Local strategies for improving the conditioning of the plane-wave Ultra-Weak Variational Formulation

Hélène Barucq, Abderrahmane Bendali, Julien Diaz, Sébastien Tordeux

Local strategies for improving the conditioning of the plane-wave Ultra-Weak Variational Formulation

Hélène Barucq[a], Abderrahmane Bendali[b,*], Julien Diaz[a], Sébastien Tordeux[a]

[a] *Makutu, Inria & E2S UPPA, CNRS UMR 5142, France*
[b] *Univ. Toulouse, INSA-Toulouse, IMT UMR CNRS 5219, 135 Avenue de Rangueil, F-31400 Toulouse (France)*

Abstract

Element-wise techniques based on SVD or QR Decomposition completely get rid of the usual ill-conditioning inherent to the plane-wave discretizations of the Ultra-Weak Variational Formulation (UWVF) for the Helmholtz equation. Associated preconditioning strategies lead to very low condition numbers of the corresponding linear system matrices, without any limitation on the number of plane waves per element. In addition, some of these procedures have the advantage of considerably reducing the size of the final system to be solved without altering the accuracy of the numerical solution.

Keywords: Helmholtz, UWVF, Plane Waves, SVD, QR
2008 MSC: 65-05, 65N30

1. Introduction

Mechanical waves can provide reliable information about the medium in which they propagate through the measurement of various related parameters such as velocities, amplitudes, phases, etc. For this reason, their numerical simulation is used in many fields of application such as medical imaging, exploration of the subsurface, damage control of devices, etc. Because of its ability to reconstruct the composition of the propagation environment, the

[*] Corresponding author

Email addresses: helene.barucq@inria.fr (Hélène Barucq), abendali@insa-toulouse.fr (Abderrahmane Bendali), julien.diaz@inria.fr (Julien Diaz), sebastien.tordeux@inria.fr (Sébastien Tordeux) solution of wave equations is therefore at the heart of numerical technologies that can possibly replace or support experimental measurements. Wave equations can be formulated in the time or in the frequency domain. For most applications, they are involved in an inversion process aimed at recovering the physical parameters of the propagation medium. Full-Waveform Inversion (FWI), as a high-definition inversion technique, promotes the use of frequency formulation because only a small number of frequencies are sufficient to solve the inverse problem. In addition, FWI is based upon the direct and adjoint problem, represented by the same linear system, with multiple right-hand sides, making the use of a direct solver highly desirable. At last, complex physical parameters depending on the frequency, like e.g. the attenuation, are then easier to handle. The frequency domain formulation of wave propagation problems has therefore attracted more and more the attention of both industrial and academic researchers (see [1] and the references therein).

Basically, in acoustics, the propagation is governed by the wave equation, which reduces to the Helmholtz equation in the frequency domain. Its solution can be tackled by various numerical methods. Continuous as well as Discontinuous Finite Elements are among the most popular because they offer the possibility of using unstructured computational grids that are well suited for considering complex and realistic media with topography changes and heterogeneous material parameters.

Although the Helmholtz equation is elliptic, it retains some characteristics of the hyperbolicity of the wave equation from which it is derived. Its numerical solution may thus be hindered by a specific instability called "*numerical pollution effect*" [2, 3]. To limit this flaw, different approaches have been proposed. They can be divided into two main classes. The first one includes the methods exploiting the fact that local high-order polynomial approximations reduce this specific instability. The second class, constituted by the generically termed wave-based methods, builds on the idea that basis functions, incorporating the physical characteristics of the problem as local solutions of the Helmholtz equation, face the numerical pollution effect more efficiently. The comparison between high-order polynomial finite element methods and wave-based methods is the subject of very interesting discussions. One of the first accepted advantages of wave-based methods is to use a priori a number of degrees of freedom that is much smaller than that required by polynomial methods. For instance, there is a numerical comparison in [4] between the use of Lobatto's shape functions and a wave-based Discontinuous Galerkin (DG) method introduced in [5]. This study shows that both methods have comparable performances provided a static condensation is used in the Finite Element Method, otherwise the conditioning of the related linear system degrades. For error levels around 1%, the conditioning of the wave-based method of [5] is comparable to the one of the polynomial method in [4] but tends also to deteriorate when more accuracy is sought. It is worth noting that conditioning tends to degrade also in the case of polynomial approximations if a high level of precision is required [6].

Ill-conditioning is an issue actually shared by all the approximation methods used to solve the Helmholtz equation. Its effective reduction is still a challenge. Here, we focus on the Ultra-Weak Variational Formulation (UWVF), which was introduced in [7], developed in [8, 9], and later extensively studied in [10, 11, 12, 13]. This method was devised building on a specific Domain Decomposition Method (DDM) for the Helmholtz equation [14]. It uses globally discontinuous basis functions that are exact local solutions of the interior PDE. In this respect, this method ranks among the Trefftz methods, and more importantly, is one of the very few approaches, which guarantees that the local problems, associated with suitable block Gaussian eliminations are well-posed (see Proposition 3 and Remark 4 below). Therefore, the solution is not affected by parasitic resonances, when solving the discrete system by an ad hoc direct solver.

Conditioning issues of plane wave discretizations have been analyzed by many authors. In [15], a numerical study clearly highlights the correlation between the mesh and the condition number. Inspired by [16], the idea of enriching the wave basis with evanescent waves is also investigated in [15]. However, if the error is improved, ill-conditioning remains present. A numerical study in [17] dedicated to the plane wave discretization of a DG formulation of the two-dimensional Helmholtz equation led to several interesting answers: (a) several factors can damage the condition number of the mass matrix related to a plane wave basis (size and aspect ratio of the element, number of plane waves, a pair of plane waves propagating in nearly the same direction, etc.); (b) the orthogonalization of the plane-wave basis (by the modified Gram-Schmidt algorithm relatively to the Hermitian part of the elemental matrix related to the sesquilinear form of the variational formulation) outstandingly improves the condition number of the final linear system matrix. Unfortunately, the orthogonalization process fails beyond a critical limit of the number of plane-wave basis functions.

In this paper, we develop two techniques at the element level, respectively based on a Singular Value Decomposition (SVD) and a Householder

QR Decomposition with pivoting, to deal with the ill-conditioning of the plane-wave UWVF. Although these approaches use a switch to orthogonal bases, actually only partly for the QR Decomposition, their basic principle is a regularization of the operator associating to each plane wave its Fourier impedance trace on the boundary of the element. This regularization consists in slightly perturbing the operator to make it having a positive lower bound, or in other words, to prevent its inverse from blowing up. The paper is organized as follows. We first introduce the UWVF for the Helmholtz equation considered in this paper. In particular, we prove that the corresponding discrete system can be inverted using block Gaussian eliminations. Next, we highlight ill-conditioning by considering a waveguide problem in which the fundamental mode is propagating. Basically, we bring out some features more or less already pointed out in the literature: (a) the error and the condition number are highly correlated; (b) the conditioning of the global system is closely related to the condition number of some specific elemental matrices. Then, we describe the two strategies employed for reducing the global condition number. We then achieve a performance assessment by conducting different numerical experiments including a difficult case where evanescent waves are created by a singularity in the geometry of the propagation domain. We end the paper by some concluding remarks including issues which are presently under consideration.

2. The plane-wave UWVF

In this section, we define the model problem we are considering. It is posed in terms of the Helmholtz equation in a bounded domain (subsection 2.1). We show how to derive the UWVF used for its numerical solution (subsection 2.2). In particular, we indicate what distinguishes this ultra weak formulation from the original one, proposed by Cessenat and Després [8]. We use plane wave basis functions to discretize the UWVF and, in addition to proving that the resulting discrete problem can be inverted by block Gaussian eliminations, we recall the main properties of the elemental and global matrices involved in the construction of the final linear system (subsection 2.3).

2.1. The boundary-value problem

For simplicity, we hereafter limit ourselves to the 2D case. However, the whole study extends to the 3D case in a straightforward way. Let Ω be a polygonal domain. We consider the following model boundary-value problem set in terms of the usual Helmholtz equation

$$\left\{ \begin{array}{l} \Delta u + \kappa^2 u = 0 \text{ in } \Omega, \\ \partial_{\boldsymbol{n}} u - i\kappa Y u = g \text{ on } \partial\Omega, \end{array} \right. \tag{1}$$

where $\partial\Omega$ stands for the boundary of Ω and $\boldsymbol{n}$ is the unit normal on $\partial\Omega$ directed towards the exterior of Ω. Constant $\kappa > 0$. It stands for the propagation wavenumber: $\kappa = \omega/c$, ω being the angular frequency and c the wave velocity. The time dependence is in $e^{-i\omega t}$ and suppressed by linearity. The piecewise constant function Y satisfies $Y \geq 0$ almost everywhere on $\partial\Omega$. It plays the role of the surface admittance of the boundary.

It can be established from many sources (see, e.g., [18, 19, 20, 21]) that, if $Y > 0$ on a neighborhood of some point of $\partial\Omega$, and if g is in $L^2(\partial\Omega)$, then problem (1) admits one and only one solution u in $H^1(\Omega)$. Despite its simplicity, this problem contains all the features related to the conditioning issues we are interested in.

2.2. The Ultra-Weak Variational Formulation

The UWVF is a special Galerkin method, which was introduced in Després's pioneering work [7]. Various equivalent versions of its statement can be found in the literature. They differ from each other according to: (a) the normalization of the incoming and the outgoing Fourier impedance traces; (b) the use of test functions or their complex conjugate; (c) the way in which the unitary operator associating incoming with outgoing traces is involved in the equations. We briefly present the formulation used here. We find its derivation more "*straightforward*" than the original one [8]. We also give the early one in a form adapted to the present context, and show that the two formulations are equivalent, both at the continuous and at the discrete level.

Contrary to other methods such as continuous finite element methods (see, e.g., [22, 23]), the UWVF, as DG methods (see, e.g., [24, 25]), requires an a priori mesh to be posed. For simplicity, we assume that the elements T of this mesh $\mathcal{T}$ (really a non-overlapping decomposition of Ω) are polygonal. To be set also, the UWVF requires a regularity of the above solution u higher than the usual variational regularity $H^1(\Omega)$. As in other studies making use of the same framework (see, e.g., [25]), we assume that Ω and g are such that $u \in H^{3/2+\eta}(\Omega)$ for an η such that $0 < \eta < 1/2$. We can thus define outgoing traces as

$$x_T = \frac{1}{2}\left(\frac{1}{i\kappa}\partial_{\boldsymbol{n}_T} u_T + u_T\right) \in L^2(\partial T), \quad T \in \mathcal{T},$$

where $\boldsymbol{n}_T$ denotes the outward unit normal on the boundary ∂T of T, and $u_T = u|_T$.

Let us now introduce the operator $\mathcal{U}_T : L^2(\partial T) \to L^2(\partial T)$, implicitly defined for a generic element $y_T \in L^2(\partial T)$ by $\mathcal{U}_T y_T = \frac{1}{2}\left(-\frac{1}{i\kappa}\partial_{\boldsymbol{n}_T} v_T + v_T\right)$ where v_T is the unique solution to the following boundary-value problem

$$\left\{ \begin{array}{l} \Delta v_T + \kappa^2 v_T = 0 \text{ in } T, \\ \partial_{\boldsymbol{n}_T} v_T + i\kappa v_T = 2i\kappa y_T \text{ on } \partial T. \end{array} \right. \tag{2}$$

Of course, for $y_T = x_T$, the uniqueness of the solution to problem (2) ensures that $v_T = u_T$. The operator $\mathcal{U}_T$ binds the outgoing trace y_T and the incoming one $\mathcal{U}_T y_T$ for solutions to the Helmholtz equation. It plays the same role as does the Dirichlet-to-Neumann operator for the Laplace equation in several issues like for instance DDMs (see, e.g., [26] for DDMs in the static case and, e.g., [27] for a DDM in the case of Helmholtz equation). Actually, it is possible to avoid the explicit use of the normal derivative $\partial_{\boldsymbol{n}_T} v_T$ and to define $\mathcal{U}_T y_T$ by means of the following relationship

$$v_T = y_T + \mathcal{U}_T y_T.$$

At this stage, it is worth recalling some well-known interesting features of the UWVF leading to its use for numerically solving problem (1). The UWVF is a Trefftz method, or as explained above in the introduction, a wave-based method. As is, it is aimed to take into account the fast oscillations of the solution more efficiently than a plain finite element approximation by gluing local solutions v_T of the Helmholtz equation on each element $T \in \mathcal{T}$ and imposing them to satisfy the boundary conditions on $\partial\Omega$. The matching conditions on the interfaces and the boundary conditions are expressed in terms of the outgoing and incoming traces and are not imposed pointwise but in a variationnal manner. To get more insight about the incoming and outgoing traces, let us consider a superposition of two plane waves $\alpha_T^+ \exp(+i\kappa x \cdot \boldsymbol{n}_T)$ and $\alpha_T^- \exp(-i\kappa x \cdot \boldsymbol{n}_T)$, propagating at normal incidence to an edge F of ∂T

$$v_T(x) = \alpha_T^+ \exp(+i\kappa x \cdot \boldsymbol{n}_T) + \alpha_T^- \exp(-i\kappa x \cdot \boldsymbol{n}_T).$$

Since the time-dependence is implicitly assumed in $e^{-i\omega t}$ and $\boldsymbol{n}_T$ is directed towards the exterior of T, $\alpha_T^+ \exp(+i\kappa x \cdot \boldsymbol{n}_T)$ and $\alpha_T^- \exp(-i\kappa x \cdot \boldsymbol{n}_T)$ are respectively propagating towards the exterior and the interior of T (Note that, for the sake of simplicity, we do not distinguish between a point x in the plane and its radius vector $\boldsymbol{r}_x$, column-vector of two components given by the cartesian coordinates of x). One can easily verify that in this particular case, the outgoing and incoming traces

$$\begin{aligned} \frac{1}{2i\kappa}\left(\partial_{\boldsymbol{n}_T} v_T + i\kappa v_T\right) &= \alpha_T^+ \exp\left(+i\kappa x \cdot \boldsymbol{n}_T\right), \\ -\frac{1}{2i\kappa}\left(\partial_{\boldsymbol{n}_T} v_T - i\kappa v_T\right) &= \alpha_T^- \exp\left(-i\kappa x \cdot \boldsymbol{n}_T\right), \end{aligned}$$

exactly coincide with respectively the trace of the outgoing wave from T and the trace of the incoming wave in T through edge F of T. This explains the outgoing and incoming trace terminology.

Let us come back to local solutions p_T and q_T to problem (2) corresponding to general outgoing traces y_T and $z_T \in L^2\left(\partial T\right)$ respectively. It is easily seen that $\mathcal{U}_T$ is symmetrical and unitary in the following sense

$$\int_{\partial T} y_T \mathcal{U}_T z_T ds = \int_{\partial T} z_T \mathcal{U}_T y_T ds, \quad \int_{\partial T} \mathcal{U}_T z_T \overline{\mathcal{U}_T y_T} ds = \int_{\partial T} z_T \overline{y_T} ds,$$

where ds is the usual Lebesgue measure on ∂T.

The basic principle of the UWVF we use consists in recasting problem (1) by imposing to local solutions of the Helmholtz equation u_T in each $T \in \mathcal{T}$ to satisfy the usual matching conditions

$$u_T = u_L, \quad \partial_{\boldsymbol{n}_T} u_T + \partial_{\boldsymbol{n}_L} u_L,$$

on each edge F shared by an element T and an element L of $\mathcal{T}$, and the boundary condition of problem (1) in terms of the outgoing and incoming traces. Then, the UWVF can be stated by expressing all these matching and boundary conditions variationally as follows (see for example [8, 28] for the details)

$$\left\{ \begin{array}{l} x = \left(x_T\right)_{T\in\mathcal{T}} \in X = \prod_{T\in\mathcal{T}} L^2\left(\partial T\right), \ \forall y = \left(y_T\right)_{T\in\mathcal{T}} \in X, \\ \sum\limits_{T\in\mathcal{T}} \displaystyle\int_{\partial T} \left(\mathcal{U}_T x_T - \mathcal{F}_T x\right) y_T \ ds = \sum\limits_{T\in\mathcal{T}} \sum\limits_{F_T^{\partial\Omega}} \displaystyle\int_{F_T^{\partial\Omega}} w_T y_T \ ds, \end{array} \right. \tag{3}$$

where

$$\int_{\partial T} \mathcal{F}_T x \ y_T \ ds = \sum_{L\in\mathcal{F}_T^{\mathcal{I}}} \int_{F_T^L} x_L y_T \ ds + \sum_{F_T^{\partial\Omega}} \int_{F_T^{\partial\Omega}} R_T x_T y_T ds.$$

Above, $\mathcal{F}_T^{\mathcal{I}}$ is the set of internal faces F_T^L shared by elements T and L. External faces of T are part of $\partial\Omega$ and denoted by $F_T^{\partial\Omega}$. The UWVF makes use of the reflection coefficient $R_T = (1-Y)/(1+Y)$, instead of the boundary admittance Y involved in the boundary condition of problem (1). Finally, to simplify the notation, we have set $w_T = (-1/i\kappa\,(1+Y))g$, on external faces $F_T^{\partial\Omega}$ and $w_T = 0$ on internal faces $F_T^L \in \mathcal{F}_T^{\mathcal{I}}$.

To write the early formulation of Cessenat and Després [8] in the present context, it is enough to take $\overline{\mathcal{U}_T y_T}$ as test function, in place of y_T in (3), and to use the fact that $\mathcal{U}_T$ is unitary

$$\begin{cases} x = (x_T)_{T\in\mathcal{T}} \in X, \ \forall y = (y_T)_{T\in\mathcal{T}} \in X, \\ \sum\limits_{T\in\mathcal{T}} \displaystyle\int_{\partial T} \left(x_T\overline{y_T} - \mathcal{F}_T x \overline{\mathcal{U}_T y_T}\right)\, ds = \sum\limits_{T\in\mathcal{T}} \sum\limits_{F_T^{\partial\Omega}} \int_{F_T^{\partial\Omega}} w_T \overline{\mathcal{U}_T y_T}\, ds. \end{cases} \tag{4}$$

By the very means of its derivation, formulation (4) is clearly equivalent to the one stated in (3). One of the advantages of formulation (4) is that it is actually set as a fixed-point problem in X

$$(I - \mathcal{U}^*\mathcal{F})\, x = w$$

where I is the identity operator, $\mathcal{U}$ and $\mathcal{F}$ are of norm ≤ 1 defined by means of the scalar product of X,

$$(\mathcal{F}x, \overline{z})_X = \sum_{T\in\mathcal{T}} \int_{\partial T} \mathcal{F}_T x \overline{z_T}\, ds, \quad (\mathcal{U}x, \overline{z})_X = \sum_{T\in\mathcal{T}} \int_{\partial T} \mathcal{U}_T x_T \overline{z_T}\, ds,$$

and w represented by

$$(w, y)_X = \sum_{T\in\mathcal{T}} \int_{\partial T} w_T y_T ds.$$

Operator $\mathcal{U}^*$ is the adjoint operator of $\mathcal{U}$, defined through $(\mathcal{U}^* x, \overline{y})_X = \left(x, \overline{\mathcal{U}y}\right)_X$, for all $x, y \in X$. We point out that $(x,y) \to (x,y)_X$ is the bilinear form underlying the scalar product $(x,\overline{y})_X$ of two elements of X.

Remark 1. It is extremely important to note that, contrary to operator $\mathcal{F}$, operator $\mathcal{U}$ is "diagonal" in the meaning that the component $(\mathcal{U}x)_T$ depends only on x_T

$$(\mathcal{U}x)_T = \mathcal{U}_T x_T, \ \forall T \in \mathcal{T}.$$

This will be fundamental to our "*local*" strategy for improving the condition number of the linear system matrix obtained from the discretization of formulation (3).

2.3. Plane wave discretization

The usual Galerkin wave-based method for solving formulation (3) consists in using a plane-wave basis

$$\left\{v_{\boldsymbol{d}_j}\right\}_{1\leq j\leq p}, \quad v_{\boldsymbol{d}_j}(x) = \exp\left(i\kappa \boldsymbol{d}_j \cdot x\right),$$

where $\boldsymbol{d}_j$ $(j = 1, \ldots, p)$ belongs to a finite subset

$$\mathbb{S}_2^p = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_p\}$$

consisting of p different directions of the unit circle $\mathbb{S}_2$ of $\mathbb{R}^2$ and x is a generic point in the plane. Parameter p thus refers to the number of plane waves used in each $T \in \mathcal{T}$.

Each $x_T \in L^2(\partial T)$ is hence approximated by

$$x_T^p = \sum_{j=1}^{p} \alpha_{T,\boldsymbol{d}_j} \frac{1}{2i\kappa}\left(\partial_{\boldsymbol{n}_T} v_{\boldsymbol{d}_j} + i\kappa v_{\boldsymbol{d}_j}\right),$$

where $\alpha_{T,\boldsymbol{d}_j} \in \mathbb{C}$ can be seen as the components of either x_T^p or the associated Trefftz function

$$u_T = \sum_{j=1}^{p} \alpha_{T,\boldsymbol{d}_j} v_{\boldsymbol{d}_j}.$$

The above expression can be written in matrix form as

$$x_T^p = \left[\begin{array}{ccc} \Lambda^+_{T,\boldsymbol{d}_1} v_{\boldsymbol{d}_1} & \cdots & \Lambda^+_{T,\boldsymbol{d}_p} v_{\boldsymbol{d}_p} \end{array}\right] \begin{bmatrix} \alpha_{T,\boldsymbol{d}_1} \\ \vdots \\ \alpha_{T,\boldsymbol{d}_p} \end{bmatrix} \in X_T^p, \ \Lambda^+_{T,\boldsymbol{d}_j} = \frac{1}{2}\left(1 + \boldsymbol{d}_j \cdot \boldsymbol{n}_T\right), \tag{5}$$

The choice of the plane-wave basis yields that the unitary operator $\mathcal{U}_T$ can be written explicitly as follows

$$\mathcal{U}_T x_T^p = \sum_{j=1}^{p} \alpha_{T,\boldsymbol{d}_j}\left(-\frac{1}{2i\kappa}\right)\left(\partial_{\boldsymbol{n}_T} v_{\boldsymbol{d}_j} - i\kappa v_{\boldsymbol{d}_j}\right),$$

either also

$$\mathcal{U}_T x_T^p = \left[\begin{array}{ccc} \Lambda^-_{T,\boldsymbol{d}_1} v_{\boldsymbol{d}_1} & \cdots & \Lambda^-_{T,\boldsymbol{d}_p} v_{\boldsymbol{d}_p} \end{array}\right] \begin{bmatrix} \alpha_{T,\boldsymbol{d}_1} \\ \vdots \\ \alpha_{T,\boldsymbol{d}_p} \end{bmatrix}, \quad \Lambda^-_{T,\boldsymbol{d}_j} = \frac{1}{2}\left(1 - \boldsymbol{d}_j \cdot \boldsymbol{n}_T\right).$$

Applying this Galerkin method in formulation (3), we get the following linear system

$$(A - F)\alpha = b, \tag{6}$$

where A, F and b are respectively defined through the following relationships

$$\int_{\partial T} y_T^p \mathcal{U}_T x_T^p ds = \beta_T^\top A_T \alpha_T, \quad \sum_{T \in \mathcal{T}} \beta_T^\top A_T \alpha_T = \beta^\top A \alpha,$$
$$\sum_{T \in \mathcal{T}} \int_{\partial T} y_T^p \mathcal{F}_T x^p ds = \beta^\top F \alpha, \quad \sum_{T \in \mathcal{T}} \int_{\partial T} y_T^p w_T ds = \beta^\top b,$$

where $x^p = (x_T^p)_{T \in \mathcal{T}}$. Column-vector α_T (resp. β_T) gathers the coefficients $\alpha_{T,\boldsymbol{d}_i}$ (resp. $\beta_{T,\boldsymbol{d}_i}$), $i = 1, \ldots, p$, and displays the coefficients of x_T^p (resp. y_T^p) according to (5). Column-vectors α and β play the same role with regard to α_T and β_T, $T \in \mathcal{T}$, and thus refer to the coefficients corresponding to x^p and y^p respectively. Finally, the transpose of a matrix is indicated by the superscript $^\top$.

We also mention that the integrals on edges of the products of two plane waves are computed exactly.

Remark 2. Clearly, matrix A is a block-diagonal matrix, the diagonal blocks of which being the matrices A_T. As its definition makes it clear, A_T is the matrix of the bilinear form associated with the operator $\mathcal{U}_T$. The fact that $\mathcal{U}_T$ is unitary by no means implies that A_T is a unitary matrix. This justifies the use of A_T instead of the more natural notation U_T. Actually, we see below that, up to a permutation of its rows, A_T is a kind of mass matrix related to outgoing traces. Operator $x^p \rightarrow \mathcal{F}_T x^p$ ensures the coupling of the degrees of freedom related to different elements of $\mathcal{T}$. It may depend on x_T^p and also on all the x_L^p with L adjacent to T. Thus, F cannot be a block-diagonal matrix like A.

Similarly, the same Galerkin method applied to formulation (4) leads to the following linear system

$$(M - E)\,\alpha = c, \tag{7}$$

where matrices M and E and column-vector c are defined through the following relationships

$$\int_{\partial T} \overline{y_T^p} x_T^p ds = \beta_T^* M_T \alpha_T, \quad \sum_{T\in\mathcal{T}} \beta_T^* M_T \alpha_T = \beta^* M \alpha$$

$$\sum_{T\in\mathcal{T}} \int_{\partial T} \overline{\mathcal{U}_T y_T^p} \mathcal{F}_T x^p ds = \beta^* E\alpha, \quad \sum_{T\in\mathcal{T}} \int_{\partial T} \overline{\mathcal{U}_T y_T^p} w_T ds = \beta^* c,$$

where β^* is the adjoint vector to β (transpose of the complex conjugate). In the same way, matrix M is a block-diagonal matrix, whose diagonal blocks are the matrices M_T related to the elements $T \in \mathcal{T}$.

The following proposition shows that, under the condition

$$\boldsymbol{d} \in \mathbb{S}_2^p \text{ if and only if } -\boldsymbol{d} \in \mathbb{S}_2^p, \tag{8}$$

system (6) and system (7) are identical except for a specific permutation of the equations. This is stated in a precise manner as follows.

Proposition 1. Under hypothesis (8), matrices A, F and column-vector b can be expressed as

$$A = PM, \quad F = PE, \quad b = Pc, \tag{9}$$

where P is a block-diagonal matrix whose diagonal blocks P_T, $T \in \mathcal{T}$, are permutation matrices, swapping the equations associated with $\boldsymbol{d}$ and $-\boldsymbol{d}$. Thus, the permutation matrices P_T are involutory, that is,

$$P_T^2 = \mathbb{I}_p, \tag{10}$$

where $\mathbb{I}_p$ is the identity matrix of order p.

Proof. Let $v_{\boldsymbol{d}}(x) = e^{i\kappa \boldsymbol{d}\cdot x}$ be a plane-wave basis function, hence with $\boldsymbol{d} \in \mathbb{S}_2^p$. The equation relative to $v_{\boldsymbol{d}}$ for the element T in system (7) is given by

$$\int_{\partial T} x_T^p \Lambda_{T,\boldsymbol{d}}^+ \overline{v_{\boldsymbol{d}}} ds - \int_{\partial T} \mathcal{F}_T x^p \Lambda_{T,\boldsymbol{d}}^- \overline{v_{\boldsymbol{d}}} ds = \int_{\partial T} w_T \Lambda_{T,\boldsymbol{d}}^- \overline{v_{\boldsymbol{d}}} ds.$$

Since $\overline{v_{\boldsymbol{d}}} = v_{-\boldsymbol{d}}$ and $\Lambda_{T,-\boldsymbol{d}}^+ = \Lambda_{T,\boldsymbol{d}}^-$, this equation can be rewritten

$$\int_{\partial T} x_T^p \Lambda_{T,-\boldsymbol{d}}^- v_{-\boldsymbol{d}} ds - \int_{\partial T} \mathcal{F}_T x^p \Lambda_{T,-\boldsymbol{d}}^+ v_{-\boldsymbol{d}} ds = \int_{\partial T} w_T \Lambda_{T,-\boldsymbol{d}}^+ v_{-\boldsymbol{d}} ds.$$

Now, noting that $\Lambda^{-}_{T,-\boldsymbol{d}} v_{-\boldsymbol{d}} = \mathcal{U}_T \Lambda^{+}_{T,-\boldsymbol{d}} v_{-\boldsymbol{d}}$ and using the symmetry of operator $\mathcal{U}_T$, we come to

$$\int_{\partial T} \mathcal{U}_T x^p_T \Lambda^{+}_{T,-\boldsymbol{d}} v_{-\boldsymbol{d}} ds - \int_{\partial T} \mathcal{F}_T x^p \Lambda^{+}_{T,-\boldsymbol{d}} v_{-\boldsymbol{d}} ds = \int_{\partial T} w_T \Lambda^{+}_{T,-\boldsymbol{d}} v_{-\boldsymbol{d}} ds.$$

This is nothing but the equation relative to $v_{-\boldsymbol{d}}$ in system (6). Hypothesis (8) completes the proof of (9). Property (10) follows from the fact that left-multiplication of a matrix by P_T swaps the rows corresponding to directions $\boldsymbol{d}$ and $-\boldsymbol{d}$. □

Remark 3. In all the numerical experiments conducted below, we have considered the case of a system of directions for the plane-wave basis uniformly distributed over $\mathbb{S}_2$

$$\boldsymbol{d}_j = (\cos\theta_j, \sin\theta_j), \theta_j = \theta_1 + (j-1)(2\pi/p), j = 1, \ldots, p,$$

with an even number p of directions. The first angle θ_1 is called the shift in the sequel. In this case,

$$P_T = \begin{bmatrix} 0 & \mathbb{I}_{p/2} \\ \mathbb{I}_{p/2} & 0 \end{bmatrix}.$$

We now go on to the following important property.

Proposition 2. Matrix M is an Hermitian positive definite matrix.

Proof. Definition of matrix M immediately yields that it is a Hermitian and positive matrix. Its definiteness follows from the uniqueness of the solution to problem (2). □

We now come to the most important property of the considered discretization. This property makes this approach among the very few methods, as for instance the DG method in [17], that are immune from internal resonances. These resonances may actually arise when inverting the related linear system by a direct solver based on block Gaussian eliminations related to the degrees of freedom corresponding to each element, when discretizing the Helmholtz equation by a usual continuous Finite Element Method. To prove this fact, we use a technique of proof introduced in [17] for a completely different issue.

Proposition 3. Let ω be a subdomain of Ω, defined by its non-overlapping domain decomposition associated with the elements T of a subset $\mathcal{T}_\omega$ of $\mathcal{T}$. Designating by α_ω any vector α of coefficients of plane waves (see (6)) such that $\alpha_T = 0$ for $T \notin \mathcal{T}_\omega$, we can state the following property: if α_ω satisfies

$$\beta_\omega^\top (A - F)\, \alpha_\omega = 0,\ \forall \beta_\omega,$$

then, $\alpha_\omega = 0$.

Proof. In view of (9), we immediately get that α_ω satisfies

$$\beta_\omega^* (M - E)\, \alpha_\omega = 0,\ \forall \beta_\omega. \tag{11}$$

In other words, the function $x_\omega^p \in X^p$ defined on ∂T for $T \in \mathcal{T}_\omega$, corresponding to α_ω, is the plane wave UWVF solution to the following boundary-value problem

$$\begin{cases} \Delta u_\omega + \kappa^2 u_\omega = 0 \text{ on } \omega, \\ \partial_{\boldsymbol{n}} u_\omega - i\kappa Y_\omega u_\omega = 0 \text{ on } \partial\omega, \end{cases}$$

with $Y_\omega = Y$ on $\partial\omega_\Omega = \partial\omega \cap \partial\Omega$, Y being the admittance of $\partial\Omega$ involved in problem (1) and $Y_\omega = 1$ on $\partial\omega \setminus \partial\omega_\Omega$; $\boldsymbol{n}$ is the unit normal on $\partial\omega$ directed towards the exterior of ω. We easily complete the proof by using the fact that the boundary condition on $\partial\omega$ cannot be a pure Neumann condition everywhere and Th 2.1 in [8]. □

Remark 4. Given any subset I of indices of the components of the above vectors α and β, the uniqueness property remains valid for (11) when considering vectors α and β such that $\alpha_i = \beta_i = 0$ for $i \notin I$. Indeed, it is enough to define $\mathcal{T}_\omega$ as the set of elements T supporting at least one plane-wave basis function corresponding to the coefficients α_i and β_i, with $i \in I$, and next to limit the trial x^p and test y^p functions to those generated by these basis functions. The uniqueness is then obtained by reusing the arguments in the proof of the above proposition. This establishes that system (7) can be processed using Gaussian eliminations without pivoting. This is a clear advantage of formulation (4) devised by Cessenat and Després [8]. However, the above proposition establishes that not everything is lost in this regard.

3. The SVD strategy

We start by displaying how ill-conditioning of the plane wave UWVF manifests itself (subsection 3.1). To remedy this shortcoming, we use a strategy based on the SVD of specific elemental matrices. In particular, we discuss some numerical issues about this approach (subsection 3.2). We then describe the two strategies based on the local SVD (subsections 3.3 and 3.4 respectively). Comparison and assessment of the two approaches are performed on two test-cases: a waveguide propagation problem (subsection 3.5) and a problem involving a singular geometry generating evanescent waves (subsection 3.6).

3.1. Breakdown of the condition number in the plane-wave UWVF

To bring out the breakdown of the condition number when using the plane-wave UWVF, we consider the simple problem of an infinite waveguide with rigid walls in which the fundamental mode $\exp(i\kappa x)$ is propagating. We generally denote by x and y the Cartesian coordinates. To numerically simulate this device, we make a fictitious truncation at $x = 0$ for the waveguide inlet and at $x = L$ for its outlet. For this simple device, it is possible to derive exact truncating conditions. As a result, the problem to be solved can be stated as follows

$$\left\{\begin{array}{l}\Delta u + \kappa^2 u = 0 \text{ in } \Omega =]0, L[\times]0, H[\,,\\ \partial_{\boldsymbol{n}} u = 0 \text{ for } y = 0 \text{ and } y = H,\\ \partial_{\boldsymbol{n}} u - i\kappa u = -2i\kappa \text{ for } x = 0,\ \partial_{\boldsymbol{n}} u - i\kappa u = 0 \text{ for } x = L.\end{array}\right.$$

All the numerical tests were performed with the following data: $L = 3$, $H = 1$, in general , and an unstructured triangular finite element mesh whose meshsize is $h = 1/6$. The dimensions are given in wavelengths (that is, $\kappa = 2\pi$). In Table 1 are reported the results for the Raw plane-wave UWVF, that is, a direct solving without any attempt to improve the condition number of the matrix related to the linear system. The shift θ_1 is taken equal to 0.2. This prevents the exact solution $\exp(i\kappa x)$ from belonging to the discrete plane-wave space generated by the plane-wave basis functions $\left\{v_{\boldsymbol{d}_j}\right\}_{1\leq j\leq p}$.

Below, as a rule, we denote by N and $\varkappa$ respectively the length of a square matrix A and an estimate of its condition number. To avoid possible confusions, we can further also write N_A and $\varkappa_A$ in a more specific way.

As reported in Table 1, the error is diminishing until an optimal number of approximating plane waves. The errors therein and below are expressed in % and relatively to the maximum norm. Then, a direct correlation of the error deterioration with the condition number degradation can be observed. The results also indicate that the local condition numbers $\varkappa_{A_T}$ are correlated

p	N_{A-F}	$\min_{T\in\mathcal{T}} \varkappa_{A_T}$	$\max_{T\in\mathcal{T}} \varkappa_{A_T}$	$\varkappa_{A-F}$	error
2	408	1.2e+00	2.2e+00	1.8e+02	**
4	816	6.5e+00	8.7e+01	1.6e+02	**
8	1632	3.1e+03	3.1e+06	1.3e+07	1.5
12	2448	1.8e+07	6.4e+11	4.3e+12	0.02
16	3265	2.8e+11	1.2e+17	2.0e+18	0.006
20	4080	8.3e+14	7.8e+17	1.1e+19	0.15
24	4896	4.2e+15	5.0e+17	1.9e+19	0.32
28	5712	1.6e+14	1.2e+18	2.7e+19	1.2
32	6528	1.2e+15	1.0e+19	1.0e+20	0.18

Table 1: Waveguide test-case dealt with by the Raw plane-wave UWVF approach. The condition numbers of small size matrices A_T are obtained by the Octave function cond while $\varkappa_{A-F}$ is estimated by the Octave function condest; double star ** indicates that the results are meaningless (the error is greater than 30 %).

with that of the global linear system $\varkappa_{A-F}$. As said in the introduction, this feature is well-known and has been formerly considered [28, 17].

In what follows, we propose procedures based either on a SVD or a QR decomposition of the matrices A_T that dramatically improve the condition number of the global linear system matrix, and consequently yields the plane-wave UWVF to be working at its full level of efficiency. After an element-wise preconditioning procedure, the condition number of the matrix of the global linear system reaches very low values, and this whatever is the number of plane waves being used. It is also shown that it is possible, in the framework of some of these procedures, to considerably reduce the order of the linear system to be solved so that it becomes somehow independent of the number p of the plane waves used for the local approximation. This approach of improving the overall conditioning of plane-wave DG discretizations of the Helmholtz condition by an element-wise strategy at the element level seems to have been introduced in [17] following previous studies dedicated to other kinds of equations as in particular [29]. Although basically different, our way of adapting this strategy to the UWVF is more in the spirit of this last reference.

3.2. The element-wise SVD

The fundamental principle of this strategy is to use a block preconditioning, built by improving the condition number of the diagonal block A_T. It is based on a SVD decomposition of the blocks A_T,

$$A_T = U_T \Sigma_T V_T^*,$$

where $\Sigma_T = \operatorname{diag}(\sigma_{1,T}, \ldots, \sigma_{n,T})$ is a diagonal matrix with coefficients $\sigma_{i,T} > 0$ for $j = 1, \ldots, p$, and U_T and V_T are unitary matrices. In this way, system (6) can be rewritten as

$$(U \Sigma V^* - F)\, \alpha = b, \tag{12}$$

where Σ is a diagonal matrix, and U and V are unitary matrices. Except F, all these matrices are block diagonal, the diagonal blocks of which are the $p \times p$ matrices Σ_T, U_T, and V_T, $T \in \mathcal{T}$, built at the element level.

Remark 5. The SVD decomposition $A_T = U_T \Sigma_T V_T^*$ of the invertible matrix A_T is computed by solving the eigenvalue problem

$$\begin{bmatrix} 0 & A_T^* \\ A_T & 0 \end{bmatrix} \begin{bmatrix} V_T & V_T \\ U_T & -U_T \end{bmatrix} = \begin{bmatrix} V_T & V_T \\ U_T & -U_T \end{bmatrix} \begin{bmatrix} \Sigma_T & 0 \\ 0 & -\Sigma_T \end{bmatrix},$$

instead of using the more direct but numerically instable approach

$$A_T^* A_T V_T = V_T \Sigma_T^2, \quad U_T = A_T V_T \Sigma_T^{-1}$$

(see, for example, [30]). Thus, property (9) seems to provide a cheaper alternative for reducing the size of the eigenvalue problem related to the SVD decomposition

$$M_T = V_T \Sigma_T V_T^*, \quad A_T = U_T \Sigma_T V_T^*, \quad U_T = P_T V_T.$$

However, Table 2 shows that the Octave function `eig` can lead to an anomalous negative value for the last eigenvalue contrary to the Octave function `svd`.

Two techniques for improving the matrix condition number of system (12) were developed in this framework. We respectively called them the Truncated SVD and the Regularized SVD. All computations are performed

rank	1	$\cdots$	11	12	13	14	15	16
svd A_T	8.29e+00	$\cdots$	4.06e-10	7.87e-13	5.54e-13	4.20e-15	2.29e-15	3.83e-16
svd M_T	8.29e+00	$\cdots$	4.06e-10	7.87e-13	5.53e-13	2.08e-15	7.03e-16	2.66e-16
eig M_T	8.29e+00	$\cdots$	4.06e-10	7.87e-13	5.53e-13	2.26e-15	7.60e-16	-3.40e-16

Table 2: Singular values of matrices A_T and M_T computed by the Octave function svd and eigenvalues of M_T computed by the Octave function eig; T is the triangle with vertices $(0,0)$, $(0.1,0)$, $(0,0.1)$; wavenumber $\kappa = 2\pi$ and $p = 16$.

at the element level during the assembly process, except for the final left and right multiplication by block-diagonal matrices. Both techniques are based on the consideration of a threshold parameter τ and require a partitioning of the indices of the diagonal of Σ_T in two sets: $I_{\tau,T}$ the set of the indices $i = 1,\ldots,p$ such that $\sigma_{i,T} \geq \tau \max_{j=1,\ldots,p} \sigma_{j,T}$ and $I^c_{\tau,T} = \{1,\ldots,p\} \setminus I_{\tau,T}$. We also use the notation $I_\tau = \cup_{T\in\mathcal{T}} I_{\tau,T}$ and $I^c_\tau = \cup_{T\in\mathcal{T}} I^c_{\tau,T}$.

3.3. The Truncated SVD

This approach is based on the decomposition of the unknown α of system (6) as follows

$$\alpha = V_{:I_\tau}\alpha_{I_\tau} + V_{:I^c_\tau}\alpha_{I^c_\tau}, \quad \alpha_{I_\tau} = V^*_{:I_\tau}\alpha,\ \alpha_{I^c_\tau} = V^*_{:I^c_\tau}\alpha,$$

where subscript of $V_{:I_\tau}$ means that this matrix is the subarray of V obtained by letting the row indices run from 1 to the length of the columns of V and column indices run in I_τ. In this way, after left-multiplying system (12) by $U^*_{:I_\tau}$ and $U^*_{:I^c_\tau}$ respectively, we get

$$\begin{bmatrix} U^*_{:I_\tau} \\ U^*_{:I^c_\tau} \end{bmatrix} (A - F) \begin{bmatrix} V_{:I_\tau} & V_{:I^c_\tau} \end{bmatrix} \begin{bmatrix} \alpha_{I_\tau} \\ \alpha_{I^c_\tau} \end{bmatrix} = \begin{bmatrix} b_{I_\tau} \\ b_{I^c_\tau} \end{bmatrix}$$

with $b_{I_\tau} = U^*_{:I_\tau} b$, $b_{I^c_\tau} = U^*_{:I^c_\tau} b$. The components $\alpha_{I^c_\tau}$ are put to 0 and the equations relative to indices I^c_τ are then removed. We then come to the reduced system

$$U^*_{:I_\tau} (A - F)\, V_{:I_\tau}\alpha_{I_\tau} = b_{I_\tau}. \tag{13}$$

This system can be inverted as is or can be further preconditioned by a left and right multiplication by $\Sigma_{I_\tau}^{-1/2}$ arriving to

$$\Sigma_{I_\tau}^{-1/2} U_{:I_\tau}^* \left(A - F\right) V_{:I_\tau} \Sigma_{I_\tau}^{-1/2} \chi_\tau = \Sigma_{I_\tau}^{-1/2} b_{I_\tau}, \tag{14}$$

with $\chi_\tau = \Sigma_{I_\tau}^{1/2} \alpha_{I_\tau}$.

Remark 6. It is worth noting that system (14) is in the form of a fixed-point system

$$\left(I - F_{\text{tr}}\right) \chi_\tau = b_{\text{tr}},$$

where "tr" is used as a diminutive of "truncated". In [8], Cessenat and Després recommended to transform (7) by left-multiplying it by M^{-1}. However, as reported in several publications (see, for example, [8, 31, 15]) this simple strategy is not enough to improve the condition number of the related system matrix.

Remark 7. Reduction to system (13) can be seen also as a filtering process to build an effective basis, the vectors of which being numerically linearly independent. Such a procedure is used in [32] for solving the same problem by the Virtual Element Method. As this can be seen in Table 3 below, such procedure gets rid of the loss of accuracy but does not efficiently improve the condition number of the linear system matrix.

3.4. Regularized SVD

One may wonder whether the previous approach, which neglects the components of the solution α related to the smallest singular values, does not lead to some loss of accuracy due to a poorer local space of trial and test functions as what is done in [28]. In this reference, the number of plane waves in each element $T \in \mathcal{T}$ is selected dynamically. It is increased or decreased until the condition number of M_T reaches a fixed value considered as acceptable. The present approach however does not reduce the number of plane waves but instead in some meaning prevents the discrete inverse of the outoing trace operator from blowing up. We were thus led to consider a second procedure which aims at keeping unchanged the space of trial and test functions. Its basic principle is to use a kind of Tichonov's regularization A_τ approximating A that can be inverted in a stabilized way

$$A_\tau = U \Sigma_\tau V^*, \; \Sigma_{\tau,T} = \text{diag}\left(\sigma_{1,T}^\tau, \ldots, \sigma_{p,T}^\tau\right), \; \sigma_{i,T}^\tau = \max\left(\sigma_{i,T}, \tau \max_{1 \le j \le p} \sigma_{j,T}\right).$$

We then obtain the following system, the unknown of which is $\alpha_\tau = V^*\alpha$

$$U^*\left(A_\tau - F\right)V\alpha_\tau = U^*b. \tag{15}$$

Here too, the system can be solved as is or preconditioned again as follows

$$\Sigma_\tau^{-1/2}U^*\left(A_\tau - F\right)V\Sigma_\tau^{-1/2}\phi_\tau = \Sigma_\tau^{-1/2}U^*b \tag{16}$$

with $\phi_\tau = \Sigma_\tau^{1/2}V^*\alpha$. Here also, system (16) is in the form of a fixed point linear system

$$\left(I - F_{\mathrm{rg}}\right)\phi_\tau = b_{\mathrm{rg}}$$

where "rg" is used as a diminutive of "regularized".

3.5. A first numerical validation

As a first example, we consider again the waveguide problem and deal with it by means of the four procedures described above: the truncated and the regularized SVDs, without preconditioning (respectively systems (13) and (15)) and with preconditioning (respectively systems (14) and (16)). Table 3 reports the condition number of the related linear system matrices and the maximum error in % for the Truncated SVD UWVF and compares them to those related to the Raw UWVF taken from Table 1. The choice $\tau = 10^{-8}$ appeared to be the most suitable for preserving accuracy.

The same results, obtained this once with the Regularized SVD UWVF (Regularized UWVF) and the Preconditioned Regularized SVD UWVF (P. Regul. SVD) are presented in Table 4.

The improvement gained by the approaches based on the local SVD speaks for itself. Even without preconditioning, the condition number stabilizes at an acceptable value inducing similarly a stabilization of the error. This is in contrast with the Raw UWVF where the error, after passing by its lowest value, starts growing with p.

Preconditioning makes it possible to stabilize the condition number at very low values for Galerkin methods associated with plane-wave basis functions. As expected above, reducing the dimension of the local approximation spaces gives rise to a slightly larger error, against the advantage of significantly reducing the size of the global linear system. As reported in Table 3 and Table 4, as p increases the SVD results are stabilizing close to the best level of accuracy, being reached by the Raw UWVF. This optimal value of p is not known a priori and depends on many factors: element shape, frequency,

	Raw UWVF			Truncated SVD			P. Trunc. SVD	
p	N	$\varkappa_{A-F}$	error	N	$\varkappa$	error	$\varkappa$	error
2	408	1.8e+02	**	408	1.8e+02	**	8.1e+01	**
4	816	1.6e+02	**	816	1.6e+02	**	4.5e+02	**
8	1632	1.3e+07	1.5	1632	1.4e+06	1.26	3.7e+02	1.5
12	2448	4.3e+12	0.02	1900	4.0e+08	0.011	3.7e+02	0.019
16	3265	2.0e+18	0.006	1902	4.0e+08	0.011	3.7e+02	0.011
20	4080	1.1e+19	0.15	1902	4.0e+08	0.011	3.7e+02	0.011
24	4896	1.9e+19	0.32	1902	4.0e+08	0.011	3.7e+02	0.011
28	5712	2.7e+19	1.2	1902	4.0e+08	0.011	3.7e+02	0.011
32	6528	1.0e+20	0.18	1902	4.0e+08	0.011	3.7e+02	0.011

Table 3: Waveguide test-case dealt with by the Truncated SVD UWVF (Truncated SVD) and Preconditioned Truncated SVD UWVF (P. Trunc. SVD). The results are compared to those of the Raw UWVF; the size of the linear system related to the Preconditioned Truncated SVD UWVF is the same than that of the truncated SVD UWVF; N and $\varkappa$ are the size and the condition number of the related linear system matrices; the other variables have been already documented above.

		Raw UWVF		Regularized SVD		P. Regul. SVD	
p	N	$\varkappa$	error	$\varkappa$	error	$\varkappa$	error
2	408	1.8e+02	**	8.0e+01	**	8.1e+01	**
4	816	1.6e+02	**	1.5e+03	**	4.4e+02	**
8	1632	1.3e+07	1.5	1.3e+07	1.26	3.7e+02	1.47
12	2448	4.3e+12	0.02	4.7e+08	0.019	3.7e+02	0.018
16	3265	2.0e+18	0.006	4.7e+08	0.006	3.7e+02	0.006
20	4080	1.1e+19	0.15	4.7e+08	0.006	3.7e+02	0.006
24	4896	1.9e+19	0.32	4.7e+08	0.006	3.7e+02	0.006
28	5712	2.7e+19	1.2	4.7e+08	0.006	3.7e+02	0.006
32	6528	1.0e+20	0.18	4.7e+08	0.006	3.7e+02	0.006

Table 4: Waveguide case-test dealt with the Regularized SVD UWVF (Regularized SVD) and Preconditionned Regularized SVD UWVF (P. Regularized SVD) approaches. The three linear systems are of the same order N; $\varkappa$ is an estimate of the matrix condition number of the related method.

etc. In this respect, the SVD approach is a powerful tool to avoid the determination of this optimal p. This is mainly explained by the fact that, beyond this optimal value, increasing p only results in smaller and smaller singular values that cannot be processed numerically. The SVD approach overcomes the difficulties associated with too small singular values, but unfortunately also rules out the associated modes that could improve accuracy.

Remark 8. Here, we make an important remark concerning the present approach and the one that can be obtained by an adaptation to system (7) of the procedure developed in [17]. When the local singular values do not go beyond the threshold τ, that is, when $\Sigma_\tau = \Sigma$ or equivalently when $I_\tau = \varnothing$, both the preconditioned truncated and the preconditioned regularized SVD coincide in arbitrary-precision arithmetic with the approach in [17]. Actually, considering that M_T is the Hermitian part of the local matrix of the variational system (4) (even if this is not true for an element T having a side $F_T^{\partial\Omega}$ on $\partial\Omega$), we are led, in the framework of this reference, to use a modified Gram-Schmidt orthogonalization relatively to the scalar product associated with M_T for constructing a matrix Q_T such that $Q_T^* M_T Q_T = \mathbb{I}_p$. It is easily seen that such a Q_T can be given by $Q_T = V_T \Sigma_T^{-1/2}$. This enables one to understand why beyond a critical value for p, the latter orthogonalization process fails.

We insist on the fact that the method is robust as long as the truncation parameter τ is taken small enough ($10^{-32} \leq \tau \leq 10^{-16}$ for the example considered here). As displayed in Fig. 1, taking τ too small, up to $\tau = 10^{-32}$ for the present example, does not damage neither the condition number nor the error. At worst, we might just end up with a non-optimal reduction of the order of the final linear system to be solved. Of course, a way to definitively settle this issue would be to develop a theoretical tool for the choice of τ. This task is difficult and deserves a separate study. However, since τ is destined to select singular values too small in a relative way, comparatively to the highest one, it turns out that the lack of such tool is far from being a prohibitive disadvantage of the method. The choice $10^{-12} \leq \tau \leq 10^{-8}$ always performed very well for all the test cases we considered.

It is worth saying some words about the above phrasing "*in arbitrary-precision arithmetic*". It emphasizes the fact that the modified Gram-Schmidt orthogonalization is less numerically favorable, especially for nearly singular matrices [30]. We were led similarly to consider a procedure that does not require to solve an eigenvalue problem. In this respect, we introduce another

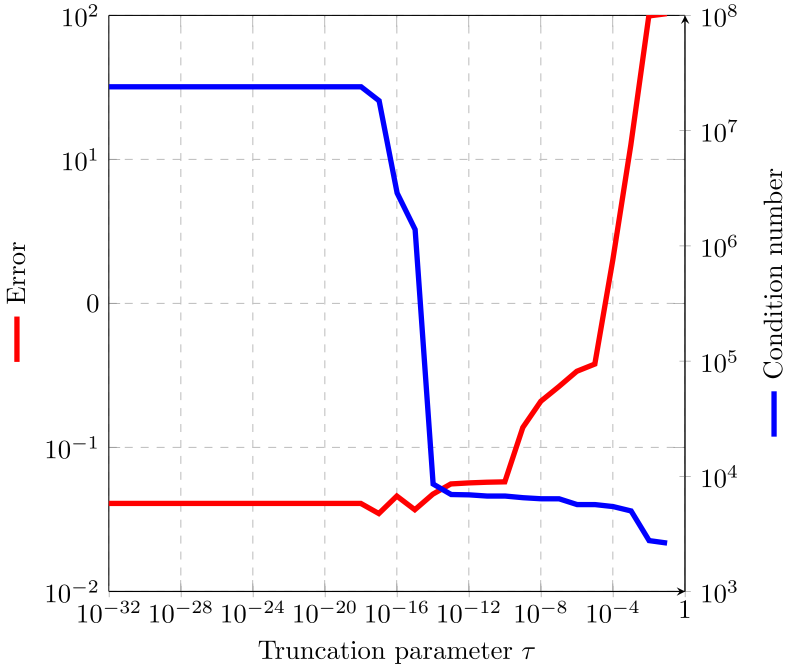
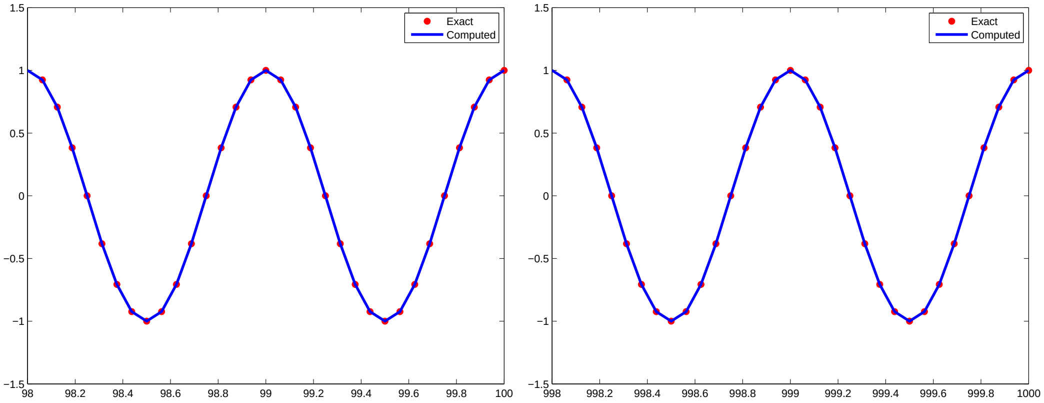


Figure 1: Parameter τ versus the error in % and an estimate of the condition number of the final linear system to be solved.



(a) First waveguide (100 wavelengths)

(b) Second waveguide (1000 wavelengths)

Figure 2: Plots of exact and computed solutions two wavelengths before the waveguide outlet.

local strategy in section 4 based on a Householder QR factorization with pivoting of A_T as an alternative to the SVD in a way completely different from the one carried out in [17].

In this first study, we did not intend to examine the dispersive and dissipative properties of the numerical methods considered in this work. We think that such a study does not fall within the scope of this paper, which is mainly devoted to the improvement of the conditioning of the plane-wave UWVF for standard problems extending over only a few wavelengths. Nonetheless, we report the results obtained for the above waveguide problem for two waveguides of length $L = 100$ wavelengths and $L = 1000$ wavelengths respectively. The tests were performed with $\tau = 10^{-12}$ and $p = 32$ and delivered results in accordance with those of the case $L = 3$ wavelengths with an error close to 0.006 %. We observed neither dissipation nor dispersion of the waves in these two cases as shown on the two plots of the exact and the computed solutions along two wavelengths before the outlet of the waveguide in figure 2.

Nevertheless, no conclusions can be drawn at this stage on this issue, which needs to be addressed, for example, as in references [33, 34].

3.6. A second example with evanescent waves

To assess the efficiency of the approach, we consider a case, in which a singularity in the geometry generates evanescent waves. These evanescent waves have to be accurately computed to get reliable numerical approximations of relevant parameters of the related physical device. The latter consists of the junction of two waveguides considered as semi-infinite (see Fig. 3).

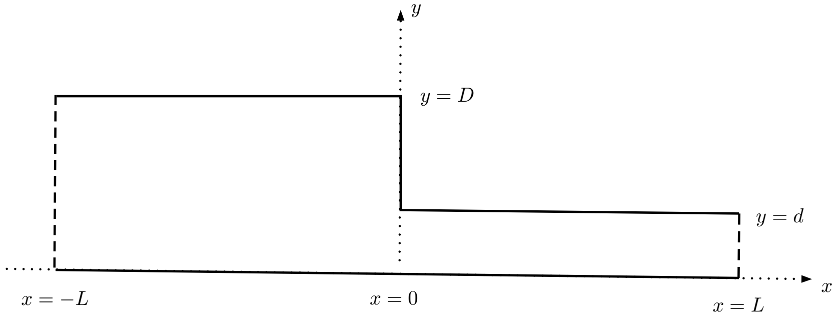


Figure 3: Schematic view of the junction of the waveguides

The problem to be solved implicitly defines the two parameters R and T, which have to be furnished by the numerical simulation

$$\begin{cases} \Delta u^- + \kappa^2 u^- = 0,\ x<0,\ 0<y<D \\ \Delta u^+ + \kappa^2 u^+ = 0,\ x>0,\ 0<y<d \\ \partial_y u^\pm = 0,\ y=0,\ \partial_y u^- = 0,\ x<0,\ y=D,\ \partial_y u^+ = 0,\ x>0,\ y=d, \\ \partial_x u^-\left(0,y\right) = 0,\ d<y<D, \\ u^-\left(0,y\right) = u^+\left(0,y\right),\ \partial_x u^-\left(0,y\right) = \partial_x u^+\left(0,y\right),\ 0<y<d, \\ u(x,y) = e^{i\kappa x} + Re^{-i\kappa x} + \varepsilon^-\left(x,y\right) \text{ for } x\to-\infty, \\ u(x,y) = Te^{i\kappa x} + \varepsilon^+\left(x,y\right) \text{ for } x\to-\infty, \end{cases}$$

with $\lim_{x\to\pm\infty}\varepsilon^\pm\left(x,y\right)=0$. The statement of the above problem is based on the fact that only the fundamental modes propagate in each of the two semi-infinite waveguides, that is, $\kappa D<\pi$. The solution of the above problem can be obtained by a mode decomposition of $u^\pm$ (see, for example, [35] for the general methodology of this solution procedure)

$$u^-\left(x,y\right) = e^{i\kappa x} + Re^{-i\kappa x} + \sum_{m\geq 1} r_m e^{\gamma_m x}\cos\left(\frac{m}{D}\pi y\right),\ \gamma_m = \sqrt{\left(\frac{m\pi}{D}\right)^2 - \kappa^2},$$

$$u^+(x,y) = Te^{i\kappa_c x} + \sum_{n\geq 1} t_n e^{-\sigma_n x}\cos\left(\frac{n}{d}\pi y\right),\ \sigma_n = \sqrt{\left(\frac{n\pi}{d}\right)^2 - \kappa^2}.$$

Lengthy but elementary calculations yield

$$R = \frac{1-\Xi_c}{1+\Xi_c}, \quad T = \frac{1}{\Xi}\frac{2\Xi_c}{1+\Xi_c}, \quad \Xi = d/D, \quad \Xi_c = \Xi/\left(1 - is^\top r\right),$$

where s is the infinite column-vector whose components are $s_m = \operatorname{sinc}\left(\frac{m\pi d}{D}\right)$, $m \geq 1$, $\operatorname{sinc} x = \sin x/x$ for $x \neq 0$, is the sinus cardinal function and r is an infinite column-vector, being a complicate function of the evanescent modes coefficients r_n and t_n. It can be expressed as

$$r = 2\kappa\Xi\left(1 + \Xi\Gamma^{-1}\Delta\Sigma\Delta^\top\right)^{-1}\Gamma^{-1}s$$

Δ being the infinite matrix whose coefficients are $\delta_{mn} = \operatorname{sinc}\left(\left(\frac{m}{D} + \frac{n}{d}\right)\pi d\right) + \operatorname{sinc}\left(\left(\frac{m}{D} - \frac{n}{d}\right)\pi d\right)$, Γ and Σ the infinite diagonal matrices whose diagonal coefficients are the above coefficients γ_m and σ_m respectively.

These formulae clearly show that coefficients R and T strongly depend on the evanescent modes through the coefficient

$$S = s^\top r.$$

By solving the discrete problem, we get approximate values of $R^p \approx R$, $T^p \approx T$ and can thus assess the quality of the approximation of the evanescent modes by means of the approximation of S by

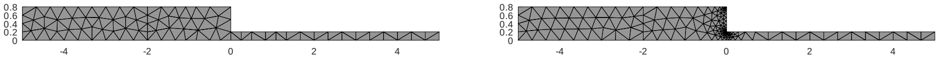
$$S^p = i\left(\Xi\frac{1+R^p}{1-R^p} - 1\right).$$

Here again, we need to introduce truncating conditions at $x = \pm L$ to apply the UWVF for effectively solving the problem. These conditions are no longer exact but only approximate and are set out as follows

$$\partial_{\boldsymbol{n}} u^- - i\kappa u^- = -2e^{i\kappa x} \text{ at } x = -L, \; \partial_{\boldsymbol{n}} u^+ - i\kappa u^+ = 0 \text{ at } x = L.$$

The computational domain is depicted in Fig. 3. They are almost exact if L is taken equal to few wavelengths.

The geometrical singularity, which here is located at the waveguides junction, is usually handled by carrying out a refinement procedure. We use two kinds of usual refinements: a refinement based on a denser mesh in the vicinity of the junction called an h-refinement and the use of a large p, that is, a large number of directions of plane waves in a unrefined mesh, called a p-refinement. The two meshes that have been used for the numerical tests are



(a) Non refined mesh

(b) Refined mesh

Figure 4: The two meshes used for the junction of waveguides case-test.

shown in Fig. 4. The unrefined mesh is composed of 132 triangles whereas the refined one is made up of 301 triangles.

Tables 5 and 6 report the obtained results. After some tests with $\tau = 10^{-8}$ as threshold, we came to the conclusion that this choice is not good enough to take the evanescent waves into account. Thus, choosing $\tau = 10^{-12}$ turned out to be as efficient as $\tau = 10^{-8}$ when there is no evanescent wave. The errors are reported in % and concern the errors

$$e_T = 100\frac{|T^p - T|}{|T|}, \quad e_R = 100\frac{|R^p - R|}{|R|}, \quad e_S = 100\frac{|S^p - S|}{|S|}.$$

The quantities m_σ and M_σ display the lowest and the highest number of singular values retained in the local SVD for the truncated SVD UWVF as well as for the construction of the Regularized SVD UWVF.

The key conclusion to be drawn from these tests is that the SVD strategy is very robust either when dealing with a p-refinement or a h-refinement. Obviously the h-refinement procedure yields the most accurate results. However, without a clue on the localization of the singularities, or when it would be too tedious or expensive to take care of them, the p-refinement procedure, when coupled with the truncated SVD UWVF can reveal itself to be a strongly efficient method. The initial objective of the SVD strategy was to get rid of the instabilities due to the blow up of the condition number of the final linear system. However, the truncated form of the approach revealed itself as a wonderful tool for reducing the size of this system in a dramatic way and this without any loss in the accuracy. One last point to note: the results delivered by the Raw plane-wave UWVF are very questionable for a large condition number. As this is reported in Table 6, they can be almost exact for the case $p = 8$ and $p = 16$ but completely wrong for $p = 20$. It is also the case for $p = 20$ for the h-refinement procedure. This seems to indicate that, despite its numerous advantages, it is highly risky to use the Raw plane-wave UWVF for effective scientific or industrial computations. We believe that the treatment given here for the ill-conditioning of the UWVF enables it to be

p		Raw	T. SVD	R. SVD	p	Raw	T.SVD	R.SVD
8	N	1056	1056	1056	32	4224	1668	4224
	$\varkappa$	9.0e+07	8.5e+02	9.1e+02		1.0e+28	1.1e+03	1.1e+03
	m_σ	–	8	–		–	11	11
	M_σ	–	8	–		–	16	16
	e_T	1.7	1.7	0.6		**	0.6	0.5
	e_R	0.9	0.9	0.3		0.3	0.3	0.3
	e_S	4	4	1.3		1.4	1.3	1.2
12	N	1584	1570	1584	64	8448	1696	8448
	$\varkappa$	5.0e+13	1.1e+03	1.1e+03		2.0e+23	1.1e+03	1.1e+03
	m_σ	–	11	–		–	11	–
	M_σ	–	12	–		–	18	–
	e_T	0.6	0.6	0.6		0.5	0.6	0.5
	e_R	0.3	0.3	0.3		0.2	0.3	0.3
	e_S	1.4	1.5	1.5		0.9	1.3	1.2
20	N	2640	1652	2640	128	16896	1760	16896
	$\varkappa$	2.0e+23	1.1e+03	1.1e+03		3.0e+24	1.1e+03	1.1e+03
	m_σ	–	11	–		–	11	11
	M_σ	–	14	–		–	22	22
	e_T	0.8	0.6	0.5		**	0.6	0.5
	e_R	0.3	0.3	0.3		0.1	0.3	0.3
	e_S	1.4	1.3	1.2		0.5	1.3	1.2

Table 5: Waveguides junction test-case dealt with by the p-refinement procedure on the non-refined mesh and the SVD approach. Raw SVD (Raw) is compared with the Preconditionned Truncated SVD (T. SVD) and the Preconditionned Regularized SVD (R. SVD); N and $\varkappa$ here refer to the size and an estimate of the condition number of the global linear system obtained by means of the Octave function condest respectively. It should be noted that the values of the transmission coefficient T provided by the Raw UWVF are completely wrong for $p = 20$ and $p = 128$.

p		Raw	T. SVD	R. SVD	p	Raw	T.SVD	R.SVD
8	N	2408	2361	2408	12	3612	2990	3612
	$\varkappa$	2.0e+17	1.5e+03	1.5e+03		9.0e+20	1.7e+03	1.5e+03
	m_σ	–	7	–		–	7	–
	M_σ	–	8	–		–	12	–
	e_T	0.03	0.06	0.06		0.03	0.06	0.06
	e_R	0.01	0.03	0.03		0.02	0.03	0.03
	e_S	0.07	0.13	0.13		0.08	0.13	0.13
16	N	4816	3059	4816	20	6020	3060	6020
	$\varkappa$	3.8e+26	1.8e+03	1.7e+03		1.0e+27	1.8e+03	1.8e+03
	m_σ	–	7	–		–	7	–
	M_σ	–	14	–		–	14	–
	e_T	0.05	0.06	0.06		**	0.06	0.06
	e_R	0.39	0.03	0.03		**	0.03	0.03
	e_S	1.7	0.13	0.13		**	0.13	0.13

Table 6: Waveguides junction dealt with by the h-refinement procedure and the SVD approach. Here too N and $\varkappa$ refer to the size and an estimate of the condition number of the global linear system respectively. Now, the values of T, R, and S are all wrong for the Raw UWVF and $p = 20$.

fully effective.

4. The QR local strategy

Here, we modify the above strategy by replacing the SVD by a QR decomposition. The study is then conducted along the same lines as in the previous section.

4.1. The element-wise QR

One issue that can be raised for the above approaches is that the SVD, even used at the element level only, may become expensive, especially when using a relatively large number p of plane waves, mainly due to the need of solving an eigenvalue problem. This is even more true in the 3D case requiring $p^2/4$ plane waves to keep the same level of accuracy than the one corresponding to p plane waves in the 2D case. Instead of a SVD, we were led to try a local decomposition in the form

$$A_T = Q_T R_T P_T^\top, \tag{17}$$

that is, a QR decomposition with pivoting (see, for example, [30]). This decomposition is also viewed as a rank revealing algorithm like SVD, but at a lower computational cost. In (17), Q_T is a unitary matrix composed of elementary Householder matrices, R_T is a upper triangular matrix, and P_T is a permutation matrix. Decomposition (17) is known to be more numerically stable than the decomposition $A_T = Q_T R_T$ obtained by the modified Gram-Schmidt algorithm [30], especially for matrices A_T like here which may become almost singular. It must be observed however that stable Householder QR does not ensure that the diagonal coefficients of R_T are positive.

We then decompose R_T as the superposition of a diagonal matrix D_T and a strict upper triangular matrix $-F_T$

$$R_T = D_T - F_T.$$

We first write R_T in the form $R_T = D_T \left(\mathbb{I}_p - D_T^{-1} F_T\right)$ and then A_T at least formally for the moment as

$$A_T = Q_T D_T V_T^{-1}, \quad V_T = P_T \left(\mathbb{I}_p - D_T^{-1} F_T\right)^{-1}. \tag{18}$$

4.2. The Truncated QR

From decomposition (18), it is possible to mimic the truncated SVD and the regularized SVD, except now that the diagonal matrix D_T can have positive and negative real coefficients and V_T is not a unitary matrix.

Let $\tau > 0$ be a given threshold. As for the SVD, we define the sets of indices I_τ and I_τ^c by

$$I_{\tau,T} = \left\{ 1 \leq i \leq p;\ |d_{i,T}| \geq \tau \max_{j=1,\dots,p} |d_{j,T}| \right\}, \quad I_{\tau,T}^c = \{1, \dots, p\} \setminus I_{\tau,T}$$

where $d_{j,T}$ are the diagonal coefficients of D_T and $I_\tau = \cup_{T \in \mathcal{T}} I_{\tau,T}$, $I_\tau^c = \cup_{T \in \mathcal{T}} I_{\tau,T}^c$. Proceeding as for the SVD, we obtain the preconditioned truncated QR UWVF

$$D_{I_\tau}^{-1/2} Q_{:I_\tau}^* \left(A - F\right) V_{:I_\tau} D_{I_\tau}^{-1/2} \chi_\tau = \Sigma_{I_\tau}^{-1/2} b_{I_\tau}$$

where $D_{I_\tau}^{-1/2}$ is obtained by inverting the square root of the diagonal coefficients of D_{I_τ}. The square root is that related to the principal branch of this function.

Remark 9. At first glance, $\left(\mathbb{I}_p - D_{\tau,T}^{-1} F_T\right)^{-1} \approx \left(\mathbb{I}_p - D_T^{-1} F_T\right)^{-1}$, with $D_{\tau,T}$ obtained by modifying the diagonal coefficients $d_{j,T}$ for $i \in I_{\tau,T}^c$ so that

$$\left(D_{\tau,T}\right)_i = d_{i,T},\ i \in I_\tau, \quad \left(D_{\tau,T}\right)_i = \tau \max_{j=1,\dots,p} |d_{j,T}|,\ i \in I_\tau^c$$

appears to be more stable than just keeping $\left(\mathbb{I}_p - D_T^{-1} F_T\right)^{-1}$. However, several tests have not confirmed this claim. In this respect, we have never observed any division by zero.

4.3. The Regularized QR

Defining Q, V and $D_\tau^{-1/2}$ similarly to the SVD matrices U, V and Σ_τ, we obtain the Preconditioned Regularized QR UWVF just as was formerly done for the case of the Regularized SVD

$$D_\tau^{-1/2} Q^* \left(A_\tau - F\right) V D_\tau^{-1/2} \phi_\tau = D_\tau^{-1/2} V^* b$$

now with $A_\tau = Q D_\tau V^{-1}$.

p		T. SVD	T. QR	F. QR	p	T.SVD	T. QR	R. QR
8	N	1056	1056	1056	32	1668	1749	4224
	$\varkappa$	8.5e+02	1.5e+03	1.5e+03		1.1e+03	1.7e+03	3.0e+07
	m_σ	8	8	–		11	11	–
	M_σ	8	8	–		16	17	–
	e_T	1.7	1.7	1.7		0.6	0.5	0.4
	e_R	0.9	0.9	0.9		0.3	0.3	0.2
	e_S	4	4	4		1.3	1.1	0.8
12	N	1570	1584	1584	64	1696	1772	8448
	$\varkappa$	1.1e+03	1.6e+03	1.6e+03		1.1e+03	1.9e+03	1.0e+09
	m_σ	11	12	–		11	11	–
	M_σ	12	12	–		22	19	–
	e_T	0.6	0.6	0.6		0.6	0.5	0.4
	e_R	0.3	0.3	0.3		0.3	0.3	0.2
	e_S	1.5	1.4	1.4		1.3	1.2	0.9
20	N	1652	1728	2640	128	1760	1836	16896
	$\varkappa$	1.1e+03	1.7e+03	5.0e+06		1.1e+03	1.8e+03	4.5e+08
	m_σ	11	11	–		11	11	–
	M_σ	14	15	–		22	23	–
	e_T	0.6	0.5	0.4		0.5	0.5	0.3
	e_R	0.3	0.3	0.2		0.3	0.3	0.2
	e_S	1.3	1.2	0.9		1.3	1.2	0.7

Table 7: Waveguides junction test-case dealt with by the p-refinement and the QR approach.

4.4. Numerical results

To assess the performances of the QR approach, we consider again the waveguide junction test-case and compare the results previously obtained in Table 5 and Table 6 to that provided by the QR approach under the same conditions, which are reported in Tables 7 and 8. Let us just specify that T. QR and R. QR respectively mean Preconditioned Truncated QR UWVF and Preconditioned Regularized QR UWVF. Instead of the results for the Raw UWVF in Table 5 and Table 6, we now compare the results of the QR approach with those of the Preconditioned Truncated SVD UWVF.

Apart from a non-decisive advantage as regards of accuracy for the Regularized QR approaches, both procedures based on the SVD or the QR de-

p		T. SVD	T. QR	F. QR	p	T.SVD	T. QR	R. QR
8	N	2361	2387	2408	12	2990	3078	3612
	$\varkappa$	1.5e+03	2.5e+03	8.3e+03		1.7e+03	2.6e+03	8.9e+05
	m_σ	7	7	–		7	7	–
	M_σ	8	8	–		12	12	–
	e_T	0.06	0.06	0.03		0.06	0.06	0.04
	e_R	0.03	0.03	0.01		0.03	0.03	0.02
	e_S	0.13	0.13	0.06		0.13	0.13	0.08
16	N	3059	3184	4816	20	3060	3198	6020
	$\varkappa$	1.8e+03	3.1e+03	2.1e+07		1.1e+03	2.8e+03	4.3e07
	m_σ	7	7	–		7	7	–
	M_σ	14	15	–		14	15	–
	e_T	0.06	0.06	0.04		0.05	0.06	0.04
	e_R	0.03	0.03	0.02		0.02	0.03	0.02
	e_S	0.13	0.12	0.09		0.12	0.13	0.09

Table 8: Waveguides junction test-case dealt with by the h-refinement and the QR approach.

compositions behave in much the same way. Nonetheless, the SVD strategy lies on a more reliable theoretical foundation.

5. Concluding remarks

We think that this study has demonstrated that the coupling of the plane-wave UWVF with one of the strategies, we have developed, based on a local SVD or a local QR is an efficient tool specially adapted for solving the boundary-value problems set in terms of the Helmholtz equation. We have seen that the truncated versions of the algorithms are particularly efficient for this purpose. Fixing a high number of plane waves per element is not at all penalizing. The method is somehow self-adapting: it keeps only the degrees of freedom that are really needed for the final solution.

Some issues remain to be addressed. We list them below.

- The most important in our opinion is to establish a more secure basis for the choice of the threshold τ. Ideally it could be a theoretical estimate based on the data of the problem under consideration. An operational answer to this issue turns out to be difficult. However, the lack of such a tool is not at all penalizing. The choice for the threshold τ is important but not critical. For a representative case of the class of problems being considered, it can be found by first taking it reasonably large and then decreasing it until the results stabilize as depicted in Figure 1.

- More numerical studies are needed to confirm that the p-refinement approach is enough to get an acceptable accuracy in the presence of the standard singularities of the Helmholtz equation. This will avoid the need for over-meshing in certain areas of the solution domain. About the overcost due to the computation of the matrices A and F for large p, we found that the assembly process takes almost all the processing time of each run, from 98 % for $p = 8$ to 78 % for $p = 128$. However, these benchmarks are not at all meaningful. Our code is only a demonstration code, running on Octave or MATLAB. The computation of the elemental matrices is performed through two nested loops, apparently hard to vectorize while the solution of the final linear system is done by means of a compiled function. One can get an idea of the disparity in processing time between the interpreted codes and compiled codes by observing that the processing time for obtaining the SVD of the elemental matrices A_T is a tiny fraction of the time of the overall assembly process, although it requires the most computational effort. In 3D, the assembly process might become a real blocking point. However, it can be addressed by an easy parallelization of the code.

- It remains to test also the approaches in 3D and for long-range propagation domains.

- We have done some attempts for solving the resulting linear systems by GMRES. The super-reduction of the condition number had not its counterpart in terms of the number of iterations for convergence. Actually, we believe that the current approach is suitable for a direct solution by Gaussian eliminations. It has to be coupled with an adapted Domain Decomposition procedure [36] when there is a need to solve this kind of problems on large-sized domains.

All these issues are under consideration and will be addressed in future works.

References

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

[2] I. M. Babuska, S. A. Sauter, Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers?, SIAM Journal on Numerical Analysis 34 (6) (1997) 2392–2423 (1997).

[3] I. Babuška, F. Ihlenburg, E. T. Paik, S. A. Sauter, A generalized Finite Element Method for solving the Helmholtz equation in two dimensions with minimal pollution, Computer Methods in Applied Mechanics and Engineering 128 (3-4) (1995) 325–359 (1995).

[4] A. Lieu, G. Gabard, H. Bériot, A comparison of high-order polynomial and wave-based methods for Helmholtz problems, Journal of Computational Physics 321 (2016) 105–125 (2016).

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

[6] K. Christodoulou, O. Laghrouche, M. S. Mohamed, J. Trevelyan, High-order finite elements for the solution of Helmholtz problems, Computers & Structures 191 (2017) 129–139 (2017).

[7] B. Després, Sur une formulation variationnelle ultra-faible, Comptes Rendus de l'Académie des Sciences Série I 318 (1994) 939–944 (1994).

[8] O. Cessenat, B. Després, Application of an ultra weak variational formulation of elliptic PDES to the two-dimensional Helmholtz problem, SIAM J. Num. Analysis 35 (1) (1998) 255–299 (1998).

[9] O. Cessenat, B. Després, Using plane waves as base functions for solving the time-harmonic equations with the ultra weak variational formulation, J. Comp. Acous. 11 (2) (2003) 227–238 (2003).

[10] A. Buffa, P. Monk, Error estimates for the ultra weak variational formulation of the Helmholtz equation, Mathematical Modelling and Numerical Analysis 42 (6) (2008) 925–940 (2008).

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

[12] R. Hiptmair, A. Moiola, I. Perugia, Plane wave discontinuous galerkin methods for the 2d Helmholtz equation: analysis of the p-version, SIAM Journal on Numerical Analysis 49 (1) (2011) 264–284 (2011).

[13] R. Hiptmair, A. Moiola, I. Perugia, Plane Wave Discontinuous Galerkin Methods: Exponential Convergence of the *hp*-version, Foundations of Computational Mathematics 16 (3) (2016) 637–675 (2016).

[14] B. Després, Domain decomposition method and the Helmholtz problem, in: G. Cohen, L. Halpern, P. Joly (Eds.), Mathematical and numerical aspects of wave propagation phenomena (Strasbourg, 1991), SIAM, Philadelphia, PA, 1991, pp. 44–52 (1991).

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

[16] E. Perrey-Debain, Plane wave decomposition in the unit disc: Convergence estimates and computational aspects, Journal of Computational and Applied Mathematics 193 (1) (2006) 140–156 (2006).

[17] S. Congreve, J. Gedicke, I. Perugia, Numerical investigation of the conditioning for plane wave discontinuous Galerkin methods, Vol. 126 of Lecture Notes in Computational Science and Engineering, Springer, 2019 (2019).

[18] J. Nečas, Direct Methods in the Theory of Elliptic Equations, Springer, Heidelberg Dordrecht London New-York, 2012 (2012).

[19] S. A. Sauter, C. Schwab, Boundary Element Methods, Springer-Verlag, Berlin-Heidelberg, 2011 (2011).

[20] W. McLean, Strongly Elliptic Systems and Boundary Integral Equations, Cambridge University Press, Cambridge, UK, and New York, USA, 2000 (2000).

[21] G. C. Hsiao, W. L. Wendland, Boundary Iintegral Equations, Springer, Berlin-Heidelberg, 2008 (2008).

[22] S.-C. Brenner, L. R. Scott, The mathematical theory of finite element methods, Springer-Verlag, 1994 (1994).

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

[24] D. N. Arnold, F. Brezzi, B. Cockburn, L. D. Marini, Unified analysis of discontinuous Galerkin methods for elliptic problems, SIAM J. Num. Analysis 39 (5) (2002) 1749–1779 (2002).

[25] R. Hiptmair, A. Moiola, I. Perugia, A Survey of Trefftz Methods for the Helmholtz Equation, Springer International Publishing, 2016, Ch. 8, pp. 237–278 (2016).

[26] P. Gosselet, C. Rey, Non-overlapping domain decomposition methods in structural mechanics, Arch. Comput. Meth. Engng. 13 (2006) 515–572 (2006).

[27] B. Després, Domain decomposition method and the Helmholtz problem II, in: Second Internationl Conference on Mathematical and Numerical Aspects of Wave Propagation, Neuwark, Delaware, SIAM, Philadelphia, 2003, pp. 197–206 (2003).

[28] T. Huttunen, J. P. Kaipio, P. Monk, The perfectly matched layer for the ultra weak variational formulation of the 3D Helmholtz equation, International Journal for Numerical Methods in Engineering 61 (2004) 1072–1092 (2004).

[29] I. Bassi, L. Botti, A. Colombo, D. A. Di Pietro, P. Tesini, On the flexibility of agglomeration based physical space discontinuous Galerkin discretizations, J. Comp. Phys. 231 (1) (2012) 45–65 (2012).

[30] N. L. Trefethen, D. Bau, Numerical Linear Algebra, SIAM, Philadelphia, 1997 (1997).

[31] T. Huttunen, P. Monk, J. P. Kaipio, Computational aspects of the ultra-weak variational formulation, J. Comput. Phys. 182 (2002) 27–46 (2002).

[32] L. Mascotto, I. Perugia, A. Pichler, A nonconforming Trefftz virtual element method for the Helmholtz problem: numerical aspects, Computer Methods in Applied Mechanics and Engineering 347 (2019) 445–476 (2019).

[33] M. Ainsworth, P. Monk, 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) (2006) 5–40 (2006).

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

[35] J. Van Bladel, Electromagnetic Fields (Revised Printing), Hemisphere Publishing Corporation, New York, Washington, Philadelphia, London, 1985 (1985).

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