \text{in } \Omega_2. \end{cases} \tag{6.21}$$

The reflected and the transmitted waves, respectively, can be expressed as

$$u_R(x,y) := R \exp(ik_1 \mathbf{d} \cdot (x, -y)), \quad u_T(x,y) := T \exp(ik_2(K_1 x + K_2 y)), \tag{6.22}$$

where the coefficients R, T, K_1, and K_2 are computed by using the transmission conditions (6.5):

$$K_1 := k_1/k_2\cos(\theta_{\text{inc}}), \quad K_2 := \sqrt{1 - k_1^2/k_2^2\cos^2(\theta_{\text{inc}})}, \quad R := \frac{k_1\sin(\theta_{\text{inc}}) - k_2K_2}{k_1\sin(\theta_{\text{inc}}) + k_2K_2}, \quad T := 1 + R.$$

Since an explicit representation of the numerical solution u_h is not available in closed form inside each element, we compute the approximate relative errors

$$\frac{\|u - \Pi u_h\|_{1,k,\mathcal{T}_n}}{\|u\|_{1,k,\Omega}}, \qquad \frac{\|u - \Pi u_h\|_{0,\mathcal{T}_n}}{\|u\|_{0,\Omega}}, \tag{6.23}$$

where $\Pi_{|K}(v_h) = \Pi^{(1),K}_{p^K}(v_h)$ for all $v_h \in V^{\Delta+\mathfrak{k}^2}(K)$ and for all $K \in \mathcal{T}_n^1$, and $\Pi_{|K}(v_h) = \Pi^{(2),K}_{p^K,\widetilde{p}^K}(v_h)$ for all $v_h \in V^{\Delta+\mathfrak{k}^2}(K)$ and for all $K \in \mathcal{T}_n^2$, are the local projectors defined in (6.18).

As stabilization $S_{\mathfrak{k}}^K(\cdot,\cdot)$ in (6.20), we employ the modified D-recipe in (5.32), namely,

$$S_{\mathfrak{k}}^K(u_h, v_h) = \sum_{s=1}^{n_K}\sum_{\ell=1}^{\widehat{p}_{e_s}} a_{\mathfrak{k}}^K(\Pi\varphi_{s,\ell}, \Pi\varphi_{s,\ell})\text{dof}_{s,\ell}(u_h)\overline{\text{dof}_{s,\ell}(v_h)}, \tag{6.24}$$

where Π^K is either $\Pi^{(1),K}_{p^K}$ or $\Pi^{(2),K}_{p^K,\widetilde{p}^K}$, depending on K.

`Test case 1` (incoming plane wave with $\theta_{\text{inc}} > \theta_{\text{crit}}$)

We firstly consider the test case of an incoming plane wave with incident angle $\theta_{\text{inc}} > \theta_{\text{crit}}$. In this case, reflection and transmission of plane waves take place.

As refraction indices, we pick $n_1 = 2$ and $n_2 = 1$. Accordingly with (6.6), the critical angle is $\theta_{\text{crit}} = 60°$. We consider $\theta_{\text{inc}} = 75°$ and $k = 7$, i.e. local wave numbers $k_1 = 14$ and $k_2 = 7$. The exact solution is given in (6.21) and its real part is depicted in Figure 6.4.

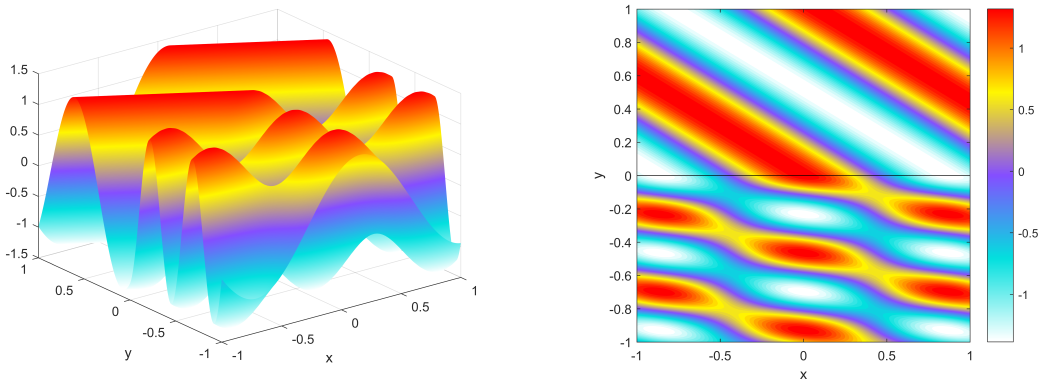


Figure 6.4: Real part of the exact solution u given by (6.21) with $k = 7$, $n_1 = 2$, $n_2 = 1$, and $\theta_{\text{inc}} = 75°$. *Left*: surface plot. *Right*: contour plot, where the black line indicates the interface Γ.

We study the h- and p-versions of the method for the problem (6.5), where the impedance datum g is computed accordingly with the exact analytical solution. Inside each subdomain Ω_1 and Ω_2, only plane waves with the same set of equidistributed directions are employed. In the following, we will always write q_1, q_2 and $\widetilde{q}_2$ when the effective plane/evanescent wave degrees do not vary elementwise within each subdomain.

For the h-version, we study the behaviour of the error curves for different values of q_1 and q_2, namely $q_1 = q_2 = 4$, $q_1 = q_2 = 6$, and $q_1 = 12$ with $q_2 = 6$. Recall that the numbers of plane waves in Ω_1 and Ω_2, respectively, are given by $p_1 = 2q_1 + 1$ and $p_2 = 2q_2 + 1$. Since no evanescent modes are expected to appear in Ω_2 and the transmitted solution is a plane wave, we do not add evanescent waves to the local spaces, i.e. we take $\widetilde{q}_2 = 0$. We employ sequences of standard regular Cartesian meshes and Voronoi meshes (reflected across the x- and y-axes), see Figure 6.5. The results are depicted in Figure 6.6.

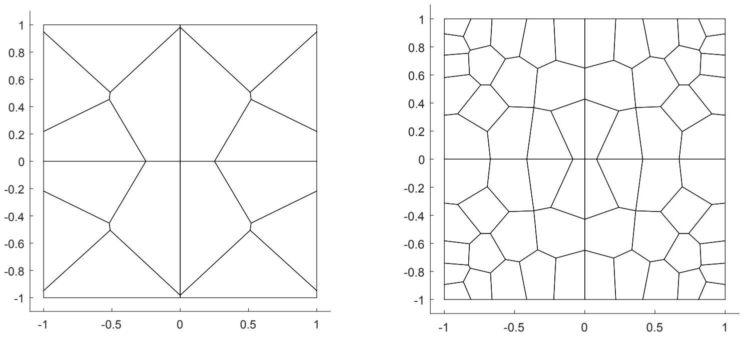


Figure 6.5: Voronoi meshes (reflected across the x- and y-axes) with 16, 64, and 128 elements; *left* to *right*.

We observe algebraic convergence in terms of the minimal effective degree $\min\{q_1, q_2\}$. The rates for the H^1 and L^2 errors are approximately given by $\min\{q_1, q_2\}$ and $\min\{q_1, q_2\} + 1$, respectively. Further, when using the Voronoi meshes, the curves are not as straight as in the Cartesian case. This can be explained by the presence of very small edges and of elements with different sizes.

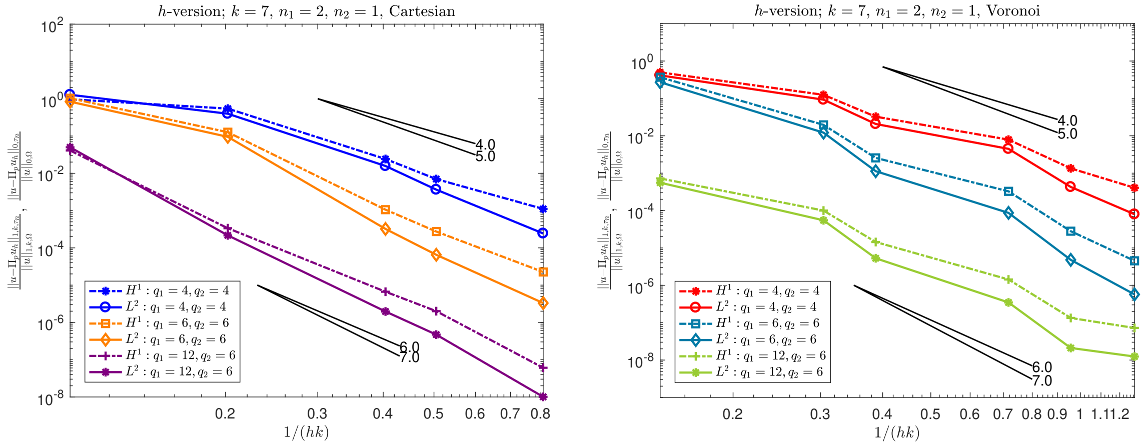


Figure 6.6: h-version of the method for u in (6.21) with $k = 7$, $n_1 = 2$, $n_2 = 1$, and $\theta_{\text{inc}} = 75°$ on a sequence of regular Cartesian meshes (*left*) and a sequence of Voronoi meshes as in Figure 6.5 (*right*). The relative errors are computed accordingly with (6.23).

Next, we investigate the p-version of the method. To this end, we fix a regular Cartesian mesh and the Voronoi mesh in Figure 6.5 with 64 elements. We vary the effective degrees q_1 and q_2, and study the behaviour for the cases $q_1 = q_2$ and $q_1 = 2q_2$. The error plots are displayed in Figure 6.7.

We observe exponential convergence with respect to the effective degree q_2, where the slope of the error curves is basically the same for $q_1 = q_2$ and $q_1 = 2q_2$, but the accuracy is a few orders higher in the latter case. Moreover, we recognize the cliff effect in Figure 6.7 (right), which is a consequence of the orthogonalization-and-filtering process in Algorithm 2. In fact, when increasing p, the growth of the number of degrees of freedom slows down; this results in a convergence rate which is effectively more than exponential. Interestingly, in the last p-refinements, the error seems to tend to zero even without an increase of the number of degrees of freedom.

It is worth to underline that the exponential convergence of the p-version is expected from the fact that we have considered so far meshes that are conforming with respect to the interface Γ and that the exact solution is piecewise analytic on the two subdomains Ω_1 and Ω_2. In Section 6.3.2, we will investigate the performance of the method employing meshes that are not conforming with respect to Γ.

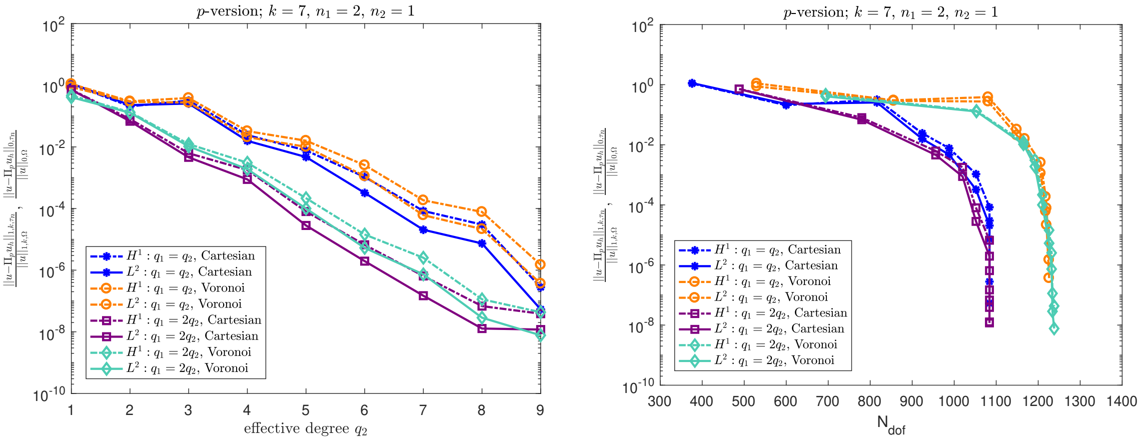


Figure 6.7: p-version of the method for u in (6.21) with $k = 7$, $n_1 = 2$, $n_2 = 1$, and $\theta_{\text{inc}} = 75°$ on a regular Cartesian mesh and the Voronoi mesh in Figure 6.5 with 64 elements each. The relative errors are computed accordingly with (6.23). *Left*: relative bulk errors against q_2. *Right*: relative bulk errors against the number of degrees of freedom.

Test case 2 (incoming plane wave with $\theta_{\text{inc}} < \theta_{\text{crit}}$)

Here, we fix the incident angle of the incoming wave $\theta_{\text{inc}} < \theta_{\text{crit}}$. This leads to total reflection of the plane wave at the interface Γ; evanescent modes occur in Ω_2. Since the evanescent modes are characterized by an exponential decay, the method could benefit from adding special functions which decay exponentially as well, that is, evanescent waves. To this purpose, inspired by [141,170], we compare the method when only plane waves are used in Ω_2 with the case when also evanescent waves in Ω_2 are added. Similarly as above, we investigate the h- and p-versions.

We pick $k = 7$, $n_1 = 2$ and $n_2 = 1$, as before, and the incoming angle $\theta_{\text{inc}} = 50°$. The real part of the corresponding exact solution computed as in (6.21) is plotted in Figure 6.8.

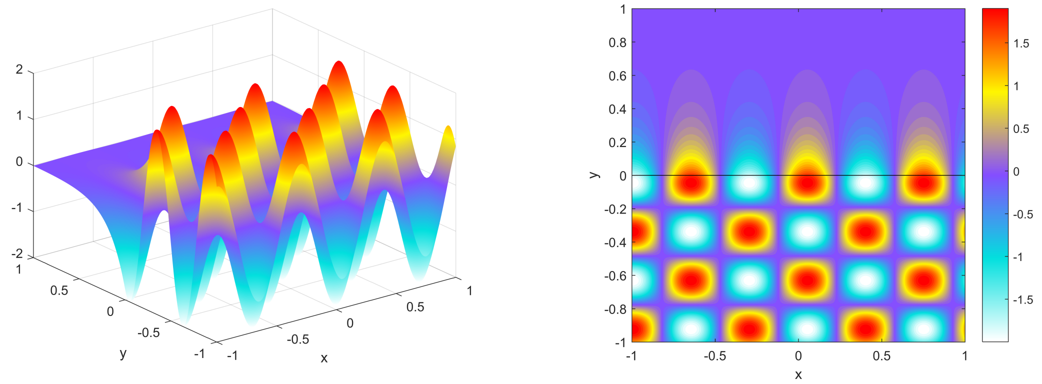


Figure 6.8: Real part of the exact solution u given by (6.21) with $k = 7$, $n_1 = 2$, $n_2 = 1$, and $\theta_{\text{inc}} = 50°$. *Left*: surface plot. *Right*: contour plot, where the black line indicates the interface Γ.

For the h-version, we assume once again that the effective plane/evanescent wave degree is the same for all elements within a subdomain. In Ω_1, we take $q_1 = 12$ (namely, 25 plane waves), whereas, in Ω_2, we consider

- $q_2 = 6$ and $\widetilde{q}_2 = 0$, i.e. 13 plane waves and 0 evanescent waves;
- $q_2 = 5$ and $\widetilde{q}_2 = 1$, i.e. 11 plane waves and 2 evanescent waves;
- $q_2 = 4$ and $\widetilde{q}_2 = 2$, i.e. 9 plane waves and 4 evanescent waves;

- $q_2 = 0$ and $\widetilde{q}_2 = 6$, i.e. 0 plane waves and 12 evanescent waves.

Note that we do not choose $q_1 = q_2 + \widetilde{q}_2$ on purpose since, in this case, the discretization error in Ω_1 dominates that in Ω_2 due to the higher local wave number. For this reason, we picked q_1 equal to the double of $q_2 + \widetilde{q}_2$.

We employ the same meshes as for the h-version in `test case 1`. The results are plotted in Figure 6.9. As already indicated in [141, Section 4], by adding evanescent waves to the local spaces, the order of convergence of the method is not changed, but the accuracy is improved by a multiplicative factor. We also underline that the convergence deteriorates when the error becomes sufficiently small (typically around 10^{-8}). This effect can be traced back to the ill-conditioning haunting the wave based methods and which can not be totally removed by Algorithm 2.

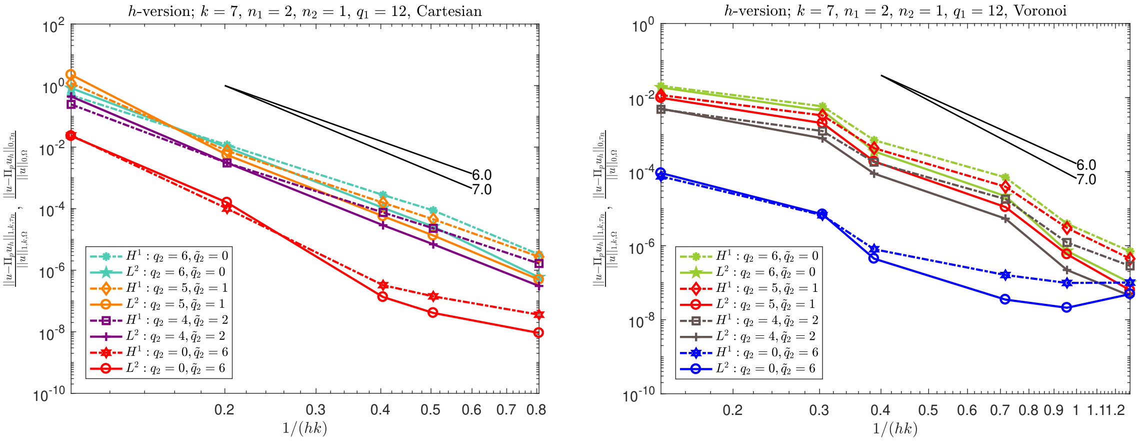


Figure 6.9: h-version of the method for u in (6.21) with $k = 7$, $n_1 = 2$, $n_2 = 1$, $q_1 = 12$, and $\theta_{\text{inc}} = 50°$ on a sequence of regular Cartesian meshes (*left*) and a sequence of Voronoi meshes as in Figure 6.5 (*right*). The relative errors are computed accordingly with (6.23).

Regarding the p-version, we fix, as before, the Voronoi mesh in Figure 6.5 with 64 elements. This time we assume that $q_1 = 2(q_2 + \widetilde{q}_2)$. We consider

- $\widetilde{q}_2 = 0$ and increase q_2;
- $\widetilde{q}_2 = 1$ and increase q_2;
- $\widetilde{q}_2 = 2$ and increase q_2;
- $q_2 = 0$ and increase $\widetilde{q}_2$.

The error plots are shown in Figure 6.10. Similar results are obtained when using a regular Cartesian mesh with 64 elements; for this reason, we omit them. As before, we observe exponential convergence in terms of the sum of the effective degrees $q_2 + \widetilde{q}_2$, where the accuracy of the method is again improved when evanescent waves are contained in the approximation spaces in Ω_2. The best performance is achieved when only evanescent waves are used in Ω_2.

`Test case 3` (non-conforming meshes and the hp-version)

So far, we have employed sequences of meshes that are conforming with respect to the interface Γ, that is, every $K \in \mathcal{T}_n$ is contained either in Ω_1 or in Ω_2. The advantage of this choice is that, since the exact solution (6.21) is piecewise analytic, the h- and the p-versions of the method have optimal order of convergence. In particular, the p-version results in exponential convergence as highlighted in Figures 6.7 and 6.10. Such an exponential convergence is however in terms of the number and not in terms of the square root of the number of degrees of freedom. This is due to the Trefftz nature of the method.

We now investigate how the method can be tuned to address the case where some elements of the mesh are cut by the interface Γ. This situation can be of interest in the following situations:

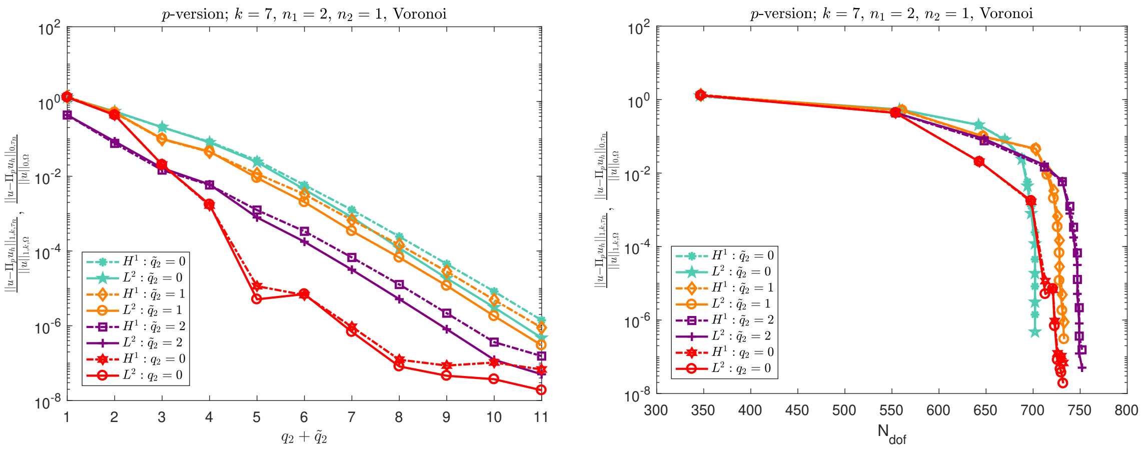


Figure 6.10: p-version (effective degrees q_1, q_2 and $\widetilde{q}_2$ with $q_1 = 2(q_2 + \widetilde{q}_2)$) of the method for u in (6.21) with $k = 7$, $n_1 = 2$, $n_2 = 1$, and $\theta_{\text{inc}} = 50°$ on the Voronoi mesh with 64 elements in Figure 6.5. The relative errors are computed accordingly with (6.23). *Left*: relative bulk errors against $q_2 + \widetilde{q}_2$. *Right*: relative bulk errors against the number of degrees of freedom.

- the interface Γ is curvilinear and one does not want to resort to curvilinear VEM [47]; in this case, some polygonal elements necessarily cut Γ;
- assuming that the parameter $\mathfrak{k}$ is subject to uncertainty, e.g. it is piecewise constant over subdomains with stochastic boundaries, one could proceed by reduced basis techniques starting from a very coarse mesh, and then, perform adaptive mesh and space refinements.

The first issue that has to be faced is the definition of the local spaces over the elements K in $\mathcal{T}_n$ such that $K^\circ \cap \Gamma \neq \emptyset$. Since on such elements, the wave number $\mathfrak{k}$ takes two different values, namely k_1 and k_2, we propose to fix the local spaces $V^{\Delta+\mathfrak{k}^2}(K)$ defined as in (6.14), with wave number either given by the maximum between k_1 and k_2 (i.e. k_1), or the average of k_1 and k_2. In both cases, the resulting method (6.7) is not Trefftz anymore.

For the forthcoming numerical tests, we focus for simplicity on the exact solution to `test case 1`, i.e. when the incident angle is larger than the critical angle. Furthermore, we do not employ evanescent waves and only consider here the case where the average of the wave number is chosen in the elements abutting Γ. Note that slightly worse results are obtained when taking the maximum between the two wave numbers.

Another issue to cope with is that, since the solution is analytic over the subdomains Ω_1 and Ω_2, but not over the complete domain Ω, the standard h- and p-versions of the method may not converge or converge suboptimally when employing non-conforming meshes. In order to overcome such a problem, we will employ hp-refinements, that is, we will construct VE spaces based on polygonal meshes that are graded geometrically towards the interface Γ and have local effective degrees possibly varying from element to element. In particular, we will resort to both isotropic and anisotropic mesh refinements.

The remainder of this section is organized as follows. Firstly, we describe the construction of VE spaces with elementwise variable effective degree on geometrically graded meshes employing isotropic and anisotropic mesh refinements, respectively. Then, we present numerical experiments, where we compare the h- and hp-versions (with isotropic mesh refinements) of the method. Finally, a comparison between hp-isotropic and anisotropic mesh refinements is discussed.

***hp*-virtual element spaces on isotropic geometrically refined meshes.** We introduce geometric isotropic mesh refinements towards the interface Γ, which can be seen as adaptations of the hp-graded meshes for point singularities introduced in Sections 3.3.3 and 5.4.3 to the case of edge singularities, and the associated hp-VE spaces.

To this purpose, we assume that a mesh $\mathcal{T}_n$ consists of $n + 1$ layers. The 0-th layer L_n^0 is the set of all polygons abutting the interface Γ, whereas the other layers are defined by induction:

$$L_\ell^n := \left\{K_1 \in \mathcal{T}_n \mid \overline{K_1} \cap \overline{K_2} \neq \emptyset \text{ for some } K_2 \in L_{\ell-1}^n,\ K_1 \not\subseteq \cup_{j=0}^{\ell-1} L_j^n\right\} \quad \forall \ell = 1, \ldots, n.$$

We say that $\{\mathcal{T}_n\}_n$ is a sequence of isotropic geometrically graded meshes if (i) $\mathcal{T}_{n+1}$ is obtained by starting from $\mathcal{T}_n$ and refining only the elements in the layer L_0^n, and (ii) there exists a grading parameter $\sigma \in (0,1)$ such that

$$h_K \approx \sigma^{n-\ell} \quad \text{if } K \in L_\ell^n. \tag{6.25}$$

In words, such isotropic geometrically graded meshes are characterized by small elements abutting the interface and elements enlarging geometrically when the distance from Γ increases. We assume that all the elements have bounded aspect ratio.

The hp-VE spaces over such meshes are defined analogously to those in Section 6.2.2, where, for some positive parameter μ and denoting by $\lceil \cdot \rceil$ the ceiling function, the following two types of distributions of the effective degrees are considered:

1. uniformly distributed effective degrees

$$\mathbf{p}_j = \lceil \mu(n+1) \rceil \quad \forall j = 1, \ldots, \operatorname{card}(\mathcal{T}_n); \tag{6.26}$$

2. graded effective degrees

$$\mathbf{p}_j = \lceil \mu(\ell+1) \rceil \text{ if } K_j \in L_\ell^n \quad \forall j = 1, \ldots, \operatorname{card}(\mathcal{T}_n). \tag{6.27}$$

The latter approach is based on effective degrees growing together with the layer index. In fact, the singularity is approximated with the aid of small elements, whereas the analytic part is approximated on large elements with high effective degrees.

In Figure 6.11, we depict the first two meshes $\mathcal{T}_1$ and $\mathcal{T}_2$ (including the graded distribution (6.27) of the effective degrees with $\mu = 1$) of a sequence of isotropic geometrically graded meshes with grading parameter σ in (6.25) given by 1/3.

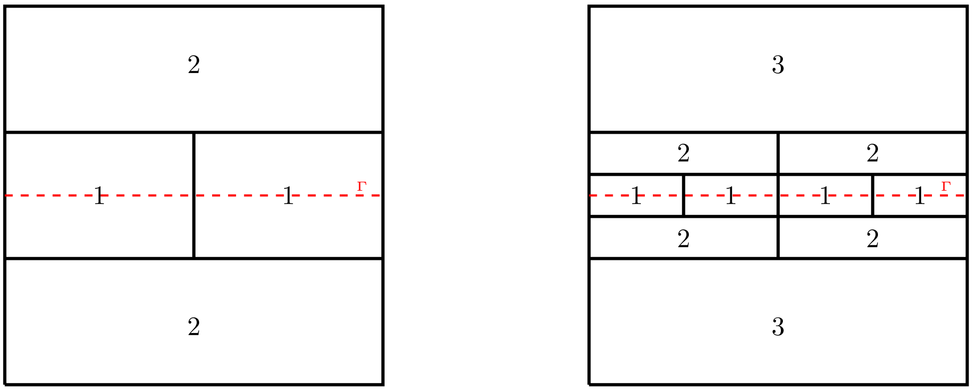


Figure 6.11: First two meshes $\mathcal{T}_1$ and $\mathcal{T}_2$ (using the graded distribution (6.27) of the effective degrees, with $\mu = 1$) of a sequence of isotropic geometrically graded meshes. The grading parameter σ in (6.25) is 1/3. The dashed red line denotes the interface Γ.

***hp*-virtual element spaces on anisotropic geometrically refined meshes.** Here, we describe anisotropic geometric mesh refinements towards the interface Γ.

The concept of layers of $\mathcal{T}_n$ is similar to that for isotropic geometric mesh refinements, however, the geometric grading is done in a slightly different way. More precisely, given $K \in \mathcal{T}_n$, let $h_{K,1}$ and $h_{K,2}$ be the lengths of the edges of the rectangle of minimal perimeter bounding K with edges parallel to Γ and its normal direction, respectively. For anisotropic geometrically graded meshes, the mesh refinement $\mathcal{T}_{n+1}$ is obtained starting from $\mathcal{T}_n$ and refining only the elements in the layer L_0^n in such a way there exists a grading parameter $\sigma \in (0,1)$ with

$$h_{K,2} \approx \sigma^{n-\ell} \quad \text{if } K \in L_\ell^n, \qquad h_{K,1} \approx 1 \qquad \forall K \in \mathcal{T}_n. \tag{6.28}$$

Thus, there are very thin elements in proximity of the interface Γ and larger elements elsewhere.

The reason why we also employ anisotropic mesh refinements is that the solution is singular only in the normal direction to Γ and not along the tangential one. Thus, roughly speaking, it suffices to refine the mesh along the normal direction to Γ. Numerically, this results in a more effective approach for approximating edge singularities. In fact, in the finite element framework, one gets exponential convergence in terms of the cubic root of the degrees of freedom (in the Trefftz setting, the cubic root becomes the square root, see e.g. [79,122,143,145]), whereas, with isotropic mesh refinements, one only obtains an algebraic rate of convergence.

Note that, for anisotropic meshes, we only employ uniformly distributed effective degrees (6.26). The graded approach (6.27) would not suffice for approximating the tangential part of the solution (here, the elements have too long edges and therefore the method would not converge properly with very few degrees of freedom).

In Figure 6.12, we depict the first two meshes $\mathcal{T}_1$ and $\mathcal{T}_2$ (including the uniform effective degrees (6.26) element by element) of a sequence of anisotropic geometrically graded meshes with grading parameter σ in (6.28) given by 1/3.

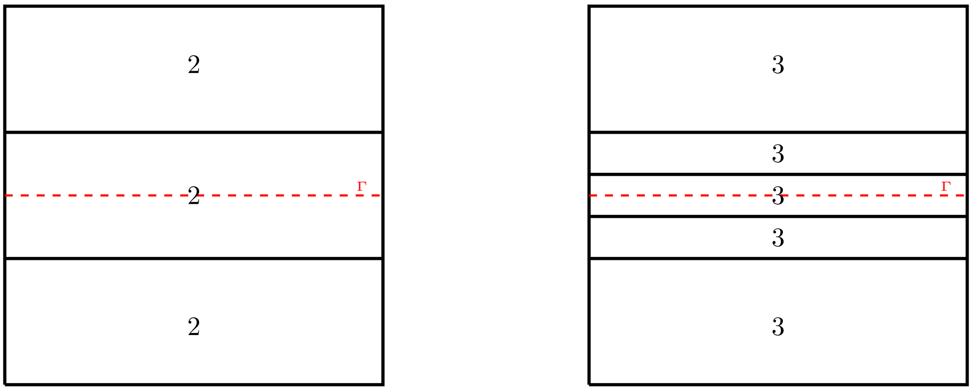


Figure 6.12: First two meshes $\mathcal{T}_1$ and $\mathcal{T}_2$ (using the uniform effective degree (6.26) element by element) of a sequence of anisotropic geometrically graded meshes. The grading parameter σ in (6.25) is 1/3. The dashed red line denotes the interface Γ.

We point out that, in the case of polynomial based methods, one could design local spaces on the elements cut by the interface Γ, which take care of the fact that the solution is smooth in the horizontal direction, but singular across the vertical one. More precisely, one could take a tensor product of polynomials of degree p on the horizontal axis, and of affine polynomials along the vertical one. By doing so, the increase in the polynomial degree and the (vertical) mesh refinement would take into account the approximation of the smooth horizontal and of the (vertical) singular behaviours, respectively, leading to exponential convergence of the error in terms of a proper root of the number of degrees of freedom, see e.g. [109] and the references therein.

To the best of our understanding, this framework does not extend to the case of plane wave based approximation spaces. In fact, plane waves do not have a tensor product structure, and by employing more basis functions almost-aligned along the direction of the interface Γ one would in general not be able to approximate the solution with less degrees of freedom. For instance, in the case of transmission of a single plane wave, a uniform p-refinement could be (for specific directions of propagation) more effective than by adding more directions along the interface.

However, it is worthwhile to underline that one may proceed with an adaptive method, taking into account the presence of a dominant propagation direction, as investigated for instance in [82].

Non-conforming meshes: comparison of the h- and the hp-isotropic versions. In this section, we compare the h-version of the method on sequences of uniform Cartesian meshes that are non-conforming with respect to the interface Γ employing $p = 15$ plane wave directions, and the hp-version of the method with isotropic geometrically graded mesh as in Figure 6.12, endowed with both the uniform and the graded effective degrees in (6.26) and (6.27), respectively. In both cases, we pick $\mu = 1$, 2, and 3. The results are displayed in Figure 6.13, where we compare the computable relative H^1 and L^2 errors in (6.23) in terms of the number of degrees of freedom.

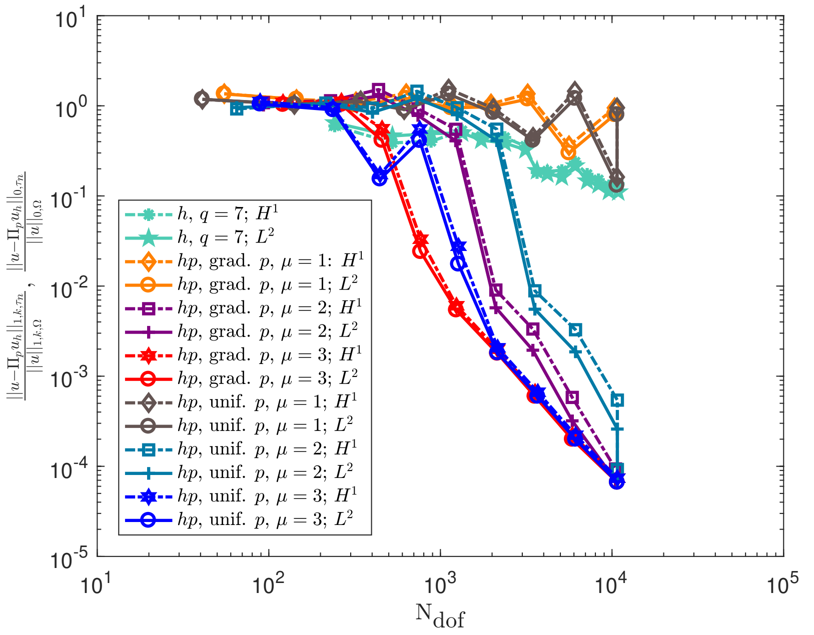


Figure 6.13: h-version employing non-conforming Cartesian meshes and hp-version with isotropic geometrically graded meshes, with grading parameter σ in (6.25) equal to 1/3, and $p = 15$ plane waves on every element. For the hp-spaces, we consider both uniform (6.26) and graded effective degrees (6.27), with $\mu = 1$, 2, and 3. The computable relative H^1 and L^2 errors in (6.23) are plotted against the number of degrees of freedom.

From Figure 6.13, we deduce that the h-version converges poorly, due to the low Sobolev regularity of the solution. The hp-version, on the other hand, performs much better. In particular, the choice of employing graded effective degrees seems to be the most effective. It has to be underlined that in order to achieve the convergence regime, the parameter μ in (6.26) and (6.27) has to be picked sufficiently large, e.g. $\mu = 2$.

Non-conforming meshes: comparison of the *hp*-isotropic and anisotropic versions. Here, we compare the behaviour of the method for the case of hp-isotropic and anisotropic mesh refinements, using the meshes depicted in Figures 6.11 and 6.12, respectively. In particular, whereas in the isotropic case, we only use the graded distribution (6.27) (since we know from Section 6.3.2 that the uniform distribution (6.26) works slightly worse), in the anisotropic case, we employ uniform effective degrees (6.26). In both cases, we take $\mu = 2$ and 3. The results are presented in Figure 6.14, where we compare the computable relative H^1 and L^2 errors in (6.23) in terms of the square root of the number of degrees of freedom.

From Figure 6.14, it is clear that employing anisotropic meshes leads to much better results. Whilst exponential convergence in terms of the square root of the number of degrees of freedom is obtained for anisotropic meshes, the rate of convergence is only algebraic for isotropic meshes.

So far, we have employed the average of the two wave numbers as an “artificial” wave number on the elements abutting the interface Γ. In Figure 6.15, we present some numerical results for the hp-version of the method when also taking the maximum between the two of them. We consider anisotropic mesh refinements and uniform effective degrees (6.26), with $\mu = 2$ and 3.

From Figure 6.15, we deduce that the choice for the “artificial” wave number does not particularly influence the method, although the performance, when picking the average, seems to be slightly better.

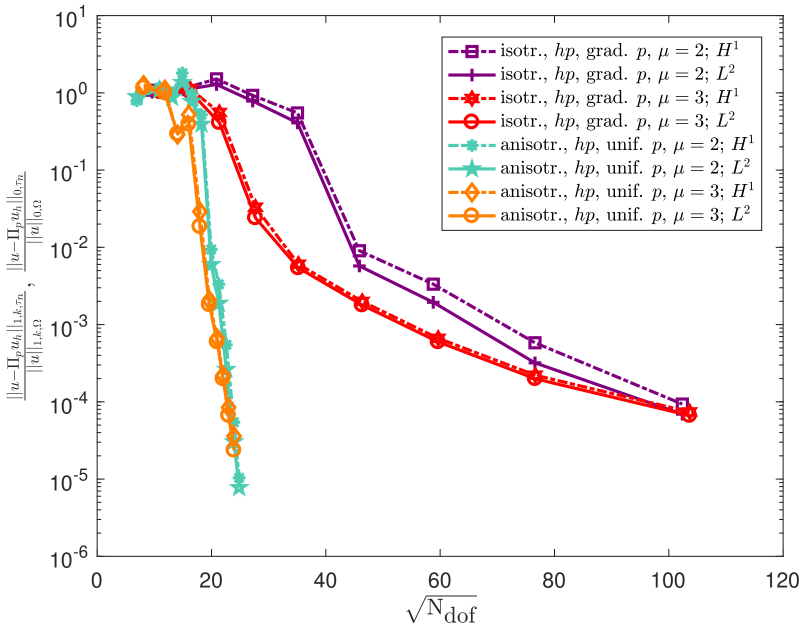


Figure 6.14: hp-versions with geometrically isotropic and anisotropic graded meshes, with grading parameter σ in (6.25) equal to 1/3. In the former case, we consider graded effective degrees (6.27), with $\mu = 2$ and 3, whereas, in the latter, the uniform one (6.26) is applied. We plot the computable relative H^1 and L^2 errors in (6.23) against the square root of the degrees of freedom.

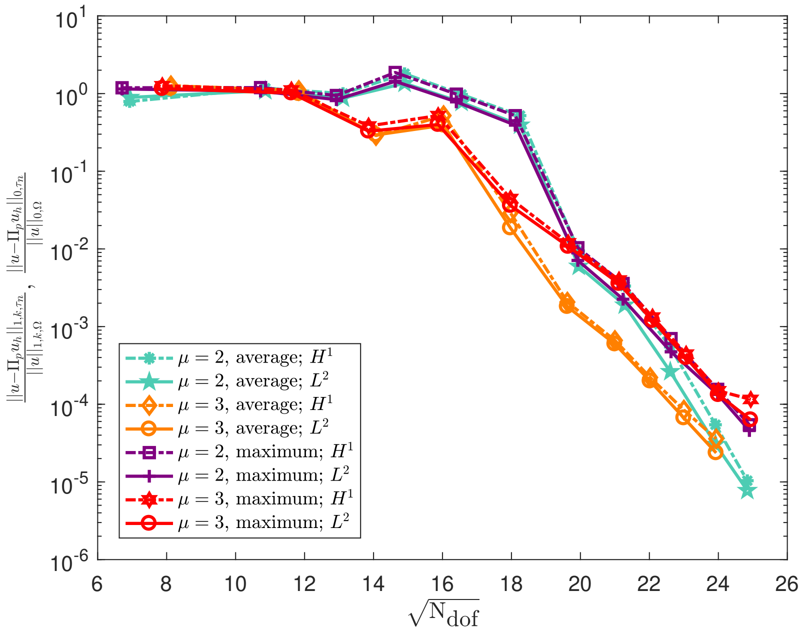


Figure 6.15: hp-version with anisotropic geometrically graded meshes, with grading parameter σ in (6.25) equal to 1/3. We consider uniform effective degrees (6.26) and compare the effects of the choice of the "artificial" wave number on the elements abutting the interface Γ; in particular, we pick the average and the maximum of the two wave numbers. On the x-axis, we plot the number of degrees of freedom; on the y-axis, we plot the computable relative H^1 and L^2 errors in (6.23).

Chapter 7

Virtual element method for the miscible displacement of incompressible fluids in porous media

In this chapter, we focus on the miscible displacement problem, given by

$$\begin{cases} \phi \dfrac{\partial c}{\partial t} + \boldsymbol{u} \cdot \nabla c - \operatorname{div}(D(\boldsymbol{u})\nabla c) = q^+(\widehat{c} - c) \\ \qquad\qquad\qquad\qquad \operatorname{div} \boldsymbol{u} = G \\ \qquad\qquad\qquad\qquad\qquad \boldsymbol{u} = -a(c)(\nabla p - \boldsymbol{\gamma}(c)), \end{cases} \tag{7.1}$$

with boundary conditions

$$\begin{cases} \qquad \boldsymbol{u} \cdot \boldsymbol{n} = 0 \quad \text{on } \partial\Omega \times J \\ D(\boldsymbol{u})\nabla c \cdot \boldsymbol{n} = 0 \quad \text{on } \partial\Omega \times J \end{cases} \tag{7.2}$$

and initial condition

$$c(\boldsymbol{x}, 0) = c_0(\boldsymbol{x}) \quad \text{in } \Omega. \tag{7.3}$$

For a detailed description of this problem and a physical interpretation of the involved quantities, we refer to Section 2.2.2. We are interested in the design and analysis of a corresponding VEM.

The outline of this chapter is as follows. In Section 7.1, VE methods corresponding to the semidiscrete (continuous in time and discrete in space) and the fully discrete formulations, respectively, are constructed. Then, in Section 7.2, an *a priori* error analysis for the fully discrete formulation is carried out and L^2 error estimates for the velocity, pressure, and concentration discretization errors, respectively, are derived. Finally, in Section 7.3, a couple of numerical experiments validating the theoretical results are presented.

The material of this section is based on the work [46], which has been submitted for publication.

7.1 The virtual element method

Starting from the weak formulation (2.19), in this section, we aim at finding VE methods for the semidiscrete and fully discrete formulations related to the model problem (7.1)-(7.3).

To this purpose, let $\mathcal{T}_n$ be a fixed mesh satisfying the mesh assumptions (**G1**)-(**G3**) in Section 2.3. Moreover, we require the following quasi-uniformity assumption on $\mathcal{T}_n$:

(**G6**) for all $n \in \mathbb{N}$ and for all $K \in \mathcal{T}_n$, it holds $h_K \geq \rho_1 h$, for some positive uniform constant ρ_1.

This assumption (that can also be found in many FEM papers on the same subject) is in fact only needed to prove bound (7.47) below.

Having this, we proceed as follows. After introducing a set of discrete spaces, discrete bilinear forms, and projectors in Section 7.1.1, we state a semidiscrete formulation which is continuous in time and discrete in space in Section 7.1.2. The fully discrete formulation is then the subject of Section 7.1.3.

7.1.1 Discrete spaces and projectors

Here, we introduce the local discrete VE spaces corresponding to $\boldsymbol{V}$, Q and Z in (2.17), a set of local projectors mapping from these VE spaces into spaces made of polynomials, and finally, the related global counterparts.

Local VE spaces

Let $K \in \mathcal{T}_n$ and let $k \in \mathbb{N}_0$ be a given *degree of accuracy.* Then, the local velocity and pressure VE spaces are defined by

$$\begin{aligned} \boldsymbol{V}_h(K) &:= \{\boldsymbol{v} \in H(\mathrm{div}; K) \cap H(\mathrm{rot}; K) : \, \boldsymbol{v} \cdot \boldsymbol{n}_{|_e} \in \mathbb{P}_k(e) \, \forall e \in \mathcal{E}^K, \\ &\qquad \mathrm{div}\, \boldsymbol{v} \in \mathbb{P}_k(K), \, \mathrm{rot}\, \boldsymbol{v} \in \mathbb{P}_{k-1}(K)\} \\ Q_h(K) &:= \{q \in L^2(K) : \, q \in \mathbb{P}_k(K)\}. \end{aligned} \tag{7.4}$$

These spaces are coupled with the *preliminary* local concentration space

$$\widetilde{Z_h}(K) := \{z \in H^1(K) : \, z_{|_{\partial K}} \in C^0(\partial K), \, z_{|_e} \in \mathbb{P}_{k+1}(e) \, \forall e \in \mathcal{E}^K, \, \Delta z \in \mathbb{P}_{k-1}(K)\}. \tag{7.5}$$

Moreover, it is important to observe that $[\mathbb{P}_k(K)]^2 \subset \boldsymbol{V}_h(K)$ and $\mathbb{P}_{k+1}(K) \subseteq \widetilde{Z_h}(K)$. Associated sets of local degrees of freedom are given as follows:

- for $\boldsymbol{V}_h(K)$, a set of degrees of freedom $\{\mathrm{dof}_j^{\boldsymbol{V}_h(K)}\}_{j=1}^{\mathrm{dim} V_h(K)}$ is defined by

$$\begin{aligned} &1. \quad \frac{1}{|e|} \int_e \boldsymbol{v} \cdot \boldsymbol{n}\, p_k \, \mathrm{d}s \qquad \forall p_k \in \mathbb{P}_k(e) \quad \forall e \in \mathcal{E}^K \\ &2. \quad \frac{1}{|K|^{\frac{1}{2}}} \int_K (\mathrm{div}\, \boldsymbol{v})\, p_k \, \mathrm{d}x \qquad \forall p_k \in \mathbb{P}_k(K)/\mathbb{R} \\ &3. \quad \frac{1}{|K|} \int_K \boldsymbol{v} \cdot \boldsymbol{x}^{\perp}\, p_{k-1} \, \mathrm{d}x \qquad \forall p_{k-1} \in \mathbb{P}_{k-1}(K), \end{aligned} \tag{7.6}$$

  with $\boldsymbol{x}^{\perp} := (\boldsymbol{x}_2, -\boldsymbol{x}_1)^T$, where we assume the coordinates to be centered at the barycenter of the element;

- for $Q_h(K)$, we consider $\{\mathrm{dof}_j^{Q_h(K)}\}_{j=1}^{\mathrm{dim} Q_h(K)}$ with

$$\frac{1}{|K|} \int_K q\, p_k \, \mathrm{d}x \qquad \forall p_k \in \mathbb{P}_k(K); \tag{7.7}$$

- for $\widetilde{Z_h}(K)$, we take $\{\mathrm{dof}_j^{\widetilde{Z_h}(K)}\}_{j=1}^{\mathrm{dim}\widetilde{Z_h}(K)}$ with

$$\begin{aligned} &1. \quad \text{pointwise values at the vertices: } \boldsymbol{v}(z) \\ &2. \quad \text{on each edge } e \in \mathcal{E}^K \text{, the values of } z \text{ at the } k \text{ internal Gauß-Lobatto points} \\ &3. \quad \frac{1}{|K|} \int_K z\, q_{k-1} \, \mathrm{d}x \qquad \forall q_{k-1} \in \mathbb{P}_{k-1}(K). \end{aligned} \tag{7.8}$$

In all three cases, unisolvency is provided. More precisely, for $\boldsymbol{V}_h(K)$, this was proven in e.g. [37], for $Q_h(K)$ it is immediate, and for $\widetilde{Z_h}(K)$, see e.g. [31].

We also highlight that $\boldsymbol{V}_h(K)$ endowed with (7.6) mimics the Raviart-Thomas element, but in fact those two elements only coincide in the special case of triangles and $k = 0$. An analogous result is true for $\widetilde{Z_h}(K)$, when compared to finite elements.

Remark 23. We note that for $k = 0$, one obtains the lowest-order local VE spaces. More precisely, in this case, the velocity space $\boldsymbol{V}_h(K)$ consists of all rotation-free vector fields with constant divergence and edgewise constant normal traces, the pressure space $Q_h(K)$ only contains the constant functions, and the concentration space $\widetilde{Z_h}(K)$ is made of all harmonic functions that are linear on each edge. This motivates the choice of the present polynomial degrees for the spaces. However, in general, it is also possible to choose a degree of accuracy k_1 for $\boldsymbol{V}_h(K)$ and $Q_h(K)$, and another strictly positive one k_2 for $\widetilde{Z_h}(K)$; see e.g. [90] for FEM. The following analysis can be extended easily to such more general case just by keeping track of the different polynomial degrees.

Remark 24. In order to really have a set of degrees of freedom in the computer code, one clearly needs to choose a basis for the polynomial test spaces appearing in (7.6) and (7.8). We here assume to take the classical choice, that is any monomial basis $\{m_1, m_2, .., m_\ell\}$ of the polynomial space satisfying $\|m_i\|_{L^\infty} \simeq 1$, $i = 1, 2, .., \ell$, where the L^∞ norm has to be taken over the corresponding edge or bulk.

Local projections

For the construction of the method, we need some tools to deal with VE functions due to the lack of their explicit knowledge in closed form. This tools are provided in the form of local operators mapping VE functions onto polynomials. To this purpose, following [31, 36], we introduce the subsequent projectors:

The projector $\boldsymbol{\Pi_k^{0,K}} : [L^2(K)]^2 \to [\mathbb{P}_k(K)]^2$ is defined as the L^2 projector onto vector-valued polynomials of degree at most k in each component: Given $\boldsymbol{f} \in [L^2(\Omega)]^2$,

$$(\boldsymbol{\Pi_k^{0,K}} \boldsymbol{f}, \boldsymbol{p}_k)_{0,K} = (\boldsymbol{f}, \boldsymbol{p}_k)_{0,K} \quad \forall \boldsymbol{p}_k \in [\mathbb{P}_k(K)]^2. \tag{7.9}$$

It can be shown, see [30], that this operator is computable for functions in $\boldsymbol{V}_h(K)$ only by knowing their values at the degrees of freedom (7.6). Moreover, one has computability also for functions of the form ∇z_h with $z_h \in \widetilde{Z_h}(K)$. This can be seen by using integration by parts:

$$\int_K (\boldsymbol{\Pi_k^{0,K}} \nabla z_h) \cdot \boldsymbol{p}_k \, \mathrm{d}s = \int_K \nabla z_h \cdot \boldsymbol{p}_k \, \mathrm{d}s = -\int_K z_h \underbrace{\operatorname{div} \boldsymbol{p}_k}_{\in \mathbb{P}_{k-1}(K)} \, \mathrm{d}s + \int_{\partial K} z_h \, \boldsymbol{p}_k \cdot \boldsymbol{n} \, \mathrm{d}s,$$

for all $\boldsymbol{p}_k \in [\mathbb{P}_k(K)]^2$, where the right-hand side is computable by means of (7.8).

The projector $\Pi_{k+1}^{\nabla,K} : H^1(K) \to \mathbb{P}_{k+1}(K)$ is given, for every $z \in H^1(K)$, by

$$\begin{cases} (\nabla \Pi_{k+1}^{\nabla,K} z, \nabla p_k)_{0,K} = (\nabla z, \nabla p_k)_{0,K} \quad \forall p_{k+1} \in \mathbb{P}_{k+1}(K) \\ \dfrac{1}{|\partial K|} \displaystyle\int_{\partial K} \Pi_{k+1}^{\nabla,K} z \, \mathrm{d}s = \dfrac{1}{|\partial K|} \displaystyle\int_{\partial K} z \, \mathrm{d}s \end{cases}$$

where the second identity is needed to fix the constants. Computability of this mapping for functions in $\widetilde{Z_h}(K)$ was shown in [31,36].

Discrete concentration space

The space in (7.5) was a preliminary space, useful to introduce the main idea of the construction. Nevertheless, we will here make use of a more advanced space for the discrete concentration variable. Indeed, one can use the operator $\Pi_{k+1}^{\nabla,K}$ to pinpoint the local enhanced space

$$\begin{aligned} Z_h(K) := \{ z \in H^1(K) : \; & z_{|\partial K} \in C^0(\partial K), \, z_{|e} \in \mathbb{P}_{k+1}(e) \, \forall e \in \mathcal{E}^K, \, \Delta z \in \mathbb{P}_{k+1}(K), \\ & \int_K z \, p_k \, \mathrm{d}x = \int_K (\Pi_{k+1}^{\nabla,K} z) \, p_k \, \mathrm{d}x \quad \forall p_k \in \mathbb{P}_{k+1}/\mathbb{P}_{k-1}(K) \}, \end{aligned}$$

where $\mathbb{P}_{k+1}/\mathbb{P}_{k-1}(K)$ is the space of polynomials in $\mathbb{P}_{k+1}(K)$ which are $L^2(K)$ orthogonal to $\mathbb{P}_{k-1}(K)$. It can be shown that the space $Z_h(K)$ has the same dimension and the same degrees of freedom (7.8) as $\widetilde{Z_h}(K)$, see [3,38]. The advantage of the space $Z_h(K)$, when compared to $\widetilde{Z_h}(K)$, is that *also* the L^2 projector $\Pi^{0,K}_{k+1} : L^2(K) \to \mathbb{P}_{k+1}(K)$ onto polynomials of degree at most $k+1$, defined analogously to (7.9), is computable [36]

Finally, we state the following approximation result for the above projectors [30, Lemma 5.1]:

Lemma 7.1.1. *Given $K \in \mathcal{T}_n$, let ψ and $\boldsymbol{\psi}$ be sufficiently smooth scalar- and vector-valued functions, respectively. Then, it holds, for all $k \in \mathbb{N}_0$,*

$$\begin{aligned} \|\psi - \Pi^{0,K}_k \psi\|_{\ell,K} &\le \zeta\, h_K^{s-\ell}\, |\psi|_{s,K}, \quad 0 \le \ell \le s \le k+1 \\ \|\boldsymbol{\psi} - \boldsymbol{\Pi}^{\mathbf{0},\boldsymbol{K}}_{\boldsymbol{k}} \boldsymbol{\psi}\|_{\ell,K} &\le \zeta\, h_K^{s-\ell}\, |\boldsymbol{\psi}|_{s,K}, \quad 0 \le \ell \le s \le k+1 \\ \|\psi - \Pi^{\nabla,K}_k \psi\|_{\ell,K} &\le \zeta\, h_K^{s-\ell}\, |\psi|_{s,K}, \quad 0 \le \ell \le s \le k+1,\ s \ge 1, \end{aligned}$$

*where $\zeta > 0$ only depends on the shape-regularity parameter ρ_0 in the mesh assumption **(G1)**, and the degree of accuracy k.*

Global VE spaces and projectors

The global discrete spaces are defined via their local counterparts:

$$\begin{aligned} \boldsymbol{V}_h &:= \{\boldsymbol{v} \in \boldsymbol{V} : \boldsymbol{v}_{|_K} \in \boldsymbol{V}_h(K)\, \forall K \in \mathcal{T}_n\} \\ Q_h &:= \{q \in Q : q_{|_K} \in Q_h(K)\, \forall K \in \mathcal{T}_n\} \\ Z_h &:= \{z \in Z : z_{|_K} \in Z_h(K)\, \forall K \in \mathcal{T}_n\}, \end{aligned}$$

with the obvious sets of global degrees of freedom.

For future use, we also introduce, for all $\boldsymbol{u}_h \in \boldsymbol{V}_h$,

$$\|\boldsymbol{u}_h\|^2_{\boldsymbol{V}_h} := \sum_{K \in \mathcal{T}_n} \|\boldsymbol{u}_h\|^2_{V,K} := \sum_{K \in \mathcal{T}_n} \left[\|\boldsymbol{u}_h\|^2_{0,K} + \|\operatorname{div} \boldsymbol{u}_h\|^2_{0,K}\right].$$

Moreover, we denote by $\boldsymbol{\Pi}^{\mathbf{0}}_{\boldsymbol{k}}$, Π^{∇}_{k+1} and Π^0_{k+1}, the global projectors which are defined elementwise as the corresponding local ones.

The sets of global degrees of freedom $\{\text{dof}_j^{\boldsymbol{V}_h}\}_{j=1}^{\dim V_h}$, $\{\text{dof}_j^{Q_h}\}_{j=1}^{\dim Q_h}$, and $\{\text{dof}_j^{Z_h}\}_{j=1}^{\dim Z_h}$ are obtained by coupling the local counterparts given in (7.6), (7.7), and (7.8), respectively.

7.1.2 Semidiscrete formulation

Our aim in this section is to find a semidiscrete formulation for (2.19) which is continuous in time and discrete in space. To this purpose, we employ the same notation for the numerical approximants $\boldsymbol{u}_h$, p_h, and c_h as in (2.18) for $\boldsymbol{u}$, p, and c, namely

$$\boldsymbol{u}_h(t)(x) := \boldsymbol{u}_h(x,t), \qquad p_h(t)(x) := p_h(x,t), \qquad c_h(t)(x) := c_h(x,t),$$

where the dependence on (t) will be suppressed again in the sequel.

A semidiscrete variational formulation for (2.19) can then be written in an abstract way as follows: for almost every $t \in J$, find $\boldsymbol{u}_h \in \boldsymbol{V}_h$, $p_h \in Q_h$, and $c_h \in Z_h$, such that

$$\begin{cases} \mathcal{M}_h\left(\dfrac{\partial c_h}{\partial t}, z_h\right) + \Theta_h(\boldsymbol{u}_h, c_h; z_h) + \mathcal{D}_h(\boldsymbol{u}_h; c_h, z_h) = \left(q^+ \widehat{c}, z_h\right)_h \\ \qquad\qquad \mathcal{A}_h(c_h; \boldsymbol{u}_h, \boldsymbol{v}_h) + B(\boldsymbol{v}_h, p_h) = (\boldsymbol{\gamma}(c_h), \boldsymbol{v}_h)_h \\ \qquad\qquad\qquad\qquad\qquad B(\boldsymbol{u}_h, q_h) = -\left(G, q_h\right)_{0,\Omega} \end{cases} \tag{7.10}$$

for all $\boldsymbol{v}_h \in \boldsymbol{V}_h$, $q_h \in Q_h$, and $z_h \in Z_h$, and the initial condition

$$c_h(0) = c_{0,h} := I_h c_0$$

is satisfied, where $I_h c_0$ is the VEM interpolant of c_0 in Z_h, and where the involved forms and terms in (7.10) are specified in the forthcoming lines.

Starting from the continuous problem (2.19), by simply replacing the continuous functions by their discrete counterparts, most of the resulting terms cannot be computed any more, owing to the fact that VE functions are not known explicitly in closed form. Thus, these terms need to be substituted by computable versions in the spirit of the VEM philosophy. To this purpose, the following replacements were made:

- The term $\mathcal{M}\left(\frac{\partial c_h}{\partial t}, z_h\right)$ in the concentration equation was replaced by

$$\mathcal{M}_h\left(\frac{\partial c_h}{\partial t}, z_h\right) := \sum_{K\in\mathcal{T}_n} \mathcal{M}_h^K\left(\frac{\partial c_h}{\partial t}, z_h\right), \tag{7.11}$$

where the local contributions are given as

$$\begin{aligned}\mathcal{M}_h^K(c_h, z_h) := \int_K \phi\,(\Pi_{k+1}^{0,K} c_h)\,(\Pi_{k+1}^{0,K} z_h)\,\mathrm{d}x \\ + \nu_{\mathcal{M}}^K(\phi) S_{\mathcal{M}}^K\left((I-\Pi_{k+1}^{0,K})c_h, (I-\Pi_{k+1}^{0,K})z_h\right),\end{aligned} \tag{7.12}$$

with $S_{\mathcal{M}}^K(\cdot,\cdot)$ denoting a stabilization term with certain properties and a constant $\nu_{\mathcal{M}}^K(\phi)$, both described in Section 7.1.2 below.

- Next, the term $\Theta(\boldsymbol{u}_h, c_h; z_h)$ was substituted by

$$\Theta_h(\boldsymbol{u}_h, c_h; z_h) := \frac{1}{2}\Big[(\boldsymbol{u}_h\cdot\nabla c_h, z_h)_h + ((q^+ + q^-)\,c_h, z_h)_h - (\boldsymbol{u}_h\,c_h, \nabla z_h)_h\Big], \tag{7.13}$$

where

$$\begin{aligned}(\boldsymbol{u}_h\cdot\nabla c_h, z_h)_h &:= \sum_{K\in\mathcal{T}_n}\int_K \boldsymbol{\Pi_k^{0,K}}\boldsymbol{u}_h\cdot\boldsymbol{\Pi_k^{0,K}}(\nabla c_h)\,\Pi_{k+1}^{0,K} z_h\,\mathrm{d}x\\ ((q^+ + q^-)\,c_h, z_h)_h &:= \sum_{K\in\mathcal{T}_n}\int_K (q^+ + q^-)\,\Pi_{k+1}^{0,K} c_h\,\Pi_{k+1}^{0,K} z_h\,\mathrm{d}x\\ (\boldsymbol{u}_h\,c_h, \nabla z_h)_h &:= \sum_{K\in\mathcal{T}_n}\int_K \boldsymbol{\Pi_k^{0,K}}\boldsymbol{u}_h\,\Pi_{k+1}^{0,K} c_h\ \cdot\boldsymbol{\Pi_k^{0,K}}(\nabla z_h)\,\mathrm{d}x.\end{aligned}$$

- Moreover, the term $\mathcal{D}(\boldsymbol{u}_h; c_h, z_h)$ was replaced by

$$\mathcal{D}_h(\boldsymbol{u}_h; c_h, z_h) := \sum_{K\in\mathcal{T}_n}\mathcal{D}_h^K(\boldsymbol{u}_h; c_h, z_h) \tag{7.14}$$

with local contributions

$$\begin{aligned}\mathcal{D}_h^K(\boldsymbol{u}_h; c_h, z_h) := \int_K D(\boldsymbol{\Pi_k^{0,K}}\boldsymbol{u}_h)\,\boldsymbol{\Pi_k^{0,K}}(\nabla c_h)\cdot\boldsymbol{\Pi_k^{0,K}}(\nabla z_h)\,\mathrm{d}x\\ + \nu_D^K(\boldsymbol{u}_h)\,S_D^K\left(((I-\Pi_{k+1}^{\nabla,K})c_h, (I-\Pi_{k+1}^{\nabla,K})z_h)\right),\end{aligned} \tag{7.15}$$

where $S_D^K(\cdot,\cdot)$ is a stabilization term with certain properties and a constant $\nu_D^K(\boldsymbol{u}_h)$, both described in Section 7.1.2 below.

- Concerning $(q^+\,\widehat{c}, z_h)_{0,\Omega}$, this term was approximated by

$$\left(q^+\,\widehat{c}, z_h\right)_h := \sum_{K\in\mathcal{T}_n}\left[\int_K q^+\,\widehat{c}\,\Pi_{k+1}^{0,K} z_h\,\mathrm{d}x\right].$$

- Regarding the mixed problem, the term $\mathcal{A}(c_h; \boldsymbol{u}_h, \boldsymbol{v}_h)$ was substituted by

$$\mathcal{A}_h(c_h; \boldsymbol{u}_h, \boldsymbol{v}_h) := \sum_{K\in\mathcal{T}_n}\mathcal{A}_h^K(c_h; \boldsymbol{u}_h, \boldsymbol{v}_h) \tag{7.16}$$

with local forms

$$\begin{aligned}\mathcal{A}_h^K(c_h; \boldsymbol{u}_h, \boldsymbol{v}_h) := \int_K A(\Pi_{k+1}^{0,K} c_h)\boldsymbol{\Pi_k^{0,K}}\boldsymbol{u}_h\cdot\boldsymbol{\Pi_k^{0,K}}\boldsymbol{v}_h\,\mathrm{d}x\\ + \nu_{\mathcal{A}}^K(c_h)\,S_{\mathcal{A}}^K((I-\boldsymbol{\Pi_k^{0,K}})\boldsymbol{u}_h, (I-\boldsymbol{\Pi_k^{0,K}})\boldsymbol{v}_h),\end{aligned} \tag{7.17}$$

where, similarly as before, $S_{\mathcal{A}}^K(\cdot,\cdot)$ is a stabilization term and $\nu_{\mathcal{A}}^K(c_h)$ a constant, both described in Section 7.1.2 below.

- Finally, the term $(\boldsymbol{\gamma}(c_h), \boldsymbol{v}_h)_{0,\Omega}$ was replaced by

$$(\boldsymbol{\gamma}(c_h), \boldsymbol{v}_h)_h := \sum_{K \in \mathcal{T}_n} \left[\int_K \boldsymbol{\gamma}(\Pi^{0,K}_{k+1} c_h) \cdot \boldsymbol{\Pi^{0,K}_{k}} \boldsymbol{v}_h \, \mathrm{d}x \right].$$

At this point, we highlight that the bilinear form $B(\cdot, \cdot)$ needs not to be substituted since it is computable for VE functions due to the choice of degrees of freedom (7.6). Further, the right-hand side term $(G, q_h)_{0,\Omega}$ remains unchanged.

Remark 25. Note that we here use the convention that terms which are written in caligraphic letters, such as $\mathcal{M}_h$, $\mathcal{D}_h$ and $\mathcal{A}_h$, include a stabilization term, whereas those in non-caligraphic fashion and those of the form $(\cdot, \cdot)_h$ with subscript h do not. In general, the terms of the type $(\cdot, \cdot)_h$ are approximations of the corresponding (possibly weighted) L^2 scalar products $(\cdot, \cdot)_{0,\Omega}$, obtained by introducing projections onto polynomials for all virtual functions, but not for the data terms that are known exactly.

Construction of the stabilizations

Here, we specify the assumptions on the stabilizations $S^K_{\mathcal{M}}(\cdot, \cdot) : Z_h \times Z_h \to \mathbb{R}$, $S^K_D(\cdot, \cdot) : Z_h \times Z_h \to \mathbb{R}$, and $S^K_{\mathcal{A}}(\cdot, \cdot) : \boldsymbol{V}_h \times \boldsymbol{V}_h \to \mathbb{R}$, in (7.11), (7.14), and (7.16), respectively.

We require that these terms represent computable, symmetric, and positive definite bilinear forms that satisfy, for all $K \in \mathcal{T}_n$, the following property: there exist positive constants $M^{\mathcal{M}}_0$, $M^{\mathcal{M}}_1$, $M^{\mathcal{D}}_0$, $M^{\mathcal{D}}_1$, $M^{\mathcal{A}}_0$, $M^{\mathcal{A}}_1$, which are independent of h and K, such that

$$\begin{aligned} M^{\mathcal{M}}_0 \|z_h\|^2_{0,K} \le S^K_{\mathcal{M}}(z_h, z_h) \le M^{\mathcal{M}}_1 \|z_h\|^2_{0,K} \qquad & \forall z_h \in Z_h \cap \ker(\Pi^{0,K}_{k+1}) \\ M^{\mathcal{D}}_0 \|\nabla z_h\|^2_{0,K} \le S^K_D(z_h, z_h) \le M^{\mathcal{D}}_1 \|\nabla z_h\|^2_{0,K} \qquad & \forall z_h \in Z_h \cap \ker(\Pi^{\nabla,K}_{k+1}) \\ M^{\mathcal{A}}_0 \|\boldsymbol{v}_h\|^2_{0,K} \le S^K_{\mathcal{A}}(\boldsymbol{v}_h, \boldsymbol{v}_h) \le M^{\mathcal{A}}_1 \|\boldsymbol{v}_h\|^2_{0,K} \qquad & \forall \boldsymbol{v}_h \in \boldsymbol{V}_h \cap \ker(\boldsymbol{\Pi^{0,K}_{k}}). \end{aligned} \tag{7.18}$$

Note that continuity follows immediately from the properties:

$$S^K_{\mathcal{M}}(z_h, \widetilde{z_h}) \le \left(S^K_{\mathcal{M}}(z_h, z_h)\right)^{\frac{1}{2}} \left(S^K_{\mathcal{M}}(\widetilde{z_h}, \widetilde{z_h})\right)^{\frac{1}{2}} \le M^{\mathcal{M}}_1 \|z_h\|_{0,K} \|\widetilde{z_h}\|_{0,K},$$

for all $z_h, \widetilde{z_h} \in Z_h \cap \ker(\Pi^{0,K}_{k+1})$; analogously for the other forms. In practice, under the mesh assumptions (**G1**)-(**G3**), one can take the following scaled stabilizations corresponding to the degrees of freedom:

$$\begin{aligned} S^K_{\mathcal{M}}(c_h, z_h) &= |K| \sum_{j=1}^{\dim Z_h(K)} \mathrm{dof}^{Z_h(K)}_j(c_h) \, \mathrm{dof}^{Z_h(K)}_j(z_h) \\ S^K_D(c_h, z_h) &= \sum_{j=1}^{\dim Z_h(K)} \mathrm{dof}^{Z_h(K)}_j(c_h) \, \mathrm{dof}^{Z_h(K)}_j(z_h) \\ S^K_{\mathcal{A}}(\boldsymbol{u}_h, \boldsymbol{v}_h) &= |K| \sum_{j=1}^{\dim V_h(K)} \mathrm{dof}^{\boldsymbol{V}_h(K)}_j(\boldsymbol{u}_h) \, \mathrm{dof}^{\boldsymbol{V}_h(K)}_j(\boldsymbol{v}_h). \end{aligned} \tag{7.19}$$

Regarding the constants appearing in front of the stabilizations in (7.11), (7.14), and (7.16), respectively, we pick:

$$\nu^K_{\mathcal{M}}(\phi) = \left|\Pi^{0,K}_0 \phi\right|, \quad \nu^K_D(\boldsymbol{u}_h) = \nu^K_{\mathcal{M}}(\phi)(d_m + d_t |\boldsymbol{\Pi^{0,K}_0} \boldsymbol{u}_h|), \quad \nu^K_{\mathcal{A}}(c_h) = |A(\Pi^{0,K}_0(c_h))|, \tag{7.20}$$

where $\Pi^{0,K}_0 : L^2(K) \to \mathbb{P}_0(K)$ and $\boldsymbol{\Pi^{0,K}_0} : [L^2(K)]^2 \to [\mathbb{P}_0(K)]^2$ are the L^2 projectors onto scalar- and vector-valued constants, respectively.

Well-posedness of the semidiscrete problem

We firstly define the constants

$$\nu_{\mathcal{M}}^{-} := \min_{K \in \mathcal{T}_n} \nu_{\mathcal{M}}^{K}, \qquad \nu_{\mathcal{M}}^{+} := \max_{K \in \mathcal{T}_n} \nu_{\mathcal{M}}^{K}.$$

Analogously, we introduce $\nu_{\mathcal{D}}^{-}$, $\nu_{\mathcal{D}}^{+}$, $\nu_{\mathcal{A}}^{-}$, and $\nu_{\mathcal{A}}^{+}$. Recalling (2.12) and (2.15), it is easy to check the following (mesh-uniform) bounds for the above constants:

$$\begin{aligned}&\phi_* \le \nu_{\mathcal{M}}^{-} \le \nu_{\mathcal{M}}^{+} \le \phi^* \ , \quad (a^*)^{-1} \le \nu_{\mathcal{A}}^{-} \le \nu_{\mathcal{A}}^{+} \le a_*^{-1},\\ &\phi_* d_m \le \nu_{\mathcal{D}}^{-} \le \nu_{\mathcal{D}}^{+} \le \phi^*(d_m + (d_\ell + d_t)\|\boldsymbol{u}_h\|_{\infty,\Omega}).\end{aligned}$$

Similarly as for their continuous counterparts, the following continuity and coercivity properties for $\mathcal{M}_h(\cdot,\cdot)$ $\mathcal{D}_h(\cdot;\cdot,\cdot)$, and $\mathcal{A}_h(\cdot;\cdot,\cdot)$, defined in (7.11), (7.14), and (7.16), respectively, hold true.

Lemma 7.1.2. *For $\mathcal{M}_h(\cdot,\cdot)$, it holds, for all $c_h, z_h \in Z_h$,*

$$\begin{aligned}\mathcal{M}_h(c_h, z_h) &\le \max\{\phi^*, \nu_{\mathcal{M}}^{+} M_1^{\mathcal{M}}\}\|c_h\|_{0,\Omega}\|z_h\|_{0,\Omega}\\ \mathcal{M}_h(z_h, z_h) &\ge \min\{\phi_*, \nu_{\mathcal{M}}^{-} M_0^{\mathcal{M}}\}\|z_h\|_{0,\Omega}^2.\end{aligned} \tag{7.21}$$

Concerning $\mathcal{D}_h(\cdot;\cdot,\cdot)$, this form satisfies, for all $\boldsymbol{u}_h \in \boldsymbol{V}_h$ and $c_h, z_h \in Z_h$,

$$\begin{aligned}\mathcal{D}_h(\boldsymbol{u}_h; c_h, z_h) &\le \left[\phi^* \left(d_m + \eta\|\boldsymbol{u}_h\|_{\infty,\Omega}(d_\ell + d_t)\right) + \nu_{\mathcal{D}}^{+} M_1^{\mathcal{D}}\right] |c_h|_{1,\mathcal{T}_n}|z_h|_{1,\mathcal{T}_n}\\ \mathcal{D}_h(\boldsymbol{u}_h; z_h, z_h) &\ge \min\{\phi_* d_m, \nu_{\mathcal{D}}^{-} M_0^{\mathcal{D}}\}|z_h|_{1,\mathcal{T}_n}^2.\end{aligned} \tag{7.22}$$

Regarding $\mathcal{A}_h(\cdot;\cdot,\cdot)$, for all $c_h \in Z_h$ and $\boldsymbol{u}_h, \boldsymbol{v}_h \in \boldsymbol{V}_h$, it yields

$$\begin{aligned}\mathcal{A}_h(c_h; \boldsymbol{u}_h, \boldsymbol{v}_h) &\le \max\left\{\frac{1}{a_*}, \nu_{\mathcal{A}}^{+} M_1^{\mathcal{A}}\right\} \|\boldsymbol{u}_h\|_{0,\Omega}\|\boldsymbol{v}_h\|_{0,\Omega}\\ \mathcal{A}_h(c_h; \boldsymbol{v}_h, \boldsymbol{v}_h) &\ge \min\left\{\frac{1}{a^*}, \nu_{\mathcal{A}}^{-} M_0^{\mathcal{A}}\right\} \|\boldsymbol{v}_h\|_{0,\Omega}^2.\end{aligned} \tag{7.23}$$

Thus, $\mathcal{A}_h(c_h;\cdot,\cdot)$ is coercive on the kernel

$$\mathcal{K}_h := \{\boldsymbol{v}_h \in \boldsymbol{V}_h : \, B(\boldsymbol{v}_h, q_h) = 0 \quad \forall q_h \in Q_h\} \subset \mathcal{K} \tag{7.24}$$

with respect to $\|\cdot\|_{\boldsymbol{V}_h}$, where $\mathcal{K}$ is given in (2.22).

Proof. The continuity bound in (7.21) follows directly by using

$$\mathcal{M}_h(c_h, z_h) \le \mathcal{M}_h(c_h, c_h)^{\frac{1}{2}} \mathcal{M}_h(z_h, z_h)^{\frac{1}{2}} \tag{7.25}$$

and then estimating

$$\begin{aligned}\mathcal{M}_h(c_h, c_h) &\le \phi^* \|\Pi_{k+1}^{0,K} c_h\|_{0,K}^2 + \nu_{\mathcal{M}}^{+} M_1^{\mathcal{M}} \|(I - \Pi_{k+1}^{0,K})c_h\|_{0,K}^2\\ &\le \max\{\phi^*, \nu_{\mathcal{M}}^{+} M_1^{\mathcal{M}}\} \left(\|\Pi_{k+1}^{0,K} c_h\|_{0,K}^2 + \|(I - \Pi_{k+1}^{0,K})c_h\|_{0,K}^2\right)\\ &= \max\{\phi^*, \nu_{\mathcal{M}}^{+} M_1^{\mathcal{M}}\} \|c_h\|_{0,K}^2,\end{aligned}$$

where the Pythagorean theorem was applied in the last equality. For the coercivity bound, one can use (2.15), (7.18), and the Pythagorean theorem.

Regarding the continuity estimate for $\mathcal{D}_h(\cdot;\cdot,\cdot)$, by using a splitting of the form (7.25), together with an estimate as in (2.23), one can deduce at the local level

$$\begin{aligned}\mathcal{D}_h^K(\boldsymbol{u}_h; c_h, c_h) &\le \phi^* \left(d_m + \eta\|\boldsymbol{\Pi}_{\boldsymbol{k}}^{\boldsymbol{0,K}} \boldsymbol{u}_h\|_{\infty,\Omega}(d_\ell + d_t)\right) \|\boldsymbol{\Pi}_{\boldsymbol{k}}^{\boldsymbol{0,K}}(\nabla c_h)\|_{0,K}^2\\ &\quad + \left(\nu_{\mathcal{D}}^{+} M_1^{\mathcal{D}}\right) \|\nabla(I - \Pi_{k+1}^{\nabla,K})c_h\|_{0,K}^2\\ &\le \left[\phi^* \left(d_m + \eta\|\boldsymbol{\Pi}_{\boldsymbol{k}}^{\boldsymbol{0,K}} \boldsymbol{u}_h\|_{\infty,\Omega}(d_\ell + d_t)\right) + \nu_{\mathcal{D}}^{+} M_1^{\mathcal{D}}\right] |c_h|_{1,\mathcal{T}_n}^2.\end{aligned} \tag{7.26}$$

By application of a polynomial inverse estimate [59, Lemma 4.5.3], the continuity of the L^2 projector, and the Hölder inequality, we further estimate

$$\|\boldsymbol{\Pi}_{\boldsymbol{k}}^{\boldsymbol{0,K}}\boldsymbol{u}_h\|_{\infty,K} \leq \eta\, h_K^{-1}\|\boldsymbol{\Pi}_{\boldsymbol{k}}^{\boldsymbol{0,K}}\boldsymbol{u}_h\|_{0,K} \leq \eta\, h_K^{-1}\|\boldsymbol{u}_h\|_{0,K} \leq \eta\|\boldsymbol{u}_h\|_{\infty,K}. \tag{7.27}$$

After inserting (7.27) into (7.26), taking the splitting into account, and summing over all elements, the stated bound follows. Concerning the coercivity bound for $\mathcal{D}_h(\cdot,\cdot)$, one can proceed similarly as in (2.25) for the consistency part, and employ (7.18) for the stabilization term, to obtain elementwise

$$\mathcal{D}_h^K(\boldsymbol{u}_h; z_h, z_h) \geq \min\{\phi_* d_m, \nu_{\mathcal{D}}^- M_0^{\mathcal{D}}\}\left[\|\boldsymbol{\Pi}_{\boldsymbol{k}}^{\boldsymbol{0,K}}\nabla z_h\|_{0,K}^2 + \|\nabla(I-\Pi_{k+1}^{\nabla,K})z_h\|_{0,K}^2\right].$$

We now note that the definitions of $\Pi_{k+1}^{\nabla,K}$ and $\boldsymbol{\Pi}_{\boldsymbol{k}}^{\boldsymbol{0,K}}$ easily yield

$$\|\nabla(I-\Pi_{k+1}^{\nabla,K})z_h\|_{0,K} \geq \|(I-\boldsymbol{\Pi}_{\boldsymbol{k}}^{\boldsymbol{0,K}})\nabla z_h\|_{0,K}. \tag{7.28}$$

Then, the estimate follows with (7.28), the Pythagorean theorem and summation over all elements.

The estimates for $\mathcal{A}_h(\cdot;\cdot,\cdot)$ are derived in a similar fashion as those for $\mathcal{M}_h(\cdot,\cdot)$, using (2.15). Coercivity on $\mathcal{K}_h$ follows from the fact

$$\mathcal{K}_h \equiv \{\boldsymbol{v}_h \in \boldsymbol{V}_h : \operatorname{div}\boldsymbol{v}_h = 0\} \subset \mathcal{K},$$

owing to the definition of $\boldsymbol{V}_h(K)$ in (7.4). □

Well-posedness of problem (7.10) can be shown by combining the results in [174] for parabolic problems with those in [30, 60] for mixed problems, using Lemma 7.1.2. More precisely, in the spirit of the two-step strategy applied in [90] for FEM, one can firstly show that, for any given $c_h(t) \in L^\infty(\Omega)$, $t \in J$, the mixed problem

$$\begin{aligned}\mathcal{A}_h(c_h; \boldsymbol{u}_h, \boldsymbol{v}_h) + B(\boldsymbol{v}_h, p_h) &= (\boldsymbol{\gamma}(c_h), \boldsymbol{v}_h)_h\\ B(\boldsymbol{u}_h, q_h) &= -\left(G, q_h\right)_{0,\Omega}\end{aligned}$$

admits a unique solution by applying the techniques in [30, 60], and then, by using the Gronwall lemma and Picard-Lindelöf (see e.g. [56, Ch.1.10]), that $c_h(t)$ is uniquely determined by the discrete concentration equation

$$\mathcal{M}_h\left(\frac{\partial c_h}{\partial t}, z_h\right) + \Theta_h(\boldsymbol{u}_h, c_h; z_h) + \mathcal{D}_h(\boldsymbol{u}_h; c_h, z_h) = \left(q^+\,\widehat{c}, z_h\right)_h,$$

see also [174]. We do not write here the details since we focus directly on the fully discrete case, see the next section.

7.1.3 Fully discrete formulation

Here, our goal is to formulate a fully discrete version of (7.10).

To start with, we introduce a sequence of time steps $t_n = n\tau$, $n = 0,\ldots,N$, with time step size τ. Next, we define $\boldsymbol{u}^n := \boldsymbol{u}(t_n)$, $p^n := p(t_n)$, $c^n := c(t_n)$, $G^n := G(t_n)$, $(q^+)^n := q^+(t_n)$, and $\widehat{c}^n := \widehat{c}(t_n)$ as the evaluations of the corresponding functions at time t_n, $n = 0,\ldots,N$. Moreover, we denote by $\boldsymbol{U}^n \approx \boldsymbol{u}_h(t_n)$, $P^n \approx p_h(t_n)$, and $C^n \approx c_h(t_n)$, the approximations of the semidiscrete solutions at those times when using a time integrator method. Among many time discretization schemes, we here make a computationally cheap choice by choosing a backward Euler method that is explicit in the nonlinear terms. The fully discrete system consequently reads as follows:

- for $n = 0$: Given $c_{0,h} \in Z_h$, solve

$$\begin{aligned}\mathcal{A}_h(c_{0,h}; \boldsymbol{U}^n, \boldsymbol{v}_h) + B(\boldsymbol{v}_h, P^n) &= (\boldsymbol{\gamma}(c_{0,h}), \boldsymbol{v}_h)_h\\ B(\boldsymbol{U}^n, q_h) &= -\left(G^n, q_h\right)_{0,\Omega}\end{aligned} \tag{7.29}$$

  for all $\boldsymbol{v}_h \in \boldsymbol{V}_h$ and $q_h \in Q_h$.

- for $n = 1, \dots, N$: Solve first the concentration equation for C^n:

$$\mathcal{M}_h\left(\frac{C^n - C^{n-1}}{\tau}, z_h\right) + \Theta_h(\boldsymbol{U}^{n-1}; C^n, z_h) + \mathcal{D}_h(\boldsymbol{U}^{n-1}; C^n, z_h) = \left((q^+)^n\, \widehat{c}^{\,n}, z_h\right)_h, \quad (7.30)$$

for all $z_h \in Z_h$, where $C^0 := c_{0,h}$. Then, solve the mixed problem for $\boldsymbol{U}^n$ and P^n:

$$\begin{aligned} \mathcal{A}_h(C^n; \boldsymbol{U}^n, \boldsymbol{v}_h) + B(\boldsymbol{v}_h, P^n) &= (\boldsymbol{\gamma}(C^n), \boldsymbol{v}_h)_h \\ B(\boldsymbol{U}^n, q_h) &= -\left(G^n, q_h\right)_{0,\Omega}, \end{aligned} \quad (7.31)$$

for all $\boldsymbol{v}_h \in \boldsymbol{V}_h$ and $q_h \in Q_h$.

Lemma 7.1.3. *Given $\tau > 0$, provided that $G^n, (q^+)^n, P^n, C^n \in L^\infty(\Omega)$, $\boldsymbol{\gamma}(C^n) \in [L^2(\Omega)]^2$, and $\boldsymbol{U}^n \in [L^\infty(\Omega)]^2$, for all $n = 0, \dots, N$, the formulation* (7.29)-(7.31) *is uniquely solvable.*

Proof. Similarly as for the semidiscrete case, well-posedness of (7.29) and (7.31) follows by using the tools of [30, 60]. Regarding (7.30), we firstly rewrite that equation as

$$\begin{aligned} \mathcal{M}_h\left(C^n, z_h\right) + \tau\left[\Theta_h(\boldsymbol{U}^{n-1}; C^n, z_h) + \mathcal{D}_h(\boldsymbol{U}^{n-1}; C^n, z_h)\right] \\ = \tau\left((q^+)^n\, \widehat{c}^{\,n}, z_h\right)_h + \mathcal{M}_h\left(C^{n-1}, z_h\right). \end{aligned} \quad (7.32)$$

We observe that all of the term are continuous with respect to the norm $\|\cdot\|_{1,\mathcal{T}_n}$. More precisely, for $\mathcal{M}_h(\cdot,\cdot)$ and $\mathcal{D}_h(\boldsymbol{U}^{n-1};\cdot,\cdot)$, continuity follows from Lemma 7.1.2 and the definition of the broken H^1 norm. Next, for the term involving $(q^+)^n$, we simply apply the Cauchy-Schwarz inequality and the stability of the L^2 projector. Finally, for the term with Θ_h, we estimate

$$\begin{aligned} \Theta_h(\boldsymbol{U}^{n-1}; C^n, z_h) &= \frac{1}{2}\left[\left(\boldsymbol{U}^{n-1}\cdot\nabla C^n, z_h\right)_h + \left((q^+ + q^-)\, C^n, z_h\right)_h - \left(\boldsymbol{U}^{n-1} C^n, \nabla z_h\right)_h\right] \\ &\leq \eta\left[\|\boldsymbol{U}^{n-1}\|_{\infty,\Omega}(|C^n|_{1,\mathcal{T}_n} + \|C^n\|_{0,\Omega}) + \|q^+ + q^-\|_{\infty,\Omega}\|C^n\|_{0,\Omega}\right]\|z_h\|_{1,\mathcal{T}_n}, \end{aligned}$$

where we also employed an inverse inequality as in (7.27). Thus, by the Lax-Milgram lemma, it only remains to show that the left-hand side of (7.32) is coercive with respect to $\|\cdot\|_{1,\mathcal{T}_n}$. This is however a direct consequence of

$$\Theta_h(\boldsymbol{U}^{n-1}; z_h, z_h) = \frac{1}{2}\left((q^+ + q^-)\, z_h, z_h\right)_h \geq 0,$$

owing to the fact that q^+ and q^- are non-negative, and the coercivity bounds (7.21) and (7.22). □

Note that both problems (7.30) and (7.31) represent linear systems of equations which are decoupled from each other in the sense that, firstly, given $\widehat{c}^{\,n}$ and $(q^+)^n$, one can determine C^n with knowledge of $\boldsymbol{U}^{n-1}$ only, and then one can use C^n to compute $\boldsymbol{U}^n$ and P^n. The quantity P^n does in fact not influence the calculation of C^n directly, but rather takes the role of a Lagrange multiplier and derived variable. This decoupling, combined with the fact that the systems to be solved at each time step are linear, makes the method quite cheap per iteration.

7.2 Error analysis for the fully discrete problem

The error analysis is performed in two steps: firstly, we estimate the discretization errors for the velocity and pressure, $\|\boldsymbol{u}^n - \boldsymbol{U}^n\|_{0,\Omega}$ and $\|p^n - P^n\|_{0,\Omega}$, respectively, and then, in the second step, the concentration error $\|c^n - C^n\|_{0,\Omega}$. In the following analysis, we assume all the needed regularity of the exact solution. Although such high regularity will not be often available in practice, the purpose of the following analysis is to give a theoretical backbone to the proposed scheme and to investigate its potential accuracy in the most favorable scenario.

7.2.1 An auxiliary result

The subsequent technical lemma will serve as an auxiliary result in the derivation of the error estimates and will be used in several occasions.

Lemma 7.2.1. *Let $r, s, t \in \mathbb{N}_0$. Denote by Π^0_r and $\boldsymbol{\Pi^0_s}$, the elementwise defined L^2 projectors onto scalar- and vector-valued polynomials of degree at most r and s, respectively. Given a scalar function $\sigma \in H^{m_r}(\mathcal{T}_n)$, $0 \le m_r \le r+1$, let $\kappa(\sigma)$ be a tensor-valued piecewise Lipschitz continuous function with respect to σ. Further, let $\widehat{\sigma} \in L^2(\Omega)$, and let $\boldsymbol{\chi}$ and $\boldsymbol{\psi}$ be vector-valued functions. We assume that $\kappa(\sigma) \in [L^\infty(\Omega)]^{2\times 2}$, $\boldsymbol{\chi} \in [H^{m_s}(\mathcal{T}_n) \cap L^\infty(\Omega)]^2$, $\boldsymbol{\psi} \in [L^2(\Omega)]^2$, and $\kappa(\sigma)\boldsymbol{\chi} \in [H^{m_t}(\mathcal{T}_n)]^2$, for some $0 \le m_s \le s+1$ and $0 \le m_t \le t+1$. Then,*

$$\begin{aligned}&(\kappa(\sigma)\boldsymbol{\chi}, \boldsymbol{\psi})_{0,\Omega} - (\kappa(\Pi^0_r\widehat{\sigma})\boldsymbol{\Pi^0_s}\boldsymbol{\chi}, \boldsymbol{\Pi^0_t}\boldsymbol{\psi})_{0,\Omega}\\ &\le \eta\big[h^{m_t}|\kappa(\sigma)\boldsymbol{\chi}|_{m_t,\mathcal{T}_n} + h^{m_s}|\boldsymbol{\chi}|_{m_s,\mathcal{T}_n}\|\kappa(\sigma)\|_{\infty,\Omega} + (h^{m_r}|\sigma|_{m_r,\mathcal{T}_n} + \|\sigma - \widehat{\sigma}\|_{0,\Omega})\|\boldsymbol{\chi}\|_{\infty,\Omega}\big]\|\boldsymbol{\psi}\|_{0,\Omega}.\end{aligned}$$

Proof. We firstly write

$$\begin{aligned}&(\kappa(\sigma)\boldsymbol{\chi}, \boldsymbol{\psi})_{0,\Omega} - (\kappa(\Pi^0_r\widehat{\sigma})\boldsymbol{\Pi^0_s}\boldsymbol{\chi}, \boldsymbol{\Pi^0_t}\boldsymbol{\psi})_{0,\Omega}\\ &= [(\kappa(\sigma)\boldsymbol{\chi}, \boldsymbol{\psi})_{0,\Omega} - (\kappa(\Pi^0_r\sigma)\boldsymbol{\Pi^0_s}\boldsymbol{\chi}, \boldsymbol{\Pi^0_t}\boldsymbol{\psi})_{0,\Omega}] + ((\kappa(\Pi^0_r\sigma) - \kappa(\Pi^0_r\widehat{\sigma}))\boldsymbol{\Pi^0_s}\boldsymbol{\chi}, \boldsymbol{\Pi^0_t}\boldsymbol{\psi})_{0,\Omega}.\end{aligned} \tag{7.33}$$

Then, for the first part on the right-hand side of (7.33), we recall that $\boldsymbol{\Pi^0_t}$ is an L^2 projection and derive, on each element $K \in \mathcal{T}_n$,

$$\begin{aligned}&(\kappa(\sigma)\boldsymbol{\chi}, \boldsymbol{\psi})_{0,K} - (\kappa(\Pi^{0,K}_r\sigma)\boldsymbol{\Pi^{0,K}_s}\boldsymbol{\chi}, \boldsymbol{\Pi^{0,K}_t}\boldsymbol{\psi})_{0,K}\\ &= [(\kappa(\sigma)\boldsymbol{\chi}, \boldsymbol{\psi})_{0,K} - (\boldsymbol{\Pi^{0,K}_t}(\kappa(\sigma)\boldsymbol{\chi}), \boldsymbol{\psi})_{0,K}]\\ &\quad + [(\boldsymbol{\Pi^{0,K}_t}(\kappa(\sigma)\boldsymbol{\chi}), \boldsymbol{\psi})_{0,K} - (\boldsymbol{\Pi^{0,K}_t}(\kappa(\sigma)\boldsymbol{\Pi^{0,K}_s}\boldsymbol{\chi}), \boldsymbol{\psi})_{0,K}]\\ &\quad + [(\boldsymbol{\Pi^{0,K}_t}(\kappa(\sigma)\boldsymbol{\Pi^{0,K}_s}\boldsymbol{\chi}), \boldsymbol{\psi})_{0,K} - (\boldsymbol{\Pi^{0,K}_t}(\kappa(\Pi^{0,K}_r\sigma)\boldsymbol{\Pi^{0,K}_s}\boldsymbol{\chi}), \boldsymbol{\psi})_{0,K}]\\ &\le \eta\big[h^{m_t}|\kappa(\sigma)\boldsymbol{\chi}|_{m_t,K} + h^{m_s}|\boldsymbol{\chi}|_{m_s,K}\|\kappa(\sigma)\|_{\infty,K} + h^{m_r}|\sigma|_{m_r,K}\|\boldsymbol{\Pi^{0,K}_s}\boldsymbol{\chi}\|_{\infty,K}\big]\|\boldsymbol{\psi}\|_{0,K},\end{aligned}$$

where in the last step we used Lemma 7.1.1 and the fact that κ is Lipschitz continuous with respect to σ. The term $\|\boldsymbol{\Pi^{0,K}_s}\boldsymbol{\chi}\|_{\infty,K}$ is estimated as in (7.27). Concerning the second part on the right-hand side of (7.33), we have, for each $K \in \mathcal{T}_n$,

$$\begin{aligned}((\kappa(\Pi^0_r\sigma) - \kappa(\Pi^0_r\widehat{\sigma}))\boldsymbol{\Pi^0_s}\boldsymbol{\chi}, \boldsymbol{\Pi^0_t}\boldsymbol{\psi})_{0,K} &\le \|(\kappa(\Pi^0_r\sigma) - \kappa(\Pi^0_r\widehat{\sigma})\|_{0,K}\|\boldsymbol{\Pi^0_s}\boldsymbol{\chi}\|_{\infty,K}\|\boldsymbol{\Pi^0_t}\boldsymbol{\psi}\|_{0,K}\\ &\le \|\sigma - \widehat{\sigma}\|_{0,K}\|\boldsymbol{\chi}\|_{\infty,K}\|\boldsymbol{\psi}\|_{0,K},\end{aligned}$$

where we used again the Lipschitz continuity of κ, the continuity properties of the L^2 projectors, and the bound (7.27). The assertion of the lemma follows after combining the estimates and summing over all elements. □

Note that the above lemma can be easily transferred to the cases where σ, $\kappa(\sigma)$, χ, and ψ are scalar, and to vector-valued $\boldsymbol{\sigma}$, $\boldsymbol{\chi}$ and scalar $\kappa(\boldsymbol{\sigma})$, ψ.

In the special case of $\chi = 1$ and vector-valued $\boldsymbol{\kappa}$, an adaptation of Lemma 7.2.1 gives

$$\begin{aligned}&(\boldsymbol{\kappa}(\sigma), \boldsymbol{\psi})_{0,\Omega} - (\boldsymbol{\kappa}(\Pi^0_r\widehat{\sigma}), \boldsymbol{\Pi^0_t}\boldsymbol{\psi})_{0,\Omega}\\ &\qquad \le \eta\big[h^{m_t}|\boldsymbol{\kappa}(\sigma)|_{m_t,\mathcal{T}_n} + h^{m_r}|\sigma|_{m_r,\mathcal{T}_n} + \|\sigma - \widehat{\sigma}\|_{0,\Omega}\big]\|\boldsymbol{\psi}\|_{0,\Omega}.\end{aligned} \tag{7.34}$$

7.2.2 Bounds on $\|\boldsymbol{u}^n - \boldsymbol{U}^n\|_{0,\Omega}$ and $\|p^n - P^n\|_{0,\Omega}$

We consider the mixed problem

$$\begin{aligned}\mathcal{A}_h(C^n; \boldsymbol{U}^n, \boldsymbol{v}_h) + B(\boldsymbol{v}_h, P^n) &= (\boldsymbol{\gamma}(C^n), \boldsymbol{v}_h)_h\\ B(\boldsymbol{U}^n, q_h) &= -(G^n, q_h)_{0,\Omega}\,,\end{aligned} \tag{7.35}$$

where $C^n \in Z_h$ is the numerical solution to the concentration equation (7.30) for $n = 1, \ldots, N$, and $C^0 = c_{0,h}$. The goal is to prove an upper bound for $\|\boldsymbol{u}^n - \boldsymbol{U}^n\|_{0,\Omega}$ and $\|p^n - P^n\|_{0,\Omega}$ with respect to $\|c^n - C^n\|_{0,\Omega}$. For the analysis, we basically follow the ideas of [30, 60] with the major differences that, here, $\mathcal{A}_h(C^n; \cdot, \cdot)$ is not consistent with respect to $[\mathbb{P}_k(K)]^2$ due to presence of C^n, and, additionally, the right-hand side of (7.35) is inhomogeneous.

Theorem 7.2.2. *Given $C^n \in Z_h$, let $(\boldsymbol{U}^n, P^n) \in \boldsymbol{V}_h \times Q_h$ be the solution to (7.35). Let us assume that, for the exact solution $(\boldsymbol{u}^n, p^n, c^n)$ to (2.19) at time t_n, it holds $\boldsymbol{u}^n \in [H^{k+1}(\mathcal{T}_n)]^2$, $p^n \in H^{k+1}(\mathcal{T}_n)$, and $c^n \in H^{k+1}(\mathcal{T}_n)$. Further, we suppose that $\boldsymbol{\gamma}(c)$ and $A(c)$ are piecewise Lipschitz continuous functions with respect to $c \in L^2(\Omega)$, and that $\boldsymbol{\gamma}(c^n), A(c^n)\boldsymbol{u}^n \in [H^{k+1}(\mathcal{T}_n)]^2$. Then, the following error estimates hold for all $k \in \mathbb{N}_0$:*

$$\begin{aligned} \|\boldsymbol{U}^n - \boldsymbol{u}^n\|_{0,\Omega} &\leq \|C^n - c^n\|_{0,\Omega}\, \zeta_1^n(\boldsymbol{u}^n) + h^{k+1}\, \zeta_2^n(\boldsymbol{u}^n, c^n, \boldsymbol{\gamma}(c^n), A(c^n)\boldsymbol{u}^n) \\ \|P^n - p^n\|_{0,\Omega} &\leq \|C^n - c^n\|_{0,\Omega}\, \zeta_3^n(\boldsymbol{u}^n) + h^{k+1}\, \zeta_4^n(\boldsymbol{u}^n, c^n, \boldsymbol{\gamma}(c^n), A(c^n)\boldsymbol{u}^n, p^n), \end{aligned}$$

where ζ_1^n-ζ_4^n are positive constants independent of h and depending only on the specified functions.

Proof. The estimate for $\|\boldsymbol{U}^n - \boldsymbol{u}^n\|_{0,\Omega}$ can be obtained as follows.

By using the second equality in (7.35), we have $\operatorname{div} \boldsymbol{U}^n = \Pi_k^0 G^n$ (use that $\operatorname{div} \boldsymbol{U}^n \in \mathbb{P}_k(K)$ for every $K \in \mathcal{T}_n$), where we recall that $(\Pi_k^0)_{|K} = \Pi_k^{0,K}$. Define now the interpolant $\boldsymbol{u}_I^n \in \boldsymbol{V}_h$ via the degrees of freedom (7.6):

$$\operatorname{dof}_i^{\boldsymbol{V}_h}(\boldsymbol{u}_I^n) = \operatorname{dof}_i^{\boldsymbol{V}_h}(\boldsymbol{u}^n), \quad i = 1, \ldots, \operatorname{dim} V_h.$$

Then, it holds [30, eq.(28)]

$$\|\boldsymbol{u}^n - \boldsymbol{u}_I^n\|_{0,\Omega} \leq \eta\, h^{k+1} \|\boldsymbol{u}^n\|_{k+1,\mathcal{T}_n}. \tag{7.36}$$

Moreover, one has $\operatorname{div} \boldsymbol{u}_I^n = \Pi_k^0 G^n$. Thus, setting $\boldsymbol{\delta}^n := \boldsymbol{U}^n - \boldsymbol{u}_I^n$, it holds that $\boldsymbol{\delta}^n \in \mathcal{K}_h \subset \mathcal{K}$, where $\mathcal{K}_h$ and $\mathcal{K}$ were defined in (7.24) and (2.22), respectively, and therefore, $\|\boldsymbol{\delta}^n\|_{\boldsymbol{V}_h} = \|\boldsymbol{\delta}^n\|_{0,\Omega}$. Owing to the assumptions on $a(\cdot)$ in (2.15) together with (7.23), we have, further using (7.35) with $\boldsymbol{v}_h = \boldsymbol{\delta}^n \in \mathcal{K}_h$ and (2.19),

$$\begin{aligned} \alpha \|\boldsymbol{\delta}^n\|_{0,\Omega}^2 \leq \mathcal{A}_h(C^n; \boldsymbol{\delta}^n, \boldsymbol{\delta}^n) &= \mathcal{A}_h(C^n; \boldsymbol{U}^n, \boldsymbol{\delta}^n) - \mathcal{A}_h(C^n; \boldsymbol{u}_I^n, \boldsymbol{\delta}^n) \\ &= (\boldsymbol{\gamma}(C^n), \boldsymbol{\delta}^n)_h - \mathcal{A}_h(C^n; \boldsymbol{u}_I^n, \boldsymbol{\delta}^n) \\ &= [(\boldsymbol{\gamma}(C^n), \boldsymbol{\delta}^n)_h - (\boldsymbol{\gamma}(c^n), \boldsymbol{\delta}^n)_{0,\Omega}] + \mathcal{A}_h(C^n; \boldsymbol{u}^n - \boldsymbol{u}_I^n, \boldsymbol{\delta}^n) \\ &\quad + \Big[\mathcal{A}(c^n; \boldsymbol{u}^n, \boldsymbol{\delta}^n) - \mathcal{A}_h(C^n; \boldsymbol{u}^n, \boldsymbol{\delta}^n) \Big] \\ &=: T_1 + T_2 + T_3. \end{aligned} \tag{7.37}$$

The terms T_1-T_3 are estimated as follows:

- term T_1: We use equation (7.34) with $\boldsymbol{\kappa} = \boldsymbol{\gamma}$, $\sigma = c^n$, $\widehat{\sigma} = C^n$, $\boldsymbol{\psi} = \boldsymbol{\delta}^n$, $r = k+1$, $t = k$, and $m_r = m_t = k+1$, and obtain

$$\begin{aligned} |T_1| &= |(\boldsymbol{\gamma}(c^n), \boldsymbol{\delta}^n)_{0,\Omega} - (\boldsymbol{\gamma}(\Pi_{k+1}^0 C^n), \boldsymbol{\Pi_k^0} \boldsymbol{\delta}^n)_{0,\Omega}| \\ &\leq \eta \big[h^{k+1} (|\boldsymbol{\gamma}(c^n)|_{k+1,\mathcal{T}_n} + |c^n|_{k+1,\mathcal{T}_n}) + \|c^n - C^n\|_{0,\Omega} \big] \|\boldsymbol{\delta}^n\|_{0,\Omega}. \end{aligned}$$

- term T_2: Owing to the continuity properties (7.23) of $\mathcal{A}_h(\cdot;\cdot,\cdot)$ and the interpolation error estimate (7.36), it holds

$$|T_2| = |\mathcal{A}_h(C^n; \boldsymbol{u}^n - \boldsymbol{u}_I^n, \boldsymbol{\delta}^n)| \leq \eta \|\boldsymbol{u}^n - \boldsymbol{u}_I^n\|_{0,\Omega} \|\boldsymbol{\delta}^n\|_{0,\Omega} \leq \eta\, h^{k+1} \|\boldsymbol{u}^n\|_{k+1,\mathcal{T}_n} \|\boldsymbol{\delta}^n\|_{0,\Omega}.$$

- term T_3: We have

$$\begin{aligned} |T_3| &= |\mathcal{A}(c^n; \boldsymbol{u}^n, \boldsymbol{\delta}^n) - \mathcal{A}_h(C^n; \boldsymbol{u}^n, \boldsymbol{\delta}^n)| \\ &\leq |(A(c^n)\boldsymbol{u}^n, \boldsymbol{\delta}^n)_{0,\Omega} - (A(\Pi_{k+1}^0 C^n) \boldsymbol{\Pi_k^0} \boldsymbol{u}^n, \boldsymbol{\Pi_k^0} \boldsymbol{\delta}^n)_{0,\Omega}| \\ &\quad + \left| \sum_{K \in \mathcal{T}_n} \nu_{\mathcal{A}}^K(C^n)\, S_{\mathcal{A}}^K((I - \boldsymbol{\Pi_k^{0,K}})\boldsymbol{u}^n, (I - \boldsymbol{\Pi_k^{0,K}})\boldsymbol{\delta}^n) \right| \\ &=: T_3^A + T_3^B. \end{aligned}$$

For the term T_3^A, we use Lemma 7.2.1 with $\kappa = A$, $\sigma = c^n$, $\widehat{\sigma} = C^n$, $\boldsymbol{\chi} = \boldsymbol{u}^n$, $\boldsymbol{\psi} = \boldsymbol{\delta}^n$, $r = k+1$, $s = t = k$, and $m_r = m_s = m_t = k+1$, to get

$$T_3^A \leq \eta \Big[h^{k+1} \big(|A(c^n)\boldsymbol{u}^n|_{k+1,\mathcal{T}_n} + |\boldsymbol{u}^n|_{k+1,\mathcal{T}_n} \|A(c^n)\|_{\infty,\Omega} + |c^n|_{k+1,\mathcal{T}_n} \|\boldsymbol{u}^n\|_{\infty,\Omega} \big) \\ + \|c^n - C^n\|_{0,\Omega} \|\boldsymbol{u}^n\|_{\infty,\Omega} \Big] \|\boldsymbol{\delta}^n\|_{0,\Omega}.$$

On the other hand, the term T_3^B can be estimated with (7.18), (2.15), and Lemma (7.1.1):

$$T_3^B \leq \eta\, h^{k+1} |\boldsymbol{u}^n|_{k+1,\mathcal{T}_n} \|\boldsymbol{\delta}^n\|_{0,\Omega}.$$

After plugging the bounds obtained for T_1-T_3 into (7.37), dividing by $\|\boldsymbol{\delta}^n\|_{0,\Omega}$, using the triangle inequality in the form

$$\|\boldsymbol{U}^n - \boldsymbol{u}^n\|_{0,\Omega} \leq \|\boldsymbol{\delta}^n\|_{0,\Omega} + \|\boldsymbol{u}^n - \boldsymbol{u}_I^n\|_{0,\Omega},$$

and employing (7.36), the convergence result follows.

The error estimate for the term $\|P^n - p^n\|_{0,\Omega}$ follows easily by combining the above ideas with the argument in [60, Theorem 6.1] and is therefore not shown. □

7.2.3 Bounds on $\|c^n - C^n\|_{0,\Omega}$

For fixed $\boldsymbol{u}(t) \in \boldsymbol{V}$ and $t \in J$, we define the projector $\mathcal{P}_c : Z \to Z_h$ (that to each $c \in Z$ associates $\mathcal{P}_c c \in Z_h$) by

$$\Gamma_{c,h}(\boldsymbol{u}(t); \mathcal{P}_c c, z_h) = \Gamma_c(\boldsymbol{u}(t); c, z_h), \tag{7.38}$$

for all $z_h \in Z_h$, where

$$\begin{aligned} \Gamma_{c,h}(\boldsymbol{u}; c, z_h) &:= \mathcal{D}_h(\boldsymbol{u}; c, z_h) + \Theta_h(\boldsymbol{u}; c, z_h) + (c, z_h)_h \\ \Gamma_c(\boldsymbol{u}; c, z_h) &:= \mathcal{D}(\boldsymbol{u}; c, z_h) + \Theta(\boldsymbol{u}; c, z_h) + (c, z_h)_{0,\Omega}, \end{aligned} \tag{7.39}$$

with

$$(c, z_h)_h := \sum_{K \in \mathcal{T}_n} \int_K c\, (\Pi_{k+1}^{0,K} z_h)\, \mathrm{d}x.$$

Lemma 7.2.3. *The projector $\mathcal{P}_c : Z \to Z_h$ given in* (7.38) *is well-defined under the assumption that $\boldsymbol{u}$, q^+, and q^- are bounded in $L^\infty(\Omega)$ for all $t \in J$.*

Proof. By the Lax-Milgram lemma, we have to show that the left-hand side of (7.38) defines a continuous and coercive bilinear form and that the right-hand side is a continuous functional with respect to $\|\cdot\|_{1,\mathcal{T}_n}$. Continuity of the latter one is obtained by combining (2.23) with

$$\begin{aligned} \Theta(\boldsymbol{u}; c, z_h) + (c, z_h)_{0,\Omega} &= \frac{1}{2} \Big[(\boldsymbol{u} \cdot \nabla c, z_h)_{0,\Omega} + ((q^+ + q^- + 2)c, z_h)_{0,\Omega} - (\boldsymbol{u}\, c, \nabla z_h)_{0,\Omega} \Big] \\ &\leq \frac{1}{2} \left[\|\boldsymbol{u}\|_{\infty,\Omega} (|c|_{1,\mathcal{T}_n} + \|c\|_{0,\Omega}) + \|q^+ + q^- + 2\|_{\infty,\Omega} \|c\|_{0,\Omega} \right] \|z_h\|_{1,\mathcal{T}_n}. \end{aligned}$$

By using (7.22) and performing similar computations as in the proof of Lemma 7.1.3, continuity of $\Gamma_{c,h}$ follows:

$$\Gamma_{c,h}(\boldsymbol{u}; c, z_h) \leq \eta\, \zeta(\boldsymbol{u}, q^+, q^-) \|c\|_{1,\mathcal{T}_n} \|z_h\|_{1,\mathcal{T}_n}, \tag{7.40}$$

where ζ only depends on the specified functions. Regarding the coercivity of $\Gamma_{c,h}$, we firstly estimate

$$\Theta_h(\boldsymbol{u}; z_h, z_h) + (z_h, z_h)_h = \sum_{K \in \mathcal{T}_n} \left(\left(\frac{1}{2}(q^+ + q^-) + 1 \right) \Pi_{k+1}^{0,K} z_h, \Pi_{k+1}^{0,K} z_h \right)_{0,K} \geq \|\Pi_{k+1}^0 z_h\|_{0,\Omega}^2,$$

where we recall that $(\Pi_{k+1}^0)_{|K} = \Pi_{k+1}^{0,K}$ for all $K \in \mathcal{T}_n$. Then, combining this result with (7.22) yields

$$\Gamma_{c,h}(\boldsymbol{u}; z_h, z_h) \geq \eta \left[|z_h|_{1,\mathcal{T}_n}^2 + \|\Pi_{k+1}^{0,K} z_h\|_{0,\Omega}^2 \right] \geq \eta \left[|z_h|_{1,\mathcal{T}_n}^2 + \|\overline{z_h}\|_{0,\Omega}^2 \right],$$

with $\overline{z_h}$ denoting the $L^2(\Omega)$ projection of z_h onto $\mathbb{P}_0(\Omega)$. Since $\overline{z_h}$ coincides with the average of z_h, one can use a Poincaré-Friedrichs inequality, see e.g. [57], to deduce

$$|z_h|^2_{1,\mathcal{T}_n} + \|\overline{z_h}\|^2_{0,\Omega} \geq C_p^{-1}\text{diam}(\Omega)^{-1}\|z_h\|^2_{1,\mathcal{T}_n}$$

and consequently the coercivity of $\Gamma_{c,h}$. □

Lemma 7.2.4. *Assume that, for all $t \in J$, $\boldsymbol{u} \in [H^{k+1}(\mathcal{T}_n) \cap L^\infty(\Omega)]^2$, $c \in H^{k+2}(\mathcal{T}_n) \cap W^{1,\infty}(\mathcal{T}_n)$, $q^+, q^- \in L^\infty(\Omega)$, $(q^+ + q^-)c \in H^{k+1}(\mathcal{T}_n)$, $\boldsymbol{u}\,c \in [H^{k+1}(\mathcal{T}_n)]^2$, $\boldsymbol{u} \cdot \nabla c \in H^{k+1}(\mathcal{T}_n)$, and $D(\boldsymbol{u})\nabla c \in [H^{k+1}(\mathcal{T}_n)]^2$. Then, the following error bounds for $c - \mathcal{P}_c c$, where $\mathcal{P}_c c$ is defined in (7.38), hold for all $k \in \mathbb{N}_0$:*

$$\begin{aligned} \|c - \mathcal{P}_c c\|_{1,\mathcal{T}_n} &\leq h^{k+1}\,\xi_1(c, \boldsymbol{u}, q^+, q^-, D(\boldsymbol{u})\nabla c, \nabla c, (q^+ + q^-)c, \boldsymbol{u} \cdot \nabla c, \boldsymbol{u}\,c), \\ \|c - \mathcal{P}_c c\|_{0,\Omega} &\leq h^{k+2}\,\xi_0(c, \boldsymbol{u}, q^+, q^-, D(\boldsymbol{u})\nabla c, \nabla c, (q^+ + q^-)c, \boldsymbol{u} \cdot \nabla c, \boldsymbol{u}\,c), \end{aligned} \tag{7.41}$$

where the constants $\xi_1, \xi_0 > 0$ only depend on the listed terms and are independent of h.

Proof. We focus on the error estimate in the broken H^1 norm at a fixed time $t \in J$. Firstly, we state the following result. Given $c \in H^{k+2}(\mathcal{T}_n)$, there exists an interpolant $c_I \in Z_h$ such that the following bounds hold true (see for instance [43, 58, 67]):

$$\|c - c_I\|_{0,\Omega} \leq \eta\, h^{k+2}\|c\|_{k+2,\mathcal{T}_n}, \quad \|c - c_I\|_{1,\mathcal{T}_n} \leq \eta\, h^{k+1}\|c\|_{k+2,\mathcal{T}_n}. \tag{7.42}$$

After denoting $\nu := \mathcal{P}_c c - c_I$, one obtains with the coercivity of $\Gamma_{c,h}$, see the proof of Lemma 7.2.3, and the definition of $\mathcal{P}_c c$ in (7.38),

$$\begin{aligned} M\|\nu\|^2_{1,\mathcal{T}_n} &\leq \Gamma_{c,h}(\boldsymbol{u}, \nu, \nu) = \Gamma_{c,h}(\boldsymbol{u}, \mathcal{P}_c c, \nu) - \Gamma_{c,h}(\boldsymbol{u}, c_I, \nu) \\ &= [\Gamma_c(\boldsymbol{u}, c, \nu) - \Gamma_{c,h}(\boldsymbol{u}, c, \nu)] + \Gamma_{c,h}(\boldsymbol{u}, c - c_I, \nu) \\ &=: S_1 + S_2, \end{aligned} \tag{7.43}$$

for a constant $M > 0$. By employing the definitions of Γ_c and $\Gamma_{c,h}$ in (7.39), the term S_1 is split as follows:

$$\begin{aligned} S_1 &= [\mathcal{D}(\boldsymbol{u}; c, \nu) - \mathcal{D}_h(\boldsymbol{u}; c, \nu)] + [\Theta(\boldsymbol{u}; c, \nu) - \Theta_h(\boldsymbol{u}; c, \nu)] + [(c, \nu)_{0,\Omega} - (c, \nu)_h] \\ &=: S_1^A + S_1^B + S_1^C. \end{aligned}$$

For S_1^A, we have

$$\begin{aligned} S_1^A &= [(D(\boldsymbol{u})\nabla c, \nabla \nu)_{0,\Omega} - (D(\boldsymbol{\Pi}^0_{\boldsymbol{k}}\boldsymbol{u})\,\boldsymbol{\Pi}^0_{\boldsymbol{k}}(\nabla c), \boldsymbol{\Pi}^0_{\boldsymbol{k}}(\nabla \nu))_{0,\Omega}] \\ &\quad + \sum_{K \in \mathcal{T}_n} \nu_D^K(\boldsymbol{u}) S_D^K((I - \Pi^{\nabla,K}_{k+1})c, (I - \Pi^{\nabla,K}_{k+1})\nu) \\ &\leq \eta\, h^{k+1}\,[|D(\boldsymbol{u})\nabla c|_{k+1,\mathcal{T}_n} + |\nabla c|_{k+1,\mathcal{T}_n}(\|D(\boldsymbol{u})\|_{\infty,\Omega} + 1) + |\boldsymbol{u}|_{k+1,\mathcal{T}_n}\|\nabla c\|_{\infty,\Omega}]\,|\nu|_{1,\mathcal{T}_n}, \end{aligned}$$

where in the inequality we applied Lemma 7.2.1 to prove an upper bound for the first part on the right-hand side of S_1^A, and made use of the continuity properties (7.18) of $S_D^K(\cdot,\cdot)$, the trivial continuity property of $\Pi^{\nabla,K}_{k+1}$ in the H^1 seminorm and its approximation properties (stated in Lemma 7.1.1) to estimate the stabilization term.

Next, for S_1^B, we compute

$$\begin{aligned} S_1^B &= \frac{1}{2}\Big\{\, [(\boldsymbol{u} \cdot \nabla c, \nu)_{0,\Omega} - (\boldsymbol{\Pi}^0_{\boldsymbol{k}}\boldsymbol{u} \cdot \boldsymbol{\Pi}^0_{\boldsymbol{k}}(\nabla c), \Pi^0_{k+1}\nu)_{0,\Omega}] \\ &\qquad + \big[((q^+ + q^-)c, \nu)_{0,\Omega} - ((q^+ + q^-)\Pi^0_{k+1}c, \Pi^0_{k+1}\nu)_{0,\Omega}\big] \\ &\qquad - \big[(\boldsymbol{u}\,c, \nabla \nu)_{0,\Omega} - (\boldsymbol{\Pi}^0_{\boldsymbol{k}}\boldsymbol{u}\,\Pi^0_{k+1}c, \boldsymbol{\Pi}^0_{\boldsymbol{k}}(\nabla \nu))_{0,\Omega}\big] \Big\} \\ &\leq \eta\, h^{k+1}\big[|\boldsymbol{u} \cdot \nabla c|_{k+1,\mathcal{T}_n} + (|\nabla c|_{k+1,\mathcal{T}_n} + |c|_{k+1,\mathcal{T}_n})\|\boldsymbol{u}\|_{\infty,\Omega} + |c|_{k+1,\mathcal{T}_n}\|q^+ + q^-\|_{\infty,\Omega} \\ &\qquad + |(q^+ + q^-)c|_{k+1,\mathcal{T}_n} + |\boldsymbol{u}\,c|_{k+1,\mathcal{T}_n} + |\boldsymbol{u}|_{k+1,\mathcal{T}_n}(\|c\|_{\infty,\Omega} + \|\nabla c\|_{\infty,\Omega})\big]\,\|\nu\|_{1,\mathcal{T}_n}, \end{aligned}$$

where in the last inequality we used Lemma 7.2.1 with $\kappa = id$ and $\sigma = \boldsymbol{u}$ for the first and third term inside the curly bracket, and $\kappa = q^+ + q^-$ and $\sigma = 1$ for the second one.

Finally, for S_1^C, it holds with the definition of the L^2 projector and Lemma 7.1.1

$$S_1^C = ((I - \Pi^0_{k+1})c, \nu)_{0,\Omega} \leq \eta\, h^{k+1} |c|_{k+1,\mathcal{T}_n} \|\nu\|_{0,\Omega}.$$

On the other hand, for S_2, we use the continuity of $\Gamma_{c,h}$ in (7.40), together with the interpolation error estimate (7.42), to derive

$$\Gamma_{c,h}(\boldsymbol{u}; c - c_I, \nu) \leq \eta\, \zeta(\boldsymbol{u}, q^+, q^-) \|c - c_I\|_{1,\mathcal{T}_n} \|\nu\|_{1,\mathcal{T}_n} \leq \eta\, \zeta(\boldsymbol{u}, q^+, q^-) h^{k+1} \|c\|_{k+2,\mathcal{T}_n} \|\nu\|_{1,\mathcal{T}_n}.$$

The error bound in the broken H^1 norm follows by plugging firstly the estimates for S_1^A, S_1^B, and S_1^C into S_1, then those obtained for S_1 and S_2 into (7.43), using the definition of the H^1 norm, dividing by $\|\nu\|_{1,\mathcal{T}_n}$, and using the triangle inequality in the form

$$\|c - \mathcal{P}_c c\|_{1,\mathcal{T}_n} \leq \|c - c_I\|_{1,\mathcal{T}_n} + \|\nu\|_{1,\mathcal{T}_n},$$

together with the approximation properties (7.42) of the interpolant c_I.

The L^2 error bound can be derived by combining the above arguments with a standard duality argument as in [174], also recalling the convexity of Ω; it is omitted here. □

By differentiation of (7.38) in time and use of similar techniques as in the proof of Lemma 7.2.4, an analogous result can be obtained for $\frac{\partial}{\partial t}(c - \mathcal{P}_c c)$, summarized in the following corollary.

Corollary 7.2.5. *Provided that the continuous data and solution are sufficiently regular in space and time, it holds*

$$\left\| \frac{\partial}{\partial t}(c - \mathcal{P}_c c) \right\|_{1,\mathcal{T}_n} \leq h^{k+1}\, \xi_{1,t}, \qquad \left\| \frac{\partial}{\partial t}(c - \mathcal{P}_c c) \right\|_{0,\Omega} \leq h^{k+2}\, \xi_{0,t},$$

where the constants $\xi_{1,t}, \xi_{0,t} > 0$ are independent of h.

Moreover, we will later on need the two subsequent bounds.

Lemma 7.2.6. *Under sufficient smoothness of the continuous data and solution, it holds:*

$$\left\| \frac{\partial c^n}{\partial t} - \frac{\mathcal{P}_c c^n - \mathcal{P}_c c^{n-1}}{\tau} \right\|_{0,\Omega} \leq \tau^{\frac{1}{2}} \left\| \frac{\partial^2 c}{\partial s^2} \right\|_{L^2(t_{n-1},t_n;L^2(\Omega))} + \tau^{-\frac{1}{2}} h^{k+2} \left(\int_{t_{n-1}}^{t_n} \xi_{0,t}^2 \,\mathrm{d}s \right)^{\frac{1}{2}},$$

where $\xi_{0,t}$ can be found in Corollary 7.2.5.

Proof. We estimate

$$\begin{aligned} \left\| \frac{\partial c}{\partial t} - \frac{\mathcal{P}_c c^n - \mathcal{P}_c c^{n-1}}{\tau} \right\|_{0,\Omega} &\leq \left\| \frac{\partial c^n}{\partial t} - \frac{c^n - c^{n-1}}{\tau} \right\|_{0,\Omega} + \left\| \frac{\mathcal{P}_c c^n - \mathcal{P}_c c^{n-1}}{\tau} - \frac{c^n - c^{n-1}}{\tau} \right\|_{0,\Omega} \\ &=: (I) + (II). \end{aligned}$$

The term (I) can be estimated exactly as for standard finite elements, see for instance [171]:

$$(I) = \left\| \frac{\partial c^n}{\partial t} - \frac{c^n - c^{n-1}}{\tau} \right\|_{0,\Omega} \leq \int_{t_{n-1}}^{t_n} \left\| \frac{\partial^2 c}{\partial s^2}(s) \right\|_{0,\Omega} \mathrm{d}s \leq \tau^{\frac{1}{2}} \left(\int_{t_{n-1}}^{t_n} \left\| \frac{\partial^2 c}{\partial s^2}(s) \right\|_{0,\Omega}^2 \mathrm{d}s \right)^{\frac{1}{2}},$$

where we also applied the Hölder inequality in the last step. Concerning (II), this term can be estimated as follows, using Corollary 7.2.5:

$$\begin{aligned} (II) = \left\| \frac{\mathcal{P}_c c^n - \mathcal{P}_c c^{n-1}}{\tau} - \frac{c^n - c^{n-1}}{\tau} \right\|_{0,\Omega} &= \frac{1}{\tau} \left\| \int_{t_{n-1}}^{t_n} \frac{\partial}{\partial s}(\mathcal{P}_c c - c)(s) \,\mathrm{d}s \right\|_{0,\Omega} \\ &\leq \tau^{-\frac{1}{2}} \left(\int_{t_{n-1}}^{t_n} \left\| \frac{\partial}{\partial s}(\mathcal{P}_c c - c)(s) \right\|_{0,\Omega}^2 \mathrm{d}s \right)^{\frac{1}{2}} \leq \tau^{-\frac{1}{2}} h^{k+2} \left(\int_{t_{n-1}}^{t_n} \xi_{0,t}^2 \,\mathrm{d}s \right)^{\frac{1}{2}}. \end{aligned}$$

The statement of the lemma follows. □

Lemma 7.2.7. *Provided that the continuous data and solution are sufficiently regular in space and time, it holds:*

$$\|\boldsymbol{u}^n - \boldsymbol{U}^{n-1}\|_{0,\Omega} \leq \tau \left\| \frac{\partial \boldsymbol{u}}{\partial t} \right\|_{L^\infty(t_{n-1},t_n;L^2(\Omega))} + \|C^{n-1} - c^{n-1}\|_{0,\Omega}\, \zeta_1^{n-1} + h^{k+1}\, \zeta_2^{n-1},$$

where ζ_1^{n-1} and ζ_2^{n-1} are the constants from Theorem 7.2.2.

Proof. By using the triangle inequality, one obtains

$$\|\boldsymbol{u}^n - \boldsymbol{U}^{n-1}\|_{0,\Omega} \leq \left\|\boldsymbol{u}^n - \boldsymbol{u}^{n-1}\right\|_{0,\Omega} + \|\boldsymbol{u}^{n-1} - \boldsymbol{U}^{n-1}\|_{0,\Omega}.$$

The first term on the right-hand side is estimated by

$$\|\boldsymbol{u}^n - \boldsymbol{u}^{n-1}\|_{0,\Omega} = \left\| \int_{t_{n-1}}^{t_n} \frac{\partial \boldsymbol{u}(s)}{\partial s}\, \mathrm{d}s \right\|_{0,\Omega} \leq \tau \left\| \frac{\partial \boldsymbol{u}}{\partial t} \right\|_{L^\infty(t_{n-1},t_n;L^2(\Omega))},$$

and the second one term is bounded with Theorem 7.2.2. □

Now, we have all the ingredients to bound $\|c^n - C^n\|_{0,\Omega}$.

Theorem 7.2.8. *Let the mesh assumptions **(G1)**-**(G3)** be satisfied. Then, provided that the continuous data and solutions are sufficiently regular, it yields*

$$\|c^n - C^n\|_{0,\Omega} \leq \eta \left[\|c_{0,h} - c^0\|_{0,\Omega} + h^{k+1}\, \varphi_1 + \tau\, \varphi_2 \right],$$

where the regularity terms φ_1, φ_2 and the positive constant η now depend on $\boldsymbol{u}$, c, q^+, q^-, $\widehat{c}$, $\frac{\partial \boldsymbol{u}}{\partial t}$, $\frac{\partial^2 \boldsymbol{u}}{\partial t^2}$, $\frac{\partial c}{\partial t}$, and $\frac{\partial^2 c}{\partial t^2}$ (and products of these functions).

Proof. To start with, we write

$$C^n - c^n = (C^n - \mathcal{P}_c c^n) + (\mathcal{P}_c c^n - c^n) =: \vartheta^n + \rho^n.$$

Equation (7.41) gives a bound on ρ^n. In order to deal with ϑ^n, we use the continuous concentration equation (2.21) with $z = \vartheta^n$, the fully discretized version (7.30) with $z_h = \vartheta^n$, and the definition of the projector $\mathcal{P}_c c^n$ in (7.38) with $z_h = \vartheta^n$:

$$\begin{aligned}
&\mathcal{M}_h\left(\frac{\vartheta^n - \vartheta^{n-1}}{\tau}, \vartheta^n\right) + \mathcal{D}_h(\boldsymbol{U}^{n-1}; \vartheta^n, \vartheta^n) \\
&= \left[\mathcal{M}\left(\frac{\partial c^n}{\partial t}, \vartheta^n\right)_{0,\Omega} - \mathcal{M}_h\left(\frac{\mathcal{P}_c c^n - \mathcal{P}_c c^{n-1}}{\tau}, \vartheta^n\right) \right] \\
&\quad + \left[\Theta_h(\boldsymbol{u}^n; \mathcal{P}_c c^n, \vartheta^n) - \Theta_h(\boldsymbol{U}^{n-1}; C^n, \vartheta^n)\right] + \left[\mathcal{D}_h\left(\boldsymbol{u}^n; \mathcal{P}_c c^n, \vartheta^n\right) - \mathcal{D}_h\left(\boldsymbol{U}^{n-1}; \mathcal{P}_c c^n, \vartheta^n\right)\right] \\
&\quad + \left[(\mathcal{P}_c c^n, \vartheta^n)_h - (c^n, \vartheta^n)_{0,\Omega}\right] + \left[((q^+)^n \widehat{c}^n, \vartheta^n)_h - ((q^+)^n \widehat{c}^n, \vartheta^n)_{0,\Omega}\right] \\
&\quad =: R_1 + R_2 + R_3 + R_4 + R_5.
\end{aligned} \tag{7.44}$$

Owing to the coercivity properties in (7.22), the second term on the left-hand side of (7.44) can be estimated by

$$\mathcal{D}_h(\boldsymbol{U}^{n-1}; \vartheta^n, \vartheta^n) \geq D_*\, |\vartheta^n|^2_{1,\mathcal{T}_n} \tag{7.45}$$

with some constant $D_* > 0$ independent of h and $\boldsymbol{U}^{n-1}$.

The terms R_1-R_5 on the right-hand side of (7.44) are estimated as follows:

- term R_1: Using the definition of $\mathcal{M}_h(\cdot,\cdot)$ in (7.11), together with (7.18), yields

$$
\begin{aligned}
R_1 &= \mathcal{M}\left(\frac{\partial c^n}{\partial t}, \vartheta^n\right)_{0,\Omega} - \mathcal{M}_h\left(\frac{\mathcal{P}_c c^n - \mathcal{P}_c c^{n-1}}{\tau}, \vartheta^n\right)\\
&= \Bigg[\left(\phi\,\frac{\partial c^n}{\partial t}, \vartheta^n\right)_{0,\Omega} - \left(\Pi^0_{k+1}\left(\phi\,\Pi^0_{k+1}\left(\frac{\mathcal{P}_c c^n - \mathcal{P}_c c^{n-1}}{\tau}\right)\right), \vartheta^n\right)_{0,\Omega}\\
&\quad - \sum_{K\in\mathcal{T}_n} \nu^K_{\mathcal{M}}(\phi) S^K_{\mathcal{M}}\left((I-\Pi^{0,K}_{k+1})\left(\frac{\mathcal{P}_c c^n - \mathcal{P}_c c^{n-1}}{\tau}\right), (I-\Pi^{0,K}_{k+1})\vartheta^n\right)\Bigg]\\
&\le \eta\Bigg[\left\|\phi\,\frac{\partial c^n}{\partial t} - \Pi^0_{k+1}\left(\phi\,\Pi^0_{k+1}\left(\frac{\mathcal{P}_c c^n - \mathcal{P}_c c^{n-1}}{\tau}\right)\right)\right\|_{0,\Omega}\\
&\quad + \left\|(I-\Pi^0_{k+1})\left(\frac{\mathcal{P}_c c^n - \mathcal{P}_c c^{n-1}}{\tau}\right)\right\|_{0,\Omega}\Bigg]\|\vartheta^n\|_{0,\Omega}\\
&=: \eta[R_1^A + R_1^B]\|\vartheta^n\|_{0,\Omega}.
\end{aligned}
$$

The term R_1^A is estimated by using the continuity of the L^2 projector, the assumption (2.15) on ϕ, and the approximation properties in Lemma (7.1.1):

$$
\begin{aligned}
R_1^A &\le \left\|(I-\Pi^0_{k+1})\left(\phi\,\frac{\partial c^n}{\partial t}\right)\right\|_{0,\Omega} + \left\|\Pi^0_{k+1}\left(\phi\,\frac{\partial c^n}{\partial t} - \phi\,\Pi^0_{k+1}\left(\frac{\partial c^n}{\partial t}\right)\right)\right\|_{0,\Omega}\\
&\quad + \left\|\Pi^0_{k+1}\left(\phi\,\Pi^0_{k+1}\left(\frac{\partial c^n}{\partial t} - \frac{\mathcal{P}_c c^n - \mathcal{P}_c c^{n-1}}{\tau}\right)\right)\right\|_{0,\Omega}\\
&\le \eta\left[h^{k+2}\left(\left|\phi\frac{\partial c^n}{\partial t}\right|_{k+2,\mathcal{T}_n} + \left|\frac{\partial c^n}{\partial t}\right|_{k+2,\mathcal{T}_n}\right) + \left\|\frac{\partial c^n}{\partial t} - \frac{\mathcal{P}_c c^n - \mathcal{P}_c c^{n-1}}{\tau}\right\|_{0,\Omega}\right].
\end{aligned}
$$

Next, we prove an upper bound for R_1^B with similar tools as for R_1^A:

$$
\begin{aligned}
R_1^B &\le \left\|(I-\Pi^0_{k+1})\left(\frac{\mathcal{P}_c c^n - \mathcal{P}_c c^{n-1}}{\tau} - \frac{\partial c^n}{\partial t}\right)\right\|_{0,\Omega} + \left\|(I-\Pi^0_{k+1})\frac{\partial c^n}{\partial t}\right\|_{0,\Omega}\\
&\le \left\|\frac{\mathcal{P}_c c^n - \mathcal{P}_c c^{n-1}}{\tau} - \frac{\partial c^n}{\partial t}\right\|_{0,\Omega} + \eta\, h^{k+2}\left|\frac{\partial c^n}{\partial t}\right|_{k+2,\mathcal{T}_n}.
\end{aligned}
$$

Thus, we deduce with Lemma 7.2.6

$$
\begin{aligned}
R_1 &\le \eta\Bigg[h^{k+2}\left(\left|\phi\frac{\partial c^n}{\partial t}\right|_{k+2,\mathcal{T}_n} + \left|\frac{\partial c^n}{\partial t}\right|_{k+2,\mathcal{T}_n}\right) + \tau^{-\frac{1}{2}} h^{k+2}\left(\int_{t_{n-1}}^{t_n} \xi^2_{0,t}\,\mathrm{d}s\right)^{\frac{1}{2}}\\
&\quad + \tau^{\frac{1}{2}}\left\|\frac{\partial^2 c}{\partial s^2}\right\|_{L^2(t_{n-1},t_n;L^2(\Omega))}\Bigg]\|\vartheta^n\|_{0,\Omega}\\
&=: \left[h^{k+2} R_1^{n,1} + \tau^{-\frac{1}{2}} h^{k+2} R_1^{n,2} + \tau^{\frac{1}{2}} R_1^{n,3}\right]\|\vartheta^n\|_{0,\Omega},
\end{aligned}
\tag{7.46}
$$

with the obvious definitions for the regularity terms $R_1^{n,1}$, $R_1^{n,2}$, and $R_1^{n,3}$.

- term R_2: By the definition of $\Theta_h(\cdot;\cdot,\cdot)$ in (7.13), the identity $\vartheta^n = C^n - \mathcal{P}_c c^n$, and the fact that $(q^+)^n$ and $(q^-)^n$ are non-negative, it holds

$$
\begin{aligned}
&\Theta_h(\boldsymbol{u}^n; \mathcal{P}_c c^n, \vartheta^n) - \Theta_h(\boldsymbol{U}^{n-1}; C^n, \vartheta^n)\\
&= \frac{1}{2}\left[(\boldsymbol{u}^n\cdot\nabla\mathcal{P}_c c^n, \vartheta^n)_h - \left(\boldsymbol{U}^{n-1}\cdot\nabla C^n, \vartheta^n\right)_h\right] - \frac{1}{2}\left((q^+ + q^-)\vartheta^n, \vartheta^n\right)_{0,\Omega}\\
&- \frac{1}{2}\left[(\boldsymbol{u}^n\mathcal{P}_c c^n, \nabla\vartheta^n)_h - \left(\boldsymbol{U}^{n-1}C^n, \nabla\vartheta^n\right)_h\right]\\
&\le \frac{1}{2}\left[(\boldsymbol{u}^n\cdot\nabla\mathcal{P}_c c^n, \vartheta^n)_h - \left(\boldsymbol{U}^{n-1}\cdot\nabla C^n, \vartheta^n\right)_h - (\boldsymbol{u}^n\mathcal{P}_c c^n, \nabla\vartheta^n)_h + \left(\boldsymbol{U}^{n-1}C^n, \nabla\vartheta^n\right)_h\right].
\end{aligned}
$$

The above equation, after adding zero in the form

$$
\begin{aligned}
0 &= (\boldsymbol{U}^{n-1} \cdot \nabla \vartheta^n, \vartheta^n)_h - (\boldsymbol{U}^{n-1} \cdot \nabla \vartheta^n, \vartheta^n)_h \\
&= (\boldsymbol{U}^{n-1} \cdot \nabla C^n, \vartheta^n)_h - (\boldsymbol{U}^{n-1} \cdot \nabla \mathcal{P}_c c^n, \vartheta^n)_h - (\boldsymbol{U}^{n-1} \cdot \nabla \vartheta^n, C^n)_h + (\boldsymbol{U}^{n-1} \cdot \nabla \vartheta^n, \mathcal{P}_c c^n)_h
\end{aligned}
$$

to the right-hand side, can be equivalently expressed as

$$
\begin{aligned}
&\Theta_h(\boldsymbol{u}^n; \mathcal{P}_c c^n, \vartheta^n) - \Theta_h(\boldsymbol{U}^{n-1}; C^n, \vartheta^n) \\
&\leq \frac{1}{2} \left[\left((\boldsymbol{u}^n - \boldsymbol{U}^{n-1}) \cdot \nabla \mathcal{P}_c c^n, \vartheta^n \right)_h - \left((\boldsymbol{u}^n - \boldsymbol{U}^{n-1}) \mathcal{P}_c c^n, \nabla \vartheta^n \right)_h \right] =: R_2^A + R_2^B.
\end{aligned}
$$

For R_2^A, we estimate

$$
R_2^A = \frac{1}{2} \left((\boldsymbol{u}^n - \boldsymbol{U}^{n-1}) \nabla \mathcal{P}_c c^n, \vartheta^n \right)_h \leq \frac{1}{2} \|\boldsymbol{u}^n - \boldsymbol{U}^{n-1}\|_{0,\Omega} \|\boldsymbol{\Pi}_{\boldsymbol{k}}^{\mathbf{0}} \nabla \mathcal{P}_c c^n\|_{\infty,\Omega} \|\vartheta^n\|_{0,\Omega}.
$$

We now use an inverse estimate [59, Lemma 4.5.3], the continuity of $\boldsymbol{\Pi}_{\boldsymbol{k}}^{\mathbf{0},\boldsymbol{K}}$, a triangle inequality, the assumption that $\mathcal{T}_n$ is quasi-regular, and Lemma 7.2.4, to deduce, for every $K \in \mathcal{T}_n$,

$$
\begin{aligned}
\|\boldsymbol{\Pi}_{\boldsymbol{k}}^{\mathbf{0},\boldsymbol{K}} \nabla \mathcal{P}_c c^n\|_{\infty,K} &\leq \eta \, h_K^{-1} \|\boldsymbol{\Pi}_{\boldsymbol{k}}^{\mathbf{0},\boldsymbol{K}} \nabla \mathcal{P}_c c^n\|_{0,K} \leq \eta \, h_K^{-1} \|\nabla \mathcal{P}_c c^n\|_{0,K} \\
&\leq \eta \, h_K^{-1} \left(\|\nabla \mathcal{P}_c c^n - \nabla c^n\|_{0,K} + \|\nabla c^n\|_{0,K} \right) \\
&\leq \eta \left(h^{-1} \|\nabla \mathcal{P}_c c^n - \nabla c^n\|_{0,\mathcal{T}_n} + \|\nabla c^n\|_{\infty,K} \right) \leq \eta.
\end{aligned} \tag{7.47}
$$

Recalling Lemma 7.2.7, the definitions of ϑ^{n-1} and ρ^{n-1}, and Lemma 7.2.4, we get

$$
\begin{aligned}
&\|\boldsymbol{u}^n - \boldsymbol{U}^{n-1}\|_{0,\Omega} \\
&\quad \leq \tau \left\| \frac{\partial \boldsymbol{u}}{\partial t} \right\|_{L^\infty(t_{n-1},t_n; L^2(\Omega))} + (\|\vartheta^{n-1}\|_{0,\Omega} + \|\rho^{n-1}\|_{0,\Omega}) \zeta_1^{n-1} + h^{k+1} \, \zeta_2^{n-1} \\
&\quad \leq \tau \left\| \frac{\partial \boldsymbol{u}}{\partial t} \right\|_{L^\infty(t_{n-1},t_n; L^2(\Omega))} + (\|\vartheta^{n-1}\|_{0,\Omega} + h^{k+2} \xi_0^{n-1}) \zeta_1^{n-1} + h^{k+1} \, \zeta_2^{n-1},
\end{aligned} \tag{7.48}
$$

thus implying

$$
R_2^A \leq \eta \left[h^{k+1} R_2^{n,1} + \tau R_2^{n,2} + \|\vartheta^{n-1}\|_{0,\Omega} R_2^{n,3} \right] \|\vartheta^n\|_{0,\Omega},
$$

with the obvious definitions for the regularity terms $R_2^{n,1}$, $R_2^{n,2}$, and $R_2^{n,3}$. The term R_2^B can be estimated analogously to R_2^A, giving

$$
R_2^B = \frac{1}{2} \left((\boldsymbol{U}^{n-1} - \boldsymbol{u}^n) \mathcal{P}_c c^n, \nabla \vartheta^n \right)_h \leq \eta \, \|\boldsymbol{u}^n - \boldsymbol{U}^{n-1}\|_{0,\Omega} |\vartheta^n|_{1,\mathcal{T}_n}.
$$

Using again the bound (7.48), one obtains

$$
R_2^B \leq \eta \left[h^{k+1} R_2^{n,1} + \tau R_2^{n,2} + \|\vartheta^{n-1}\|_{0,\Omega} R_2^{n,3} \right] |\vartheta^n|_{1,\mathcal{T}_n}.
$$

Thus,

$$
R_2 \leq \eta \left[h^{k+1} R_2^{n,1} + \tau R_2^{n,2} + \|\vartheta^{n-1}\|_{0,\Omega} R_2^{n,3} \right] \left(\|\vartheta^n\|_{0,\Omega} + |\vartheta^n|_{1,\mathcal{T}_n} \right).
$$

- term R_3: We use the definition of $\mathcal{D}_h(\cdot;\cdot,\cdot)$ in (7.14), a standard Hölder inequality in the spirit of (2.24), the estimate (7.47), the scaling properties of the stabilization in (7.18), the Lipschitz continuity of $D(\cdot;\cdot,\cdot)$, and ν_D^K in (7.20), to deduce

$$
\begin{aligned}
R_3 &= \mathcal{D}_h \left(\boldsymbol{u}^n; \mathcal{P}_c c^n, \vartheta^n \right) - \mathcal{D}_h \left(\boldsymbol{U}^{n-1}; \mathcal{P}_c c^n, \vartheta^n \right) \\
&= \left((D(\boldsymbol{\Pi}_{\boldsymbol{k}}^{\mathbf{0}} \boldsymbol{u}^n) - D(\boldsymbol{\Pi}_{\boldsymbol{k}}^{\mathbf{0}} \boldsymbol{U}^{n-1})) \, \boldsymbol{\Pi}_{\boldsymbol{k}}^{\mathbf{0}} (\nabla \mathcal{P}_c c^n) \cdot \boldsymbol{\Pi}_{\boldsymbol{k}}^{\mathbf{0}} (\nabla \vartheta^n) \right)_{0,\Omega} \\
&\quad + \sum_{K \in \mathcal{T}_n} (\nu_D^K(\boldsymbol{u}^n) - \nu_D^K(\boldsymbol{U}^{n-1})) S_D^K \left((I - \Pi_{k+1}^{\nabla,K}) \mathcal{P}_c c^n, (I - \Pi_{k+1}^{\nabla,K}) \vartheta^n \right) \\
&\leq \eta \|\boldsymbol{u}^n - \boldsymbol{U}^{n-1}\|_{0,\Omega} |\vartheta^n|_{1,\mathcal{T}_n}.
\end{aligned}
$$

Hence, with (7.48) we have

$$R_3 \leq \eta \Bigg[h^{k+1} R_2^{n,1} + \tau R_2^{n,2} + \|\vartheta^{n-1}\|_{0,\Omega} R_2^{n,3} \Bigg] |\vartheta^n|_{1,\mathcal{T}_n}.$$

- term R_4: The use of Lemma 7.2.4 yields

$$\begin{aligned} R_4 &= -[(c^n, \vartheta^n)_{0,\Omega} - (\mathcal{P}_c c^n, \vartheta^n)_h] = -[((I - \Pi^0_{k+1}) c^n, \vartheta^n)_{0,\Omega} + (\Pi^0_{k+1}(c^n - \mathcal{P}_c c^n), \vartheta^n)_{0,\Omega}] \\ &\leq \eta\, h^{k+2} \left[|c^n|_{k+2,\mathcal{T}_n} + \xi_0^n \right] \|\vartheta^n\|_{0,\Omega} =: \eta\, h^{k+2} R_4^{n,1} \|\vartheta^n\|_{0,\Omega} \end{aligned}$$

  with the obvious definition of $R_4^{n,1}$.

- term R_5: The approximation properties in Lemma 7.1.1 yield

$$\begin{aligned} R_5 &= -\left((I - \Pi^0_{k+1})((q^+)^n \widehat{c}^n), \vartheta^n \right)_{0,\Omega} \leq \eta\, h^{k+2} \, |(q^+)^n \widehat{c}^n|_{k+2,\mathcal{T}_n} \|\vartheta^n\|_{0,\Omega} \\ &=: \eta\, h^{k+2} R_5^{n,1} \|\vartheta^n\|_{0,\Omega} \end{aligned}$$

  with the obvious definition of $R_5^{n,1}$.

We now insert (7.45) and the bounds on R_1-R_5 into (7.44). Afterwards, we observe that all regularity terms $\{R_J^{n,i}\}$ above only depend on the continuous solution and can be assumed to be bounded uniformly in h. We only keep track of the terms $R_1^{n,2}$ and $R_1^{n,3}$. This yields

$$\begin{aligned} &\frac{1}{\tau} \mathcal{M}_h \left(\vartheta^n - \vartheta^{n-1}, \vartheta^n \right) + D_* \, |\vartheta^n|^2_{1,\mathcal{T}_n} \\ &\leq \|\vartheta^{n-1}\|_{0,\Omega} \|\vartheta^n\|_{0,\Omega}\, \omega_1^n + \|\vartheta^{n-1}\|_{0,\Omega} |\vartheta^n|_{1,\mathcal{T}_n} \omega_2^n + \|\vartheta^n\|_{0,\Omega} \omega_3^n + |\vartheta^n|_{1,\mathcal{T}_n} \omega_4^n \\ &= \|\vartheta^n\|_{0,\Omega} \left[\omega_3^n + \|\vartheta^{n-1}\|_{0,\Omega}\, \omega_1^n \right] + |\vartheta^n|_{1,\mathcal{T}_n} \left[\omega_4^n + \|\vartheta^{n-1}\|_{0,\Omega}\, \omega_2^n \right], \end{aligned} \tag{7.49}$$

with the positive scalars

$$\omega_i^n \leq \eta, \quad i = 1, 2, \quad \omega_3^n \leq \eta \left(\tau + h^{k+1} + \tau^{-\frac{1}{2}} h^{k+2} R_1^{n,2} + \tau^{\frac{1}{2}} R_1^{n,3} \right), \quad \omega_4^n \leq \eta \left(\tau + h^{k+1} \right). \tag{7.50}$$

Next, we introduce, for all $w_h \in Z_h$, the discrete norm

$$\|w_h\|^2_{0,h} := \mathcal{M}_h(w_h, w_h). \tag{7.51}$$

Owing to Lemma 7.1.2, there exist positive constants c_* and c^*, such that, for all $w_h \in Z_h$, it holds

$$c_* \|w_h\|_{0,h} \leq \|w_h\|_{0,\Omega} \leq c^* \|w_h\|_{0,h}. \tag{7.52}$$

Reshaping (7.49), and employing (7.51) and (7.52), then gives

$$\begin{aligned} \|\vartheta^n\|^2_{0,h} + \tau D_* \, |\vartheta^n|^2_{1,\mathcal{T}_n} &\leq \mathcal{M}_h(\vartheta^{n-1}, \vartheta^n) + \tau \|\vartheta^n\|_{0,h} \left[c^* \omega_3^n + \|\vartheta^{n-1}\|_{0,h} (c^*)^2 \omega_1^n \right] \\ &\quad + \tau |\vartheta^n|_{1,\mathcal{T}_n} \left[\omega_4^n + \|\vartheta^{n-1}\|_{0,h}\, c^* \omega_2^n \right] \\ &=: T_1 + T_2 + T_3. \end{aligned} \tag{7.53}$$

The terms T_1 and T_2 are estimated as follows:

$$\begin{aligned} T_1 + T_2 &\leq \|\vartheta^n\|_{0,h} \left[(1 + \tau\eta) \|\vartheta^{n-1}\|_{0,h} + \tau c^* \omega_3^n \right] \\ &\leq \frac{1}{2} \left(\|\vartheta^n\|^2_{0,h} + \left[(1 + \tau\eta) \|\vartheta^{n-1}\|_{0,h} + \tau c^* \omega_3^n \right]^2 \right), \end{aligned} \tag{7.54}$$

where we used (7.25) and (7.51) in the first step. The term T_3 is estimated as follows:

$$\begin{aligned} T_3 &\leq \tau D_* |\vartheta^n|^2_{1,\mathcal{T}_n} + \frac{\tau}{4 D_*} \left[\omega_4^n + \|\vartheta^{n-1}\|_{0,h}\, c^* \omega_2^n \right]^2 \\ &\leq \tau D_* |\vartheta^n|^2_{1,\mathcal{T}_n} + \frac{\tau}{2} \eta \left[(\omega_4^n)^2 + \|\vartheta^{n-1}\|^2_{0,h} \right]^2. \end{aligned} \tag{7.55}$$

Next, we plug (7.54) and (7.55) into (7.53), cancel the terms $\tau D_* |\vartheta^n|^2_{1,\mathcal{T}_n}$ and manipulate the resulting inequality, to obtain

$$\|\vartheta^n\|^2_{0,h} \leq \left[(1+\tau\eta)\|\vartheta^{n-1}\|_{0,h} + \tau c^* \omega_3^n\right]^2 + \tau\eta\left[(\omega_4^n)^2 + \|\vartheta^{n-1}\|^2_{0,h}\right]^2.$$

Moreover, we estimate

$$\begin{aligned}
&\left[(1+\tau\eta)\|\vartheta^{n-1}\|_{0,h} + \tau c^* \omega_3^n\right]^2 \\
&\quad = (1+\tau\eta)^2\|\vartheta^{n-1}\|^2_{0,h} + 2\tau^{\frac{1}{2}}\|\vartheta^{n-1}\|_{0,h}\tau^{\frac{1}{2}}(1+\tau\eta)c^*\omega_3^n + \tau^2(c^*)^2(\omega_3^n)^2 \\
&\quad \leq \left[(1+\tau\eta)^2 + \tau\right]\|\vartheta^{n-1}\|^2_{0,h} + \left[\tau(1+\tau\eta)^2 + \tau^2\right](c^*)^2(\omega_3^n)^2 \\
&\quad \leq (1+\tau\eta)\|\vartheta^{n-1}\|^2_{0,h} + \tau\eta(\omega_3^n)^2.
\end{aligned}$$

Hence,

$$\|\vartheta^n\|^2_{0,h} \leq (1+\tau\eta)\|\vartheta^{n-1}\|^2_{0,h} + \tau\eta\left[(\omega_3^n)^2 + (\omega_4^n)^2\right].$$

Defining

$$\gamma^n := (\omega_3^n)^2 + (\omega_4^n)^2$$

and solving the recursion then leads to

$$\|\vartheta^n\|^2_{0,h} \leq (1+\tau\eta)^n\|\vartheta^0\|^2_{0,h} + \tau\eta\sum_{j=1}^n \gamma^j \leq \eta\|\vartheta^0\|^2_{0,h} + \tau\eta\sum_{j=1}^n \gamma^j,$$

where we recall that $n \leq T/\tau$ with T the final time instant. With (7.52), the estimate in the L^2 norm is a direct consequence:

$$\|\vartheta^n\|^2_{0,\Omega} \leq \eta\|\vartheta^0\|^2_{0,\Omega} + \tau\eta\sum_{j=1}^n \gamma^j. \tag{7.56}$$

The initial term $\|\vartheta^0\|^2_{0,\Omega}$ is estimated by

$$\|\vartheta^0\|_{0,\Omega} = \|c_{0,h} - \mathcal{P}_c c^0\|_{0,\Omega} \leq \|c_{0,h} - c^0\|_{0,\Omega} + \|c^0 - \mathcal{P}_c c^0\|_{0,\Omega} \leq \|c_{0,h} - c^0\|_{0,\Omega} + h^{k+2}\,\xi_0^0, \tag{7.57}$$

where we applied Lemma 7.2.4. Moreover, using (7.50), the fact that $\sum_{j=1}^n \tau \leq T$, and the definitions of $R_1^{j,2}$ and $R_1^{j,3}$ in (7.46), after some simple manipulations, we obtain

$$\begin{aligned}
\tau\eta\sum_{j=1}^n \gamma_j &\leq \eta\left(\sum_{j=1}^n \tau(\omega_3^j)^2 + \sum_{j=1}^n \tau(\omega_4^j)^2\right) \\
&\leq \eta\left[\sum_{j=1}^n \tau(\tau + h^{k+1})^2 + (h^{k+2})^2\sum_{j=1}^n (R_1^{j,2})^2 + \tau^2\sum_{j=1}^n (R_1^{j,3})^2\right] \\
&\leq \eta\left[(\tau + h^{k+1})^2 + (h^{k+2})^2\sum_{j=1}^n (R_1^{j,2})^2 + \tau^2\sum_{j=1}^n (R_1^{j,3})^2\right] \\
&\leq \eta\left[(\tau + h^{k+1})^2 + (h^{k+2})^2\int_0^{t_n} \xi_{0,t}^2\,\mathrm{d}s + \tau^2\int_0^{t_n}\left\|\frac{\partial^2 c}{\partial s^2}(s)\right\|^2_{0,\Omega}\mathrm{d}s\right].
\end{aligned} \tag{7.58}$$

The assertion of the theorem follows by combining (7.56) with (7.58) and (7.57). □

7.3 Numerical experiments

In this section, we demonstrate the performance of the method on the basis of numerical experiments, focusing on the lowest-order case $k = 0$. To this purpose, we firstly consider an ideal test case (*Example 1*), and then a more realistic one (*Example 2*). The aim of the first test is to validate (also numerically) the convergence of the method on a problem with regular known solution, whereas those of the second test is to check the method's performance on a well-known benchmark that mimics a more realistic situation.

Example 1: Here, we study a generalized version of (2.10), given by

$$\begin{cases} \phi\, \dfrac{\partial c}{\partial t} + \boldsymbol{u}\cdot\nabla c - \operatorname{div}(D(\boldsymbol{u})\nabla c) = f \\ \qquad\qquad\qquad\qquad\qquad \operatorname{div}\, \boldsymbol{u} = g \\ \qquad\qquad\qquad\qquad\qquad\quad \boldsymbol{u} = -a(c)(\nabla p - \boldsymbol{\gamma}(c)), \end{cases}$$

endowed with the boundary and initial conditions in (2.13) and (2.14), respectively. We fix $\Omega = (0,1)^2$ and pick the same choice of parameters as in [123], namely $T = 0.01$, $\phi = 1$, $D(\boldsymbol{u}) = |\boldsymbol{u}| + 0.02$, $d_m = 0.02$, $d_\ell = d_t = 1$, $c_0 = 0$, $\boldsymbol{\gamma}(c) = 0$, and $a(c) = (c+2)^{-1}$, where f and g are taken in accordance with the analytical solutions

$$\begin{aligned} c(x,y,t) &= t^2\left[x^2(x-1)^2 + y^2(y-1)^2\right] \\ \boldsymbol{u}(x,y,t) &= 2t^2\begin{pmatrix} x(x-1)(2x-1) \\ y(y-1)(2y-1)\end{pmatrix} \\ p(x,y,t) &= -\frac{1}{2}c^2 - 2c + \frac{17}{6300}t^4 + \frac{2}{15}t^2. \end{aligned} \tag{7.59}$$

Plots of the exact solution at the final time T are shown in Figures 7.1 and 7.2.

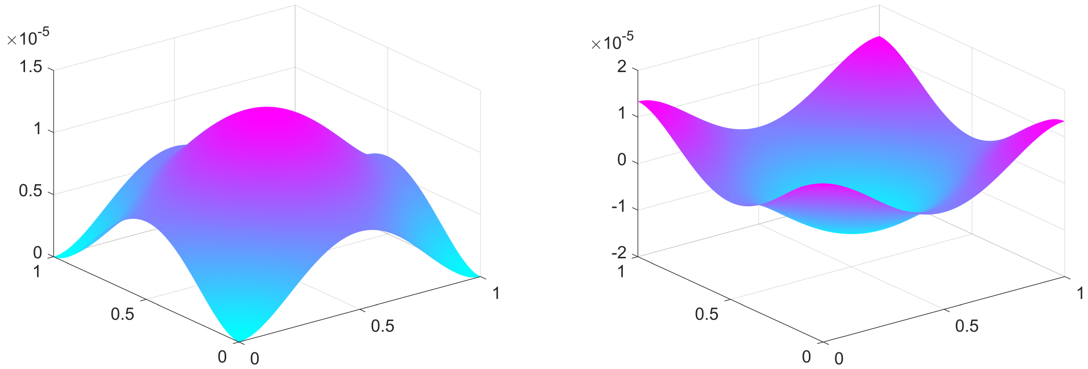


Figure 7.1: Exact concentration c (*left*) and pressure p (*right*) of example 1, given by (7.59), at the final time $T = 0.01$.

We employ a sequence of regular Cartesian meshes and Voronoi meshes, as portrayed in Figure 7.3. In addition to the current version, we also test the method when replacing the stabilization terms in (7.12), (7.15), and (7.17) by alternative ones:

$$\begin{aligned} \nu_{\mathcal{M}}^K(\phi) S_{\mathcal{M}}^K\left((I-\Pi_{k+1}^{0,K})c_h, (I-\Pi_{k+1}^{0,K})z_h\right) &\quad\rightsquigarrow\quad \widetilde{S_{\mathcal{M}}^K}\left((I-\Pi_{k+1}^{0,K})c_h, (I-\Pi_{k+1}^{0,K})z_h\right) \\ \nu_{D}^K(\boldsymbol{u}_h)\, S_{D}^K\left((I-\Pi_{k+1}^{\nabla,K})c_h, (I-\Pi_{k+1}^{\nabla,K})z_h)\right) &\quad\rightsquigarrow\quad \widetilde{S_{D}^K}\left(\boldsymbol{u}_h; (I-\Pi_{k+1}^{\nabla,K})c_h, (I-\Pi_{k+1}^{\nabla,K})z_h)\right) \\ \nu_{\mathcal{A}}^K(c_h)\, S_{\mathcal{A}}^K((I-\boldsymbol{\Pi}_{\boldsymbol{k}}^{\boldsymbol{0,K}})\boldsymbol{u}_h, (I-\boldsymbol{\Pi}_{\boldsymbol{k}}^{\boldsymbol{0,K}})\boldsymbol{v}_h) &\quad\rightsquigarrow\quad \widetilde{S_{\mathcal{A}}^K}(c_h; (I-\boldsymbol{\Pi}_{\boldsymbol{k}}^{\boldsymbol{0,K}})\boldsymbol{u}_h, (I-\boldsymbol{\Pi}_{\boldsymbol{k}}^{\boldsymbol{0,K}})\boldsymbol{v}_h). \end{aligned}$$

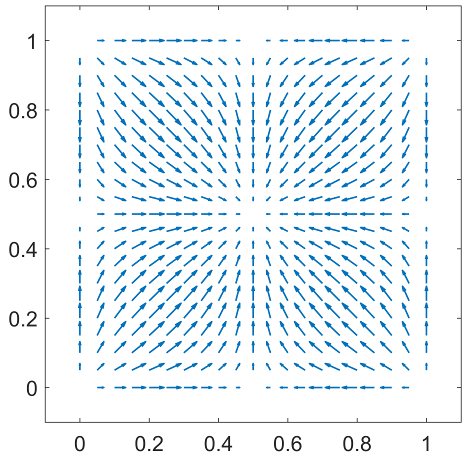


Figure 7.2: Exact vector field $\boldsymbol{u}$ of example 1, given by (7.59), at the final time $T = 0.01$.

The alternative (diagonal) stabilizations are given by

$$
\begin{aligned}
\widetilde{S^K_{\mathcal{M}}}(c_h, z_h) &= |K| \sum_{j=1}^{\dim Z_h(K)} d_j^{\mathcal{M}} \operatorname{dof}_j^{Z_h(K)}(c_h) \operatorname{dof}_j^{Z_h(K)}(z_h) \\
\widetilde{S^K_{D}}(c_h, z_h) &= \sum_{j=1}^{\dim Z_h(K)} d_j^{D} \operatorname{dof}_j^{Z_h(K)}(c_h) \operatorname{dof}_j^{Z_h(K)}(z_h) \\
\widetilde{S^K_{\mathcal{A}}}(\boldsymbol{u}_h, \boldsymbol{v}_h) &= |K| \sum_{j=1}^{\dim V_h(K)} d_j^{\mathcal{A}} \operatorname{dof}_j^{\boldsymbol{V}_h(K)}(\boldsymbol{u}_h) \operatorname{dof}_j^{\boldsymbol{V}_h(K)}(\boldsymbol{v}_h)
\end{aligned}
\tag{7.60}
$$

with

$$
\begin{aligned}
d_j^{\mathcal{M}} &:= \max\left\{ \frac{1}{|K|} \int_K \phi \, (\Pi^{0,K}_{k+1} \varphi_j^K)^2 \, \mathrm{d}x, \, \sigma \nu^K_{\mathcal{M}}(\phi) \right\} \\
d_j^{D} &:= \max\left\{ \int_K D(\boldsymbol{\Pi}^{\boldsymbol{0},\boldsymbol{K}}_{\boldsymbol{k}} \boldsymbol{u}_h) \, |\boldsymbol{\Pi}^{\boldsymbol{0},\boldsymbol{K}}_{\boldsymbol{k}} (\nabla \varphi_j^K)|^2 \, \mathrm{d}x, \, \sigma \nu^K_{D}(\boldsymbol{u}_h) \right\} \\
d_j^{\mathcal{A}} &:= \max\left\{ \frac{1}{|K|} \int_K A(\Pi^{0,K}_{k+1} c_h) \, |\boldsymbol{\Pi}^{\boldsymbol{0},\boldsymbol{K}}_{\boldsymbol{k}} \boldsymbol{\psi}_j^K|^2 \, \mathrm{d}x, \, \sigma \nu^K_{\mathcal{A}}(c_h) \right\},
\end{aligned}
\tag{7.61}
$$

where $\{\varphi_j^K\}_{\ell=1}^{\dim Z_h(K)}$ and $\{\boldsymbol{\psi}_j^K\}_{\ell=1}^{\dim V_h(K)}$ denote the local canonical basis functions for $Z_h(K)$ and $\boldsymbol{V}_h(K)$, respectively, and $\sigma > 0$ is a safety parameter. In the forthcoming experiments, we set $\sigma = 1e - 3$. We highlight that these stabilizations are in fact modifications of the D-recipe, which was introduced in [41] and we have already used in a slightly different version for the Helmholtz problem, see Chapters 5 and 6. The first entry inside the max is simply the "diagonal part" of the consistency term of the local approximate forms in (7.12), (7.15), and (7.17), respectively, whereas the second terms correspond to the original stabilizations associated to the degrees of freedom in (7.19) multiplied by σ, which acts as a positivity safeguard. Importantly, it is easy to check that the error analysis can be easily extended to the new choice of stabilizations.

Due to the virtuality of the basis functions, we measure the following relative L^2 errors:

$$
\frac{\|c - \Pi^0_1 C^n\|_{0,\Omega}}{\|c\|_{0,\Omega}}, \quad \frac{\|\boldsymbol{u} - \boldsymbol{\Pi}^0_0 \boldsymbol{U}^n\|_{0,\Omega}}{\|\boldsymbol{U}^n\|_{0,\Omega}}, \quad \frac{\|p - \Pi^0_0 P^n\|_{0,\Omega}}{\|p\|_{0,\Omega}},
$$

where C^n, $\boldsymbol{U}^n$, and P^n are the numerical solutions at the final time T.

The relative L^2 discretization errors for the concentration are plotted in Figure 7.4 in terms of the mesh size h, for both families of meshes and both variants of stabilizations. In order to better

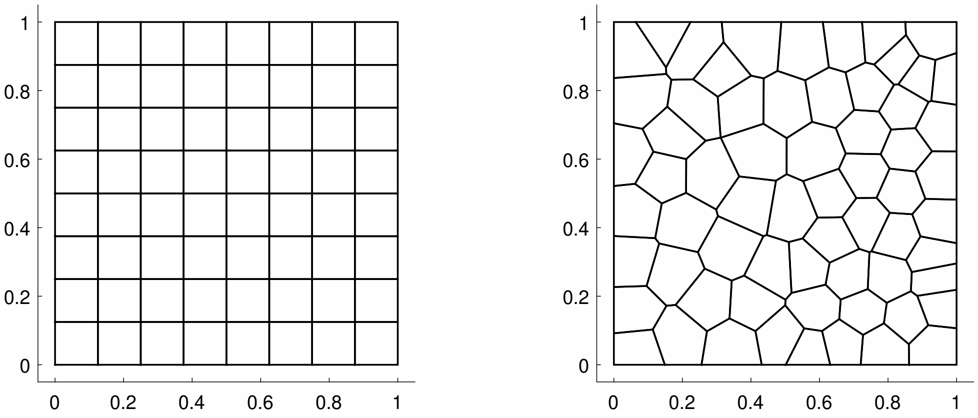


Figure 7.3: Meshes: regular 8x8 Cartesian mesh (*left*); Voronoi mesh with 64 elements (*right*).

underline the expected linear convergence of the method both in h and τ (see Theorem 7.2.8, recalling that $k = 0$), the time step τ is chosen proportional to h. In other words, starting with the coarsest mesh and $\tau = T/5$, each subsequent case is obtained by dividing both h (adopting a finer mesh) and τ by a factor of 2. Analogous plots are shown for the velocity and pressure variable errors in Figures 7.5 and 7.6, respectively. In all the cases, the linear convergence rates are in accordance with Theorem 7.2.2 and Theorem 7.2.8. For the pressure discretization error, since the initial meshes are very coarse, we observe some pre-asymptotic regime when employing the original stabilizations in (7.19). This effect, however, is not present for the alternative stabilizations in (7.60). Both variants lead to similar results for the concentration and velocity errors.

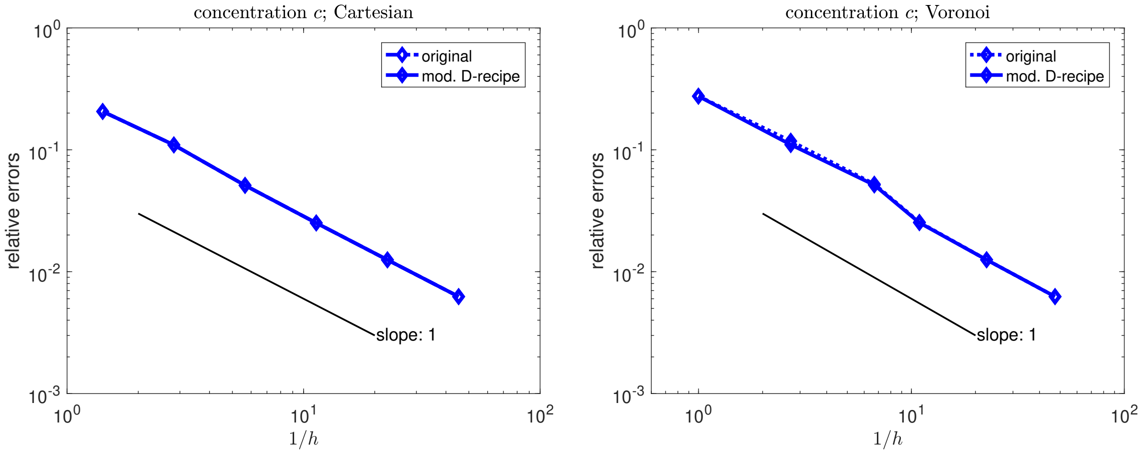


Figure 7.4: Relative L^2 errors for the concentration in example 1 at the final time T on regular Cartesian meshes (*left*) and Voronoi meshes (*right*). The original stabilization (7.19) and the D-recipe stabilization (7.60) are employed.

Since the concentration often evolves more rapidly than the velocity and pressure, it could be worth to consider a cheaper variant of the discrete scheme (7.29)-(7.31), where the discrete velocity-pressure pair is updated only every R time steps (with $R \in \mathbb{N}$). This leads to a smaller number of linear system resolutions (possibly with a small reduction in accuracy) since only the system (7.30) is solved at every time step, while (7.31) is solved only every R steps. In order to test this, we tried to run the same test above and compare the original version with the cheaper version with $R = 5$. The difference in error was only at the fourth meaningful digit; we do not plot the graphs because these would completely overlap the ones of the original method.

Example 2: Next, we investigate the behavior of the method for Test 1 and Test 2 in [73, 176].

The problem is given in the form (2.10) with boundary conditions (2.13) and initial condition (2.14) over the spatial domain $\Omega = (0, 1000)^2$ ft^2. Moreover, $T = 3600$ days and $\tau = 36$ days. At the upper right corner, i.e. at $[1000, 1000]$, fluid with concentration $\widehat{c} = 1.0$ is injected with

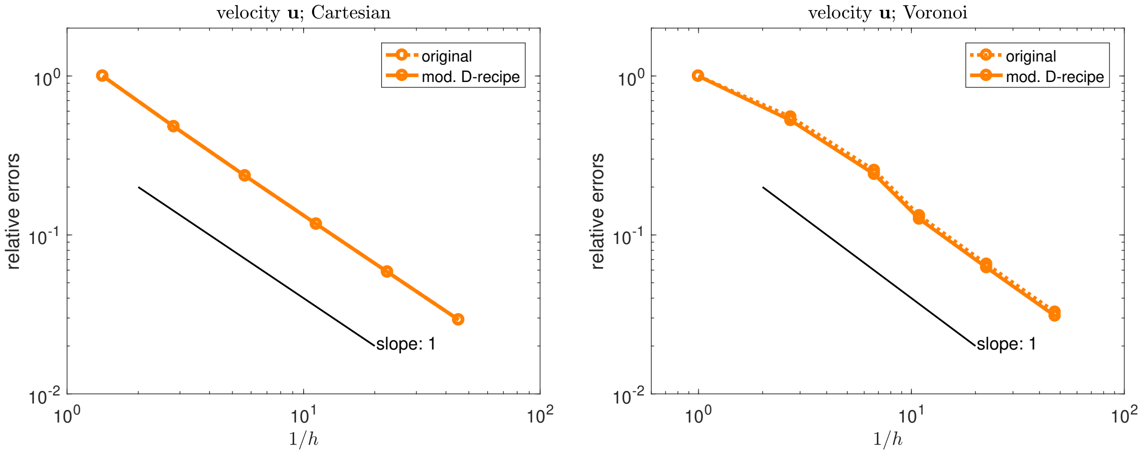


Figure 7.5: Relative L^2 errors for the velocity field in example 1 at the final time T on regular Cartesian meshes (*left*) and Voronoi meshes (*right*). The original stabilization (7.19) and the D-recipe stabilization (7.60) are employed.

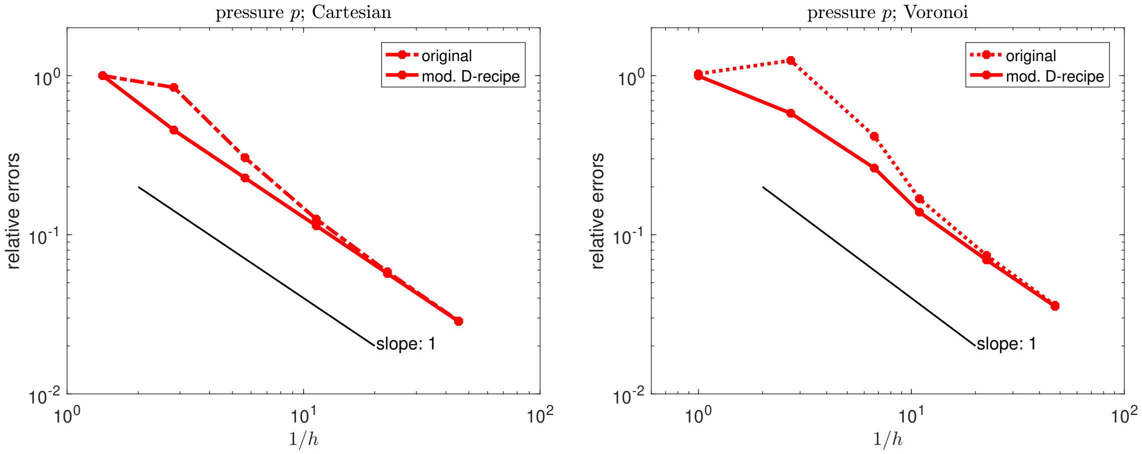


Figure 7.6: Relative L^2 errors for the pressure in example 1 at the final time T on regular Cartesian meshes (*left*) and Voronoi meshes (*right*). The original stabilization (7.19) and the D-recipe stabilization (7.60) are employed.

rate $q^+ = 30$ ft^2/day, whereas at the lower left corner, i.e. at $[0, 0]$, material is absorbed with rate $q^- = 30$ ft^2/day. Both wells are henceforth treated as Dirac masses, which is admissible at the discrete level since the discrete functions are piecewise regular (which can be interpreted as an approximation of the Dirac delta by a localized function with support within the corner element and unitary integral). Furthermore, the following choices for the parameters are picked: $\phi = 0.1$, $d_\ell = 50$, $d_t = 5$, $c_0 = 0$, $\gamma(c) = 0$, and $a(c) = 80(1 + (M^{\frac{1}{4}} - 1)c)^4$, where

$$\text{Test A} : d_m = 10, M = 1; \qquad\qquad \text{Test B} : d_m = 0, M = 41.$$

Whereas $a(c)$ is constant for Test A, it changes rapidly across the fluid interface for Test B (which is in fact not covered by the theoretical analysis since $d_m = 0$, but is interesting to study numerically) resulting in a much faster propagation of the fluid concentration front along the diagonal direction ($d_\ell \gg d_t$). This effect is known as *macroscopic fingering phenomenon* [98].

For this example, we used a regular 25x25 Cartesian mesh and we employed the more sophisticated stabilization in (7.60). Since Test B is highly convection-dominated, pure application of our method leads to local disturbances in the form of *overshoots* and *undershoots* of the numerical solution for the concentration, typical in the context of convection-dominated problem. To this purpose, for this test case, we employ the flux-corrected transport (FCT) algorithm with linearization [135, 136]. The FCT scheme with linearization for convection-dominated flow problems operates in two steps: (1) advance the solution in time by a low-order overly diffusive scheme to suppress spurious oscillations, (2) correct the solution using (linear) antidiffusive fluxes. In that way, the computed solution does not show spurious oscillations and layers are not smeared.

Due to the fact that no analytical solutions are available for Test A and Test B, we plot the numerical solutions (and the corresponding contour plots) for the concentration after 3 and 10 years. These times correspond to $n = 30$ and $n = 100$, respectively. For visualization of the results, since the numerical solution is virtual, but the nodal values are known, we simply add, inside each square, the barycenter with associated mean value of the nodal values, then create a triangulation based upon these points, and finally interpolate the function values linearly inside each triangle. In Figures 7.7 and 7.8, the results for Test A are portrayed, and in Figures 7.9 and 7.10 those for Test B. The results are similar to those obtained in [73, 176].

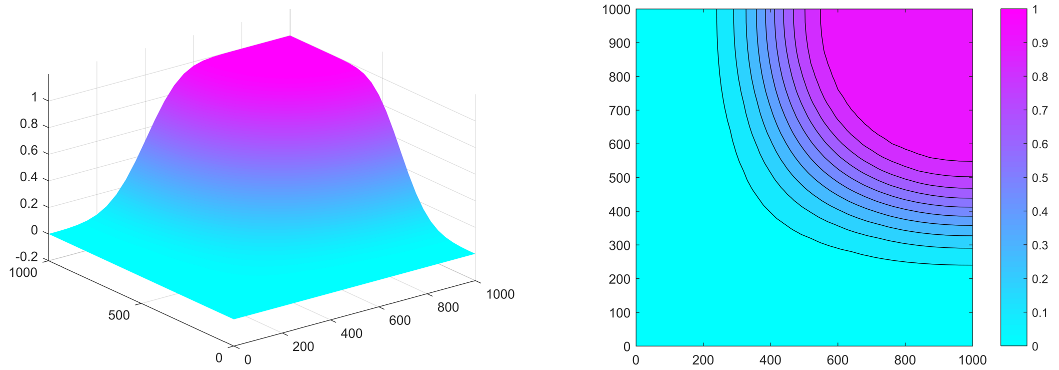


Figure 7.7: Numerical solution for the concentration (*left*) and contour plot (*right*) after 3 years in Test A.

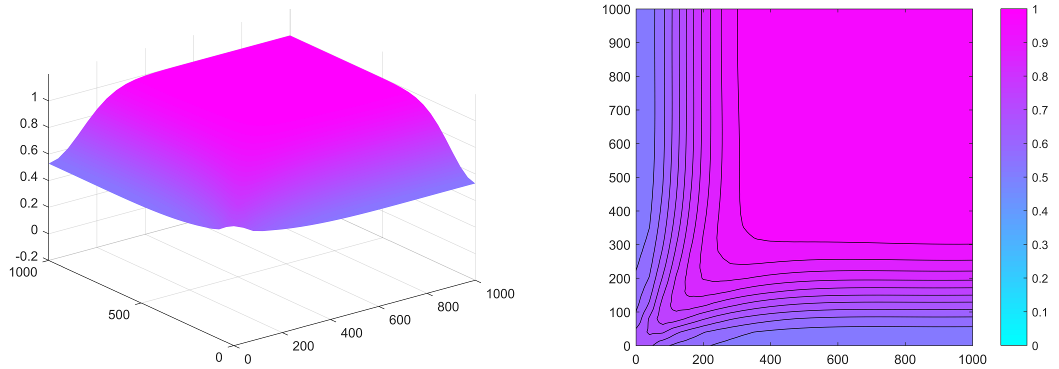


Figure 7.8: Numerical solution for the concentration (*left*) and contour plot (*right*) after 10 years in Test A.

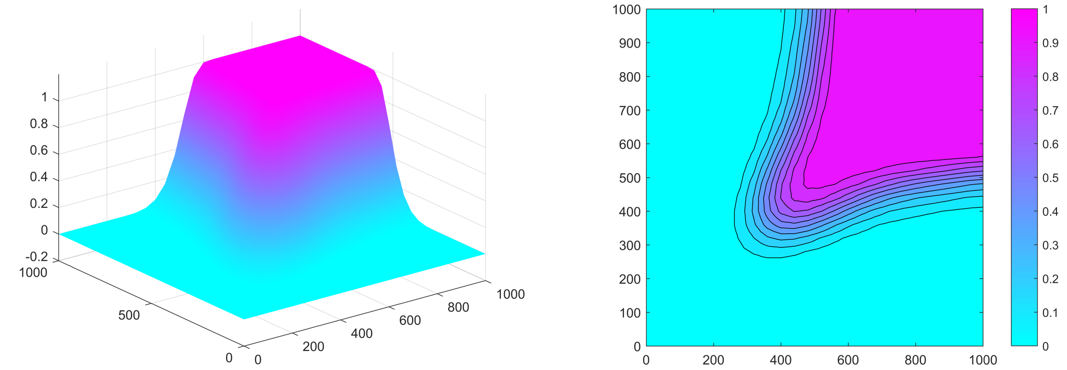


Figure 7.9: Numerical solution for the concentration (*left*) and contour plot (*right*) after 3 years in Test B.

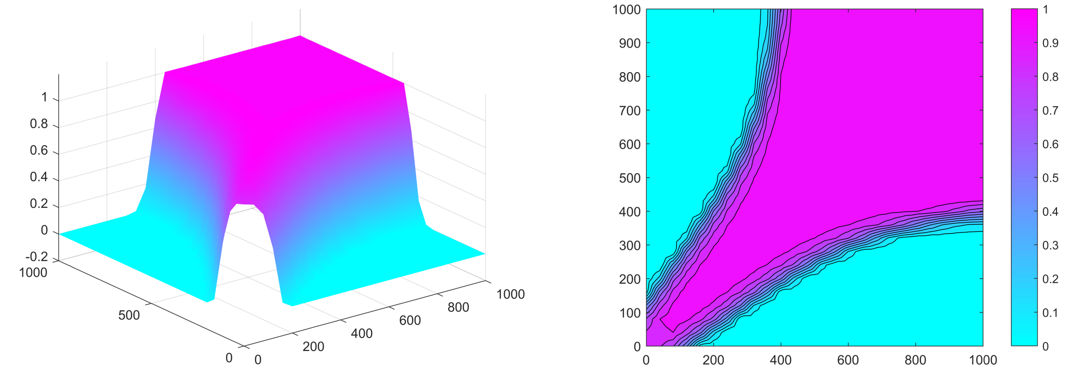


Figure 7.10: Numerical solution for the concentration (*left*) and contour plot (*right*) after 10 years in Test B.

Chapter 8

Outlook and open questions

Here, we summarize some open questions that are presently under investigation and are or could be the topic of future research.

The first one is related to the theoretical analysis of the p- and hp-versions of the nonconforming Trefftz VEM for the Helmholtz problem introduced in Chapters 4 and 5. We deem that corresponding error estimates could be proven, provided that one is able to derive, for a given stabilization, the explicit dependence of the discrete Gårding inequality and continuity constants in (4.31) in terms of p. We refer to the works [14,39,40,84,142] on p- and hp-versions of Poisson-type boundary value problems in the framework of VEM.

Another issue is the extension of the nonconforming Trefftz VEM to the 3D case, which, owing to the nonconforming setting and the fact that no internal moments are involved, should be rather straightforwardly obtained from the 2D case, see also the hints on the construction of a 3D Trefftz VEM for the Laplace problem in Section 3.2.5.

Regarding Chapter 6, the generalization of the nonconforming Trefftz VEM to the case of Helmholtz boundary value problems with smoothly varying wave number is a hot topic. We highlight that such problems have already been tackled, for Trefftz DG, in [127,128,130], where *generalized plane waves* (GPW), which are exponential functions of complex polynomials, computed by minimizing the residual in the fulfillment of the Helmholtz equation, were introduced. Recently, in [129], the construction of GPW was also described for a wider set of PDEs, and corresponding local interpolation properties were derived.

Finally, an open issue is also the question of the generalization of the nonconforming Trefftz VEM and its analysis to the case of the time-harmonic Maxwell problem. So far, the only VEM works for Maxwell are for static problems [32–34].

Concerning the miscible displacement problem in Chapter 7, several issues are of interest. Among them is, similarly as for the Helmholtz problem, the extension to the 3D case. Moreover, the theoretical analysis when using time integration schemes of higher order than order one, such as the second-order Crank-Nicolson method, is worth studying. Finally, the combination of the method with upwinding schemes (instead of the flux-correcting transport schemes at the algebraic level) in order to mitigate the oscillations occurring for strongly convection-dominated problems could be future work.

Bibliography

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55. Courier Corporation, 1964.

[2] R. A. Adams and J. J. F. Fournier. *Sobolev Spaces*, volume 140. Academic Press, 2003.

[3] B. Ahmad, A. Alsaedi, F. Brezzi, L. D. Marini, and A. Russo. Equivalent projectors for virtual element methods. *Comput. Math. Appl.*, 66(3):376–391, 2013.

[4] M. Ainsworth. Discrete dispersion relation for hp-version finite element approximation at high wave number. *SIAM J. Num. Anal.*, 42:553–575, 2004.

[5] M. Ainsworth. Dispersive and dissipative behaviour of high order discontinuous Galerkin finite element methods. *J. Comput. Phys.*, 198(1):106–130, 2004.

[6] M. Ainsworth, P. Monk, and W. Muniz. Dispersive and dissipative properties of discontinuous Galerkin finite element methods for the second-order wave equation. *J. Sci. Comput.*, 27(1-3):5–40, 2006.

[7] Y. Amirat and A. Ziani. Asymptotic behavior of the solutions of an elliptic-parabolic system arising in flow in porous media. *Zeitschrift für Analysis und ihre Anwendungen*, 23(2):335–351, 2004.

[8] P. F. Antonietti, L. Beirão da Veiga, D. Mora, and M. Verani. A stream virtual element formulation of the Stokes problem on polygonal meshes. *SIAM J. Numer. Anal.*, 52(1):386–404, 2014.

[9] P. F. Antonietti, L. Beirão da Veiga, S. Scacchi, and M. Verani. A $\mathcal{C}^1$ virtual element method for the Cahn–Hilliard equation with polygonal meshes. *SIAM J. Numer. Anal.*, 54(1):34–56, 2016.

[10] P. F. Antonietti, M. Bruggi, S. Scacchi, and M. Verani. On the virtual element method for topology optimization on polygonal meshes: a numerical study. *Comput. Math. Appl.*, 74(5):1091–1109, 2017.

[11] P. F. Antonietti, A. Cangiani, J. Collis, Z. Dong, E. H. Georgoulis, S. Giani, and P. Houston. Review of discontinuous Galerkin finite element methods for partial differential equations on complicated domains. In *Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations*, pages 279–308. Springer, 2016.

[12] P. F. Antonietti, G. Manzini, and M. Verani. The conforming virtual element method for polyharmonic problems. `http://arXiv:1811.04317`, 2018.

[13] P. F. Antonietti, G. Manzini, and M. Verani. The fully nonconforming virtual element method for biharmonic problems. *Math. Models Methods Appl. Sci.*, 28(02):387–407, 2018.

[14] P. F. Antonietti, L. Mascotto, and M. Verani. A multigrid algorithm for the p-version of the virtual element method. *ESAIM Math. Model. Numer. Anal.*, 52(1):337–364, 2018.

[15] N. Aronszajn. A unique continuation theorem for solutions of elliptic partial differential equations or inequalities of second order. *J. Math. Pures Appl.*, 36(9):235–249, 1957.

[16] J. Asakura, T. Sakurai, H. Tadano, T. Ikegami, and K. Kimura. A numerical method for nonlinear eigenvalue problems using contour integrals. *JSIAM Letters*, 1:52–55, 2009.

[17] S. Axler, P. Bourdon, and R. Wade. *Harmonic Function Theory*, volume 137. Springer Science & Business Media, 2013.

[18] B. Ayuso, K. Lipnikov, and G. Manzini. The nonconforming virtual element method. *ESAIM Math. Model. Numer. Anal.*, 50(3):879–904, 2016.

[19] I. Babuška and B. Guo. The *hp* version of the finite element method. *Comput. Mech.*, 1(1):21–41, 1986.

[20] I. Babuška and B. Guo. The *hp* version of the finite element method for domains with curved boundaries. *SIAM J. Numer. Anal.*, 25(4):837–861, 1988.

[21] I. Babuška and B. Guo. Regularity of the solution of elliptic problems with piecewise analytic data. Part I. Boundary value problems for linear elliptic equation of second order. *SIAM J. Math. Anal.*, 19(1):172–203, 1988.

[22] I. Babuška and J. M. Melenk. The partition of unity finite element method: basic theory and applications. *Comput. Methods Appl. Mech. Engrg.*, 139(1-4):289–314, 1996.

[23] I. Babuška and J. M. Melenk. Approximation with harmonic and generalized harmonic polynomials in the partition of unity method. *Comput. Assist. Methods Eng. Sci.*, 4:607–632, 1997.

[24] I. Babuška and S. Sauter. Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers? *SIAM J. Numer. Anal.*, 34(6):2392–2423, 1997.

[25] L. Banjai, E. H. Georgoulis, and O. Lijoka. A Trefftz polynomial space-time discontinuous Galerkin method for the second order wave equation. *SIAM J. Numer. Anal.*, 55(1):63–86, 2017.

[26] R. E. Bank, A. Parsania, and S. Sauter. Saturation estimates for *hp*-finite element methods. *Comput. Vis. Sci.*, 16(5):195–217, 2013.

[27] R. Bañuelos and T. Carroll. Brownian motion and the fundamental frequency of a drum. *Duke Math. J.*, 75(3):575–602, 1994.

[28] S. Bartels, M. Jensen, and R. Müller. Discontinuous Galerkin finite element convergence for incompressible miscible displacement problems of low regularity. *SIAM J. Numer. Anal.*, 47(5):3720–3743, 2009.

[29] J. Bear. *Dynamics of fluids in porous media.* Courier Corporation, 2013.

[30] L. Beirão da Veiga, F. Brezzi, L. D. Marini, and A. Russo. Mixed virtual element methods for general second order elliptic problems on polygonal meshes. *ESAIM Math. Model. Numer. Anal.*, 50(3):727–747, 2016.

[31] L. Beirão da Veiga, F. Brezzi, A. Cangiani, G. Manzini, L. D. Marini, and A. Russo. Basic principles of virtual element methods. *Math. Models Methods Appl. Sci.*, 23(01):199–214, 2013.

[32] L. Beirão da Veiga, F. Brezzi, F. Dassi, L. D. Marini, and A. Russo. Virtual element approximation of 2D magnetostatic problems. *Comput. Methods Appl. Mech. Engrg.*, 327:173–195, 2017.

[33] L. Beirão da Veiga, F. Brezzi, F. Dassi, L. D. Marini, and A. Russo. A family of three-dimensional virtual elements with applications to magnetostatics. *SIAM J. Numer. Anal.*, 56(5):2940–2962, 2018.

[34] L. Beirão da Veiga, F. Brezzi, F. Dassi, L. D. Marini, and A. Russo. Lowest order virtual element approximation of magnetostatic problems. *Comput. Methods Appl. Mech. Engrg.*, 332:343–362, 2018.

[35] L. Beirão da Veiga, F. Brezzi, and L. D. Marini. Virtual elements for linear elasticity problems. *SIAM J. Numer. Anal.*, 51:794–812, 2013.

[36] L. Beirão da Veiga, F. Brezzi, L. D. Marini, and A. Russo. The hitchhiker's guide to the virtual element method. *Math. Models Methods Appl. Sci.*, 24(8):1541–1573, 2014.

[37] L. Beirão da Veiga, F. Brezzi, L. D. Marini, and A. Russo. H(div) and H(curl)-conforming virtual element methods. *Numerische Mathematik*, 133(2):303–332, 2016.

[38] L. Beirão da Veiga, F. Brezzi, L. D. Marini, and A. Russo. Virtual element method for general second-order elliptic problems on polygonal meshes. *Math. Models Methods Appl. Sci.*, 26(4):729–750, 2016.

[39] L. Beirão da Veiga, A. Chernov, L. Mascotto, and A. Russo. Basic principles of *hp* virtual elements on quasiuniform meshes. *Math. Models Methods Appl. Sci.*, 26(8):1567–1598, 2016.

[40] L. Beirão da Veiga, A. Chernov, L. Mascotto, and A. Russo. Exponential convergence of the *hp* virtual element method with corner singularity. *Numer. Math.*, 138(3):581–613, 2018.

[41] L. Beirão da Veiga, F. Dassi, and A. Russo. High-order virtual element method on polyhedral meshes. *Comput. Math. Appl.*, 74:1110–1122, 2017.

[42] L. Beirão da Veiga, K. Lipnikov, and G. Manzini. *The Mimetic Finite Difference Method for elliptic problems*, volume 11. Springer, 2014.

[43] L. Beirão da Veiga, C. Lovadina, and A. Russo. Stability analysis for the virtual element method. *Math. Models Methods Appl. Sci.*, 27(13):2557–2594, 2017.

[44] L. Beirão da Veiga, C. Lovadina, and G. Vacca. Divergence free virtual elements for the Stokes problem on polygonal meshes. *ESAIM Math. Model. Numer. Anal.*, 51(2):509–535, 2017.

[45] L. Beirão da Veiga and G. Manzini. A virtual element method with arbitrary regularity. *IMA J. Numer. Anal.*, 34(2):759–781, 2014.

[46] L. Beirão da Veiga, A. Pichler, and G. Vacca. A virtual element method for the miscible displacement of incompressible fluids in porous media. `https://arxiv.org/abs/1907.13080`, 2019. Submitted.

[47] L. Beirão da Veiga, A. Russo, and G. Vacca. The virtual element method with curved edges. *ESAIM Math. Model. Numer. Anal.*, 53(2):375–404, 2019.

[48] M. F. Benedetto, S. Berrone, A. Borio, S. Pieraccini, and S. Scialò. Order preserving SUPG stabilization for the virtual element formulation of advection–diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, 311:18–40, 2016.

[49] M. F. Benedetto, S. Berrone, S. Pieraccini, and S. Scialò. The virtual element method for discrete fracture network simulations. *Comput. Meth. Appl. Mech. Engrg.*, 280:135–156, 2014.

[50] C. Bernardi, N. Fiétier, and R. G. Owens. An error indicator for mortar element solutions to the Stokes problem. *IMA J. Numer. Anal.*, 21(4):857–886, 2001.

[51] C. Bernardi and Y. Maday. Polynomial interpolation results in Sobolev spaces. *J. Comput. Appl. Math.*, 43(1):53–80, 1992.

[52] S. Berrone, A. Borio, and G. Manzini. SUPG stabilization for the nonconforming virtual element method for advection–diffusion–reaction equations. *Comput. Methods Appl. Mech. Engrg.*, 340:500–529, 2018.

[53] S. Bertoluzza, M. Pennacchio, and D. Prada. BDDC and FETI-DP for the virtual element method. *Calcolo*, 54(4):1565–1593, 2017.

[54] W.-J. Beyn. An integral method for solving nonlinear eigenvalue problems. *Linear Algebra and its Applications*, 436(10):3839–3863, 2012.

[55] J. Bramble and L. Payne. Bounds in the Neumann problem for second order uniformly elliptic operators. *Pac. J. Math.*, 12(3):823–833, 1962.

[56] M. Braun and M. Golubitsky. *Differential Equations and Their Applications*, volume 4. Springer, 1983.

[57] S. C. Brenner. Poincaré–Friedrichs inequalities for piecewise H^1 functions. *SIAM J. Numer. Anal.*, 41(1):306–324, 2003.

[58] S. C. Brenner, Q. Guan, and L. Y. Sung. Some estimates for virtual element methods. *Comput. Methods Appl. Math.*, 17(4):553–574, 2017.

[59] S. C. Brenner and L. R. Scott. *The mathematical theory of Finite Element Methods*, volume 15. Texts in Applied Mathematics, Springer-Verlag, New York, third edition, 2008.

[60] F. Brezzi, R. S. Falk, and L. D. Marini. Basic principles of mixed virtual element methods. *Math. Mod. Num. Anal.*, 48(4):1227–1240, 2014.

[61] F. Brezzi, L. P. Franca, and A. Russo. Further considerations on residual-free bubbles for advective-diffusive equations. *Comput. Methods Appl. Mech. Engrg.*, 166(1-2):25–33, 1998.

[62] A. N. Brooks and T. J. R. Hughes. Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, 32(1-3):199–259, 1982.

[63] E. Burman. Consistent SUPG-method for transient transport problems: Stability and convergence. *Comput. Methods Appl. Mech. Engrg.*, 199(17-20):1114–1123, 2010.

[64] A. Cangiani, P. Chatzipantelidis, G. Diwan, and E. H. Georgoulis. Virtual element method for quasilinear elliptic problems. `http://arXiv:1707.01592`, 2018.

[65] A. Cangiani, Z. Dong, E. H. Georgoulis, and P. Houston. *hp*-version discontinuous Galerkin methods for advection-diffusion-reaction problems on polytopic meshes. *ESAIM Math. Model. Numer. Anal.*, 50(3):699–725, 2016.

[66] A. Cangiani, Z. Dong, E. H. Georgoulis, and P. Houston. *hp-Version discontinuous Galerkin methods on polygonal and polyhedral meshes.* Springer, 2017.

[67] A. Cangiani, E. H. Georgoulis, T. Pryer, and O. J. Sutton. A posteriori error estimates for the virtual element method. *Numer. Math.*, 137:857–893, 2017.

[68] A. Cangiani, V. Gyrya, and G. Manzini. The non-conforming virtual element method for the Stokes equations. *SIAM J. Numer. Anal.*, 54(6):3411–3435, 2016.

[69] A. Cangiani, G. Manzini, and O. J. Sutton. Conforming and nonconforming virtual element methods for elliptic problems. *IMA J. Numer. Anal.*, 37:1317–1354, 2016.

[70] O. Cessenat. *Application d'une nouvelle formulation variationnelle aux équations d'ondes harmoniques: problèmes de Helmholtz 2D et de Maxwell 3D.* PhD thesis, Paris 9, 1996.

[71] O. Cessenat and B. Després. Application of an ultra weak variational formulation of elliptic PDEs to the two-dimensional Helmholtz problem. *SIAM J. Numer. Anal.*, 35(1):255–299, 1998.

[72] O. Cessenat and B. Després. Using plane waves as base functions for solving time harmonic equations with the ultra weak variational formulation. *J. Comput. Acoust.*, 11(02):227–238, 2003.

[73] C. Chainais-Hillairet and J. Droniou. Convergence analysis of a mixed finite volume scheme for an elliptic-parabolic system modeling miscible fluid flows in porous media. *SIAM J. Numer. Anal.*, 45(5):2228–2258, 2007.

[74] C. Chainais-Hillairet, S. Krell, and A. Mouton. Study of discrete duality finite volume schemes for the Peaceman model. *SIAM J. Sci. Comput.*, 35(6):A2928–A2952, 2013.

[75] C. Chainais-Hillairet, S. Krell, and A. Mouton. Convergence analysis of a DDFV scheme for a system describing miscible fluid flows in porous media. *Numer. Methods Partial Differ. Equ.*, 31(3):723–760, 2015.

[76] G. Chavent and J. Jaffré. *Mathematical models and finite elements for reservoir simulation: single phase, multiphase and multicomponent flows through porous media*, volume 17. Elsevier, 1986.

[77] Z. Chen and R. Ewing. Mathematical analysis for reservoir models. *SIAM J. Math. Anal.*, 30(2):431–453, 1999.

[78] A. Chernov. Optimal convergence estimates for the trace of the polynomial L^2-projection operator on a simplex. *Math. Comp.*, 81(278):765–787, 2012.

[79] A. Chernov and L. Mascotto. The harmonic virtual element method: stabilization and exponential convergence for the Laplace problem on polygonal domains, 2018. doi: `https://doi.org/10.1093/imanum/dry038`.

[80] P. G. Ciarlet, P. Ciarlet, S. A. Sauter, and C. Simian. Intrinsic finite element methods for the computation of fluxes for Poisson's equation. *Numerische Mathematik*, 132(3):433–462, 2016.

[81] B. Cockburn, J. Gopalakrishnan, and R. Lazarov. Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems. *SIAM J. Numer. Anal.*, 47(2):1319–1365, 2009.

[82] S. Congreve, P. Houston, and I. Perugia. Adaptive refinement for *hp*–version Trefftz discontinuous Galerkin methods for the homogeneous Helmholtz problem. *Adv. Comput. Math.*, 45(1):361–393, 2019.

[83] M. Crouzeix and P.-A. Raviart. Conforming and nonconforming finite element methods for solving the stationary Stokes equations. *RAIRO Anal. Numér.*, 7(R3):33–75, 1973.

[84] F. Dassi and L. Mascotto. Exploring high-order three dimensional virtual elements: bases and stabilizations. *Comput. Math. Appl.*, 75(9):3379–3401, 2018.

[85] E. Deckers, O. Atak, L. Coox, R. D'Amico, H. Devriendt, S. Jonckheere, K. Koo, B. Pluymers, D. Vandepitte, and W. Desmet. The wave based method: An overview of 15 years of research. *Wave Motion*, 51(4):550–565, 2014.

[86] A. Deraemaeker, I. Babuška, and P. Bouillard. Dispersion and pollution of the FEM solution for the Helmholtz equation in one, two and three dimensions. *Int. J. Numer. Meth. Engng.*, 46:471–499, 1999.

[87] B. Després. Sur une formulation variationnelle de type ultra-faible. *C. R. Acad. Sci. Paris, Ser. I*, 318(10):939–944, 1994.

[88] D. A. Di Pietro and A. Ern. Hybrid high-order methods for variable-diffusion problems on general meshes. *C. R. Math. Acad. Sci. Paris*, 353(1):31–34, 2015.

[89] J. Douglas, Jr. Finite difference methods for two-phase incompressible flow in porous media. *SIAM J. Numer. Anal.*, 20(4):681–696, 1983.

[90] J. Douglas, Jr., R. E. Ewing, and M. F. Wheeler. The approximation of the pressure by a mixed method in the simulation of miscible displacement. *RAIRO. Analyse numérique*, 17(1):17–33, 1983.

[91] J. Droniou, R. Eymard, A. Prignet, and K. S. Talbot. Unified convergence analysis of numerical schemes for a miscible displacement problem. *Found. Comput. Math.*, 19(2):333–374, 2019.

[92] J. Droniou and K. S. Talbot. On a miscible displacement model in porous media flow with measure data. *SIAM J. Numer. Anal.*, 46(5):3158–3175, 2014.

[93] Y. Du and H. Wu. Preasymptotic error analysis of higher order FEM and CIP-FEM for Helmholtz equation with high wave number. *SIAM J. Numer. Anal.*, 53(2):782–804, 2015.

[94] H. Egger, F. Kretzschmar, S. M. Schnepp, and T. Weiland. A space-time discontinuous Galerkin Trefftz method for time dependent Maxwell's equations. *SIAM J. Sci. Comput.*, 37(5):B689–B711, 2015.

[95] C. L. Epstein and M. O'Neil. Smoothed corners and scattered waves. *SIAM J. Sci. Comput.*, 38(5):A2665–A2698, 2016.

[96] S. Esterhazy and J. M. Melenk. On stability of discretizations of the Helmholtz equation. In I. Graham, T. Hou, O. Lakkis, and R. Scheichl, editors, *Numerical analysis of multiscale problems*, volume 83 of *Lecture Notes in Computational Science and Engineering*, pages 285–324. Springer, 2011.

[97] L. C. Evans. *Partial Differential Equations.* American Mathematical Society, 2010.

[98] R. E. Ewing. *The mathematics of reservoir simulation.* SIAM, 1983.

[99] R. E. Ewing, T. F. Russell, and M. F. Wheeler. Simulation of miscible displacement using mixed methods and a modified method of characteristics. In *SPE Reservoir Simulation Symposium.* Society of Petroleum Engineers, 1983.

[100] R. E. Ewing, T. F. Russell, and M. F. Wheeler. Convergence analysis of an approximation of miscible displacement in porous media by mixed finite elements and a modified method of characteristics. *Comput. Methods Appl. Mech. Engrg.*, 47(1-2):73–92, 1984.

[101] C. Farhat, I. Harari, and L. P. Franca. The discontinuous enrichment method. *Comput. Methods Appl. Mech. Engrg.*, 190(48):6455–6479, 2001.

[102] X. B. Feng. On existence and uniqueness results for a coupled system modeling miscible displacement in porous media. *J. Math. Anal. Appl.*, 194(3):883–910, 1995.

[103] L. P. Franca and L. Tobiska. Stability of the residual free bubble method for bilinear finite elements on rectangular grids. *IMA J. Numer. Anal.*, 22(1):73–87, 2002.

[104] E. Gagliardo. Caratterizzazioni delle tracce sulla frontiera relative ad alcune classi di funzioni in n variabili. *Rend. Sem. Mat. Univ. Padova*, 27(405):284–305, 1957.

[105] A. L. Gain, G. H. Paulino, S. D. Leonardo, and I. F. M. Menezes. Topology optimization using polytopes. *Comput. Methods Appl. Mech. Engrg.*, 293:411–430, 2015.

[106] S. Ganesan and L. Tobiska. Stabilization by local projection for convection–diffusion and incompressible flow problems. *J. Sci. Comput.*, 43(3):326–342, 2010.

[107] F. Gardini and G. Vacca. Virtual element method for second order elliptic eigenvalue problems. *IMA J. Numer. Anal.*, 2017.

[108] T. Gelhard, G. Lube, M. A. Olshanskii, and J.-H. Starcke. Stabilized finite element schemes with LBB-stable elements for incompressible flows. *J. Comput. Appl. Math.*, 177(2):243–267, 2005.

[109] E. H. Georgoulis, E. Hall, and P. Houston. Discontinuous Galerkin methods on *hp*–anisotropic meshes II: A posteriori error analysis and adaptivity. *Appl. Numer. Math.*, 59(9):2179–2194, 2009.

[110] C. J. Gittelson. Plane wave discontinuous Galerkin methods. Master's thesis, SAM-ETH Zürich, 2008.

[111] C. J. Gittelson and R. Hiptmair. Dispersion analysis of plane wave discontinuous Galerkin methods. *Int. J. Numer. Meth. Engng.*, 98(5):313–323, 2014.

[112] C. J. Gittelson, R. Hiptmair, and I. Perugia. Plane wave discontinuous Galerkin methods: analysis of the *h*-version. *ESAIM Math. Model. Numer. Anal.*, 43(2):297–331, 2009.

[113] J. Gopalakrishnan, I. Muga, and N. Olivares. Dispersive and dissipative errors in the DPG method with scaled norms for Helmholtz equation. *SIAM J. Sci. Comput.*, 36(1):A20–A39, 2014.

[114] J. Gopalakrishnan, M. Solano, and F. Vargas. Dispersion Analysis of HDG Methods. *J. Sci. Comput.*, pages 1–33, 2018.

[115] I. Graham and S. Sauter. Stability and finite element error analysis for the Helmholtz equation with variable coefficients. *Math. Comp.*, 2019. doi: `https://doi.org/10.1090/mcom/3457`.

[116] P. Grisvard. *Elliptic problems in nonsmooth domains.* SIAM, 2011.

[117] R. Hiptmair, A. Moiola, and I Perugia. Plane wave approximation of homogeneous Helmholtz solutions. *Z. Angew. Math. Phys.*, 62(5):809, 2011.

[118] R. Hiptmair, A. Moiola, and I. Perugia. Plane wave discontinuous Galerkin methods for the 2D Helmholtz equation: analysis of the p-version. *SIAM J. Numer. Anal.*, 49(1):264–284, 2011.

[119] R. Hiptmair, A. Moiola, and I. Perugia. Error analysis of Trefftz-discontinuous Galerkin methods for the time-harmonic Maxwell equations. *Math. Comp.*, 82(281):247–268, 2013.

[120] R. Hiptmair, A. Moiola, and I. Perugia. Plane wave-discontinuous Galerkin methods: exponential convergence of the *hp*-version. *Found. Comput. Math.*, 16(3):637–675, 2016.

[121] R. Hiptmair, A. Moiola, and I. Perugia. A survey of Trefftz methods for the Helmholtz equation. In *Building bridges: connections and challenges in modern approaches to numerical partial differential equations*, pages 237–279. Springer, 2016.

[122] R. Hiptmair, A. Moiola, I. Perugia, and C. Schwab. Approximation by harmonic polynomials in star-shaped domains and exponential convergence of Trefftz *hp*-dGFEM. *ESAIM Math. Model. Numer. Anal.*, 48(3):727–752, 2014.

[123] H. Hu, Y. Fu, and J. Zhou. Numerical solution of a miscible displacement problem with dispersion term using a two-grid mixed finite element approach. *J. Numer. Algor.*, 2018. doi: `https://doi.org/10.1007/s11075-018-0575-2`.

[124] T. Huttunen, P. Monk, and J. P. Kaipio. Computational aspects of the ultra-weak variational formulation. *J. Comp. Phys.*, 182(1):27–46, 2002.

[125] F. Ihlenburg and I. Babuška. Dispersion analysis and error estimation of Galerkin finite element methods for the Helmholtz equation. *Int. J. Numer. Meth. Eng.*, 38(22):3745–3774, 1995.

[126] F. Ihlenburg and I. Babuška. Finite element solution of the Helmholtz equation with high wave number Part I: The h-version of the FEM. *Comput. Math. Appl.*, 30(9):9–37, 1995.

[127] L.-M. Imbert-Gérard. Interpolation properties of generalized plane waves. *Numerische Mathematik*, 131(4):683–711, 2015.

[128] L.-M. Imbert-Gérard and B. Després. A generalized plane-wave numerical method for smooth nonconstant coefficients. *IMA J. Numer. Anal.*, 34(3):1072–1103, 2013.

[129] L.-M. Imbert-Gérard and Sylvand G. A roadmap for generalized plane waves and their interpolation properties. `http://arXiv:1907.08309`, 2019.

[130] L.-M. Imbert-Gérard and P. Monk. Numerical simulation of wave propagation in inhomogeneous media using generalized plane waves. *ESAIM Math. Model. Numer. Anal.*, 51(4):1387–1406, 2017.

[131] I. Kalashnikova, C. Farhat, and R. Tezaur. A discontinuous enrichment method for the finite element solution of high Péclet advection–diffusion problems. *Finite Elem. Anal. Des.*, 45(4):238–250, 2009.

[132] I. Kalashnikova, R. Tezaur, and C. Farhat. A discontinuous enrichment method for variable-coefficient advection–diffusion at high Péclet number. *Int. J. Numer. Meth. Eng.*, 87(1-5):309–335, 2011.

[133] F. Kretzschmar, A. Moiola, I. Perugia, and S. M. Schnepp. A priori error analysis of space–time Trefftz discontinuous Galerkin methods for wave problems. *IMA J. Numer. Anal.*, 36(4):1599–1635, 2015.

[134] F. Kretzschmar, S. M. Schnepp, I. Tsukerman, and T. Weiland. Discontinuous Galerkin methods with Trefftz approximations. *J. Comput. Appl. Math.*, 270:211–222, 2014.

[135] D. Kuzmin. Explicit and implicit FEM-FCT algorithms with flux linearization. *J. Comput. Phys.*, 228(7):2517–2534, 2009.

[136] D. Kuzmin and M. Möller. Algebraic flux correction. I. Scalar conservation laws. In *Flux-corrected transport*, Sci. Comput., pages 155–206. Springer, Berlin, 2005.

[137] J. Ladevèze and P. Ladevèze. Bounds of the Poincaré constant with respect to the problem of star-shaped membrane regions. *Z. Angew. Math. Phys.*, 29(4):670–683, 1978.

[138] F. Li. On the negative-order norm accuracy of a local-structure-preserving LDG method. *J. Sci. Comput.*, 51(1):213–223, 2012.

[139] F. Li and C. Shu. A local-structure-preserving local discontinuous Galerkin method for the Laplace equation. *Methods Appl. Anal.*, 13(2):215, 2006.

[140] K. Lipnikov, G. Manzini, and M. Shashkov. Mimetic finite difference method. *J. Comput. Phys.*, 257:1163–1227, 2014.

[141] T. Luostari, T. Huttunen, and P. Monk. Improvements for the ultra weak variational formulation. *Internat. J. Numer. Methods Engrg.*, 94(6):598–624, 2013.

[142] L. Mascotto. Ill-conditioning in the virtual element method: Stabilizations and bases. *Numer. Methods Partial Differential Equations*, 34(4):1258–1281, 2018.

[143] L. Mascotto, I. Perugia, and A. Pichler. Non-conforming harmonic virtual element method: h- and p-versions. *J. Sci. Comput.*, 77(3):1874–1908, 2018.

[144] L. Mascotto, I. Perugia, and A. Pichler. A nonconforming Trefftz virtual element method for the Helmholtz problem. *Math. Models Methods Appl. Sci.*, 2019. `https://doi.org/10.1142/S0218202519500301`.

[145] L. Mascotto, I. Perugia, and A. Pichler. A nonconforming Trefftz virtual element method for the Helmholtz problem: numerical aspects. *Comput. Methods Appl. Mech. Engrg.*, 347:445–476, 2019.

[146] L. Mascotto and A. Pichler. Extension of the nonconforming Trefftz virtual element method to the Helmholtz problem with piecewise constant wave number. *Appl. Num. Math.*, 2019. `https://doi.org/10.1016/j.apnum.2019.04.005`.

[147] W. C. H. McLean. *Strongly Elliptic Systems and Boundary Integral Equations.* Cambridge University Press, 2000.

[148] J. M. Melenk. Operator adapted spectral element methods I: harmonic and generalized harmonic polynomials. *Numer. Math.*, 84(1):35–69, 1999.

[149] M. Melenk. *On Generalized Finite Element Methods.* PhD thesis, University of Maryland, 1995.

[150] A. Moiola. *Trefftz-discontinuous Galerkin methods for time-harmonic wave problems.* PhD thesis, ETH Zürich, 2011.

[151] P. Monk and D.-Q. Wang. A least-squares method for the Helmholtz equation. *Comput. Methods Appl. Mech. Engrg.*, 175(1-2):121–136, 1999.

[152] D. Mora, G. Rivera, and R. Rodríguez. A virtual element method for the Steklov eigenvalue problem. *Math. Models Methods Appl. Sci.*, 25(08):1421–1445, 2015.

[153] G. Morel, C. Buet, and B. Després. Trefftz discontinuous Galerkin method for Friedrichs systems with linear relaxation: application to the P1 model. *Comput. Methods Appl. Math.*, 18(3):521–557, 2018.

[154] L. E. Payne and H. F. Weinberger. An optimal Poincaré inequality for convex domains. *Arch. Ration. Mech. Anal.*, 5(1):286–292, 1960.

[155] D. W. Peaceman. Improved treatment of dispersion in numerical calculation of multidimensional miscible displacement. *Society of Petroleum Engineers Journal*, 6(03):213–216, 1966.

[156] D. W. Peaceman. *Fundamentals of Numerical Reservoir Simulation.* Elsevier, 1977.

[157] D. W. Peaceman and H. H. Rachford Jr. Numerical calculation of multidimensional miscible displacement. *Society of Petroleum Engineers Journal*, 2(04):327–339, 1962.

[158] I. Perugia and A. Pichler. A numerical study of the dispersion and dissipation properties of virtual element methods for the Helmholtz problem. `https://arxiv.org/abs/1906.09965`, 2019. Submitted.

[159] I. Perugia, P. Pietra, and A. Russo. A plane wave virtual element method for the Helmholtz problem. *ESAIM Math. Model. Numer. Anal.*, 50(3):783–808, 2016.

[160] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*, volume 23. Springer, 2008.

[161] H. Riou, P. Ladeveze, and B. Sourcis. The multiscale VTCR approach applied to acoustics problems. *J. Comput. Acoust.*, 16(04):487–505, 2008.

[162] B. M. Rivière and N. J. Walkington. Convergence of a discontinuous Galerkin method for the miscible displacement equation under low regularity. *SIAM J. Numer. Anal.*, 49(3):1085–1110, 2011.

[163] S. Rjasanow and S. Weißer. Higher order BEM-based FEM on polygonal meshes. *SIAM J. Numer. Anal.*, 50(5):2357–2378, 2012.

[164] H.-G. Roos, M. Stynes, and L. Tobiska. *Robust numerical methods for singularly perturbed differential equations: convection-diffusion-reaction and flow problems*, volume 24. Springer, 2008.

[165] A. H. Schatz. An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. *Math. Comp.*, 28(128):959–962, 1974.

[166] C. Schwab. *p- and hp- Finite Element Methods: Theory and Applications in Solid and Fluid Mechanics.* Clarendon Press Oxford, 1998.

[167] O. Steinbach. *Numerical approximation methods for elliptic boundary value problems: Finite and Boundary Elements.* Springer, 2007.

[168] S. Sun, B. Rivière, and M. F. Wheeler. A combined mixed finite element and discontinuous Galerkin method for miscible displacement problem in porous media. In *Recent progress in computational and applied PDEs*, pages 323–351. Springer, 2002.

[169] C. Talischi, G. H. Paulino, A. Pereira, and I. F. M. Menezes. PolyMesher: a general-purpose mesh generator for polygonal elements written in Matlab. *Struct. Multidiscip. Optim.*, 45:309–328, 2012.

[170] R. Tezaur, L. Zhang, and C. Farhat. A discontinuous enrichment method for capturing evanescent waves in multiscale fluid and fluid/solid problems. *Comput. Methods Appl. Mech. Engrg.*, 197(19-20):1680–1698, 2008.

[171] V. Thomée. *Galerkin Finite Element Methods for Parabolic Problems*, volume 1054. Springer, 1984.

[172] E. Trefftz. Ein Gegenstück zum Ritzschen Verfahren. In *Proc. 2nd Int. Cong. Appl. Mech. Zurich*, pages 131–137, 1926.

[173] H. Triebel. *Interpolation theory, function spaces, differential operators.* North-Holland, 1978.

[174] G. Vacca and L. Beirão da Veiga. Virtual element methods for parabolic problems on polygonal meshes. *Numer. Methods Partial Differ. Equ.*, 31(6):2110–2134, 2015.

[175] A. Veeser and P. Zanotti. Quasi-optimal nonconforming methods for symmetric elliptic problems. II–Overconsistency and classical nonconforming elements. *SIAM J. Numer. Anal.*, 57(1):266–292, 2019.

[176] H. Wang, D. Liang, R. E. Ewing, S. L. Lyons, and G. Qin. An approximation to miscible fluid flows in porous media with point sources and sinks by an Eulerian–Lagrangian localized adjoint method and mixed finite element methods. *SIAM J. Sci. Comput.*, 22(2):561–581, 2000.

[177] M. F. Wheeler and B. L. Darlow. Interior penalty Galerkin procedures for miscible displacement problems in porous media. In *Computational methods in nonlinear mechanics (Proc. Second Internat. Conf., Univ. Texas, Austin, Tex., 1979)*, pages 485–506. North-Holland Amsterdam, 1980.

[178] J. Xiao, C. Zhang, T.-M. Huang, and T. Sakurai. Solving large-scale nonlinear eigenvalue problems by rational interpolation and resolvent sampling based Rayleigh–Ritz method. *Int. J. Numer. Meth. Eng.*, 110(8):776–800, 2017.