SIAM J. NUMER. ANAL.
Vol. 35, No. 1, pp. 255–299, February 1998

014

APPLICATION OF AN ULTRA WEAK VARIATIONAL FORMULATION OF ELLIPTIC PDES TO THE TWO-DIMENSIONAL HELMHOLTZ PROBLEM*

OLIVIER CESSENAT† AND BRUNO DESPRES†

Abstract. A new technique to solve elliptic linear PDEs, called ultra weak variational formulation (UWVF) in this paper, is introduced in [B. Després, *C. R. Acad. Sci. Paris*, 318 (1994), pp. 939–944]. This paper is devoted to an evaluation of the potentialities of this technique. It is applied to a model wave problem, the two-dimensional Helmholtz problem. The new method is presented in three parts following the same style of presentation as the classical one of the finite elements method, even though they are definitely conceptually different methods. The first part is committed to the variational formulation and to the continuous problem. The second part defines the discretization process using a Galerkin procedure. The third part actually studies the efficiency of the technique from the order of convergence point of view. This is achieved using theoretical proofs and a series of numerical experiments. In particular, it is proven and shown the order of convergence is lower bounded by a linear function of the number of degrees of freedom. An application to scattering problems is presented in a fourth part.

Key words. ultra weak, Helmholtz problem, Galerkin procedure, convergence order

AMS subject classifications. 65P05, 65N12

PII. S0036142995285873

Introduction. This work deals with the application of a new UWVF of elliptic linear partial differential equations to a model harmonic wave equation: the two-dimensional Helmholtz problem. This was developed for the first time in [13]. In this paper, Ω is a bounded domain with boundary Γ which is assumed to be sufficiently regular. This problem derives from the wave equation after suppressing the time dependence. This is formalized, normalizing the wave's velocity to 1, by

$$\begin{cases} -\Delta u - \omega^2 u = f & \text{in } \Omega\ , \\ (\partial_\nu + i\omega)u = t(-\partial_\nu + i\omega)u + g & \text{on } \Gamma\ , \\ |t| < 1,\ t \in \mathbb{C}\ . \end{cases} \tag{0.1}$$

The outer normal derivative is referred to by ∂_ν and the angular frequency by ω. A wide range of boundary conditions is provided by a complex constant t. For instance, taking $t = 0$ corresponds to an absorbing boundary condition of zero order. We shall assume $|t| \leq \delta < 1$ so that (0.1) is a well-posed problem in the $H^1(\Omega)$ framework.

The main difficulties in numerically solving harmonic wave problems lie in the noncoercive nature of the problem and in the fact that the solution is oscillatory with a wavelength $\lambda = 2\pi/\omega$. The classical methods in numerical analysis used to solve harmonic wave problems are mainly the following:

1. The integral equations [20], [9], [3] that are well adapted to scattering problems by obstacles in an infinite space since they take full account of the exact radiation condition. Also, it reduces an N-dimensional problem to an $(N-1)$-dimensional one since it restricts the calculations to the boundary of the obstacle. This technique is

*Received by the editors May 10, 1995; accepted for publication (in revised form) September 6, 1996.

http://www.siam.org/journals/sinum/35-1/28587.html

†Commissariat à l'Energie Atomique, CEL-V, F-94 195 Villeneuve St. Georges Cedex, France (cessenat@limeil.cea.fr, despres@limeil.cea.fr).

limited to the study of objects whose features can be modeled by a surface operator, such as a Leontovich (or impedance) condition for instance.

2. The finite element method (FEM), finite volume [16], and finite difference [26] methods are able to solve problems in nonhomogeneous media. For a problem in an unbounded domain they demand either to truncate it and set free absorbing conditions, or to be coupled with an integral equation method. The numerical advantage of these techniques is that they lead to a sparse matrix. Their main drawback is to demand meshing a volume around the diffracting object. This dramatically increases the number of unknowns and forces the implementation of high order absorbing boundary conditions that may be very costly as well.

To our opinion, the main setback of both methods is rooted in the fact that they all lead to a system of the form

$$(F - K)u = b \tag{0.2}$$

in which K is a compact perturbation of the coercive operator F. The injectivity of the discretized matrix constructed from $F - K$ is conditioned by the discretization step. For instance, in the FEM, a

$$h \approx \lambda/10 \tag{0.3}$$

discretization step is found to be practically required. The integral equations method is also constrained by a relation like (0.3), with a $\lambda/5$ surface mesh.

We present here a new approach to solving harmonic wave problems. This method originates from the domain decomposition techniques for which we refer the reader to [11]. In contrast to classical methods, the main elements of the new formulation are the following.

1. The possible use of discontinuous basis functions (not H^1 basis functions as generally in the FEM). The test functions are solutions of the adjoint problem (see [7]). This explains the "ultra weak formulation" terminology. The basis functions belong to the L^2 space of the boundaries of the elements of the mesh. They are constructed by Robin mixed boundary conditions. The equivalent of the "Neumann–Dirichlet" operator is bounded in L^2, while the classical "Neumann–Dirichlet" operator (also called the "Steklov–Poincaré" operator [10]) works in $H^{-1/2}$ on $H^{1/2}$; see also [24] and [15] in the domain decomposition context, [23] for hybrid methods.

2. The discretized problem is unconditionally well posed. There is no relation like (0.3) binding the stability of the linear system to the mesh size parameter.

3. The linear system is solved using a linear iterative algorithm which is similar to a fixed point algorithm, namely the Richardson algorithm [21]. This type of relaxed algorithm has been studied in a domain decomposition context dealing with coercive operators by [19] and [18].

4. The order of convergence of the UWVF method appears to be lower bounded by a linear function of the number of basis functions per elements. This is not the case in the FEM where the order of convergence is a square root of the number of degrees of freedom [8]. This means that, for the same accuracy level, our UWVF method needs less computational storage than the FEM. Numerical quadratures [22], [1] may reduce the computational storage demanded by the FEM. However, such a technique is essentially useful when the mesh is made of very simple shapes of elements (such as square or cubic shapes). For other types of elements (such as triangles, nonlinear elements), quadrature formulas are much more difficult to implement and analyze. In our technique there is no need to use quadrature techniques: the order of convergence

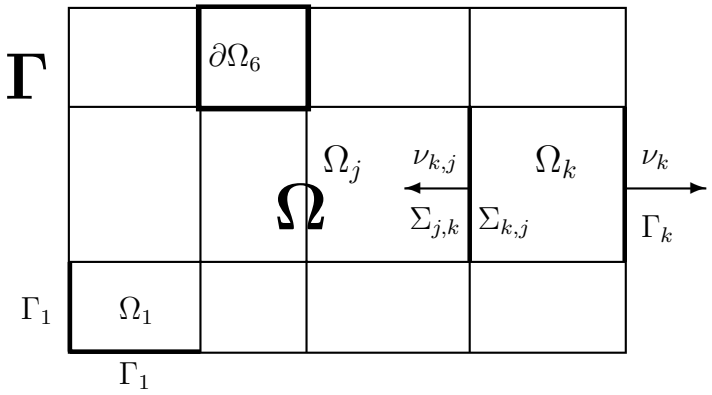


FIG. 0.1. *Partition of an* Ω *domain into elements* Ω_k.

is reached on any mesh, provided it is uniform regular. Furthermore, quadrature techniques applied to the FEM do not decrease the number of degrees of freedom: they decrease only the size of the matrix.

It could also be interesting to compare the accuracy of our method with that of spectral approximations [1] which are also high order methods. A summary of these techniques can be found in [22].

The basic idea of the ultra weak formulation is to consider a new unknown as well as a new continuous problem from which we will be able to go back to the original problem (0.1). To achieve this, let us consider a partition of Ω in the sense that

$$\begin{aligned} &\overline{\Omega} = \cup \overline{\Omega_k},\ \Omega_k \cap \Omega_j = \emptyset \text{ for } k \neq j\ , \\ &\Gamma_k = \overline{\Omega_k} \cap \Gamma\ , \\ &\Sigma_{kj} = \overline{\Omega_k} \cap \overline{\Omega_j} \text{ oriented from } \Omega_k \text{ to } \Omega_j\ , \\ &\partial\Omega_k = (\cup_j \Sigma_{kj}) \cup \Gamma_k\ . \end{aligned} \tag{0.4}$$

Figure 0.1 helps to visualize these definitions. In practice, the partition is a mesh of domain Ω. The sets Ω_k are the elements. We denote by edge any of the interfaces Σ_{kj} or free edges Γ_k. Let us define the framework V of the UWVF formulation as

$$V = \prod_{k=1}^{K} L^2(\partial\Omega_k) \tag{0.5}$$

with the natural scalar product

$$(x, y) = \sum_k \int_{\partial\Omega_k} x_{|\partial\Omega_k}\, \overline{y_{|\partial\Omega_k}} \tag{0.6}$$

that defines the norm on $||.||_V$ and an induced norm of any operator $A \in \mathcal{V}$ by

$$||A|| = \sup_{x \neq 0} \frac{||Ax||_V}{||x||_V}\ . \tag{0.7}$$

Remark 1. The framework V depends on the mesh, but it is not a discretization space of finite dimension. We should refer to it using a subscript K to designate the

mesh. Nevertheless, we keep the V notation, adopted in [13] and [6], in order to have shorter notation. Let ∂_{ν_k} denote the outer normal derivative to $\partial\Omega_k$. The value of the unknown x of the UWVF formulation will be defined from u solution of (0.1) as being

$$x_{|\partial\Omega_k} = ((-\partial_{\nu_k} + i\omega)u_{|\Omega_k})_{|\partial\Omega_k} \ , \tag{0.8}$$

assuming the regularity hypothesis $x \in V$. This hypothesis is valid as soon as u is smooth enough. So, the first important feature of the UWVF formulation is to lead to a system whose unknowns are functions defined on interfaces. The resulting problem will appear as a linear system of the form

$$\begin{cases} \text{for } b \in V, \text{ find } x \in V \ , \\ (I - A)x = b \end{cases} \tag{0.9}$$

where A denotes a linear operator in V satisfying $||A|| \leq 1$ and $(I - A)$ is invertible. The form of this operator will be made precise later.

Let us point out that this formulation is fairly general as it holds for any linear PDE involving an elliptic operator (see [14]). Also, it is still an exact formulation of the continuous problem even if it is linked with a mesh of Ω (via the subdomains or elements Ω_k). The full discretization of the problem will appear as a second step of the method. This is the first time this formulation is used for the numerical solution of PDEs.

The use of a mesh makes the practical framework of our method similar to that of the FEM. We, therefore, concentrate on comparing our formulation to that of the FEM. This naturally leads us to follow the classical presentation of the FEM.

• Section 1 presents the new formulation. Under the regularity hypothesis $\forall k \ \partial u/\partial\nu_k \in L^2(\partial\Omega_k)$, we define x and V as in (0.8) and (0.5) and we prove (Theorem 1.3) that the new variational formulation in V is equivalent to the original Helmholtz problem (0.1) in $H^1(\Omega)$ using an extension operator. This shows the UWVF problem is well posed. The dual space of the formulation is made of z functions whose restrictions z_k to $\partial\Omega_k$ are constructed from e functions that are solutions of the dual homogeneous Helmholtz problem as follows

$$\begin{cases} (-\Delta - \omega^2)e_k = 0 & \text{in } \Omega_k \\ (-\partial_\nu + i\omega)e_k = z_k & \text{on } \partial\Omega_k \ . \end{cases} \tag{0.10}$$

PROPOSITION 0.1. *There exists an operator A in $\mathcal{L}(V)$ and b in V such that $(I - A)x = b$ and*

$$\begin{cases} ||A|| \leq 1 & \text{(Proposition 1.10)} \\ (I - A) \text{ is injective} & \text{(Proposition 1.11).} \end{cases} \tag{0.11}$$

• Section 2 presents the discretization process that follows from a Galerkin procedure. The approximate solution x_h will be looked for in a finite dimensional space V_h subspace of V and will induce the approximate solution of u denoted by u_h. The Galerkin space V_h is constructed using p basis functions $z_{kl,l=1..p}$ on any element Ω_k (with k from 1 to K) that are solutions of the dual homogeneous problem (0.10) with a support included in Ω_k. The formal discretized problem writes

$$(I - P_h A)x_h = P_h b \tag{0.12}$$

and leads to the pK-finite dimensional linear system

$$(D-C)X = b \tag{0.13}$$

where $X = (x_{kl})_{kl} \in \mathbb{C}^{pK}$ and $(x_h)_{|\partial\Omega_k} = \sum_{l=1}^{p} x_{kl} z_{kl}$. We can state the following proposition.

PROPOSITION 0.2. *The system* (0.13) *is invertible for any mesh size parameter* h *(Theorem* 2.1*). The linear system's matrix* $(D-C)$ *is sparse.*

The use of plane waves for the construction of basis functions allow an analytical calculation of the matrices D (A.1) and C ((A.2) and (A.3)). Since $||A|| \leq 1$, a convergent linear iterative algorithm is proposed to solve it ((2.33) or (2.34)). Notice that one controls the quality of the approximation using two parameters: the mesh size on one hand (defining V), the number of basis functions on the other hand.

• Section 3 is dedicated to the order of convergence of the method. It is divided into two parts: A numerical part to show the order of convergence is bounded by linear functions of the number of basis functions per element, a theoretical part to explain these computational results, leading to the two essential statements, Proposition 0.3 and Proposition 0.4 which are Corollaries 3.8 and 3.9 of Theorem 3.7. The first result concerns the homogeneous problem.

PROPOSITION 0.3. *Let* u *be a solution of* (0.1) *with* $f = 0$ *(this defines the "homogeneous" terminology). Let* x *satisfy* (0.5) *and* x_h *be a solution of* (0.12). *Let* $[\alpha]$ *stand for the integer part of* α. *We assume* u *is of class* $C^{[(p+1)/2]}(\Omega)$ *where* $p \geq 3$ *is the number of basis functions per element. Let us assume the basis functions are plane waves with given directions. Then,*

$$||x - x_h||_{L^2(\Gamma)} \leq C h^{[(p-1)/2]-1/2} ||u||_{C^{[(p+1)/2]}(\Omega)}. \tag{0.14}$$

The order of convergence being defined as the exponent of h *is thus* $[(p-1)/2] - 1/2$. *It is a linear function of* $[p/2]$. *For instance* $p = 3$ *or* $p = 4$ *gives* $h^{1/2}$.

Using a duality technique, a linear function lower bounds the order of convergence in nonhomogeneous problems, as is stated in Proposition 0.4.

PROPOSITION 0.4. *Let* u *be a solution of the general* (0.1) *problem. Let* x *satisfy* (0.5) *and* x_h *be a solution of* (0.12). *We assume* u *is of class* $C^2(\Omega)$ *and* $p \geq 3$ *is the number of basis functions per element and that the basis functions are plane waves with given directions, then*

$$\forall s > [(p-1)/2] - 1/2,\ ||x - x_h||_{H^{-s}(\Gamma)} \leq C h^{[(p-1)/2]} ||u||_{C^2(\Omega)}\ . \tag{0.15}$$

The order of convergence is thus $[(p-1)/2]$. *It is a linear function of* $[p/2]$. *For instance* $p = 3$ *or* $p = 4$ *gives* h^1.

These results are achieved using only

$$\begin{cases} ||A|| \leq 1, \\ (I - A) \text{ is injective}, \\ \text{the basis functions are constructed from plane waves.} \end{cases} \tag{0.16}$$

This is why we believe these results could be generalized to all linear PDEs whose discretized formulation is written in the form of (0.9).

• Section 4 presents two scattering test cases with radar cross section computations since this is one of the grounds for studying frequency wave problems.

1. Basic classical results. Let us give some classical results concerning the solutions of (0.1): existence and uniqueness theorem (1.1) and regularity theorem (1.2). As usual, the regularity theorem assumes more regularity on the boundary than the existence and uniqueness theorem.

1.1. Existence and uniqueness.

THEOREM 1.1. *Let Ω be an open bounded set, and Γ be its boundary assuming it is of class C^1 nearly everywhere. Let $f \in L^2(\Omega)$ and $g \in L^2(\Gamma)$. We let $\zeta = 1 - t/1 + t$ and assume t to be constant, $|t| < 1$ (then $\Re(\zeta) > 0$). Then, there exists a unique $u \in H^1(\Omega)$ satisfying*

$$\left\{ \begin{array}{l} \forall v \in H^1(\Omega) \\ \displaystyle\int_\Omega \nabla u \, \overline{\nabla v} - \omega^2 \int_\Omega u \, \overline{v} + i\omega\zeta \int_\Gamma u \, \overline{v} = \int_\Omega f \, \overline{v} + \frac{1}{1+t} \int_\Gamma g \, \overline{v} \ , \end{array} \right. \tag{1.1}$$

or equivalently

$$\left\{ \begin{array}{ll} -(\Delta + \omega^2) u = f & \text{in } \Omega \ , \\ (\partial_\nu + i\omega) u = t(-\partial_\nu + i\omega) u + g & \text{on } \Gamma. \end{array} \right. \tag{1.2}$$

Proof. See [11]. ∎

1.2. A standard regularity result associated with problem (1.2).

THEOREM 1.2. *Let Ω be an open bounded set of $\mathbb{R}^n$ of boundary Γ. Let us assume that Γ is an infinitely differentiable manifold in $\mathbb{R}^{n-1}$ and that Ω is locally on one side of Γ. Let $s \in \mathbb{R}, s \geq 2$ and $f \in H^{s-2}(\Omega)$ and $g \in H^{s-3/2}(\Gamma)$. Then, the solution u of (0.1) has the regularity $u \in H^s(\Omega)$ and one has the following estimate:*

$$||u||_{H^s(\Omega)} \leq C \left\{ ||f||_{H^{s-2}(\Omega)} + ||g||_{H^{s-3/2}(\Gamma)} \right\} \ . \tag{1.3}$$

Also, for $f = 0$, the estimate (1.3) remains true $\forall s \in \mathbb{R}$.

Proof. See [17, p. 202, Remark 7.2]. ∎

Remark 2. If $u \in H^1(\Omega)$, $u_{|\Gamma} \in H^{1/2}(\Gamma)$. If the trace of u on Γ is $L^2(\Gamma)$, and if g is also at least $L^2(\Gamma)$ then $\partial_\nu u \in L^2(\Gamma)$.

1.3. A new variational formulation. In Theorem 1.3 we establish a new problem which is equivalent, in the sense that we give in the statement of the theorem, to the original one and whose unknowns are defined on the edges $\partial\Omega_k$. Theorem 1.9 rewrites this problem, using formal operators, in the new variational form of (1.21) that we referred to in the introduction as the ultra weak one. The fundamental properties of system (1.23) are presented by Propositions 1.10 and 1.11.

THEOREM 1.3. *Let $u \in H^1(\Omega)$ satisfy the regularity hypothesis $\partial_{\nu_k} u \in L^2(\partial\Omega_k)$ for any k and be a solution of the Helmholtz problem (0.1). Then x defined by $x_{|\partial\Omega_k} = x_k$ with $x_k = ((-\partial_{\nu_k} + i\omega) u_{|\Omega_k})_{|\partial\Omega_k}$ satisfies*

$$\begin{aligned} &\left(\sum_k \int_{\partial\Omega_k} x_k \, \overline{(-\partial_{\nu_k} + i\omega) e_k} \right) \\ &- \left(\sum_{k,j} \int_{\Sigma_{kj}} x_j \, \overline{(\partial_{\nu_k} + i\omega) e_k} + \sum_k \int_{\Gamma_k} t x_k \, \overline{(\partial_{\nu_k} + i\omega) e_k} \right) \\ &= -2i\omega \sum_k \int_{\Omega_k} f \, \overline{e_k} + \sum_k \int_{\Gamma_k} g \, \overline{(\partial_{\nu_k} + i\omega) e_k} \end{aligned} \tag{1.4}$$

for any

$$e \in H,\ e = (e_k)_{k=1,\dots,K}, H = \prod_{k=1}^{K} H_k \tag{1.5}$$

with

$$H_k = \left\{ v_k \in H^1(\Omega_k) \,\middle|\, \begin{array}{l} (-\Delta - \omega^2) v_k = 0 \ in\ \Omega_k \\ (-\partial_{\nu_k} + i\omega)(v_k)_{|\partial\Omega_k} \in L^2(\partial\Omega_k) \end{array} \right\} . \tag{1.6}$$

Conversely, if x is solution of (1.4) *then the function u defined by*

$$\left| \begin{array}{l} u_{|\Omega_k} = u_k \ , \\ (-\Delta - \omega^2) u_k = f_{|\Omega_k} \ , \\ (-\partial_\nu + i\omega) u_k = x_k \ . \end{array} \right. \tag{1.7}$$

where $u_k \in H^1(\Omega_k)$ is the unique solution of (0.1).

Proof. By hypotheses, $u \in H^1(\Omega)$, so the trace of u on the boundary $\partial\Omega_k$ of any element Ω_k is $u_{|\partial\Omega_k} \in H^{1/2}(\partial\Omega_k)$, since the boundary is of class C^1 nearly everywhere. It is also assumed that $\partial_\nu u \in V$. This allows us to write

$$\begin{aligned} &\int_{\partial\Omega_k} (-\partial_{\nu_k} + i\omega) u \, \overline{(-\partial_{\nu_k} + i\omega) e_k} \\ &= \int_{\partial\Omega_k} (+\partial_{\nu_k} + i\omega) u \, \overline{(+\partial_{\nu_k} + i\omega) e_k} - 2i\omega \int_{\partial\Omega_k} \left(u \, (\overline{\partial_{\nu_k} e_k}) - (\partial_{\nu_k} u) \, \overline{e_k} \right) . \end{aligned} \tag{1.8}$$

Recall that, according to (0.1) and (1.6),

$$\left\{ \begin{array}{ll} (-\Delta - \omega^2) \, \overline{e}_k = 0 & \text{in } \Omega_k \ , \\ (-\Delta - \omega^2) u = f & \text{in } \Omega \ ; \end{array} \right. \tag{1.9}$$

thus, with suitable integrations by parts in (1.9),

$$\left\{ \begin{array}{l} \displaystyle\int_{\Omega_k} \nabla u \, \overline{\nabla e_k} - \omega^2 u \, \overline{e}_k = \int_{\partial\Omega_k} u (\overline{\partial_{\nu_k} e_k}) \ , \\ \displaystyle\int_{\Omega_k} \nabla u \, \overline{\nabla e_k} - \omega^2 u \, \overline{e}_k = \int_{\partial\Omega_k} (\partial_{\nu_k} u) \, \overline{e}_k + \int_{\Omega_k} f \, \overline{e}_k \ . \end{array} \right. \tag{1.10}$$

From (1.10) and (1.8) we have

$$\begin{aligned} &\int_{\partial\Omega_k} (-\partial_{\nu_k} + i\omega) u \, \overline{(-\partial_{\nu_k} + i\omega) e_k} \\ &\qquad - \int_{\partial\Omega_k} (+\partial_{\nu_k} + i\omega) u \, \overline{(+\partial_{\nu_k} + i\omega) e_k} = -2i\omega \int_{\Omega_k} f \, \overline{e}_k \ . \end{aligned} \tag{1.11}$$

The continuity of u on Σ_{kj} and the boundary condition of system (0.1) write

$$\left\{ \begin{array}{l} (+\partial_{\nu_k} + i\omega) u_{|\Sigma_{kj}} = (-\partial_{\nu_j} + i\omega) u_{|\Sigma_{jk}} \ , \\ (+\partial_{\nu_k} + i\omega) u_{|\Gamma_k} = t(-\partial_{\nu_k} + i\omega) u_{|\Gamma_k} + g \ . \end{array} \right. \tag{1.12}$$

In the second term $(+\partial_{\nu_k}+i\omega)u\,\overline{(+\partial_{\nu_k}+i\omega)e_k}$ of equation (1.11) we replace the summation over $\partial\Omega_k$ by a summation over Γ_k and Σ_{kj} using the relations (1.12). Then, summing for all k, we obtain equation (1.4).

Conversely, let x be solution of (1.4) and u defined by (1.7) for all its restrictions to Ω_k. By hypotheses on u and e, we have (1.9) that leads to (1.11). Summing on all the elements, we have

$$\begin{aligned}\sum_k \int_{\partial\Omega_k} x_k\,\overline{(-\partial_{\nu_k}+i\omega)e_k}\\ -\sum_k \int_{\partial\Omega_k}(+\partial_{\nu_k}+i\omega)u\,\overline{(+\partial_{\nu_k}+i\omega)e_k} = -2i\omega\sum_k\int_{\Omega_k} f\,\overline{e}_k\ .\end{aligned}\tag{1.13}$$

Since x satisfies (1.4) and combining with (1.13), we have

$$\begin{cases}\forall\ (+\partial_{\nu_k}+i\omega)e_k\in V,\ e_k\in H_k\ ,\\ \displaystyle\sum_{k,j}\int_{\Sigma_{kj}}(+\partial_{\nu_k}+i\omega)u\,\overline{(+\partial_{\nu_k}+i\omega)e_k}+\sum_k\int_{\Gamma_k}(+\partial_{\nu_k}+i\omega)u\,\overline{(+\partial_{\nu_k}+i\omega)e_k}\\ \qquad\displaystyle=\sum_{k,j}\int_{\Sigma_{kj}} x_j\,\overline{(\partial_{\nu_k}+i\omega)e_k}+\sum_k\int_{\Gamma_k}(tx_k+g)\,\overline{(\partial_{\nu_k}+i\omega)e_k}\ .\end{cases}\tag{1.14}$$

This yields (1.12). It is then clear that a function whose restrictions are $H^1(\Omega_k)$ solutions of (1.7) and that satisfies the continuity relations (1.12) is the solution of the Helmholtz problem (0.1)—such a function is admittable and it is the only one according to the uniqueness Theorem 1.1. ∎

Formulation (1.4) is summarized in Theorem 1.9 after introducing the operators E, F, Π, and A.

DEFINITION 1.4. *Let us define the extension mappings E and E_f by*

$$E_f=\begin{cases}V\to H=\displaystyle\prod_{k=1}^K H_k\ ,\\ z\mapsto e=(e_k),\ e_k=e_{|\Omega_k}\end{cases}\tag{1.15}$$

where e_k is the unique solution of

$$\begin{cases}(-\partial_{\nu_k}+i\omega)e_k=z_{|\partial\Omega_k} & \text{on } \partial\Omega_k\ ,\\ (-\Delta-\omega^2)e_k=f_{|\Omega_k} & \text{in } \Omega_k\ ,\end{cases}$$

and

$$E=E_0\ .\tag{1.16}$$

Remark 3. The mapping E is a linear operator, whereas E_f for $f\neq 0$ is an affine operator.

DEFINITION 1.5. *Let us define the operator $F\in\mathcal{L}(V)$ mapping the outgoing trace $(-\partial_{\nu_k}+i\omega)e_k$ to the incoming one $(\partial_{\nu_k}+i\omega)e_k$*

$$F=\begin{cases}V\to V\ ,\\ z\mapsto Fz=((\partial_{\nu_k}+i\omega)(E(z)_{|\Omega_k})_{|\partial\Omega_k})_k\ .\end{cases}\tag{1.17}$$

As previously pointed out, F is analogous to a "Neumann–Dirichlet" operator, with the originality to be applied to Robin conditions on the boundaries of the elements of the mesh. Also, we shall point out it is continuous in L^2 and not from $H^{1/2}$ on $H^{-1/2}$.

Remark 4. These definitions make sense according to Theorem 1.1. It implies that $(E(z)_{|\Omega_k})_{|\partial\Omega_k}$ exists and is unique in H. From $Fz = -z + 2i\omega(E(z)_{|\Omega_k})_{|\partial\Omega_k}$, we obtain $Fz \in \prod_k (L^2(\partial\Omega_k)) = V$. Let us note that

(1.18)
$$\prod_{k=1}^{K} H_k \text{ is isomorphic to } Y = \{(-\partial_{\nu_k} + i\omega)(v_k)_{|\partial\Omega_k}, v_k \in H_k, k = 1, \dots, K\} \ .$$

DEFINITION 1.6. *Considering a complex valued function t defined on Γ satisfying $|t| \leq 1$, let us define the linear operator Π by*

$$\Pi \in \mathcal{L}(\mathcal{V}) \begin{cases} \Pi z_{|\Sigma_{kj}} = z_{|\Sigma_{jk}} \ , \\ \Pi z_{|\Gamma_k} = t z_{|\Gamma_k} \ . \end{cases} \tag{1.19}$$

LEMMA 1.7. *The above operator Π obviously satisfies $||\Pi||_V \leq 1$ for $|t| \leq 1$.*

DEFINITION 1.8. *Let $F^* \in \mathcal{L}(\mathcal{V})$ (V is identified to its dual space) denote the adjoint of F. Let $A \in \mathcal{L}(\mathcal{V})$ be defined by*

$$A = F^* \Pi \ . \tag{1.20}$$

In equation (1.4) we can identify the first integral term with the scalar product in V, in the second term we recognize on the left-hand side the operator Π applied to x, and on the right-hand side the operator F applied to the basis function y. So we can state the existence and uniqueness theorem of the UWVF formulation, assuming the regularity hypothesis $x \in V$.

THEOREM 1.9.

a) *Problem* (1.4) *is equivalent to*

$$\begin{cases} \text{Find } x \in V \text{ such that } \forall y \in V \\ (x, y)_V - (\Pi x, Fy)_V = (b, y)_V \end{cases} \tag{1.21}$$

where the right-hand side $b \in V$ is defined, via the Riesz representation theorem, by

$$\forall y \in V \quad (b, y)_V = -2i\omega \sum_k \int_{\Omega_k} f \, \overline{E(y)}_{\Omega_k} + \sum_k \int_{\Gamma_k} g \, \overline{F(y)}_{\Gamma_k} \ . \tag{1.22}$$

b) *If u is solution of* (0.1) *then $x = (\partial_{\nu_k} + i\omega)u$ is solution of* (1.21).

c) *Conversely, if x is solution of* (1.21) *then $u = E_f(x)$ is the unique solution of* (0.1). *Problem* (1.21) *is equivalent to*

$$\begin{cases} \text{Given } b \in V, \text{ find } x \in V \ , \\ \boxed{(I - A)x = b} \end{cases} \tag{1.23}$$

Proof. Let $x = (-\partial_{\nu_k} + i\omega)u$ (as in (0.8)), and $y \in V$ be given. From y, we define e by $E(y) = e$ (using the existence and uniqueness Theorem 1.1 for $f = 0$). Then

$y = (-\partial_{\nu_k} + i\omega)e_k$. Using the equality (1.4) and the definitions of F (1.17) and Π (1.19) we obtain (1.22). Since (1.22) is valid for all $y \in V$ (E is defined on V), we can state (1.23). The converse is given by Theorem 1.1. ∎

The operator A satisfies the essential following properties.

PROPOSITION 1.10. *The induced norm of A satisfies $||A|| \leq 1$.*

PROPOSITION 1.11. *The operator $(I - A)$ is injective.*

The proofs are easy using the following fundamental lemma.

LEMMA 1.12. *The operator F is a bijective isometry in V.*

Proof. Let y satisfy $e = E(y)$ (as in (1.15)). Let $\Im(\alpha)$ denote the imaginary part of a complex number α. Then,

$$(Fy, Fy) = \int_{\partial\Omega_k} \left| \left(\frac{\partial}{\partial \nu_k} + i\omega \right) e_k \right|^2 = \int_{\partial\Omega_k} \left| \frac{\partial}{\partial \nu_k} e_k \right|^2 + \omega^2 |e_k|^2 - 2\omega \Im \left(\int_{\partial\Omega_k} \frac{\partial}{\partial \nu_k} e_k . \overline{e_k} \right) .$$

Integrating by parts in $\int_{\Omega_k} (-\Delta - \omega^2) e_k \overline{e}_k = 0$, we have

$$\int_{\Omega_k} |\nabla e_k|^2 - \omega^2 |e_k|^2 = \int_{\partial\Omega_k} \frac{\partial}{\partial \nu_k} e_k . \overline{e_k} \ \in \mathbb{R},$$

hence,

$$\int_{\partial\Omega_k} \left| \left(\frac{\partial}{\partial \nu_k} + i\omega \right) e_k \right|^2 = \int_{\partial\Omega_k} \left| \frac{\partial}{\partial \nu_k} e_k \right|^2 + \omega^2 |e_k|^2 = \int_{\partial\Omega_k} \left| \left(-\frac{\partial}{\partial \nu_k} + i\omega \right) e_k \right|^2 .$$

This implies that

$$\|Fy\|^2 = \sum_k \int_{\partial\Omega_k} \left| \left(\frac{\partial}{\partial \nu_k} + i\omega \right) e_k \right|^2 = \sum_k \int_{\partial\Omega_k} \left| \left(-\frac{\partial}{\partial \nu_k} + i\omega \right) e_k \right|^2 = \|y\|^2 .$$

Since this is true for any y of V, we have $F^*F = I$. The adjoint F^* is the function that to $(\partial_{\nu_k} + i\omega)e_k$ associates $(-\partial_{\nu_k} + i\omega)e_k$. It is thus obviously an injective isometry, so we have $FF^* = I$. ∎

Proof (of Proposition 1.10). The operator Π defined by (1.19) satisfies $\|\Pi\| \leq 1$ (Lemma 1.7). This, combined with Lemma 1.12 leads straight to $||A|| \leq 1$. ∎

Proof (of Proposition 1.11). This proof can be seen as a consequence of Theorem (1.9) that states the equivalence between the UWVF problem and the Helmholtz problem in $H^1(\Omega)$ (that ensures the regularity hypothesis $\partial_{\nu_k} u \in V$). It can be done explicitly as follows. We assume the existence of x such that $x = Ax$ and we check $x = 0$. Let $w = E(x)$ (E defined by (1.15)), in other words $x_{|\partial\Omega_k} = (-\partial_{\nu_k} + i\omega) w_{|\Omega_k}$ with

$$(-\Delta - \omega^2) w_{|\Omega_k} = 0 \ . \tag{1.24}$$

Equation $x = Ax$ multiplied to the left by operator F becomes $FF^*\Pi x = Fx$. This, using $FF^* = I$ since F is a bijective isometry (Lemma 1.12), rewrites on w:

$$(+\partial_{\nu_k} + i\omega) w_{|\Sigma_{kj}} = (-\partial_{\nu_j} + i\omega) w_{|\Sigma_{jk}}, \tag{1.25}$$

$$(+\partial_{\nu_k} + i\omega) w_{|\Gamma_k} = t(-\partial_{\nu_k} + i\omega) w_{|\Gamma_k} \ . \tag{1.26}$$

Equations (1.24) and (1.25) mean that w is a solution of a Helmholtz problem (0.1). Equation (1.24) means that the second member f is zero on Ω. Equation (1.26) means that g is zero on Γ. Under the hypotheses of the existence and uniqueness theorem 1.1, we know that $w = 0$. Thus, we have $x = 0$. ∎

2. Discretization of the variational formulation. In this section, we deal with the discretization of the variational problem (1.21). We proceed with a Galerkin-like method. We introduce a discretization space $V_h \subset V$ of finite dimension to obtain formulation (2.1) below.

$$\text{Find } x_h \in V_h, \text{ so that}$$

$$\begin{cases} \forall y_h \in V_h \\ (x_h, y_h)_V - (\Pi x_h, F y_h)_V = (b, y_h)_V \end{cases} \tag{2.1}$$

THEOREM 2.1. *Problem* (2.1) *has a unique solution.*

Proof. Let $P_h : V \to V_h$ be the orthogonal projection operator from V to V_h. Equation (2.1) means

$$(I - P_h A)x_h = P_h b \ . \tag{2.2}$$

Let us prove only the injectivity of (2.2) as V_h being a finite dimension space, the injectivity is equivalent to the bijectivity. That is to say that formulation (2.2) with $P_h b = 0$ leads to $x_h = 0$. Using P_h is an orthogonal projector, $(I - P_h A)x_h = 0$ is equivalent to

$$||x_h||^2 = ||P_h A x_h||^2 = ||A x_h||^2 - ||(I - P_h)A x_h||^2 \ . \tag{2.3}$$

Since $||A|| \leq 1$ (see Proposition 1.10) we have $||x_h||^2 \leq ||x_h||^2 - ||(I - P_h)A x_h||^2$ so $(I - P_h)A x_h = 0$. Using $(I - P_h A)x_h = 0$ this leads to $(I - A)x_h = 0$. Since $(I - A)$ is injective (see Proposition 1.11), we finally obtain $x_h = 0$. ∎

Let us draw the reader's attention on the fact Theorem 2.1 holds unconditionally. There is no assumption made on the mesh size as in the FEM or integral equations method.

2.1. Note on the V and V_h notation. Following from Remark 1, the notation V_h is borrowed from the finite element terminology (see [8]) and from the uniform regularity hypotheses H1, H2, and H3 applied to the mesh.

H1. $\partial\Omega_k$ is of class C^1 nearly everywhere.

H2. Let h_k be the diameter of Ω_k and ρ_k be the maximum of the diameters of the spheres inscribed in Ω_k. We demand that

$$\exists \sigma \text{ such that } h_k \leq \sigma \rho_k \ .$$

We then define the refinement parameter or mesh size parameter h by (2.4).

$$h = \max_k h_k \ . \tag{2.4}$$

The terminology "regular mesh" comes from [8].

H3. Σ_{kj} is a common edge of Ω_k and Ω_j. In the finite element method, the data of the degree of the polynomials (that are the basis functions) and of the mesh entirely define the approximation space V_h. In our method, generating the mesh is part of the study of the continuous problem, a preliminary to the study of the discrete problem. So, generating the mesh is equivalent to constructing the continuous space V, and the choice of the basis functions remains free. Let V_h denote the subspace of V generated by a finite number of known basis functions. The subscript notation h stands for both the finite dimension and, in the case of a uniformly regular mesh, the mesh size parameter.

2.2. Construction of V_h. In this paper, when speaking of a homogeneous problem, we refer the reader to a problem where $f = 0$ in $L^2(\Omega)$. We refer to a nonhomogeneous problem otherwise. In each element Ω_k, we introduce a finite number of functions e_{kl} supported in Ω_k and that are independent solutions of the homogeneous Helmholtz equation in the element Ω_k. A particular set of e_{kl} functions is given by those of compact support in Ω_k.

$$\begin{cases} (e_{kl})_{|\Omega_j} = 0 \text{ if } k \neq j \ , \\ (-\Delta - \omega^2)(e_{kl})_{|\Omega_k} = 0 \text{ and } e_{kl} \neq 0 \ . \end{cases} \tag{2.5}$$

The use of basis functions with a compact support, as usually done in the finite element method, leads to a sparse matrix.

The subscript k (varying from 1 to K the overall number of elements in the mesh) stands for the number of the element in which e_{kl} is not identically zero. The second subscript l is ranged from 1 to p the local number of basis functions in the element Ω_k. To simplify, we consider some constant number p of basis functions for all elements Ω_k. The functions z_{kl} that will form the basis of V_h are uniquely defined by $z_{kl} = (-\partial_{\nu_k} + i\omega)e_{kl}$, e_{kl} being a solution of system (2.5). Thus we have

$$\begin{cases} (z_{kl})_{|\partial\Omega_j} = 0 \text{ if } k \neq j \ , \\ (z_{kl})_{|\partial\Omega_k} = (-\partial_{\nu_k} + i\omega)e_{kl} \ . \end{cases} \tag{2.6}$$

One checks that $\{z_{kl}\}_{k,l}$ is a basis of V_h under the hypotheses of the following lemma.

LEMMA 2.2. *The functions $\{z_{kl}\}_{1\leq l\leq p}$ defined by* (2.6) *are linearly independent in V if and only if $\{e_{kl}\}_{1\leq l\leq p}$ are linearly independent in H_k.*

Proof. As in the proof of Proposition 1.11, this is a consequence of Theorem (1.9). More precisely, if the $\{e_{kl}\}_{1\leq l\leq p}$ family is not independent, it is obvious that the $\{z_{kl}\}_{1\leq l\leq p}$ functions will not form an independent set. Conversely, let us prove that if the e_{kl} functions are independent, then so are the z_{kl} functions. We shall assume the contrapositive statement. Let us define $y_h \in V$ and $w_h \in H$ (as in (1.5)) and $Y = (y_{kl})_{k,l} \in \mathbb{C}^{pK}$ by

$$\begin{cases} y_h = \displaystyle\sum_{l=1}^{p} y_{kl}\, z_{kl} \ , \\ w_h = E(y_h) \ , \end{cases} \tag{2.7}$$

assuming

$$||y_h||_V = 0 \ , \tag{2.8}$$

or, equivalently

$$\sum_k \int_{\partial\Omega_k} |(-\partial_{\nu_k} + i\omega)w_h|^2 = 0 \ . \tag{2.9}$$

Thus, w_h satisfies for all element Ω_k

$$\begin{cases} (-\Delta - \omega^2)w_h = 0 \text{ in } \Omega_k \ , \\ (-\partial_{\nu_k} + i\omega)w_h = 0 \text{ on } \partial\Omega_k \ . \end{cases} \tag{2.10}$$

According to Theorem 1.1, equation (2.10) implies that $w_h = 0$ on $L^2(\Omega_k)$ for any element Ω_k. The definition (2.7) yields $w_h = \sum_{l=1}^{p} y_{kl}\, e_{kl}$. This is summarized by

$$\forall k, \forall \vec{x} \in \Omega_k \ \sum_{l=1}^{p} y_{kl} e_{kl} = 0 \ , \tag{2.11}$$

which is the independence condition of the e_{kl} functions. □

Let us then define the finite dimensional space V_h as

$$V_h = \ \text{Vect} \ \left\{ (z_{kl})_{1\le k\le K}^{1\le l\le p} \right\} \ . \tag{2.12}$$

An approximate solution x_h is defined by its pK complex valued coefficients x_{kl} on the given basis functions z_{kl} as

$$(x_h)_{|\partial\Omega_k} = \sum_l x_{kl}\, z_{kl} \ , \tag{2.13}$$

or written using the e_{kl} functions

$$(x_h)_{|\partial\Omega_k} = \sum_l x_{kl}(-\partial_{\nu_k} + i\omega) e_{kl} \ . \tag{2.14}$$

2.2.1. Reconstruction of u_h from x_h. Theorem (1.3) assumes the equivalence between the UWVF formulation on x and the original Helmholtz problem (0.1) posed in $H^1(\Omega)$ with a partition into elements Ω_k that will make the extension operator E invertible. Here we describe how to compute an approximation of u (the solution of the Helmholtz problem (0.1)), denoted by u_h. There are two independent steps in the reconstruction of u_h.

i) Reconstruction of u_h on $\partial\Omega_k$. Since the UWVF method uses functions defined on edges as unknowns, it is not surprising that it is well adapted to compute the values of the Helmholtz problem (0.1) on the edges of the mesh. More precisely, let us distinguish between the edges that are interfaces (inside edges) and the ones that are free edges (boundary parts).

On Σ_{kj}, using $u_{|\Sigma_{kj}} = u_{|\Sigma_{jk}}$ and $\partial_{\nu_j} u_{|\Sigma_{jk}} = -\partial_{\nu_k} u_{|\Sigma_{kj}}$

$$\begin{aligned}(I+\Pi)x &= (-\partial_{\nu_k} + i\omega)u_{|\Sigma_{kj}} + (-\partial_{\nu_j} + i\omega)u_{|\Sigma_{jk}} \\ &= (-\partial_{\nu_k} + i\omega)u_{|\Sigma_{kj}} + (+\partial_{\nu_k} + i\omega)u_{|\Sigma_{kj}} \\ &= 2i\omega u \ .\end{aligned}$$

On Γ_k, using $(\partial_{\nu_k} + i\omega)u_{|\Gamma_k} = t(-\partial_{\nu_k} + i\omega)u_{|\Gamma_k} + g$

$$\begin{aligned}(I+\Pi)x + g &= (-\partial_{\nu_k} + i\omega)u_{|\Gamma_k} + t(-\partial_{\nu_k} + i\omega)u_{|\Gamma_k} + g \\ &= (-\partial_{\nu_k} + i\omega)u_{|\Gamma_k} + (+\partial_{\nu_k} + i\omega)u_{|\Gamma_k} \\ &= 2i\omega u \ .\end{aligned}$$

This is summed up by

$$\begin{cases} u = \dfrac{1}{2i\omega}[(I+\Pi)x] & \text{on } \Sigma_{kj} \ , \\ u = \dfrac{1}{2i\omega}[(I+\Pi)x + g] & \text{on } \Gamma_k \ . \end{cases} \tag{2.15}$$

It is therefore natural to define u_h on the edges of the mesh by

$$\begin{cases} u_h = \dfrac{1}{2i\omega}[(I+\Pi)x_h] & \text{on } \partial\Omega_k \ , \\ u_h = \dfrac{1}{2i\omega}[(I+\Pi)x_h + g] & \text{on } \Gamma_k \ . \end{cases} \tag{2.16}$$

Practically, equations (2.16) mean that we calculate u_h on V and Γ as follows:

1. We construct u_h on Γ by

$$2i\omega(u_h)_{|\Gamma_k} = g + (1+t_k)\sum_l x_{kl}(-\partial_{\nu_k} + i\omega)e_{kl} \ . \tag{2.17}$$

2. We construct u_h on Σ_{kj} by

$$2i\omega(u_h)_{|\Sigma_{kj}} = \sum_{l(k)} x_{kl}(-\partial_{\nu_k} + i\omega)e_{kl} + \sum_{l(j)} x_{jl}(-\partial_{\nu_j} + i\omega)e_{jl} \ . \tag{2.18}$$

ii) Reconstruction of u_h in Ω. As already stated by Theorem 1.3, we know that it is theoretically possible to invert the operator E_f restricted to Ω_k for all elements of the mesh. In practice, inverting

$$\begin{cases} -(\Delta + \omega^2)u = f & \text{in } \Omega_k \ , \\ (-\partial_\nu + i\omega)u = x & \text{on } \partial\Omega_k \ , \end{cases}$$

or in the discretized form

$$\begin{cases} -(\Delta + \omega^2)u_h = f & \text{in } \Omega_k \ , \\ (-\partial_\nu + i\omega)u_h = x_h & \text{on } \partial\Omega_k \ , \end{cases} \tag{2.19}$$

is *a priori* equivalent to solving the original Helmholtz problem. Nevertheless, it can be interesting to solve many problems in small domains rather than only one in a large domain. This is the original idea of the domain decomposition techniques. Nevertheless, in the particular case when $f = 0$ on Ω_k, equation (2.19) is easy to solve. It is clear that

$$(u_h)_{|\Omega_k} = \sum_{l=1}^{p} x_{kl}\, e_{kl} \tag{2.20}$$

is solution of problem (2.19) for all Ω_k. Let us draw the reader's attention to the fact that formula (2.20) can be used on the edges of the mesh instead of (2.17) and (2.18) if required.

Remark 5. In the nonhomogeneous problem, i.e., f is not identical to zero in the element Ω_k, we cannot solve equation (2.19) in a simple way. Formula (2.20) does not provide us with the solution of (2.19). It is, nevertheless, still possible to solve (2.19) using other methods, such as the finite element method. We can also accept a linear approximation from the edge values of u given by (2.16). On a practical point of view, graphic representation softwares usually require the node values of the mesh. This is obtained using formula (2.16). Another possibility is to continue using formula (2.20). This is indeed also a mere linear approximation of u on Ω_k knowing u on $\partial\Omega_k$. The convergence order in this case will be limited by this first-order linear approximate and possibly will not exceed h.

2.2.2. Construction of the discrete operators. We have to calculate the terms corresponding to the formulation (1.23). The discretized formulation (2.1) of (1.21) leads to the system (2.21), discrete form of (1.23).

$$\begin{cases} \text{Find } X \in \mathbb{C}^{pK} \ , \\ (D-C)X = b \ . \end{cases} \tag{2.21}$$

The matrix D is the matrix of the scalar product in V_h. The matrix referred to by the symbol C is the matrix of the bilinear form $(\Pi x_h, F y_h)$. For convenience, the right-hand side of the second equality in (2.21) is still denoted by b.

i) The coefficients of the matrix D defined by $D_{k,j}^{l,m} = (z_{jm}, z_{kl})_V$ are

$$D_{k,j}^{l,m} = \delta_{kj} \int_{\partial\Omega_k} (-\partial_{\nu_k} + i\omega) e_{jm} \, \overline{(-\partial_{\nu_k} + i\omega) e_{kl}} \ . \tag{2.22}$$

ii) The coefficients of the matrix C defined by $C_{k,j}^{l,m} = (\Pi z_{jm}, F z_{kl})_V$ are

$$\begin{cases} C_{k,j}^{l,m} = \displaystyle\int_{\Sigma_{kj}} (+\partial_{\nu_k} + i\omega) e_{jm} \, \overline{(+\partial_{\nu_k} + i\omega) e_{kl}} \\ \qquad\qquad + \displaystyle\int_{\Gamma_k} t \, (-\partial_{\nu_k} + i\omega) e_{jm} \, \overline{(+\partial_{\nu_k} + i\omega) e_{kl}} \end{cases} \tag{2.23}$$

iii) Let b be the second member defined by $b_{k,l} = (b, z_{kl})_V$

$$b_{k,l} = -2i\omega \int_{\Omega_k} f \, \overline{(e_{kl})} + \int_{\Gamma_k} g \, \overline{(+\partial_{\nu_k} + i\omega) e_{kl}} \ . \tag{2.24}$$

2.3. Example of numerical construction of V_h.

2.3.1. A practical choice of the approximation space. We have chosen to build our approximation space to the following specifications:

• A regular triangulation of Ω. The "regular" terminology refers to hypotheses H1, H2, and H3. Our interest in this is that it is easily done by a mesh generator.

• Basis functions derived from plane waves on Ω_k, i.e., of the form

$$\begin{cases} e_{kl} = e^{(i\omega \vec{v}_{kl} . \vec{x})} \ , \\ \vec{v}_{kl} \, \vec{v}_{kl} = 1 \ , \\ l \neq m \rightarrow \vec{v}_{kl} \neq \vec{v}_{km} \ . \end{cases} \tag{2.25}$$

Lemma 2.3 states the linear independence of such basis functions.

• During numerical simulations, the directions of the wave vectors of these plane waves have been chosen equidistributed in the plane

$$\forall k \in \{1, \ldots, K\}, \ \vec{v}_{kl, \ l=1,\ldots,p} = \begin{pmatrix} \cos\left(2\pi \dfrac{l-1}{p}\right) \\ \sin\left(2\pi \dfrac{l-1}{p}\right) \end{pmatrix} \ . \tag{2.26}$$

LEMMA 2.3. *The functions* $\{z_{kl}\}_{1 \leq l \leq p}$ *defined by* $z_{kl} = (-\partial_{\nu_k} + i\omega) e_{kl}$ *with* e_{kl} *satisfying* (2.25) *form a basis of* V_h.

Proof. Lemma 2.2 states the equivalence between the linear independence of z_{kl} in V_h and the one of e_{kl} in H (see (1.5)). The linear independence of plane waves is a well-known property of exponential functions on $\mathbb{R}^2$. It is anyway easily proven assuming there are nonzero coefficients that combine the e_{kl} functions so as to form the zero function

$$\sum_{l=1}^{p} \alpha_l e_{kl} = 0 \ . \tag{2.27}$$

Differentiating an exponential with respect to the two-space directions gives the original exponential multiplied by a constant, depending only on the exponential wave vector. Using the differential operator $(D_x + iD_y)$ (where the symbol D stands here only for the derivative operator) and applying it p times to (2.27), we obtain a linear system whose matrix has a Vandermonde determinant which is nonzero if and only if all the vectors $\vec{v}_{km}$ are different. ∎

Remark 6. Choosing a triangular mesh is not a prerequisite of the technique. We could equally have used a quadrangular or a mixed mesh. We restricted ourselves to this choice for the sake of the simplicity of the computation. Another choice would not have exhibited any fundamental difference. For the same reason of simplicity, we decided to consider the same number of basis functions per element.

Remark 7. The use of plane waves is justified by the fact that the calculation of the matrices is then analytic. Taking equidistributed directions for the plane waves is *a priori* justified since it is the simplest choice we could make, as long as the type of solution is unknown. We could have chosen the normals to the mesh triangles' edges as wave vectors of the plane waves, or used some particular directions corresponding to the asymptotic behavior of the solution. This choice has been experimented on some test cases and compared with the equidistributed choice. No major improvement has been found in taking these different possibilities. We found two reasons to take equidistributed plane waves.

1. It is slightly preferable for the numerical cost to have basis functions being all equal from one element to another. This makes the determination of the basis functions easier, and may be used in reducing the computer storage by $K_{int} \times p^2$, where K_{int} is the number of inner elements (elements having no free edge) and p is the number of basis functions per element (still assumed to be constant).

2. From a pure numerical point of view, note that the equidistributed choice maximizes the determinant of the matrices D_k when the mesh size parameter tends to 0 and when taking $p = 3$ basis functions per element (see [6]). This allows us to give a lower bound to the condition number.

2.3.2. Generating the matrices. The reader can refer to Appendix 5 for the analytic formulas that derive from (2.22), (2.23), and (2.24).

2.4. Inverting the linear system (2.21). System (2.21) is solved in two steps:
1. Invertibility of D.
2. Use of a converging linear iterative algorithm.

2.4.1. Invertibility of D, property of $M = D^{-1}C$.

LEMMA 2.4. *Let the matrix D be constructed from functions e_{kl} solutions of a homogeneous Helmholtz problem in each element Ω_k and linearly independent on the considered element. Then, D is Hermitian positive definite. If the basis functions derive from plane waves of directions $\vec{v}_{kl}$ in the element Ω_k, the definiteness of D is*

equivalent to the fact that the directions of the plane waves e_{kl} are all distinct (on each element), or formally $\forall k$, $\forall (i,j)$, $\vec{v}_{ki} \neq \vec{v}_{kj}$).

Proof. The matrix D represents the scalar product in V_h (since it inherits the properties of the continuous problem). It is thus Hermitian. Let us prove it is definite, i.e., find the vector independence condition of the basis functions z_{kl} for the scalar product of V. This is stated by Lemma 2.3. ∎

Let us define then

$$\begin{cases} b^{'} = D^{-1} b \ , \\ M = D^{-1} C \ . \end{cases} \tag{2.28}$$

LEMMA 2.5. *The eigenvalues of the matrix M defined by* (2.28) *are located inside the unit complex disk, the value* 1 *is excluded.*

Proof. Let $\lambda \in \mathbb{C}$ be an eigenvalue of M and $Y \in \mathbb{C}^N - \{0\}$ be an associated eigenvector, with y_{kl} its complex coefficients: $Y = (y_{kl})_{k,l}$.

i) Let us prove first that $|\lambda| \leq 1$. The equality $MY = \lambda Y$ and the definition (2.28) giving M yield

$$(CY, Y) = \lambda (DY, Y) \ . \tag{2.29}$$

Let $y_h \in V_h$ be the function whose coefficients y_{kl} are given by $(y_h)_{|\partial\Omega_k} = \sum_l y_{kl} z_{kl}$. By definition of C and D, we have

$$\begin{cases} (CY, Y) = (Ay_h, y_h)_V \ , \\ (DY, Y) = (y_h, y_h)_V \ . \end{cases} \tag{2.30}$$

Hence, using (2.29) and (2.30) we obtain

$$\lambda = \frac{(Ay_h, y_h)_V}{(y_h, y_h)_V} \ . \tag{2.31}$$

As $y_h \in V_h \subset V$ and $||A|| \leq 1$, equality (2.31) yields $|\lambda| \leq 1$.

ii) Let us prove now that the value $\lambda = 1$ is excluded. Supposing the opposite (that is to say $\lambda = 1$), equation (2.31) writes

$$(Ay_h, y_h)_V = (y_h, y_h)_V \ ,$$

but

$$\begin{aligned} ||(I - A) y_h||^2 &= ||y_h||^2 + ||Ay_h||^2 - 2\Re e \ (Ay_h, y_h)_V \\ &= ||Ay_h||^2 - ||y_h||^2 \leq 0 \ ; \end{aligned}$$

thus, $(I - A) y_h = 0$. The injectivity of $I - A$ yields $y_h = 0$, thus Y is zero. This is incompatible with Y being an eigenvector. ∎

2.4.2. Construction of an iterative algorithm. We shall invert system (2.21) using the underrelaxed discrete version of the Neumann series expansion of $(I - A)$. Despite the fact that such an expansion does not necessarily converge for any continuous operator A even with $||A|| \leq 1$, it does with the associated discrete underrelaxed operator. We solve system (2.32) using iterative algorithms derived from the domain decomposition method (see [11]). These algorithms are analogous to underrelaxed

Jacobi methods, already studied by [19] and [18]. System (2.21) can be expressed in the form of system (2.32) where M is defined in (2.28)

$$(I - M)X = b^{'} \ . \tag{2.32}$$

We have used two similar algorithms.

i) Let $\beta \in]0.5; 1[$ be fixed. The algorithm is

$$\begin{cases} X_1 = \beta b' \ , \\ X_{n+1} = \beta b' + [(1-\beta)I + \beta M]X_n \ . \end{cases} \tag{2.33}$$

ii) Let $\beta_n \in]0.5; 1-\varepsilon[, \ \varepsilon > 0$ be a sequence of random numbers. The algorithm is

$$\beta_n \in]0,5; 1-\varepsilon[, \ \varepsilon > 0 \begin{cases} X_1 = \beta_1 b' \ , \\ X_{n+1} = \beta_n b' + [(1-\beta_n)I + \beta_n M]X_n \ . \end{cases} \tag{2.34}$$

This algorithm is "the Richardson's iterative algorithm" [21].

Remark 8. The introduction of a correction factor $\beta \in]0; 1[$ tends to globally gather the eigenvalues of M to the origin. This suggests an improvement in the convergence from a numerical point of view.

For the proof of the convergence of the algorithms, let us refer the reader to [25] for (2.33). For the proof of (2.34), the reader can refer to [21] who shows the convergence of the Richardson algorithm in very general situations. The proof can also be done as an extension of the proof made by [25] in the constant β_n coefficients case [5].

2.4.3. Numerical remarks on the iterative algorithm.

Remark 9. Algorithm (2.34) is more efficient than the simpler release (2.33). In practice, we took β_n randomly between 0.5 and 0.99 (using a uniform distribution law on this interval). We observed that a 10 to 20% reduction in the number of iterations resulted in a similar accuracy to algorithm (2.33). Let us note there could be a better choice of these coefficients than the purely random choice. This point has not been studied any further in this paper, but a discussion was made in [12].

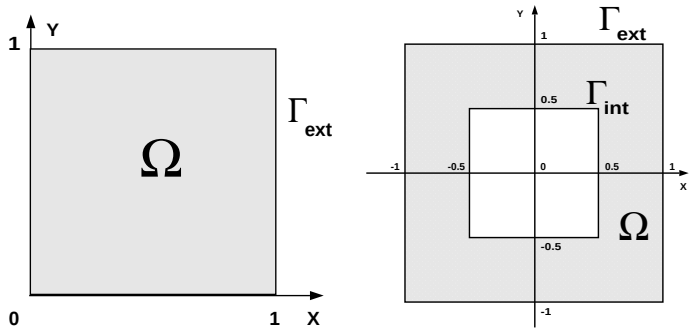
Remark 10. The computation time depends in decreasing order on the following routines.

i) The iterative algorithm (2.34) that consists of computing matrix to vector products.

ii) Computing the matrices D and C.

iii) Inverting matrices D_k (of reduced size, lower, in practice, than 15×15).

Remark 11. This type of algorithm is fairly robust, but it requires inverting D for the calculation of $M = D^{-1}C$. During the numerical tests to compute the order of the method of convergence, we have used many basis functions per element and small elements simultaneously. This situation leads to very accurate numerical results, accurate up to the square root of the machine's precision. Unfortunately, this situation also brings about serious condition number problems on the matrix D. The classical procedures we have implemented in our code for inverting a matrix (such as the LU decomposition, Cholesky ...) then happened to be insufficient and forced us to stop our asymptotic study at this point. Fortunately, in most practical cases, such accuracy is not needed.



a) Simply connected domain b) Nonsimply connected domain

FIG. 3.1. *Types of* Ω *domain.*

3. Study of the order of the method. The aim of this large section is to indicate how accurate this UWVF can be. It is organized as follows.

1. Recall of classical results of the FEM, as a reference for a comparison with our method (section 3.1).

2. Series of numerical tests to find out a heuristic convergence law (section 3.2).

3. Main estimates on the order of convergence (Corollaries 3.8 and 3.9 (section 3.3.6)).

4. Estimates of the condition number of the linear system resulting from the method (Theorem 3.10 (section 3.4)).

3.1. Notion of order for the FEM. Let Ω be a polygonal two-dimensional bounded set. The classical finite element theory (see [8]) proves that, when discretizing the problem

$$\begin{cases} -\Delta u = f & \text{in } \Omega \ , \\ u = 0 & \text{on } \Gamma \end{cases}$$

using finite P_k elements, we obtain the following estimate, for $u \in H^{k+1}(\Omega)$ with $k+1 \geq m$ (see [22]):

$$||u - u_h||_{H^m(\Omega)} \leq Ch^{k+1-m}||u||_{H^{k+1}(\Omega)} \ .$$

In the above, h is the mesh size parameter of the triangulation (we assume it is regular, as defined by hypotheses H1, H2, and H3, in section 2.1). The exponent of h is called the order of the method. For instance, for $k = 1$, we have

$$||u - u_h||_{L^2(\Omega)} \leq Ch^2||u||_{H^2(\Omega)} \ .$$

For $k = 2$, we have

$$||u - u_h||_{L^2(\Omega)} \leq Ch^3||u||_{H^3(\Omega)} \ .$$

3.2. Numerical study of the order of convergence. For the numerical simulations, we first considered two types of domain of Ω as indicated in Figure 3.1. In order to be able to easily compute an exact solution of the Helmholtz problem, we considered a simple solution, that is a plane wave propagating following the law (3.1)

where $\vec{v}_0$ is a complex wave vector, independent of the basis functions wave vectors.

(3.1)
$u = e^{i\omega \vec{v}_0 . \vec{x}}$ with $\vec{v}_0 = (v_0^1, v_0^2)$ and $\vec{x} = (x_1, x_2)$ denotes the position in the plane.

This type of solution, according to the values taken by $\vec{v}_0$, results from two different problems for which we analytically compute the second member (given the boundary conditions specified in sections 3.2.3 and 3.2.4). These two types of problems are

i) the homogeneous Helmholtz problem ($(v_0^1)^2 + (v_0^2)^2 = 1$),

ii) the nonhomogeneous Helmholtz problem ($(v_0^1)^2 + (v_0^2)^2 \neq 1$).

3.2.1. Estimate of the order of convergence. We computed, on test cases that are described in the following, the difference between the exact and the approximate solution using various norms, essentially $||x - x_h||_V$, $||x - x_h||_\Gamma$, $||u - u_h||_\Gamma$, $||u - u_h||_\Omega$. The number p of basis functions per element was fixed, while the mesh size parameter of the regular triangulation h tended to 0. We then computed the error between the exact solution x of (1.23) and the approximate x_h defined by (2.13): $x_h = \sum_{l=1}^{L} x_{kl}\, z_{kl}$. In the homogeneous problem, we shall also compute the errors between the exact solution u and the approximate u_h obtained using (2.20): $u_h = \sum_{l=1}^{L} x_{kl}\, e_{kl}$. In the nonhomogeneous problem, we can calculate u_h by (2.16). According to Lemma 3.4, section 3.3.3, we know that the order of u's trace at the boundary ($||u - u_h||_\Gamma$) will be at least equal to the order of $||x - x_h||_\Gamma$ (estimate (3.33)). We shall present a computation made using formula (2.20) (legitimate in the homogeneous problem) in order to demonstrate the drawbacks of this formula. We shall not write here how to compute the errors in L^2 norms, but let us point out that these computations are no more difficult than constructing the matrix D, and are still made using exact integrals.

3.2.2. Computational example of the order of convergence. We give Figure 3.2 the value of the relative error on u in $L^2(\Omega)$ norm as a function of $1/h$ and for various numbers of basis functions p per element. Let us recall that h is the mesh size parameter that we define by

$$h = \sqrt{\frac{\text{Surface of } \Omega}{\text{Total number of elements in the triangulation}}}. \tag{3.2}$$

The parameters of the example are given in Table 3.1. The procedure adopted for calculating the order of convergence consists in measuring the maximum negative slope of these curves (in logarithmic scale). Let us note that, by inspection, the pencil of Figure 3.2 is made of approximately parallel lines, for an odd number p and for $p + 1$.

Remark 12. We need to remove the points that give a positive slope. These points appear when the discretization becomes too fine (in space with the mesh size parameter h tending to zero, as in the number of basis functions p tending to infinity). This is explained by the link between the order and the condition number of the diagonal matrices D_k (A.1). The condition number of the diagonal matrices is greater than $h^{-[p/2]}$ for $p \geq 4$ as proven in Theorem 3.10, section 3.4 (and happened to be $h^{-2[p/2]}$ during these numerical tests). The matrices D_k^{-1} are then inaccurately calculated, spoiling our attempt to invert the linear system. This negates our gain in accuracy due to the large number of basis functions. Note that this is not a hindrance to the use of this method, since this occurs only when demanding a high accuracy level,

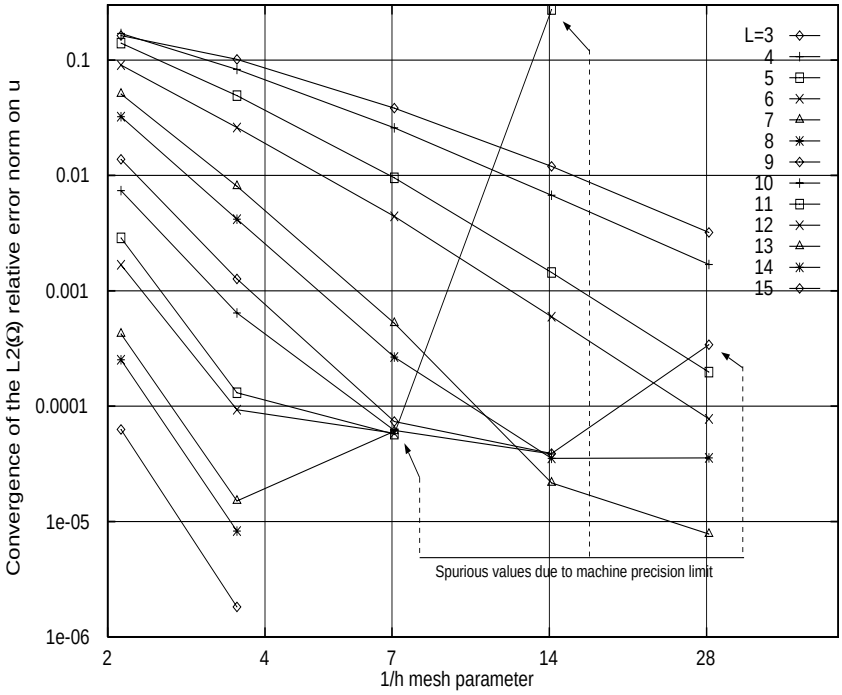


FIG. 3.2. *Pencil of convergence curves as functions of* $1/h$. *Number of basis functions per element varying from* 3 *to* 15.

TABLE 3.1
Parameters of the homogeneous example.

Variables	**Value**
f	0
g	see equation (3.3)
t	0.1
$\vec{v}_0$	$(1.009946454058, i \times 0.1413925035682)$
ω	4π

which is not needed in most practical cases. Nevertheless, for three basis functions per element, the condition number remains independent of h [6].

3.2.3. Homogeneous problem (Figures 3.3 and 3.4, Table 3.1). In equation (0.1) the boundary condition g is

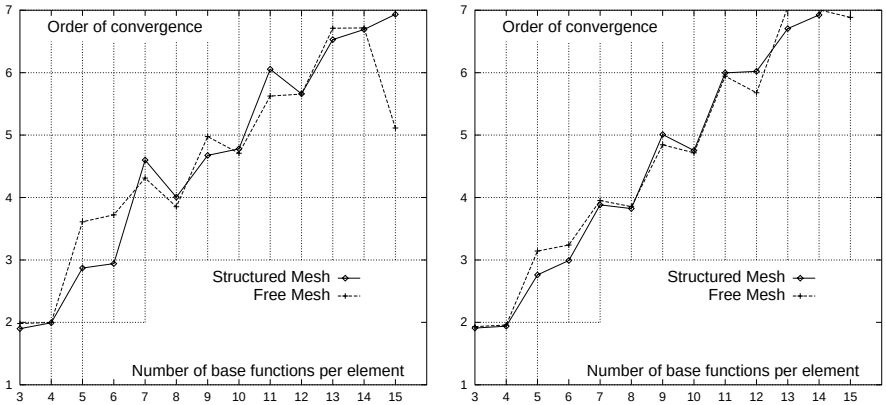
$$g = i\omega((1+t)\vec{\nu}\,\vec{v}_0 + (1-t))e^{i\omega\vec{v}_0.\vec{x}}\ , \tag{3.3}$$

with $\vec{v}_0 = (v_0^1, v_0^2)$ satisfying $(v_0^1)^2 + (v_0^2)^2 = 1$. The exact solution is $e^{i\omega\vec{v}_0.\vec{x}}$.

For $p = 3$ basis functions per element, we observe on the numerical test of Figure 3.3 (Table 3.1) that

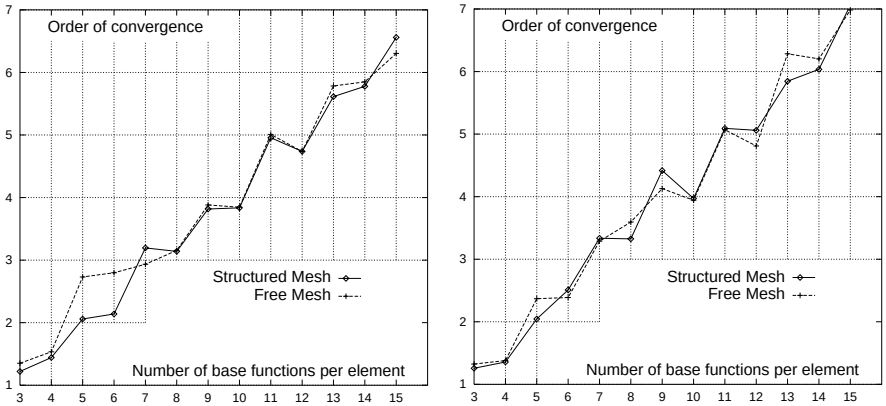
$$\begin{aligned} ||u - u_h||_{L^2(\Omega)} &\leq Ch^2, \\ ||u - u_h||_{L^2(\Gamma)} &\leq Ch^2\ . \end{aligned} \tag{3.4}$$

The estimate on Ω seems to be optimal. The estimate on Γ is better than expected by the trace theorems.



$L^2(\Omega)$ norm on u $L^2(\Gamma)$ norm on u

FIG. 3.3. *Experimental orders of convergence on u for $f = 0$.*



V norm on x $L^2(\Gamma)$ norm on x

FIG. 3.4. *Experimental orders of convergence on x for $f = 0$.*

3.2.4. Nonhomogeneous problem (Figure 3.5, Table 3.2). With respect to the previous section (i.e., 3.2.3), we consider $\vec{v}_0 = (v_0^1, v_0^2)$ satisfying

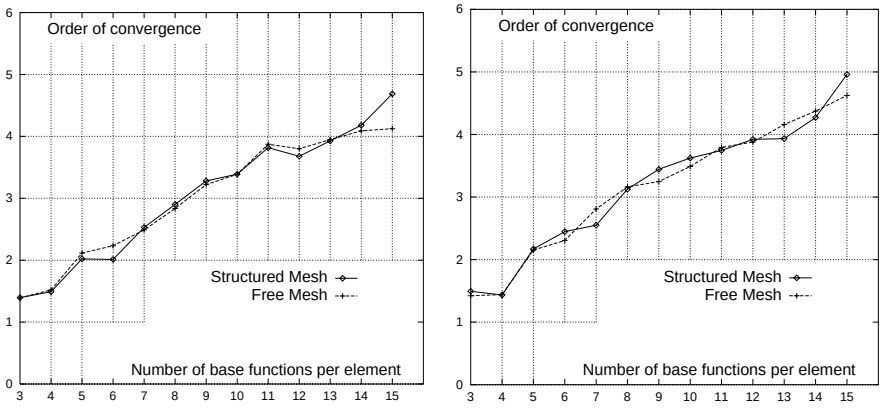
$$(v_0^1)^2 + (v_0^2)^2 = 1 + \mu,\ \mu \in \mathbb{C},\ \mu \neq 0\ .$$

This introduces a second nonzero right-hand side f. Thus, we have

$$\left\{ \begin{array}{l} f = \mu\omega^2 e^{(i\omega \vec{v}_0 . \vec{x})} \\ g = ((1+t)\partial_\nu + (1-t)i\omega) e^{(i\omega \vec{v}_0 . \vec{x})}\ . \end{array} \right. \tag{3.5}$$

The exact solution is $e^{i\omega \vec{v}_0 . \vec{x}}$.

Remark 13. According to Lemma 3.4, section 3.3.3 that gives the estimate (3.33), we know that the estimates of u on Γ remain fairly good so long as we compute u_h on Γ by (2.16), section 2.2.1. So the computation of the relative error on u_h in the $L^2(\Gamma)$ norm cannot lead to a poorer result than Figure 3.5b). Nevertheless, let us inquire what happens when we try to approximate u (on Γ and Ω) by a linear combination of the functions e_{kl} using formula (2.20), section 2.2.1.



a) V norm b) $L^2(\Gamma)$ norm

FIG. 3.5. *Experimental orders of convergence on x, nonhomogeneous example.*

TABLE 3.2
Parameters of the nonhomogeneous example.

Variables	**Value**
f	see Equation (3.5)
g	see Equation (3.5)
t	0.1
$\vec{v}_0$	$(1.4282799726, i \times 0.199959196164)$
ω	4π
μ	1

In such a case, the norms

$$\frac{||u-u_h||_{L^2(\Omega)}}{||u||_{L^2(\Omega)}} \text{ and } \frac{||u-u_h||_{L^2(\Gamma)}}{||u||_{L^2(\Gamma)}}$$

happened to be constant functions of p, the number of basis functions per element. They obey a law of convergence in h^1, the mesh size parameter.

This constant order of convergence was expected, for we attempted to approximate a solution of a nonhomogeneous problem by a homogeneous function. So far, the approximation of u in Ω remains rather poor. We should keep in mind that on Γ, we need to be sure to use the right approximation formula (2.16) and not the easiest to compute (2.20).

3.2.5. Approximate numerical convergence laws. We can roughly estimate that the norms that follow obey the laws of their respective tables

1. Both norms $\frac{||x-x_h||_V}{||x||_V}$ and $\frac{||x-x_h||_{L^2(\Gamma)}}{||x||_{L^2(\Gamma)}}$: Table 3.3.

2. The norm $\frac{||u-u_h||_{L^2(\Gamma)}}{||u||_{L^2(\Gamma)}}$: Table 3.3, using the general approximation formula (2.17).

3. The norm $\frac{||u-u_h||_{L^2(\Omega)}}{||u||_{L^2(\Omega)}}$: Table 3.4, using the approximation formula (2.20) that should be used in the homogeneous case only.

Remark 14. We observe a linear paired growth of the order of convergence as a function of the number of basis functions p per element.

TABLE 3.3
Experimental orders of convergence at the boundary.

p basis functions numb./elt		3 4	5 6	7 8	9 10	11 12	13 14
problem's order	Non homogeneous	1.5	2	2.5	3	3.5	4
	Homogeneous	1.5	2.5	3.5	4.5	5.5	?

TABLE 3.4
Experimental orders of convergence in the Ω domain.

p basis functions numb./elt		3 4	5 6	7 8	9 10	11 12	13 14
problem's order	Non homogeneous	1	1	1	1	1	1
	Homogeneous	2	3	4	5	6	?

Remark 15. As a first estimate, we can say that the slope of the law of convergence of the L^2 norms in the nonhomogeneous test problems is about half that in the homogeneous one (see section 3.2.3).

Remark 16. We have more complete numerical results in the nonhomogeneous study (see the question mark in Tables 3.3 and 3.4). Whereas the condition number does not depend on the second member f of problem (0.1), the convergence order of the nonhomogeneous problem increases at a slower rate than that of the homogeneous problem, by inspection about half. The loss of numerical accuracy due to the condition number matters more in a highly accurate computation (homogeneous problem) than in a lower one (nonhomogeneous problem). These inspections are theoretically explained in sections 3.3.6 and 3.4 with Corollaries 3.8 and 3.9 and Theorem 3.10.

Remark 17. It may be useful to read Remark 13 again to understand the differences in the nonhomogeneous example between Tables 3.3 and 3.4 concerning the trace of u on Γ. In the homogeneous case, we find an improvement on the norm of u on Γ using approximation (2.20) instead of the approximation (2.16).

3.2.6. Hypotheses drawn from the numerical study. In this numerical study the order of the method is bounded by linear functions of the number of degrees of freedom and not lower bounded by a square root as in the FEM. The storage in both methods remains a square function. This property shows the asymptotically greater accuracy of our method that will be checked later.

3.3. Theoretical study. The aim of this section is to explain, from a theoretical point of view, why the order of convergence behaves as a linear function of the integer part of half the number of basis functions per element. This is rooted in the use of basis functions that are solutions of the homogeneous Helmholtz problem. A duality technique, equivalent to the Aubin–Nitsche lemma, extends this law to nonhomogeneous problems. To simplify this study, we made theoretical estimates only for the boundary traces on Γ. These estimates are probably not optimal.

The study is split into the following steps.

1. We establish the fundamental estimate which relates the error $(x - x_h)$ to the so-called interpolation error $||(I - P_h)x||$ (section 3.3.1, Lemma 3.1):

$$||(I - A)(x - x_h)|| \leq 2||(I - P_h)x|| \ ,$$

where P_h denotes the orthogonal projector on the discretization space V_h.

2. Using a duality technique, we prove an estimate using negative Sobolev norms on Γ, here supposed to be C^∞ (section 3.3.2, Theorem 3.2) for $s > 1/2$

$$||x - x_h||_{H^{-s}(\Gamma)} \leq 2||(I - P_h)x|| \sup_{\psi \in H^s(\Gamma)} \frac{||(I - P_h)y(\psi)||}{||\psi||_{H^s(\Gamma)}} ,$$

where $y(\psi)$ derives from a solution of the homogeneous Helmholtz.

3. Assuming $|t| \leq \delta < 1$, we obtain bounds of the error on the boundary (Lemma 3.3, section 3.3.3) in the $L^2(\Omega)$ norm

$$||x - x_h||_{L^2(\Gamma)} \leq \frac{2}{\sqrt{1 - \delta^2}}||(I - P_h)x|| .$$

4. We estimate the error on $u - u_h$ at the boundary from the error on $x - x_h$ (Lemma 3.4, section 3.3.4) in any Sobolev norm, whatever real positive s:

$$||u - u_h||_{H^{-s}(\Gamma)} \leq \frac{1 + \delta}{2\omega}||x - x_h||_{H^{-s}(\Gamma)} .$$

5. In the particular case when x corresponds to a homogeneous Helmholtz problem ($f = 0$ in (0.1)) with u of class $C^{[(p+1)]/2}$, we estimate the order of convergence of the interpolation term $||(I - P_h)x||_V$ (section 3.3.5, Theorem 3.7). Using p basis functions per element constructed from given plane waves on a regular mesh, we have

$$||(I - P_h)x||_V \leq Ch^{[(p-1)]/2-1/2}||u||_{C^{[(p+1)]/2}(\Omega)} .$$

6. Using Theorem 3.7 we obtain the final results concerning the estimates of the orders of convergence on boundary norms (section 3.3.6). In the homogeneous case specifically, we have estimates in the $L^2(\Gamma)$ norm as stated in Corollary 3.8 using Lemma 3.3. A general statement (no assumption on f) completes this estimate with Corollary 3.9 that is based on Theorem 3.2. This estimate concerns a negative Sobolev norm and is thus of higher order than the $L^2(\Gamma)$ one. Lemma 3.4 states that the estimates are at least of the same order on $u - u_h$ and on $x - x_h$.

This study applies only to norms on the boundary. No inner estimate is being done. This could be achieved in a further study using an appropriate duality technique. Remark 19 explains why estimates on the boundary are already satisfying for us.

3.3.1. Estimate of the residue.

LEMMA 3.1. *Let us consider $x \in V$ the solution of* (1.23)*, section* 1.3*, and $x_h \in V_h$ the solution of* (2.1)*. Let P_h denote the orthogonal projector onto V_h. We have*

$$(I - A)(x - x_h) \in V_h^\perp , \tag{3.6}$$

$$\boxed{||(I - A)(x - x_h)|| \leq 2||(I - P_h)x||} . \tag{3.7}$$

Proof. Applying P_h to the equalities

$$\begin{aligned} &(I - A)x = b , \\ &(I - P_h A)x_h = b_h = P_h b \end{aligned} \tag{3.8}$$

we obtain

$$P_h(x - x_h) = P_h A(x - x_h) , \tag{3.9}$$

which is equivalent to (3.6). For the proof of (3.7), the reader may refer to [13]. Since this proof is important to make this paper self-contained, we give here a variant of it. There is no fundamental difference between the two proofs. Using equation (3.9), we have

$$\begin{aligned} x - x_h &= x - P_h x + P_h x - x_h \\ &= (I - P_h)x + P_h A(x - x_h) \\ &= (I - P_h)x - (I - P_h)A(x - x_h) + A(x - x_h) \ ; \end{aligned} \tag{3.10}$$

hence,

$$(I - A)(x - x_h) = (I - P_h)x - (I - P_h)A(x - x_h) \ . \tag{3.11}$$

The proof now consists of estimating the right-hand side in the above equation (3.11). Since P_h is an orthogonal projector, we have

$$||(I - P_h)A(x - x_h)||^2 = ||A(x - x_h)||^2 - ||P_h A(x - x_h)||^2 \ . \tag{3.12}$$

Plugging (3.9) in the right-hand side of (3.12) gives

$$||(I - P_h)A(x - x_h)||^2 = ||A(x - x_h)||^2 - ||P_h(x - x_h)||^2 \ , \tag{3.13}$$

and using $||A|| \leq 1$

$$||(I - P_h)A(x - x_h)||^2 \leq ||(x - x_h)||^2 - ||P_h(x - x_h)||^2 \ . \tag{3.14}$$

Since P_h is an orthogonal projector, we obtain

$$||(I - P_h)A(x - x_h)||^2 \leq ||(I - P_h)(x - x_h)||^2 \ , \tag{3.15}$$

and as $(I - P_h)x_h = 0$, we have

$$||(I - P_h)A(x - x_h)||^2 \leq ||(I - P_h)x||^2 \ . \tag{3.16}$$

So, using equations (3.11) and (3.16), we finally write

$$||(I - A)(x - x_h)|| \leq 2||(I - P_h)x|| \ . \qquad \square \tag{3.17}$$

3.3.2. A negative Sobolev norm estimate on the boundary. This estimate is achieved using a duality technique. It is the equivalent of the Aubin–Nitsche lemma. This will be useful in the study of the nonhomogeneous problem.

THEOREM 3.2. *Let t be a complex valued function on Γ with $|t| \leq 1$. Let us consider $x \in V$ the solution of* (1.23)*, section* 1.3*, and $x_h \in V_h$ the solution of* (2.1)*. Let P_h be the orthogonal projector onto space V_h. Let $s > 1/2$ and $\psi \in H^s(\Gamma)$. We assume that Γ is C^∞. Let w be defined by*

$$\begin{cases} (-\Delta - \omega^2)w = 0 & \text{in } \Omega \ , \\ (-\partial_\nu + i\omega)w = \overline{t}(+\partial_\nu + i\omega)w + \psi & \text{on } \Gamma \ . \end{cases} \tag{3.18}$$

and $y(\psi) \in V$ by

$$y(\psi)_{|\partial\Omega_k} = \big((-\partial_{\nu_k} + i\omega)w_{|\Omega_k}\big)_{|\partial\Omega_k} \ .$$

Then, for all $s > 1/2$,we have

$$\boxed{||x - x_h||_{H^{-s}(\Gamma)} \leq 2||(I - P_h)x|| \sup_{\psi \in H^s(\Gamma)} \frac{||(I - P_h)y(\psi)||}{||\psi||_{H^s(\Gamma)}} .} \tag{3.19}$$

Proof.

i) Let us show first that $y(\psi)$ is defined on V and $w \in C^m(\overline{\Omega})$ with $m = [s + 1/2]$.

• According to classical regularity results on elliptic problems (see Theorem 1.2, section 1 or see [17]), we know that if $g \in H^s(\Gamma)$ for $s > 1/2$ then $w \in H^{s+3/2}(\Omega)$. Trace theorems ensure the regularity at the boundary of $y(\psi)$.

• We use the Sobolev inequalities in the case of a two-dimensional space with square integrable functions (see [4, p. 169]). Then, for any $s = m - 1/2 + \epsilon$ with $1 > \epsilon > 0$, we have a continuous injection from $H^{s+3/2}(\Omega)$ to $C^m(\overline{\Omega})$. For $s > 1/2$ we have $m \geq 1$.

• From the two points above, we have $w \in C^m(\overline{\Omega})$ with $m \geq 1$. This implies that its gradient is continuous on $\overline{\Omega}$. Since Ω is bounded, $y(\psi) \in V$ for any $s > 1/2$.

ii) Let us check that $y(\psi)$ satisfies $(I - A^*)y(\psi) = (\psi)_{|\Gamma}$.

Let A^* stand for the adjoint of A (A is defined in (1.20)). By inspection, w satisfies

$$\begin{cases} (\partial_{\nu_j} + i\omega)w_j = (-\partial_{\nu_k} + i\omega)w_k & \text{on } \Sigma_{kj} , \\ (-\partial_{\nu_k} + i\omega)w_k = \overline{t}(+\partial_\nu + i\omega)w + \psi & \text{on } \Gamma_k ; \end{cases}$$

hence for all z in V we have

$$\sum_k \int_{\Gamma_k} \psi\overline{z} = \sum_k \int_{\partial\Omega_k} (-\partial_{\nu_k} + i\omega)w_k\overline{z} - \left(\sum_{kj} \int_{\Sigma_{kj}} (\partial_{\nu_j} + i\omega)w_j\overline{z} + \sum_k \int_{\Gamma_k} (+\partial_{\nu_k} + i\omega)w_k\overline{tz} \right) .$$

In other words

$$\forall z \in V,\ (y(\psi), z)_V - (Fy(\psi), \Pi z)_V = (\psi, z)_{L^2(\Gamma)} . \tag{3.20}$$

iii) Let us prove the inequality $|(x - x_h, \psi)_\Gamma| \leq 2||(I - P_h)x||\ ||(I - P_h)y(\psi)||$.

According to (3.20) we have

$$\begin{aligned} (x - x_h, \psi)_\Gamma &= (x - x_h, (I - A^*)y(\psi))_V \\ &= ((I - A)(x - x_h), y(\psi))_V \\ &= ((I - A)(x - x_h), (I - P_h)y(\psi))_V \end{aligned} \tag{3.21}$$

since $P_h(I - A)(x - x_h) = 0$ (see Lemma 3.1). Finally, from (3.21) and Lemma 3.1, we have

$$|(x - x_h, \psi)_\Gamma| \leq 2||(I - P_h)x||\ ||(I - P_h)y(\psi)|| . \tag{3.22}$$

iv) Let us conclude using (3.22) and

$$||x - x_h||_{H^{-s}(\Gamma)} = \sup_{\psi \in H^s(\Gamma)} \frac{(x - x_h, \psi)_{L^2(\Gamma)}}{||\psi||_{H^s(\Gamma)}} . \tag{3.23}$$

∎

Remark 18. The interest of (3.19) comes from the fact that $\sup_{\psi\in H^s(\Gamma)} ||(I-P_h)y(\psi)||/||\psi||_{H^s(\Gamma)}$ tends to zero with h as is stated in Theorem (3.7). We get a better estimate in this negative Sobolev norm $H^{-s}(\Gamma)$ than in the classical $L^2(\Gamma)$ norm.

Remark 19. This kind of upper bounding in a Sobolev space such as $H^{-s}(\Gamma)$ with $s > 1/2$ is still of a practical interest. The calculation of the amplitude of diffusion, denoted by a (θ), which is needed by engineers for radar cross section computations, is obtained using (3.24).

$$a(\theta) = \frac{1}{\sqrt{\omega}} \int_{\Gamma_a} e^{i\omega \vec{x}\vec{e}_\theta} (-i\omega \vec{\nu}\, \vec{e}_\theta u + \partial_\nu u) \ . \tag{3.24}$$

We consider $\Gamma_a = \Gamma_{int}$ the inner boundary of Ω and use the upper bounding (3.33) of Lemma 3.4. We estimate $|a(\theta) - a_h(\theta)|$ from $||x - x_h||_{H^{-s}(\Gamma)}$ using the Cauchy–Schwarz inequality.

3.3.3. An energy boundary error estimate. We study here the relationship between the residue's norm on the boundary $||x - x_h||_{L^2(\Gamma)}$, and the interpolation error $||(I - P_h)x||$. This is the equivalent of Cea's lemma.

LEMMA 3.3. *Let t (the boundary operator of the Helmholtz problem* (0.1)*) be constant and $|t| \le \delta < 1$. Let us consider $x \in V$ the solution of* (1.23)*, section* 1.3*, and $x_h \in V_h$ the solution of* (2.1)*. Let P_h be the orthogonal projector onto the space V_h. We have* (3.25)*.*

$$\boxed{||x - x_h||_{L^2(\Gamma)} \le \frac{2}{\sqrt{1-\delta^2}} ||(I - P_h)x||_V \ .} \tag{3.25}$$

Proof. Let $\varepsilon_h = (x - x_h)$ and $e_h = u - u_h$. Using the definition of A, the Cauchy–Schwarz inequality, and the fact that F is an isometry, we have

$$\begin{aligned} ((I-A)\varepsilon_h, \varepsilon_h)_V &= ||\varepsilon_h||^2 - (\Pi\varepsilon_h, F\varepsilon_h)_V \\ &\ge ||\varepsilon_h||^2 - ||\Pi\varepsilon_h||\ ||F\varepsilon_h|| \\ &\ge ||\varepsilon_h||^2 \left(1 - \frac{||\Pi\varepsilon_h||}{||\varepsilon_h||}\right) \ . \end{aligned} \tag{3.26}$$

By definition of Π

$$\begin{aligned} ||\Pi\varepsilon_h||^2 &= + \sum_k \int_{\Gamma_k} |t|^2\ |(-\partial_{\nu_k} + i\omega)e_h|^2 + \sum_{kj} \int_{\Sigma_{kj}} |(-\partial_{\nu_k} + i\omega)e_h|^2 \\ &= + \sum_k \int_{\Gamma_k} (|t|^2 - 1)\ |(-\partial_{\nu_k} + i\omega)e_h|^2 + \sum_k \int_{\partial\Omega_k} |(-\partial_{\nu_k} + i\omega)e_h|^2 \ , \end{aligned} \tag{3.27}$$

that is, defining $||\varepsilon_h||_\Gamma^2 = \int_\Gamma |\varepsilon_h|^2$, we obtain $||\Pi\varepsilon_h||^2 \le ||\varepsilon_h||^2 - (1-|\delta|^2)||\varepsilon_h||_\Gamma^2$. Hence,

$$||\Pi\varepsilon_h|| \le ||\varepsilon_h|| \left(1 - \frac{1-|\delta|^2}{2} \frac{||\varepsilon_h||_\Gamma^2}{||\varepsilon_h||^2}\right) \ . \tag{3.28}$$

From inequalities (3.28) and (3.26) we have

$$|((I-A)\varepsilon_h, \varepsilon_h)_V| \ge ||\varepsilon_h||^2 \left[1 - (1 - \frac{1-|\delta|^2}{2} \frac{||\varepsilon_h||_\Gamma^2}{||\varepsilon_h||^2})\right] \ , \tag{3.29}$$

which yields the upper bounding on $||\varepsilon_h||_\Gamma$

$$|((I-A)\varepsilon_h,\varepsilon_h)_V| \geq \frac{1-|\delta|^2}{2}||\varepsilon_h||_\Gamma^2 \ . \tag{3.30}$$

According to Lemma 3.1 and the Cauchy–Schwarz inequality, we can write

$$|((I-A)\varepsilon_h,\varepsilon_h)|_V = |((I-A)\varepsilon_h,(I-P_h)x)|_V \leq 2||(I-P_h)x||^2 \ . \tag{3.31}$$

Thus, finally from (3.30) and (3.31), we get

$$||(I-P_h)x|| \geq \frac{\sqrt{1-\delta^2}}{2}||\varepsilon_h||_\Gamma \ . \qquad \square \tag{3.32}$$

3.3.4. Boundary error estimates on $u-u_h$.

LEMMA 3.4. *Let t (the boundary operator of the Helmholtz equation* (0.1)*) be constant and $|t| \leq \delta < 1$. Let us consider $x \in V$ the solution of* (1.23)*, section* 1.3*, and $x_h \in V_h$ the solution of* (2.1)*, u the solution of* (0.1)*, and u_h defined by* (2.16)*. We have for any real positive s* (3.33)*.*

$$\boxed{||u-u_h||_{H^{-s}(\Gamma)} \leq \frac{1+\delta}{2\omega}||x-x_h||_{H^{-s}(\Gamma)} \ .} \tag{3.33}$$

Proof. Recall the relations (2.16) and (2.15), section 2.2.1

$$\begin{cases} u = \dfrac{1}{2i\omega}[(I+\Pi)x+g] & \text{on } \Gamma_k \ , \\ u_h = \dfrac{1}{2i\omega}[(I+\Pi)x_h+g] & \text{on } \Gamma \ . \end{cases}$$

Thus, for all $s \in \mathbb{R}^+$, we have $||u-u_h||_{H^s(\Gamma)} = 1/2\omega||(I+\Pi)(x-x_h)||_{H^s(\Gamma)}$, hence, using the definition of Π and the hypothesis $|t| \leq \delta$, we have $||u-u_h||_{H^s(\Gamma)} \leq 1+\delta/2\omega||x-x_h||_{H^s(\Gamma)}$. $\square$

Remark 20. We have observed numerically that the upper bounding (3.33) of Lemma 3.4 that uses the approximation (2.20) for u_h is not optimal in the homogeneous problem. A sharper analysis should give a gain of around a half on the order of the error on u at the boundary Γ with respect to the error on x. Refer to Figure 3.3, section 3.2.3.

COROLLARY 3.5. *Assuming the same hypothesis as in Lemma* (3.3)*, we derive from* (3.33) *and* (3.25) *the following bound on the $L^2(\Gamma)$ norm of the error on u with respect to the interpolation error $||(I-P_h)x||_V$*

$$||u-u_h||_{L^2(\Gamma)} \leq \frac{1}{\omega}\sqrt{\frac{1+\delta}{1-\delta}}||(I-P_h)x||_V \ . \tag{3.34}$$

COROLLARY 3.6. *Assuming the same hypothesis as in Theorem* (3.2)*, we derive from* (3.33) *and* (3.19) *the following bound on the negative Sobolev norm $H^{-s}(\Gamma)$ for any $s > 1/2$ of the error on u*

$$||u-u_h||_{H^{-s}(\Gamma)} \leq \frac{1+\delta}{\omega}||(I-P_h)x|| \sup_{\psi \in H^s(\Gamma)} \frac{||(I-P_h)y(\psi)||}{||\psi||_{H^s(\Gamma)}} \ . \tag{3.35}$$

3.3.5. Study of the interpolation error. We have $\forall x_a \in V_h$, $||(I-P_h)x|| \leq ||x-x_a||_V$. The key to the upper bounding is finding a particular x_a for which we have a "good" estimate.

THEOREM 3.7. *Let u be a solution of a homogeneous Helmholtz problem. We assume that u is of class C^{n+1} with $n \geq 1$. Let $x \in V$ satisfy*

$$x_{|\partial\Omega_k} = (-\partial\nu_k + i\omega)u_{|\partial\Omega_k} \ .$$

We assume the mesh $(\Omega_k)_{k=1,\ldots,K}$ satisfies the hypotheses of uniform regularity (H1, H2, *and* H3 *of section* 2.1). *The space of approximation V_h is constructed by $p = 2n+1$ functions z_{kl} by element Ω_k such that $z_{kl} = (-\partial_{\nu_k} + i\omega)e_{kl}$ and $(e_{kl})_{l=1\ldots p}$ being an independent family of plane waves. Let us assume, for the sake of the simplicity of the proof, that their directions are fixed once for all. Then*

$$\begin{cases} \exists C > 0 \text{ depending on } n \text{ and the problem's data } (0.1) \\ ||(I-P_h)x||_V \leq Ch^{n-1/2}||u||_{C^{n+1}(\Omega)} \ . \end{cases} \tag{3.36}$$

Proof.

a) *Step* 1. Let us suppose the existence of a function u_a in the vector space spanned by the plane waves e_{kl} such that, on each element Ω_k, u_a approaches u at the order n. We shall leave the k subscript of the element Ω_k. To be precise, if we denote by $\vec{x} = (x_1, x_2)$ a point of Ω_k in a coordinate system centered on the barycenter of this element, we assume that there exists the following estimates:

$$|u(\vec{x}) - u_a(\vec{x})| \leq C_1 h^{n+1}||u||_{C^{n+1}(\Omega)} \tag{3.37}$$

and

$$|\nabla u(\vec{x}) - \nabla u_a(\vec{x}))| \leq C_1 h^n||u||_{C^{n+1}(\Omega)} \tag{3.38}$$

where C_1 depends on ω and Ω only (not on h or u) with u_a satisfying

$$u_a = \sum_{l=1}^{p} x_l^n e_l \ . \tag{3.39}$$

Since u_a is assumed to be at least $C^{n+1}(\Omega)$ we can define x_a by

$$x_a = (-\partial_\nu + i\omega)u_a \ . \tag{3.40}$$

By definition of $||x-x_a||_{L^2(\partial\Omega_k)}$, we have

$$\begin{aligned} ||x-x_a||^2_{L^2(\partial\Omega_k)} &= \int_{\partial\Omega_k} |(-\partial_{\nu_k} + i\omega)(u-u_a)|^2 \\ &\leq 2\int_{\partial\Omega_k} \left\{ |\left(\nabla u(\vec{x}) - \nabla u_a(\vec{x})\right).\vec{\nu}|^2 + \omega^2|u(\vec{x}) - u_a(\vec{x})|^2 \right\} \\ &\leq 2C_1^2 h^{2n} \left(\int_{\partial\Omega_k} (1+\omega^2h^2) \right) ||u||^2_{C^{n+1}(\Omega)} \ . \end{aligned}$$

By definition, for h small enough, the integral on $\partial\Omega_k$ can be bounded above as follows ($C' > 0$):

$$\int_{\partial\Omega_k} \left(1+\omega^2h^2\right) \leq C'h \ .$$

Therefore, letting $C^2 = 2C_1^2 C^{'} \omega^2$, we have

$$||x - x_a||_{L^2(\partial\Omega_k)} \leq Ch^{n+1/2}||u||_{C^{n+1}(\Omega)} \ . \tag{3.41}$$

Knowing that, for a regular mesh of size parameter h, the total number of elements is majorized by C/h^2, we deduce

$$||x - x_a||_V^2 \leq Ch^{n-1/2}||u||_{C^{n+1}(\Omega)} \ . \tag{3.42}$$

The constant C in (3.42) does not depend on k since the directions of the plane waves are fixed.

b) *Step* 2. We look for the existence of u_a satisfying the condition (3.37), omitting the index k to simplify the notation.

Since u is of class C^{n+1}, the Taylor polynomial of u (at a point of Ω_k denoted by (x_1, x_2) in a coordinate system centered on the barycenter of the considered element) at the order n exists in the sense that

$$\left| u(\vec{x}) - \sum_{m=0}^{n} \sum_{q_1+q_2=m}^{q_1\geq 0, q_2 \geq 0} B_{q_1,q_2} x_1^{q_1} x_2^{q_2} \right| \leq Ch^{n+1}||u||_{C^{n+1}(\Omega)} \ . \tag{3.43}$$

The basis functions e_l are C^∞ in the interior of Ω_k (and even analytic). Let e_l^n be the truncated Taylor expansion at order n of e_l, i.e., $|e_l - e_l^n| \leq C_{l,n} h^{n+1}||e_l||_{C^{n+1}(\Omega)}$. Hence, we write

$$e_l^n(\vec{x}) = \sum_{m=0}^{n} \sum_{q_1+q_2=m}^{q_1\geq 0, q_2 \geq 0} M_{q_1,q_2}^l x_1^{q_1} x_2^{q_2} \ . \tag{3.44}$$

It is easy to prove that we have (using the Taylor Lagrange extension of a $C^{n+1}(\Omega)$ function and using the basic definition of the derivative as $lim_{h\to 0} e(x+h) - e(x)/h$)

$$|\Delta e_l(\vec{x}) - \Delta e_l^n(\vec{x})| \leq C_{l,n} h^{n-1}||e_l||_{C^{n+1}(\Omega)} \ . \tag{3.45}$$

The condition of existence of u_a such that (3.37) holds is equivalent to the existence of p complex coefficients x_l^n such that

$$\begin{cases} B_{q_1,q_2} - \displaystyle\sum_{l=1}^{p} x_l^n M_{q_1,q_2}^l = 0 \ , \\ \\ \forall (q_1, q_2) \in \mathbb{N}^2, \ q_1 \geq 0, \ q_2 \geq 0, \ 0 \leq q_2 + q_1 \leq n \ . \end{cases} \tag{3.46}$$

This system of linear equations (3.47) can be gathered into a matrix problem

$$\begin{cases} \text{Find } X^n \in \mathbb{C}^{2n+1} \text{ such that} \\ M_n X^n = B_n \ . \end{cases} \tag{3.47}$$

The matrix M_n has $p = 2n + 1$ columns and $(n + 2)(n + 1)/2$ rows, defining a linear

function from $\mathbb{C}^{2n+1}$ to $\mathbb{C}(n+1)(n+2)/2$. Recalling $p = 2n+1$, we have

$$M_n = \begin{bmatrix} M^1_{0,0} & M^2_{0,0} & \dots & M^p_{0,0} \\ M^1_{1,0} & M^2_{1,0} & \dots & M^p_{1,0} \\ M^1_{0,1} & M^2_{0,1} & \dots & M^p_{0,1} \\ M^1_{2,0} & M^2_{2,0} & \dots & M^p_{2,0} \\ M^1_{1,1} & M^2_{1,1} & \dots & M^p_{1,1} \\ M^1_{0,2} & M^2_{0,2} & \dots & M^p_{0,2} \\ \vdots & \vdots & \vdots & \vdots \\ M^1_{n,0} & M^2_{n,0} & \dots & M^p_{n,0} \\ \vdots & \vdots & \vdots & \vdots \\ M^1_{0,n} & M^2_{0,n} & \dots & M^p_{0,n} \end{bmatrix} \text{ and } B_n = \begin{bmatrix} B_{0,0} \\ B_{1,0} \\ B_{0,1} \\ B_{2,0} \\ B_{1,1} \\ B_{0,2} \\ \vdots \\ B_{n,0} \\ \vdots \\ B_{0,n} \end{bmatrix},$$

$$^{\top}X^n = \left[\; x_1^n \; , \; x_2^n \; , \; \dots \; , \; x_p^n \; \right] .$$

The matrix M_n depends only on the plane waves directions. The second member B_n depends only on the problem's data (0.1). We shall prove this system admits a solution.

i) Let us prove that the image of M_n is included in a subspace K of dimension $2n+1$. We also prove that the right-hand side in the above linear system is an element of K. This is summarized by

$$\begin{cases} \mathrm{Im}(M_n) \subset K \; , \\ B_n \in K \; , \\ \dim(K) = 2n+1 \; . \end{cases}$$

Let us calculate Δe_l^n, through first developing $\partial^2 e_l^n / \partial^2 x_1$:

$$\begin{aligned} \frac{\partial^2 e_l^n}{\partial^2 x_1}(\vec{x}) &= \frac{\partial^2}{\partial^2 x_1} \sum_{m=0}^{n} \sum_{q_1+q_2=m}^{q_1 \geq 0, q_2 \geq 0} M^l_{q_1,q_2} x_1^{q_1} x_2^{q_2} \\ &= \sum_{m=0}^{n} \sum_{q_1+q_2=m}^{q_1 \geq 0, q_2 \geq 0} M^l_{q_1,q_2} q_1(q_1-1) x_1^{q_1-2} x_2^{q_2} \\ &= \sum_{m=0}^{n} \sum_{q_1+q_2=m}^{q_1 \geq 0, q_2 \geq 0} M^l_{q_1+2,q_2} (q_1+1)(q_1+2) x_1^{q_1} x_2^{q_2} \; . \end{aligned} \tag{3.48}$$

Using $-(\Delta+\omega^2)e_l = 0$, relation (3.45), and replacing $\Delta e_l = \partial^2 e_l/\partial^2 x_1 + \partial^2 e_l/\partial^2 x_2$ by its expansion (3.48) twice, once substituting x_1 by x_2, we obtain the relations (3.49) between the coefficients of the Taylor polynomial of e_l:

$$\begin{cases} \forall l \in \{1, \dots, p\} \\ \forall (q_1, q_2) \in \mathbb{N}^2, \; q_1 \geq 0, \; q_2 \geq 0, \; 0 \leq q_1 + q_2 \leq n-2 \; , \\ (q_1+1)(q_1+2) M^l_{q_1+2,q_2} + (q_2+1)(q_2+2) M^l_{q_1,q_2+2} = \omega^2 M^l_{q_1,q_2} \; . \end{cases} \tag{3.49}$$

Since u also satisfies the homogeneous Helmholtz equations, we have

$$\begin{cases} \forall (q_1,q_2) \in \mathbb{N}^2,\ q_1 \geq 0,\ q_2 \geq 0,\ 0 \leq q_1+q_2 \leq n-2\ , \\ (q_1+1)(q_1+2)B_{q_1+2,q_2} + (q_2+1)(q_2+2)B_{q_1,q_2+2} = \omega^2 B_{q_1,q_2}\ . \end{cases} \tag{3.50}$$

Let K be the vector space defined by

$$\begin{aligned} K = \{ (B_{q_1,q_2}) \in \mathbb{C}^{\frac{n(n+1)}{2}}, & q_1 \geq 0,\ q_2 \geq 0,\ q_1+q_2 \leq n \text{ with} \\ & \forall (q_1,q_2) \in \mathbb{N}^2,\ q_1 \geq 0,\ q_2 \geq 0,\ 0 \leq q_1+q_2 \leq n-2 \\ & (q_1+1)(q_1+2)B_{q_1+2,q_2} + (q_2+1)(q_2+2)B_{q_1,q_2+2} = \omega^2 B_{q_1,q_2} \} \ . \end{aligned} \tag{3.51}$$

It is clear that $B_n \in K$ and $\mathrm{Im}(M_n) \subset K$. The column vectors of M_n belong to K, thus the vector space spanned by the columns of M_n is included in K. Let us prove that $\dim(K) \leq 2n+1$. The vector space K is defined in the form of $n(n-1)/2$ relations that we write in the form of

$$N(B_{q_1,q_2}) = 0\ ,$$

with N a $(n(n-1)/2, n(n+1)/2)$ matrix whose first $n(n-1)/2$ columns are

$$\begin{bmatrix} -\omega^2 & 0 & 0 & 2 & 0 & 2 & 0 & 0 & 0 & 0 & \dots \\ 0 & -\omega^2 & 0 & 0 & 0 & 0 & 6 & 0 & 6 & 0 & \dots \\ 0 & 0 & -\omega^2 & 0 & 0 & 0 & 0 & 6 & 0 & 6 & \dots \\ 0 & 0 & 0 & -\omega^2 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & -\omega^2 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & -\omega^2 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & -\omega^2 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\omega^2 & 0 & 0 & \dots \\ \vdots & & & & & & & & \ddots & & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\omega^2 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\omega^2 \end{bmatrix}$$

forming an upper triangular square matrix that is obviously invertible. The $n(n-1)/2$ relations that defined K are independent. This means that the dimension of K is lower than $n(n+1)/2 - n(n-1)/2$, that is $2n+1$.

ii) Let us prove that M_n has a rank greater than $2n+1$.

The previous part of the proof used only the fact that the functions e_l were exact solutions of the homogeneous Helmholtz equation. This part uses the exact form of the functions and requires, for technical reasons, that the basis functions are plane waves. Here we need to extract from the matrix M_n or from a linear transform of M_n a square invertible submatrix of rank $2n+1$.

Let G stand for the barycenter of the element Ω_k. The matrix M_n is defined by the coefficients $M^l_{q,s} = (i\omega)^{q+s}/q!s!\partial^{q+s}/\partial^q_{x_1}\partial^s_{x_2} e_l(G)$. Let S_n be the matrix defined by $S_{q,l} = (\partial_{x_1} + i\partial_{x_2})^q e_l(G)$ for $q=0$ to $q=n$ and $S_{n+q,l} = (\partial_{x_1} - i\partial_{x_2})^q e_l(G)$ for $q=1$ to $q=n$. This matrix is a linear transform of M_n since the mixed derivative operators used to define it are linear combinations of the derivative operators that

give the matrix M_n. More precisely,

$$q \leq n \Rightarrow \begin{cases} S_{q,l} = \dfrac{q!}{(i\omega)^q} \displaystyle\sum_{s=0}^{q} i^s M^l_{s-q,s} \ , \\ \\ S_{n+q,l} = \dfrac{q!}{(i\omega)^q} \displaystyle\sum_{s=0}^{q} (-i)^s M^l_{s-q,s} \ . \end{cases} \tag{3.52}$$

The above relations show that there exists a $(n(n+1)/2 \times 2n+1)$ matrix P_n that transforms M_n to S_n: $S_n = P_n \ M_n$. The transform matrix P_n depends on ω and n only. It does not depend on u. The coefficients of P_n are the $q!/(i\omega)^q i^s$ and $q!/(i\omega)^q(-i)^s$ in the right places. The matrix S_n is a square matrix of size $2n+1$. For e_l a plane wave, that is

$$e_l = e^{i\omega(u_l x_1 + v_l x_2)}, \tag{3.53}$$

we have

$$\begin{cases} (\partial_x + i\partial_y)^q e_l = (u_l + iv_l)^q \, e_l = z_l^q \ , \\ (\partial_x - i\partial_y)^q e_l = (u_l - iv_l)^q \, e_l = z_l^{-q} \end{cases} \tag{3.54}$$

recalling that $(u_l - iv_l) = 1/(u_l + iv_l)$ since $u_l^2 + v_l^2 = 1$ (e_l is a solution of the homogeneous Helmholtz problem (0.1)) and that $e_l = 1$ at the barycenter. The constructed matrix S_n is thus

$$S_n = \begin{bmatrix} 1 & \dots & 1 \\ z_1 & \dots & z_p \\ z_1^2 & \dots & z_p^2 \\ z_1^3 & \dots & z_p^3 \\ \vdots & \dots & \vdots \\ z_1^n & \dots & z_p^n \\ z_1^{-1} & \dots & z_p^{-1} \\ z_1^{-2} & \dots & z_p^{-2} \\ z_1^{-3} & \dots & z_p^{-3} \\ \vdots & \dots & \vdots \\ z_1^{-n} & \dots & z_p^{-n} \end{bmatrix} . \tag{3.55}$$

The matrix S_q^n has a Vandermonde determinant. Its value is

$$\prod_{i=1}^{n} z_i^{-n} \prod_{i<j} (z_i - z_j) \ . \tag{3.56}$$

The product as above (3.56) is not zero if and only if $\forall (i,j) z_i \neq z_j$. To recapitulate, S_q^n is a square matrix of dimensions $2n+1$ rows, $p = 2n+1$ columns and is invertible. We have thus $rank(M_n) \geq rank(S_q^n) = 2n+1$.

iii) Let us recapitulate the previous points i) and ii). Point i) shows the inclusion of the image of M_n in K a vector space of dimension at most $2n+1$. Point ii) teaches

us that the dimension of $\mathrm{Im}(M_n)$ is at least $2n+1$. So $\mathrm{Im}(M_n) = K$. Furthermore, the second member B_n is an element of K. We may sum up

$$\left.\begin{array}{l} K = \text{ Im } (M_n) \\ B_n \in K \end{array}\right\} \Rightarrow B_n \in \text{ Im } (M_n) \ .$$

This proves that system (3.47) has a unique solution and that u_a (defined by (3.39)) exists. From the Taylor approximation of $u - u_a$, we have

$$\begin{aligned} |u - u_a| \leq & Ch^{n+1} \sup ||u - u_a||_{C^{n+1}(\Omega)} \\ \leq & Ch^{n+1} \left(||u||_{C^{n+1}(\Omega)} + ||u_a||_{C^{n+1}(\Omega)} \right) \ , \end{aligned} \tag{3.57}$$

and using (3.39)

$$||u_a||_{C^{n+1}(\Omega)} \leq \sum_{l=1}^{p} |x_l^n| ||e_l||_{C^{n+1}(\Omega)} \ . \tag{3.58}$$

From $S_n X_n = P_n M_n X_n = P_n B_n$ and S_n is a given matrix, invertible, and of finite dimension, we obtain

$$X^n = (S_n)^{-1} P_n B_n \ .$$

Since M_n depends only on the given basis functions and thereby has its norm uniformly bounded above by $\sup_{l=1}^{p} ||e_l||_{C^{n+1}(\Omega)}$, and since the B_n coefficients are bounded above by the $C^{n+1}(\Omega)$ norm of u, we state there exists a positive constant C_2 such that for any $l = 1$ to $p = 2n+1$ we have

$$|x_l^n| \leq C_2 ||u||_{C^{n+1}(\Omega)} \ . \tag{3.59}$$

Using (3.58) and (3.59) in (3.57), we have, letting $C_1 = C(pC_2 + 1)$,

$$|u - u_a| \leq C_1 h^{n+1} ||u||_{C^{n+1}(\Omega)} \ , \tag{3.60}$$

so we satisfy the hypothesis (3.37) of Step 1. Now, let us check that we satisfy hypothesis (3.38). From the Taylor approximation of $u - u_a$, we have

$$|\nabla u(\vec{x}) - \nabla u_a(\vec{x}))| \leq Ch^n ||u - u_a||_{C^{n+1}(\Omega)} + \left| \nabla u_n(\vec{x}) - \sum_{l=1}^{2n+1} x_l^n \nabla e_l^n(\vec{x})) \right| \tag{3.61}$$

where u_n stands for the Taylor polynomial at order n of u as in (3.43). By construction,

$$u_n(\vec{x}) - \sum_{l=1}^{2n+1} x_l^n e_l^n(\vec{x})) = 0$$

so

$$\nabla u_n(\vec{x}) - \sum_{l=1}^{2n+1} x_l^n \nabla e_l^n(\vec{x})) = 0.$$

So we have, as in (3.60),

$$|\nabla u(\vec{x}) - \nabla u_a(\vec{x}))| \leq C_3 h^n ||u||_{C^{n+1}(\Omega)} \ , \tag{3.62}$$

which is hypothesis (3.38) of Step 1. □

Remark 21. We can calculate the determinant of M_n for three basis functions. This determinant is maximum for equidistributed basis functions (see [6] or better [5]).

Remark 22. The proof of Theorem (3.7) demands the use of plane waves to extract a free $(2n+1)$ submatrix from M_n. It should be possible to generalize this result to other free sets of given homogeneous Helmholtz equations. For instance, in a 3-dimensional space, spherical wave functions like

$$\frac{e^{i\omega kr}}{kr} \tag{3.63}$$

are $C^{n+1}(\Omega)$ for $r = |\vec{x}| \neq 0$. We believe it is possible to extend Theorem (3.7) to such basis functions.

Remark 23. This result could be established in a Sobolev space directly using the integral remainder of the Taylor extension.

Remark 24. In the proof of Theorem 3.7, the fact that u is solution of a homogeneous problem is essential as soon as $n \geq 2$ since the basis functions satisfy the homogeneous Helmholtz equation. Nonetheless, in the case where $n = 1$, the result remains true for a nonhomogeneous problem. We have $p = 2n+1 = 3 = (n+2)(n+1)/2$ basis functions satisfying the hypotheses of Theorem 3.7. The vector space K (see (3.51)) is 3-dimensional since $n(n-1)/2 = 0$. More simply, for $p = 3$ and thus $n = 1$, there is no use of the fact that u and the basis functions satisfy any Helmholtz equation. This is why, for all $p \geq 3$, x defined by (0.8) with u the solution of (0.1), we have

$$||(I - P_h)x||_V \leq Ch^{1/2} ||u||_{C^2(\Omega)} \ . \tag{3.64}$$

3.3.6. Orders of convergence estimates. Here are our main theoretical results about the order of convergence. All these results are mere consequences of Theorem 3.7.

According to Theorem 3.7 and Lemma 3.3, we have Corollary 3.8.

COROLLARY 3.8. *Let the basis functions satisfy the hypotheses of Theorem* 3.7. *Let u be a solution of* (0.1) *homogeneous ($f = 0$) and t be constant, (x defined by* (0.8)*) and x_h be a solution of* (2.21) *with $p \geq 3$ basis functions per element. Let $[\alpha]$ denote the integral part of α. We assume u is of class $C^{[(p+1)/2]}(\Omega)$; then*

$$\begin{cases} ||x - x_h||_{L^2(\Gamma)} \leq Ch^{[(p-1)/2]-1/2} ||u||_{C^{[(p+1)/2]}(\Omega)} \ , \\ ||u - u_h||_{L^2(\Gamma)} \leq C' h^{[(p-1)/2]-1/2} ||u||_{C^{[(p+1)/2]}(\Omega)} \ . \end{cases} \tag{3.65}$$

For instance $p = 3$ or $p = 4$ gives $h^{1/2}$. Using (3.33)*, we have the same law of convergence on $||u - u_h||_{L^2(\Gamma)}$.*

COROLLARY 3.9. *Let the p basis functions be plane waves satisfying the hypotheses of Theorem* 3.7. *Let u be a solution of* (0.1) *nonhomogeneous (f is not necessarily identically zero on $L^2(\Omega)$) and t be constant, (x defined by* (0.8)*) and x_h be a solution*

of (2.21) *with* $p \geq 3$ *basis functions per element. Let* $[\alpha]$ *denote the integral part of* α. *We assume* u *is of class* $C^2(\Omega)$ *and* Γ *is* C^∞*; then*

$$\forall s > [(p-1)/2] - 1/2 \begin{cases} ||x - x_h||_{H^{-s}(\Gamma)} \leq Ch^{[(p-1)/2]}||u||_{C^2(\Omega)} \ , \\ ||u - u_h||_{H^{-s}(\Gamma)} \leq C^{'} h^{[(p-1)/2]}||u||_{C^2(\Omega)} \ . \end{cases} \tag{3.66}$$

For instance $p = 3$ *or* $p = 4$ *gives* h^1. *Using* (3.33), *we have the same law of convergence on* $||u - u_h||_{H^{-s}(\Gamma)}$.

Proof. Recall the upper bounding obtained using a duality technique (3.19), section 3.3.2, Theorem 3.2. It is stated that, $\forall s > 1/2$, we have

$$||x - x_h||_{H^{-s}(\Gamma)} \leq 2||(I - P_h)x||_V \sup_{||\psi||_{H^s(\Gamma)}=1} ||(I - P_h)y(\psi)||_V \ . \tag{3.67}$$

For $p \geq 3$ and $u \in C^1(\Omega)$, we use the upper bounding (3.64) of Remark 24:

$$||(I - P_h)x||_V \leq Ch^{1/2}||u||_{C^2(\Omega)} \ . \tag{3.68}$$

Furthermore, w (see (3.18)) the solution of the homogeneous Helmholtz problem (see (0.1) with $f = 0$, $t = 0$ and $g = \psi$), is $C^{[s+1/2]}(\Omega)$ (see section i) of the proof of Theorem 3.2). Thus, for any $s > [(p-1)/2] - 1/2$, we can apply the upper bounding (3.65) of Corollary 3.8

$$\sup_{||\psi||_{H^s(\Gamma)}=1} ||(I - P_h)y(\psi)||_V \leq Ch^{[(p-1)/2]-1/2} \ . \tag{3.69}$$

Finally, from (3.69) and (3.68), we have the upper bounding (3.66). ∎

3.4. Study of the condition number. Let us define the order q of the condition number as the exponent of h the mesh size parameter supposing the condition number is bounded below by h^q. To complete the study of the numerical accuracy of the UWVF method, we prove in Theorem 3.10 the linear lower bound law of the order of the condition number as a function of the number of basis functions (section 3.4). We also present some numerical computations of the condition number.

3.4.1. Theoretical evolution of the condition number.

THEOREM 3.10. *Let the* p *basis functions* ($p \geq 4$) *satisfy the hypotheses of Theorem* 3.7. *We have a linear growth* (3.70) *of the order of the condition number of* D_k *as a function of* $[p/2]$. *Let* h_k *be the diameter of* Ω_k *and* $[\alpha]$ *denote the integral part of* α. *There exists* C *positive such that*

$$\frac{\lambda_{max}}{\lambda_{min}} \geq Ch_k^{-2[p/2]+2} \ . \tag{3.70}$$

For instance $p = 4$ *or* $p = 5$ *gives* h_k^{-2}.

Proof. For the proof, let us consider D_k is a square matrix of dimension p. Let λ_{min} be the smallest eigenvalue of D_k and λ_{max} be the biggest one. We have

$$\begin{cases} \forall Y \in \mathbb{C}^p \ , \\ \lambda_{min} \leq \dfrac{Y^\top D_k Y}{||Y||^2} \leq \lambda_{max} \ . \end{cases} \tag{3.71}$$

i) Estimate of the lowest eigenvalue. Recall the basis functions z_l derive from the plane waves e_l by $z_l = (-\partial_\nu + i\omega)e_l$. Let us consider Y so that $Y^\top = [X', -1]$ where X' is the $\mathbb{C}^{p-1}$ vector that approximates the function $(-\partial_\nu + i\omega)e_p$ using the $(p-1)$ functions $(z_1, \ldots, z_{p-1})$. We can apply Theorem 3.7 with $p-1$ basis functions and $u = e_p$ since the plane waves are solutions of the homogeneous Helmholtz problem. Let us assume $p - 1 = 2n + 1$. Inequality (3.41) of Theorem 3.7 states

$$Y^\top D_k Y = \left(||z_p - \sum_{l=1}^{p-1} x_l^{'} z_l||_{L^2(\partial\Omega_k)} \right)^2 \leq C(h_k^{n+1/2})^2 \ .$$

This implies, using $||Y||^2 = 1 + \sum_{l=1}^{p} x_l'^2 \geq 1$ and (3.71) that

$$\lambda_{min} \leq C h_k^{2n+1} \ .$$

ii) Estimate of the biggest eigenvalue. Let us consider Y so that $Y^\top = [0, -1]$. Replacing into equation (3.71), we obtain

$$\begin{aligned} \lambda_{max} \geq Y^\top D_k Y &\geq \omega^2 \sum_{i=1}^{3} L_i (1 - \vec{\nu}_i . \vec{e}_{p+1})^2 \\ &\geq \omega^2 \left[(L_1 + L_2 + L_3) - 2.\vec{e}_{p+1} \sum_{i=1}^{3} L_i \vec{\nu}_i \right] \\ &\geq \omega^2 (L_1 + L_2 + L_3) \geq \omega^2 \sup(L_1, L_2, L_3) \geq \omega^2 h_k \end{aligned}$$

since $L_1\vec{\nu}_1 + L_2\vec{\nu}_2 + L_3\vec{\nu}_3 = 0$ and the normals ν_n are $\pi/2$ rotations of $\vec{x}_{n+1} - \vec{x}_n / |\vec{x}_{n+1} - \vec{x}_n|$ as well as $L_1\vec{\nu}_1 + L_2\vec{\nu}_2 + L_3\vec{\nu}_3$ is a $\pi/2$ rotation of $(\vec{x}_2 - \vec{x}_1) + (\vec{x}_3 - \vec{x}_2) + (\vec{x}_1 - \vec{x}_3)$ which is the zero vector.

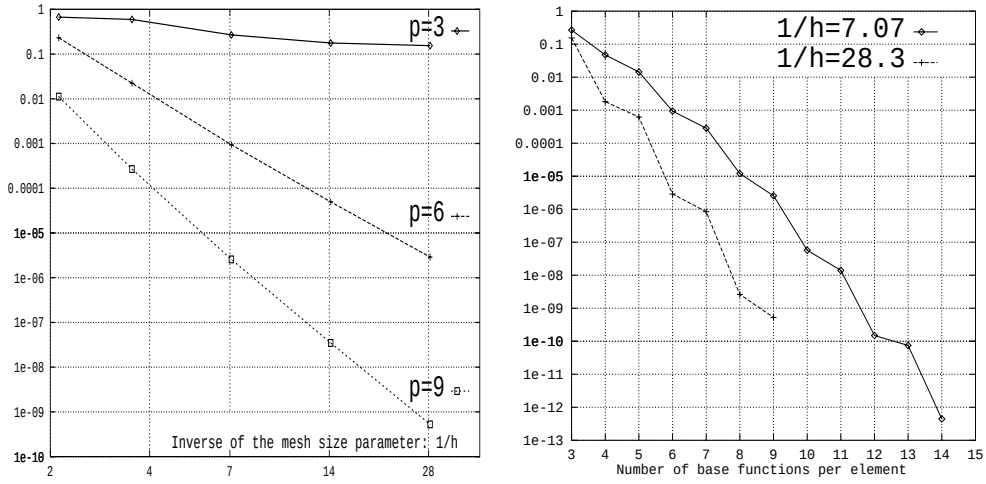
iii) Conclusion. From the above points i) and ii), we obtain a condition number whose asymptotic lower bound is given by (recalling $n = [p/2] - 1$ and $p \geq 4$)

$$C h_k^{1-(2n+1)} = C h_k^{-2[p/2]+2} \ . \qquad \square$$

This law seems to be optimal as the same law is obtained in the numerical simulation of Figure 3.6. Indeed, for $p = 6$ we computed from Figure 3.6 that the law was $h_k^{-4.2}$; for $p = 9$ we computed 6.1. Both values are very closed to the integer values given by the law (3.70). $u - u_h$ in $L^2(\Gamma)$ (for $u - u_h$ this is the case only when u is solution of the homogeneous problem). Indeed, the numerical study showed the estimate (3.70) is not optimal. We found a law in $-2[p/2]+1$ (Figure 3.6 (left); compare, for instance, the slope for $p = 6$ (midcurve) with the equivalent curve of Figure 3.2, section 3.2.2).

Remark 25. The iterative algorithm (2.34) demands, as a prerequisite, the inversion of the matrix D. Since the matrix D is block-diagonal, it can be inverted inverting the matrices D_k separately. This is why Theorem 3.10 is numerically interesting. It can be extended to the matrix D: we obtain the same law of the condition number of D, with the mesh size parameter h. Indication: use relation (2.4).

3.4.2. Computed evolution of the condition number (of D). Let us present a numerical test (Figure 3.6) of the evolution of the condition number as a function of the number of basis functions per element, and as a function of the mesh size parameter h, all the other parameters being fixed (see Table 3.2, section 3.2). The mesh in the example is structured, but this has no influence. By inspection, it is clear that



as a function of $1/h$ as a function of p

FIG. 3.6. *Inverse of the conditioning.*

a) The logarithm of the condition number is by inspection proportional to $2[p/2]$. This shows that the estimate (3.70) is not optimal.

b) The condition number problems do not arise for $p = 3$ basis functions per elements.

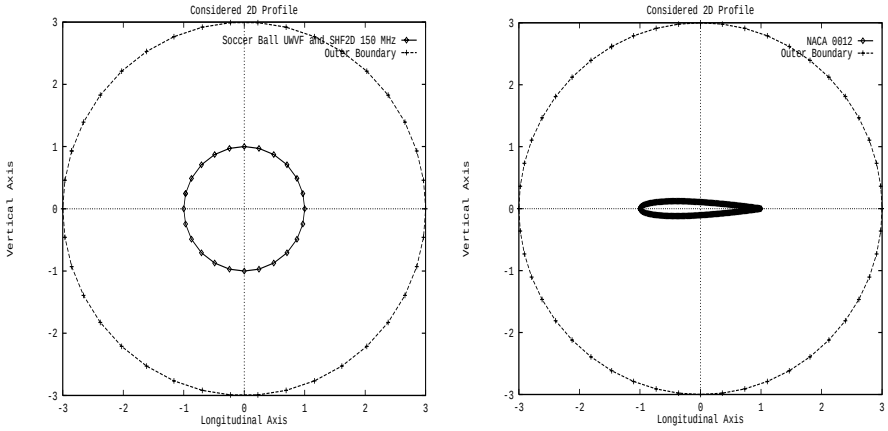
In [6] we show that for $p = 3$ basis functions per element the condition number of D_k is independent of h_k. It is proved that the equidistribution of the plane waves maximizes the determinant of D_k independently of the geometry of Ω_k.

4. Application to scattering problems. A $2D$ scattering problem is a two-dimensional acoustic wave diffraction that can be modelized by the approximate equation (4.1) as follows.

$$\begin{cases} (-\Delta - \omega^2)u = 0 & \text{in } \Omega \ , \\ (\partial_\nu + i\omega)u = 0 & \text{on } \Gamma_{ext} \ , \\ u = -u_{inc} & \text{on } \Gamma_{int} \end{cases} \tag{4.1}$$

where $\Gamma = \Gamma_{int} \cup \Gamma_{ext}$ the inner and outer boundaries of Ω, the inner boundary Γ_{int} being the frontier of the studied diffracting object (see, for instance, Figure 3.1b), section 3.2, or Figure 4.1). The u_{inc} notation stands for the incident wave. This formulation remains a simpler form of (0.1); it is covered by the theory as soon as the outer boundary Γ_{ext} actually exists. We are here to solve two scattering problems: a disk that is very roughly meshed so that it looks like a soccer ball, and a NACA profile. These two profiles are drawn in Figure 4.1 with their computation domains Ω. Table 4.1 sums up the characteristics of the two test cases. Let L denote the wing's length or the ball's diameter. Our two objects are so that $L = 2$. Bear in mind how we defined the mesh size parameter h by (3.2) section 3.2.2.

4.1. Radar cross section calculations (RCS). RCS computations are indeed one of the major purposes of frequency wave analysis. This was also a way to compare our UWVF with an already reliable code, namely an Integral Equations (IE) code called SHF2D and developed at CEA-CEL-V by Pierre Bonnemason and Bruno



Soccer Ball NACA 0012

FIG. 4.1. *Domains with boundary meshes.*

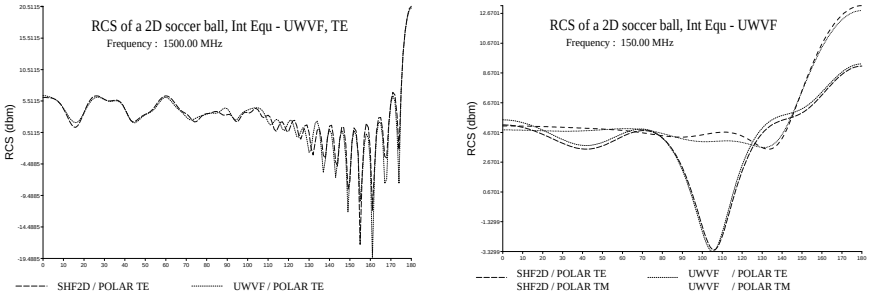
TABLE 4.1
Meshes' features.

f	0	
Refinement parameters	Value	
p: basis functions numb./elt	5	
NACA 0012	mesh size	
	L/h Area 1	100
	L/h Area 2	20
	L/h Area 3	5
Soccer Ball	mesh size	
	L/h	7.4
frequency	150 MHz	1500 MHz
L/λ	1	10
Boundary conditions	Dirichlet diffracted field, TM mode	Neumann diffracted field, TE mode
g_{int}	see Equation (3.3)	see Equation (3.3)
t_{int}	-1	$+1$
g_{ext}	0.0	0.0
t_{ext}	0.0	0.0
Incident wave	trailing edge	leading edge
$\vec{v}_0$	$(-1, 0)$	$(+1, 0)$

Stupfel [2]. This code simulates Neumann and Dirichlet boundary conditions using a perfectly isolating or conducting material, i.e., of impedance equal to zero or infinite. The RCS is $20 \log(|a(\theta)|^2)$ where $a(\theta)$ is given by formula (3.24) of section 3.3.2 in which $\vec{e}_\theta = (\cos\theta, \sin\theta)$, θ being the bistatic angle of observation with regard to the incident wave $\vec{v}_0$:

$$a(\theta) = \frac{1}{\sqrt{\omega}} \int_{\Gamma_{int}} e^{i\omega \vec{x} \vec{e}_\theta} (-i\omega \vec{\nu}\, \vec{e}_\theta u_h + \partial_\nu u_h) \ . \tag{4.2}$$

Using the approximation (2.14), section 2.2.1 of x_h, the relation satisfied by g (3.3) and then the definition of x_h (2.2) and the boundary condition (2.16) defining u_h



1500 MHz. TE mode 150 MHz. TM and TE modes

FIG. 4.2. *Soccer ball RCS, UWVF-IE.*

using x_h, we state

$$\begin{cases} (x_h)_{|\Gamma_k} = i\omega \sum_l x_{kl}(1 - \vec{\nu}_k\, \vec{v}_{kl}) e^{i\omega \vec{v}_{kl}\, \vec{x}}\ , \\ \\ (g)_{|\Gamma_k} = i\omega((1+t_k)\vec{\nu}_k\, \vec{v}_0 + (1-t_k)) e^{i\omega \vec{v}_0\, \vec{x}}\ , \end{cases} \quad \begin{cases} (-\partial_{\nu_k} + i\omega)(u_h)_{|\Gamma_k} = (x_h)_{|\Gamma_k}\ , \\ (+\partial_{\nu_k} + i\omega)(u_h)_{|\Gamma_k} = g + t_k (x_h)_{|\Gamma_k}. \end{cases}$$

So we can calculate the integral term of equation (4.2) on Γ_k

(4.3)
$$\begin{cases} 2i\omega(u_h)_{|\Gamma_k} = g + (1+t_k)(x_h)_{|\Gamma_k} \ \text{ and } \ \partial_{\nu_k}(u_h)_{|\Gamma_k} = g + (t_k - 1)(x_h)_{|\Gamma_k}\ , \\ 2(-i\omega\vec{\nu}_k\, \vec{e}_\theta u_h + \partial_{\nu_k} u_h) = (-\vec{\nu}_k\, \vec{e}_\theta (g + (1+t_k)(x_h)_{|\Gamma_k}) + (g + (t_k - 1)(x_h)_{|\Gamma_k})\ . \end{cases}$$

This replaced in equation (4.2) allows an exact computation of the RCS. According to the polarization, we multiply $a(\theta)$ of equation (4.2) by -1 in polarization mode TM, Dirichlet boundary conditions on Γ_{int}, and by i in polarization mode TE, Neumann boundary conditions on Γ_{int}.

4.2. Soccer ball. The soccer ball is meshed in an area enclosed in a circle of diameter three times the ball's diameter. The number of elements of the mesh is 346. Free outer edges (artificial absorbing boundary) are 38, inner edges 24. We obtain Figure 4.2 the RCS calculations. Let us point out that our discretization step of the soccer ball problem at 1500 MHz is very large: the mesh size parameter is around the same as the wavelength. Comparing with the SHF2D code, we had to multiply the number of elements around the profile by about 5 so as to reach the classical discretization law of IE methods $h \approx \lambda/5$.

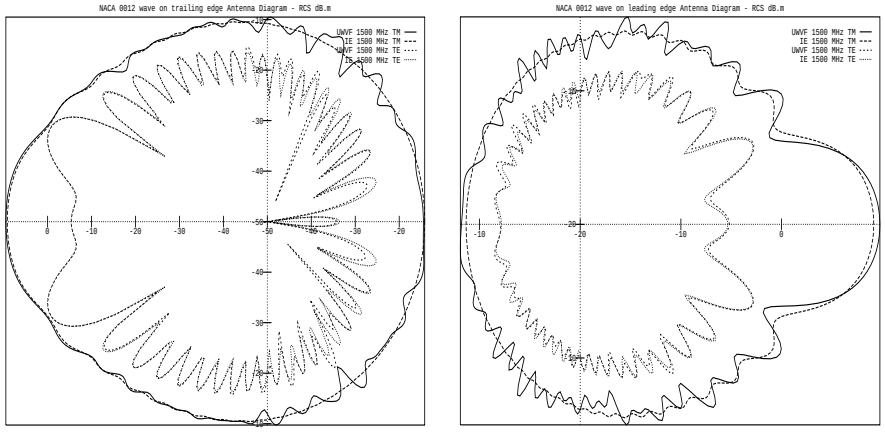
4.3. NACA 0012. The mesh used for the field computation is made of three areas as follows.

1st Area: along the profile, looks like a boundary layer, about 3% of the wing's length wide. Mesh size parameter = 1% of the wing's length.

2nd Area: enclosing the boundary layer in a limited area about the same surface as the profile. Mesh size parameter = 5% of the wing's length.

3rd Area: main domain as regards the surface. Limited by the previous area and a circle of diameter three times the profile's length. Mesh size parameter = 20% of the wing's length. Let us point out, this mesh is extremely large as at 1500 MHz it corresponds to $h \approx 2\lambda$, which is very far from the empirical law (0.3) between the discretization step and the frequency.

FIG. 4.3. *Dirichlet TM imaginary diffracted field NACA* 0012 *on trailing edge at* 1500 *MHz.*



trailing edge leading edge

FIG. 4.4. *Antenna diagram NACA* 0012 *RCS, UWVF-IE comparison at* 1500 *MHz.*

We obtain an overall number of 2976 elements of 1615 nodes. Free outer edges (artificial absorbing boundary) are 48, inner edges 202. For the NACA wing we give a representation of the field u around the profile. This is achieved using a finer grid than the discretization mesh. We compute u_h using formula (2.20), section 2.2.1. The mesh used for the field representation (i.e., the domain of Figure 4.3) is made of the areas 1 and 2 described above, but with a uniform mesh size parameter of 1% of the wing's length. We obtain an overall number of 3572 elements of 2010 nodes. Free outer edges (artificial graphic boundary) are 246, inner edges 202.

We obtain Figure 4.4 the RCS calculations.

5. Conclusion. In the model problem of the two-dimensional Helmholtz problem, this study has proven the two essential properties of our method.

1. The order of convergence is lower bounded by a linear function of the number of basis functions per element, whereas in the FEM, the order behaves as the square root of the number of basis functions.

2. The setup of the method is similar to that of the FEM: we assemble a sparse matrix from a regular mesh.

3. The linear system is invertible unconditionally on the discretization step. An iterative algorithm easily inverts the linear system.

It thus seems that the UWVF method is an interesting alternative to other discretization methods, such as the FEM, the finite difference method, and the finite volume method. The next step would involve studying what happens when dealing with large elements. This is likely to be a good way of solving frequency wave problems that is still difficult to achieve nowadays, due to the massive storage requirements

of common methods. As seen in section 4 with scattering problems, our method is stable even with a very large discretization step. Furthermore, we will be able to practice frequency sweeps on the same mesh or grid, merely by raising the number of basis functions or degrees of freedom. We will generalize these techniques to the harmonic Maxwell's equations [5].

Note that the suggested formulation is also applicable to a wide range of partial differential equations (PDEs) (see [14]), that could lead to a forthcoming study. The formulation holds in nonhomogeneous media and will adapt to diffraction problems on materials. We should obtain similar results concerning the rate of convergence to those obtained in this paper.

As it is to be studied in [5], it is also applicable to 3-dimensional problems, for instance Maxwell's harmonic problem. The order of convergence is expected to behave as the square root of the number of degrees of freedom [5] and not as to the power third as in the FEM.

Appendix A. Generating the matrices. We analytically calculate the terms of the discrete formulation's matrices (2.21), and the second member in some particular cases. This section is rather technical. It is important for the practical point of view, but is of no theoretical interest. The authors believed it had to be joined to this paper to make it self-contained.

The notation required for this section has the same labeling convention as the one of the ideas mesh generator. It is not a natural labeling on a triangle, since a mesh generator can be operated using other types of polyhedra where their labeling is simpler.

- Geometry (vertices, edges, normals, lengths):
  1. Let (x_1^k, x_2^k, x_3^k) be the positions of the three vertices of triangle Ω_k.
  2. Let (x_1^{kj}, x_2^{kj}) be the vertices of edge Σ_{kj}.
  3. Let $(\nu_1^k, \nu_2^k, \nu_3^k)$ be the three outer normals to triangle Ω_k, to edges (x_1^k, x_2^k), (x_2^k, x_3^k), (x_3^k, x_1^k), respectively.
  4. Let (ν_{kj}) be the outer normal of Ω_k towards triangle Ω_j. This is also the normal to edge (x_1^{kj}, x_2^{kj}).
  5. Let (L_1^k, L_2^k, L_3^k) be the lengths of the three edges of triangle Ω_k, i.e., $|x_1^k - x_2^k|, |x_2^k - x_3^k|, |x_3^k - x_1^k|$.
  6. Let L_{kj} be the length of edge Σ_{kj}: $|x_1^{kj} - x_2^{kj}|$.
- Quantities defined specifically to generate the terms of the matrices:
  1. For $n = 1$ to 3, $h_n^k = \omega(\vec{v}_{km} - \vec{v}_{kl})\,(\vec{x}_{n+1}^k - \vec{x}_n^k)/2$ with $x_4^k = x_1^k$.
  2. For $n = 1$ to 3, $Z_n^k = e^{(i\omega(\vec{v}_{km} - \vec{v}_{kl})\,\vec{x}_n^k)}$.
  3. Let $h_{kj} = \omega(\vec{v}_{jm} - \vec{v}_{kl})\,(\vec{x}_2^{kj} - \vec{x}_1^{kj})/2$.
  4. Let $Z_{kj} = e^{(i\omega(\vec{v}_{jm} - \vec{v}_{kl})\,\vec{x}_1^k)}$.

Using the above notation, we are able to calculate the values of the terms of the matrices D and C as well as, in some cases, the second member b.

i) Let $D_k^{l,m} = D_{k,k}^{l,m}$ be the sum of terms corresponding to the contributions of each edge of the triangle Ω_k. Recall that $(k \neq j) \Rightarrow (\forall (l,m)\ D_{k,j}^{l,m} = 0)$

$$D_k^{l,m} = \omega^2 \sum_{n=1}^{3} L_n^k Z_n^k (1 - \vec{\nu}_n^k\,\vec{v}_{km})(1 - \vec{\nu}_n^k\,\vec{v}_{kl}) \frac{\sin h_n^k}{h_n^k} e^{ih_n^k}\ . \tag{A.1}$$

ii) The matrix C is composed of coupling terms between two neighboring elements and of coupling diagonal terms on the boundary Γ.

1. In the case where Ω_k is a neighbor of Ω_j, $C_{k,j}^{l,m}$ is the contribution of the interface Σ_{kj}

$$C_{k,j}^{l,m} = \omega^2 L_{kj} Z_{kj} (1 + \vec{\nu}_{kj}\, \vec{v}_{jm})(1 + \vec{\nu}_{kj}\, \vec{v}_{kl}) e^{ih_{kj}} \frac{\sin h_{kj}}{h_{kj}} \ . \tag{A.2}$$

2. In the case where Γ_k is not an empty set, $C_{k,k}^{l,m}$ is the contribution of the boundary condition on Γ_k. Assuming t_k is constant on Γ_k, we have

$$C_{k,k}^{l,m} = \sum_{n/[x_n^k, x_{n+1}^k] \in \Gamma_k} t_k \omega^2 L_n^k Z_n^k (1 - \vec{\nu}_n^k\, \vec{v}_{km})(1 + \vec{\nu}_n^k\, \vec{v}_{kl}) e^{ih_n^k} \frac{\sin h_n^k}{h_n^k} \ . \tag{A.3}$$

3. If $\Gamma_k = \emptyset$ or $\overline{\Omega_k} \cap \overline{\Omega_j} = \emptyset$ there is no coupling term: $C_{k,j}^{l,m} = 0$.

iii) Let us give the values of the right-hand side b in the following test cases. It is still assumed t is a constant function on any free edge Γ_k. According to (2.24) we have

$$b_{k,l} = -2i\omega \int_{\Omega_k} f . \overline{e^{(i\omega \vec{v}_{kl} . \vec{x})}} + \int_{\Gamma_k} g . \overline{(+\partial_{\nu_k} + i\omega) e^{(i\omega \vec{v}_{kl} . \vec{x})}} \ . \tag{A.4}$$

Note the linearity of $b_{k,l}$ as a function of f and g.

1. For $f = \delta_{x_0}$, we have

$$b_{k,l} = \begin{cases} -2i\omega e^{-i\omega(\vec{v}_{kl} . \vec{x_0})} & \text{if } x_0 \in \Omega_k \ , \\ 0 & \text{otherwise.} \end{cases} \tag{A.5}$$

2. For $g = [(1+t_k)(\partial_{\nu_k}) + i\omega(1-t_k)] e^{i\omega(\vec{v}_0 . \vec{x})}$, we have, defining $\xi = (1+t_k)\vec{\nu}_n \vec{v}_0 + 1 - t_k$

$$b_{k,l} = \sum_{n/[x_n^k, x_{n+1}^k] \in \Gamma_k} \omega^2 L_n Z_n^k \xi (1 + \vec{\nu}_n\, \vec{v}_{kl}) e^{ih_n^k} \frac{\sin h_n^k}{h_n^k}. \tag{A.6}$$

3. For $f = \mu\omega^2 e^{i\omega(\vec{v}_0 . \vec{x})}$, we have

$$b_{k,l} = (-4i\omega^3 \mu S) . \left(Z_2 \frac{e^{i\alpha} \dfrac{\sin \alpha}{\alpha} - e^{i\beta} \dfrac{\sin \beta}{\beta}}{2i(\alpha - \beta)} \right) \tag{A.7}$$

where $\alpha = \omega(\vec{v}_0 - \vec{v}_{kl})\,(\vec{x}_1^k - \vec{x}_2^k)/2$, $\beta = \omega(\vec{v}_0 - \vec{v}_{kl})\,(\vec{x}_3^k - \vec{x}_2^k)/2$, $Z_2 = e^{(i\omega(\vec{v}_0 - \vec{v}_{kl})\, \vec{x}_2^k)}$, and $S = \frac{1}{2} L_1^k L_3^k \sin(x_2^k x_1^k x_3^k) = \text{surface}\ (\Omega_k)$.

The detailed calculations are to be found in [5]. There can be found details of Taylor polynomials around singular points (i.e., $\alpha = 0$, $\beta = 0$, $\alpha = \beta$ in formula (A.7)), and a precise explanation on the vectorized numerical implementation on a Cray supercomputer.

Acknowledgments. The authors are greatly indebted to Professor Patrick Joly for both his reviewings of this paper and his many rewordings in fundamental parts of this article. Professor Patrick Joly restructured some sections as well as bringing his expertise and rigor in reasoning. This paper would never have been published without his help. The authors would like to express their thanks to Pierre Bonnemason for his help in computing and using the SHF2D code and to Dr. Olivier Lafitte for his many rereadings.

REFERENCES

[1] C. Bernardi and Y. Maday, *Approximations spectrales de problèmes aux limites elliptiques*, Mathématiques et Applications 10, Springer-Verlag, Berlin, New York, 1992.

[2] P. Bonnemason and B. Stupfel, *Résolution par formulations intégrales du problème de la diffraction d'une onde électromagnétique monochromatique par des cylindres infinis: code SHF2D*, Tech. Report, CEA/CEL-V, BP 27, F-94 195 Villeneuve St. Georges Cx., November 1992.

[3] P. Bonnemason and B. Stupfel, *Modeling high frequency scattering by axisymmetric perfectly or imperfectly conducting scatterers*, Electromagnetics, 13 (1993), pp. 111–129.

[4] H. Brézis, *Analyse fonctionnelle - Théorie et applications*, Masson, Paris, 1987.

[5] O. Cessenat, *Application d'une nouvelle formulation variationnelle aux équations d'ondes harmoniques, Problèmes de Helmholtz 2D et de Maxwell 3D*, Ph.D. thesis, FRANCE/Paris IX Dauphine, 1996, to appear.

[6] O. Cessenat and B. Després, *Une nouvelle formulation variationnelle des équations d'onde en fréquence, Application au problème de Helmholtz 2D*, Tech. report 2779, CEA/CEL-V, BP 27, F-94 195 Villeneuve St Georges Cx., December 1994.

[7] J. Chabrowski, *The Dirichlet Problem with L^2-Boundary Data for Elliptic Linear Equations*, Lecture Notes in Mathematics 1482, Springer-Verlag, New York, 1991.

[8] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1979.

[9] D. Colton and R. Kreiss, *Integral Equation Methods in Scattering Theory*, Wiley-Interscience, New York, 1983.

[10] R. Dautray and J. Lions, *Mathematical Analysis and Numerical Calculus*, Masson, Paris, 1987.

[11] B. Després, *Méthode de décomposition de domaine pour les problèmes de propagation d'ondes en régime harmonique, Le théorème de Borg pour l'équation de Hill vectorielle*, Ph.D. thesis, FRANCE/Paris IX Dauphine, 1991.

[12] B. Després, *Domain decomposition method and the Helmholtz problem (Part* ii*)*, in 2nd International Conference on Mathematical and Numerical Aspects of Wave Propagation Phenoma, Newark, DE, SIAM, Philadelphia, 1993.

[13] B. Després, *Un procédé de discrétisation des équations d'ondes en fréquence*, Tech. report 2726, CEA/CEL-V, BP 27, F-94 195 Villeneuve St Georges Cx., May 1993.

[14] B. Després, *Sur une formulation variationnelle de type ultra-faible*, C.R. Acad. Sci. Paris, 318 (1994), pp. 939–944.

[15] M. Dryja and O. Widlund, *Some recent results on Schwarz type domain decomposition algorithms*, in Domain Decomposition Methods in Science and Engineering, Q. P. K. Widlund, ed., Amer. Math. Soc., Providence, RI, 1994.

[16] C. Le Potier and R. Le Martret, *Finite volume solution of Maxwell's equations in nonsteady mode*, La Recherche Aérospatiale, 5 (1994), pp. 329–342.

[17] J. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Vol. 1 et 2, Dunod, Paris, 1968.

[18] L. Marini and A. Quarteroni, *An iterative procedure for domain decomposition methods: A finite element approach*, in Domain Decomposition Methods for Partial Differential Equations, G. G. M. Périaux, ed., SIAM, Philadelphia, PA, 1988, pp. 129–143.

[19] L. Marini and A. Quarteroni, *A relaxation procedure for domain decomposition methods using finite elements*, Numer. Math., 55 (1989), pp. 575–598.

[20] J. Nédelec, *Approximation des équations intégrales en mécanique et en physique*, cours de l'école d'été d'analyse numérique CEA-EDF-INRIA, Ecole Polytechnique, 1977.

[21] G. Opfer and G. Schober, *Richardson's iteration for nonsymmetric matrices*, Linear Algebra Appl., 58 (1984), pp. 343–361.

[22] P. Raviart and J. Thomas, *Introduction à l'analyse numérique des équations aux dérivées partielles*, in Collection Mathématiques Appliquées pour la maîtrise, 3rd ed., Masson, Paris, 1992.

[23] J. E. Roberts and J. M. Thomas, *Mixed and hybrid elements*, North–Holland, Amsterdam, 1991, pp. 521–633.

[24] P. L. Tallec, *Domain decomposition methods in computional mechanics*, Comput. Mech. Advances, 2 (1994), pp. 121–220.

[25] R. Theodor and P. Lascaux, *Analyse numérique matricielle appliquée à l'art de l'ingénieur*, Vol. 2, Masson, Paris, 1986.

[26] K. Yee, *Numerical solution of initial boundary value problem in isotropic media*, IEEE Trans. Antennas and Propagation, AP-14 (1966), pp. 302–307.