# Ultra-weak variational formulation for heterogeneous maxwell problem in the context of high performance computing

Sébastien Pernet, Margot Sirdey, Sébastien Tordeux

**ESAIM: PROCEEDINGS AND SURVEYS, Vol. ?, 2023, 1-10**
Editors: Will be set by the publisher

# ULTRA-WEAK VARIATIONAL FORMULATION FOR HETEROGENEOUS MAXWELL PROBLEM IN THE CONTEXT OF HIGH PERFORMANCE COMPUTING

Sébastien Pernet[1], Margot Sirdey[2] and Sébastien Tordeux[3]

**Abstract.** Electromagnetic simulations on large domains require a huge memory consumption. Domain decomposition methods, based on Trefftz methods, could be an answer to this issue. In this paper, we associate to heterogeneous three-dimensional Maxwell equations one variational formulation which can be obtained either by upwind fluxes or Riemann traces. We associate to this variational formulation an iterative Trefftz Krylov solver. The poor conditioning due to the use of plane wave basis functions is bypassed thanks to a compression strategy. Moreover, the developed iterative solver is accelerated thanks to a left preconditioner. The considered numerical cases illustrate the performance of this basis reduction, which leads to the consideration of an industrial case of more than 750 millions of degrees of freedom.

## Introduction

Electromagnetic waves are present in a wide panel of applications, such as transports, high-technology or medicine. Their numerical modeling remains a challenging topic when considering complex industrial geometries. Some numerical methods can simulate accurately electromagnetic fields such as Finite Element Methods (FEM) [29, 37, 42], or Finite Difference (FD) [54]. However, they face two main issues: the numerical dispersion and the memory limitation.

First, numerical dispersion, see [1, 2, 35, 36], appears when the size of the domain is large with respect to the wavelength. In such cases, wave phenomena are oscillating and non dissipative. Then, the number of discretisation points per wavelength increases with the size of the domain to accurately approximate the wave phase. Many methods may partially counter this problem, such as high-order FEM [13, 44], polynomial Discontinuous Galerkin (DG) [14, 19, 22, 28, 33, 59], Boundary Element Methods (BEM) [6, 45, 53, 58], integral equation collocation method [5], or Trefftz type methods [31, 49]. The latter are at main interest in this paper. They take into account local basis functions adapted to the electromagnetic phenomena such that they reduce numerical dispersion. In three dimensions, these methods are associated to large linear systems which are extremely costly to invert when using LU factorization. It is then needed to choose the method depending on the studied geometry. Condensation methods, such as Hybrid DG (HDG) methods [12, 43, 46] have been introduced to reduce the number of degrees of freedom. Nevertheless, these optimisations are not sufficient to simulate electromagnetic waves on domains whose size is larger than 30 wavelengths. For BEM, this issue has been solved thanks to compression methods like Fast Multipole [16, 17] and Adaptive Cross Approximation

[1] DTIS, ONERA
[2] Makutu, EPC INRIA, e2s-UPPA, LMAP UMR CNRS 5142 - DTIS, ONERA
[3] Makutu, EPC INRIA, e2s-UPPA, LMAP UMR CNRS 5142

methods [4, 38], for example. But, implementation of BEM in the case of heterogeneous structures is a really hard task.

When dealing with larger heterogeneous geometries, the indefinite property of the matrix associated to most volumic methods prevents the use of algebraic iterative inversion. Many authors propose Domain Decomposition Methods (DDM), see [20,57] for example, but DDM are often difficult to carry out in the context of High-Parallel Computing (HPC) architecture. The Ultra-Weak Variational Formulation (UWVF) methods [7,10,18,34,40,56], developed by Cessenat-Després in [9], are conversely naturally adapted to domain decomposition since they involve definite positive matrices. Consequently, UWVF is easily implementable in HPC framework. Trefftz methods often use plane wave local basis functions, see [24, 26, 27, 41], but we could have chose spherical Bessel functions [39]. Practically, the advantage of using such basis results in an easy implementation. But plane wave basis functions difficultly approximate high-complexity electromagnetic waves such as point effects at corners or trapped waves in elaborated heterogeneous configurations. Moreover, they often provide ill-conditioned linear systems, due to rounding errors caused by plane waves linear-dependency, see [15]. This may conduct to a lack of accuracy for the numerical solution. To avoid such issue, previous works have developed strategies, for which we propose some improvements here, see [15, 40]. For homogeneous cases, the convergence of the method is well established, see [30].

Here, we aim at developing an iterative Trefftz solver using few memory and leading to an accurate numerical solution. The developed method is called the heterogeneous Cessenat-Després UWVF. This numerical method resorts to Maxwell equations to model electromagnetic waves. In the present paper we study a dimensionless Maxwell problem which needs to be constructed. We recall the general Maxwell formulas in absence of charges and currents and for an isotropic linear medium

$$\left\{\begin{array}{lllll} \nabla\cdot\mathbf{d} &=& 0, & \nabla\times\mathbf{e} &=& -\dfrac{\partial\mathbf{b}}{\partial t}, \qquad \mathbf{d}=\varepsilon_0\varepsilon_r\mathbf{e},\\[2ex] \nabla\cdot\mathbf{b} &=& 0, & \nabla\times\mathbf{h} &=& \dfrac{\partial\mathbf{d}}{\partial t}, \qquad \mathbf{b}=\mu_0\mu_r\mathbf{h}, \end{array}\right.$$

where $\varepsilon_0$ (*resp.* $\varepsilon_r$ ) and $\mu_0$ (*resp.* $\mu_r$) are the permittivity (*resp.* relative permittivity) and the permeability (*resp.* relative permeability) of the vacuum (*resp.* of the medium). This system involves the electric and magnetic field intensities $\mathbf{e}$ and $\mathbf{h}$, the electric displacement $\mathbf{d}$ and the magnetic induction $\mathbf{b}$. We suppose they are all time-harmonic fields. Therefore, they can be represented by four complex valued normalised functions $\mathbf{E}$, $\mathbf{H}$, $\mathbf{D}$ and $\mathbf{B}$ which are associated to their normalisation amplitudes $e_0$, $d_0=\varepsilon_0 e_0$, $h_0=\sqrt{\varepsilon_0/\mu_0}\,e_0$ and $b_0=\sqrt{\varepsilon_0\mu_0}\,e_0$,

$$\left\{\begin{array}{ll} \mathbf{e}(\mathbf{x},t)= & e_0\mathcal{R}\left(e^{i\omega t}\mathbf{E}(\mathbf{x})\right), \qquad \mathbf{h}(\mathbf{x},t)= \quad h_0\mathcal{R}\left(e^{i\omega t}\mathbf{H}(\mathbf{x})\right),\\[2ex] \mathbf{d}(\mathbf{x},t)= & d_0\mathcal{R}\left(e^{i\omega t}\mathbf{D}(\mathbf{x})\right), \qquad \mathbf{b}(\mathbf{x},t)= \quad b_0\mathcal{R}\left(e^{i\omega t}\mathbf{B}(\mathbf{x})\right), \end{array}\right.$$

where $\omega$ is the angular frequency that accounts for time-harmonic dependency. We define the wavenumber $k_0=\omega/c_0$, where $c_0=(\epsilon_0\mu_0)^{-\frac{1}{2}}$ is the velocity in the vacuum. We get the following Maxwell normalised problem defined on a connex Lipschitz domain $\Omega\subset\mathbb{R}^3$

$$\nabla\times\mathbf{H}=ik_0\,\varepsilon_r\mathbf{E} \quad\text{and}\quad \nabla\times\mathbf{E}=-ik_0\,\mu_r\mathbf{H}, \qquad \text{on } \Omega, \tag{1}$$

where both fields are in the space $H(\nabla\times,\Omega)$ defined by

$$H(\nabla\times,\Omega):=\left\{\mathbf{u}:\Omega\to\mathbb{C}^3,\int_\Omega|\mathbf{u}|^2\,\mathrm{d}\mathbf{x}<\infty,\int_\Omega|\nabla\times\mathbf{u}|^2\,\mathrm{d}\mathbf{x}<\infty\right\}.$$

The unknown of the problem is denoted by $\mathbb{E} := (\mathbf{E}, \mathbf{H})$. We complete by an impedance boundary condition defined on the domain boundary $\partial\Omega$

$$\gamma_t \mathbf{E} + Z_{\partial\Omega}\, \mathbf{n} \times \gamma_t \mathbf{H} = \mathbf{g} \quad \text{on } \partial\Omega,$$

where we have the following definitions and notations:

- $\mathbf{n} \in \mathbb{R}^3$ is the outward unit normal to $\partial\Omega$,
- $Z_{\partial\Omega} : \partial\Omega \to \mathbb{R}^+$ is a piecewise constant function, with a strictly non-positive real part,
- $\gamma_t \mathbf{E} = \mathbf{E} - (\mathbf{E} \cdot \mathbf{n})\, \mathbf{n} = (\mathbf{n} \times \mathbf{E}) \times \mathbf{n}$ is the tangential component of the electric field,
- $\gamma_t \mathbf{H} = \mathbf{H} - (\mathbf{H} \cdot \mathbf{n})\, \mathbf{n} = (\mathbf{n} \times \mathbf{H}) \times \mathbf{n}$ is the tangential component of the magnetic field,
- $\varepsilon_r$ and $\mu_r$ are piecewise constant functions defined on a partition of $\Omega$. The material parameters are assumed to be constant over each mesh cell.
- $\mathbf{g} : \partial\Omega \to \mathbb{C}^3$ is a purely tangential field in the following functional space

$$L^2_t(\partial\Omega) := \Big\{ \mathbf{u} \in \big(L^2(\partial\Omega)\big)^3, \quad \mathbf{u} \cdot \mathbf{n} = 0 \Big\}.$$

There exists a unique solution of problem (1), see [42]. The main goal of this paper is to develop a numerical method approximating accurately the solution to the Maxwell problem, with a low memory cost. In Section 1, we introduce the mesh associated to the domain $\Omega$ and its properties. Afterwards, we design an heterogeneous plane wave mixed DG Trefftz variational formulation based on upwind fluxes; *mixed* in the sense that it involves both the electric and the magnetic fields. The well-posedness of the associated problem is proved thanks to a weak coercivity property. The well-posedness of the associated problem relies on the positivity of the variational formulation. In Section 2, we bring to the fore the equivalence between the mixed DG Trefftz formulation and the heterogeneous Cessenat-Després UWVF. Then, the Trefftz UWVF problem is discretised, leading to a matricial linear system, which can be inverted with LU solver. However, due to the memory challenge, we opt for an iterative Trefftz UWVF algorithm based on a singular regular decomposition of the matrix. In Section 4, we associate to the constructed iterative scheme a Krylov type method. The latter Krylov method, combined with a preconditioning strategy, is proved to be convergent thanks to Galerkin theory. Thereafter, we set up a compression strategy reducing the number of plane wave basis functions in the discrete Galerkin space, leading to a memory cost reduction. Finally, a global preconditioning aims at accelerating the Trefftz Krylov UWVF solver.

## 1. Mixed discontinuous Galerkin Trefftz formulation

In this section, we devise a mixed DG Trefftz problem based on upwind fluxes. This goes through the introduction of Trefftz spaces and the reciprocity formula. Finally, we establish a weak coercivity property and show the well-posedness of the associated variational formulation.

### 1.1. Mesh properties and Trefftz spaces

We consider a three-dimensional mesh made of non-overlapping polyhedral elements $T$ meshing the computational domain $\Omega$. Indeed, we choose a mesh that follows the partitions $\Omega_p$, $1 \leq p \leq P$, of $\Omega$, in which $\varepsilon_r$ and $\mu_r$ are constant. It means that there exists a unique $1 \leq p_0 \leq P$ such that $T \subset \Omega_{p_0}$, leading to piecewise constant functions $\varepsilon_r$ and $\mu_r$. We denote by $\mathcal{T}$ the set of elements $T$ and by $\mathcal{F}$ the set of faces $F$ in $\Omega$. We define the following sets of faces:

- The set $\mathcal{F}_{\text{int}}$ of interior faces

$$\mathcal{F}_{\text{int}} := \{\partial T \cap \partial K \,:\, T, K \in \mathcal{T} \text{ with } T \neq K \text{ and } \mathit{area}(\partial T \cap \partial K) \neq 0\},$$

  where $\mathit{area}(I)$ refers to the area of the face $I$ (that is zero for edges and vertices).

- The set $\mathcal{F}_{\rm ext}$ of exterior faces

$$\mathcal{F}_{\rm ext} := \{\partial T \cap \partial\Omega \,:\, T \in \mathcal{T} \text{ and } \mathit{area}(\partial T \cap \partial\Omega) \neq 0\}.$$

- The set $\mathcal{F}_T$ of faces associated to an element $T \in \mathcal{T}$

$$\mathcal{F}_T := \{F \in \mathcal{F}_{\rm int} \cup \mathcal{F}_{\rm ext} \,:\, \mathit{area}(F \cap \partial T) \neq 0\}.$$

We denote by $\mathbb{X}_T$ the local Trefftz space defined for $T \in \mathcal{T}$ as the set of functions

$$\mathbb{E}^T := (\mathbf{E}^T, \mathbf{H}^T) \in H(\nabla\times, T) \times H(\nabla\times, T) \text{ satisfying} \tag{2a}$$

$$\nabla \times \mathbf{H}^T = ik_0\,\varepsilon_r \mathbf{E}^T \quad \text{and} \quad \nabla \times \mathbf{E}^T = -ik_0\,\mu_r \mathbf{H}^T, \quad \text{on } T, \tag{2b}$$

$$\gamma_t \mathbf{E}^T \in L^2_t(\partial T) \text{ and } \gamma_t \mathbf{H}^T \in L^2_t(\partial T), \tag{2c}$$

where $\mathbf{E}^T$ and $\mathbf{H}^T$ are the restrictions of the electric field $\mathbf{E}$ and the magnetic field $\mathbf{H}$ on the element $T$.

**Remark 1.1.** *The elliptic regularity theory seems to restrict the numerical method of this paper to regular solutions, ie $\mathbf{E}^T, \mathbf{H}^T \in H^{\frac{3}{2}}(T)$. However, we did not face any difficulty to apply the present method to less regular solutions.*

The space $\mathbb{X}_T$ is equipped with the inner product parametrised by $\alpha > 0$ and $\beta > 0$

$$(\mathbb{E}^T, \mathbb{E}'^T)_{\mathbb{X}_T} := \int_{\partial T} \left(\alpha\,\gamma_t \mathbf{E}^T \cdot \overline{\gamma_t \mathbf{E}'^T} + \beta\,\gamma_t \mathbf{H}^T \cdot \overline{\gamma_t \mathbf{H}'^T}\right) \mathsf{ds}_{\mathbf{x}},$$

where $\mathsf{ds}_{\mathbf{x}}$ is the Lebesgue measure associated to surfaces of the three-dimensional mesh.

**Remark 1.2.** *Due to the unique continuation theorem, the pairing $(\cdot,\cdot)_{\mathbb{X}_T}$ is an inner product which means*

$$(\mathbb{E}^T, \mathbb{E}^T)_{\mathbb{X}_T} \;=\; 0 \quad \Longrightarrow \quad \gamma_t \mathbf{E}^T = 0 \text{ and } \gamma_t \mathbf{H}^T = 0 \text{ on } \partial T \quad \underset{\text{(2b)}}{\Longrightarrow} \quad \mathbb{E}^T = (\mathbf{E}^T, \mathbf{H}^T) = 0 \text{ in } T.$$

It leads to the global discontinuous Trefftz space $\mathbb{X}_{\mathcal{T}}$ defined element by element by

$$\mathbb{X}_{\mathcal{T}} := \prod_{T\in\mathcal{T}} \mathbb{X}_T, \text{ equipped with the norm } (\mathbb{E}, \mathbb{E}')_{\mathbb{X}_{\mathcal{T}}} := \sum_{T\in\mathcal{T}} (\mathbb{E}^T, \mathbb{E}'^T)_{\mathbb{X}_T}.$$

**Remark 1.3.** *Every function $\mathbb{E} = (\mathbb{E}^T)_{T\in\mathcal{T}} \in \mathbb{X}_{\mathcal{T}}$ is associated to a function defined on $\Omega$ whose restriction to $T$ is $\mathbb{E}^T$. Each $\mathbb{E}^T = (\mathbf{E}^T, \mathbf{H}^T)$ satisfies* (2b) *and $\mathbb{E}$ is generally discontinuous across faces.*

## 1.2. Trefftz reciprocity formula

We introduce the reciprocity formula, also called the virtual work formula, which involves the restrictions on $T \in \mathcal{T}$ of the tangential component of the electric field and the tangential trace of the magnetic field

$$\gamma_t \mathbf{E}^T := \left(\mathbf{n}_T \times \mathbf{E}^T\right) \times \mathbf{n}_T \quad \text{and} \quad \gamma_\times^T \mathbf{H}^T := \mathbf{n}_T \times \gamma_t \mathbf{H}^T, \tag{3}$$

where $\mathbf{n}_T \in \mathbb{R}^3$ is the outward unit normal to $\partial T$.

**Proposition 1.1.** *For $\mathbb{E} \in \mathbb{X}_{\mathcal{T}}$ and $\mathbb{E}' \in \mathbb{X}_{\mathcal{T}}$, the global reciprocity formula is*

$$r(\mathbb{E}, \mathbb{E}') := \sum_{T\in\mathcal{T}} r_T(\mathbb{E}, \mathbb{E}') \quad \textit{with} \quad r_T(\mathbb{E}, \mathbb{E}') := \int_{\partial T} \left(\gamma_\times^T \mathbf{H}^T \cdot \overline{\gamma_t \mathbf{E}'^T} + \gamma_t \mathbf{E}^T \cdot \gamma_\times^T \overline{\mathbf{H}'^T}\right) \mathsf{ds}_{\mathbf{x}} = 0. \tag{4}$$

*Proof.* The reciprocity formula takes the following form, with $\mathbb{E}^T = \left(\mathbf{E}^T, \mathbf{H}^T\right)$ and $\mathbb{E}'^T = \left(\mathbf{E}'^T, \mathbf{H}'^T\right)$ in $\mathbb{X}_T$,

$$W_T = ik_0 \int_T \varepsilon_r \mathbf{E}^T \cdot \overline{\mathbf{E}'^T} + \mu_r \mathbf{H}^T \cdot \overline{\mathbf{H}'^T} \, \mathrm{d}T, \quad \text{for } T \in \mathcal{T}.$$

Since both the solution $\mathbb{E}^T \in \mathbb{X}_T$ and the test function $\mathbb{E}'^T \in \mathbb{X}_T$ satisfy (2), we have

$$W_T = \int_T \nabla \times \mathbf{H}^T \cdot \overline{\mathbf{E}'^T} - \nabla \times \mathbf{E}^T \cdot \overline{\mathbf{H}'^T} \, \mathrm{d}T = \int_T \mathbf{H}^T \cdot \overline{\nabla \times \mathbf{E}'^T} - \mathbf{E}^T \cdot \overline{\nabla \times \mathbf{H}'^T} \, \mathrm{d}T.$$

By subtracting these two last expressions of $W_T$, we have

$$\int_T \nabla \times \mathbf{H}^T \cdot \overline{\mathbf{E}'^T} - \mathbf{H}^T \cdot \overline{\nabla \times \mathbf{E}'^T} \, \mathrm{d}T + \int_T \mathbf{E}^T \cdot \overline{\nabla \times \mathbf{H}'^T} - \nabla \times \mathbf{E}^T \cdot \overline{\mathbf{H}'^T} \, \mathrm{d}T = 0.$$

Due to the Stokes formula (see [42]), we get the local reciprocity formula

$$r_T(\mathbb{E}, \mathbb{E}') = \int_{\partial T} (\mathbf{n}_T \times \mathbf{H}^T) \cdot \overline{\mathbf{E}'^T} + \mathbf{E}^T \cdot \left(\mathbf{n}_T \times \overline{\mathbf{H}'^T}\right) \mathsf{ds}_{\mathbf{x}} = 0, \quad \text{for } \mathbb{E} \text{ and } \mathbb{E}' \text{ in } \mathbb{X}_{\mathcal{T}}.$$

We end the proof by using definitions (3). □

**Remark 1.4.** *We remark that $\partial T$ can be decomposed into faces $F \in \mathcal{F}_T$ leading to*

$$\sum_{T \in \mathcal{T}} \int_{\partial T} f^T \, \mathsf{ds}_{\mathbf{x}} \; = \; \sum_{T \in \mathcal{T}} \sum_{F \in \mathcal{F}_T} \int_F f^T \, \mathsf{ds}_{\mathbf{x}}, \tag{5}$$

*for any piecewise continuous regular function $f$ whose restriction on one element $T \in \mathcal{T}$ is denoted by $f^T$. The reciprocity formula* (4) *is then equivalent to*

$$r_T(\mathbb{E}, \mathbb{E}') := \sum_{F \in \mathcal{F}_T} \int_F \left(\gamma_\times^T \mathbf{H}^T \cdot \overline{\gamma_t \mathbf{E}'^T} + \gamma_t \mathbf{E}^T \cdot \gamma_\times^T \overline{\mathbf{H}'^T}\right) \mathsf{ds}_{\mathbf{x}} = 0.$$

### 1.3. **Mixed Discontinuous Galerkin Trefftz variational formulation**

Following the same approach than for polynomial mixed DG methods, see [22, 32], the mixed DG Trefftz formulation is deduced by inserting upwind fluxes, see [47], into the reciprocity formula (4).

**Problem 1.** *Find $\mathbb{E} \in \mathbb{X}_{\mathcal{T}}$ such that for all $\mathbb{E}' \in \mathbb{X}_{\mathcal{T}}$*

$$\sum_{T \in \mathcal{T}} \sum_{F \in \mathcal{F}_T} \int_F \left(\widehat{\gamma_t \mathbf{E}} \cdot \gamma_\times^T \overline{\mathbf{H}'^T} + \widehat{\gamma_\times^T \mathbf{H}} \cdot \overline{\gamma_t \mathbf{E}'^T}\right) \mathsf{ds}_{\mathbf{x}} = 0, \tag{6}$$

*with interior upwind fluxes for $F \in \mathcal{F}_{\text{int}}$ separating two elements $T$ and $K$ defined as*

$$\left(\widehat{\gamma_t \mathbf{E}}\right)_{|F} := \frac{Z_K}{Z_K + Z_T} \gamma_{\text{out}}^T \mathbb{E}^T + \frac{Z_T}{Z_K + Z_T} \gamma_{\text{out}}^K \mathbb{E}^K \quad \textit{and} \quad \left(\widehat{\gamma_\times^T \mathbf{H}}\right)_{|F} := -\frac{1}{Z_K + Z_T} \gamma_{\text{out}}^T \mathbb{E}^T + \frac{1}{Z_K + Z_T} \gamma_{\text{out}}^K \mathbb{E}^K, \tag{7}$$

*and boundary upwind fluxes for $F \in \mathcal{F}_{\text{ext}}$ defined as*

$$\left(\widehat{\gamma_t \mathbf{E}}\right)_{|F} := \frac{Z_{\partial\Omega}}{Z_{\partial\Omega} + Z_T} \gamma_{\text{out}}^T \mathbb{E}^T + \frac{Z_T}{Z_{\partial\Omega} + Z_T} \mathbf{g}^T \quad \textit{and} \quad \left(\widehat{\gamma_\times^T \mathbf{H}}\right)_{|F} := -\frac{1}{Z_{\partial\Omega} + Z_T} \gamma_{\text{out}}^T \mathbb{E}^T + \frac{1}{Z_{\partial\Omega} + Z_T} \mathbf{g}^T, \tag{8}$$

*where the general incoming and outgoing wave trace operators* $\gamma^T_{\mathrm{in/out}} : \mathbb{X}_T \to L^2_t(\partial T)$ *are defined by*

$$\gamma^T_{\mathrm{in}}\mathbb{E}^T := \gamma_t\mathbf{E}^T + Z_T\gamma^T_\times\mathbf{H}^T \quad \text{and} \quad \gamma^T_{\mathrm{out}}\mathbb{E}^T := \gamma_t\mathbf{E}^T - Z_T\gamma^T_\times\mathbf{H}^T. \tag{9}$$

**Remark 1.5.** *For this particular choice of numerical fluxes* (7) *and* (8)*, the exact solution satisfies*

$$\widehat{(\gamma_t\mathbf{E})}_{|F} = (\gamma_t\mathbf{E}^T)_{|F} = (\gamma_t\mathbf{E}^K)_{|F} \quad \text{and} \quad \widehat{(\gamma^T_\times\mathbf{H})}_{|F} = (\gamma^T_\times\mathbf{H}^T)_{|F} = -(\gamma^K_\times\mathbf{H}^K)_{|F}.$$

*Therefore, the variational formulation of Problem 1 encodes the continuity of its solution.*

**Remark 1.6.** *An exterior face* $F \in \mathcal{F}_{\mathrm{ext}}$ *can be seen as a face interfacing an element* $T$ *and a virtual element exterior to the domain. More precisely, denoting by* $Z_{\partial\Omega} = Z_K$ *and* $\mathbf{g}^T = \gamma^K_{\mathrm{out}}\mathbb{E}^K$*,* (8) *can be seen as* (7)*.*

Problem 1 can be interpreted as a variational problem and leads to the following proposition.

**Proposition 1.2.** *Problem 1 can be written as*

$$\text{Find } \mathbb{E} \in \mathbb{X}_{\mathcal{T}} \text{ such that for all } \mathbb{E}' \in \mathbb{X}_{\mathcal{T}} \text{ we have} \quad a(\mathbb{E},\mathbb{E}') = \ell(\mathbb{E}'), \tag{10}$$

*where the sesquilinear form* $a$ *and the linear form* $\ell$ *are given by*

$$a(\mathbb{E},\mathbb{E}') := \sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T}\int_F \left(a^{EE'}_{T,F} + a^{EH'}_{T,F} + a^{HE'}_{T,F} + a^{HH'}_{T,F}\right)\mathsf{ds}_{\mathbf{x}}, \tag{11a}$$

$$\ell(\mathbb{E}') := \sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T\cap\mathcal{F}_{\mathrm{ext}}}\int_F \left(\ell^{E'}_{T,F} + \ell^{H'}_{T,F}\right)\mathsf{ds}_{\mathbf{x}}, \tag{11b}$$

*where we have if* $F \in \mathcal{F}_{\mathrm{int}}$ *separating two neighboring elements* $T$ *and* $K$

$$\begin{cases} a^{EE'}_{T,F} := \dfrac{1}{Z_T+Z_K}(\gamma_t\mathbf{E}^T - \gamma_t\mathbf{E}^K)\cdot\overline{\gamma_t\mathbf{E}'^T}, & a^{EH'}_{T,F} := \dfrac{Z_T}{Z_T+Z_K}(\gamma_t\mathbf{E}^T - \gamma_t\mathbf{E}^K)\cdot\overline{\gamma^T_\times\mathbf{H}'^T}, \\ a^{HE'}_{T,F} := \dfrac{Z_K}{Z_T+Z_K}(\gamma^T_\times\mathbf{H}^T - \gamma^T_\times\mathbf{H}^K)\cdot\overline{\gamma_t\mathbf{E}'^T}, & a^{HH'}_{T,F} := \dfrac{Z_T\,Z_K}{Z_T+Z_K}(\gamma^T_\times\mathbf{H}^T - \gamma^T_\times\mathbf{H}^K)\cdot\overline{\gamma^T_\times\mathbf{H}'^T}, \end{cases} \tag{12a}$$

*and if* $F \in \mathcal{F}_{\mathrm{ext}}$

$$\begin{cases} a^{EE'}_{T,F} := \dfrac{1}{Z_T+Z_{\partial\Omega}}\gamma_t\mathbf{E}^T\cdot\overline{\gamma_t\mathbf{E}'^T}, & a^{EH'}_{T,F} := \dfrac{Z_T}{Z_T+Z_{\partial\Omega}}\gamma_t\mathbf{E}^T\cdot\overline{\gamma^T_\times\mathbf{H}'^T}, \\ a^{HE'}_{T,F} := \dfrac{Z_{\partial\Omega}}{Z_T+Z_{\partial\Omega}}\gamma^T_\times\mathbf{H}^T\cdot\overline{\gamma_t\mathbf{E}'^T}, & a^{HH'}_{T,F} := \dfrac{Z_T\,Z_{\partial\Omega}}{Z_T+Z_{\partial\Omega}}\gamma^T_\times\mathbf{H}^T\cdot\overline{\gamma^T_\times\mathbf{H}'^T}, \\ \ell^{E'}_{T,F} := \dfrac{1}{Z_T+Z_{\partial\Omega}}\mathbf{g}^T\cdot\overline{\gamma_t\mathbf{E}'^T}, & \ell^{H'}_{T,F} := \dfrac{Z_T}{Z_T+Z_{\partial\Omega}}\mathbf{g}^T\cdot\overline{\gamma^T_\times\mathbf{H}'^T}. \end{cases} \tag{12b}$$

*Proof.* Problem 1 reads

$$\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T}\int_F \mathcal{I}_{T,F}\,\mathsf{ds}_{\mathbf{x}} = 0, \quad \text{with} \quad \mathcal{I}_{T,F} := \widehat{\gamma_t\mathbf{E}}\cdot\overline{\gamma^T_\times\mathbf{H}'^T} + \widehat{\gamma^T_\times\mathbf{H}}\cdot\overline{\gamma_t\mathbf{E}'^T}.$$

$(a)$ If $F \in \mathcal{F}_{\mathrm{int}}$, using the definition of the numerical traces (7), we have

$$\mathcal{I}_{T,F} = \Big(\frac{Z_K}{Z_K+Z_T}\gamma^T_{\mathrm{out}}\mathbb{E}^T + \frac{Z_T}{Z_T+Z_K}\gamma^K_{\mathrm{out}}\mathbb{E}^K\Big)\cdot\overline{\gamma^T_\times\mathbf{H}'^T} + \Big(-\frac{1}{Z_K+Z_T}\gamma^T_{\mathrm{out}}\mathbb{E}^T + \frac{1}{Z_K+Z_T}\gamma^K_{\mathrm{out}}\mathbb{E}^K\Big)\cdot\overline{\gamma_t\mathbf{E}'^T}.$$

Taking into account the definition of the outgoing trace operator in (9), we get

$$\mathcal{I}_{T,F} = \Big[\frac{Z_K}{Z_K+Z_T}\big(\gamma_t\mathbf{E}^T - Z_T\gamma_\times^T\mathbf{H}^T\big) + \frac{Z_T}{Z_T+Z_K}\big(\gamma_t\mathbf{E}^K + Z_K\gamma_\times^T\mathbf{H}^K\big)\Big]\cdot\overline{\gamma_\times^T\mathbf{H}'^T} + \Big[-\frac{1}{Z_K+Z_T}\big(\gamma_t\mathbf{E}^T - Z_T\gamma_\times^T\mathbf{H}^T\big) + \frac{1}{Z_K+Z_T}\big(\gamma_t\mathbf{E}^K + Z_K\gamma_\times^T\mathbf{H}^K\big)\Big]\cdot\overline{\gamma_t\mathbf{E}'^T}.$$

Expanding, we have

$$\mathcal{I}_{T,F} = \frac{\big(Z_K\gamma_t\mathbf{E}^T + Z_T\gamma_t\mathbf{E}^K\big)}{Z_K+Z_T}\cdot\overline{\gamma_\times^T\mathbf{H}'^T} + \frac{\big(Z_T\gamma_\times^T\mathbf{H}^T + Z_K\gamma_\times^T\mathbf{H}^K\big)}{Z_K+Z_T}\cdot\overline{\gamma_t\mathbf{E}'^T} - a_{T,F}^{EE'} - a_{T,F}^{HH'}.$$

We then remark that $\dfrac{Z_K}{Z_K+Z_T} + \dfrac{Z_T}{Z_K+Z_T} = 1$ and $\dfrac{Z_{\partial\Omega}}{Z_{\partial\Omega}+Z_T} + \dfrac{Z_T}{Z_{\partial\Omega}+Z_T} = 1$. This leads to

$$\mathcal{I}_{T,F} = \gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}'^T} + \gamma_\times^T\mathbf{H}^T\cdot\overline{\gamma_t\mathbf{E}'^T} - a_{T,F}^{EE'} - a_{T,F}^{EH'} - a_{T,F}^{HE'} - a_{T,F}^{HH'}. \tag{13}$$

($b$) If $F\in\mathcal{F}_{\rm ext}$, we have in the same way

$$\mathcal{I}_{T,F} = \gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}'^T} + \gamma_\times^T\mathbf{H}^T\cdot\overline{\gamma_t\mathbf{E}'^T} \;+\; \ell_{T,F}^{E'} \;+\; \ell_{T,F}^{H'} - a_{T,F}^{EE'} - a_{T,F}^{EH'} - a_{T,F}^{HE'} - a_{T,F}^{HH'}. \tag{14}$$

Proposition 1.2 follows from (13), (14) and the reciprocity formula (4). □

### 1.4. **Positivity and injectivity properties**

The well-posedness of Problem 1 relies on the positivity of the sesquilinear form $a$ defined by (11a). It implies the study of the problem stability through interfaces $F\in\mathcal{F}_{\rm int}$ involving the jump of the tangential trace (*resp.* component) of the electric (*resp.* magnetic) field, respectively defined as

$$[\![\gamma_t\mathbf{E}]\!]_F := \gamma_t\mathbf{E}^T - \gamma_t\mathbf{E}^K \quad\text{and}\quad [\![\gamma_\times\mathbf{H}]\!]_F := \gamma_\times^T\mathbf{H}^T - \gamma_\times^T\mathbf{H}^K. \tag{15}$$

**Proposition 1.3.** *The sesquilinear form $a$ is positive since $\Re(a(\mathbb{E},\mathbb{E}))\geq 0$. We define the* DG *norm as*

$$\|\mathbb{E}\|_{\rm DG} = \sqrt{\Re(a(\mathbb{E},\mathbb{E}))}, \quad \textit{for all} \quad \mathbb{E}\in\mathbb{X}_{\mathcal{T}}, \tag{16}$$

*for which we have the following properties*

(i) $\|\mathbb{E}\|_{\rm DG}^2 = \|\mathbb{E}\|_{\rm int}^2 + \|\mathbb{E}\|_{\rm ext}^2$ *with*

$$\|\mathbb{E}\|_{\rm int}^2 = \sum_{F\in\mathcal{F}_{\rm int}}\int_F\Big(\frac{1}{Z_T+Z_K}[\![\gamma_t\mathbf{E}]\!]_F\cdot\overline{[\![\gamma_t\mathbf{E}]\!]_F} \;+\; \frac{Z_T\,Z_K}{Z_T+Z_K}[\![\gamma_\times\mathbf{H}]\!]_F\cdot\overline{[\![\gamma_\times\mathbf{H}]\!]_F}\Big)\,\mathsf{ds}_{\mathbf{x}},$$

$$\|\mathbb{E}\|_{\rm ext}^2 = \sum_{F\in\mathcal{F}_{\rm ext}}\int_F\Big(\frac{1}{Z_T+Z_{\partial\Omega}}\gamma_t\mathbf{E}\cdot\overline{\gamma_t\mathbf{E}} \;+\; \frac{Z_T\,Z_{\partial\Omega}}{Z_T+Z_{\partial\Omega}}\gamma_t\mathbf{H}\cdot\overline{\gamma_t\mathbf{H}}\Big)\,\mathsf{ds}_{\mathbf{x}}.$$

(ii) $a(\mathbb{E},\mathbb{E}) = 0 \quad\Longrightarrow\quad \mathbb{E}\equiv 0.$

The proof of Proposition 1.3 resorts to the following remark.

**Remark 1.7.** *Each face $F\in\mathcal{F}_{\rm int}$ is involved twice in the sum* (5) *of Remark 1.4. Thus, we can decompose the integral on $\partial T$ by using a "face point of view"*

$$\sum_{T\in\mathcal{T}}\int_{\partial T} f^T\,\mathsf{ds}_{\mathbf{x}} = \sum_{F\in\mathcal{F}_{\rm ext}}\int_F f^T\,\mathsf{ds}_{\mathbf{x}} + \sum_{F\in\mathcal{F}_{\rm int}}\int_F f^T + f^K\,\mathsf{ds}_{\mathbf{x}},$$

*where, in the right-hand side, $T$ (resp. $T$ and $K$) is one element (resp. are two elements) with one face $F$.*

*Proof.* (i) Using (11a), we can write the real part of $a$ as

$$\Re\big(a(\mathbb{E},\mathbb{E})\big) = \sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T}\int_F \Re\Big(a_{T,F}^{EE}+a_{T,F}^{EH}+a_{T,F}^{HE}+a_{T,F}^{HH}\Big)\,\mathsf{ds}_{\mathbf{x}}.$$

($a$) We first show that

$$\mathcal{I}^1 \;=\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T}\int_F \underbrace{\Re\big(a_{T,F}^{EH}+a_{T,F}^{HE}\big)}_{:=\mathcal{I}^1_{T,F}}\,\mathsf{ds}_{\mathbf{x}} = 0. \tag{17}$$

We evaluate $\mathcal{I}^1_{T,F}$ by distinguishing the cases of an exterior face $F\in\mathcal{F}_{\rm ext}$ and an interior face $F\in\mathcal{F}_{\rm int}$. For $F\in\mathcal{F}_{\rm int}$, using the definitions of forms (12a) and taking into account that $\Re(x\overline{y})=\Re(y\overline{x})$, we have

$$\begin{aligned}\mathcal{I}^1_{T,F} &= \Re\Big(\frac{Z_T}{Z_T+Z_K}\big(\gamma_t\mathbf{E}^T-\gamma_t\mathbf{E}^K\big)\cdot\overline{\gamma_\times^T\mathbf{H}^T}+\frac{Z_K}{Z_T+Z_K}\gamma_t\mathbf{E}^T\cdot\overline{\big(\gamma_\times^T\mathbf{H}^T-\gamma_\times^T\mathbf{H}^K\big)}\Big),\\ &= \Re\Big(\gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}^T}-\frac{Z_T}{Z_T+Z_K}\gamma_t\mathbf{E}^K\cdot\overline{\gamma_\times^T\mathbf{H}^T}-\frac{Z_K}{Z_T+Z_K}\gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}^K}\Big).\end{aligned}$$

In a similar way, using (12b) for $F\in\mathcal{F}_{\rm ext}$, we get

$$\mathcal{I}^1_{T,F} \;=\; \Re\Big(\frac{Z_T}{Z_T+Z_{\partial\Omega}}\gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}^T}+\frac{Z_{\partial\Omega}}{Z_T+Z_{\partial\Omega}}\gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}^T}\Big) \;=\; \Re\big(\gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}^T}\big).$$

Since $\Re\big(r_T(\mathbb{E},\mathbb{E})\big)=2\displaystyle\sum_{F\in\mathcal{F}_T}\int_F\Re\big(\gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}^T}\big)$, see Remark 1.4, and summing over $F\in\mathcal{F}_T$ we get

$$\sum_{F\in\mathcal{F}_T}\int_F\mathcal{I}^1_{T,F}\,\mathsf{ds}_{\mathbf{x}} \;=\; \frac{1}{2}\underbrace{\Re\big(r_T(\mathbb{E},\mathbb{E})\big)}_{=0\text{ see (4)}}-\sum_{F\in\mathcal{F}_T\cap\mathcal{F}_{\rm int}}\int_F\Re\Big(\frac{Z_T}{Z_T+Z_K}\gamma_t\mathbf{E}^K\cdot\overline{\gamma_\times^T\mathbf{H}^T}+\frac{Z_K}{Z_T+Z_K}\gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}^K}\Big)\,\mathsf{ds}_{\mathbf{x}}.$$

Summing over elements, we then take the face point of view of Remark 1.7

$$\mathcal{I}^1=-\sum_{F\in\mathcal{F}_{\rm int}}\int_F\Re\Big(\frac{Z_T}{Z_T+Z_K}\big(\gamma_t\mathbf{E}^K\cdot\overline{\gamma_\times^T\mathbf{H}^T}+\gamma_t\mathbf{E}^K\cdot\overline{\gamma_\times^K\mathbf{H}^T}\big)+\frac{Z_K}{Z_T+Z_K}\big(\gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}^K}+\gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^K\mathbf{H}^K}\big)\Big)\,\mathsf{ds}_{\mathbf{x}}.$$

Remarking that $\gamma_\times^T+\gamma_\times^K=0$, we finally have (17).

($b$) It remains to evaluate

$$\mathcal{I}^2 \;=\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T}\int_F\Re(a_{T,F}^{EE})\,\mathsf{ds}_{\mathbf{x}}\quad\text{and}\quad\mathcal{I}^3\;=\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T}\int_F\Re(a_{T,F}^{HH})\,\mathsf{ds}_{\mathbf{x}}.$$

We use again (12) and sum over elements, thus differentiating interior faces and exterior faces

$$\mathcal{I}^2=\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T\cap\mathcal{F}_{\rm int}}\int_F\Re\Big(\frac{1}{Z_T+Z_K}\big(\gamma_t\mathbf{E}^T-\gamma_t\mathbf{E}^K\big)\cdot\overline{\gamma_t\mathbf{E}^T}\Big)\,\mathsf{ds}_{\mathbf{x}}+\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T\cap\mathcal{F}_{\rm ext}}\int_F\Re\Big(\frac{1}{Z_T+Z_{\partial\Omega}}\gamma_t\mathbf{E}^T\cdot\overline{\gamma_t\mathbf{E}^T}\Big)\,\mathsf{ds}_{\mathbf{x}}.$$

Assembling following a face point of view (see Remark 1.7), we have

$$\begin{aligned}\mathcal{I}^2 \quad &= \quad \sum_{F\in\mathcal{F}_{\rm int}}\int_F \Re\Big(\frac{1}{Z_T+Z_K}\big(\gamma_t\mathbf{E}^T-\gamma_t\mathbf{E}^K\big)\cdot\overline{\gamma_t\mathbf{E}^T}+\frac{1}{Z_K+Z_T}\big(\gamma_t\mathbf{E}^K-\gamma_t\mathbf{E}^T\big)\cdot\overline{\gamma_t\mathbf{E}^K}\Big)\,\mathsf{ds}_{\mathbf{x}}\\ &+ \quad \sum_{F\in\mathcal{F}_{\rm ext}}\int_F \Re\Big(\frac{1}{Z_T+Z_{\partial\Omega}}\gamma_t\mathbf{E}^T\cdot\overline{\gamma_t\mathbf{E}^T}\Big)\,\mathsf{ds}_{\mathbf{x}}.\end{aligned}$$

Using the electric field jump definition in (15), we get the following formula for $\mathcal{I}^2$, which vanishes for $F\in\mathcal{F}_{\rm int}$,

$$\mathcal{I}^2=\sum_{F\in\mathcal{F}_{\rm int}}\int_F\frac{1}{Z_T+Z_K}[\![\gamma_t\mathbf{E}]\!]_F\cdot\overline{[\![\gamma_t\mathbf{E}]\!]_F}\,\mathsf{ds}_{\mathbf{x}}+\sum_{F\in\mathcal{F}_{\rm ext}}\int_F\frac{1}{Z_T+Z_{\partial\Omega}}\gamma_t\mathbf{E}^T\cdot\overline{\gamma_t\mathbf{E}^T}\,\mathsf{ds}_{\mathbf{x}}.$$

With a similar reasoning and the definition of the magnetic field jump in (15), we also get

$$\mathcal{I}^3=\sum_{F\in\mathcal{F}_{\rm int}}\int_F\frac{Z_TZ_K}{Z_T+Z_K}[\![\gamma_\times\mathbf{H}]\!]_F\cdot\overline{[\![\gamma_\times\mathbf{H}]\!]_F}\,\mathsf{ds}_{\mathbf{x}}+\sum_{F\in\mathcal{F}_{\rm ext}}\int_F\frac{Z_TZ_{\partial\Omega}}{Z_T+Z_{\partial\Omega}}\gamma_\times^T\mathbf{H}^T\cdot\overline{\gamma_\times^T\mathbf{H}^T}\,\mathsf{ds}_{\mathbf{x}}.$$

We have finally proved the proposition since $\|\mathbb{E}\|_{\rm DG}^2=\Re\big(a(\mathbb{E},\mathbb{E})\big)=\mathcal{I}^2+\mathcal{I}^3$. □

*Proof.* (ii) Let us prove the injectivity. Since $a(\mathbb{E},\mathbb{E})=0$, we have $\|\mathbb{E}\|_{\rm DG}^2=0$ and therefore

$$[\![\gamma_t\mathbf{E}]\!]_F=[\![\gamma_\times\mathbf{H}]\!]_F=0,\quad \text{on } F\in\mathcal{F}_{\rm int}\quad\text{and}\quad \gamma_t\mathbf{E}=\gamma_\times^T\mathbf{H}=0,\quad \text{on } F\in\mathcal{F}_{\rm ext}. \tag{18}$$

We recall that if $\mathbb{E}^T$ satisfies (2) and $\exists F\in\mathcal{F}_T$ such that $\gamma_t\mathbf{E}^T=0$ and $\gamma_\times^T\mathbf{H}^T=0$ on $F$, then by unique continuation $\mathbb{E}^T\equiv 0$. Let $\widetilde{\mathcal{T}}\subset\mathcal{T}$ be the set of elements satisfying $\widetilde{\mathcal{T}}:=\Big\{T\in\mathcal{T}\;\Big|\;\mathbb{E}^T\equiv 0\Big\}$.
We will show that $\widetilde{\mathcal{T}}=\mathcal{T}$.
($a$) The set $\widetilde{\mathcal{T}}$ is non-empty since $\mathbb{E}^T$ is vanishing in any element $T$ with one face $F\subset\mathcal{F}_T$ included in the exterior boundary $\mathcal{F}_{\rm ext}$, see (18).
($b$) Let $T\in\mathcal{T}$ and $K\in\widetilde{\mathcal{T}}$ with a common face $F\in\mathcal{F}_T\cap\mathcal{F}_K$. Let us show that $T\in\widetilde{\mathcal{T}}$. Due to (18), we have

$$\gamma_t\mathbf{E}^T=\gamma_t\mathbf{E}^K=0 \text{ and } \gamma_\times^T\mathbf{H}^T=-\gamma_\times^K\mathbf{H}^K=0 \text{ on } F\in\mathcal{F}_T\cap\mathcal{F}_K \underset{\text{unique continuation}}{\implies} \mathbb{E}^T=0 \text{ in T} \implies T\in\widetilde{\mathcal{T}}.$$

($c$) It follows from the completeness of the neighborhood graph that $\mathcal{T}=\widetilde{\mathcal{T}}$. Thus, we have $\mathbb{E}^T=0$. □

The uniqueness of the solution to Problem 1 is proved in (ii) of Proposition 1.3. Its existence is ensured by the fact that the solution to (2), see [42], is also solution to Problem 1. However, it does not mean that the variational formulation (10), with the form $a$ given by (11a), is well-posed for every $\ell\in(\mathbb{X}_{\mathcal{T}})^*$. This question remains an open problem.

**Remark 1.8.** *The point (ii) of Proposition 1.3 implies that any finite dimensional Galerkin approximation of Problem 1 is well-posed.*

## 2. Classical Trefftz solver

The following section aims at devising a generalisation of the UWVF for heterogeneous media, based on Cessenat-Després trace operators (see [9]). Moreover, we show the equivalence between the Trefftz variational formulation based on upwind fluxes, see Problem 1, and the generalised Cessenat-Després UWVF.

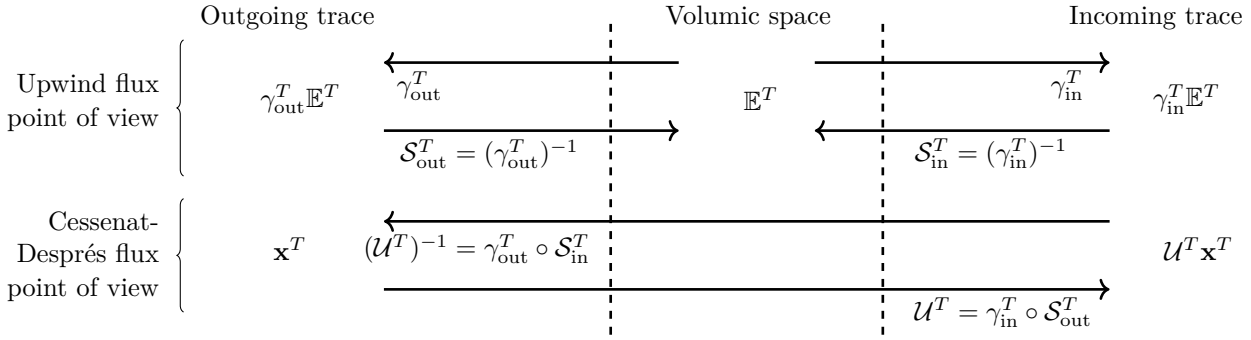


FIGURE 1. Comparison between the upwind flux and the Cessenat-Després flux points of view.

Let us first introduce, for each $T \in \mathcal{T}$, the Cessenat-Després operator $\mathcal{U}^T$ associating to the outgoing trace $\mathbf{x}^T$ of an electromagnetic field its incoming trace $\mathcal{U}^T\mathbf{x}^T$ (see Figure 1)

$$\mathbf{x}^T := \gamma^T_{\rm out}\mathbb{E}^T \underset{(9)}{=} \gamma_t\mathbf{E}^T - Z_T\gamma^T_\times\mathbf{H}^T \quad \text{and} \quad \mathcal{U}^T\mathbf{x}^T := \gamma^T_{\rm in}\mathbb{E}^T \underset{(9)}{=} \gamma_t\mathbf{E}^T + Z_T\gamma^T_\times\mathbf{H}^T. \tag{19}$$

More precisely, the operator $\mathcal{U}^T : L^2_t(\partial T) \to L^2_t(\partial T)$ is defined by $\mathcal{U}^T := \gamma^T_{\rm in} \circ \mathcal{S}^T_{\rm out}$, with the solution operator $\mathcal{S}^T_{\rm out} : L^2_t(\partial T) \to \mathbb{X}_T$, $\mathbf{x}^T \mapsto \mathbb{E}^T$ satisfying (2) and $\gamma^T_{\rm out}\mathbb{E}^T = \mathbf{x}^T$. Using above definitions, the electric tangential component and the magnetic tangential trace can then be deduced

$$\gamma_t\mathbf{E}^T = \frac{1}{2}\left(\mathcal{U}^T\mathbf{x}^T + \mathbf{x}^T\right) \quad \text{and} \quad \gamma^T_\times\mathbf{H}^T = \frac{1}{2Z_T}\left(\mathcal{U}^T\mathbf{x}^T - \mathbf{x}^T\right). \tag{20}$$

Problem 1 can be rephrased with Cessenat-Després notations. Indeed, by artificially adding and subtracting $\mathbf{x}^T$ to upwind numerical fluxes (7) and (8), we have for $F \in \mathcal{F}_T$

$$\widehat{(\gamma_t\mathbf{E})}_{|F} = \frac{1}{2}\Big(\widehat{(\mathcal{U}^T\mathbf{x})}_{|F} + \mathbf{x}^T\Big) \quad \text{and} \quad \widehat{(\gamma^T_\times\mathbf{H})}_{|F} = \frac{1}{2\,Z_T}\Big(\widehat{(\mathcal{U}^T\mathbf{x})}_{|F} - \mathbf{x}^T\Big), \tag{21}$$

where $\widehat{\mathcal{U}^T\mathbf{x}}$ is the Cessenat-Després numerical flux associated to the element $T$ defined by

$$\widehat{(\mathcal{U}^T\mathbf{x})}_{|F} := \frac{Z_K - Z_T}{Z_K + Z_T}\mathbf{x}^T + \frac{2Z_T}{Z_K + Z_T}\mathbf{x}^K \quad \text{for } F \in \mathcal{F}_{\rm int}, \tag{22a}$$

$$\widehat{(\mathcal{U}^T\mathbf{x})}_{|F} := \frac{Z_{\partial\Omega} - Z_T}{Z_{\partial\Omega} + Z_T}\mathbf{x}^T + \frac{2Z_T}{Z_{\partial\Omega} + Z_T}\mathbf{g}^T \quad \text{for } F \in \mathcal{F}_{\rm ext}. \tag{22b}$$

**Remark 2.1.** *As for Remark 1.6, the formula* (22b) *can be seen as the interaction with a virtual element $K$, when setting $Z_{\partial\Omega} = Z_K$ and $\mathbf{g}^T = \mathbf{x}^K$.*

When replacing numerical fluxes (21) in Problem 1, we get for all $\mathbf{x}' \in L^2_t(\partial\mathcal{T}) := \prod_{T\in\mathcal{T}} L^2_t(\partial T)$

$$\sum_{T\in\mathcal{T}}\int_{\partial T} Z_T^{-1}\left(\mathbf{x}^T \cdot \overline{\mathbf{x}'^T} - \widehat{\mathcal{U}^T\mathbf{x}} \cdot \overline{\mathcal{U}^T\mathbf{x}'^T}\right) \mathsf{ds}_{\mathbf{x}} = 0.$$

This leads to an alternative formulation called the Cessenat-Després heterogeneous UWVF.

**Problem 2.** *Find* $\mathbf{x} \in L^2_t(\partial\mathcal{T})$ *such that for all* $\mathbf{x}' \in L^2_t(\partial\mathcal{T})$ *we have*

$$\big(\mathbf{x},\mathbf{x}'\big)_{L^2_t(\partial\mathcal{T})} - \Big(\widehat{\mathcal{U}\mathbf{x}},\mathcal{U}\mathbf{x}'\Big)_{L^2_t(\partial\mathcal{T})} = 0, \tag{23}$$

*where* $\Big(\widehat{\mathcal{U}\mathbf{x}}\Big)^T := \widehat{\mathcal{U}^T\mathbf{x}}$, $(\mathcal{U}\mathbf{x}')^T := \mathcal{U}^T\mathbf{x}'^T$ *and the weighted scalar product on* $L^2_t(\partial\mathcal{T})$ *is defined as*

$$(\mathbf{x},\mathbf{x}')_{L^2_t(\partial\mathcal{T})} \;:=\; \sum_{T\in\mathcal{T}} (\mathbf{x}^T,\mathbf{x}'^T)_{L^2_t(\partial T)} = \sum_{T\in\mathcal{T}} \int_{\partial T} Z_T^{-1}\mathbf{x}^T\cdot\overline{\mathbf{x}'^T}\,\mathsf{ds}_{\mathbf{x}}. \tag{24}$$

As illustrated in Figure 1, a link is estalished between Problem 2 based on heterogeneous Cessenat-Després fluxes, see (19) and (21), and Problem 1 based on upwind fluxes, see (7) and (8).

**Theorem 2.1.** *Problem 1 is equivalent to Problem 2 in the following sense*

- *If* $\mathbf{x}$ *is solution to Problem 2 then* $\mathbb{E}$ *defined by* $\mathbb{E}^T = \mathcal{S}^T_{\rm out}\mathbf{x}^T$ *is solution to Problem 1.*
- *If* $\mathbb{E}$ *is solution to Problem 1 then* $\mathbf{x}$ *defined by* $\mathbf{x}^T = \gamma^T_{\rm out}\mathbb{E}^T$ *is solution to Problem 2.*

*Moreover if* $\mathbf{x}$ *is solution to Problem 2 then it is also solution to the variational formulation* (10).

Taking into account (22), we obtain the variational formulation of Problem 2.

**Proposition 2.1.** *Find* $\mathbf{x} \in L^2_t(\partial\mathcal{T})$ *such that for all* $\mathbf{x}' \in L^2_t(\partial\mathcal{T})$

$$\mathbf{a}(\mathbf{x},\mathbf{x}') \;=\; \mathbf{l}(\mathbf{x}'), \quad \textit{with} \quad \mathbf{a}(\mathbf{x},\mathbf{x}') = (\mathbf{x},\mathbf{x}')_{L^2_t(\partial\mathcal{T})} - \mathbf{k}(\mathbf{x},\mathbf{x}'), \tag{25a}$$

*where the sesquilinear form* $\mathbf{k} : L^2_t(\partial\mathcal{T})\times L^2_t(\partial\mathcal{T}) \to \mathbb{C}$ *and the antilinear form* $\mathbf{l} : L^2_t(\partial\mathcal{T}) \to \mathbb{C}$ *are given by*

$$\mathbf{k}(\mathbf{x},\mathbf{x}') := \Big(\Pi_{\mathcal{U}}\mathbf{x},\mathcal{U}\mathbf{x}'\Big)_{L^2_t(\partial\mathcal{T})} \quad \textit{and} \quad \mathbf{l}(\mathbf{x}') := \Big(\mathbf{g}_{\mathcal{U}},\mathcal{U}\mathbf{x}'\Big)_{L^2_t(\partial\mathcal{T})}, \tag{25b}$$

*with* $\Pi_{\mathcal{U}} : L^2_t(\partial\mathcal{T}) \to L^2_t(\partial\mathcal{T})$ *the flux operator and* $\mathbf{g}_{\mathcal{U}} \in L^2_t(\partial\mathcal{T})$ *the second-member, both defined on each* $T \in \mathcal{T}$ *and for all* $F \in \mathcal{F}_T$, *by*

$$\big(\Pi^T_{\mathcal{U}}\mathbf{x}\big)_{|F} := \big(\Pi_{\mathcal{U}}\mathbf{x}\big)^T_{|F} = \begin{cases} \Big(\widehat{\mathcal{U}^T\mathbf{x}}\Big)_{|F} & \textit{if } F\in\mathcal{F}_{\rm int},\\ \dfrac{Z_{\partial\Omega}-Z_T}{Z_{\partial\Omega}+Z_T}\mathbf{x}^T & \textit{if } F\in\mathcal{F}_{\rm ext}, \end{cases} \qquad \big(\mathbf{g}^T_{\mathcal{U}}\big)_{|F} := \begin{cases} 0 & \textit{if } F\in\mathcal{F}_{\rm int},\\ \dfrac{2Z_T}{Z_{\partial\Omega}+Z_T}\mathbf{g}^T & \textit{if } F\in\mathcal{F}_{\rm ext}. \end{cases} \tag{26}$$

The next proposition will assert that $\mathbf{k}$ is contractant. Thus, (25) leads to a fixed point problem.

**Proposition 2.2.** *We have*

(i) *The operator* $\Pi_{\mathcal{U}} : L^2_t(\partial\mathcal{T}) \to L^2_t(\partial\mathcal{T})$ *satisfies* $\|\Pi_{\mathcal{U}}\|_{L^2_t(\partial\mathcal{T})} \le 1$.
(ii) *The operator* $\mathcal{U} : L^2_t(\partial\mathcal{T}) \to L^2_t(\partial\mathcal{T})$ *satisfies* $\|\mathcal{U}\|_{L^2_t(\partial\mathcal{T})} = 1$.
(iii) *The operator* $\mathbf{k} : L^2_t(\partial\mathcal{T})\times L^2_t(\partial\mathcal{T}) \to \mathbb{C}$ *is contractant, ie* $\|\mathbf{k}\|_{L^2_t(\partial\mathcal{T})} \le 1$,

*where, if* $V$ *is an Hilbert space, the norm of the linear operator* $\mathcal{A} : V \to V$ *and the sesquilinear form* $\mathsf{a} : V\times V \longrightarrow \mathbb{C}$ *are defined by*

$$\|\mathcal{A}\|_V := \sup_{\mathbf{x}\in V\setminus\{0\}}\sup_{\mathbf{x}'\in V\setminus\{0\}} \frac{|(\mathcal{A}\,\mathbf{x},\mathbf{x}')_V|}{\|\mathbf{x}\|_V\;\|\mathbf{x}'\|_V} \quad \textit{and} \quad \|\mathsf{a}\|_V := \sup_{\mathbf{x}\in V\setminus\{0\}}\sup_{\mathbf{x}'\in V\setminus\{0\}} \frac{|\mathsf{a}(\mathbf{x},\mathbf{x}')_V|}{\|\mathbf{x}\|_V\;\|\mathbf{x}'\|_V}. \tag{27}$$

*Proof.* (i) Taking the face point of view from Remark 1.7, we have for all $\mathbf{x} \in L_t^2(\partial\mathcal{T})$

$$\sum_{T\in\mathcal{T}}\int_{\partial T} Z_T^{-1}\big|\Pi_{\mathcal{U}}^T\mathbf{x}\big|^2\,\mathsf{ds}_{\mathbf{x}} = \sum_{F\in\mathcal{F}}\mathcal{K}_F \text{ with } \mathcal{K}_F = \begin{cases} \displaystyle\int_F Z_T^{-1}\big|\Pi_{\mathcal{U}}^T\mathbf{x}\big|^2 + Z_K^{-1}\big|\Pi_{\mathcal{U}}^K\mathbf{x}\big|^2\,\mathsf{ds}_{\mathbf{x}} \text{ for } F\in\mathcal{F}_{\rm int},\\ \displaystyle\int_F Z_T^{-1}\big|\Pi_{\mathcal{U}}^T\mathbf{x}\big|^2\,\mathsf{ds}_{\mathbf{x}} \text{ for } F\in\mathcal{F}_{\rm ext}. \end{cases}$$

For $F\in\mathcal{F}_{\rm int}$ interfacing two elements $T$ and $K$, we have by (26)

$$\begin{cases} \mathcal{K}_F &= \displaystyle\int_F Z_T^{-1}\Big(\frac{Z_K-Z_T}{Z_K+Z_T}\Big)^2\big|\mathbf{x}^T\big|^2 + \frac{2(Z_T-Z_K)}{(Z_K+Z_T)^2}(\mathbf{x}^T\cdot\overline{\mathbf{x}^K}+\mathbf{x}^K\cdot\overline{\mathbf{x}^T}) + \frac{4Z_T}{(Z_K+Z_T)^2}\big|\mathbf{x}^K\big|^2\,\mathsf{ds}_{\mathbf{x}}\\ &+ \displaystyle\int_F Z_K^{-1}\Big(\frac{Z_T-Z_K}{Z_T+Z_K}\Big)^2\big|\mathbf{x}^K\big|^2 + \frac{2(Z_K-Z_T)}{(Z_T+Z_K)^2}(\mathbf{x}^K\cdot\overline{\mathbf{x}^T}+\mathbf{x}^T\cdot\overline{\mathbf{x}^K}) + \frac{4Z_K}{(Z_T+Z_K)^2}\big|\mathbf{x}^T\big|^2\,\mathsf{ds}_{\mathbf{x}}, \quad (28)\\ &= \displaystyle\int_F Z_T^{-1}\big|\mathbf{x}^T\big|^2 + Z_K^{-1}\big|\mathbf{x}^K\big|^2\,\mathsf{ds}_{\mathbf{x}}. \end{cases}$$

Remarking that $\dfrac{Z_{\partial\Omega}-Z_T}{Z_{\partial\Omega}+Z_T}\le 1$, we obtain for $F\in\mathcal{F}_{\rm ext}$

$$\int_F Z_T^{-1}\big|\Pi_{\mathcal{U}}^T\mathbf{x}^T\big|^2\,\mathsf{ds}_{\mathbf{x}} = \int_F Z_T^{-1}\Big(\frac{Z_{\partial\Omega}-Z_T}{Z_{\partial\Omega}+Z_T}\Big)^2\big|\mathbf{x}^T\big|^2\,\mathsf{ds}_{\mathbf{x}} = \int_F Z_T^{-1}\big|\mathbf{x}^T\big|^2\,\mathsf{ds}_{\mathbf{x}} - \int_F\frac{4Z_{\partial\Omega}}{(Z_{\partial\Omega}+Z_T)^2}\big|\mathbf{x}^T\big|^2\,\mathsf{ds}_{\mathbf{x}}. \quad (29)$$

Using the definition of the global scalar product (24), it follows

$$\|\Pi_{\mathcal{U}}\mathbf{x}\|_{L_t^2(\partial\mathcal{T})}^2 = \sum_{F\in\mathcal{F}}\mathcal{K}_F \text{ with } \mathcal{K}_F \le \begin{cases} \displaystyle\int_F Z_T^{-1}\big|\mathbf{x}^T\big|^2 + Z_K^{-1}\big|\mathbf{x}^K\big|^2\,\mathsf{ds}_{\mathbf{x}} \text{ for } F\in\mathcal{F}_{\rm int},\\ \displaystyle\int_F Z_T^{-1}\big|\mathbf{x}^T\big|^2\,\mathsf{ds}_{\mathbf{x}} \text{ for } F\in\mathcal{F}_{\rm ext}. \end{cases}$$

Applying Remark 1.7, we get an element assembling point of view leading to

$$\|\Pi_{\mathcal{U}}\mathbf{x}\|_{L_t^2(\partial\mathcal{T})}^2 \le \sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T}\int_F Z_T^{-1}\big|\mathbf{x}^T\big|^2\,\mathsf{ds}_{\mathbf{x}} = \|\mathbf{x}\|_{L_t^2(\partial\mathcal{T})}^2 \quad\Longrightarrow\quad \|\Pi_{\mathcal{U}}\|_{L_t^2(\partial\mathcal{T})}\le 1.$$

□

*Proof.* (ii) Recalling that we have the reciprocity formula (4), let us remark that

$$r_T(\mathbb{E},\mathbb{E}') = \int_{\partial T}\Big(\gamma_\times^T\mathbf{H}^T\cdot\overline{\gamma_t\mathbf{E}'^T} + \gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}'^T}\Big)\,\mathsf{ds}_{\mathbf{x}} = \frac{\mathcal{I}_1-\mathcal{I}_2}{2} = 0,$$

with $\mathcal{I}_1$ and $\mathcal{I}_2$ defined here as

$$\mathcal{I}_1 := \int_{\partial T} Z_T^{-1}\gamma_{\rm in}^T\mathbb{E}^T\cdot\overline{\gamma_{\rm in}^T\mathbb{E}'^T}\,\mathsf{ds}_{\mathbf{x}} = \int_{\partial T} Z_T^{-1}\big(\gamma_t\mathbf{E}^T + Z_T\gamma_\times^T\mathbf{H}^T\big)\cdot\Big(\overline{\gamma_t\mathbf{E}'^T} + Z_T\overline{\gamma_\times^T\mathbf{H}'^T}\Big)\,\mathsf{ds}_{\mathbf{x}},$$

$$\mathcal{I}_2 := \int_{\partial T} Z_T^{-1}\gamma_{\rm out}^T\mathbb{E}^T\cdot\overline{\gamma_{\rm out}^T\mathbb{E}'^T}\,\mathsf{ds}_{\mathbf{x}} = \int_{\partial T} Z_T^{-1}\big(\gamma_t\mathbf{E}^T - Z_T\gamma_\times^T\mathbf{H}^T\big)\cdot\Big(\overline{\gamma_t\mathbf{E}'^T} - Z_T\overline{\gamma_\times^T\mathbf{H}'^T}\Big)\,\mathsf{ds}_{\mathbf{x}}.$$

Consequently, for all $\mathbb{E}^T\in\mathbb{X}_T$, $\mathbb{E}'^T\in\mathbb{X}_T$ and for all $\mathbf{x}^T\in L_t^2(\partial T)$, $\mathbf{x}'^T\in L_t^2(\partial T)$, we have

$$\int_{\partial T} Z_T^{-1}\gamma_{\rm out}^T\mathbb{E}^T\cdot\overline{\gamma_{\rm out}^T\mathbb{E}'^T}\,\mathsf{ds}_{\mathbf{x}} = \int_{\partial T} Z_T^{-1}\gamma_{\rm in}^T\mathbb{E}^T\cdot\overline{\gamma_{\rm in}^T\mathbb{E}'^T}\,\mathsf{ds}_{\mathbf{x}} = (\mathbb{E}^T,\mathbb{E}'^T)_{\mathbb{X}_T} = (\mathcal{S}_{\rm out}^T\mathbf{x}^T,\mathcal{S}_{\rm out}^T\mathbf{x}'^T)_{\mathbb{X}_T}.$$

Since $\mathcal{S}^T_{\text{out}} : L^2_t(\partial T) \to \mathbb{X}_T$ is bijective, we notice that the space $\mathbb{X}_T$ can be parametrised by $L^2_t(\partial T)$. Summing over elements $T \in \mathcal{T}$, it leads to $\|\mathbf{x}\|^2_{L^2_t(\partial\mathcal{T})} = \|\mathcal{U}\mathbf{x}\|^2_{L^2_t(\partial\mathcal{T})}$. Therefore, we have $\|\!|\mathcal{U}|\!\|_{L^2_t(\partial\mathcal{T})} = 1$. □

*Proof.* (iii) Using the Cessenat-Després point of view, Problem 2 and the Cauchy-Schwarz inequality, we have

$$\mathbf{k}(\mathbf{x},\mathbf{x}') = \Big(\Pi_{\mathcal{U}}\mathbf{x}, \mathcal{U}\mathbf{x}'\Big)_{L^2_t(\partial\mathcal{T})} \;\leq\; \big\|\Pi_{\mathcal{U}}\mathbf{x}\big\|_{L^2_t(\partial\mathcal{T})}\, \big\|\mathcal{U}\mathbf{x}'\big\|_{L^2_t(\partial\mathcal{T})} \underset{\text{(i),(ii)}}{\leq} \big\|\mathbf{x}\big\|_{L^2_t(\partial\mathcal{T})}\, \big\|\mathbf{x}'\big\|_{L^2_t(\partial\mathcal{T})}.$$

This leads to $\|\!|\mathbf{k}|\!\|_{L^2_t(\partial\mathcal{T})} \leq 1$ and ends the proof. □

**Remark 2.2.** *Property* (ii) *in Proposition 2.2 can be applied to the Cessenat-Després Problem 2. More precisely, we can replace the scalar product in the left-hand member of equation* (23)*. We then we get another formulation for the heterogeneous UWVF*

$$\textit{Find } \mathbf{x} \in L^2_t(\partial\mathcal{T}) \textit{ such that for all } \mathbf{x}' \in L^2_t(\partial\mathcal{T}), \quad \textit{we have} \quad \Big(\mathcal{U}\mathbf{x} - \widehat{\mathcal{U}\mathbf{x}}, \mathcal{U}\mathbf{x}'\Big)_{L^2_t(\partial\mathcal{T})} = 0.$$

*This variational formulation encodes the continuity of the solution, ie* $\mathcal{U}^T\mathbf{x}^T = \big(\widehat{\mathcal{U}^T\mathbf{x}}\big)$.

Proposition 2.2 does not ensure that the operator $\mathbf{k}$ is strictly contractant. Therefore, we have chosen to introduce an iterative Trefftz problem only at a discrete level.

## 3. Iterative Trefftz solver

An iterative Trefftz method based on Cessenat-Després numerical fluxes is introduced. The convergence is theoretically established in Proposition 3.1.

We define the global discrete Trefftz space $\mathbb{X}^h_{\mathcal{T}}$, that is a finite dimensional linear subspace of $\mathbb{X}_{\mathcal{T}}$, as

$$\mathbb{X}^h_{\mathcal{T}} := \Big(\prod_{T\in\mathcal{T}} \mathbb{X}^h_T\Big) \subset \mathbb{X}_{\mathcal{T}}.$$

The local Trefftz space $\mathbb{X}^h_T := \text{span}\Big\{\mathbf{v}^\ell_T \in \mathbb{X}_T, \ell = 1, N_T\Big\}$, where the functions $\mathbf{v}^\ell_T$ are electromagnetic plane waves, with direction $\mathbf{d}^\ell_T$ and polarisation $\mathbf{p}^\ell_T$, which can be chosen as

$$\mathbf{v}^\ell_T := \big(\mathbf{E}^T, \mathbf{H}^T\big) \in \mathbb{X}_T \quad \text{with} \quad \mathbf{E}^T := \mathbf{p}^\ell_T\, e^{ik_0\sqrt{\varepsilon_r\mu_r}\mathbf{d}^\ell_T\cdot\mathbf{x}} \quad \text{and} \quad \mathbf{H}^T := Z_T(\mathbf{p}^\ell_T \times \mathbf{d}^\ell_T)\, e^{ik_0\sqrt{\varepsilon_r\mu_r}\mathbf{d}^\ell_T\cdot\mathbf{x}}, \tag{30}$$

with $Z_T = \sqrt{\mu^T_r/\varepsilon^T_r}$ a normalised impedance defined on $T$. More precisely, $\mathbf{d}^\ell_T \in \mathcal{D}$, a finite discrete subspace of the unit sphere, and $\mathbf{p}^\ell_T \in S_{\mathbf{d}}$, an orthonormal basis of the two-dimensional linear subspace $(\mathbf{d}^\ell_T)^\perp \subset \mathbb{R}^3$. Many different choices for the definition of $\mathcal{D}$ exist. Directions are given by the vertices of the triangular mesh surface of the unit sphere, or the surface mesh of a unit cube. For a complete plane waves approximation theory, one can refer to [32, 41].

Problem 2 can be discretised on the finite dimensional space $\mathbb{Y}^h_{\mathcal{T}} \subset L^2_t(\partial\mathcal{T})$

$$\mathbb{Y}^h_{\mathcal{T}} := \prod_{T\in\mathcal{T}} \mathbb{Y}^h_T\,, \quad \text{with } \mathbb{Y}^h_T := \gamma^T_{\text{out}}\mathbb{X}^h_T.$$

More precisely, for each $T \in \mathcal{T}$, the local finite discrete space $\mathbb{Y}^h_T$ is spanned by the incoming traces of plane waves in $\mathbb{X}^h_T$, which are defined by (30). Since $\gamma^T_{\text{out}}$ is a bijective operator, the space $\mathbb{Y}^h_T$ is of dimension $N := \mathsf{dim}\big(\mathbb{Y}^h_T\big) = \mathsf{dim}\big(\mathbb{X}^h_T\big) = N_T$. Consequently, the basis $\mathbb{Y}^h_T$ is given by

$$\mathbb{Y}^h_T := \Big\{\mathbf{w}_T = \gamma^T_{\text{out}}\mathbf{v}_T \text{ such that } \mathbf{v}_T \in \mathbb{X}^h_T\Big\} = \underset{\ell=1,N}{\text{span}}\Big\{\mathbf{w}^\ell_T \text{ such that } \mathbf{w}^\ell_T = \gamma^T_{\text{out}}\mathbf{v}^\ell_T\Big\}.$$

The space $\mathbb{Y}^h_{\mathcal{T}}$ has the dimension $\#\text{dof} := N\,\#\text{elem}$, with $\#\text{elem} := \mathsf{card}(\mathcal{T})$ the total number of elements in the mesh. Any element $\mathbf{x} \in \mathbb{Y}^h_{\mathcal{T}}$ will be represented by a complex column vector $[\mathbf{x}]$ of dimension $\#\text{dof}$. Their components $[\mathbf{x}]_{\text{iglob}}$ are the amplitudes of the plane wave traces $\mathbf{w}^\ell_{\text{i}} \in \mathbb{Y}^h_T$ in the $\text{i}^{\text{th}}$ element $T$. It leads to

$$\mathbf{x} = \sum_{\text{iglob}=1}^{\#\text{dof}} [\mathbf{x}]_{\text{iglob}} \mathbf{w}^{\text{iglob}} = \sum_{\text{i}=1}^{\#\text{elem}} \sum_{\ell=1}^{N} [\mathbf{x}]^\ell_{\text{i}} \mathbf{w}^\ell_{\text{i}}, \quad \text{with iglob} := (\text{i}-1)\,N + \ell, \text{ and } \begin{cases} [\mathbf{x}]_{\text{iglob}} := [\mathbf{x}]^\ell_{\text{i}}, \\ \mathbf{w}^{\text{iglob}} := \mathbf{w}^\ell_{\text{i}}. \end{cases} \tag{31}$$

The discrete direct Cessenat-Després heterogeneous UWVF is formulated as follows.

**Problem 3.** *Find $\mathbf{x}^h \in \mathbb{Y}^h_{\mathcal{T}}$ such that for all $\mathbf{x}' \in \mathbb{Y}^h_{\mathcal{T}}$, we have*

$$\mathbf{a}(\mathbf{x}^h, \mathbf{x}') = \mathbf{l}(\mathbf{x}') \quad \Longleftrightarrow \quad \mathbf{A}[\mathbf{x}] = \mathbf{F}, \tag{32}$$

*with $\mathbf{A} \in \mathbb{C}^{\#\text{dof}\times\#\text{dof}}$ and $\mathbf{F} \in \mathbb{C}^{\#\text{dof}}$ defined component by component by*

$$\mathbf{A}_{\text{iglob,jglob}} := \mathbf{a}(\mathbf{w}^{\text{jglob}}, \mathbf{w}^{\text{iglob}}) \quad \text{and} \quad \mathbf{F}_{\text{iglob}} := \mathbf{l}(\mathbf{w}^{\text{iglob}}) \quad \text{for iglob, jglob} = 1, \#\text{dof}. \tag{33}$$

**Remark 3.1.** *The global matrix $\mathbf{A}$ is composed of block matrices $\mathbf{A}_{\text{i,j}} := \left(\mathbf{A}^{\ell,k}_{\text{i,j}}\right)_{\ell,k=1}^{N} \in \mathbb{C}^{N\times N}$ for $\text{i,j} = 1, \#\text{elem}$ defined as*

$$\mathbf{A}^{\ell,k}_{\text{i,j}} = \mathbf{a}(\mathbf{w}^k_{\text{j}}, \mathbf{w}^\ell_{\text{i}}) \quad \text{with} \quad \ell, k = 1, N.$$

Proposition 2.2 leads to a singular regular decomposition of matrix $\mathbf{A} = \mathbf{M} - \mathbf{N}$. In the homogeneous case, it corresponds to a Jacobi-block decomposition, see [11]. We then obtain the discrete iterative UWVF problem.

**Problem 4.** *Compute the sequence $\mathbf{x}^{\text{jac}}_n \in \mathbb{Y}^h_{\mathcal{T}}$, $n \in \mathbb{N}^*$, from the recurrence*

$$(\mathbf{x}^{\text{jac}}_{n+1}, \mathbf{x}')_{L^2_t(\partial\mathcal{T})} - \mathbf{k}(\mathbf{x}^{\text{jac}}_n, \mathbf{x}') = \mathbf{l}(\mathbf{x}'),\ \forall \mathbf{x}' \in \mathbb{Y}^h_{\mathcal{T}} \quad \Longleftrightarrow \quad \mathbf{M}[\mathbf{x}^{\text{jac}}_{n+1}] = \mathbf{N}[\mathbf{x}^{\text{jac}}_n] + \mathbf{F}, \quad \text{with } [\mathbf{x}^{\text{jac}}_0] = 0, \tag{34}$$

*with $\mathbf{M}/\mathbf{N} \in \mathbb{C}^{\#\text{dof}\times\#\text{dof}}$ defined by*

$$\mathbf{M}_{\text{iglob,jglob}} := \left(\mathbf{w}^{\text{jglob}}, \mathbf{w}^{\text{iglob}}\right)_{L^2_t(\partial\mathcal{T})} \quad \text{and} \quad \mathbf{N}_{\text{iglob,jglob}} := \mathbf{k}(\mathbf{w}^{\text{jglob}}, \mathbf{w}^{\text{iglob}}), \quad \text{for iglob, jglob} = 1, \#\text{dof}. \tag{35}$$

**Remark 3.2.** *Support properties of DG basis functions lead to a hermitian positive block-diagonal matrix $\mathbf{M}$*

$$\mathbf{M}^{\ell,k}_{\text{i,j}} = (\mathbf{w}^k_{\text{i}}, \mathbf{w}^\ell_{\text{i}})_{L^2_t(\partial\mathcal{T})}\,\delta_{\text{i,j}} \qquad \text{for} \quad \text{i,j} = 1, \#\text{elem} \quad \text{and} \quad \ell, k = 1, N,$$

*which allows a fast direct inversion of the linear system* (34).

To ensure the convergence of the discrete iterative Problem 4, the matrix $\mathbf{M}^{-1}\mathbf{N}$ associated to the form $\mathbf{k}$ needs to be strictly contractant. In other terms, its spectral radius, denoted by $\rho(\mathbf{M}^{-1}\mathbf{N})$, has to verify $\rho(\mathbf{M}^{-1}\mathbf{N}) < 1$. Due to the conformal nature of the discretisation, *ie* $\mathbb{Y}^h_{\mathcal{T}} \subset L^2_t(\partial\mathcal{T})$, we have

$$\rho(\mathbf{M}^{-1}\mathbf{N}) \;\le\; |\!|\!|\mathbf{k}|\!|\!|_{\mathbb{Y}^h_{\mathcal{T}}} \;=\; \sup_{\mathbf{x}\in\mathbb{Y}^h_{\mathcal{T}}\setminus\{0\}} \sup_{\mathbf{x}'\in\mathbb{Y}^h_{\mathcal{T}}\setminus\{0\}} \frac{|\mathbf{k}(\mathbf{x},\mathbf{x}')|}{\|\mathbf{x}\|_{\mathbb{Y}^h_{\mathcal{T}}}\,\|\mathbf{x}'\|_{\mathbb{Y}^h_{\mathcal{T}}}} \;\le |\!|\!|\mathbf{k}|\!|\!|_{L^2_t(\partial\mathcal{T})} \le 1, \quad \text{with } \|\mathbf{x}\|_{\mathbb{Y}^h_{\mathcal{T}}} = \|\mathbf{x}\|_{L^2_t(\partial\mathcal{T})}.$$

It then remains to exclude the eigenvalues on the unit circle.

**Proposition 3.1.** *The matrix $\mathbf{M}^{-1}\mathbf{N}$ is strictly contractant, ie $\rho(\mathbf{M}^{-1}\mathbf{N}) < 1$.*

*Proof.* We suppose that it exists $\mathbf{x} \in \mathbb{Y}^h_{\mathcal{T}}$ and $\lambda \in \mathbb{C}$, such that $\mathbf{x} \neq 0$ and $|\lambda| = 1$. We have

$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = \lambda\,(\mathbf{x}, \mathbf{x}')_{L^2_t(\partial\mathcal{T})}, \quad \forall \mathbf{x}' \in \mathbb{Y}^h_{\mathcal{T}} \quad \Longleftrightarrow \quad \mathbf{N}[\mathbf{x}] = \lambda\,\mathbf{M}[\mathbf{x}], \text{ with } \mathbf{x} \text{ represented by } [\mathbf{x}] \text{ through (31)}.$$

It leads to

$$\|\Pi_{\mathcal{U}}\mathbf{x}-\lambda\mathcal{U}\mathbf{x}\|^2_{L^2_t(\partial\mathcal{T})}=\|\Pi_{\mathcal{U}}\mathbf{x}\|^2_{L^2_t(\partial\mathcal{T})}-\lambda\,(\mathcal{U}\mathbf{x},\Pi_{\mathcal{U}}\mathbf{x})_{L^2_t(\partial\mathcal{T})}-\overline{\lambda}(\Pi_{\mathcal{U}}\mathbf{x},\mathcal{U}\mathbf{x})_{L^2_t(\partial\mathcal{T})}+|\lambda|^2\,\|\mathcal{U}\mathbf{x}\|_{L^2_t(\partial\mathcal{T})}.$$

Thanks to Proposition 2.2 and the definition of $\mathbf{k}$, see (25b), we obtain

$$\|\Pi_{\mathcal{U}}\mathbf{x}-\lambda\mathcal{U}\mathbf{x}\|_{L^2_t(\partial\mathcal{T})}\leq\|\mathbf{x}\|^2_{L^2_t(\partial\mathcal{T})}-\lambda\overline{\lambda}\,\|\mathbf{x}\|^2_{L^2_t(\partial\mathcal{T})}-\overline{\lambda}\lambda\,\|\mathbf{x}\|^2_{L^2_t(\partial\mathcal{T})}+|\lambda|^2\,\|\mathcal{U}\mathbf{x}\|^2_{L^2_t(\partial\mathcal{T})}\underset{|\lambda|=1}{\leq}0\implies\Pi_{\mathcal{U}}\mathbf{x}=\lambda\mathcal{U}\mathbf{x}. \quad (36)$$

Then, we act by contradiction. Let us show that $\mathbf{x}=0$. Let $\widetilde{\mathcal{T}}:=\left\{T\in\mathcal{T}\;\middle|\;\mathbf{x}^T\equiv0\right\}\subset\mathcal{T}$. It remains to prove that $\widetilde{\mathcal{T}}=\mathcal{T}$ to obtain Proposition 3.1.
($a$) Let us show that $\widetilde{\mathcal{T}}$ is non-empty. From (36), we have $\|\Pi_{\mathcal{U}}\mathbf{x}\|_{L^2_t(\partial\mathcal{T})}=\|\mathcal{U}\mathbf{x}\|_{L^2_t(\partial\mathcal{T})}=\|\mathbf{x}\|_{L^2_t(\partial\mathcal{T})}$ since $\mathcal{U}$ is unitary from Proposition 2.2. From (28) and (29), it follows that

$$0\;=\;\|\mathbf{x}\|^2_{L^2_t(\partial\mathcal{T})}-\|\Pi_{\mathcal{U}}\mathbf{x}\|^2_{L^2_t(\partial\mathcal{T})}\;=\;\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T\cap\mathcal{F}_{ext}}\int_F\underbrace{\frac{4Z_{\partial\Omega}}{(Z_{\partial\Omega}+Z_T)^2}}_{>0}\left|\mathbf{x}^T\right|^2\mathsf{ds}_{\mathbf{x}}\implies\mathbf{x}^T=0\quad\text{on }F\in\mathcal{F}_T\cap\mathcal{F}_{\text{ext}}.$$

Moreover, recalling the definition (26) and due to (36), we also have

$$\mathcal{U}^T\mathbf{x}^T=\frac{1}{\lambda}\,\frac{Z_{\partial\Omega}-Z_T}{Z_{\partial\Omega}+Z_T}\mathbf{x}^T=0\quad\text{on }F\in\mathcal{F}_T\cap\mathcal{F}_{\text{ext}}.$$

Combining the two last results and using (20), we obtain $\gamma_t\mathbf{E}^T=\gamma_\times\mathbf{H}^T=0$ on $F$, leading to $\mathbb{E}^T\equiv0$ and $\mathbf{x}^T\equiv0$ in $T$ by the unique continuation.
($b$) Let $T\in\mathcal{T}$ and $K\in\widetilde{\mathcal{T}}$ with a common face $F\in\mathcal{F}_T\cap\mathcal{F}_K$. Let us show that $T\in\widetilde{\mathcal{T}}$. Since $K\in\widetilde{\mathcal{T}}$, we have $\mathbf{x}^K=0$ and $\mathcal{U}^K\mathbf{x}^K=0$. From (36), we have $\Pi^T_{\mathcal{U}}\mathbf{x}=\lambda\mathcal{U}^T\mathbf{x}^T$ and $\Pi^K_{\mathcal{U}}\mathbf{x}=\lambda\mathcal{U}^K\mathbf{x}^K$ on $F$. From (22a) and (26), we obtain

$$\frac{Z_T-Z_K}{Z_K+Z_T}\mathbf{x}^K+\frac{2Z_K}{Z_K+Z_T}\mathbf{x}^T=\lambda\mathcal{U}^K\mathbf{x}^K\quad\text{and}\quad\frac{Z_K-Z_T}{Z_K+Z_T}\mathbf{x}^T+\frac{2Z_T}{Z_K+Z_T}\mathbf{x}^K=\lambda\mathcal{U}^T\mathbf{x}^T,$$

leading to $\mathbf{x}^T=0$ and $\mathcal{U}^T\mathbf{x}^T=0$ on $F$, since $Z_{T/K}>0$ and $\lambda\neq0$. Thus, $T\in\widetilde{\mathcal{T}}$. Due to (20), we also obtain $\gamma_t\mathbf{E}^T=\gamma_\times\mathbf{H}^T=0$ on $F$, leading to $\mathbb{E}^T\equiv0$ and $\mathbf{x}^T\equiv0$ in $T$ by the unique continuation theorem.
($c$) From the completness and the connexity of the neighborhood graph, we then have $\mathcal{T}=\widetilde{\mathcal{T}}$. □

**Remark 3.3.** *A weaker contraction result has been proved in the context of an under-relaxation iterative method by Cessenat-Després (see [10]). This result takes the form of $\rho\left((1-\beta)\mathbf{I}+\beta\,\mathbf{M}^{-1}\mathbf{N}\right)<1$ for all $\beta\in]0,1[$. In particular, the Proposition 3.1 refines this result by including the limit case $\beta=1$.*

**Remark 3.4.** *This proof ensures the uniqueness of the solution to the variational formulation* (32)*. Combined with the rank-nullity theorem, it leads to the well-posedness of Problem 4 which is finite dimensional.*

This result makes it possible to use a Jacobi like iterative method to solve the linear system. However, the spectral radius of the iterative matrix $\mathbf{M}^{-1}\mathbf{N}$ can be very close of 1 and can significantly slow down the iterative algorithm as illustrated in Table 1 and Fig. 2.

Improvements proposed in Section 4 will show that it is related to the Cessenat-Després method.

| $\mathcal{D}_\Omega$ | 8 | 20 | 40 | 80 |
|---|---|---|---|---|
| $\rho\left(\mathbf{M}^{-1}\mathbf{N}\right)$ | $1-0.0064$ | $1-0.0022$ | $1-0.0014$ | $1-0.00054$ |

TABLE 1. Comparison of $\rho\left(\mathbf{M}^{-1}\mathbf{N}\right)$ thanks to the power method, on different $\mathcal{D}_\Omega$ in wavelength.

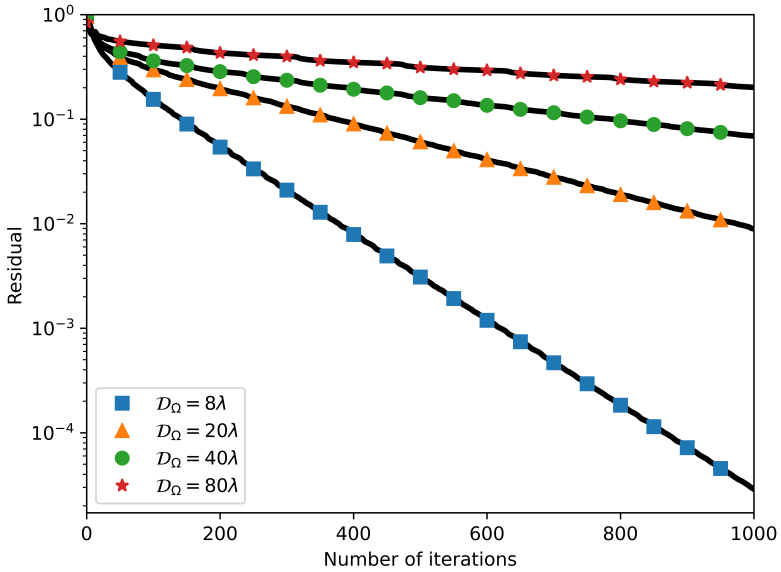


FIGURE 2. Comparison of the convergence rate of the iterative Cessenat-Despres method in function of the size of the domain.

**Remark 3.5.** *Unless otherwise specified, the following configuration is considered for numerical experiments*

- *All distances or lengths are given in wavelengths, ie* $k = 2\pi$.
- *The domain is of size* $\mathcal{D}_\Omega$ *and is decomposed into elements* $T$ *which are cubes of size* $\mathcal{D}_T = 1$.
- *The reflexion coefficient on* $\partial\Omega$ *is denoted by* $R_{\partial\Omega}$. *In particular, we set* $Z_{\partial\Omega} = (1 - R_{\partial\Omega})/(1 + R_{\partial\Omega})$ *with* $R_{\partial\Omega} = 0.9$.
- *The number of plane wave basis functions is* $N = 52$, *that are homogeneously distributed in space.*

## 4. New Iterative Trefftz solver

In the present section, we propose alternatives to ensure the development of a robust Trefftz iterative method, even with the presence of rounding errors. First, we will guarantee its convergence thanks to a preconditioned Krylov method. Then, we resort to a reduction strategy to consider less basis functions providing a lower numerical cost. Finally, we set up a global preconditioner to improve the convergence rate of the method.

### 4.1. Preconditioned iterative Trefftz solver

A large amount of Krylov type methods exist, see [50]. In this paper, we focus on one based on the positivity of the matrix $\mathbf{A}$, *ie* $[\mathbf{x}]^*\mathbf{A}[\mathbf{x}] > 0$. It is a variational method that resorts to Krylov spaces to reduce the dimension of the Galerkin space, see [51, 52]. It can be interpreted as an iterative numerical method which does not require the computation of the inverse of the matrix $\mathbf{A}$. When associated to a restart strategy, this method has a low memory cost and is appropriated to large numerical cases. Moreover, it is perfectly suited to real positive matrices, thus adapted to the inversion of Problem 3 which satisfies $[\mathbf{x}]^*\mathbf{A}[\mathbf{x}] > 0$. The Galerkin framework of the UWVF method leads to the introduction of the following iterative Krylov UWVF problem.

**Problem 5.** *Setting* $\mathbf{x}_0 = 0$, *compute the sequence* $(\mathbf{x}_n)_{n\in\mathbb{N}^*}$ *such that* $\mathbf{x}_n \in \mathbb{K}_n$ *is the solution of*

$$\textit{Find } \mathbf{x}_n \in \mathbb{K}_n,\ \mathbf{a}(\mathbf{x}_n, \mathbf{x}') = \mathbf{l}(\mathbf{x}'),\ \forall \mathbf{x}' \in \mathbb{K}_n \quad \Longleftrightarrow \quad \textit{Find } [\mathbf{x}_n] \in [\mathbb{K}_n],\ [\mathbf{x}']^*\mathbf{A}[\mathbf{x}_n] = [\mathbf{x}']^*\mathbf{F},\ \forall [\mathbf{x}'] \in [\mathbb{K}_n], \quad (37)$$

*with $\mathbb{K}_n$ a linear subspace of $\mathbb{Y}_{\mathcal{T}}^h$ defined as: $\mathbf{x}_n \in \mathbb{K}_n$ is represented by $[\mathbf{x}_n] \in \big[\mathbb{K}_n\big]$ through the bijection defined by (31), where $\big[\mathbb{K}_n\big]$ is the Krylov space associated to $\mathbf{A}$ defined as*

$$\big[\mathbb{K}_n\big] := \underset{0\le k\le n-1}{\text{span}}\big(\mathbf{A}^k\mathbf{F}\big) = \text{span}\left\{\mathbf{F},\, \mathbf{AF},\, \mathbf{A}^2\mathbf{F},\, \ldots,\, \mathbf{A}^{n-1}\mathbf{F}\right\}. \tag{38}$$

The convergence theory of Problem 5 is established thanks to the Galerkin theory of the UWVF Problem 2.

**Proposition 4.1.** *Let $\mathbf{x}^h \in \mathbb{Y}_{\mathcal{T}}^h$ and $\mathbf{x}_n \in \mathbb{K}_n$ be the solutions of Problems 2 and 5, respectively. The convergence of the iterative Krylov UWVF method is ensured by*

$$\|\mathbf{x}^h - \mathbf{x}_n\|_{\text{DG}} \;\le\; \sqrt{2}\min_{\mathbf{y}\in\mathbb{K}_n}\|\mathbf{x}^h - \mathbf{y}\|_{L_t^2(\partial\mathcal{T})}. \tag{39}$$

*Proof.* Let us recall that we have $\|\mathbf{x}^h - \mathbf{x}_n\|_{\text{DG}}^2 = \Re\big(\mathbf{a}(\mathbf{x}^h - \mathbf{x}_n, \mathbf{x}^h - \mathbf{x}_n)\big)$. Since $\mathbf{a}(\mathbf{x}^h - \mathbf{x}_n, \mathbf{x}_n - \mathbf{y}) = 0$ for all $\mathbf{y} \in \mathbb{K}_n$, we have

$$\|\mathbf{x}^h - \mathbf{x}_n\|_{\text{DG}}^2 = \Re\big(\mathbf{a}(\mathbf{x}^h - \mathbf{x}_n, \mathbf{x}^h - \mathbf{y})\big) = \Re\big([\mathbf{x}^h - \mathbf{y}]^*\mathbf{A}[\mathbf{x}^h - \mathbf{x}_n]\big).$$

Since $\mathbf{M}$ is symmetric positive definite, we have

$$[\mathbf{x}^h - \mathbf{y}]^*\mathbf{A}[\mathbf{x}^h - \mathbf{x}_n] = [\mathbf{x}^h - \mathbf{y}]^*\mathbf{M}^{\frac12}\mathbf{M}^{-\frac12}\mathbf{A}[\mathbf{x}^h - \mathbf{x}_n] = \big(\mathbf{M}^{\frac12}[\mathbf{x}^h - \mathbf{y}]\big)^*\mathbf{M}^{-\frac12}\mathbf{A}[\mathbf{x}^h - \mathbf{x}_n].$$

Then, we apply the Cauchy-Schwarz inequality and we get

$$\left\{\begin{array}{rcl} \|\mathbf{x}^h - \mathbf{x}_n\|_{\text{DG}}^2 & \le & \sqrt{\big(\mathbf{M}^{\frac12}[\mathbf{x}^h - \mathbf{y}]\big)^*\mathbf{M}^{\frac12}[\mathbf{x}^h - \mathbf{y}]}\sqrt{\big(\mathbf{M}^{-\frac12}\mathbf{A}[\mathbf{x}^h - \mathbf{x}_n]\big)^*\mathbf{M}^{-\frac12}\mathbf{A}[\mathbf{x}^h - \mathbf{x}_n]}, \\ & \le & \sqrt{[\mathbf{x}^h - \mathbf{y}]^*\mathbf{M}[\mathbf{x}^h - \mathbf{y}]}\sqrt{\big(\mathbf{A}[\mathbf{x}^h - \mathbf{x}_n]\big)^*\mathbf{M}^{-1}\mathbf{A}[\mathbf{x}^h - \mathbf{x}_n]}. \end{array}\right.$$

Let us now prove that, for all $[\mathbf{x}] \in \mathbb{C}^{\#\text{dof}}$, $\big(\mathbf{A}[\mathbf{x}]\big)^*\mathbf{M}^{-1}\big(\mathbf{A}[\mathbf{x}]\big) \le 2\,\Re\big([\mathbf{x}]^*\mathbf{A}[\mathbf{x}]\big)$. Using the Cessenat-Després decomposition $\mathbf{A} = \mathbf{M} - \mathbf{N}$, we get

$$\big(\mathbf{A}[\mathbf{x}]\big)^*\mathbf{M}^{-1}\big(\mathbf{A}[\mathbf{x}]\big) = [\mathbf{x}]^*\mathbf{M}[\mathbf{x}] - [\mathbf{x}]^*\mathbf{N}[\mathbf{x}] - [\mathbf{x}]^*\mathbf{N}^*[\mathbf{x}] + [\mathbf{x}]^*\mathbf{N}^*\mathbf{M}^{-1}\mathbf{N}[\mathbf{x}].$$

Let us take $[\mathbf{z}] \in \mathbb{C}^{\#\text{dof}}$, set as $[\mathbf{z}] = \mathbf{M}^{-1}\mathbf{N}[\mathbf{x}]$. It leads to $[\mathbf{x}]^*\mathbf{N}^*\mathbf{M}^{-1}\mathbf{N}[\mathbf{x}] = [\mathbf{z}]^*\mathbf{N}[\mathbf{x}] = \mathbf{k}(\mathbf{x},\mathbf{z})$. Due to (iii) of Proposition 2.2, we have

$$[\mathbf{x}]^*\mathbf{N}^*\mathbf{M}^{-1}\mathbf{N}[\mathbf{x}] \;\le\; \|\mathbf{x}\|_{L_t^2(\partial\mathcal{T})}\|\mathbf{z}\|_{L_t^2(\partial\mathcal{T})} = \sqrt{[\mathbf{z}]^*\mathbf{M}[\mathbf{z}]}\sqrt{[\mathbf{x}]^*\mathbf{M}[\mathbf{x}]} \le \sqrt{[\mathbf{x}]^*\mathbf{N}^*\mathbf{M}^{-1}\mathbf{N}[\mathbf{x}]}\sqrt{[\mathbf{x}]^*\mathbf{M}[\mathbf{x}]}.$$

Therefore, we get $[\mathbf{x}]^*\mathbf{N}^*\mathbf{M}^{-1}\mathbf{N}[\mathbf{x}] \le [\mathbf{x}]^*\mathbf{M}[\mathbf{x}]$, leading to

$$\big(\mathbf{A}[\mathbf{x}]\big)^*\mathbf{M}^{-1}\big(\mathbf{A}[\mathbf{x}]\big) \;\le\; 2\,[\mathbf{x}]^*\mathbf{M}[\mathbf{x}] - 2\,\Re\big([\mathbf{x}]^*\mathbf{N}[\mathbf{x}]\big) = 2\,\Re\big([\mathbf{x}]^*\mathbf{A}[\mathbf{x}]\big) = 2\,\|\mathbf{x}\|_{\text{DG}}^2.$$

The result follows from $\|\mathbf{x}^h - \mathbf{x}_n\|_{\text{DG}} \;\le\; \sqrt{2}\,\|\mathbf{x}^h - \mathbf{y}\|_{L_t^2(\partial\mathcal{T})}$. □

We have not succeed to obtain an optimal estimate for the right-hand side of (39). Many convergence bounds of the Krylov residual exist in the literature, see [21]. But these bounds are generally pessimistic. Practically, Krylov solver often has a much better behaviour. Although we are aware of this aspect, we use the general convergence theory of Krylov methods, see [52], to estimate the right-hand side of (39).

**Proposition 4.2.** *Suppose that $\mathbf{A}$ is diagonalisable, ie $\mathbf{A} = \mathbf{XDX}^{-1}$, we have for $\mathbf{x}^h \in \mathbb{Y}_{\mathcal{T}}^h$*

$$\min_{\mathbf{y}\in\mathbb{K}_n}\|\mathbf{x}^h - \mathbf{y}\|_{L^2(\partial\mathcal{T})} \;\le\; \min_{p\in\mathcal{P}_n^0}\max_{\lambda\in\sigma^{\mathbf{A}}}\left|\frac{p(\lambda)}{\lambda}\right|\sqrt{\kappa(\mathbf{M})}\,\kappa(\mathbf{X})\,\|\mathbf{l}\|_{(\mathbb{Y}_{\mathcal{T}}^h)^*}, \tag{40}$$

*where $\mathcal{P}_n^0$ is the set of polynomials of degree $\leq n$ satisfying $p(0)=1$, $\sigma^{\mathbf{A}}$ is the spectrum of $\mathbf{A}$, $\kappa(\cdot)$ is the condition number and $(\mathbb{Y}_{\mathcal{T}}^h)^*$ is the dual space of $\mathbb{Y}_{\mathcal{T}}^h$.*

*Proof.* Let $\mathbf{y} \in \mathbb{K}_n$. Thus, it exists $q$ a polynomial of degree $n-1$ such that $\mathbf{y} = q(\mathbf{A})\mathbf{F}$. Since $\mathbf{A}$ is assumed to be diagonalisable, *ie* $\mathbf{A} = \mathbf{X}\mathbf{D}\mathbf{X}^{-1}$, we have

$$\mathbf{X}^{-1}\left(\left[\mathbf{x}^h - \mathbf{y}\right]\right) = (\mathbf{D}^{-1} - q(\mathbf{D}))\mathbf{X}^{-1}\mathbf{F}.$$

Noting that the Euclidian norm of a diagonal matrix is the highest absolute value of its coefficients:

$$\left\|\mathbf{X}^{-1}\left[\mathbf{x}^h - \mathbf{y}\right]\right\| \leq \max_{\lambda\in\sigma^{\mathbf{A}}}\left|\frac{1}{\lambda} - q(\lambda)\right| \left\|\mathbf{X}^{-1}\mathbf{F}\right\|,$$

and using the estimate:

$$\left\|\left[\mathbf{x}^h - \mathbf{y}\right]\right\| \leq \|\mathbf{X}\| \left\|\mathbf{X}^{-1}\left[\mathbf{x}^h - \mathbf{y}\right]\right\|,$$

we can write

$$\left\|\left[\mathbf{x}^h - \mathbf{y}\right]\right\| \leq \max_{\lambda\in\sigma^{\mathbf{A}}}\left|\frac{1-\lambda q(\lambda)}{\lambda}\right| \kappa(\mathbf{X})\ \|\mathbf{F}\|,$$

with $\kappa(\mathbf{X}) := \|\mathbf{X}\|\ \left\|\mathbf{X}^{-1}\right\|$ the condition number of $\mathbf{X}$.
We end the proof by using these following statements:

- the bijection $q \in \mathcal{P}_{n-1} \mapsto p(x) = 1 - xq(x) \in \mathcal{P}_n^0$,
- $\forall \mathbf{z} \in \mathbb{Y}_{\mathcal{T}}^h$,

$$\left\|\mathbf{M}^{-1}\right\|^{-1}\ \|[z]\|^2 \leq \|\mathbf{z}\|_{L^2(\partial\mathcal{T})}^2 = [z]^*\ \mathbf{M}\ [z] \leq \|\mathbf{M}\|\ \|[z]\|^2.$$

□

The convergence rate depends on the spectrum of $\mathbf{A}$ and draws a particular attention to eigenvalues close to 0, see in red in Figure 3. The fact that $\mathbf{A}$ is diagonalisable is theoretically discutable, even if it appears to be in practice.

In the continuation, we introduce a preconditioned iterative Krylov UWVF method.

**Problem 6.** *Setting $\mathbf{x}_0^{\mathrm{prec}} = 0$, compute the sequence $(\mathbf{x}_n^{\mathrm{prec}})_{n\in\mathbb{N}^*}$ solution to: Find $\mathbf{x}_n^{\mathrm{prec}} \in \mathbb{K}_n^{\mathrm{prec}}$, or equivalently $[\mathbf{x}_n^{\mathrm{prec}}] \in \left[\mathbb{K}_n^{\mathrm{prec}}\right]$, such that we have*

$$\mathbf{a}(\mathbf{x}_n^{\mathrm{prec}}, \mathbf{x}') = \mathbf{l}(\mathbf{x}'), \quad \forall \mathbf{x}' \in \mathbb{K}_n^{\mathrm{prec}} \quad \Longleftrightarrow \quad [\mathbf{x}']^*\mathbf{A}[\mathbf{x}_n^{\mathrm{prec}}] = [\mathbf{x}']^*\mathbf{F}, \quad \forall [\mathbf{x}'] \in \left[\mathbb{K}_n^{\mathrm{prec}}\right], \tag{41}$$

*where the finite dimensional space $\mathbb{K}_n^{\mathrm{prec}}$ is the image of $[\mathbb{K}_n^{\mathrm{prec}}]$ through (31). The space $[\mathbb{K}_n^{\mathrm{prec}}]$ is defined by*

$$[\mathbb{K}_n^{\mathrm{prec}}] := \mathbf{M}^{-\frac{1}{2}}[\widetilde{\mathbb{K}}_n^{\mathrm{prec}}] \quad \textit{with} \quad [\widetilde{\mathbb{K}}_n^{\mathrm{prec}}] := \operatorname*{span}_{0\leq k\leq n-1}\left(\widetilde{\mathbf{A}}^k\widetilde{\mathbf{F}}\right) = \operatorname{span}\left\{\widetilde{\mathbf{F}}, \widetilde{\mathbf{A}}\widetilde{\mathbf{F}}, \ldots, \widetilde{\mathbf{A}}^{n-1}\widetilde{\mathbf{F}}\right\}, \tag{42}$$

*where $\widetilde{\mathbf{A}} = \mathbf{M}^{-\frac{1}{2}}\mathbf{A}\mathbf{M}^{-\frac{1}{2}}$ and $\widetilde{\mathbf{F}} = \mathbf{M}^{-\frac{1}{2}}\mathbf{F}$, leading to $[\mathbb{K}_n^{\mathrm{prec}}] = \operatorname{span}\left\{\mathbf{M}^{-1}\mathbf{F}, \ldots, \left(\mathbf{M}^{-1}\mathbf{N}\right)^{n-1}\mathbf{M}^{-1}\mathbf{F}\right\}$.*

We remark that $[\mathbf{x}_n^{\mathrm{jac}}] \in \mathbb{C}^{\#\mathrm{dof}}$, satisfying Problem 4, also belongs to $[\mathbb{K}_n^{\mathrm{prec}}]$, *ie* $\mathbf{x}_n^{\mathrm{jac}} \in \mathbb{K}_n^{\mathrm{prec}}$. Thus, we obtain a convergence theorem for the preconditioned Krylov UWVF problem.

**Theorem 4.1.** *The preconditioned Krylov UWVF method converges minimally at the same rate than the iterative UWVF method. We have*

$$\|\mathbf{x}^h - \mathbf{x}_n^{\mathrm{prec}}\|_{\mathrm{DG}} \ \leq\ \sqrt{2}\|\mathbf{x}^h - \mathbf{x}_n^{\mathrm{jac}}\|_{L_t^2(\partial\mathcal{T})} \underset{n\to+\infty}{\longrightarrow} 0. \tag{43}$$

**Remark 4.1.** *From Proposition 3.1, the convergence rate of $\mathbf{x}_n^{\mathrm{jac}} \in \mathbb{Y}_{\mathcal{T}}^h$ to $\mathbf{x}^h \in \mathbb{Y}_{\mathcal{T}}^h$ is $\rho(\mathbf{M}^{-\frac{1}{2}}\mathbf{N}\mathbf{M}^{-\frac{1}{2}}) < 1$.*

**Remark 4.2.** *The definition of* $\mathbb{K}_n^{\text{prec}}$ *in* (42) *and the Theorem 4.1 put forward the advantage to perform a symmetric preconditioning of* $\mathbf{A}$ *with* $\mathbf{M}^{-\frac{1}{2}}$.

The convergence rate of the preconditioned Krylov UWVF (*resp.* iterative UWVF ) solver depends on the spectrum of $\widetilde{\mathbf{A}}$ (*resp.* $\mathbf{A}$), see Figure 4 (*resp.* Figure 3). The spectrum of $\widetilde{\mathbf{A}}$ is contained into the unity circle and graphically justifies the use of the convergence Proposition 3.1 on Problem 6. We believe that the presence of fewer eigenvalues close to 0 results in a faster convergence of the Krylov solver, see Figure 4. Moreover, $\widetilde{\mathbf{A}}$ has a lower condition number than $\mathbf{A}$, see Table 2, which testifies its better convergence properties.

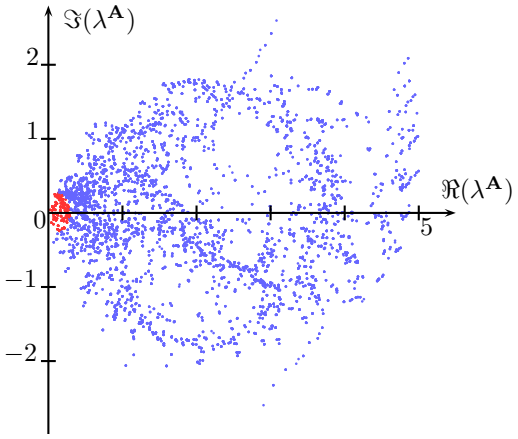


FIGURE 3. Real and complex parts, *resp.* $\Re(\lambda^{\mathbf{A}})$ and $\Im(\lambda^{\mathbf{A}})$, of the spectrum $\lambda^{\mathbf{A}}$ of matrix $\mathbf{A}$ for $\mathcal{D}_\Omega = 6$.

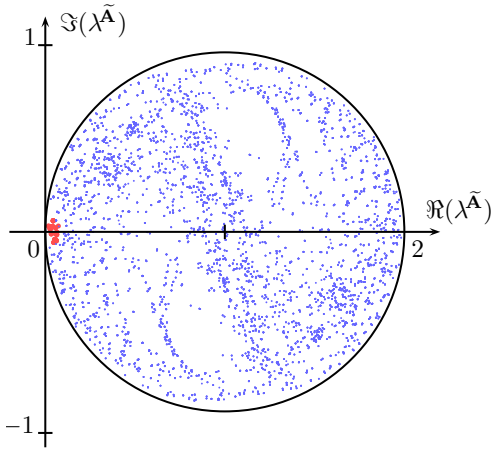


FIGURE 4. Real and complex parts, *resp.* $\Re(\lambda^{\widetilde{\mathbf{A}}})$ and $\Im(\lambda^{\widetilde{\mathbf{A}}})$, of the spectrum $\lambda^{\widetilde{\mathbf{A}}}$ of matrix $\widetilde{\mathbf{A}}$ for $\mathcal{D}_\Omega = 6$.

| $\mathcal{D}_\Omega$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\kappa(\mathbf{A})$ | 18.5 | 23.4 | 45.1 | 60.7 | 88.8 | 113 |
| $\kappa(\widetilde{\mathbf{A}})$ | 4.82 | 10.9 | 25.6 | 29.9 | 59.8 | 60.9 |

TABLE 2. Comparison of the condition number of $\mathbf{A}$ and $\widetilde{\mathbf{A}}$ according to $\mathcal{D}_\Omega$ in wavelength.

## 4.2. Basis reduction strategy

The preconditioning of the UWVF matrix $\mathbf{A}$ ensures the convergence of the preconditioned Krylov solver, see Theorem 4.1. Previous work, see [3, 15, 40, 48], have shown that plane wave basis functions can be numerically linearly-dependant in the sense that $\mathbf{M}$ has small eigenvalues. A rounding error on smallest eigenvalues of $\mathbf{M}$ gives rise to a large error on $\mathbf{M}^{-\frac{1}{2}}$ and impacts the convergence of the Krylov solver. Thus, we reduce the Galerkin space $\mathbb{Y}_{\mathcal{T}}^h$ by keeping only the eigenvectors associated to the highest eigenvalues. This will ensure a representation of $\mathbf{x} \in \mathbb{Y}_{\mathcal{T}}^h$ filtering the numerical noise in the plane wave basis, through (31). The diagonalisation of the hermitian matrix $\mathbf{M}$ representing the $L^2(\partial\mathcal{T})$ scalar product, see (35), is

$$\mathbf{M} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}^* \quad \text{with } \mathbf{T} \in \mathbb{C}^{\#\text{dof}\times\#\text{dof}} \text{ and } \mathbf{\Lambda} \in \mathbb{C}^{\#\text{dof}\times\#\text{dof}},$$

where $\mathbf{T}$ is the orthogonal eigenvector matrix and $\mathbf{\Lambda}$ is the diagonal eigenvalue matrix satisfying

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{\#\text{dof}} \quad \text{with } \lambda_{\text{i}} = \mathbf{\Lambda}_{\text{i,i}}.$$

When performing the basis reduction, the orthogonal matrix then becomes rectangular

$$(\mathbf{T}_{\rm red})_{\rm i,j} = (\mathbf{T})_{\rm i,j} \quad \text{for} \quad {\rm i} = 1, \#{\rm dof} \text{ and } {\rm j} = 1, \#{\rm dof}_{\rm red}, \quad \text{with} \quad \lambda_{\#{\rm dof}_{\rm red}} \geq \lambda_1 \varepsilon > \lambda_{\#{\rm dof}_{\rm red}+1},$$

where the small threshold coefficient $\varepsilon$ selects the largest eigenvalues. It leads to the following reduced iterative UWVF Galerkin problem. This problem involves less unknowns and is therefore less costly to solve numerically.

**Problem 7.** *Let $\mathbb{Y}^h_{\mathcal{T},\rm red}$ be the linear subspace of $\mathbb{Y}^h_{\mathcal{T}}$ parametrised by $[\mathbf{x}_{\rm red}] \in \mathbb{C}^{\#{\rm dof}_{\rm red}}$. Any element $\mathbf{x}_{\rm red} \in \mathbb{Y}^h_{\mathcal{T},\rm red}$ is given through*

$$\mathbf{x} = \sum_{\rm iglob=1}^{\#{\rm dof}} [\mathbf{x}]_{\rm iglob} \mathbf{w}^{\rm iglob}, \quad \textit{with} \quad [\mathbf{x}] := \mathbf{T}_{\rm red}\, \mathbf{\Lambda}_{\rm red}^{-\frac{1}{2}}\, [\mathbf{x}_{\rm red}]. \tag{44}$$

*Find $\mathbf{x}_{\rm red} \in \mathbb{Y}^h_{\mathcal{T},\rm red}$, or equivalently $[\mathbf{x}_{\rm red}] \in \mathbb{C}^{\#{\rm dof}_{\rm red}}$, such that we have*

$$\mathbf{a}(\mathbf{x}_{\rm red}, \mathbf{x}'_{\rm red}) = \mathbf{l}(\mathbf{x}'_{\rm red}), \quad \forall \mathbf{x}'_{\rm red} \in \mathbb{Y}^h_{\mathcal{T},\rm red} \iff [\mathbf{x}'_{\rm red}]^* \mathbf{A}_{\rm red} [\mathbf{x}_{\rm red}] = [\mathbf{x}'_{\rm red}]^* \mathbf{F}_{\rm red}, \quad \forall [\mathbf{x}'_{\rm red}] \in \mathbb{C}^{\#{\rm dof}_{\rm red}}, \tag{45}$$

*with*

$$\mathbf{A}_{\rm red} := \mathbf{\Lambda}_{\rm red}^{-\frac{1}{2}}\, \mathbf{T}^*_{\rm red}\, \mathbf{A}\, \mathbf{T}_{\rm red}\, \mathbf{\Lambda}_{red}^{-\frac{1}{2}} \quad \textit{and} \quad \mathbf{F}_{\rm red} := \mathbf{\Lambda}_{\rm red}^{-\frac{1}{2}}\, \mathbf{T}^*_{\rm red}\, \mathbf{F}. \tag{46}$$

Since $\mathbb{Y}^h_{\mathcal{T},\rm red} \subset \mathbb{Y}^h_{\mathcal{T}}$, the convergence theory established in Proposition 3.1 remains true for the reduced system (45). Therefore, Theorem 4.1 leads to a convergent iterative reduced preconditioned Krylov UWVF method.

**Remark 4.3.** *Let us notice that $\mathbf{M}$ becomes the identity matrix of dimension $\#{\rm dof}_{\rm red}$*

$$\mathbf{I}_{\rm red} := \mathbf{\Lambda}_{\rm red}^{-\frac{1}{2}} \mathbf{T}^*_{\rm red} \mathbf{M} \mathbf{T}_{\rm red} \mathbf{\Lambda}_{\rm red}^{-\frac{1}{2}},$$

*which leads to a significant memory gain, particularly for large cases ie $\left(\mathcal{D}_\Omega\right)^3 > 100^3$ in wavelength cube, since the iterative algorithm* (34) *becomes*

$$[\mathbf{x}^{\rm jac}_{n+1}] = \mathbf{N}_{\rm red} [\mathbf{x}^{\rm jac}_n] + \mathbf{F}_{\rm red}, \quad \textit{with} \quad \mathbf{N}_{\rm red} := \mathbf{\Lambda}_{\rm red}^{-\frac{1}{2}} \mathbf{T}^*_{\rm red} \mathbf{N} \mathbf{T}_{\rm red} \mathbf{\Lambda}_{\rm red}^{-\frac{1}{2}} \quad \textit{and} \quad [\mathbf{x}^{\rm jac}_0] = 0.$$

The number of basis functions after reduction, denoted by $N_{\rm red}$, depends on the value of the threshold $\varepsilon$ and on the size $\mathcal{D}_T$ of each element, see Table 3. When the threshold $\varepsilon$ is sufficiently small, *ie* $\varepsilon \leq 10^{-4}$, the convergence rate is the same, see Figure 7, leading to the same accuracy of the numerical solution visualised in Figure 6. As soon as $\varepsilon \geq 10^{-3}$, the convergence is faster in terms of iterations. But, the obtained numerical solution does not approximate the physical phenomenon anymore, see Figure 5 where $\varepsilon = 10^{-2}$. Consequently, to get a numerical solution of a given accuracy, the solver can use 16.5 GB instead of 275 GB without reduction, *ie* $N_{\rm red} = 48$ instead of $N = 196$ basis functions, see Table 4. In any reduced configuration, the iterative or the direct solver is also faster in terms of time since we perform less costly matrix vector products, see Table 4.

**Remark 4.4.** *Figures 5, 6, 7 and 12 are given for $\mathcal{D}_\Omega = 40$ and $\mathcal{D}_T = 0.25$ in wavelength, and respect the configuration mentioned in Remark 3.5.*

### 4.3. **Global algebric left-preconditioner**

The non-local aspect of wave propagation encourages to set up a global preconditioner. A popular approach has been introduced in [55]. We propose an alternative which is based on the different subsets of faces of the cubic elements.

**Definition 4.1.** *Let us consider an element $T \in \mathcal{T}$. We denote by*

| $\mathcal{D}_T$ \ $\varepsilon$ | $10^{-16}$ | $10^{-15}$ | $10^{-13}$ | $10^{-11}$ | $10^{-9}$ | $10^{-7}$ | $10^{-5}$ | $10^{-4}$ | $10^{-3}$ | $10^{-2}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.25 | 180 | 175 | 154 | 126 | 96 | 70 | 48 | 36 | 30 | 16 |
| 0.5 | 196 | 196 | 190 | 186 | 174 | 132 | 96 | 84 | 70 | 48 |
| 1 | 196 | 196 | 196 | 196 | 196 | 190 | 180 | 174 | 148 | 114 |

TABLE 3. Values of $N_{\text{red}}$ when reducing the basis of size $N = 196$ according to $\varepsilon$ and $\mathcal{D}_T$ in wavelength.

| $\varepsilon$ | $10^{-16}$ | $10^{-13}$ | $10^{-11}$ | $10^{-9}$ | $10^{-7}$ | $10^{-5}$ | $10^{-4}$ | $10^{-3}$ | $10^{-2}$ |
|---|---|---|---|---|---|---|---|---|---|
| $N_{\text{red}}$ | 180 | 154 | 126 | 96 | 70 | 48 | 36 | 30 | 16 |
| Time (s) | 4121 | 2936 | 1888 | 1085 | 496.1 | 254.9 | 141.8 | 99.68 | 25.61 |
| Memory cost (Gb) | 0.151 | 0.129 | 0.105 | 0.080 | 0.058 | 0.04 | 0.03 | 0.02 | 0.01 |

TABLE 4. Numerical data: basis reduction treshold $\varepsilon$, number of basis functions after reduction $N_{\text{red}}$, time to get the numerical solution, for the reduced Krylov solution, with a size of Krylov space set to 100, in function of the chosen basis reduction treshold $\varepsilon$ for $\mathcal{D}_\Omega = 5\lambda$, $\mathcal{D}_T = 0.25$ and $N = 196$.

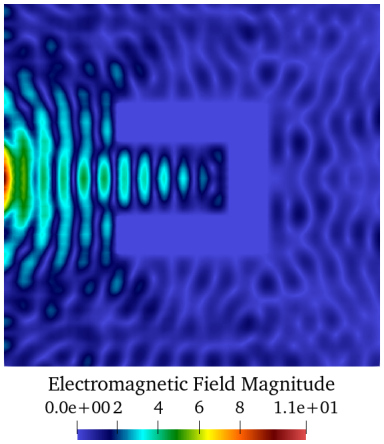


FIGURE 5. Magnitude of the electromagnetic field in a slice view of a three-dimensional cup for $N_{\text{red}} = 24$.

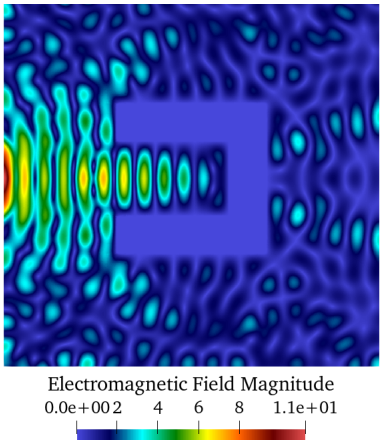


FIGURE 6. Magnitude of the electromagnetic field in a slice view of a three-dimensional cup for $N_{\text{red}} = 48$.

(i) $\mathcal{F}_{\mathsf{x}}^T$ *the set of the left and right faces of* $T$,
(ii) $\mathcal{F}_{\mathsf{y}}^T$ *the set of the front and back faces of* $T$,
(iii) $\mathcal{F}_{\mathsf{z}}^T$ *the set of the bottom and top faces of* $T$.

The set $\mathcal{T}$ is divided into one-dimensional subsets, either in $\mathsf{x}$, $\mathsf{y}$ or $\mathsf{z}$ direction, see Figure 9. These one-dimensional subdomains lead to three singular regular decompositions $\mathbf{M} - \mathbf{N}$ of matrix $\mathbf{A}$ (see [11]) which are associated to the following sesquilinear forms.

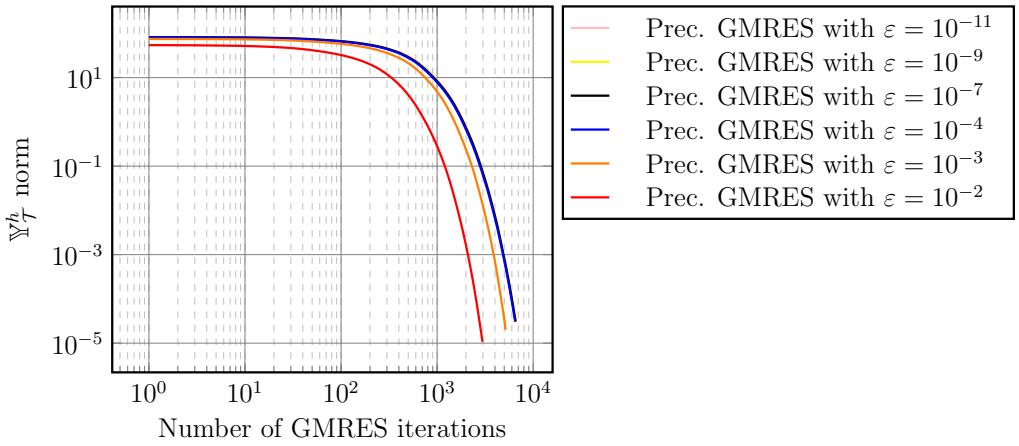


FIGURE 7. Comparison of the convergence rate of the preconditioned Krylov UWVF with basis reduction for different values of the threshold $\varepsilon$.

**Definition 4.2.** *For all* $\mathbf{x} \in \mathbb{Y}^h_{\mathcal{T}}$ *and for all* $\mathbf{x}' \in \mathbb{Y}^h_{\mathcal{T}}$*, we define the three sesquilinear forms* $\mathbf{k}^{\mathsf{x/y/z}}$ *and their associated matrices* $\mathbf{N}^{\mathsf{x/y/z}} \in \mathbb{C}^{\#\mathrm{dof}\times\#\mathrm{dof}}$ *as*

$$\mathbf{k}^{\mathsf{x/y/z}}(\mathbf{x},\mathbf{x}') := \sum_{T\in\mathcal{T}} \sum_{F\in\mathcal{F}^T_{\mathsf{x/y/z}}\cap\mathcal{F}_{\mathrm{int}}} \left(\Pi_{\mathcal{U}}\mathbf{x}, \mathcal{U}^T\mathbf{x}'\right)_{L^2_t(F)} = [\mathbf{x}']^*\mathbf{N}^{\mathsf{x/y/z}}[\mathbf{x}].$$

*Moreover, we define the associated three regular matrices as* $\mathbf{M}^{\mathsf{x/y/z}} := \mathbf{A} + \mathbf{N}^{\mathsf{x/y/z}}$.

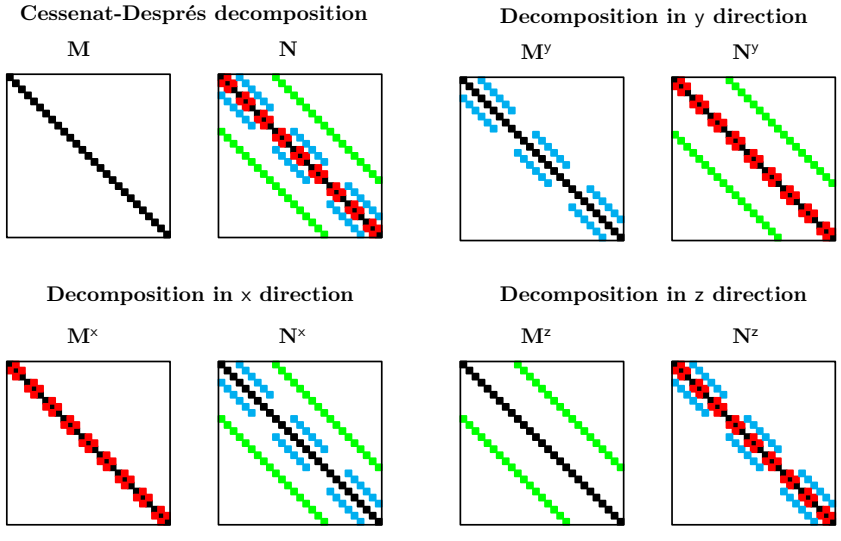


FIGURE 8. Matrices structures for the Cessenat-Després decomposition and the decompositions in $\mathsf{x}, \mathsf{y}$ or $\mathsf{z}$ direction, for $\mathcal{D}_\Omega = 3$ in wavelength.

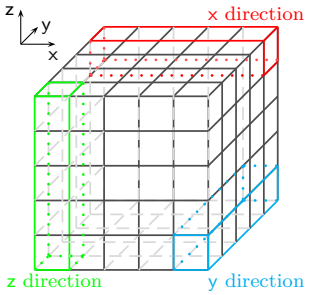


FIGURE 9. Global one-dimensional subdomains in a cube, for $\mathcal{D}_\Omega = 5$ in wavelength.

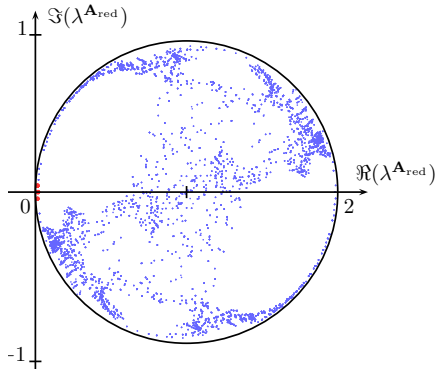


FIGURE 10. Real and complex parts, *resp.* $\Re(\lambda^{\mathbf{A}_{\text{red}}})$ and $\Im(\lambda^{\mathbf{A}_{\text{red}}})$, of the spectrum $\lambda^{\mathbf{A}_{\text{red}}}$ of $\mathbf{A}_{\text{red}}$, for $\mathcal{D}_\Omega = 6$.

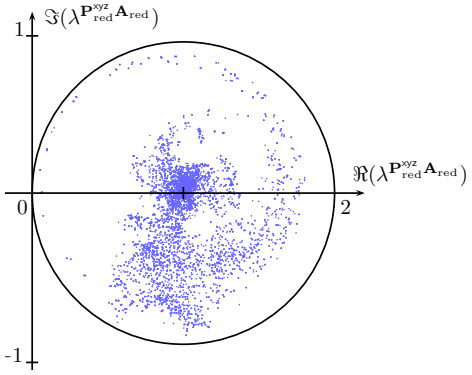


FIGURE 11. Real and complex parts, *resp.* $\Re(\lambda^{\mathbf{P}^{\text{xyz}}_{\text{red}}\mathbf{A}_{\text{red}}})$ and $\Im(\lambda^{\mathbf{P}^{\text{xyz}}_{\text{red}}\mathbf{A}_{\text{red}}})$, of the spectrum $\lambda^{\mathbf{P}^{\text{xyz}}_{\text{red}}\mathbf{A}_{\text{red}}}$ of $\mathbf{P}^{\text{xyz}}_{\text{red}}\mathbf{A}_{\text{red}}$, for $\mathcal{D}_\Omega = 6$.

Then we get three iterative schemes, leading to left preconditioners

$$\mathbf{M}^{\mathsf{x/y/z}}[\mathbf{x}_{n+1}] = \mathbf{F} + \mathbf{N}^{\mathsf{x/y/z}}[\mathbf{x}_n].$$

By applying successively each of these one-dimensional preconditioners, we get a global preconditioner involving the three directions $\mathsf{x}, \mathsf{y}$ and $\mathsf{z}$. The latter associates $\mathbf{F}$ to $[\mathbf{y}]$ thanks to intermediate solution vectors $[\mathbf{x}^0_{n+1}]$ and $[\mathbf{x}^1_{n+1}]$ through the following iterative scheme

$$\begin{cases} \mathbf{M}^{\mathsf{x}}[\mathbf{x}^0_{n+1}] = \mathbf{F} + \mathbf{N}^{\mathsf{x}}\mathbf{F}, \\ \mathbf{M}^{\mathsf{y}}[\mathbf{x}^1_{n+1}] = \mathbf{F} + \mathbf{N}^{\mathsf{y}}[\mathbf{x}^0_{n+1}], \\ \mathbf{M}^{\mathsf{z}}\mathbf{y} = \mathbf{F} + \mathbf{N}^{\mathsf{z}}[\mathbf{x}^1_{n+1}]. \end{cases}$$

Thus, we obtain an iterative method of the form $[\mathbf{x}_{n+1}] = \mathbf{P}^{\mathsf{xyz}}\mathbf{F} + \mathbf{Q}^{\mathsf{xyz}}[\mathbf{x}_n]$. Taking $[\mathbf{x}_0] = 0$, it leads to an approximation of the solution to the problem $\mathbf{A}[\mathbf{x}] = \mathbf{F}$. Therefore, this iterative method defines a global preconditioner that can be applied to the non preconditioned Problem 5,

$$\mathbf{P}^{\mathsf{xyz}}\mathbf{A}[\mathbf{x}] = \mathbf{P}^{\mathsf{xyz}}\mathbf{F}, \quad \text{where} \quad \mathbf{P}^{\mathsf{xyz}} := \left(\mathbf{M}^{\mathsf{z}}\right)^{-1} - \left(\mathbf{M}^{\mathsf{z}}\right)^{-1}\mathbf{N}^{\mathsf{z}}\left(\mathbf{M}^{\mathsf{y}}\right)^{-1} + \left(\mathbf{M}^{\mathsf{z}}\right)^{-1}\mathbf{N}^{\mathsf{z}}\left(\mathbf{M}^{\mathsf{y}}\right)^{-1}\mathbf{N}^{\mathsf{y}}\left(\mathbf{M}^{\mathsf{x}}\right)^{-1},$$

where we choose to use left-preconditioning. We can resort to the same method for the preconditioned Problem 6 or the reduced preconditioned Problem 7. We do not provide any theory ensuring the convergence of the reduced left-preconditioned Krylov solver associated to $\mathbf{P}^{\mathsf{xyz}}_{\text{red}}$. However, it is well-known that few isolated eigenvalues do not cause any problem to the Krylov solver. In the numerous numerical cases that we considered, the spectrum of $\mathbf{P}^{\mathsf{xyz}}_{\text{red}}\mathbf{A}_{\text{red}}$ is concentrated around 1 (see Figure 11), removing the small eigenvalues of $\mathbf{A}_{\text{red}}$ (in red in Figure 10) which are slowing down the convergence (see the blue curve in Figure 12). The results shown in Figure 12 point out the smaller amount of Krylov iterations when using the reduced left-preconditioned matrix $\mathbf{P}^{\mathsf{xyz}}_{\text{red}}\mathbf{A}_{\text{red}}$.

## CONCLUSION

In this paper, we have introduced an heterogeneous iterative Trefftz method solving three-dimensional Maxwell equations. As all iterative methods, the present solver does not need to store the LU factorisation of $\mathbf{A}$ and can be matrix-free. The latter speed-up accelerates the convergence such that it takes a few minutes

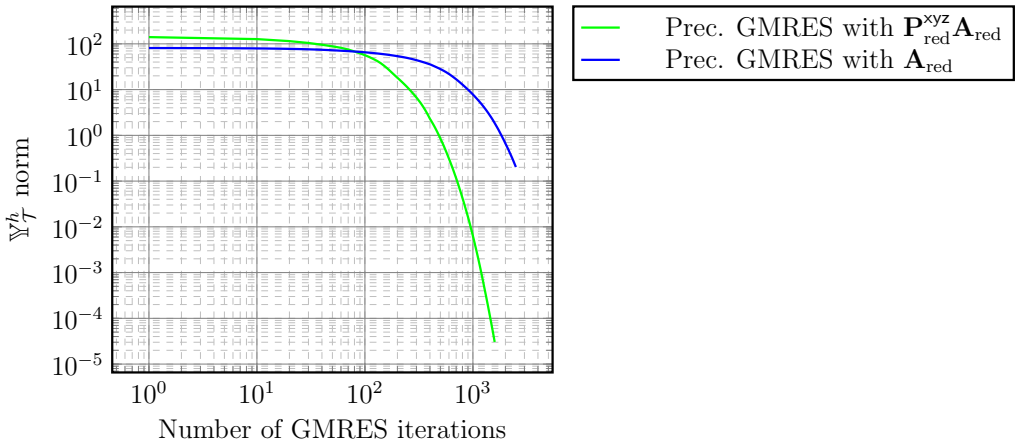


FIGURE 12. Comparison of the convergence rate of the preconditioned Krylov UWVF using either $\mathbf{A}_{\text{red}}$ or $\mathbf{P}_{\text{red}}^{\text{xyz}}\mathbf{A}_{\text{red}}$.

to solve Maxwell equations on small heterogeneous domains, *i.e.* up to $\mathcal{D}_\Omega = 40$. This memory gain has been strengthened thanks to a basis reduction proposed in Section 4.2. This compression method turned out to be a successful development, reducing the memory cost of the heterogeneous preconditioned Krylov UWVF solver. It enables to consider large complex geometries, see Figure 13. Nevertheless, this basis reduction should be applied reasonably (*ie* a good choice is $\varepsilon = 10^{-6}$) to use enough basis functions to describe the numerical solution. Moreover, the introduced Trefftz Krylov method is accelerated thanks to a new global preconditioner.

However, on some configurations, direct methods outperform iterative solvers since they can deal with many right-hand sides, see [23]. Judicious choices should then be made depending on the numerical case. But research on Krylov solvers efficiently dealing with multiple right-hand sides remains an active topic, see [8, 25]. It could then be preferable to use Krylov methods instead of direct solvers in the future.

To end this paper, let us consider an industrial application which consists of a radar echo on a boat. This large heterogeneous configuration is a boat surrounded by an homogeneous domain of size $\mathcal{D}_\Omega = 24 \times 61 \times 154$ in wavelength and satisfying $Z_{\partial\Omega} = 1$. The boat has a perfect metal surface, such that the incident electromagnetic wave perfectly reflects when striking its top, see Figure 13. This large three-dimensional case is sped up thanks to a Cessenat-Després preconditioner. It also requires a restart strategy to fit into a 1 Terabyte Broadwell Intel Xeon CPU E5-2650 node. Our ways of checking the accuracy of this result are limited due to its dimensions, see Table 5. We ensure that the Krylov residual is $10^{-3}$. The large dimensions, and the heterogeneity of this numerical case both emphasize the robustness of the developed iterative Trefftz method.

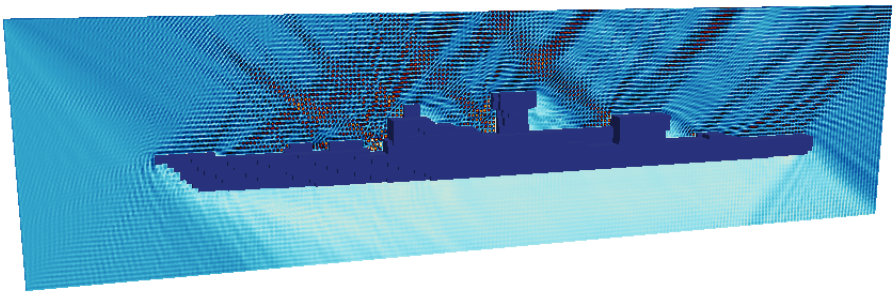

FIGURE 13. Visualisation of an electromagnetic wave striking the top of a boat.

| #elem | $N_{\text{red}}$ | $N$ | $\#\text{dof}_{\text{red}}$ | #dof | Krylov iterations | Memory cost | Total duration |
|---|---|---|---|---|---|---|---|
| $> 14.4 \times 10^6$ | 46 | 52 | $> 663 \times 10^6$ | $> 750 \times 10^6$ | 473 | $\approx 836$ GB | 37.8 hours |

TABLE 5. Numerical data and results associated to the boat case, see Figure 13.

## REFERENCES


[1] M. Ainsworth. Dispersive properties of high-order Nédélec/edge Element approximation of the time-harmonic Maxwell equations. Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 362(1816):471–491, 2004.

[2] 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(1–3):5–40, 2006.

[3] 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, 441:110449, September 2021.

[4] M. Bebendorf. Approximation of Boundary Element matrices. Numerische Mathematik, 86(4):565–589, 2000.

[5] O. P. Bruno and L. A. Kunyansky. A fast, high-order algorithm for the solution of surface scattering problems: basic implementation, tests, and applications. Journal of Computational Physics, 169(1):80–110, 2001.

[6] A. Buffa and R. Hiptmair. Galerkin Boundary Element Methods for electromagnetic scattering. In Topics in computational wave propagation, pages 83–124. Springer, 2003.

[7] A. Buffa and P. Monk. Error estimates for the Ultra Weak Variational Formulation of the Helmholtz equation. Mathematical Modelling and Numerical Analysis, 42(6):925–940, 2008.

[8] H. Calandra, S. Gratton, R. Lago, X. Vasseur, and L. M. Carvalho. A modified block flexible GMRES method with deflation at each iteration for the solution of non-Hermitian linear systems with multiple right-hand sides. SIAM Journal on Scientific Computing, 35(5):S345–S367, 2013.

[9] 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.

[10] O. Cessenat and B. Després. Application of an Ultra Weak Variational Formulation of elliptic PDE to the two-dimensional Helmholtz problem. SIAM J. Num. Analysis, 35(1):255–299, 1998.

[11] P. G. Ciarlet. Introduction à l'analyse numerique matricielle et à l'optimisation PG Ciarlet. Masson, 1990.

[12] B. Cockburn, J. Gopalakrishnan, and R. Lazarov. Unified Hybridization of Discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems. SIAM Journal on Numerical Analysis, 47(2):1319–1365, 2009.

[13] G. Cohen. Higher-order numerical methods for transient wave equations. Springer Science & Business Media, 2001.

[14] G. Cohen and S. Pernet. Finite Element and Discontinuous Galerkin methods for transient wave equations. Springer, 2017.

[15] S. Congreve, J. Gedicke, and I. Perugia. Numerical investigation of the conditioning for plane wave Discontinuous Galerkin methods. In European Conference on Numerical Mathematics and Advanced Applications, pages 493–500. Springer, 2017.

[16] E. Darve. The Fast Multipole Method: numerical implementation. Journal of Computational Physics, 160(1):195–240, 2000.

[17] E. Darve and P. Havé. Efficient Fast Multipole Method for low-frequency scattering. Journal of Computational Physics, 197(1):341–363, 2004.

[18] B. Després. Sur une Formulation Variationnelle Ultra Faible. Comptes Rendus de l'Académie des Sciences, Série I 318:939–944, 1994.

[19] V. Dolean, H. Fol, S. Lanteri, and R. Perrussel. Solution of the time-harmonic Maxwell equations using Discontinuous Galerkin methods. Journal of computational and applied mathematics, 218(2):435–445, 2008.

[20] V. Dolean, M. J. Gander, S. Lanteri, J.-F. Lee, and Z. Peng. Effective transmission conditions for Domain Decomposition Methods applied to the time-harmonic curl–curl Maxwell's equations. Journal of computational physics, 280:232–247, 2015.

[21] M. Embree. How descriptive are GMRES convergence bounds. NA Report 99, 8, 1999.

[22] C. Farhat, R. Tezaur, and P. Weidemann-Goiran. Higher-order extensions of a Discontinuous Galerkin method for mid-frequency Helmholtz problems. International journal for numerical methods in engineering, 61(11):1938–1956, 2004.

[23] F. Faucher and O. Scherzer. Adjoint-state method for Hybridizable Discontinuous Galerkin discretization, application to the inverse acoustic wave problem. Computer Methods in Applied Mechanics and Engineering, 372:113406, 2020.

[24] G. Gabard. Discontinuous Galerkin methods with plane waves for time-harmonic problems. Journal of Computational Physics, 225:1961–1984, 2007.

[25] L. Giraud, Y.-F. Jing, and Y.-F. Xiang. A block minimum residual norm subspace solver with partial convergence management for sequences of linear systems. SIAM Journal on Matrix Analysis and Applications, 2022.

[26] C. Gittelson and R. Hiptmair. Dispersion analysis of plane wave discontinuous methods. International Journal for Numerical Methods in Engineering, 98(5):313–323, 2014.

[27] C. Gittelson, R. Hiptmair, and I. Perugia. Plane wave Discontinuous Galerkin methods: analysis of the $h$-version. Mathematical Modelling and Numerical Analysis, 43:297–331, 2009.

[28] J. S. Hesthaven and T. Warburton. Nodal Discontinuous Galerkin methods: algorithms, analysis, and applications. Springer Science & Business Media, 2007.

[29] R. Hiptmair. Finite Elements in computational electromagnetism. Acta Numerica, 11:237–339, 2002.
[30] R. Hiptmair, A. Moiola, and I. Perugia. Plane wave Discontinuous Galerkin methods for the 2-D Helmholtz equation: analysis of the p-version. SIAM Journal on Numerical Analysis, 49(1):264–284, 2011.
[31] R. Hiptmair, A. Moiola, and I. Perugia. A survey of Trefftz methods for the Helmholtz equation. In Building bridges: connections and challenges in modern approaches to numerical partial differential equations, pages 237–279. Springer, 2016.
[32] R. Hiptmair, A. Moiola, I. Perugia, and C. Schwab. Approximation by harmonic polynomials in star-shaped domains and exponential convergence of Trefftz *hp*-DGFEM. Mathematical Modelling and Numerical Analysis, 48:727–752, 2014.
[33] P. Houston, I. Perugia, A. Schneebeli, and D. Schötzau. Interior Penalty method for the indefinite time-harmonic Maxwell equations. Numerische Mathematik, 100(3):485–518, 2005.
[34] T. Huttunen, M. Malinen, and P. Monk. Solving Maxwell's equations using the Ultra Weak Variational Formulation. Journal of Computational Physics, 223(2):731–758, 2007.
[35] F. Ihlenburg and I. Babuska. Finite Element solution of the Helmholtz equation with high wave number – part i: the h-version of the FEM. Computers Math. Applic., 30(9):9–37, 1995.
[36] F. Ihlenburg and I. Babuska. Finite Element solution of the Helmholtz equation with high wave number – part ii: the h-p version of the FEM. SIAM J. Numer. Anal., 34(1):315–358, 1997.
[37] J. M. Jin. The Finite Element Method in Electromagnetics, Second Edition. John Wiley & Sons, New York, 2002.
[38] S. Kurz, O. Rain, and S. Rjasanow. The adaptive cross-approximation technique for the 3-D Boundary Element Method. IEEE transactions on Magnetics, 38(2):421–424, 2002.
[39] T. Luostari, T. Huttunen, and P. Monk. The Ultra Weak Variational Formulation using bessel basis functions. Communications in Computational Physics, 11(2):400–414, 2012.
[40] T. Luostari, T. Huttunen, and P. Monk. Improvements for the Ultra Weak Variational Formulation. International Journal for Numerical Methods in Engineering, 94(6):598–624, 2013.
[41] A. Moiola, R. Hiptmair, and I. Perugia. Plane wave approximation of homogeneous Helmholtz solutions. Z. Angew. Math. Phys., 62:809–837, 2011.
[42] P. Monk. Finite Element Methods for Maxwell's Equations. Numerical Analysis and Scientific Computations. Clarendon Press, 2003.
[43] D. Moro, N. C. Nguyen, and J. Peraire. A Hybridized Discontinuous Petrov-Galerkin scheme for scalar conservation laws. International journal for numerical methods in engineering, 91(9):950–970, 2012.
[44] J.-C. Nédélec. A new family of mixed Finite Elements in $\mathbb{R}^3$. Numerische Mathematik, 50(1):57–81, 1986.
[45] J.-C. Nédélec. Acoustic and electromagnetic equations: integral representations for harmonic problems. Springer Science & Business Media, 2001.
[46] N. C. Nguyen, J. Peraire, and B. Cockburn. Hybridizable Discontinuous Galerkin methods for the time-harmonic Maxwell's equations. Journal of Computational Physics, 230(19):7151–7175, 2011.
[47] S. Pernet, N. Serdiuk, M. Sirdey, and S. Tordeux. Discontinuous Galerkin method based on Riemann fluxes for the time domain Maxwell system, 2021.
[48] E. Perrey-Debain. Plane wave decomposition in the unit disc: convergence estimates and computational aspects. Journal of Computational and Applied Mathematics, 193(1):140–156, 2006.
[49] B. Pluymers, B. Hal, D. Vandepitte, and W. Desmet. Trefftz-based methods for time-harmonic acoustics. Archives of Computational Methods in Engineering, 14:343–381, 01 2007.
[50] Y. Saad. Krylov subspace methods for solving large unsymmetric linear systems. Mathematics of computation, 37(155):105–126, 1981.
[51] Y. Saad. Iterative Methods for Sparse Linear Systems. PWS Pblishing Company, Boston, 1996.
[52] Y. Saad and M. H. Schultz. GMRES: A Generalized Minimal RESidual algorithm for solving non-symmetric linear systems. SIAM Journal on scientific and statistical computing, 7(3):856–869, 1986.
[53] S. A. Sauter and C. Schwab. Boundary Element Methods. Springer-Verlag, Berlin-Heidelberg, 2011.
[54] A. Taflove. Computational Electrodynamics. The Finite-Difference Time-Domain Method. Artech house, Inc., Norwood, MA 02062, 1995.
[55] A. Vion and C. Geuzaine. Double sweep preconditioner for optimized Schwarz methods applied to the Helmholtz problem. J. Comput. Phys., 266:171–190, 2014.
[56] D. Wang, R. Tezaur, J. Toivanen, and C. Ferhat. Overview of the Discontinuous Enrichment Method, the Ultra Weak Variational Formulation, and the Partition of Unity Method for the acoustic scattering in the medium frequency regime and performance comparisons. International Journal for Numerical Methods in Engineering, 89:403–417, 2012.
[57] N. Zerbib. Méthodes de Sous-Structuration et de Décomposition de Domaine pour la Résolution des Équations de Maxwell : Application au Rayonnement d'antenne dans un Environnement Complexe. PhD thesis, National Institute for Applied Sciences (INSA), INSA Toulouse, 2006.
[58] K. Zhao, M. N. Vouvakis, and J.-F. Lee. The adaptive cross approximation algorithm for accelerated method of moments computations of EMC problems. IEEE transactions on electromagnetic compatibility, 47(4):763–773, 2005.

[59] L. Zhu, E. Burman, and H. Wu. Continuous Interior Penalty Finite Element Method for Helmholtz equation with high wave number: One dimensional analysis. preprint available at arXiv:1211.1424.