Revisiting the Plane-Wave Ultra-Weak Variational Formulation

Hélène Barucq, Abderrahmane Bendali, Julien Diaz, Sébastien Tordeux

▶ **To cite this version:**

Hélène Barucq, Abderrahmane Bendali, Julien Diaz, Sébastien Tordeux. Revisiting the Plane-Wave Ultra-Weak Variational Formulation. 2024. hal-04591459

HAL Id: hal-04591459
https://hal.science/hal-04591459v1

Preprint submitted on 28 May 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Revisiting the Plane-Wave Ultra-Weak Variational Formulation

H. Barucq[†], A. Bendali[‡], J. Diaz[†], and S. Tordeux[†]

[‡]Univ. Toulouse, INSA-Toulouse, IMT UMR CNRS 5219, Toulouse (France)
[†]EPI Makutu, Inria, Université de Pau et des Pays de l'Adour, TotalEnergies, CNRS UMR 5142, Pau (France)

May 28, 2024

Abstract

Several topics of the plane-wave ultra-weak variational formulations for numerically solving the Helmholtz equation are revisited: treatment of the case of piecewise constant anisotropic coefficients, derivation of the formulation, coercivity properties, etc. The construction of one of these formulations, compatible with the transmission and reflection of plane waves, at normal incidence on an interface shared by two elements of the mesh, leads to a general framework which covers all the previously used ultra-weak formulations. We thus show that any ultra-weak formulation can be characterized by its equivalence with a unique discontinuous Galerkin method for which the numerical fluxes are expressed by outgoing traces. It is also shown that the particular ultra-weak formulation, which provided the general framework, can be considered as an upwind scheme in the sense that these numerical fluxes can be obtained from a Riemann solver. Based on the theory of elliptic interface boundary-value problems, conditions on the coefficients and the geometry ensuring the coercivity properties of the formulation are brought out in the 2D case. The identification of two ways of describing plane waves in the anisotropic case and an appropriate change of variables reduce the estimates of the convergence error to those related to the usual Helmholtz equation. This also allows us to derive a theoretical basis for the choice of local plane-wave bases for an efficient coverage of the anisotropic case. Some numerical experiments illustrating the efficiency of the approach complete the study.

1 Introduction

As reported in many publications (see, e.g., [29, 34, 15, etc.]), the numerical solution of the Helmholtz equation is at the heart of a wide variety of scientific

and engineering activities. However, usual approaches based on local polynomial approximations like continuous or discontinuous Finite Element Methods (FEMs) can quickly become prohibitively expensive when the frequency gets higher due to the corresponding increase of the solution oscillations (see, e.g., [29, 19, etc.]). Trefftz methods (cf., e.g., [27]), which use linear expansions of exact solutions of the interior equation as test and trial functions on each element of the mesh, are aimed to face this difficulty. The plane-wave Ultra-Weak Variational Formulation (UWVF), which was introduced in [11, 12], is currently considered to be among the most effective of such methods [34]. We refer for instance to [29] for a comprehensive discussion about this issue.

To clearly bring out our purpose in this study, we review some of the features of the UWVF in the framework of the usual Helmholtz equation. For an isotropic homogeneous medium of propagation, possibly after a normalization of the physical units, the time-harmonic wave equation is nothing but the usual Helmholtz equation

$$\Delta p + \omega^2 p = 0 \text{ in } \Omega \tag{1}$$

where here, for simplicity, Ω is supposed to be an open polygonal/polyhedral domain of $\mathbb{R}^d$, $d = 2, 3$, and ω is the angular frequency. Equation (1) can be seen as the equation of propagation of sound, written in terms of the phasor p of the acoustic disturbance $p^{\#}$ of the pressure given by the following real part

$$p^{\#}(x,t) = \Re\left(p(x)\, e^{-i\omega t}\right),\ x \in \Omega,\ t \in \mathbb{R}, \tag{2}$$

with a sound speed c furthermore normalized to 1.

The UWVF is set out as follows. Let $\mathcal{T}$ be a finite non-overlapping decomposition of Ω, the elements of which being for simplicity polygonal/polyhedral open subdomains of Ω, generically denoted by T; more precisely

$$\overline{\Omega} = \bigcup_{T\in\mathcal{T}} \overline{T}, \quad T \cap L = \varnothing \text{ if } T \neq L.$$

The non-overlapping decomposition $\mathcal{T}$ plays the role of the mesh even if it is not constrained by the usual restrictive matching conditions of the FEM. As a rule $\boldsymbol{n}_T$ will denote the unit normal on the boundary ∂T of T, directed towards the exterior of T and p_T stands for the restriction $p|_T$ of p to T.

One fundamental characteristic of the UWVF is to write the standard transmission conditions

$$p_T = p_L, \quad \boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T + \boldsymbol{\nabla} p_L \cdot \boldsymbol{n}_L = 0, \tag{3}$$

at any internal edge/face $F_{\{T,L\}}$, shared by the boundaries ∂T and ∂L of two elements T and L of $\mathcal{T}$, in the following equivalent form,

$$\begin{cases} -\boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T + i\omega\eta p_T = \boldsymbol{\nabla} p_L \cdot \boldsymbol{n}_L + i\omega\eta p_L, \\ -\boldsymbol{\nabla} p_L \cdot \boldsymbol{n}_L + i\omega\eta p_L = \boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T + i\omega\eta p_T, \end{cases} \tag{4}$$

possibly except for a multiplicative constant, where η is some normalization positive constant. It is worth mentioning that the notation $F_{\{T,L\}}$ for the internal edge/face aims to indicate that there is no prescribed order on the pairing

$\{T, L\}$. Parameter η characterizes the particular ultra-weak formulation being considered. For example, this is the framework in [11] with $\eta = 1$ and after the substitution of $-i$ for i to be in accordance with the time-dependency in (2) of the phasors. This is also the framework in [39], that can be retrieved by substituting ω for κ.

To be handled with the UWVF, the boundary conditions are put in the following form

$$-\boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T + i\omega\eta p_T = Q_T \left(\boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T + i\omega\eta p_T\right) + g_T \text{ on } F_T^{\partial\Omega} \tag{5}$$

where $F_T^{\partial\Omega}$ is any external edge/face of ∂T lying on $\partial\Omega$, with $Q_T = -1$ for a Dirichlet, $Q_T = +1$ for a Neumann, and $|Q_T| < 1$ for a Fourier-Robin condition.

To get more insight about transmission conditions (4), let us consider the total energy, the sum of kinetic and potential energy, contained in element T, where $p_T^{\#}$ is the time-dependent function defined in (2)

$$\mathcal{E}_T\left(p_T^{\#}\right)(t) = \frac{1}{2}\int_T \left|\partial_t p_T^{\#}\right|^2 dx + \frac{1}{2}\int_T \left|\boldsymbol{\nabla} p_T^{\#}\right|^2 dx.$$

The derivative of this energy yields

$$\partial_t \mathcal{E}_T\left(p_T^{\#}\right)(t) = \int_T \partial_t^2 p_T^{\#} \partial_t p_T^{\#} dx + \int_T \boldsymbol{\nabla} p_T^{\#} \cdot \boldsymbol{\nabla} \partial_t p_T^{\#} dx.$$

Since $\pm\boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T + i\omega\eta p_T = 0$ corresponds to

$$\boldsymbol{\nabla} p_T^{\#} \cdot \boldsymbol{n}_T \mp \eta \partial_t p_T^{\#} = 0 \text{ on } \partial T,$$

Green's formula then yields

$$\partial_t \mathcal{E}_T\left(p_T^{\#}\right)(t) = \int_T \left(\underbrace{\partial_t^2 p_T^{\#} - \Delta p_T^{\#}}_{-=0}\right) \partial_t p_T^{\#} dx \pm \int_{\partial T} \eta \left|\partial_t p_T^{\#}\right|^2 ds,$$

where ds is the Lebesgue measure on ∂T. Condition $+\boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T + i\omega\eta p_T = 0$ therefore causes that the net flux of energy to go from the outside of T to the interior of T and condition $-\boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T + i\omega\eta p_T = 0$ the opposite. It is thus relevant to call these traces respectively the "*outgoing trace*" and the "*incoming trace*" as this is done for example in [39].

This terminology is also supported by a plane-wave analysis. Helmholtz equation (1), corresponds to a sound speed $c = 1$, and hence to a wavenumber $\kappa = \omega/c = \omega$. Assume that $p_T = p_T^+ + p_T^-$ and $p_L = p_L^+ + p_L^-$ where $p_T^{\pm}$ and $p_L^{\pm}$ are plane waves propagating respectively in T and L and in the respective directions of $\pm\boldsymbol{n}_T$ and $\pm\boldsymbol{n}_L$

$$p_T^{\pm}(x) = \alpha_T^{\pm} \exp\left(\pm i\omega \boldsymbol{n}_T \cdot x\right), \quad p_L^{\pm}(x) = \alpha_L^{\pm} \exp\left(\pm i\omega \boldsymbol{n}_L \cdot x\right).$$

Taking $\eta = 1$, we find that the outgoing and incoming traces are then given by

$$\pm\boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T + i\omega p_T = 2i\omega p_T^{\pm}, \quad \pm\boldsymbol{\nabla} p_L \cdot \boldsymbol{n}_L + i\omega p_L = 2i\omega p_L^{\pm}$$

By normalizing the expressions of these traces, we get

$$\pm\frac{1}{2i\omega}\left(\boldsymbol{\nabla} p_T\cdot\boldsymbol{n}_T\pm i\omega p_T\right)=p_T^{\pm}.\quad\pm\frac{1}{2i\omega}\left(\boldsymbol{\nabla} p_L\cdot\boldsymbol{n}_L\pm i\omega p_L\right)=p_L^{\pm}\qquad(6)$$

The transmission conditions (4) just express in this case that

$$p_T^{-}=p_L^{+},\quad p_L^{-}=p_T^{+},$$

or in other words, that the trace of the outgoing wave from L gives rise to that of the incoming wave in T and reciprocally the trace of the outgoing wave from T gives rise to that of the incoming wave in L. More importantly, these relations are here written without appealing to the decomposition of the waves p_T and p_L in their respective incoming p_T^- and p_L^- and outgoing parts p_T^+ and p_L^+. We now pass to boundary condition (5). Thanks to this interpretation, we can see that this statement can be translated from its mathematical expression as follows: "*the incoming trace results from the reflection of the outgoing trace and the trace of an incoming wave created by the source terms*". The terminology is thus particularly meaningful in this case and the expression of the boundary conditions by (5) specifically adequate.

Our objective in this work is to deal with more general Helmholtz equations

$$\boldsymbol{\nabla}\cdot A\boldsymbol{\nabla} p+\omega^2\chi p=0\text{ in }\Omega,\qquad(7)$$

in which A and χ are piecewise constant real functions such that $A(x)$ is a $d\times d$ symmetric definite positive matrix and $\chi(x)>0$ for almost all $x\in\Omega$. The isotropic, or scalar, case corresponds to $A=a$ with a a real-valued piecewise constant function positive almost everywhere. The anisotropic case corresponds to non scalar matrices A. If the standard Helmholtz equation was treated thoroughly in the literature (cf., e.g. [11, 8, 39, etc.]), in the opinion of the authors, the piecewise constant coefficients equation (7) was only partly addressed even in the isotropic case [29, 30, 34, 10]. In all these studies indeed, the principle, adopted for the boundary condition that the incoming trace partly consists of the reflection of the outgoing trace, is no more respected at an interface where the coefficients of the Helmholtz equation are discontinuous. One of the objectives of this work was to design a ultra-weak formulation that respects the reflection and transmission of plane waves propagating at normal incidence to an interface between two elements T and L with distinct or identical coefficients. Meanwhile, we found that it is possible to write a general formulation, covering this UWVF as well as all the previous ones, for which the matching conditions at an interface are stated from the principle that the incoming trace into an element T results from a reflection of the outgoing trace from T and the transmission of the outgoing trace from the adjacent element L. For the derivation of the UWVF itself, we found more clear to use a kind of *"reciprocity principle"* based on the fact that the outgoing-incoming trace operator is unitary (i.e., whose adjoint is the inverse operator) and an interpretation of the interface and boundary conditions as *"numerical fluxes"* instead of the usual

integration by parts approach [11, etc.]. We also give an alternative formulation to UWVF, with less advantageous mathematical properties, but which allows to isolate the incoming outgoing-incoming trace operator from the *"numerical fluxes operator"*. The separation of these operators greatly simplifies the computer programming of the method and the implementation of techniques improving the conditioning of the final linear system to be solved [6]. It also makes it easier to couple the plane-wave UWVF considered in this study with a polynomial UWVF which will be addressed in a future article. [papier-PUWVF]. We then extend the approach in [8] to set up a framework in which the stability properties underlying the approximation of all these UWVFs, at the level of each element using a plane-wave basis, are established. As a result, we thus prove that all of these UWVFs are among the very few methods leading to a linear system which can be solved by Gauss eliminations without pivoting. As in [8], this is done by establishing that the general UWVF is equivalent to a Trefftz-DG method. However, if for the usual Helmholtz equation this can be done from simple calculations, a more systematic handling is required in the present context. In summary, the plane wave approximation of both the newly introduced and all the previous UWVFs can thus be analyzed using the same framework. This enables us then to prove the convergence of these methods for realistic geometries and equations, at least in the 2D case. The convergence analysis also provides us with a theoretical basis for the choice of directions for the construction of plane-wave bases in the anisotropic case.

After introducing the UWVF compatible with the reflection and transmission of plane waves at normal incidence on an interface between two mesh elements, we present in section 2 a general framework covering all previous UWVFs. In section 3, we establish some coercivity properties for the general ultra-weak formulation. Towards this end, we first revisit the notion of a Trefftz Discontinuous Galerkin (Trefftz-DG) method. We then show that the UWVF of the previous general framework is equivalent to a Trefftz-DG and prove that the latter is the only one of these methods whose numerical fluxes are expressed in terms of the outgoing traces. We next show that the particular UWVF, compatible with the reflection and transmission of plane waves, is actually an upwind scheme in the meaning that the numerical fluxes of its equivalent Trefftz-DG method can be obtained by means of a Riemann solver. In section 4, we analyze the plane-wave approximation of the general framwork UWVF. In particular, we establish the existence-uniqueness of the discrete problem, its solvability with a direct solver based on Gauss eliminations without pivoting, and convergence properties of the related discretization. The section is completed by some numerical experiments, including in particular an assessment of the efficiency of the approach in the general case of equation (7) and an extension to this case of the techniques developed in our previous work [6] for improving the well-known bad conditioning of problems based on a palne-wave discretization. Section 5 is dedicated to some concluding remarks concerning some issues raised in the previous sections and to the discussion to some topics which are currently under study.

2 The general ultra-weak formulation

2.1 The boundary-value problem

We consider the following boundary-value problem, related to the interior PDE (7) and to mixed boundary conditions

$$\begin{cases} \boldsymbol{\nabla} \cdot A \boldsymbol{\nabla} p + \omega^2 \chi p = 0 \text{ in } \Omega, \\ p = g_D \text{ on } \partial\Omega_D, \\ A \boldsymbol{\nabla} p \cdot \boldsymbol{n} - i\omega Y^{\partial\Omega} p = g_N \text{ on } \partial\Omega_N, \end{cases} \tag{8}$$

with $Y^{\partial\Omega} \geq 0$ ds-almost everywhere on $\partial\Omega_N$.

The assumptions on the coefficients A, χ, and $Y^{\partial\Omega}$ are stated in a more precise way as follows. There exists a non-overlaping decomposition of Ω in polygonal/polyhedral subdomains $\Omega^{(\ell)}$, $\ell \in \mathcal{L}$, $\mathcal{L}$ being a finite set of indices, such that

- $A^{(\ell)} = A|_{\Omega^{(\ell)}}$ is a real constant symmetrical positive definite matrix,
- $\chi^{(\ell)} = \chi|_{\Omega^{(\ell)}}$ is a real positive constant,
- $Y^{(\ell)} = Y^{\partial\Omega}|_{\partial\Omega^{(\ell)} \cap \partial\Omega}$ is a real non-negative constant whenever the ds-measure of $\partial\Omega^{(\ell)} \cap \partial\Omega_N$ is nonzero.

We also assume that the following compatibility condition: if the ds-measure of $\partial\Omega \cap \partial\Omega^{(\ell)}$ is nonzero, to a set of ds-measure zero, either $\partial\Omega_N \cap \partial\Omega^{(\ell)} = \varnothing$ or $\partial\Omega_D \cap \partial\Omega^{(\ell)} = \varnothing$, in other words each common part of $\partial\Omega$ and $\partial\Omega^{(\ell)}$ supports either a Dirichlet or an impedance boundary condition.

As above, let be given a mesh $\mathcal{T}$ of Ω, which is now compatible with the decomposition $\left\{\Omega^{(\ell)}\right\}_{\ell \in \mathcal{L}}$ in the meaning that every $T \in \mathcal{T}$ is contained in a subdomain $\Omega^{(\ell)}$. As a result, we assume that $A_T = A|_T$, $\chi_T = \chi|_T$, and $Y_T^{\partial\Omega} = Y^{\partial\Omega}|_{\partial T \cap \partial\Omega}$ (when the ds-measure of $\partial T \cap \partial\Omega > 0$) are all constant.

In order to solve this problem by an UWVF, the boundary conditions are expressed in the following form

$$-\frac{1}{\eta}\left(A \boldsymbol{\nabla} p \cdot \boldsymbol{n} - i\omega\sigma p\right) = Q \frac{1}{\eta}\left(A \boldsymbol{\nabla} p \cdot \boldsymbol{n} + i\omega\sigma p\right) + g$$

with $Q = 1$ and $g = (2i\omega\sigma/\eta)\, g_D$ for a Dirichlet condition on $\partial\Omega_D$, and $Q = \left(1 - Y^{\partial\Omega}/\sigma\right) / \left(1 + Y^{\partial\Omega}/\sigma\right)$ and $g = -(1+Q)\, g_N/\eta$ for a Neumann or a Fourier-Robin condition on $\partial\Omega_N$.

It can be seen that all the boundary-value problems considered in the following references [34, 39, 29] fall within this framework, as well as those dealt with in [11, 10, 30] after substituting $-i$ for i to switch to the above convention (2) on the phasors.

2.2 Plane waves

In the approach of this study, plane waves are fundamental in two respects. They provide local solutions bases used as trial and test functions for this Trefftz-DG method. They also enter as an essential tool in the design of matching conditions at the interface of two elements of the mesh $\mathcal{T}$ that are compatible with the reflection and transmission of plane waves impinging at normal incidence on this interface.

Plane waves can be defined for constant coefficients only. We therefore look for solutions inside element $T \in \mathcal{T}$ for which coefficients A and χ of Eq. (7) are equal to A_T and χ_T respectively. There are two ways for defining plane waves propagating in the direction of the unit vector $\boldsymbol{\nu}$ of $\mathbb{R}^d$. The first one is based on the reduction of Eq. (7) to a first-order system [16] and yields as solutions

$$p^{\pm}(x) = \alpha^{\pm} \exp\left(\pm i \kappa_{T,\boldsymbol{\nu}} \boldsymbol{\nu} \cdot x\right) \tag{9}$$

where

$$\kappa_{T,\boldsymbol{\nu}} = \omega \sqrt{\frac{\chi_T}{\boldsymbol{\nu} \cdot A_T \boldsymbol{\nu}}}$$

can be called the *directional wavenumber* in the direction of vector $\boldsymbol{\nu}$, and $\alpha^{\pm}$ play the role of the complex amplitude of the related plane wave. One can indeed easily verify that functions (9) are solutions to Eq. (7) in T from the following relations

$$\begin{aligned} A_T \boldsymbol{\nabla} p^{\pm} &= \pm i \kappa_{T,\boldsymbol{\nu}} p^{\pm} A_T \boldsymbol{\nu} \\ \boldsymbol{\nabla} \cdot A_T \boldsymbol{\nabla} p^{\pm} = \pm i \kappa_{T,\boldsymbol{\nu}} A_T \boldsymbol{\nu} \cdot \boldsymbol{\nabla} p^{\pm} &= -\kappa_{T,\boldsymbol{\nu}}^2 \boldsymbol{\nu} \cdot A_T \boldsymbol{\nu} p^{\pm} = -\omega^2 \chi_T p^{\pm}. \end{aligned} \tag{10}$$

It is worth mentioning that for an isotropic medium of propagation, that is, for $A_T = a_T$, the directional wavenumber reduces to the usual wavenumber

$$\kappa_{T,\boldsymbol{\nu}} = \kappa_T = \omega \sqrt{\frac{\chi_T}{a_T}}.$$

A second way for defining plane waves uses the square root $A_T^{1/2}$ of A_T which can be defined from the eigenvector decomposition $A_T = U_T^{\top} D_T U_T$, where U_T is a real unitary matrix, D_T is a diagonal matrix with positive coefficients on its diagonal [10]. This other way of defining plane waves is in fact equivalent to the first one, as we will see later. To distinguish them, we call the former directional plane waves and the latter Cessenat's plane waves.

2.3 Plane-wave reflection and transmission at an interface

As above mentioned, the UWVF, which is developed here, is compatible with the reflection and transmission of plane waves propagating at normal incidence to an interface at which coefficients A and χ have a jump. To begin with, we introduce the notion of what can be called here a directional admittance $Y_{T,\boldsymbol{\nu}}$ in T along the direction of $\boldsymbol{\nu}$. It is defined through the following relation

$$A_T \boldsymbol{\nabla} p^{\pm} \cdot \boldsymbol{\nu} = \pm i \omega \sqrt{\frac{\chi_T}{\boldsymbol{\nu} \cdot A_T \boldsymbol{\nu}}} \boldsymbol{\nu} \cdot A_T \boldsymbol{\nu}\ p^{\pm} = \pm i \omega Y_{T,\boldsymbol{\nu}} p^{\pm} \tag{11}$$

with

$$Y_{T,\boldsymbol{\nu}} = \sqrt{\chi_T \boldsymbol{\nu} \cdot A_T \boldsymbol{\nu}} \tag{12}$$

and $p^{\pm}$ given by (9). The directional admittance will be involved here through the following relations

$$\pm \frac{1}{2i\omega Y_{T,\boldsymbol{\nu}}} \left(A_T \boldsymbol{\nabla} p \cdot \boldsymbol{\nu} \pm i\omega Y_{T,\boldsymbol{\nu}} p \right) = p^{\pm} \tag{13}$$

where $p = p^+ + p^-$ is the superposition of the two plane waves p^+ and p^-. These relations generalize those for an isotropic medium. They therefore achieve a filtering of the superposition $p = p^+ + p^-$of a system of two plane waves traveling in the direction of vector $+\boldsymbol{\nu}$ and its opposite $-\boldsymbol{\nu}$ respectively.

Then let an interface $F_{\{T,L\}}$ of two elements T and L of $\mathcal{T}$. We denote by $\boldsymbol{n}_T$ and $\boldsymbol{n}_L$ the unit vectors on $F_{\{T,L\}}$ respectively directed towards the exterior of T and L. Let also two superpositions of plane waves

$$p_T = p_T^+ + p_T^- \text{ and } p_L = p_L^+ + p_L^- \tag{14}$$

be propagating respectively along the directions of $\pm\boldsymbol{n}_T$ and $\pm\boldsymbol{n}_L$

$$p_T^{\pm} = \alpha_T^{\pm} \exp\left(\pm i\kappa_{\boldsymbol{n}_T} \boldsymbol{n}_T \cdot (x - x_0)\right), \quad p_L^{\pm} = \alpha_L^{\pm} \exp\left(\pm i\kappa_{\boldsymbol{n}_L} \boldsymbol{n}_L \cdot (x - x_0)\right)$$

with $x_0 \in F_{\{T,L\}}$. For simplicity, we denote by $\kappa_{\boldsymbol{n}_T} = \kappa_{T,\boldsymbol{n}_T}$, and

$$Y_T = Y_{T,\boldsymbol{n}_T}. \tag{15}$$

A similar notation is used for $\kappa_{\boldsymbol{n}_L}$ and Y_L assuming that the coefficients of Eq. (7) are respectively A_L and χ_L in L. The usual matching conditions on p_T and p_L, to yield a solution to Eq. (7) in the interior of $\overline{T \cup L}$, are

$$p_T = p_L, \quad A_T \boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T + A_L \boldsymbol{\nabla} p_L \cdot \boldsymbol{n}_L = 0 \text{ on } F_{\{T,L\}}. \tag{16}$$

The following proposition translates these conditions in terms of reflection and transmission of outgoing waves p_T^+ and p_L^+.

Proposition 1 *Above superposition (14) of plane waves in respectively T and L satisfies the usual matching conditions (16) if and only if the following relations hold true*

$$\begin{cases} p_T^- = \frac{Y_T - Y_L}{Y_T + Y_L} p_T^+ + \frac{2Y_L}{Y_T + Y_L} p_L^+, \\ p_L^- = \frac{Y_L - Y_T}{Y_L + Y_T} p_L^+ + \frac{2Y_T}{Y_L + Y_T} p_T^+, \end{cases} \quad \text{on } F_{\{T,L\}}. \tag{17}$$

Proof. Since $(x - x_0) \cdot \boldsymbol{n}_T = -(x - x_0) \cdot \boldsymbol{n}_L = 0$ for all $x \in F_{\{T,L\}}$, using (11) we can write (16) in the form

$$\begin{cases} \alpha_T^+ + \alpha_T^- - \alpha_L^+ - \alpha_L^- = 0, \\ Y_T \alpha_T^+ - Y_T \alpha_T^- + Y_L \alpha_L^+ - Y_L \alpha_L^- = 0. \end{cases}$$

Solving this system in α_T^- and α_L^-, we get (17). ■

An important step towards the derivation of the UWVF, which is developed here, is the expression of (17) without resorting to the decomposition of the superposition of plane waves p_T and p_L in their incoming and outgoing parts $p_T^{\pm}$ and $p_L^{\pm}$ by means of the following operators,

$$\Lambda^{\pm}_{T,Y_T} p_T = \pm \frac{1}{2i\omega Y_T} \left(A_T \boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T \pm i\omega Y_T p_T \right) \tag{18}$$

and the same expressions for L. In view of (13), matching conditions (17) can be expressed as

$$\left\{ \begin{array}{l} \Lambda^{-}_{T,Y_T} p_T = \frac{Y_T - Y_L}{Y_T + Y_L} \Lambda^{+}_{T,Y_T} p_T + \frac{2Y_L}{Y_T + Y_L} \Lambda^{+}_{L,Y_L} p_L, \\ \Lambda^{-}_{L,Y_L} p_L = \frac{Y_L - Y_T}{Y_L + Y_T} \Lambda^{+}_{L,Y_L} p_L + \frac{2Y_T}{Y_L + Y_T} \Lambda^{+}_{T,Y_T} p_T, \end{array} \right. \quad \text{on } F_{\{T,L\}}. \tag{19}$$

Actually as this is established below, these matching conditions are equivalent to the usual ones (16) and give rise to the UWVF, which is developed here.

We now derive a general UWVF covering the one developed here, as well as all those considered previously. The main tool is a generalization of operators (18).

2.4 The generalized outgoing and incoming traces

The generalized outgoing and incoming traces mimic those given in (18) by considering for each $T \in \mathcal{T}$ a function η_T on ∂T, constant and positive on each side/face of ∂T, which will play the role of a "*fictitious admittance*". Below, we refer to the admittances defined in (15) as "*actual admittances*" to distinguish them from the fictitious ones. Then, similarly to the above traces defined in terms of the actual admittances, the outgoing and incoming traces associated with the fictitious admittances are defined as follows

$$\Lambda^{\pm}_{T,\eta_T} p_T = \pm \frac{1}{2i\omega\eta_T} \left(A_T \boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T \pm i\omega\eta_T p_T \right). \tag{20}$$

From now on, except explicitly stated otherwise, generic function p_T (as well as p_L when it is involved) is a sufficiently smooth function, defined on T, on L for p_L, so that the above outgoing and incoming traces make sense.

A simple but important feature of these traces lies on the fact that they are equivalent to the usual Cauchy traces for Eq. (7) from the following relations

$$p_T = \Lambda^{+}_{T,\eta_T} p_T + \Lambda^{-}_{T,\eta_T} p_T, \quad \frac{1}{i\omega\eta_T} A_T \boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T = \Lambda^{+}_{T,\eta_T} p_T - \Lambda^{-}_{T,\eta_T} p_T. \tag{21}$$

The following proposition provides the first ingredient in the derivation of the general UWVF.

Proposition 2 *Usual matching conditions (16) are equivalent to the following ones*

$$\left\{ \begin{array}{l} \Lambda^{-}_{T,\eta_T} p_T = \frac{\eta_T - \eta_L}{\eta_T + \eta_L} \Lambda^{+}_{T,\eta_T} p_T + \frac{2\eta_L}{\eta_T + \eta_L} \Lambda^{+}_{L,\eta_L} p_L, \\ \Lambda^{-}_{L,\eta_L} p_L = \frac{\eta_L - \eta_T}{\eta_L + \eta_T} \Lambda^{+}_{L,\eta_L} p_L + \frac{2\eta_T}{\eta_L + \eta_T} \Lambda^{+}_{T,\eta_T} p_T. \end{array} \right. \quad \textit{on } F_{\{T,L\}}, \tag{22}$$

In the same way, the boundary conditions in problem (8) can be equivalently expressed by

$$\Lambda^{-}_{T,\eta_T} p_T = Q_T \Lambda^{+}_{T,\eta_T} p_T + g_T \ \text{on}\ F_T^{\partial\Omega} \tag{23}$$

with

$$\begin{cases} Q_T = -1,\ g_T = g_D \ \text{on}\ \partial\Omega_D \cap \partial T, \\ Q_T = \frac{\eta_T - Y^{\partial\Omega}}{\eta_T + Y^{\partial\Omega}},\ g_T = -\frac{1}{i\omega(\eta_T + Y^{\partial\Omega})} g_N \ \text{on}\ \partial\Omega_N \cap \partial T. \end{cases}$$

Proof. First condition of (22) can also be written

$$\frac{1}{i\omega}\frac{1}{\eta_T + \eta_L}\left(A_T \boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T + A_L \boldsymbol{\nabla} p_L \cdot \boldsymbol{n}_L\right) = \frac{\eta_L}{\eta_T + \eta_L}\left(p_T - p_L\right).$$

In the same way, we can put the second condition in the form

$$\frac{1}{i\omega}\frac{1}{\eta_T + \eta_L}\left(A_T \boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T + A_L \boldsymbol{\nabla} p_L \cdot \boldsymbol{n}_L\right) = \frac{\eta_T}{\eta_T + \eta_L}\left(p_L - p_T\right).$$

It is then easily seen that (22) and (16) are equivalent. The rest of the proof is next directly obtained from (21). ■

We now come to the second ingredient in the derivation of the general UWVF.

2.5 The unitary operator

This is the fundamental tool in the design of an UWVF (see, e.g., [11]). We first establish that any $x_T \in L^2_{\eta_T}(\partial T)$, the space of L^2 functions on ∂T with weight η_T, can be considered as the outgoing trace $\Lambda^{+}_{\eta_T,T} p_T$ of a well-defined function p_T in

$$W_T = \left\{ q_T \in H^1(T);\ \boldsymbol{\nabla} \cdot A_T \boldsymbol{\nabla} q_T + \omega^2 \chi_T q_T = 0 \text{ in } T,\ \Lambda^{+}_{\eta_T,T} q_T \in L^2_{\eta_T}(\partial T) \right\} \tag{24}$$

This function can be retrieved by solving the following boundary-value problem set in T

$$\begin{cases} p_T \in H^1(T), \\ \boldsymbol{\nabla} \cdot A_T \boldsymbol{\nabla} p_T + \omega^2 \chi_T p_T = 0 \text{ in } T, \\ A_T \boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T + i\omega\eta_T p_T = 2i\omega\eta_T x_T. \end{cases}$$

The existence and uniqueness of this problem can be obtained in an elementary way from its variational formulation and the Fredholm alternative

$$\begin{cases} p_T \in H^1(T),\ \forall q_T \in H^1(T), \\ a_T(p_T, q_T) + i\omega \displaystyle\int_{\partial T} p_T q_T \eta_T ds = 2i\omega \int_{\partial T} x_T q_T \eta_T ds, \end{cases} \tag{25}$$

where

$$a_T(p_T, q_T) = \int_T A_T \boldsymbol{\nabla} p_T \cdot \boldsymbol{\nabla} q_T - \omega^2 \int_T p_T q_T \chi_T dx. \tag{26}$$

Setting $\mathcal{X}_T x_T = p_T$, we thus define a bounded operator $\mathcal{X}_T$ from $L^2_{\eta_T}(\partial T)$ into W_T equipped with the graph norm

$$\|q_T\|_{W_T} = \sqrt{\|q_T\|^2_{H^1(T)} + \left\|\Lambda^+_{\eta_T,T} q_T\right\|^2_{L^2_{\eta_T}(\partial T)}}.$$

Let us define $\mathcal{U}_T$ from $L^2_{\eta_T}(\partial T)$ into $L^2_{\eta_T}(\partial T)$ by

$$\mathcal{U}_T x_T = \Lambda^-_{\eta_T,T} \mathcal{X}_T x_T.$$

Actually, $\mathcal{U}_T x_T$ can be defined directly from (21) without resorting to $\Lambda^-_{\eta_T,T}$, thus showing that it is a bounded operator from $L^2_{\eta_T}(\partial T)$ into $L^2_{\eta_T}(\partial T)$. Further properties of this operator are established in the following proposition.

Proposition 3 *Operator $\mathcal{U}_T$ is a symmetric operator in $L^2_{\eta_T}(\partial T)$, i.e.,*

$$(\mathcal{U}_T x_T, y_T)_{L^2_{\eta_T}(\partial T)} = (x_T, \mathcal{U}_T y_T)_{L^2_{\eta_T}(\partial T)}, \quad \forall x_T, y_T \text{ in } L^2_{\eta_T}(\partial T),$$

where

$$(x_T, y_T)_{L^2_{\eta_T}(\partial T)} = \int_{\partial T} x_T y_T \eta_T ds$$

is the bilinear form underlying the scalar-product $(x_T, \overline{y_T})_{L^2_{\eta_T}(\partial T)}$ of $L^2_{\eta_T}(\partial T)$. It is also a unitary operator in $L^2_{\eta_T}(\partial T)$ in the meaning

$$\mathcal{U}_T^* \mathcal{U}_T = I_{L^2_{\eta_T}(\partial T)}$$

where $I_{L^2_{\eta_T}(\partial T)}$ is the identity operator in $L^2_{\eta_T}(\partial T)$.

Proof. From the variational definition of $\mathcal{X}_T x_T$ and $\mathcal{X}_T y_T$, we can write

$$a_T(\mathcal{X}_T x_T, \mathcal{X}_T y_T) + i\omega (\mathcal{X}_T x_T, \mathcal{X}_T y_T)_{L^2_{\eta_T}(\partial T)} = 2i\omega (x_T, \mathcal{X}_T y_T)_{L^2_{\eta_T}(\partial T)}$$
$$a_T(\mathcal{X}_T y_T, \mathcal{X}_T x_T) + i\omega (\mathcal{X}_T y_T, \mathcal{X}_T x_T)_{L^2_{\eta_T}(\partial T)} = 2i\omega (y_T, \mathcal{X}_T x_T)_{L^2_{\eta_T}(\partial T)}$$

and directly get that

$$(x_T, \mathcal{X}_T y_T)_{L^2_{\eta_T}(\partial T)} = (\mathcal{X}_T x_T, y_T)_{L^2_{\eta_T}(\partial T)}.$$

Substracting $(x_T, y_T)_{L^2_{\eta_T}(\partial T)}$ from the two sides of this equality, and making use of (21) directly yields that $\mathcal{U}_T$ is symmetric.

Now, knowing that $\overline{q_T}$ is a valid test function, by conjugating the related equation we obtain

$$a\left(\overline{\mathcal{X}_T x_T}, q_T\right) - i\omega \left(\overline{\mathcal{X} x_T}, q_T\right)_{L^2_{\eta_T}(\partial T)} = -2i\omega (\overline{x_T}, q_T)_{L^2_{\eta_T}(\partial T)}.$$

Relation (21) first gives

$$a\left(\overline{\mathcal{X}_T x_T}, q_T\right) + i\omega \left(\overline{\mathcal{X}_T x_T}, q_T\right)_{L^2_{\eta_T}(\partial T)} = 2i\omega \left(\overline{\mathcal{U}_T x_T}, q_T\right)_{L^2_{\eta_T}(\partial T)},$$

and then

$$\mathcal{U}_T\overline{\mathcal{U}_T x_T} = \overline{\mathcal{X}_T x_T} - \overline{\mathcal{U}_T x_T} = \overline{\mathcal{X}_T x_T - \mathcal{U}_T x_T} = \overline{x_T}. \tag{27}$$

The symmetry of $\mathcal{U}_T$ yields

$$\overline{\left(\mathcal{U}_T\overline{\mathcal{U}_T x_T}, y_T\right)_{L^2_{\eta_T}(\partial T)}} = \overline{\left(\overline{\mathcal{U}_T x_T}, \mathcal{U}_T y_T\right)_{L^2_{\eta_T}(\partial T)}} = \left(\mathcal{U}_T x_T, \overline{\mathcal{U}_T y_T}\right)_{L^2_{\eta_T}(\partial T)}.$$

This completes the proof since then

$$\left(\mathcal{U}_T x_T, \overline{\mathcal{U}_T y_T}\right)_{L^2_{\eta_T}(\partial T)} = \left(x_T, \overline{y_T}\right)_{L^2_{\eta_T}(\partial T)} = \left(\mathcal{U}_T^*\mathcal{U}_T x_T, \overline{y_T}\right)_{L^2_{\eta_T}(\partial T)}.$$

■

We have now completed the necessary set-up for the derivation of the general UWVF.

2.6 The general UWVF

Usually the derivation of the UWVF is carried out "*on-the-fly*" by integration by parts [11, 29] and using matching conditions (22) and boundary conditions (23). Basically, the same approach is followed here. However, the presentation we make of it is, in our opinion, simpler to understand. The unitary property of operator $\mathcal{U}_T$ is first operated variationally, yielding a kind of *"reciprocity principle"* relatively to the outgoing and incoming traces

$$\left(x_T, \overline{y_T}\right)_{L^2_{\eta_T}(\partial T)} - \left(\mathcal{U}_T x_T, \overline{\mathcal{U}_T y_T}\right)_{L^2_{\eta_T}(\partial T)} = 0.$$

The matching conditions (22) and the boundary conditions (23) are then interpreted as numerical fluxes and plugged into the *"reciprocity relation"* yielding

$$\left(x_T, \overline{y_T}\right)_{L^2_{\eta_T}(\partial T)} - \left(\mathcal{F}_T x, \overline{\mathcal{U}_T y_T}\right)_{L^2_{\eta_T}(\partial T)} = \left(g_T, \overline{\mathcal{U}_T y_T}\right)_{L^2_{\eta_T}(\partial T)}$$

where x is in the product space

$$X_{\mathcal{T}} = \prod_{T\in\mathcal{T}} L^2_{\eta_T}(\partial T)$$

and $\mathcal{F}_T x$ is variationally defined by

$$\left(\mathcal{F}_T x, y_T\right)_{L^2_{\eta_T}(\partial T)} = \sum_{F_{\{T,L\}}\in\mathcal{A}_T} \int_{F_{\{T,L\}}} \left(\frac{\eta_T - \eta_L}{\eta_T + \eta_L} x_T + \frac{2\eta_L}{\eta_T + \eta_L} x_L\right) y_T \eta_T ds + \sum_{F_T^{\partial\Omega}\in\mathcal{B}_T} \int_{F_T^{\partial\Omega}} Q_T x_T y_T \eta_T ds$$

$\mathcal{A}_T$ being the set of sides/faces of ∂T shared by another element L adjacent to T, $\mathcal{B}_T$ being the set of sides/faces of ∂T contained in ∂T.

The variational space is equipped with its canonical bilinear form

$$(x,y) \to (x,y)_{X_\mathcal{T}} = \sum_{T\in\mathcal{T}} (x_T, y_T)_{L^2_{\eta_T}(\partial T)}$$

underlying its scalar product $\sum_{T\in\mathcal{T}} (x_T, \overline{y_T})_{L^2_{\eta_T}(\partial T)}$, x_T and y_T being the respective components of x and y. We denote by $\|x\|_{X_\mathcal{T}}$ the related norm of $x \in X_\mathcal{T}$. Defining operators $\mathcal{U}$ and $\mathcal{F}$ by their respective components

$$(\mathcal{U}x)_T = \mathcal{U}_T x_T, \quad (\mathcal{F}x)_T = \mathcal{F}_T x,$$

we then come to the general UVWF

$$\left\{ \begin{array}{l} x \in X_\mathcal{T},\ \forall y \in X_\mathcal{T}, \\ (x - \mathcal{U}^*\mathcal{F}x, \overline{y})_{X_\mathcal{T}} = \left(g, \overline{\mathcal{U}y}\right)_{X_\mathcal{T}}, \end{array} \right. \tag{28}$$

where $\mathcal{U}^*$ is the adjoint operator to $\mathcal{U}$. The component g_T of $g \in X_\mathcal{T}$ is the g_T involved in the boundary condition (23) for $F_T^{\partial\Omega} \in \mathcal{B}_T$ and is 0 on $F_{\{T,L\}} \in \mathcal{A}_T$.

By setting $y = \overline{\mathcal{U}z}$ and noting that $\mathcal{U}$ is symmetric and unitary, we readily get an equivalent formulation to the above general UWVF

$$\left\{ \begin{array}{l} x \in X_\mathcal{T},\ \forall z \in X_\mathcal{T}, \\ (\mathcal{U}x - \mathcal{F}x, z)_{X_\mathcal{T}} = (g, z)_{X_\mathcal{T}}. \end{array} \right. \tag{29}$$

Formulation (29) can be called Direct-UWVF because it can be also obtained by simply expressing the matching conditions (22) and the boundary conditions (23) variationally. Even though it has less advantageous mathematical properties than the general UWVF, it owns the interesting property of isolating the outgoing-incoming trace operator $\mathcal{U}$ from what can be called the *"scattering operator"* $\mathcal{F}$. This feature plays an important role for improving the conditioning of the linear systems resulting from the Galerkin plane-wave approximation of the formulation [B2DT-JCP].

The ultra-weak formulation (28) keeps a fundamental property of the usual UWVFs, as stated in the following proposition.

Proposition 4 *Operator $\mathcal{F}$ satisfies*

$$\|\mathcal{F}x\|_{X_\mathcal{T}} \leq \|x\|_{X_\mathcal{T}}, \quad \forall x \in X_\mathcal{T}. \tag{30}$$

As a result, problem (28) is a fixed point problem in $X_\mathcal{T}$ in the sense that

$$x - \mathcal{U}^*\mathcal{F} = \mathcal{U}^* g$$

with $\|\mathcal{U}^*\mathcal{F}x\|_{X_\mathcal{T}} \leq \|x\|_{X_\mathcal{T}}$, $\forall x \in X_\mathcal{T}$.

Proof. Using the definition of $\mathcal{F}$, we can write

$$\|\mathcal{F}x\|^2_{X_\mathcal{T}} = \sum_{T\in\mathcal{T}} \left(\begin{array}{c} \displaystyle\sum_{F_{\{T,L\}}\in\mathcal{A}_T} \int_{F_{\{T,L\}}} \left| \tfrac{\eta_T-\eta_L}{\eta_T+\eta_L} x_T + \tfrac{2\eta_L}{\eta_T+\eta_L} x_L \right|^2 \eta_T ds \\ \displaystyle + \sum_{F_T^{\partial\Omega}\in\mathcal{B}_T} \int_{F_T^{\partial\Omega}} |Q_T x_T|^2 \eta_T ds \end{array} \right).$$

Passing from the sum on the triangles to the sum on the sides/faces $F_{\{T,L\}}$ and $F_T^{\partial\Omega}$, we get

$$\|\mathcal{F}x\|^2_{X_{\mathcal{T}}} = \sum_{F_{\{T,L\}}\in\mathcal{A}} \int_{F_{\{T,L\}}} \left(\begin{array}{c} \left|\frac{\eta_T-\eta_L}{\eta_T+\eta_L}x_T + \frac{2\eta_L}{\eta_T+\eta_L}x_L\right|^2 \eta_T \\ + \left|\frac{\eta_L-\eta_T}{\eta_T+\eta_L}x_L + \frac{2\eta_T}{\eta_T+\eta_L}x_T\right|^2 \eta_L \end{array} \right) ds \\ + \sum_{F_T^{\partial\Omega}\in\mathcal{B}} \int_{F_T^{\partial\Omega}} |Q_T x_T|^2 \eta_T ds,$$

where $\mathcal{A}$ is the set of all internal edges/faces $F_{\{T,L\}}$ of $\mathcal{T}$ and $\mathcal{B}$ is the set of all boundary edges/faces $F_T^{\partial\Omega}$ of $\mathcal{T}$. The proof is completed by noting that

$$\left|\frac{\eta_T-\eta_L}{\eta_T+\eta_L}x_T + \frac{2\eta_L}{\eta_T+\eta_L}x_L\right|^2 \eta_T + \\ \left|\frac{\eta_L-\eta_T}{\eta_T+\eta_L}x_L + \frac{2\eta_T}{\eta_T+\eta_L}x_T\right|^2 \eta_L = |x_T|^2 \eta_T + |x_L|^2 \eta_L$$

and $|Q_T| \leq 1$. ■

The fundamental property of the above proposition distinguishes the plane-wave UWVF from a standard plane-wave DG method. It will be at the basis of the coerciveness properties of the general UVWF.

We now see how the general UWVF covers all those previously considered as well as the one developed here.

2.7 Two classes of ultra-weak variational formulations

Recall that each UWVF is characterized by a positive function η_T defined on the boundary ∂T of each $T \in \mathcal{T}$, and being constant on each side/face of ∂T. The function η, associated with the collection $\{\eta_T\}_{T\in\mathcal{T}}$, is a priori double-valued on the interior interfaces $F_{\{T,L\}}$. We are naturally led to distinguish two classes of UWVFs.

- **Single-valued admittance function η.** Explicitly, this means that

  $$\eta_T = \eta_L \text{ on the interfaces } F_{\{T,L\}}.$$

  In this way, the reflection and the transmission coefficients are nothing else but

  $$\frac{\eta_T-\eta_L}{\eta_T+\eta_L} = 0, \quad \frac{2\eta_L}{\eta_T+\eta_L} = 1,$$

  and, in particular, express that the matching conditions correspond to writing that the incoming trace from an element T at an interface $F_{\{T,L\}}$ is the outgoing trace from the adjacent element L through this interface. To the authors' knowledge, all previous UWVFs can be retrieved in this framework by properly choosing the single-valued function η and with

incoming and outgoing traces that differ from the current ones by a multiplicative constant (cf., e.g., [11, 10, 30, 34, 39, 29]). For example, for the case of the acoustics system, with piecewise constant density ϱ of the fluid at rest and wavenumber κ, the ultra-weak formulation considered by Kaipio, Huttunen, and Monk in [29], is obtained with the following substitutions $A \to 1/\varrho$, $\omega \to 1$, $\chi \to \kappa^2/\varrho$. To be in the framework of the present study, we have also to limit here the wavenumbers κ to be real. Function η for the UWVF considered by these authors is single-valued and equal to

$$\eta_T = \frac{1}{2}\left(\frac{\kappa_T}{\varrho_T} + \frac{\kappa_L}{\varrho_L}\right)$$

which is nothing else than the mean value of the actual admittances $Y_T = \kappa_T/\varrho_T$ and $Y_L = \kappa_L/\varrho_L$ on both sides of $F_{\{T,L\}}$. The outgoing and incoming traces used in this reference, respectively x_T^{HKM} and $\mathcal{U}_T^{\mathrm{HKM}} x_T^{\mathrm{HKM}}$, are relatedthose in this paper as follows

$$x_T^{\mathrm{HKM}} = -2i\eta_T x_T, \quad \mathcal{U}_T^{\mathrm{HKM}} x_T^{\mathrm{HKM}} = -2i\eta_T \mathcal{U}_T x_T.$$

As a result, we simply have $\mathcal{U}_T^{\mathrm{HKM}} = \mathcal{U}_T$. Such a UWVF can be seen somehow as a "*centered scheme*" as opposed to the "*upwind schemes*" that are introduced later.

- **Double-valued admittance function** η. Presently, the only representative case of this class of UWVFs consists in choosing $\eta_T = Y_T$, the actual admittances. For the acoustics system considered in [29], this leads to take $\eta_T = Y_T = \kappa_T/\varrho_T$. This amounts to consider as a "*upwind scheme*" the UWVF, which was developed above.

Since operators $\mathcal{U}$ and $\mathcal{F}$ are of norm ≤ 1, the operator in which is posed the general UWVF (28) satisfies the following bound

$$\|x - \mathcal{U}^*\mathcal{F}x\|_{X_\mathcal{T}} \leq 2\,\|x\|_{X_\mathcal{T}}\,, \quad \forall x \in X_\mathcal{T}. \tag{31}$$

Unfortunately, the formulation does not seem to be coercive or even to satisfy an inf-sup condition of Brezzi-Babuška-Ladyzenskya type (cf., e.g., [7]) in the norm of $X_\mathcal{T}$. Here, we follow the procedure devised by Buffa and Monk [8] for the standard Helmholtz equation (1). The idea consists in establishing that the UWVF is coercive in a weaker norm, that of a Trefftz-DG method to which it is equivalent. Of course, this coercivity is not enough to prove that the plane-wave Galerkin approximation of the variational system (28) is stable but, as seen below, the formulation has further properties that implies this stability.

3 Coercivity properties

3.1 Another construction of Trefftz-DG methods

Usually, a Trefftz-DG method for solving (8) is designed starting from a DG method and then assuming that the test and trial functions are solution respectively to the internal EDP and to the formal adjoint of this EDP (in fact the

same EDP here since it is formally self-adjoint) (cf., e.g., [27, 21, 25, etc.]). We see below that we can introduce Trefftz-DG type formulations in a more rapid and direct way by using the classical reciprocity relation stated in the following proposition.

Proposition 5 *Let T be in $\mathcal{T}$ and p_T and q_T in W_T, defined in (24) above. Then, p_T and q_T satisfy the following reciprocity relation*

$$\int_{\partial T} (q_T A_T \boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T - p_T A_T \boldsymbol{\nabla} q_T \cdot \boldsymbol{n}_T)\, ds = 0. \tag{32}$$

Proof. Green's formula yields

$$\int_{\partial T} (q_T A_T \boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T - p_T A_T \boldsymbol{\nabla} q_T \cdot \boldsymbol{n}_T)\, ds = \int_T (\boldsymbol{\nabla} \cdot (q_T A_T \boldsymbol{\nabla} p_T) - \boldsymbol{\nabla} \cdot (p_T A_T \boldsymbol{\nabla} q_T))\, ds.$$

The proof is completed by expanding the two divergence terms. ■

We then introduce the following Trefftz-space

$$W_{\mathcal{T}} = \left\{ p \in L^2(\Omega)\,;\; p|_T \in W_T,\ \forall T \in \mathcal{T} \right\}$$

which can be identified to the Hilbert product-space $W_{\mathcal{T}} = \prod_{T\in\mathcal{T}} W_T$. It is immediate that the following map $x \in X_{\mathcal{T}} \to p = \mathcal{X} x \in W_{\mathcal{T}}$ with $p_T = \mathcal{X}_T x_T$, $\forall T \in \mathcal{T}$, is bijective and bicontinuous. In particular, this application constructs a solution to the boundary-value problem (8) by solving either problem (28) or problem (29).

A Trefftz-DG method is obtained by choosing a system of *numerical fluxes* $\widehat{\boldsymbol{v}} \in \mathbb{C}^d$ and $\widehat{p} \in \mathbb{C}$, defined on the skeleton of the mesh $\mathcal{T}$, composed of the internal edges/faces $F_{\{T,L\}} \in \mathcal{A}$ and the boundary edges/faces $F_T^{\partial\Omega} \in \mathcal{B}$. For the UWVFs, we can restrict the numerical fluxes to be expressed as linear combinations of the Cauchy traces p_T, p_L, $A_T \boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T$, $A_L \boldsymbol{\nabla} p_L \cdot \boldsymbol{n}_L$ on internal edges/faces $F_{\{T,L\}} \in \mathcal{A}_T$ and in the same way as linear combinations of p_T, $A_T \boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T$, g_T on boundary edges/faces $F_T^{\partial\Omega} \in \mathcal{B}_T$. Adopting the point of view in [3], we can consider these expressions as a linear bounded operator

$$\begin{cases} W_{\mathcal{T}} \times L^2(\partial\Omega) \to L^2_{\mathrm{N}}\left(\Gamma; \mathbb{C}^d\right) \times L^2(\Gamma), \\ (p, g) \to (\widehat{\boldsymbol{v}}, \widehat{p}), \end{cases}$$

where Γ is the union $\left(\cup_{F_{\{T,L\}}\in\mathcal{A}} F_{\{T,L\}}\right) \cup \left(\cup_{F_T^{\partial\Omega}\in\mathcal{B}} F_T^{\partial\Omega}\right)$ (in the meaning of union of subsets of point of $\mathbb{R}^d$) of the internal as well as the external edges/faces of decomposition $\mathcal{T}$, and $L^2_{\mathrm{N}}\left(\Gamma; \mathbb{C}^d\right)$ is the subspace of the vector fields on Γ with d complex components ds-almost everywhere normal to Γ.

Remark 6 *We prefer to include here the conservativity conditions on the numerical fluxes of [3] in the definition by assuming that the numerical fluxes are*

single-valued functions on Γ. *We depart also from [3] by imposing to the vectorial numerical fluxes to be normal to* Γ. *In this reference, this restriction is not enforced even if it is explicitly stated that it is only the normal components of the fluxes that are involved in the formulation.*

The consistency of the numerical fluxes is obtained by adapting the conditions brought out in [3]

$$i\omega\left(\widehat{\boldsymbol{v}}\left(p,g\right)\right)|_{\partial T}\cdot\boldsymbol{n}_T=A_T\boldsymbol{\nabla}p_T\cdot\boldsymbol{n}_T,\quad\left(\widehat{p}\left(p,g\right)\right)|_{\partial T}=p_T,\ \forall T\in\mathcal{T},\tag{33}$$

if p is a solution to problem (8) relatively to the right-hand side data g.

The Trefftz-DG method is then obtained from the reciprocity relation (32) by replacing the traces of the exact solution by the numerical fluxes according to the consistency relation (33)

$$\int_{\partial T}\left(i\omega\widehat{\boldsymbol{v}}\cdot\boldsymbol{n}_Tq_T-\widehat{p}A_T\boldsymbol{\nabla}q_T\cdot\boldsymbol{n}_T\right)ds=0,\ \forall q_T\in W_T,\ \forall T\in\mathcal{T}.\tag{34}$$

It is clear that if the numerical fluxes are chosen in such a way that the variational equation (34) is stable relatively to the small perturbations of p, the method is then straightforwardly convergent.

Since the complex conjugation $q_T\rightarrow\overline{q_T}$ is an inner operation in W_T, for all $T\in\mathcal{T}$, we can as well consider the equivalent formulation obtained by taking $\overline{q_T}$ *as* test function in (34)

$$\int_{\partial T}\left(i\omega\widehat{\boldsymbol{v}}\cdot\boldsymbol{n}_T\overline{q_T}-\widehat{p}A_T\overline{\boldsymbol{\nabla}q_T\cdot\boldsymbol{n}_T}\right)ds=0,\ \forall q_T\in W_T,\ \forall T\in\mathcal{T}.\tag{35}$$

Apparently, formulation (35) is the exclusive one considered in the literature, the consistency condition (33) being otherwise implicitly included in the definition of numerical fluxes (see [27] and the references therein).

In the other hand, it is actually more convenient to express the internal numerical fluxes in terms of averages and jumps

$$\left\{\begin{array}{ll}\left[\!\left[p\right]\!\right]=p_T\boldsymbol{n}_T+p_L\boldsymbol{n}_L & \left[\!\left[\boldsymbol{v}\right]\!\right]=\boldsymbol{v}_T\cdot\boldsymbol{n}_T+\boldsymbol{v}_L\cdot\boldsymbol{n}_L\\ \left\{\!\left\{p\right\}\!\right\}=\left(p_T+p_L\right)/2 & \left\{\!\left\{\boldsymbol{v}\right\}\!\right\}=\left(\left(\boldsymbol{v}_T\cdot\boldsymbol{n}_T\right)\boldsymbol{n}_T+\left(\boldsymbol{v}_L\cdot\boldsymbol{n}_L\right)\boldsymbol{n}_L\right)/2\end{array}\right.\tag{36}$$

at any internal edge/face $F_{\{T,L\}}\in\mathcal{A}$.

Remark 7 *Contrary to the general approach adopted in [3], it is more convenient in this study to define the jump* $\left\{\!\left\{\boldsymbol{v}\right\}\!\right\}$ *of* $\boldsymbol{v}$ *in (36) in terms of the normal components* $\boldsymbol{v}_T\cdot\boldsymbol{n}_T$ *of* $\boldsymbol{v}_T$ *and* $\boldsymbol{v}_L\cdot\boldsymbol{n}_L$ *of* $\boldsymbol{v}_L$ *instead of the values* $\boldsymbol{v}_T$ *and* $\boldsymbol{v}_L$ *on* $F_{\{T,L\}}$

$$\left\{\!\left\{\boldsymbol{v}\right\}\!\right\}=\left(\boldsymbol{v}_T+\boldsymbol{v}_L\right)/2.$$

With respect to the expression of the internal numerical fluxes in terms of averages and jumps, we have the following proposition.

Proposition 8 *Any system of internal numerical fluxes*

$$\left\{\begin{array}{l} \widehat{\boldsymbol{v}} = \alpha_T^{(1)} p_T \boldsymbol{n}_T + \alpha_L^{(1)} p_L \boldsymbol{n}_L + \beta_T^{(1)} \boldsymbol{v}_T \cdot \boldsymbol{n}_T \ \boldsymbol{n}_T + \beta_L^{(1)} \boldsymbol{v}_L \cdot \boldsymbol{n}_L \ \boldsymbol{n}_L, \\ \widehat{p} = \alpha_T^{(2)} p_T + \alpha_L^{(2)} p_L + \beta_T^{(2)} \boldsymbol{v}_T \cdot \boldsymbol{n}_T + \beta_L^{(2)} \boldsymbol{v}_L \cdot \boldsymbol{n}_L. \end{array}\right. \tag{37}$$

related to $F_{\{T,L\}} \in \mathcal{A}$ satisfying the consistency condition (33) can be expressed by means of two scalar constants and two vector constants normal to Γ as follows

$$\widehat{\boldsymbol{v}} = \{\!\{\boldsymbol{v}\}\!\} + \alpha [\![p]\!] + \boldsymbol{\delta} [\![\boldsymbol{v}]\!], \quad \widehat{p} = \{\!\{p\}\!\} + \boldsymbol{\gamma} \cdot [\![p]\!] + \beta [\![\boldsymbol{v}]\!], \tag{38}$$

where $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ are two vectors normal to $F_{\{T,L\}}$.

Proof. We start from the following expressions

$$\begin{aligned} p_T \boldsymbol{n}_T &= \tfrac{1}{2}\left(p_T + p_L\right) \boldsymbol{n}_T + \tfrac{1}{2}\left(p_T - p_L\right) \boldsymbol{n}_T = \{\!\{p\}\!\} \boldsymbol{n}_T + \tfrac{1}{2} [\![p]\!], \\ \boldsymbol{v}_T \cdot \boldsymbol{n}_T &= \tfrac{1}{2}\left(\boldsymbol{v}_T \cdot \boldsymbol{n}_T + \boldsymbol{v}_L \cdot \boldsymbol{n}_L\right) + \tfrac{1}{2}\left(\boldsymbol{v}_T \cdot \boldsymbol{n}_T - \boldsymbol{v}_L \cdot \boldsymbol{n}_L\right) \\ &= \tfrac{1}{2} [\![\boldsymbol{v}]\!] + \{\!\{\boldsymbol{v}\}\!\} \cdot \boldsymbol{n}_T, \\ p_T &= \{\!\{p\}\!\} + \tfrac{1}{2} [\![p]\!] \cdot \boldsymbol{n}_T, \\ \left(\boldsymbol{v}_T \cdot \boldsymbol{n}_T\right) \boldsymbol{n}_T &= \{\!\{\boldsymbol{v}\}\!\} + \tfrac{1}{2} [\![\boldsymbol{v}]\!] \boldsymbol{n}_T. \end{aligned}$$

Inserting these expressions in (37), we get

$$\begin{aligned} \widehat{\boldsymbol{v}} = \left(\beta_T^{(1)} + \beta_L^{(1)}\right) \{\!\{\boldsymbol{v}\}\!\} + \left(\alpha_T^{(1)} \boldsymbol{n}_T + \alpha_L^{(1)} \boldsymbol{n}_L\right) \{\!\{p\}\!\} + \\ \frac{1}{2}\left(\alpha_T^{(1)} + \alpha_L^{(1)}\right) [\![p]\!] + \frac{1}{2}\left(\beta_T^{(1)} \boldsymbol{n}_T + \beta_L^{(1)} \boldsymbol{n}_L\right) [\![\boldsymbol{v}]\!], \end{aligned}$$

$$\begin{aligned} \widehat{p} = \left(\alpha_T^{(2)} + \alpha_L^{(2)}\right) \{\!\{p\}\!\} + \left(\beta_T^{(2)} \boldsymbol{n}_T + \beta_L^{(2)} \boldsymbol{n}_L\right) \{\!\{\boldsymbol{v}\}\!\} + \\ \frac{1}{2}\left(\alpha_T^{(2)} \boldsymbol{n}_T + \alpha_L^{(2)} \boldsymbol{n}_L\right) [\![p]\!] + \frac{1}{2}\left(\beta_T^{(2)} + \beta_L^{(2)}\right) [\![\boldsymbol{v}]\!]. \end{aligned}$$

We readily then obtain that internal numerical fluxes (37) are in the form (38). ■

Remark 9 *It seems that numerical fluxes in the form (38) have been introduced in [9, Eq. (2.4)] for the Laplace equation. The aim of these authors was to define the LDG methods, i.e., those corresponding to $\beta = 0$. For this choice, $\widehat{p}$ does not depend on $\boldsymbol{v}$ so that the latter can be eliminated at the level of the assembly process. We will see below that the Trefftz-DG formulations, that are equivalent to the general UWVF, in no way can be an LDG method.*

3.2 Equivalence of UWVF and Trefftz-DG formulation

Hereafter, $p = \mathcal{X} x$ and $q = \mathcal{X} y$ denote two elements of $W_{\mathcal{T}}$ respectively associated with x and y in $X_{\mathcal{T}}$ by means of the bijective map χ. The following proposition links ultra-weak formulations (28) and (29) with Trefftz-DG methods (35) and (34) respectively.

Proposition 10 *Denoting by $(\mathcal{F}x+g)_T$ the component related to $T\in\mathcal{T}$ of the element $\mathcal{F}x+g$ in the product space $X_{\mathcal{T}}$, the following relations hold true*

$$\int_{\partial T}\left(x_T\overline{y_T}-\mathcal{U}_T^*\left(\mathcal{F}x+g\right)_T\overline{y_T}\right)\eta_T ds = \frac{1}{2i\omega}\int_{\partial T}\left(i\omega\widehat{\boldsymbol{v}}\cdot\boldsymbol{n}_T\overline{q_T}-\widehat{p}\overline{A_T\boldsymbol{\nabla}q_T\cdot\boldsymbol{n}_T}\right)ds \tag{40}$$

$$\int_{\partial T}\left(\mathcal{U}_T x_T y_T-\left(\mathcal{F}x+g\right)_T y_T\right)\eta_T ds = \frac{1}{2i\omega}\int_{\partial T}\left(i\omega\widehat{\boldsymbol{v}}\cdot\boldsymbol{n}_T q_T-\widehat{p}A_T\boldsymbol{\nabla}q_T\cdot\boldsymbol{n}_T\right)ds \tag{41}$$

where

$$\begin{cases}\widehat{\boldsymbol{v}}=\frac{2\eta_T\eta_L}{\eta_T+\eta_L}\left(x_T\boldsymbol{n}_T+x_L\boldsymbol{n}_L\right),\\ \widehat{p}=\frac{2\eta_T}{\eta_T+\eta_L}x_T+\frac{2\eta_L}{\eta_T+\eta_L}x_L,\end{cases} \tag{42}$$

on internal edges/faces $F_{\{T,L\}}\in\mathcal{A}$ and

$$\begin{cases}\widehat{\boldsymbol{v}}=\eta_T\left(\left(1-Q_T\right)x_T-g_T\right)\boldsymbol{n}_T,\\ \widehat{p}=\left(1+Q_T\right)x_T+g_T,\end{cases} \tag{43}$$

on boundary edges/faces $F_T^{\partial\Omega}\in\mathcal{B}$.

Proof. Using the symmetry of $\mathcal{U}_T$, we first put (29) in the form

$$\int_{\partial T}\left(x_T\mathcal{U}_T y_T-\left(\mathcal{F}x+g\right)_T y_T\right)\eta_T ds = \frac{1}{2i\omega}\int_{\partial T}\left(x_T 2i\omega\eta_T\mathcal{U}_T y_T-\left(\mathcal{F}x+g\right)_T 2i\omega\eta_T y_T\right)ds.$$

Noting then that $y_T=\Lambda^+_{T,\eta_T}q_T$ and $\mathcal{U}_T y_T=\Lambda^-_{T,\eta_T}q_T$, we immediately get from (20)

$$\int_{\partial T}\left(x_T\mathcal{U}_T y_T-\left(\mathcal{F}x+g\right)_T y_T\right)\eta_T ds = \frac{1}{2i\omega}\int_{\partial T}i\omega\eta_T\left(x_T-\left(\mathcal{F}x+g\right)_T\right)-\left(x_T+\left(\mathcal{F}x+g\right)_T\right)A_T\boldsymbol{\nabla}q_T\cdot\boldsymbol{n}_T ds.$$

Relation (41) is then obtained in a straightforward way.

Let us now start from (41) with $\overline{\mathcal{U}_T y_T}$ as test function. Denoting by $r_T=\mathcal{X}_T\overline{\mathcal{U}_T y_T}$, we get

$$\int_{\partial T}\left(\mathcal{U}_T x_T\overline{\mathcal{U}_T y_T}-\left(\mathcal{F}x+g\right)_T\overline{\mathcal{U}_T y_T}\right)\eta_T ds = \frac{1}{2i\omega}\int_{\partial T}\left(i\omega\widehat{\boldsymbol{v}}\cdot\boldsymbol{n}_T r_T-\widehat{p}A_T\boldsymbol{\nabla}r_T\cdot\boldsymbol{n}_T\right)ds.$$

From (27), we readily obtain that $r_T = \overline{q_T}$ and $A_T \boldsymbol{\nabla} r_T \cdot \boldsymbol{n}_T = \overline{A_T \boldsymbol{\nabla} q_T \cdot \boldsymbol{n}_T}$ and then (40) from

$$\int_{\partial T} \left(\mathcal{U}_T x_T \overline{\mathcal{U}_T y_T} - (\mathcal{F}x + g)_T \overline{\mathcal{U}_T y_T} \right) \eta_T ds = \int_{\partial T} \left(x_T \overline{y_T} - \mathcal{U}_T^* \left(\mathcal{F}x + g \right)_T \overline{y_T} \right) \eta_T ds.$$

since $\mathcal{U}_T$ is a unitary operator. ■

We are in position to state the following theorem, at the basis of the analysis of the plane-wave discretization of the above ultra-weak formulations.

Theorem 11 *The only internal numerical fluxes in the form (38) and which are expressed through the traces x_T and x_L outgoing on both sides of the internal edge/face $F_{\{T,L\}} \in \mathcal{A}$ are those corresponding to the following coefficients*

$$\alpha = \frac{\eta_T \eta_L}{\eta_T + \eta_L}, \quad \beta = \frac{1}{\eta_T + \eta_L}, \quad \boldsymbol{\gamma} = -\boldsymbol{\delta} = \frac{1}{2} \frac{1}{\eta_T + \eta_L} \left(\eta_T \boldsymbol{n}_T + \eta_L \boldsymbol{n}_L \right). \quad (44)$$

These numerical fluxes are actually those given in (42).

Proof. In view of (38), we write $\widehat{\boldsymbol{v}}$ and $\widehat{p}$ as follows

$$\begin{cases} \widehat{\boldsymbol{v}} = \delta_T \left(\boldsymbol{v}_T \cdot \boldsymbol{n}_T + \frac{\alpha}{\delta_T} p_T \right) \boldsymbol{n}_T + \delta_L \left(\boldsymbol{v}_L \cdot \boldsymbol{n}_L + \frac{\alpha}{\delta_L} p_L \right) \boldsymbol{n}_L, \\ \widehat{p} = \beta \left(\left(\boldsymbol{v}_T \cdot \boldsymbol{n}_T + \frac{\gamma_T}{\beta} p_T \right) + \left(\boldsymbol{v}_L \cdot \boldsymbol{n}_L + \frac{\gamma_L}{\beta} p_L \right) \right), \end{cases}$$

with

$$\delta_T = \frac{1}{2} + \boldsymbol{\delta} \cdot \boldsymbol{n}_T, \quad \delta_L = \frac{1}{2} + \boldsymbol{\delta} \cdot \boldsymbol{n}_L, \quad \gamma_T = \frac{1}{2} + \boldsymbol{\gamma} \cdot \boldsymbol{n}_T, \quad \gamma_L = \frac{1}{2} + \boldsymbol{\gamma} \cdot \boldsymbol{n}_L,$$

The numerical fluxes $\widehat{\boldsymbol{v}}$ and $\widehat{p}$ are expressed through x_T and x_L if and only if their related coefficients are solution to the following linear system

$$\alpha / Y_T - \delta_T = 0, \quad \alpha / Y_L - \delta_L = 0, \quad \gamma_T - \beta Y_T = 0, \quad \gamma_L - \beta Y_L = 0.$$

The end of the proof can then be obtained by a simple check. ■

3.3 UWVF and upwind schemes

Some authors classify the UWVF among the methods issued from a upwind scheme for hyperbolic systems [8, 16, 26]. This way of seeing this formulation can be used to bring out its dispersive properties [1]. This also casts the UWVF in the same framework as Discontinuous Galerkin Methods and the Least Square Methods [17], giving rise to specific numerical approaches. There are however several ways to define a upwind scheme. We first try below to make this clearer. The first step is to write the general Helmholtz equation (7) in the form of a

first-order system. This can be obtained by first setting $\boldsymbol{v} = (1/i\omega)\, A\boldsymbol{\nabla} p$ coming to

$$\begin{cases} -i\omega p + (1/\chi)\, \boldsymbol{\nabla}\cdot \boldsymbol{v} = 0, \\ -i\omega \boldsymbol{v} + A\boldsymbol{\nabla} p = 0. \end{cases}$$

Reminding that multiplication by $-i\omega$, when one does not distinguish between a field and its phasor, is the time derivative ∂_t, we arrive to the following system

$$\begin{cases} \partial_t p + (1/\chi)\, \boldsymbol{\nabla}\cdot \boldsymbol{v} = 0, \\ \partial_t \boldsymbol{v} + A\boldsymbol{\nabla} p = 0. \end{cases} \tag{45}$$

It is important to note that this system is not in a conservative form.

Remark 12 *There are several other ways to reduce Helmholtz equation (7) to a first-order linear system. For instance, by setting $\boldsymbol{m} = (1/i\omega)\, \boldsymbol{\nabla} p$, we would have arrived to the following system*

$$\begin{cases} \partial_t p + (1/\chi)\, \boldsymbol{\nabla}\cdot A\boldsymbol{m} = 0, \\ \partial_t \boldsymbol{m} + \boldsymbol{\nabla} p = 0. \end{cases} \tag{46}$$

By considering system (45), we have favoured the fact that it is the normal component $\boldsymbol{v}\cdot\boldsymbol{n}$ to the interface between T and L which is continuous when passing from a subdomain T to another subdomain L of Ω. Here, $\boldsymbol{n}$ is the unit normal to the interface directed towards the exterior of T. For the second system (46), it is rather the normal component $A\boldsymbol{m}\cdot\boldsymbol{n}$ of $A\boldsymbol{m}$ which is preferred to be continuous. Assuming that χ is constant, we can put system (46) in the conservative form

$$\begin{cases} \partial_t p + \boldsymbol{\nabla}\cdot (1/\chi)\, A\boldsymbol{m} = 0, \\ \partial_t \boldsymbol{m} + \boldsymbol{\nabla} p = 0. \end{cases} \tag{47}$$

For the acoustic system, $\boldsymbol{m}$ is the acoustic disturbances of the momentum in system (46) while $\boldsymbol{v}$ is the acoustic disturbances of the velocity in system (46). Similarly, coefficient $\chi = 1/c^2\varrho$ is then the adiabatic bulk compressibility, denoted K_s in [28, p. 28], and $A = 1/\varrho$ where c and ϱ are respectively the sound velocity and the density of the fluid at rest in which the sound is propagating. The conservative form is advantageous because the fluxes function is expressed through the matrix characterizing the hyperbolicity of system (47) in the direction of $\boldsymbol{n}$

$$F\begin{bmatrix} p \\ \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} \boldsymbol{n}\cdot \frac{1}{\chi} A\boldsymbol{v} \\ p\boldsymbol{n} \end{bmatrix} = \begin{bmatrix} 0 & \boldsymbol{n}\cdot \frac{1}{\chi} A \\ \boldsymbol{n} & 0_{d\times d} \end{bmatrix} \begin{bmatrix} p \\ \boldsymbol{v} \end{bmatrix},$$

formally obtained by replacing the operator $\boldsymbol{\nabla}$ by the unit normal $\boldsymbol{n}$ to the interface. For constant coefficients, i.e. A and χ constant, the diagonalization of this matrix directly leads to the construction of numerical fluxes (cf., e.g., [16]) in the form of a flux splitting

$$\widehat{F\begin{bmatrix} p \\ \boldsymbol{v} \end{bmatrix}} = F^+ \begin{bmatrix} p_T \\ \boldsymbol{v}_T \end{bmatrix} + F^- \begin{bmatrix} p_L \\ \boldsymbol{v}_L \end{bmatrix} \tag{48}$$

$$F = F^+ + F^- \tag{49}$$

where F^+ and F^- are defined through the eigenvalue decomposition of F, where, as above, subscripts T or L refer respectively to values of the related functions from T and L. An extension of the decomposition (49) and associated numerical fluxes (48) is given in [17] for the acoustic system in conservative form assuming only piecewise constant coefficients, i.e. for χ constant and $(1/\chi)\,A = (1/\chi)\,A_T = c_T^2$ in T and $(1/\chi)\,A = (1/\chi)\,A_L = c_L^2$ in L. It is obtained by combining the two decompositions $F_T = F_T^+ + F_T^-$ and $F_L = F_L^+ + F_L^-$, assuming in each case that the coefficients are constant and equal respectively to χ, A_T and χ, A_L. This combination is based on an analogy with the boundary conditions ensuring the well-posedness of a mixed hyperbolic system (hyperbolic system with initial and boundary conditions).. Since for non constant χ, system (45) is not in conservative form, we have here to follow a different path. Adapting the techniques developed in [33, Chap. 9], we solve the associated Riemann problem to directly get the numerical fluxes.

The Riemann problem is inherently a one-dimensional problem [33, p. 6]. We thus assume that T and L are the two half-spaces

$$T = \left\{x \in \mathbb{R}^d;\ (x-m)\cdot\boldsymbol{n} < 0\right\}, \quad L = \left\{x \in \mathbb{R}^d;\ (x-m)\cdot\boldsymbol{n} > 0\right\}$$

and we look for solutions to system (45) in the form

$$(s,t) \to \left[\begin{array}{c} p \\ \boldsymbol{v} \end{array}\right](m + s\boldsymbol{n}, t)$$

and an initial value constant for $s < 0$ and $s > 0$. This can be rewritten as

$$\left\{\begin{array}{l} \left\{\begin{array}{l} \partial_t p + (1/\chi_T)\,\partial_s \boldsymbol{n}\cdot\boldsymbol{v} = 0 \\ \partial_t \boldsymbol{v} + A_T \boldsymbol{n}\partial_s p = 0 \end{array}\right. (s<0), \ \left\{\begin{array}{l} \partial_t p + (1/\chi_L)\,\partial_s \boldsymbol{n}\cdot\boldsymbol{v} = 0 \\ \partial_t \boldsymbol{v} + A_L \boldsymbol{n}\partial_s p = 0 \end{array}\right. (s>0), \\ p|_{s=0^-} = p|_{s=0^+},\ \boldsymbol{n}\cdot\boldsymbol{v}|_{s=0^-} = \boldsymbol{n}\cdot\boldsymbol{v}|_{s=0^+} \text{ for } t>0, \\ p|_{s<0,t=0^+} = p_T,\ \boldsymbol{v}|_{s<0,t=0^+} = \boldsymbol{v}_T, \\ p|_{s>0,t=0^+} = p_L,\ \boldsymbol{v}|_{s>0,t=0^+} = \boldsymbol{v}_L. \end{array}\right.$$

Actually, the solution of this Riemann problem can be obtained from the following decomposition

$$\left[\begin{array}{c} p \\ \boldsymbol{v} \end{array}\right] = \left[\begin{array}{c} 0 \\ \boldsymbol{v}_{||} \end{array}\right] + \left[\begin{array}{c} p \\ (\boldsymbol{n}\cdot\boldsymbol{v})\,\boldsymbol{n} \end{array}\right]$$

and $\left[\begin{array}{cc} p & \boldsymbol{n}\cdot\boldsymbol{v} \end{array}\right]^\top$ solution to the following Riemann problem

$$\left\{\begin{array}{l} \left\{\begin{array}{l} \partial_t p + (1/\chi_T)\,\partial_s \boldsymbol{n}\cdot\boldsymbol{v} = 0 \\ \partial_t \boldsymbol{n}\cdot\boldsymbol{v} + a_T^2 \partial_s p = 0 \end{array}\right. (s<0), \ \left\{\begin{array}{l} \partial_t p + (1/\chi_L)\,\partial_s \boldsymbol{n}\cdot\boldsymbol{v} = 0 \\ \partial_t \boldsymbol{n}\cdot\boldsymbol{v} + a_L^2 \partial_s p = 0 \end{array}\right. (s>0), \\ p|_{s=0^-} = p|_{s=0^+},\ \boldsymbol{n}\cdot\boldsymbol{v}|_{s=0^-} = \boldsymbol{n}\cdot\boldsymbol{v}|_{s=0^+} \text{ for } t>0, \\ p|_{s<0,t=0^+} = p_T,\ \boldsymbol{n}\cdot\boldsymbol{v}|_{s<0,t=0^+} = \boldsymbol{n}\cdot\boldsymbol{v}_T, \\ p|_{s>0,t=0^+} = p_L,\ \boldsymbol{n}\cdot\boldsymbol{v}|_{s>0,t=0^+} = \boldsymbol{n}\cdot\boldsymbol{v}_L, \end{array}\right. \tag{50}$$

with $a_T = \sqrt{\boldsymbol{n} \cdot A_T \boldsymbol{n}}$ and $a_L = \sqrt{\boldsymbol{n} \cdot A_L \boldsymbol{n}}$. To solve system (50), we use the method of characteristics, similarly to, but in a more straightforward way than, [33, Sect. 9.9]. To this end, we first make the following diagonalization

$$\begin{bmatrix} 0 & 1/\chi \\ a^2 & 0 \end{bmatrix} = \underbrace{\frac{1}{\sqrt{2}} \begin{bmatrix} -1/Y & 1/Y \\ 1 & 1 \end{bmatrix}}_{R} \begin{bmatrix} -c & 0 \\ 0 & c \end{bmatrix} \underbrace{\frac{1}{\sqrt{2}} \begin{bmatrix} -Y & 1 \\ Y & 1 \end{bmatrix}}_{R^{-1}}$$

where a and χ stand for either a_T and χ_T or a_L and χ_L and Y and c are the related directional admittance $Y = a\sqrt{\chi}$, defined in (15) and what can be called the wave directional velocity

$$c = \frac{a}{\sqrt{\chi}}.$$

Directional is related here to the direction of vector $\boldsymbol{n}$. Setting

$$\boldsymbol{w}(s,t) = R_T^{-1} \begin{bmatrix} p|_{s<0} \\ \boldsymbol{n} \cdot \boldsymbol{v}|_{s<0} \end{bmatrix} (s<0), \quad \boldsymbol{w}(s,t) = R_L^{-1} \begin{bmatrix} p|_{s>0} \\ \boldsymbol{n} \cdot \boldsymbol{v}|_{s>0} \end{bmatrix} (s>0),$$

and

$$\boldsymbol{w}(s,t) = \begin{bmatrix} w^-(s,t) \\ w^+(s,t) \end{bmatrix},$$

we get in particular

$$\begin{cases} \partial_t w^+ + c_T \partial_s w^+ = 0 \\ w^+(s,0^+) = \left(1/\sqrt{2}\right)(\boldsymbol{n} \cdot \boldsymbol{v}_T + Y_T p_T) \end{cases} (s<0),$$

$$\begin{cases} \partial_t w^- - c_T \partial_s w^- = 0 \\ w^-(s,0^+) = \left(1/\sqrt{2}\right)(\boldsymbol{n} \cdot \boldsymbol{v}_T - Y_T p_T) \end{cases} (s>0).$$

The method of characteristics (see Fig.) then yields

$$w^+(0^-,t) = \left(1/\sqrt{2}\right)(\boldsymbol{n} \cdot \boldsymbol{v}_T + Y_T p_T), \; w^-(0^+,t) = \left(1/\sqrt{2}\right)(\boldsymbol{n} \cdot \boldsymbol{v}_L - Y_L p_L).$$

The interface conditions, 2nd set of equations in system (50), can be rewritten in terms the boundary values of function $\boldsymbol{w}$

$$\frac{1}{\sqrt{2}} \begin{bmatrix} -1/Y_T & 1/Y_T \\ 1 & 1 \end{bmatrix} \begin{bmatrix} w^-(0^-,t) \\ w^+(0^-,t) \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1/Y_L & 1/Y_L \\ 1 & 1 \end{bmatrix} \begin{bmatrix} w^-(0^+,t) \\ w^+(0^+,t) \end{bmatrix}.$$

This shows that $w^-(0^-,t)$ and $w^+(0^+,t)$ can be expressed in terms of $w^+(0^-,t)$ and $w^-(0^+,t)$ by solving the following linear system

$$\begin{bmatrix} -1/Y_T & -1/Y_L \\ 1 & -1 \end{bmatrix} \begin{bmatrix} w^-(0^-,t) \\ w^+(0^+,t) \end{bmatrix} = \begin{bmatrix} -1/Y_L & -1/Y_T \\ 1 & -1 \end{bmatrix} \begin{bmatrix} w^-(0^+,t) \\ w^+(0^-,t) \end{bmatrix}.$$

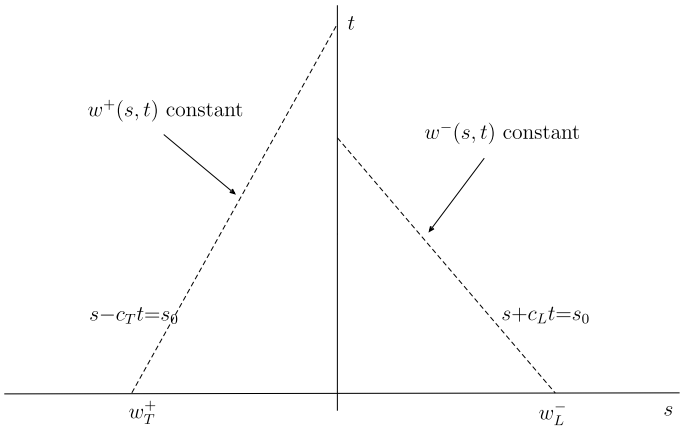

Figure 1: Determination of $w^-(0^+,t)$ and $w^+(0^-,t)$ by the method of the characteristics

Remark 13 *The elimination of the tangential component is fundamental here: it leads to an invertible linear system in $w^-(0^-,t)$ and $w^+(0^-,t)$. Without this elimination, we would have arrived to a linear system of three equations in the two unknowns $w^-(0^-,t)$ and $w^+(0^-,t)$, equivalent to the above 2×2 system. This would also have been the case when using the Riemann solver for the system in conservative form (47). This observation clearly shows that the Riemann solver cannot be used for general hyperbolic systems with constant but different coefficients in T and L. At least, for hyperbolic systems in conservative form, one can always resort to the upwind flux splitting of [17]. However, there is also aconcern with this way of proceeding: it is not clear whether the flows thus generated remain conservative.*

Thus

$$\begin{bmatrix} w^-(0^-,t) \\ w^+(0^+,t) \end{bmatrix} = \begin{bmatrix} \frac{2Y_T}{Y_L+Y_T} & \frac{Y_L-Y_T}{Y_T+Y_L} \\ \frac{Y_T-Y_L}{Y_L+Y_T} & \frac{2Y_L}{Y_L+Y_T} \end{bmatrix} \begin{bmatrix} w^-(0^+,t) \\ w^+(0^-,t) \end{bmatrix}$$
$$= \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{2Y_T}{Y_L+Y_T} & \frac{Y_L-Y_T}{Y_T+Y_L} \\ \frac{Y_T-Y_L}{Y_L+Y_T} & \frac{2Y_L}{Y_L+Y_T} \end{bmatrix} \begin{bmatrix} -Y_L p_L + \boldsymbol{n}\cdot\boldsymbol{v}_L \\ Y_T p_T + \boldsymbol{n}\cdot\boldsymbol{v}_T \end{bmatrix}.$$

It then follows that $p(0^-,t) = p(0^+,t) = p(0,t)$ and $\boldsymbol{v}(0^-,t)\cdot\boldsymbol{n} = \boldsymbol{v}(0^+,t)\cdot\boldsymbol{n} = \boldsymbol{v}(0,t)\cdot\boldsymbol{n}$ are constant for $t>0$ and given by

$$\begin{bmatrix} p(0,t) \\ \boldsymbol{v}(0,t)\cdot\boldsymbol{n} \end{bmatrix} = R_T \begin{bmatrix} w^-(0^-,t) \\ w^+(0^-,t) \end{bmatrix}$$

which can be written noting that $\boldsymbol{n} = \boldsymbol{n}_T = -\boldsymbol{n}_L$ the unit normal directed

towards the exterior of T and L respectively

$$\left\{\begin{array}{l} p\left(0,t\right)=\frac{2Y_T}{Y_T+Y_L}\frac{1}{2Y_T}\left(\boldsymbol{n}_T\cdot\boldsymbol{v}_T+Y_Tp_T\right)+\frac{2Y_L}{Y_T+Y_L}\frac{1}{2Y_L}\left(\boldsymbol{n}_L\cdot\boldsymbol{v}_L+Y_Lp_L\right),\\ \boldsymbol{v}\left(0,t\right)\cdot\boldsymbol{n}=\frac{2Y_TY_L}{Y_T+Y_L}\left(\frac{1}{2Y_T}\left(\boldsymbol{n}_T\cdot\boldsymbol{v}_T+Y_Tp_T\right)-\frac{1}{2Y_L}\left(\boldsymbol{n}_L\cdot\boldsymbol{v}_L+Y_Lp_L\right)\right). \end{array}\right.$$

These values can be used to define the numerical fluxes at the interfaces built through the Riemann solver

$$\left\{\begin{array}{l} \widehat{p}=\frac{2Y_T}{Y_T+Y_L}\frac{1}{2Y_T}\left(\boldsymbol{n}_T\cdot\boldsymbol{v}_T+Y_Tp_T\right)+\frac{2Y_L}{Y_T+Y_L}\frac{1}{2Y_L}\left(\boldsymbol{n}_L\cdot\boldsymbol{v}_L+Y_Lp_L\right),\\ \widehat{\boldsymbol{v}}=\frac{2Y_TY_L}{Y_T+Y_L}\left(\frac{1}{2Y_T}\left(\boldsymbol{n}_T\cdot\boldsymbol{v}_T+Y_Tp_T\right)\boldsymbol{n}_T+\frac{1}{2Y_L}\left(\boldsymbol{n}_L\cdot\boldsymbol{v}_L+Y_Lp_L\right)\boldsymbol{n}_L\right). \end{array}\right.$$

These fluxes are exactly the numerical fluxes (42) defining the DG method equivalent to the UWVF related to the actual admittances (15). We have thus proved the following theorem.

Theorem 14 *Of all the UWVFs considered in the general framework of subsection 2.4, the one for actual admittances (15) is the only one that is equivalent to a DG method whose numerical interior fluxes are obtained by a Riemann solver, and is an upwind scheme in this meaning.*

3.4 Coercivity in the DG norm

Bilinear and sesquilinear forms, respectively $(x,y)\rightarrow\left(\mathcal{U}x-\mathcal{F}x,y\right)_{X_\mathcal{T}}$ and $(x,y)\rightarrow\left(x-\mathcal{U}^*\mathcal{F}x,\overline{y}\right)_{X_\mathcal{T}}$, do not seem to satisfy a coercivity estimate in the norm of $X_\mathcal{T}$, even in the sense of an inf-sup condition. For the isotropic and constant coefficients case (standard Helmholtz equation $\Delta p+\kappa^2p=0$), Buffa and Monk [8] have shown that the ultra-weak formulation devised by Cessenat and Després [11] has a coercivity property relatively to a weaker norm, the one related to the equivalent Trefftz-DG method. This property is extended below to the ultra-weak formulation (35).

From (35), (40) and theorem 11, we get

$$\begin{array}{l} 2\left(x-\mathcal{U}^*\mathcal{F}x,\overline{y}\right)_{X_\mathcal{T}}\\ =\frac{1}{i\omega}\sum\limits_{F_{\{T,L\}}\in\mathcal{A}}\int_{F_{\{T,L\}}}\left(\begin{array}{c} i\omega\left(\{\!\{\boldsymbol{v}\}\!\}+\alpha[\![p]\!]-\boldsymbol{\gamma}[\![\boldsymbol{v}]\!]\right)\overline{[\![q]\!]}\\ -\left(\{\!\{p\}\!\}+\boldsymbol{\gamma}\cdot[\![p]\!]+\beta[\![\boldsymbol{v}]\!]\right)\overline{[\![A\boldsymbol{\nabla}q]\!]}\end{array}\right)ds\\ +\frac{1}{i\omega}\sum\limits_{F_T^{\partial\Omega}\in\mathcal{B}}\int_{F_T^{\partial\Omega}}\left(\boldsymbol{v}_T\cdot\boldsymbol{n}_T+\eta_Tp_T\right)\left(i\omega\frac{1-Q_T}{2}\overline{q_T}-\frac{1+Q_T}{2\eta_T}\overline{A_T\boldsymbol{\nabla}q_T\cdot\boldsymbol{n}_T}\right)ds. \end{array}$$

Returning to $A\boldsymbol{\nabla}p=i\omega\boldsymbol{v}$, we can write this expression as

$$\begin{array}{l} 2\left(x-\mathcal{U}^*\mathcal{F}x,\overline{y}\right)_{X_\mathcal{T}}\\ =\frac{1}{i\omega}\sum\limits_{F_{\{T,L\}}\in\mathcal{A}}\int_{F_{\{T,L\}}}\left(\begin{array}{c} \left(\{\!\{A\boldsymbol{\nabla}p\}\!\}+i\omega\alpha[\![p]\!]-\boldsymbol{\gamma}[\![A\boldsymbol{\nabla}p]\!]\right)\overline{[\![q]\!]}\\ -\left(\{\!\{p\}\!\}+\boldsymbol{\gamma}\cdot[\![p]\!]+\frac{\beta}{i\omega}[\![A\boldsymbol{\nabla}p]\!]\right)\overline{[\![A\boldsymbol{\nabla}q]\!]}\end{array}\right)ds\\ +\frac{1}{i\omega}\sum\limits_{F_T^{\partial\Omega}\in\mathcal{B}}\int_{F_T^{\partial\Omega}}\left(\frac{1}{i\omega}A_T\boldsymbol{\nabla}p_T\cdot\boldsymbol{n}_T+\eta_Tp_T\right)\left(i\omega\frac{1-Q_T}{2}\overline{q_T}-\frac{1+Q_T}{2\eta_T}\overline{A_T\boldsymbol{\nabla}q_T\cdot\boldsymbol{n}_T}\right)ds, \end{array}$$

which is then organized as follows

$$
\begin{aligned}
&2\left(x-\mathcal{U}^{*}\mathcal{F}x,\overline{y}\right)_{X_{\mathcal{T}}}\\
&=\sum_{F_{\{T,L\}}\in\mathcal{A}}\int_{F_{\{T,L\}}}\left(\begin{array}{c}\left(\alpha[\![p]\!]\overline{[\![q]\!]}+\frac{\beta}{\omega^{2}}[\![A\boldsymbol{\nabla}p]\!]\overline{[\![A\boldsymbol{\nabla}q]\!]}\right)\\ -\frac{1}{i\omega}\left([\![A\boldsymbol{\nabla}p]\!]\boldsymbol{\gamma}\cdot\overline{[\![q]\!]}+\boldsymbol{\gamma}\cdot[\![p]\!]\overline{[\![A\boldsymbol{\nabla}q]\!]}\right)\end{array}\right)ds\\
&+\sum_{F_{T}^{\partial\Omega}\in\mathcal{B}}\int_{F_{T}^{\partial\Omega}}\left(\eta_{T}\tfrac{1-Q_{T}}{2}p_{T}q_{T}+\tfrac{1}{\omega^{2}\eta_{T}}\tfrac{1+Q_{T}}{2}A\boldsymbol{\nabla}p_{T}\cdot\boldsymbol{n}_{T}A\boldsymbol{\nabla}q_{T}\cdot\boldsymbol{n}_{T}\right)ds\\
&+\frac{1}{i\omega}\sum_{F_{\{T,L\}}\in\mathcal{A}}\int_{F_{\{T,L\}}}\left(\{\!\{A\boldsymbol{\nabla}p\}\!\}\overline{[\![q]\!]}-\{\!\{p\}\!\}\overline{[\![A\boldsymbol{\nabla}q]\!]}\right)ds\\
&+\frac{1}{i\omega}\sum_{F_{T}^{\partial\Omega}\in\mathcal{B}}\int_{F_{T}^{\partial\Omega}}\tfrac{1}{2}\left(A_{T}\boldsymbol{\nabla}p_{T}\cdot\boldsymbol{n}_{T}\overline{q_{T}}-p_{T}\overline{A_{T}\boldsymbol{\nabla}q_{T}\cdot\boldsymbol{n}_{T}}\right)ds\\
&+\frac{1}{i\omega}\sum_{F_{T}^{\partial\Omega}\in\mathcal{B}}\int_{F_{T}^{\partial\Omega}}\tfrac{Q_{T}}{2}\left(A\boldsymbol{\nabla}p_{T}\cdot\boldsymbol{n}_{T}\overline{q_{T}}+p_{T}\overline{A_{T}\boldsymbol{\nabla}q_{T}\cdot\boldsymbol{n}_{T}}\right)ds. \qquad (51)
\end{aligned}
$$

We first prove the following proposition.

Proposition 15 *The expression*

$$
\begin{aligned}
|x|_{DG}=\Bigg(&\sum_{F_{\{T,L\}}\in\mathcal{A}}\int_{F_{\{T,L\}}}\left(\alpha\left|[\![p]\!]\right|^{2}+\frac{\beta}{\omega^{2}}\left|[\![A\boldsymbol{\nabla}p]\!]\right|^{2}\right)ds\\
&+\sum_{F_{T}^{\partial\Omega}\in\mathcal{B}}\int_{F_{T}^{\partial\Omega}}\left(\eta_{T}\tfrac{1-Q_{T}}{2}\left|p_{T}\right|^{2}+\tfrac{1}{\omega^{2}\eta_{T}}\tfrac{1+Q_{T}}{2}\left|A_{T}\boldsymbol{\nabla}p_{T}\cdot\boldsymbol{n}_{T}\right|^{2}\right)ds\Bigg)^{1/2}
\end{aligned}
$$

defines a norm on $X_{\mathcal{T}}$.

Proof. Since the map $p\to x$ is a bijective map from $W_{\mathcal{T}}$ onto $X_{\mathcal{T}}$, it is enough to prove that $|x|_{\mathrm{DG}}=0$ implies $p=0$. Thus, assuming that $|x|_{\mathrm{DG}}=0$, we first get that $[\![p]\!]=[\![A\boldsymbol{\nabla}p]\!]=0$, which implies that $p\in H^{1}(\Omega)$ and verifies $\boldsymbol{\nabla}\cdot A\boldsymbol{\nabla}p+\omega^{2}\chi p=0$ in Ω. Since $|Q|\leq 1$, p satisfies $(1-Q)\,p=0$ and $(1+Q)\,A\boldsymbol{\nabla}p\cdot\boldsymbol{n}=0$ on $\partial\Omega$. Hence, $\mathcal{U}x=Qx$. The uniqueness of problem (8) enables us to conclude that $p=0$. ■

We thus arrive at the following fundamental result, extending that of Buffa and Monk [8] for the standard Helmholtz equation.

Theorem 16 *Ultra-weak formulation (28) satisfies*

$$
2\Re\left(x-\mathcal{U}^{*}\mathcal{F}x,\overline{x}\right)_{X_{\mathcal{T}}}=|x|_{DG}^{2},\quad\forall x\in X_{\mathcal{T}}. \qquad (52)
$$

Proof. Taking $q = p$ in (51) and passing to the real part of the obtained expression, we get

$$\begin{aligned}2\Re\left(x-\mathcal{U}^*\mathcal{F}x,\overline{x}\right)_{X_\mathcal{T}} &= |x|^2_{\mathrm{DG}} \\ &+ \Re\frac{1}{i\omega}\sum_{F_{\{T,L\}}\in\mathcal{A}}\int_{F_T^L}\left(\{\!\{A\boldsymbol{\nabla}p\}\!\}\overline{[\![p]\!]}-\{\!\{p\}\!\}\overline{[\![A\boldsymbol{\nabla}p]\!]}\right)ds \\ &+ \Re\frac{1}{i\omega}\sum_{F_T^{\partial\Omega}\in\mathcal{B}}\int_{F_T^{\partial\Omega}}\tfrac{1}{2}\left(A\boldsymbol{\nabla}p_T\cdot\boldsymbol{n}_T\overline{p_T}-p_T\overline{A\boldsymbol{\nabla}p_T\cdot\boldsymbol{n}_T}\right)ds\end{aligned}$$

Adding and substracting $\overline{\{\!\{p\}\!\}}[\![A\boldsymbol{\nabla}p]\!]$ and $\overline{p_T}A\boldsymbol{\nabla}p_T\cdot\boldsymbol{n}_T$, we come to the following equality

$$\begin{aligned}2\Re\left(x-\mathcal{U}^*\mathcal{F}x,\overline{x}\right)_{X_\mathcal{T}} &= |x|^2_{\mathrm{DG}} \\ &+ \Re\frac{1}{i\omega}\sum_{F_{\{T,L\}}\in\mathcal{A}}\int_{F_T^L}\left(\{\!\{A\boldsymbol{\nabla}p\}\!\}\overline{[\![p]\!]}+\overline{\{\!\{p\}\!\}}[\![A\boldsymbol{\nabla}p]\!]\right)ds \\ &+ \Re\frac{1}{i\omega}\sum_{F_T^{\partial\Omega}\in\mathcal{B}}\int_{F_T^{\partial\Omega}}A\boldsymbol{\nabla}p_T\cdot\boldsymbol{n}_T\overline{p_T}ds.\end{aligned}$$

Developing the averages and the jumps, we directly put the above equality in the form

$$2\Re\left(x-\mathcal{U}^*\mathcal{F}x,\overline{x}\right)_{X_\mathcal{T}} = |x|^2_{\mathrm{DG}} + \Re\frac{1}{i\omega}\sum_{T\in\mathcal{T}}\int_{\partial T}A\boldsymbol{\nabla}p_T\cdot\boldsymbol{n}_T\overline{p_T}ds.$$

Green's formula and the fact that $p\in W_\mathcal{T}$ make it possible to readily complete the proof. ■

In particular, the above theorem immediately yields that the DG-norm in $X_\mathcal{T}$ is dominated by the canonical norm of this space

$$|x|_{\mathrm{DG}} \le 2\left\|x\right\|_{X_\mathcal{T}}, \quad \forall x\in X_\mathcal{T},$$

and that the direct UWVF (29) satisfies a kind of inf-sup condition

$$2\Re\left(\mathcal{U}x-\mathcal{F}x,\overline{\mathcal{U}x}\right)_{X_\mathcal{T}} = |x|^2_{\mathrm{DG}}.$$

In general, almost all properties of the ultra-weak formulation (28) can be transposed to the ultra-weak formulation (29) as soon as the considered subspaces of $X_\mathcal{T}$ are stable by the following correspondence $x\to\overline{\mathcal{U}x}$. As a result, throughout the rest of this paper, we mention only those properties that are not common to both formulations.

At this point, a difficulty stems since the sesquilinear form $(x,y)\to(x-\mathcal{U}^*\mathcal{F}x,\overline{y})_{X_\mathcal{T}}$ has an upper bound in the norm of $X_\mathcal{T}$, that is, $\left|(x-\mathcal{U}^*\mathcal{F}x,\overline{y})_{X_\mathcal{T}}\right| \le 2\left\|x\right\|_{X_\mathcal{T}}\left\|y\right\|_{X_\mathcal{T}}$, and the coercivity property (52) in the DG-norm only. Such difficulty is well-known in the analysis of DG-methods. It is overcome by establishing a sharper upper bound of the sesquilinear form by a term of the form

$C\,|x|_{\mathrm{DG}}\,\|y\|_{\mathrm{DG}+}$, where $|x|_{\mathrm{DG}}$ is the norm for coercivity and $\|y\|_{\mathrm{DG}+}$ is that in which is expressed the continuity of the sesquilinear form (cf., e.g., [27, 36]). We follow the same path here, the role of the DG+ norm being played by the natural norm of $X_{\mathcal{T}}$. The main argument is a simple but powerful trick devised by Imbert-Gérard and Després [31].

Theorem 17 *The following upper-bound holds true*

$$\left|(x-\mathcal{U}^*\mathcal{F}x,\overline{y})_{X_{\mathcal{T}}}\right| \leq |x|_{DG}\,\|y\|_{X_{\mathcal{T}}} \tag{53}$$

for all x and y in $X_{\mathcal{T}}$.

Proof. Cauchy-Schwarz inequality yields

$$\left|(x-\mathcal{U}^*\mathcal{F}x,\overline{y})_{X_{\mathcal{T}}}\right| \leq \|x-\mathcal{U}^*\mathcal{F}x\|_{X_{\mathcal{T}}}\,\|y\|_{X_{\mathcal{T}}}\,.$$

Since $\mathcal{U}$ is a unitary operator and the norm of $\mathcal{F}$ is ≤ 1, we can use Imbert-Gérard and Després trick to get

$$\begin{aligned}\|x-\mathcal{U}^*\mathcal{F}x\|^2_{X_{\mathcal{T}}} &= \|x\|^2_{X_{\mathcal{T}}} + \|\mathcal{U}^*\mathcal{F}x\|^2_{X_{\mathcal{T}}} - 2\Re\,(\mathcal{U}^*\mathcal{F}x,\overline{x})_{X_{\mathcal{T}}}\\ &\leq 2\Re\,(x-\mathcal{U}^*\mathcal{F}x,\overline{x})_{X_{\mathcal{T}}} = |x|^2_{\mathrm{DG}}\,.\end{aligned}$$

This completes the proof of the theorem. ■

3.5 Duality estimates

Coercivity and continuity properties given above will enable us to set out error estimates in the DG-norm $|\cdot|_{\mathrm{DG}}$. More significant estimates can be obtained in the $L^2\,(\Omega)$-norm by the duality techniques introduced in [40]. Towards this end, we assume that elements T of $\mathcal{T}$ can be obtained from a reference element $\widehat{T}$ by means of a linear-affine map and that they satisfy the following uniformity condition

$$\min_{T\in\mathcal{T}} d_T \geq h/C,\ C>0, \quad h=\max_{T\in\mathcal{T}} h_T \tag{54}$$

d_T being the diameter of the inscribed ball in T, and h_T the diameter of this element. This condition implies that elements T satisfy the following Finite Element non-degeneracy condition

$$h_T/d_T \leq Ch_T/h \leq C.$$

We also make use of the following spaces of functions, that are piecewise in a Sobolev space of order $s>0$

$$\mathcal{H}^s_{\mathcal{L}} = \left\{p\in L^2\,(\Omega)\,;\ p|_{\Omega^{(\ell)}}\in H^s\left(\Omega^{(\ell)}\right),\ \ell\in\mathcal{L}\right\}.$$

The non-overlapping decomposition $\left\{\Omega^{(\ell)}\right\}_{\ell\in\mathcal{L}}$ of Ω is defined within the statement of problem (8).

We now state and prove the following theorem adapting the result of Monk and Wang [40] to the present context.

Theorem 18 *Under the above general conditions on the non-overlaping decomposition $\mathcal{T}$ of Ω, if there exists $0 < \eta \leq 1/2$ such that the bounded operator $\varphi \in L^2(\Omega) \to u \in H^1(\Omega)$, with*

$$\begin{cases} \boldsymbol{\nabla} \cdot A\boldsymbol{\nabla} u + \omega^2 \chi u = -\varphi \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega_D, \\ A\boldsymbol{\nabla} u \cdot \boldsymbol{n} - i\omega Y^{\partial\Omega} u = 0 \text{ on } \partial\Omega_N, \end{cases} \tag{55}$$

is also bounded as operator $\varphi \in L^2(\Omega) \to u \in \mathcal{H}^{3/2+\eta}_{\mathcal{L}}$, then the following bound holds with a constant C independent of $x \in X_{\mathcal{T}}$ and of $\mathcal{T}$

$$\|\mathcal{X}x\|_{0,\Omega} \leq Ch^{-1/2} |x|_{DG}. \tag{56}$$

Proof. Set $p = \mathcal{X}x$ as above and consider φ and u linked by (55). Using that $p \in W_{\mathcal{T}}$, and Green's formula, we get

$$\begin{aligned} \int_\Omega \varphi p dx &= \sum_{T\in\mathcal{T}} \int_T -\left(\boldsymbol{\nabla} \cdot A\boldsymbol{\nabla} u + \omega^2 \chi u\right) p dx \\ &= \sum_{T\in\mathcal{T}} \int_T \left(u\boldsymbol{\nabla} \cdot A_T \boldsymbol{\nabla} p_T - p_T \boldsymbol{\nabla} \cdot A_T \boldsymbol{\nabla} u\right) dx \\ &= \sum_{T\in\mathcal{T}} \int_{\partial T} \left(uA_T \boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T - p_T A_T \boldsymbol{\nabla} u \cdot \boldsymbol{n}_T\right) dx. \end{aligned}$$

Switching to a sum on the edges/faces rather than on the elements T, we can write

$$\begin{aligned} \int_\Omega \varphi p dx = \sum_{F_{\{T,L\}}\in\mathcal{A}} \int_{F_{\{T,L\}}} &\left(u[\![A\boldsymbol{\nabla} p]\!] - \{\!\{A\boldsymbol{\nabla} u\}\!\}[\![p]\!]\right) ds \\ &+ \sum_{F_T^{\partial\Omega}\in\mathcal{B}} \int_{F_T^{\partial\Omega}} \left(uA_T \boldsymbol{\nabla} p_T \cdot \boldsymbol{n}_T - p_T A_T \boldsymbol{\nabla} u \cdot \boldsymbol{n}_T\right) ds. \end{aligned}$$

Since $u = 0$ when $1 + Q_T = 0$ and $A_T \boldsymbol{\nabla} u \cdot \boldsymbol{n}_T = 0$ when $1 - Q_T = 0$, we can use the Cauchy-Schwarz inequality to obtain

$$\left|\int_\Omega \varphi p dx\right| \leq C |x|_{\mathrm{DG}} \left(\sum_{T\in\mathcal{T}} \left(\|u\|^2_{L^2(\partial T)} + \|\boldsymbol{\nabla} u\|^2_{L^2(\partial T)}\right)\right)^{1/2}.$$

Estimate [37, Est. (7.11)]

$$\|v\|^2_{L^2(\partial T)} \leq C\left(h^{-1}\|v\|^2_{L^2(T)} + h^{2s-1}|v|^2_{s,T}\right) \tag{57}$$

with C a constant independent of T and v yield

$$\sum_{T\subset\Omega^{(\ell)}} \|v\|^2_{L^2(\partial T)} \leq C\left(h^{-1}\|v\|^2_{L^2\left(\Omega^{(\ell)}\right)} + h^{2s-1}|v|^2_{s,\Omega^{(\ell)}}\right),\ \ell \in \mathcal{L},$$

where C is a constant independent of $v \in \mathcal{H}^s_{\mathcal{L}}$ with $1/2 < s \leq 1$, and

$$|v|^2_{s,D} = \int_{D\times D} \frac{|v(x)-v(y)|^2}{|x-y|^{d-1+2s}} dxdy \text{ if } s<1 \text{ and } |v|^2_{1,D} = \|\boldsymbol{\nabla} v\|^2_{L^2(D)},$$

D being either T or $\Omega^{(\ell)}$. Proceeding as in [37, Rem. 4.3.9 and 4.4.11], we readily complete the proof. ■

It remains to give conditions on the geometry and coefficients of problem (55) ensuring the extra-regularity assumed on its solution u. The three spatial dimensions problem seems elusive [42, 43]. We thus focus below on the 2D problem ($d = 2$).

Nicaise and Sändig theory of "*general interface problems*" [42, 43], in particular when restricted to problem (55) in two dimensions, shows that, under a condition, that will be given below, u is equal to the superposition of $u_0 \in \mathcal{H}^2_{\mathcal{L}}$ and a finite expansion in singular functions at each vertex S of the decomposition $\left\{\Omega^{(\ell)}\right\}_{\ell\in\mathcal{L}}$ of Ω. It should be noted that a point S on $\partial\Omega$ at the junction of $\partial\Omega_D$ and $\partial\Omega_N$ is considered as a vertex even if it is located in the interior of an edge of $\partial\Omega$. Both u_0 and the finite expansion coefficients continuously depend on φ in the $L^2(\Omega)$-norm. The singular functions $u_{S,\lambda}$ at a vertex S are in the following form

$$u_{S,\lambda} = \eta_S r^\lambda \sum_{j=1}^{N_{S,\lambda}} \sum_{k_j=0}^{m_{S,\lambda,j}} c_{S,\lambda,j,k_j} v^{S,\lambda}_{j,k_j}(\theta) \log^{k_j} r \tag{58}$$

where (r,θ) are the polar coordinates at S

$$x = S + r(\cos\theta, \sin\theta) \tag{59}$$

and η_S is cutoff function. The power λ corresponds to the solutions in the form $u_\lambda = r^\lambda v(\theta)$ of the following problem

$$\begin{cases} \boldsymbol{\nabla}\cdot A\boldsymbol{\nabla} u_\lambda = 0 \text{ in } C_S, \\ u_\lambda = 0 \text{ on } \partial\Omega_D \cap \partial C_S, \\ A\boldsymbol{\nabla} u_\lambda \cdot \boldsymbol{n} = 0 \text{ on } \partial\Omega_N \cap \partial C_S, \end{cases} \tag{60}$$

with $C_S = \{x \in \Omega;\ |x-S| < \varrho\}$ for a small enough ϱ, with the further condition

$$0 < \Re\lambda < 1. \tag{61}$$

For an interior vertex S, C_S is a disk and the boundary conditions are not applicable.

Elementary calculations yield that $u_\lambda = r^\lambda v(\theta)$ is solution to (60) if and only if (λ, v) is an eigenpair of the following quadratic eigenvalue problem

$$\begin{cases} \lambda \in \mathbb{C},\ v \in V,\ v \neq 0,\ \forall \phi \in V, \\ \displaystyle\int_{\theta_0}^{\theta_N} a_{\theta\theta}\partial_\theta v \partial_\theta \phi d\theta - \lambda \int_{\theta_0}^{\theta_N} a_{r\theta}\left(\partial_\theta v \phi - v\partial_\theta\phi\right) d\theta \\ \qquad\qquad -\lambda^2 \displaystyle\int_{\theta_0}^{\theta_N} a_{rr} v\phi d\theta = 0, \end{cases} \tag{62}$$

with $\boldsymbol{e}_r = \partial_r x$, $\boldsymbol{e}_\theta = (1/r)\partial_\theta x$, x given by (59), $a_{rr} = \boldsymbol{e}_r \cdot A\boldsymbol{e}_r$, $a_{r\theta} = a_{\theta r} = \boldsymbol{e}_r \cdot A\boldsymbol{e}_\theta$, $a_{\theta\theta} = \boldsymbol{e}_\theta \cdot A\boldsymbol{e}_\theta$,

$$C_S = \left\{(r,\theta) \in \mathbb{R}^2;\ 0 < r < \varrho,\ \theta_0 < \theta < \theta_N\right\}$$

and the usual adaptations when C_S is a disk and V being the subspace of those $v \in H^1\left(]\theta_0, \theta_N[\right)$ satisfying

- $v(\theta_0) = 0$ if $S + (r\cos\theta_0, r\sin\theta_0) \in \partial C_S \cap \partial\Omega_D$ for $0 < r < \varrho$,
- $v(\theta_N) = 0$ if $s + (r\cos\theta_N, r\sin\theta_N) \in \partial C_S \cap \partial\Omega_D$ for $0 < r < \varrho$,
- $v\left(0\right) = v\left(2\pi\right)$ if C_S is a disk (in this case, $\theta_0 = 0$ and $\theta_N = 2\pi$ and S is an interior point of Ω).

We will also consider the special instance of problem (62) obtained for $A = 1$

$$\left\{\begin{array}{l} \lambda \in \mathbb{C},\ v \in V,\ v \neq 0,\ \forall \phi \in V, \\ \displaystyle\int_{\theta_0}^{\theta_N} \partial_\theta v \partial_\theta \phi d\theta - \lambda^2 \int_{\theta_0}^{\theta_N} v\phi d\theta = 0, \end{array}\right. \tag{63}$$

and in particular its first positive eigenvalue denoted below λ_S.
Assuming that

$$C_S \cap \Omega^{(\ell_j)} = \left\{x \in \mathbb{R}^2;\ x = S + r\left(\cos\theta, \sin\theta\right),\ 0 < r < \varrho,\ \theta_{j-1} < \theta < \theta_j\right\} \tag{64}$$

for $j = 1, \ldots, N$, and $\theta_0 < \theta_1 < \cdots < \theta_{N-1} < \theta_N$, coefficients a_{rr}, $a_{\theta y}$ and $a_{r\theta}$ are thus in $\mathcal{C}^\infty\left([\theta_{j-1}, \theta_j]\right)$.

The general results in [42, 43], based on the ellipticity of the boundary-value interface problem (55), ensure that

- there is at most a finite number of eigenvalues in the strip defined by (61),
- the eigenspace associated with each eigenvalue is finite-dimensional,
- the family $\left\{v_{j,0}^{S,\lambda}\right\}_{j=1}^{j=N_{S,\lambda}}$ (see (58)) constitutes a basis of this eigenspace and $\left\{v_{j,k}^{S,\lambda}\right\}_{k=0}^{k=m_{S,\lambda,j}}$ is a Jordan chain associated with λ.

This clearly shows how the singularities (58) at the vertices of the solution u to (55) are connected to the eigenvalues λ and associated Jordan chains of the eigenvalue problem (62).

We are thus able to state the following fundamental theorem, which is a particular version of Nicaise and Sändig's more general results [42, 43].

Lemma 19 *If $\left\{\Omega^{(\ell)}\right\}_{\ell\in\mathcal{L}}$ is a non-overlaping decomposition of the 2D domain Ω in polygonal subdomains and under the following condition*

$$\lambda = 1 + i\xi,\ \xi \in \mathbb{R},\quad \textit{is not an eigenvalue of (62)} \tag{65}$$

the solution $u \in H^1(\Omega)$ to problem (55) is in the form

$$u = u_0 + \sum_S \sum_{0<\Re\lambda<1} u_{S,\lambda} \tag{66}$$

where λ stands for an eigenvalue of (62), $u_{S,\lambda}$ is a singular function given in (58), and u_0 a piecewise H^2 function in $\mathcal{H}^2_{\mathcal{L}}$. Furthermore, u_0 and the coefficients c_{S,λ,j,k_j} continuously depend on φ.

Proof. We are almost in the conditions of application of Theorem 8.6 in [43]. Actually, condition (65) is required in a milder form, for $\xi \neq 0$ only. However, the theorem assumes that a system of operators related to problem (60) is injective modulo an adequate space of homogeneous polynomials. It can be easily checked that this condition is ensured here by requiring that the condition (65) is still valid for $\xi = 0$. ■

Remark 20 *The conclusions of Lemma 19 remain valid without requiring that $\lambda = 1$ is not an eigenvalue of problem (62) provided that problem (60) owns a further property related to its solutions in an appropriate space of homogeneous polynomials [43, Thm. 8.6, Assumption (H2), and Def. 7.2].*

We now state and prove an important result establishing that the singular functions $u_{S,\lambda}$ belong to $\mathcal{H}^{1+\mu}_{\mathcal{L}}$ with $\mu > 0$ depending on λ such that $0 < \Re\lambda < 1$. In particular, this result establishes that it is sufficient to determine the eigenvalues λ of problem (62), without having to care about the associated eigenfunctions or Jordan chains, to check whether the assumption of theorem 18 on problem (55) is satisfied. For this purpose, we use an approach introduced by Makhlouf [35], based on the Sobolev embedding theorem.

To do so, we first remark that, since A is constant in each $\Omega^{(\ell)}$, $\ell \in \mathcal{L}$, there exists $\theta_0 < \theta_1 < \cdots < \theta_N$ such that the above functions $v^{S,\lambda}_{j,k_j}$ are such that $v^{S,\lambda}_{j,k_j}|_{]\theta_i,\theta_{i+1}[} \in \mathcal{C}^\infty([\theta_i, \theta_{i+1}])$. As a result, by extending each of these restrictions to a 2π-periodic function in $\mathcal{C}^\infty(\mathbb{R})$, we can focus on the regularity in a neighborhood of 0 of functions of the following type

$$w(x) = r^\lambda p_m(\theta, \ln r)\,,\ p_m(\theta, \ln r) = r^\lambda \sum_{k=0}^{m} v_k(\theta) \ln^k r$$

with v_k a 2π-periodic function in $\mathcal{C}^\infty(\mathbb{R})$. For convenience, we say that these functions are pseudo-homogeneous of degree λ. Clearly then, the derivatives $\partial_{x_j} w$, $j = 1, 2$, of a pseudo-homogeneous function of degree λ is a pseudo-homogeneous function of degree $\lambda - 1$, and any pseudo-homogeneous function of degree λ such that $\Re\lambda > 0$ is bounded on each disk B_ϱ centered in 0 and of finite radius $\varrho > 0$.

We first prove the following intermediate result.

Lemma 21 *Any pseudo-homogeneous function w of degree λ such that $-2 < \Re\lambda < 0$ is in $L^P(B_\varrho)$ for*

$$1 \leq p < -\frac{2}{\Re\lambda}$$

Proof. The pseudo-homogeneous function $w \in L^p(B_\varrho)$ if

$$\int_0^\varrho r^{p\Re\lambda+1} dr < +\infty,$$

that is, if $-1 < p\Re\lambda + 1$. ■

The above lemma in particular ensures that the singularities in the decomposition (66) of the solution u to (55) are compatible with its $H^1(\Omega)$ variational regularity.

We now come to the main result allowing to link the singularities of the variational solution of problem(55) to an extra-regularity of this solution.

Lemma 22 *Every pseudo-homogeneous function w of degree λ such that $0 < \Re\lambda < 1$ satisfies*

$$w \in H^{1+s}(B_\varrho) \text{ for } 0 < s < \Re\lambda. \tag{67}$$

Proof. It is enough to prove that $w_j = \partial_{x_j} w$ is in $H^s(B_\varrho)$. We already know that $w_j \in L^p(B_\varrho)$ for $1 \le p < 2/(1-\Re\lambda)$ and $\partial_{x_i} w_j \in L^p(B_\varrho)$ for

$$1 \le p < \frac{2}{2-\Re\lambda}.$$

Therefore, for p satisfying this second condition $w_j \in W^{1,p}(B_\varrho)$. The Sobolev embedding theorems [41, Th. 1] yield that

$$W^{1,p}(B_\varrho) \subset_\to W^{s,q}(B_\varrho) \overset{q=2}{=} H^s(B_\varrho) \text{ for } 1 - \frac{2}{p} > s - \underbrace{\frac{2}{q}}_{=1 \text{ for } q=2}.$$

This last condition can also be written as $s < 2(1-1/p)$ and, on account on the above second condition on p, leads to

$$s < 2(1-1/p) < \Re\lambda.$$

■

The following proposition shows that the result of the above Lemma cannot be improved.

Proposition 23 *For pseudo-homogeneous functions w in the above Lemma in the form $w = r^\lambda v(\theta)$, the condition $0 < s < \Re\lambda$ is also necessary for w to be in $H^{1+s}(B_\varrho)$.*

Proof. Let φ be a cut-off function in $\mathcal{D}(\mathbb{R}^2)$ such that $\varphi \equiv 1$ on B_ϱ. It is proven in [44, p.211] that the Fourier transform $\xi \to \widehat{\varphi w}(\xi)$ has the following behavior

$$\widehat{\varphi w}(\xi) = \mathcal{O}\left(|\xi|^{-\Re\lambda-2}\right)$$

uniformly in $\xi/|\xi|$ as $|\xi| \to \infty$. This shows that φw is in $H^{1+s}(\mathbb{R}^2)$ if and only if

$$\int_a^{+\infty} |\xi|^{2(1+s)} |\xi|^{-2(\Re\lambda+2)} |\xi| \, d|\xi| < +\infty.$$

This establishes the proposition. ■

We thus arrive to the main result of this section.

Theorem 24 *Under the general assumptions of Theorem 19, if furthermore, for each vertex S of the non-overlaping decomposition $\left\{\Omega^{(\ell)}\right\}_{\ell\in\mathcal{L}}$ the eigenvalues λ of problem (62) such that $\Re\lambda > 0$ satisfy*

$$\Re\lambda > \frac{1}{2}, \tag{68}$$

then the condition on the regularity of solutions u to problem (55) is fulfilled.

Proof. The proof immediately follows from Lemmas 19 and 22. ■

In general, the quadratic eigenvalue problem can be solved only numerically. We have implemented a high order finite element method to approximately solve it. However, in the isotropic case, i.e., when A is a scalar, problem (62) is a Sturm-Liouville problem, for which lower bounds of the relevant eigenvalues ensuring (68) can be obtained from the min-max theorem of Courant-Fisher.

Theorem 25 *If A is scalar and denoted a, problem (62) reduces to an eigenvalue problem for a self-adjoint positive operator of compact resolvent. Its eigenvalues with positive real part verify*

$$\Re\lambda \geq \lambda_S \sqrt{\frac{a_{min}}{a_{\max}}} \tag{69}$$

where a_{min} and a_{max} are the minimum and maximum of the piecewise constant function a, and λ_S is the lowest positive eigenvalue of problem (63).

Proof. If $A = a$, $a_{rr} = a_{\theta\theta} = a$, $a_{r\theta} = 0$. Clearly, problem (62) is then an eigenvalue problem for a positive self-adjoint operator of compact resolvent, indeed a Sturm-Liouville problem. Its eigenvalues are hence non negative $0 \leq \mu_1 = \lambda_1^2 \leq \mu_2 = \lambda_2^2 \leq \cdots$, and are such that $\lim_{n\to\infty} \mu_n = +\infty$. The eigenvalues of (62) are actually the square roots $\pm\lambda_n$ of the μ_n. Let us deal with the instance where $V = H^1\left(]\theta_0, \theta_N[\right)$, the other cases can be treated similarly. It is well-known then that the first eigenvalue $\mu_1 = 0$ is simple and that the associated eigenvalue is a constant function $c \neq 0$. On the other hand, for any $v \neq 0$ in V, the following inequality holds true

$$\frac{a_{\min}}{a_{\max}} \frac{\int_{\theta_0}^{\theta_N} |\partial_\theta v|^2 \, d\theta}{\int_{\theta_0}^{\theta_N} |v|^2 \, d\theta} \leq \frac{\int_{\theta_0}^{\theta_N} a\,|\partial_\theta v|^2 \, d\theta}{\int_{\theta_0}^{\theta_N} a\,|v|^2 \, d\theta}.$$

In particular, in view of the characterization of the second eigenvalue λ_S^2 of problem (63), we obtain the following bound

$$\frac{a_{\min}}{a_{\max}} \lambda_S^2 = \frac{a_{\min}}{a_{\max}} \min_{u\neq 0, \int_{\theta_0}^{\theta_N} u d\theta = 0} \frac{\int_{\theta_0}^{\theta_N} |\partial_\theta u|^2 \, d\theta}{\int_{\theta_0}^{\theta_N} |u|^2 \, d\theta} \leq \frac{\int_{\theta_0}^{\theta_N} a\,|\partial_\theta v|^2 \, d\theta}{\int_{\theta_0}^{\theta_N} a\,|v|^2 \, d\theta}$$

for v satisfying the condition

$$\int_{\theta_0}^{\theta_N} v d\theta = 0. \tag{70}$$

To complete the proof, we now mimic that of the max-min theorem [14, p. 406]. We denote by e_1 and e_2 eigenfunctions of problem (62) associated respectively with the eigenvalues $\mu_1 = 0$ and $\mu_2 > 0$ and normalize them as follows

$$\begin{array}{ll} \int_{\theta_0}^{\theta_N} a\partial_\theta e_1 \overline{\partial_\theta v} d\theta = 0, \ \forall v \in V, & \int_{\theta_0}^{\theta_N} a \left|e_1\right|^2 d\theta = 1, \\ \int_{\theta_0}^{\theta_N} a\partial_\theta e_2 \overline{\partial_\theta v} d\theta = \mu_2 \int_{\theta_0}^{\theta_N} a e_2 \overline{v} d\theta, \ \forall v \in V, & \int_{\theta_0}^{\theta_N} a \left|e_2\right|^2 d\theta = 1. \end{array}$$

As well-known, these eigenvalues satisfy also the usual orthogonality property

$$\int_{\theta_0}^{\theta_N} a e_1 \overline{e_2} d\theta = 0.$$

For $v = c_1 e_1 + c_2 e_2 \in \text{span}\,(e_1, e_2)$ to satisfy (70), coefficients c_1 and c_2 must be such that

$$c_1 \int_{\theta_0}^{\theta_N} e_1 d\theta + c_2 \int_{\theta_0}^{\theta_N} e_2 d\theta = 0.$$

Since the coefficient of c_1 in the above equation is not equal to zero, we can choose $v \in \text{span}\,(e_1, e_2)$ satisfying (70) such that

$$\int_{\theta_0}^{\theta_N} a \left|v\right|^2 d\theta = 1.$$

For this v, we have therefore

$$\frac{a_{\min}}{a_{\max}} \lambda_S^2 \leq \frac{\int_{\theta_0}^{\theta_N} a \left|\partial_\theta v\right|^2 d\theta}{\int_{\theta_0}^{\theta_N} a \left|v\right|^2 d\theta} = \left|c_2\right|^2 \lambda_2^2 \leq \lambda_2^2.$$

This proves the theorem. ■

The eigenvalues λ_S are well-known and easy to compute [22, p. 50]

- $\lambda_S = 1$ when S is an interior point, $\theta_0 = 0$, $\theta_N = 2\pi$, and $v\,(0) = v\,(2\pi)$, (one should be aware that, if A is constant in C_S, problem (60) is not involving any special singularity around S);
- $\lambda_S = \pi / (\theta_N - \theta_0)$ when S is a boundary point, $0 \leq \theta_0 < \theta_N \leq 2\pi$, and either $v\,(\theta_0) = v\,(\theta_N) = 0$ or no condition on $v\,(\theta_0)$ and $v\,(\theta_N)$ (Dirichlet or Neumann condition on both sides $\{\theta = \theta_0\}$ and $\{\theta = \theta_N\}$ of C_S);
- $\lambda_S = \pi/2\,(\theta_N - \theta_0)$ when S is a boundary point, $0 \leq \theta_0 < \theta_N \leq 2\pi$, and either $v\,(\theta_0) = 0$ and no condition on $v\,(\theta_N)$ or the opposite (Dirichlet condition on one side and Neumann condition on the other side of C_S).

To illustrate the general theory above, we will now examine in detail a special example related to a single change in the characteristics of an isotropic propagation medium at a straight boundary. This case will show in particular how to deal with the exception of an eigenvalue $\lambda = 1$ for problem (62). The involved truncated cone C_S is depicted in Fig. 2, in particular indicating the

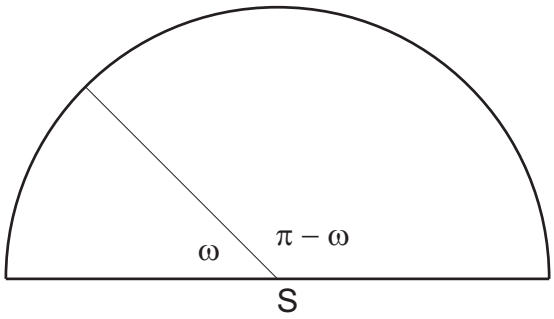

Figure 2: The truncated cone of the dealt with special case.

domains where coefficient a is constant

$$a(x) = \begin{cases} a_1 \text{ for } 0 < \theta < \omega, \\ a_2 \text{ for } \omega < \theta < \pi - \omega, \end{cases}$$

with $x = S + r(\cos\theta, \sin\theta)$. Here we consider a homogeneous Neumann boundary condition on the straight boundaries $\{\theta = 0\}$ and $\{\theta = \pi\}$ so that the problem of regularity can also be addressed by adapting Lemrabet's approach [32] who instead considered Dirichlet conditions. For symmetry reasons, we can assume that $\omega \leq \pi/2$. An easy calculation first shows that any eigenpair (λ, v) of problem (62) with $\lambda > 0$ satisfies

$$v(x) = \begin{cases} \alpha_1 \cos\lambda\theta \text{ for } 0 < \theta < \omega, \\ \alpha_2 \cos\lambda(\theta - \pi) \text{ for } 0 < \theta < \omega, \end{cases}$$

with $|\alpha_1| + |\alpha_2| \neq 0$. The interface conditions at $\{\theta = \omega\}$ are then reduced to $F(\lambda) = 0$ with

$$F(\lambda) = a_1 \sin(\lambda\omega)\cos(\lambda(\pi - \omega)) + a_2 \cos(\lambda\omega)\sin(\lambda(\pi - \omega)).$$

Since $\omega \leq \pi/2$ and we are interested only on eigenvalues such that $0 < \lambda \leq 1$, we can assume that $\cos(\lambda\omega)\cos(\lambda(\pi - \omega)) \neq 0$ for $\lambda \neq \lambda_c$ with $\lambda_c(\pi - \omega) = \pi/2$. A simple discussion then reveals that, for $\omega < \pi/2$ and $0 < \lambda \leq 1$, $F(\lambda) = 0$ if and only if $f(\lambda) = 0$ with

$$f(\lambda) = a_1 \tan(\lambda\omega) + a_2 \tan(\lambda(\pi - \omega)).$$

An examination of the variations of $\lambda \to f(\lambda)$ for $0 < \lambda \leq 1$ then gives that $f(\lambda) = 0$ has a simple zero $0 < \lambda < 1$ if and only if $a_1 > a_2$ (with the a

priori assumption that $a_1 \neq a_2$) and that it always satisfies condition (68). Let us now turn our attention to the limiting case $\omega = \pi - \omega = \pi/2$. We have $F(\lambda) = ((a_1 + a_2)/2) \sin \lambda\pi$. Clearly, we were then in the excluded case $\lambda = 1$. As mentioned in Remark 20, the conditions, ensuring that the conclusions of Lemma 19 remain valid, read as follows: every solution u to problem

$$\left\{ \begin{array}{l} u_1 = u|_{0<\theta<\pi/2},\ u_2 = u|_{\pi/2<\theta<\pi}, \\ \left(\frac{1}{r}\partial_r r \partial_r + \frac{1}{r^2}\partial_\theta^2\right) u_j = 0,\ j = 1, 2, \\ -(a_1/r)\,\partial_\theta u_1|_{\theta=0} = \alpha_1,\ (a_2/r)\,\partial_\theta u_2|_{\theta=\pi} = \alpha_2, \\ (u_1 - u_2)\,|_{\theta=\pi/2} = \beta_1 r,\ \ \frac{1}{r}\partial_\theta\,(a_1 u_1 - a_2 u_2)\,|_{\theta=\pi/2} = \beta_2, \end{array} \right.$$

where $\alpha_1,\ \alpha_2,\ \beta_1,\ \beta_2$ are arbitrary constants, which is furthermore pseudo-homogeneous of degree 1, is such that its restrictions to the truncated cones $\{0 < r < \varrho,\ 0 < \theta < \omega\}$ and $\{0 < r < \varrho,\ \omega < \theta < \pi\}$ are homogeneous polynomials of degree 1. This property is directly checked by elementary calculations. Finally, we solved problem (62) using the above mentioned Finite Element code with data $\omega = \pi/4$, $a_1 = 2$, $a_2 = 1$, using a uniform mesh with a meshsize $h = 1/60$ and a polynomial approximation of degree $m = 6$ and compared the eigenvalue $0 < \lambda < 1$ it furnishes with the zero of the above function $\lambda \to f(\lambda)$ obtained by the MATLAB function `fzero`. Both results gave $\lambda = 0.8933992419\ldots$ and coincided up to 10 decimal digits. This case also shows that estimate (69) is generally too pessimistic.

4 Numerical approximation of the UWVF

4.1 Galerkin approximation of the UWVF

For each $T \in \mathcal{T}$, let X_T^h be a finite-dimensional subspace of the space $L^2_{\eta_T}(\partial T)$. This gives us a finite-dimensional subspace $X_{\mathcal{T}}^h$ of $X_{\mathcal{T}}$, which in turn leads to the Galerkin discretization of both the UWVF (28)

$$\left\{ \begin{array}{l} x^h \in X_{\mathcal{T}}^h,\ \forall y^h \in X_{\mathcal{T}}^h, \\ \left(x^h - \mathcal{U}^* \mathcal{F} x^h, \overline{y}^h\right)_{X_{\mathcal{T}}} = \left(g, \overline{\mathcal{U} y^h}\right)_{X_{\mathcal{T}}}, \end{array} \right. \tag{71}$$

and the Direct-UWVF (29)

$$\left\{ \begin{array}{l} x^h \in X_{\mathcal{T}}^h,\ \forall z^h \in X_{\mathcal{T}}^h, \\ \left(\mathcal{U} x^h - \mathcal{F} x^h, z^h\right)_{X_{\mathcal{T}}} = \left(g, z^h\right)_{X_{\mathcal{T}}}. \end{array} \right. \tag{72}$$

It will be convenient below to refer to a basis of X_T^h as a UWVF-basis. Numbering the elements of a UWVF-basis $\left\{B_{i,T}^h\right\}_{1 \leq i \leq N_T}$ and the elements $T \in \mathcal{T}$ (actually to lighten the notation, we do not distinguish between an element T and its number also denoted T), we readily get that the variational problems (71) and (72) can be respectively put in the form of the following linear systems

$$(M - E)\alpha = c \tag{73}$$

and

$$(A - F)\alpha = b, \tag{74}$$

where α is the block column-vector the blocks of which being the column-vector constituted by the coefficients of $x_T^h \in X_T^h$

$$x_T^h = \sum_{j=1}^{N_T} \alpha_{j,T} B_{j,T}^h, \tag{75}$$

M, E, A, and F are respectively the associated matrices relatively to the sesquilinear and bilinear forms $\left(x^h, z^h\right) \to (x^h, \overline{z^h})_{X_\mathcal{T}}$, $\left(x^h, z^h\right) \to (\mathcal{F}x^h, \overline{\mathcal{U}z^h})_{X_\mathcal{T}}$, $\left(x^h, y^h\right) \to (\mathcal{U}x^h, y^h)_{X_\mathcal{T}}$, $\left(x^h, y^h\right) \to (\mathcal{F}x^h, y^h)_{X_\mathcal{T}}$, and c and b the column-vectors respectively associated to the anti-linear and linear forms $z^h \to (g, \overline{\mathcal{U}z^h})_{X_\mathcal{T}}$ and $y^h \to (g, y^h)_{X_\mathcal{T}}$. The numbering of each row i, T or each column j, L of the above matrices A, ..., or vectors α, b, c, is done accordingly to the notation in (75) of the components of a vector $x^h = \left\{x_T^h\right\}_{T \in \mathcal{T}}$. We will also denote by $A_{T,L} = (A_{i,T;j,L})_{1 \le i \le N_T; 1 \le j \le N_L}$, ..., the various blocks of matrices A, ... From now on, we say that such a matrix is block-diagonal when only the diagonal blocks $A_{T,T}$ are non zero. Clearly, A and M are block-diagonal but neither F nor E are.

The following proposition establishes that, under the assumption that each of the mappings $y_T \in L^2_{\eta_T}(\partial T) \to \overline{\mathcal{U}_T y_T} \in L^2_{\eta_T}(\partial T)$ just permutes the UWVF-basis vectors $\left\{B_{j,T}^h\right\}_{1 \le j \le N_T}$, $T \in \mathcal{T}$, linear systems (73) and (74) are identical except for a permutation of the equations relative to each block $T \in \mathcal{T}$.

Proposition 26 *Under the assumption that for each $T \in \mathcal{T}$, there exists a permutation σ_T of the N_T first positive integers $\{1, \ldots, N_T\}$ such that*

$$\overline{\mathcal{U}_T B_{\sigma_T(i),T}^h} = B_{i,T}^h, \quad i = 1, \ldots, N_T, \tag{76}$$

then, matrices A, F and vector b can be respectively expressed as

$$A = PM, \; F = PE, \; c = Pb, \tag{77}$$

where P is a block-diagonal matrix whose diagonal blocks are permutation matrices of the rows given by

$$P_{i,T;j,T} = \delta_{\sigma_T(i),j}, \quad i, j = 1, \ldots, N_T. \tag{78}$$

Proof. Clearly, the existence of such a permutation is equivalent to the fact that the mapping $y_T \in L^2_{\eta_T}(\partial T) \to \overline{\mathcal{U}_T y_T} \in L^2_{\eta_T}(\partial T)$ transforms each vector of the basis $\left\{B_{j,T}^h\right\}_{1 \le j \le N_T}$ in another vector of this basis. Let us show that $A = PM$, the same proof establishes the other relations. The very definition of M yields

$$M_{\sigma_T(i),T;j,T} = \left(B_{j,T}^h, \overline{B_{\sigma_T(i),T}^h}\right)_{L^2_{\eta_T}(\partial T)}.$$

Using then the fact that $\mathcal{U}_T$ is a unitary operator, we can write

$$M_{\sigma_T(i),T;j,T} = \left(\mathcal{U}_T B^h_{j,T}, \overline{\mathcal{U}_T B^h_{\sigma_T(i),T}}\right)_{L^2_{\eta_T}(\partial T)}$$

and in view of (76)

$$M_{\sigma_T(i),T;j,T} = \left(\mathcal{U}_T B^h_{j,T}, B^h_{i,T}\right)_{L^2_{\eta_T}(\partial T)} = A_{i,T;j,T}.$$

To complete the proof, it just remains to note that

$$M_{\sigma_T(i),T;j,T} = \sum_{l=1}^{N_T} \delta_{\sigma_T(i),l} M_{l,T;j,T},$$

$\delta_{i,j}$ being the Kronecker symbol. ■

We now come to one of the important properties that makes the UWVF particularly suitable for solving the Helmholtz equation, and among the very few methods that can do so [36].

Theorem 27 *Linear system (73) is invertible and it can be solved with Gaussian eliminations without pivoting. As a result, discrete problem (71) has one and only one solution x^h leading to the following general error estimate*

$$\left|x - x^h\right|_{DG} \leq 2 \left\|x - y^h\right\|_{X_\mathcal{T}}, \quad \forall y^h \in X^h_\mathcal{T}, \tag{79}$$

where x is the solution to variational problem (28).

Proof. The proof follows from the following property

$$\alpha^* \left(M - F\right) \alpha = 0 \text{ implies } \alpha = 0,$$

where $\alpha^* = \overline{\alpha}^\top$ is the transpose of the conjugate of α. This property is in turn a direct consequence of the following way to express $\alpha^* \left(M - F\right) \alpha$

$$\alpha^* \left(M - F\right) \alpha = \left(x^h - \mathcal{U}^* \mathcal{F} x^h, \overline{x^h}\right)_{X_\mathcal{T}}, \quad x^h_T = \sum_{j=1}^{N_T} \alpha_{j,T} B^h_{j,T}$$

and the coercivity equality (52).

The second part of the theorem is obtained along the adaptation to DG formulations of the theory of Galerkin approximation of variational problems inspired from [24, 36]. By Galerkin orthogonality, in the terminology of [24], or by Cea's lemma [13] in a more usual FEM vocabulary, we readily get

$$\left(x - x^h - \mathcal{U}^* \mathcal{F}\left(x - x^h\right), \overline{x - x^h}\right)_{X_\mathcal{T}} = \left(x - x^h - \mathcal{U}^* \mathcal{F}\left(x - x^h\right), \overline{x - y^h}\right)_{X_\mathcal{T}}.$$

Then, coercivity (52) and bound (53) yield

$$\frac{1}{2} \left|x - x^h\right|^2_{\mathrm{DG}} \leq \left|x - x^h\right|_{\mathrm{DG}} \left\|x - y^h\right\|_{X_\mathcal{T}}$$

and hence complete the proof. ■

Remark 28 *From a computer programming point of view, it is more advantageous to assemble the matrices of system (74) and, after permuting their rows, store the result as the matrices of system (73). In this way, the operators* $\mathcal{U}$ *and* $\mathcal{F}$ *are completely isolated from each other in the assembly process. Also, one does not have to care about the term* $\overline{\mathcal{U}y^h}$ *when assembling the right hand-side.*

4.2 Directional plane waves and Cessenat's plane waves

The UWVF-basis $\left\{B_{j,T}^h\right\}_{1\leq j\leq N_T}$ must be chosen so that the outgoing-incoming trace operator is made explicit. Furthermore, we require here that the condition (76) must be satisfied. The choice of plane waves as basis vectors meets the first part of this requirement. The plane-wave basis must also have good stability and approximation properties to be usefully implemented in an efficient numerical solution process. In the isotropic case, without additional requirements, these directions can be chosen equally distributed over the unit ball. By highlighting the link of Cessenat's plane waves with directional plane waves and then with usual plane waves with respect to the standard Helmholtz equation, we will be in position to adequatly address this issue in the anisotropic case too.

Let us first recall the definition of Cessenat's plane waves. In all this section, we use the general notation and framework of sub-section 2.2. A Cessenat plane wave, propagating in the direction of unit vector $\boldsymbol{\nu}$, is a particular solution to Eq. (7) in T (recall that $A_T = A|_T$ and $\chi_T = \chi|_T$ in T) in the form

$$p\left(x\right) = \alpha \exp\left(i\omega\sqrt{\chi_T}A_T^{-1/2}\boldsymbol{\nu}\cdot x\right), \tag{80}$$

where $A_T^{-1/2}$ is computed from the eigenvalue decomposition of A_T.

The following proposition states that each directional plane wave is actually a Cessenat plane wave.

Proposition 29 *Each directional plane wave* $p\left(x\right) = \alpha\exp(i\omega\sqrt{\chi_T/\boldsymbol{\nu}\cdot A_T\boldsymbol{\nu}}\,\boldsymbol{\nu}\cdot x)$*, propagating along the unit vector* $\boldsymbol{\nu}$*, is a Cessenat plane wave* $p\left(x\right) = \alpha\exp(i\omega\sqrt{\chi_T}A_T^{-1/2}\widetilde{\boldsymbol{\nu}}\cdot x)$*, propagating along the unit vector*

$$\widetilde{\boldsymbol{\nu}} = \Upsilon_T\left(\boldsymbol{\nu}\right) = \frac{1}{\left|A_T^{1/2}\boldsymbol{\nu}\right|}A_T^{1/2}\boldsymbol{\nu}. \tag{81}$$

The mapping $\boldsymbol{\nu}\rightarrow\Upsilon\left(\boldsymbol{\nu}\right)$ *from the unit sphere* $\mathbb{S}^{d-1}$ *into itself has an explicit inverse given by*

$$\Upsilon_T^{-1}\left(\widetilde{\boldsymbol{\nu}}\right) = \frac{1}{\left|A_T^{-1/2}\widetilde{\boldsymbol{\nu}}\right|}A_T^{-1/2}\widetilde{\boldsymbol{\nu}}. \tag{82}$$

Proof. The first part of the proposition follows by observing that

$$\boldsymbol{\nu}\cdot A_T\boldsymbol{\nu} = \boldsymbol{\nu}\cdot A_T^{1/2}A_T^{1/2}\boldsymbol{\nu} = \left|A_T^{1/2}\boldsymbol{\nu}\right|^2,$$

$$\frac{1}{\left|A^{1/2}\boldsymbol{\nu}\right|}\boldsymbol{\nu}\cdot x = A_T^{-1/2}\left(\frac{1}{\left|A_T^{1/2}\boldsymbol{\nu}\right|}A_T^{1/2}\boldsymbol{\nu}\right)\cdot x.$$

The second part is obtained by first applying $A_T^{-1/2}$ to both sides of (82)

$$A_T^{-1/2}\widetilde{\boldsymbol{\nu}} = \frac{1}{\left|A_T^{1/2}\boldsymbol{\nu}\right|}\boldsymbol{\nu}.$$

Since $\boldsymbol{\nu}$ is a unit vector, we then get

$$\left|A_T^{-1/2}\widetilde{\boldsymbol{\nu}}Th\right| = \frac{1}{\left|A_T^{1/2}\boldsymbol{\nu}\right|}$$

and thus complete the proof. ■

Thanks to the symmetry of $A_T^{-1/2}$, the linear change of variables

$$\widetilde{x} = \mathcal{G}_T x, \text{ with } \mathcal{G}_T x = \sqrt{\chi_T}A_T^{-1/2}x, \tag{83}$$

pulls back a Cessenat plane wave $p(x) = \alpha\exp\left(i\omega\sqrt{\chi_T}A_T^{-1/2}\boldsymbol{\nu}\cdot x\right)$ to a usual plane wave

$$\widetilde{p}(\widetilde{x}) = \alpha\exp\left(i\omega\boldsymbol{\nu}\cdot\widetilde{x}\right). \tag{84}$$

It indeed transforms the anisotropic Helmholtz equation $\boldsymbol{\nabla}_x\cdot A_T\boldsymbol{\nabla}_x p(x) + \omega^2\chi_T p(x) = 0$ into the standard Helmholtz equation times the multiplicative factor χ_T: $\chi_T\left(\Delta_{\widetilde{x}}\widetilde{p}(\widetilde{x}) + \omega^2\widetilde{p}(\widetilde{x})\right) = 0$. This linear change of variables is well known and attributed to I'lin. However, we have not been able to find a reference where this is explicitly mentioned.

We are now in position to prove the following proposition.

Proposition 30 *Each system of N_T directional plane waves*

$$\exp\left(i\omega\sqrt{\frac{\chi_T}{\boldsymbol{\nu}_j\cdot A_T\boldsymbol{\nu}_j}}\boldsymbol{\nu}_j\cdot x\right), \quad 1\leq j\leq N_T, \tag{85}$$

relative to a system of N_T distincts unit vectors $\boldsymbol{\nu}_j$, $j = 1,\ldots,N_T$, is linearly independent.

Proof. Since each directional plane wave $\exp\left(i\omega\sqrt{\chi_T/\boldsymbol{\nu}_j\cdot A_T\boldsymbol{\nu}_j}\boldsymbol{\nu}_j\cdot x\right)$ is a Cessenat's plane wave relative to the unit vector $\boldsymbol{\nu}_j^{\#} = \Upsilon_T\boldsymbol{\nu}_j$ defined by (81), change of variables (83) clearly reduces the proof to the case where $\chi_T = 1$ and $A_T = 1$. This case is well-known (cf., e.g., [11, 2]). However, the proof in [11] is valid in the two-dimensional case only and is based on a complicate argument, while that in [2] requires the use of the Fourier transform in the distributional sense. We show here that this property can be established much more simply. The proof is by induction. The base case is proved by noting that

$\exp(i\omega\boldsymbol{\nu}_1\cdot x)$ is indeed linearly independent. For the induction step, we assume that $\exp(i\omega\boldsymbol{\nu}_l\cdot x)$, $l=1,\ldots,j$, with $j<N_T$, are linearly independent. Let $v(x)=\sum_{l=1}^{j+1}\lambda_l\exp(i\omega\boldsymbol{\nu}_l\cdot x)=0$. Thus

$$\left(\frac{1}{i\omega}\partial_{\boldsymbol{\nu}_{j+1}}-1\right)v(x)=\sum_{l=1}^{j}(\boldsymbol{\nu}_{j+1}\cdot\boldsymbol{\nu}_l-1)\,\lambda_l\exp(i\omega\boldsymbol{\nu}_l\cdot x)=0.$$

Since $\boldsymbol{\nu}_{j+1}\cdot\boldsymbol{\nu}_l-1<0$ for $l=1,\ldots,j$ (limiting case of the Cauchy-Schwarz inequality), we can thus conclude that $\lambda_1=\cdots=\lambda_j=0$. Therefore, $v(x)=\lambda_{j+1}\exp(i\omega\boldsymbol{\nu}_{j+1}\cdot x)=0$ implying that $\lambda_{j+1}=0$. ■

4.3 UWVF-basis associated with a plane-wave basis

To each directional plane-wave basis

$$\left\{e_{j,T}(x)=\exp\left(i\omega\sqrt{\frac{\chi_T}{\boldsymbol{\nu}_j\cdot A_T\boldsymbol{\nu}_j}}\boldsymbol{\nu}_j\cdot x\right)\right\}_{1\le j\le N_T} \tag{86}$$

we can associate a finite family of vectors of X_T by

$$\begin{aligned}\Lambda^+_{T,\eta_T}e_{j,T}&=\frac{1}{2i\omega\eta_T}\left(A_T\boldsymbol{\nabla}e_{j,T}\cdot\boldsymbol{n}_T+i\omega\eta_T e_{j,T}\right)\\&=\frac{1}{2\eta_T}\left(\sqrt{\frac{\chi_T}{\boldsymbol{\nu}_j\cdot A_T\boldsymbol{\nu}_j}}\boldsymbol{n}_T\cdot A_T\boldsymbol{\nu}_j+\eta_T\right)e_{j,T}\end{aligned} \tag{87}$$

for $j=1,\ldots,N_T$. Well-posedness of problem (25) implies that Λ^+_{T,η_T} is an algebraic and topological isomorphism from X_T onto W_T. Therefore, $\{B_{j,T}=\Lambda^+_{T,\eta_T}e_{j,T}\}_{1\le j\le N_T}$ is indeed a UWVF-basis of a finite-dimensional subspace X_T^h of X_T. It is the UWVF-basis associated with the directional plane-wave basis $\{e_{T,j}\}_{1\le j\le N_T}$. Observe that for this UWVF-basis operators $\mathcal{X}_T$ and $\mathcal{U}_T$ are explicit and respectively given by

$$\mathcal{X}_T B_{j,T}=e_{j,T},$$

and

$$\mathcal{U}_T B_{j,T}=\Lambda^-_{T,\eta_T}e_{j,T}=-\frac{1}{2\eta_T}\left(\sqrt{\frac{\chi_T}{\boldsymbol{\nu}_j\cdot A_T\boldsymbol{\nu}_j}}\boldsymbol{n}_T\cdot A_T\boldsymbol{\nu}_j-\eta_T\right)e_{j,T}$$

for $j=1,\ldots,N_T$.

The following proposition shows that it is easy to build directional plane-wave basis such that the associated UWVF-basis satisfies property (76).

Proposition 31 *If the directional plane-wave basis (86) is associated with a family of distinct unit vectors $\{\boldsymbol{\nu}_j\}_{1\le j\le N_T}$ such that*

$$\boldsymbol{\nu}_{\sigma_T(j)}=-\boldsymbol{\nu}_j,\quad j=1,\ldots,N_T,$$

where σ_T is a permutation of the N_T first positive integers, then the associated UWVF-basis $\left\{B_{j,T}=\Lambda^+_{T,\eta_T}e_{T,j}\right\}_{1\le j\le N_T}$ satisfies property (76).

Proof. The proof results from the following equalities

$$\begin{aligned}
\overline{\mathcal{U}_T B_{\sigma_T(j),T}} &= \overline{\Lambda^+_{T,\eta_T} e_{\sigma_T(j),T}} \\
&= \tfrac{1}{2\eta_T} \left(\sqrt{\tfrac{\chi_T}{\boldsymbol{\nu}_{\sigma_T(j)} \cdot A_T \boldsymbol{\nu}_{\sigma_T(j)}}} \boldsymbol{n}_T \cdot A_T \boldsymbol{\nu}_{\sigma_T(j)} + \eta_T \right) \\
&\qquad \overline{\exp\left(i\omega \sqrt{\tfrac{\chi_T}{\boldsymbol{\nu}_{\sigma_T(j)} \cdot A_T \boldsymbol{\nu}_{\sigma_T(j)}}} \boldsymbol{\nu}_{\sigma_T(j)} \cdot x \right)} \\
&= \tfrac{1}{2\eta_T} \left(\sqrt{\tfrac{\chi_T}{(-\boldsymbol{\nu}_j) \cdot A_T (-\boldsymbol{\nu}_j)}} \boldsymbol{n}_T \cdot A_T (-\boldsymbol{\nu}_j) + \eta_T \right) \\
&\qquad \overline{\exp\left(i\omega \sqrt{\tfrac{\chi_T}{(-\boldsymbol{\nu}_j) \cdot A_T (-\boldsymbol{\nu}_j)}} (-\boldsymbol{\nu}_j) \cdot x \right)} \\
&= \Lambda^-_{T,\eta_T} \exp\left(i\omega \sqrt{\tfrac{\chi_T}{\boldsymbol{\nu}_j \cdot A_T \boldsymbol{\nu}_j}} \boldsymbol{\nu}_j \cdot x \right).
\end{aligned}$$

■

Finally, the following proposition establishes that Λ^+_{T,η_T} and Λ^-_{T,η_T} take an extremely simple expression for a directional plane-wave basis when these operators are those related to the actual admittances.

Proposition 32 *Assume that* $\eta_T = Y_T = \sqrt{\chi_T \boldsymbol{n}_T \cdot A_T \boldsymbol{n}_T}$*, then*

$$\Lambda^+_{T,Y_T} e_{j,T} = \cos^2 \theta_{j,A_T} e_{j,T}, \quad \Lambda^-_{T,Y_T} e_{j,T} = \sin^2 \theta_{j,A_T} e_{j,T}$$

where θ_{j,A_T} *is the half-angle relatively to the metric defined by* A_T *of vectors* $\boldsymbol{n}_T$ *and* $\boldsymbol{\nu}_j$*, i.e.,*

$$\theta_{j,A_T} = \frac{1}{2} \arccos\left(\frac{\boldsymbol{n}_T \cdot A_T \boldsymbol{\nu}_j}{\sqrt{\boldsymbol{n}_T \cdot A_T \boldsymbol{n}_T} \sqrt{\boldsymbol{\nu}_j \cdot A_T \boldsymbol{\nu}_j}} \right).$$

Proof. We first put $\Lambda^\pm_{T,Y_T} e_{j,T}$ under the form

$$\begin{aligned}
\Lambda^\pm_{T,Y_T} e_{j,T} &= \frac{1}{2} \left(1 \pm \frac{1}{Y_T} \sqrt{\frac{\chi_T}{\boldsymbol{\nu}_j \cdot A_T \boldsymbol{\nu}_j}} \boldsymbol{n}_T \cdot A_T \boldsymbol{\nu}_j \right) e_{j,T} \\
&= \frac{1}{2} \left(1 \pm \frac{1}{\sqrt{\chi_T \boldsymbol{n}_T \cdot A_T \boldsymbol{n}_T}} \sqrt{\frac{\chi_T}{\boldsymbol{\nu}_j \cdot A_T \boldsymbol{\nu}_j}} \boldsymbol{n}_T \cdot A_T \boldsymbol{\nu}_j \right) e_{j,T} \\
&= \frac{1}{2} \left(1 \pm \frac{1}{\sqrt{\boldsymbol{n}_T \cdot A_T \boldsymbol{n}_T}} \frac{1}{\sqrt{\boldsymbol{\nu}_j \cdot A_T \boldsymbol{\nu}_j}} \boldsymbol{n}_T \cdot A_T \boldsymbol{\nu}_j \right) e_{j,T}.
\end{aligned}$$

Next noting that

$$\left| \frac{1}{\sqrt{\boldsymbol{n}_T \cdot A_T \boldsymbol{n}_T}} \frac{1}{\sqrt{\boldsymbol{\nu}_j \cdot A_T \boldsymbol{\nu}_j}} \boldsymbol{n}_T \cdot A_T \boldsymbol{\nu}_j \right| \leq 1,$$

we can define θ_{j,A_T} by

$$2\theta_{j,A_T} = \arccos\left(\frac{1}{\sqrt{\boldsymbol{n}_T \cdot A_T \boldsymbol{n}_T}} \frac{1}{\sqrt{\boldsymbol{\nu}_j \cdot A_T \boldsymbol{\nu}_j}} \boldsymbol{n}_T \cdot A_T \boldsymbol{\nu}_j \right).$$

The rest of the proof results from the elementary equalities

$$\frac{1}{2}\left(1+\cos\left(2\theta_{j,T}\right)\right)=\cos^2\theta_{j,T},\quad \frac{1}{2}\left(1-\cos\left(2\theta_{j,T}\right)\right)=\sin^2\theta_{j,T}.$$

■

4.4 Approximation by directional plane waves

Proposition 29 establishes that Cessenat's plane waves are in some sense aliases of directional plane waves. The change of variables (83) thus makes it possible to reduce the study of approximating properties of directional plane waves to those of usual plane waves.

From now on, we limit ourselves to the two-dimensional case. In three spatial dimensions, the approximation properties of usual plane waves are much more intricate to describe, and require more strigent conditions on the elements T [37, p. 113] and the choice of the system of approximating plane waves. We also limit ourselves to the h-convergence properties of the plane wave discretization of the UWVF and assume also for simplicity that $\mathcal{T}$ consists of a triangular mesh. Note however that, contrary to the case of a usual finite element method, it is not necessary here to exclude hanging nodes in the mesh.

We are now in position to establish that the use of local bases of directional plane waves for the discretization of the UWVF for the anistropic Helmholtz equation can be carried out in a similar manner than that of the UWVF for the usual Helmholtz equation (see, e.g., [37]), and leads to the same convergence properties for the associated approximation process. The following proposition provides the primary tool for this purpose. Let us recall that $\|u\|_{m,T}$ and $|u|_{j,T}$ respectively denote the usual norm and semi-norm of order $j\leq m$ of an element u in the Sobolev space $H^m(T)$.

Proposition 33 *The transform $u\rightarrow\widetilde{u}$ with $\widetilde{u}=u\circ\mathcal{G}_T$, $\mathcal{G}_T$ being given in (83), induces an isomorphism between $H^m(T)$ and $H^m\left(\widetilde{T}\right)$, with $\widetilde{T}=\mathcal{G}_T(T)$ such that*

$$C_m\left|u\right|_{j,T}\leq\left|\widetilde{u}\right|_{j,\widetilde{T}}\leq C_M\left|u\right|_{j,T},\quad j=0,\ldots,m, \tag{88}$$

where C_m and C_M are two positive constants depending only on m.

Proof. It results from elementary calculations and estimates based on the fact that A_T and χ_T are constant and take only a finite number of values for different $T\in\mathcal{T}$. ■

The directional plane-wave basis (86) is built from a family of N_T equidistributed vectors

$$\widetilde{\boldsymbol{\nu}}_{j,T}=\left[\begin{array}{c}\cos\theta_{j,T}\\ \sin\theta_{j,T}\end{array}\right]$$

on the unit circle with

$$\theta_{j,T}=\theta_{1,T}+(j-1)\,\frac{2\pi}{N_T},\quad j=1,\ldots,N_T. \tag{89}$$

It is then obtained from the unit vectors

$$\boldsymbol{\nu}_{j,T} = \Upsilon_T^{-1} \widetilde{\boldsymbol{\nu}}_{j,T}, \quad j = 1, \ldots, N_T. \tag{90}$$

Since $\Upsilon_T^{-1}(-\widetilde{\boldsymbol{\nu}}_{j,T}) = -\Upsilon_T^{-1}\widetilde{\boldsymbol{\nu}}_{j,T}$, for basis (86) to satisfy (76), N_T must be an even positive integer.

The following proposition thus extends the bounds of the error relative to the best approximation by usual plane waves to that by directional plane waves.

Proposition 34 *Let m be a positive integer and $u \in W_T \cap H^{m+1}(T)$. For $N_T = 2(m+1)$, there exists a system of N_T coefficients $\alpha_{j,T}$, $j = 1, \ldots, N_T$ such that*

$$\left| u - \sum_{j=1}^{N_T} \alpha_{j,T} e_{j,T} \right|_{\ell,T} \leq C h^{m+1-\ell} \|u\|_{m+1,T}, \quad \ell = 0, \ldots, m,$$

where $\{e_{j,T}\}_{1 \leq j \leq N_T}$ is the directional plane-wave basis (86) associated with the directions related to the system of unit vectors (90), and C is a constant independent of h and u.

Proof. In view of definition (90), variable change (83), proposition 29, and estimates (88), seeking a convergent approximation of u in the space spanned by the directional plane-wave basis (86) reduces to that of $\widetilde{u} = \mathcal{G}_T u$ in the space spanned by usual plane wave basis $\{\widetilde{e}_{j,T}\}_{1 \leq j \leq N_T}$

$$\left| u - \sum_{j=1}^{N_T} \alpha_{j,T} e_{j,T} \right|_{\ell,T} \lesssim \left| \widetilde{u} - \sum_{j=1}^{N_T} \alpha_{j,T} \widetilde{e}_{j,\widetilde{T}} \right|_{\ell,\widetilde{T}}, \ \ell = 0, \ldots, m,$$

where $\widetilde{e}_{j,T}(x) = \exp(i\omega \widetilde{\boldsymbol{\nu}}_{j,T} \cdot x)$, $j = 1, \ldots, N_T$ is the usual plane wave basis related to the directions of unit vectors $\widetilde{\boldsymbol{\nu}}_{j,\widetilde{T}} = \Upsilon_T \boldsymbol{\nu}_{j,T}$, $j = 1, \ldots, N_T$, and $\widetilde{T} = \mathcal{G}_T T$. In all the sequel, symbol $\lesssim$ stands for right bounds with a constant depending at most on ω only. Parameter N_T must be even to comply with requirement (76). Unfortunately, the error bounds for the usual-plane wave approximation are only stated for an odd number of such waves [38]. By discarding one basis function, we are led to approximate $\widetilde{u}$ by the family of unit vectors $\{\widetilde{\boldsymbol{\nu}}_{j,T}\}_{1 \leq j \leq N_T - 1}$. In this case, the condition on the quasi asymptotic repartition of the unit vectors on the unit circle [38, cond. (18)]

$$\min_{1 \leq i \neq j \leq N_T - 1} |\theta_{i,T} - \theta_{j,T}| \geq \frac{2\pi}{(N_T - 1)} \delta$$

is satisfied with $\delta = (N_T - 1)/N_T = 1 - 1/N_T$ verifying $3/4 \leq \delta \leq 1$. Using then [38, Estimate (46)] and taking $\alpha_{N_T,T} = 0$, we readily complete the proof by (88). ■

We are now able to prove the following theorem about some error estimates related to the numerical solution of problem (8) by the plane-wave UWVF.

Implicitly, for each for $T \in \mathcal{T}$, X_T^h is the sub-space span $(B_{1,T}, \ldots, B_{N_T,T})$ of $X_{\mathcal{T}}$, with $B_{j,T}$ corresponding to $e_{j,T}$ by (87), $\{e_{j,T}\}_{1 \leq j \leq N_T}$ being the directional plane-wave basis (86) associated with the directions related to the system of unit vectors (90) with N_T chosen as in Proposition 34.

Theorem 35 *Under the general above assumptions, in particular those of Theorem 19, and furthermore condition (68), if the solution p to problem (8) is in $\mathcal{H}_{\mathcal{L}}^{m+1}$ for some integer $m \geq 1$, then the solution x to problem (28) and that x^h to problem (71) satisfy the following error bounds*

$$\left|x - x^h\right|_{DG} \leq C h^{m-1/2} \left\|p\right\|_{\mathcal{H}_{\mathcal{L}}^{m+1}},$$

and

$$\left\|p - p^h\right\|_{L^2(\Omega)} \leq C h^{m-1} \left\|p\right\|_{\mathcal{H}_{\mathcal{L}}^{m+1}},$$

where C is a constant independent of h and p.

Proof. The proof is a simple adaptation of that related to the standard Helmholtz equation discretized with the usual plane-wave UWVF (cf., e.g. [37, Th. 4.4.4]). For the convenience of the reader, the main steps are reproduced below. In all the estimates, C will denote a constant independent of h and p not the same in all instances.

Using (79) and (57), we get

$$\left|x - x^h\right|_{\mathrm{DG}}^2 \leq C \sum_{T \in \mathcal{T}} \left(h^{-1} \left\|p - q^h\right\|_{1,T}^2 + h \left\|p - q^h\right\|_{2,T}^2 \right),$$

with $q^h = \mathcal{X} y^h$, y^h being an arbitrary element of X^h. Now, choosing y^h such that q^h approximates p as in Proposition 34, we come to

$$\left|x - x^h\right|_{\mathrm{DG}}^2 \leq C h^{2m-1} \underbrace{\sum_{T \in \mathcal{T}} \left\|p\right\|_{m+1,T}^2}_{= \left\|p\right\|_{\mathcal{H}_{\mathcal{L}}^{m+1}}^2}.$$

The second error estimate then follows from (56). ■

4.5 Numerical experiments

We focus here on two problems involving an anisotropic material. The first one is related to the reflection and the transmission of the fundamental mode of a waveguide by a piece of anisotropic material. When the anisotropy is directed along the axial and transverse directions of the guide, the problem can be solved analytically. We can therefore compare the plane-wave UWVF (PW-UWVF), i.e. the above UWVF corresponding to the admittances defined in (15) and a plane-wave approximation per element, with a more standard FE polynomial solution of degree 4 on each element on a refined mesh, called hereafter the FE solution. We next examine the efficiency of the PW-UWVF approach for a

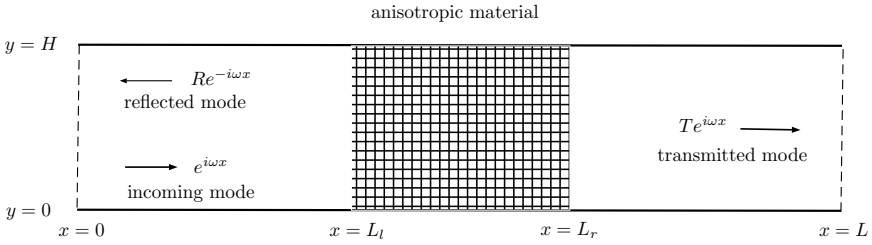

Figure 3: Schematic and data of the waveguide problem.

coarse model of a sub-soil, adapted from a benchmark example designed in [45], in comparison with this FE solver.

Figure 3 presents a schematic of the waveguide problem. The anisotropic material is characterized by the matrix of anisotropy coefficients

$$A_M = \begin{bmatrix} \alpha^2 & \gamma \\ \gamma & \beta^2 \end{bmatrix}, \quad \alpha, \beta > 0, \quad \alpha\beta > \gamma \geq 0, \tag{91}$$

and a refractive index $n_{\text{ind}} \geq 1$

The wave speed c is normalized to 1 so that the wavenumber $\kappa = \omega$. The section size H of the waveguide and ω are fixed so that H is equal to a half-wavelength. In this way, in the parts of the waveguide without material, only the fundamental mode $e^{\pm i\omega x}$ can propagate. The walls of the waveguide $\left\{(x, y) \in \mathbb{R}^2;\ 0 < x < L,\ y = 0, H\right\}$ are assumed to be impenetrable, here with a pure Neumann condition

$$A\boldsymbol{\nabla} p \cdot \boldsymbol{n} = 0 \text{ for } y = 0,\ y = H,$$

with $A = 1$ outside and $A = A_M$ inside the material. The truncating conditions

$$\begin{cases} A\boldsymbol{\nabla} p \cdot \boldsymbol{n} - i\omega p = -2i\omega \text{ for } x = 0, \\ A\boldsymbol{\nabla} p \cdot \boldsymbol{n} - i\omega p = 0 \text{ for } x = L, \end{cases}$$

are quasi exact if L_l and $L - L_r$ are greater than a few wavelengths, typically 2 or 3. In the notation of problem (8), $A = 1$, $\chi = 1$, for $x < L_l$ and $x > L_r$, $A = A_M$ and $\chi = n_{\text{ind}}^2$ for $L_l < x < L_r$, $\partial\Omega_D = \varnothing$, $Y^{\partial\Omega} = 0$ for $y = 0$ and $y = H$, and $Y^{\partial\Omega} = 1$ for $x = 0$ and $x = L$. For $\gamma = 0$, the truncating conditions at $x = 0$ and $x = L$ are exact. Coefficients R and T can then be exactly computed by solving the following system obtained by noting that p has the following expression in terms of R and T and two further coefficients R_M and T_M

$$p(x) = \begin{cases} \exp(i\omega x) + R\exp(-i\omega x), & x < L_l \\ T_M \exp(i\omega x n_{\text{ind}}/\alpha) + R_M \exp(-i\omega x n_{\text{ind}}/\alpha), & L_l < x < L_r, \\ T\exp(i\omega x), & x > L_r, \end{cases} ,$$

and writing the related transmission conditions (16) at $x = L_l$ and $x = L_f$

$$\begin{bmatrix} e^{i\omega L_r n_{\rm ind}/\alpha} & e^{-i\omega L_r n_{\rm ind}/\alpha} & -e^{i\omega L_r} & 0 \\ \alpha n_{\rm ind} e^{i\omega L_r n_{\rm ind}/\alpha} & -\alpha n_{\rm ind} e^{-i\omega L_r n_{\rm ind}/\alpha} & -e^{i\omega L_r} & 0 \\ e^{i\omega L_l n_{\rm ind}/\alpha} & e^{-i\omega L_l n_{\rm ind}/\alpha} & 0 & -e^{-i\omega L_l} \\ \alpha n_{\rm ind} e^{i\omega L_l n_{\rm ind}/\alpha} & -\alpha n_{\rm ind} e^{-i\omega L_l n_{\rm ind}/\alpha} & 0 & e^{-i\omega L_l} \end{bmatrix} \begin{bmatrix} T_M \\ R_M \\ T \\ R \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ e^{i\omega L_l} \\ e^{i\omega L_l} \end{bmatrix}.$$

For $\alpha \neq \beta$ and $\gamma = 0$, the material is still anisotropic with the specificity that the main direction of the anistropy are those of the cartesian axes. For $\gamma \neq 0$, exact expressions for R and T are no more available.

Only the four vertices $\{(L_l, 0), (L_r, 0), (L_l, H), (L_r, H)\}$ are involved in the decomposition (66) of the solution u to problem (55). The decompositions at these points are all the same. They can therefore be obtained by solving the same quadratic eigenvalue problem (62). For $\alpha = 2$ and $\beta = 3$, the above mentionned FE code, running with a uniform decomposition of the interval $]0, \pi[$ in $N_s = 60$ segments and a local polynomial FE approximation of degree 6, gave that the first eigenvalue λ with $\Re\lambda > 0$ is $\lambda = 1.0274$. As a result, the decomposition (66) reduces to the term u_0 so that the solution u to problem (55) is in $\mathcal{H}^2_{\mathcal{L}}$. The mesh $\mathcal{T}$, used for the PW-UWVF, is obtained by a FE mesher in triangles with a meshsize $h = 1$. The units are in half-wavelengths. The FE solution is obtained by refining the mesh $\mathcal{T}$ by subdividing each edge of $\mathcal{T}$ in N_h segments uniformely and correspondingly subdividing each triangle $T \in \mathcal{T}$ as depicted in Fig. 4. Parameter N_h and the number N_T of plane waves on each element $T \in \mathcal{T}$ are set so that the orders of the corresponding global linear systems to be solved are approximately the same.

In order to check the robustness of the PW-UWVF relatively to the "*numerical pollution error*" (roughly speaking, the need to increase the density of degrees of freedom when solving problems set on domains of larger size; see, e.g., [5, 4]), we considered three waveguides of length $L = 10, 100, 1000$ wavelengths respectively, with anisotropic material of two wavelengths thickness placed at their middle. Parameter $\chi = 1$ ouside the material and $\chi = 4$ inside. Parameters of anisotropy were set to $\alpha = 2$, $\beta = 3$, and $\gamma = 0$.

The results are reported in Tab. 1 in terms of the error in % on the reflection and transmission coefficients, R and T respectively, and have been obtained by three methods: the above FE solution with a refinement parameter $N_h = 2$, a plain PW-UWVF with $N_T = 14$ plane waves per element, and the same PW-UWVF but with a local strategy based on a SVD for improving the conditioning [6]. We also mention in the Tab. 1 the related condition number of the global linear system to be solved and its order. Parameter τ of the PW-UWVF with the SVD strategy [6] was fixed to 10^{-8}.

These results well illustrate the improvement brought by the PW-UWVF

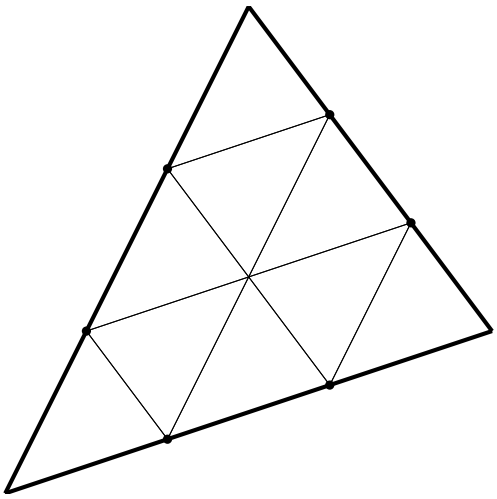

Figure 4: Partitions of edges and triangles corresponding to a refinement parameter $N_h = 3$.

method with respect to the "*numerical pollution error*" and also the improvement of the conditioning gained by the local SVD procedure [6].

We now pass to the simplified model of underground. The main difference with the model considered in [45] is that instead of considering that $A = 1$ everywhere, we assume that it is in the form (91) with $\alpha = 2$, $\beta = \sqrt{10}$, $\gamma = 1$, in domain 2 depicted in Fig. 5 The remaining data are kept from those in [45]. In the present notation, they give $\omega = 2\pi \times \nu$, $\nu = 30$ being the frequency, $\chi = 1/c^2$, c being the velocity, $c = 3000$ in domain 1, $c = 1500$ in domain 2, and $c = 2000$ in domain 1. However, instead of a point source, we prefered here to consider a wide gaussian $g_N(x, y) = \exp\left(-(x/a)^2\right)$, with $a = 10$ to avoid a too singular solution in the vicinity of the point source at the top part of the boundary of the solution domain. We will come back to this point in the concluding remarks below. The other parts of the boundary are endowed with the lowest order absorbing condition $A\boldsymbol{\nabla} u \cdot p + i\omega\sqrt{\chi}\, p = 0$. Using the above FE code to solve the nonlinear eigenvalue problem 62 (with a uniform mesh of 64 segments and a Lagrange FEM of degree 6), we found that any solution to Problem (55) is of maximum regularity except at Points 1 and Point 4. The two first eigenvalues such that $\Re\lambda > 0$ are respectively $\lambda_1 = 0.92$ and $\lambda_2 = 2.38$ for Point 1 and $\lambda_1 = 0.92$ and $\lambda_2 = 2.39$ for Point 2. It then follows from the connection between the real part of these eigenvalues and decomposition (66) of any solution to the problem (55), regularity lemma 22, and Theorem 18 that the coercivity bound (56) holds true for this problem. Mesh $\mathcal{T}$ was obtained using approximately a meshsize $h = 2$, the units being in wavelengths. The reference solution is obtained as described above using a refined mesh by subdividing

L	**Method**	E_R	E_T	N	$\varkappa$
10	PW-UWVF	4.5E-03	8.2E-04	840	1.4E+12
	PW-UWVF+	4.5E-03	8.2E-04	784	1.6E+03
	FE	4.5E-03	1.2E-03	869	9.0E+04
100	PW-UWVF	3.9E-03	5.4E-04	7224	1.4E+12
	PW-UWVF+	3.9E-03	5.3E-04	7020	1.9E+05
	FE	2.2E-02	9.0E-03	7404	9.7E+06
1000	PW-UWVF	4.0E-03	2.2E-03	71680	1.5E+12
	PW-UWVF+	4.0E-03	2.3E-03	69978	2.0E+07
	FE	1.8E-01	8.6E-02	73489	4.5E+08

Table 1: Errors on the reflection E_R and transmission E_T coefficients in %. Parameters N and $\varkappa$ are respectively the order of the final linear system to be solved and an estimate obtained by the MATLAB function `condest`. Methods PW-UWVF, PW-UWVF+, and FE are respectively the plain plane-wave UWVF, the same method but with a local SVD strategy for improving the conditioning, and a FE solution on a refined mesh.

each edge into 10 segments. This results in an approximation of about 20 nodes per wavelength. The PW-UWVF solution is obtained with 32 plane waves per triangle, with and without the local SVD technique for improving the conditioning. The observation is the values of the solution p in the top edge of the domain. The corresponding results are reported in Table 2.

Several observations can be drawn from these results. The first of these is the dramatic impovement of the conditioning of the final linear system to be solved when using the SVD approach of [6]. The second one concerns the low number of degrees of freedom per wavelength required by the PW-UWVF, here only 4 per wavelength, to deliver an accuracy similar to that of the FE solution with 20 nodes per wavelength and a local polynomial approximation of degree 4. The cost of the PW-UWVF solution can even be reduced by a significant factor by the SVD strategy. Table 3 reports the deviation of this PW-UWVF solution from the FE ones obtained with various densities of nodes per wavelength. It shows in particular that the results become to be seriously damaged with the classical FE approach below 8 nodes per wavelength, a density of degrees of freedom which is the double of that related to the PW-UWVF.

5 Concluding remarks

5.1 Advantages and limitations of the PW-UWVF

The study revealed several advantages of the PW-UWVF for the general Helmholtz equation (7)

- It operates like a DG method and, in this respect, may be posed on meshes

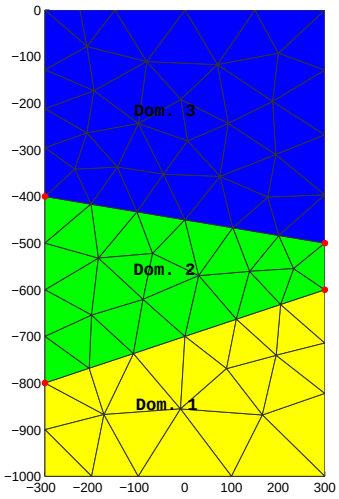

Figure 5: Representation and used mesh $\mathcal{T}$ for the simplified model of underground. The only points in the vicinity of which the solution to Problem (55) may not be of maximum regularity are represented by small red disks. These points are numbered as follows: Point 1 $(-300, -400)$, Point 2 $(300, -500)$, Point 3 $(300, 600)$, Point 4 $(-300, -800)$.

presenting hanging nodes.

- It can be extended without any loss in its stability or approximating properties for the anisotropic case both theoretically and numerically.
- Unlike the FE solution, which may be hampered by fictitious local resonances when solving the final linear system corresponding to Gaussian eliminations without pivoting, the PW-UWVF is theoretically guaranteed not to suffer from this flaw.
- The PW-UWVF requires much less degrees of freedom than a FE solution for delivering similar accuracy.
- It seems less sensitive to the specific instabilities linked to the numerical solution of wave propagation problems known as "*numerical pollution effect*" as reported for the above waveguide problem above. However, the validity of such a claim has to be confirmed by further investigations both at the theoretical level as in [1] and at the numerical one as in [20].

The PW-UWVF has however some limitations.

- Approximations by plane wave bases can lead to very poor conditioning. For example, for the above underground problem, the linear system matrix

PW-UWVF	$\tau = 10^{-8}$	$\tau = 10^{-12}$	$\tau = 10^{-16}$	Plain
# dofs	2241	2767	3273	3296
Cond.	$1.8\ 10^3$	$2.0\ 10^3$	$2.6\ 10^3$	$1.33\ 10^{23}$
Deviat.	2 %	1.17 %	1.17 %	1.17 %

Table 2: Parameter τ refers to the threshold used in the SVD approach in [JCP-paper] to improve the conditioning. Plain indicates that the PW-UWVF is used as is, without improving the conditioning by the SVD approach. Cond. is an estimate of the condition number of the matrix of the final linear system to be solved delivered by the MATLAB function `condest` and #dofs is its order. Deviat. is the deviation in % relative to the L^2-norm.

Density	8	12	16	20
Deviation	3.2 %	1.45 %	1.20 %	1.17 %

Table 3: Deviations of the PW-UWVF solution from the FEM solutions obtained by lower densities. The densities are expressed in number of nodes per wavelength.

resulting from the PW-UWVF with 48 plane waves per element has a condition number of 10^{27}, which in turn leads to numerical outlier results. But this breakdown can now be overcome with effective local strategies [6].

- Plane-wave approximations are also unsuitable for the approximation of singular functions. For example, considering again the underground problem, for a boundary data given by a Gaussian with a narrower peak, $g_N = \exp\left(-x^2\right)$, corresponding to $a = 1$, we still use 48 plane waves per element for the approximation of the PW-UWVF but circumvent the previous poor conditioning by using the SVD procedure of [6]. We now obtain a 15% deviation from the results provided by a FE computation on a mesh that is twice as refined as the one used for the wider peak Gaussian ($a = 10$). The real part of the observation obtained by the FE solution on a refined mesh and the one by the PW-UWVF are plotted in Fig. 7. One can observe that the spirious oscillations of the PW-UWVF solution concentrate near the location of the narrow peak. An obvious way to circumvent this flaw is to use polynomial approximations in the vicinity of the singularities and keep the PW-UWVF approach elsewhere. This is the spirit of the strategy adopted in [18] which consists in coupling a FEM with plane-wave Discontinuous Galerkin method. A criticism that could be made regarding this procedure is that local resonances could be generated when solving the corresponding linear system by Gaussian eliminations without pivoting. A way to extend the UWVF to the polynomial setting to face such flaws is presently ongoing.

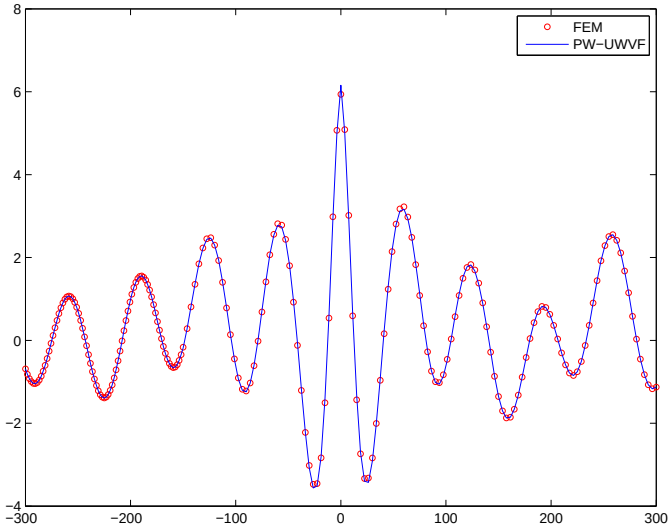

Figure 6: Plots of the real part of p on the top of the domain obtained from the FEM on a refined mesh and the PW-UWVF with 32 plane waves per triangle.

5.2 Final comments

The framework developed in this study is not only compatible with the transmission and reflection of plane waves at normal incidence at the interface between two mesh elements but also allows to cover all the ultra-weak formulations that have been considered for the usual or anisotropic Helmholtz equation with piecewise constant coefficients. We have seen that this framework highlights the similarity in the treatment of interface conditions between mesh elements and boundary conditions. Moreover, it enabled us to establish that the UWVF is equivalent to a DG method and to characterize this method as the unique consistent DG one whose numerical fluxes can be expressed in terms of the outgoing traces. We have also seen that the UWVF compatible with the reflection and transmission of waves at an interface between two elements, is the one whose numerical fluxes are obtained as an upwind scheme resulting from the application of a Riemann solver to the first order hyperbolic system equivalent to the considered Helmholtz equation. This UWVF is also that for which the expression of the outgoing and incoming traces of a plane wave have the simplest expression. Finally, through two non-trivial problems, we have highlighted the exceptional ability of the plane-wave UWVF to efficiently handle difficult numerical simulations. Of course, some issues still need to be addressed. We list below some of them.

- The plane-wave UWVF is specially designed to solve problems over long propagation distances. A specific study of its dispersion and attenuation

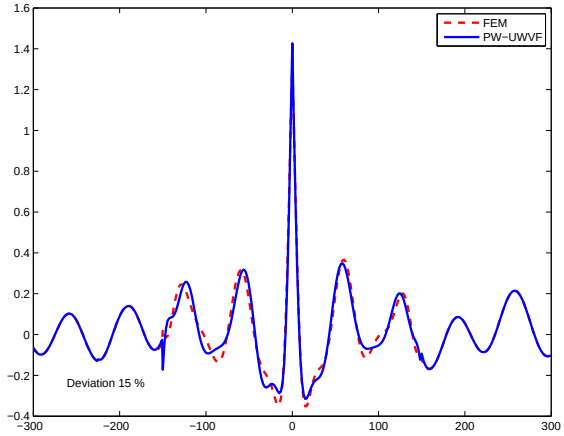

Figure 7: Plot of the real part of p on the top of the domain corresponding to the narrower peak gaussian.

properties in the same manner as in [1] is still to be done. From a theoretical point of view, explicit estimates in ω, for χ and A varying only in a "small" part of the domain Ω, like those for the standard Helmholtz equation in [21, 36], must be established. Such results would be based on an extension of the estimates established in [23].

- To face the instabilities raised by a singularity of the solution like the one displayed in Fig. 7, an approach, when the singularity is localized in a well-defined part of the domain, is to couple a polynomial UWVF in the vicinity of the singularity with the plane-wave UWVF elsewhere. This study is presently going on and will be presented in a fourthcoming publication. A second approach, based on using the plane-wave UWVF as a precontionner of a global polynomial UWVF can also be tried.

- It would be interesting to examine if the approach can be extended to other wave propagation problems such as in electromagnetism or elasticity. If the extension to elastic waves does not seem to be particularly problematic, this is not the case for electromagnetic waves due to the difficulty of adequately defining outgoing and incoming traces.

References

[1] M. Ainsworth, P. Monk, and W. Muniz, *Dispersive and dissipative properties of discontinuous Galerkin finite element methods for the second-order wave equation.*, Journal of Scientific Computing, 27 (2006), pp. 5–40.

[2] C. J. Alves and S. S. Valtchev, *Numerical comparison of two meshfree methods for acoustic wave scattering*, Engineering Analysis with Boundary Elements, 29 (2005), pp. 371–383.

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

[4] I. Babuška, F. Ihlenburg, E. T. Paik, and S. A. Sauter, *A generalized Finite Element Method for solving the Helmholtz equation in two dimensions with minimal pollution*, Comput. Meth. Appl. Mech. Engrg., 128 (1995), pp. 325–359.

[5] I. Babuška and S. Sauter, *Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers?*, SIAM Journal on Numerical Analysis, 34 (1997), pp. 2392–2423.

[6] H. Barucq, A. Bendali, J. Diaz, and S. Tordeux, *Local strategies for improving the conditioning of the plane-wave ultra-weak variational formulation*, Journal of Computational Physics, 441 (2021), p. 110449.

[7] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[8] A. Buffa and P. Monk, *Error estimates for the ultra weak variational formulation of the Helmholtz equation*, Mathematical Modelling and Numerical Analysis, 42 (2008), pp. 925–940.

[9] P. Castillo, B. Cockburn, I. Perugia, and D. Schötzau, *An a priori error analysis of the Local Discontinuous Galerkin method for elliptic problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1676–1706.

[10] O. Cessenat, *Application d'une nouvelle formulation variationnelle aux équations d'ondes harmoniques. Problèmes d'Helmholtz 2D et de Maxwell 3D.*, PhD thesis, University of Paris XI Dauphine, 1996.

[11] 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 (1998), pp. 255–299.

[12] O. Cessenat and B. Després, *Using plane waves as base functions for solving the time-harmonic equations with the ultra weak variational formulation*, J. Comp. Acous., 11 (2003), pp. 227–238.

[13] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North Holland, Amsterdam, 1978.

[14] R. Courant and D. Hilbert, *Methods of Mathematical Physics, Vol. 1*, Interscience John Wiley & Sons, 1953.

[15] F. Faucher, *Contributions to seismic full waveform inversion for time harmonic wave equations: stability estimates, convergence analysis, numerical experiments involving large scale optimization algorithms*, PhD thesis, Université de Pau et des Pays de l'Adour, 2017.

[16] G. Gabard, *Discontinuous Galerkin methods with plane waves for time-harmonic problems*, Journal of Computational Physics, 225 (2007), pp. 1961–1984.

[17] G. Gabard, P. Gamallo, and T. Huttunen, *A comparison of wave-based discontinuous Galerkin, ultra-weak and least-square methods for wave problems*, Int. J. Numer. Meth. Engng, 85 (2011), pp. 380–402.

[18] M. Gaborit, O. Dazel, P. Göranson, and G. Gabard, *Coupling of finite-element and plane waves discontinuous Galerkin methods for time-harmonic problems*, Int. J. Numer. Methods Eng., 116 (2018), pp. 487–503.

[19] G. Giorgiani, D. Modesto, S. Fernández-Méndez, and A. Huerta, *High-order continuous and discontinuous galerkin methods for wave problems*, Int. J. Numer. Meth. Fluids, 73 (2013), pp. 883–903.

[20] C. Gittelson and R. Hiptmair, *Dispersion analysis of plane wave discontinuous methods*, Int. J. Numer. Methods Eng., 98 (2014), pp. 313–323.

[21] C. J. Gittelson, R. Hiptmair, and I. Perugia, *Plane wave discontinuous Galerkin methods: analysis of the h-version*, Mathematical Modelling and Numerical Analysis, 43 (2009), pp. 297–331.

[22] P. Grisvard, *Singularities in boundary value problems*, Masson and Spinger-Verlag, 1992.

[23] U. Hetmaniuk, *Stability estimates for a class of Helmholtz problems*, Commun. Math. Sci., 5 (2007), pp. 665–678.

[24] 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 (2011), pp. 264–284.

[25] R. Hiptmair, A. Moiola, and I. Perugia, *Trefftz discontinuous Galerkin methods for acoustic scattering on locally refined meshes*, Applied Numerical Mathematics, 79 (2014), pp. 79–91.

[26] R. Hiptmair, A. Moiola, and I. Perugia, *Plane Wave Discontinuous Galerkin Methods: Exponential Convergence of the hp-version*, Foundations of Computational Mathematics, 16 (2016), pp. 637–675.

[27] R. Hiptmair, A. Moiola, and I. Perugia, *A Survey of Trefftz Methods for the Helmholtz Equation*, Springer International Publishing, 2016, ch. 8, pp. 237–278.

[28] M. S. Howe, *Acoustics of Fluid-Structure Interactions.*, Cambridge University Press., Cambridge, 1998.

[29] T. Huttunen, J. P. Kaipio, and P. Monk, *The perfectly matched layer for the ultra weak variational formulation of the 3D Helmholtz equation*, Int. J. Numer. Methods Eng., 61 (2004), pp. 1072–1092.

[30] T. Huttunen and P. Monk, *The use of plane waves to approximate wave propagation in anisotropic media*, Journal of Computational Mathematics, 25 (2007), pp. 350–367.

[31] L.-M. Imbert-Gérard and B. Després, *A generalized plane-wave numerical method fort smooth nonconstant coefficients*, IMA Journal of Numerical Analysis, 34 (2014), pp. 1072–1103.

[32] K. Lemrabet, *Régularité de la solution d'un problème de transmission*, J. Math. pures et appl., 56 (1977), pp. 1–38.

[33] R. J. LeVeque, *Finite-Volume Methods for Hyperbolic Problems*, Cambridge University Press, Cambridge, 2004.

[34] T. Luostari, T. Huttunen, and P. Monk, *Improvements for the ultra weak variational formulation*, Int. J. Numer. Meth. Engng, 94 (2013), pp. 598–624.

[35] A. Makhlouf, *Justification et Amélioration de Modèles d'Antennes Patch par la Méthode des Développements Asymptotiques Raccordés*, PhD thesis, INSA Toulouse, 2008.

[36] J. M. Melenk, A. Parsania, and S. Sauter, *General DG-Methods for highly indefinite Helmholtz problems*, J. Sci. Comput., 57 (2013), pp. 536–581.

[37] A. Moiola, *Trefftz-Discontinuous Galerkin Methods for Time-Harmonic Wave Problems*, PhD thesis, ETH Zurich, 2011.

[38] A. Moiola, R. Hiptmair, and I. Perugia, *Plane wave approximation of homogeneous Helmholtz solutions*, Z. Angew. Math. Phys., 62 (2011), pp. 809–837.

[39] P. Monk, J. Schöberl, and A. Sinwel, *Hybridizing Raviart-Thomas elements for the Helmholtz equation*, Electromagnetics, 30 (2010), pp. 149–176.

[40] P. Monk and D.-Q. Wang, *A least-squares method for the Helmholtz equation*, Comput. Meth. Appl. Mech. Engrg., 175 (1999), pp. 121–136.

[41] T. Muramatu, *On imbedding theorems for Sobolev spaces and some of their generalization*, Publ. RIMS, Kyoto Univ. Ser. A, 3 (1968), pp. 393–416.

[42] S. Nicaise and A.-M. Sändig, *General interface problems—i*, Mathematical Methods in the Applied Sciences, 17 (1994), pp. 395–429.

[43] ——, *General interface problems—ii*, Mathematical Methods in the Applied Sciences, 17 (1994), pp. 431–450.

[44] R. Seeley, *Pseudo-differential Operators, C.I.M.E. Summer School*, vol. 47, Springer, Berlin, Heidelberg, 2010.

[45] A. Vion and C. Geuzaine, *Double sweep preconditioner for optimized Schwarz methods applied to the Helmholtz problem.*, J. Comput. Phys., 266 (2014), pp. 171–190.