A generalized plane wave numerical method for smooth non constant coeffIcients

Lise-Marie Imbert-Gérard, Bruno Després

Journal (2012) Page 1 of 34
p

A generalized plane wave numerical method for smooth non constant coefficients

LISE-MARIE IMBERT-GÉRARD† AND BRUNO DESPRÉS‡
Laboratoire Jacuqes-Louis LIONS, Université Pierre et Marie Curie,
Boîte courrier 187, 75252 Paris Cedex 05 France

[Received on 29 August 2012; revised on ??]

We propose an original method based on generalized plane waves and approximated coefficients for the numerical approximation of the Helmholtz equation with a smooth constant coefficient. This is justified by a high order convergence estimate rate. Our motivation stems from the Maxwell's equations with Hermitian dielectric tensor ε which are used to model reflectometry in fusion plasma. Simplified models split them into two different propagation modes. Some numerical results are presented in dimension one and two.

Keywords: generalized plane wave; wave equations; high order.

1. Introduction

Our aim is to describe a new numerical method with generalized plane waves for the numerical approximation of time harmonic wave equations with smooth non constant coefficients. Our model problem is the Helmholtz problem with a smooth non constant coefficient

$$\left\{ \begin{array}{rcll} -\Delta u + \alpha u & = & f, & x \in \Omega, \\ (\partial_\nu + i\gamma)\, u & = & Q\,(-\partial_\nu + i\gamma)\, u + g, & x \in \Gamma. \end{array} \right. \tag{1.1}$$

Here the smooth coefficient is real, $\alpha \in \mathbb{R}$, but everything could be adapted replacing the homogeneous problem by the adjoint problem. The γ function can be a variable physical parameter satisfying $0 < \gamma_m \leqslant \gamma \leqslant \gamma_M$, but for the sake of simplicity we will consider it constant and positive. The sign of α may change even if we restrict the presentation to real coefficients for simplicity. The method can be used for complex valued coefficients as well. The unknown $u(\mathbf{x}) \in \mathbb{C}$ is sought in the space of complex valued functions.

1.1 *Physical motivations*

Our motivation comes from the need of efficient numerical methods for certain Maxwell's harmonic equations appearing in plasma physics. These equations read

$$\mathrm{curl}\,(\mathrm{curl} E) - \frac{\omega^2}{c^2}\varepsilon(\mathbf{x})E = 0, \quad \mathbf{x} = (x,y,z), \tag{1.2}$$

where E denotes the electric field, ω is the pulsation, c the sound speed and ε the dielectric tensor. The dielectric tensor represents the electromagnetic behavior of the media. The cold plasma theory Swanson

†Corresponding author. Email: imbert@ann.jussieu.fr
‡Email: despres@ann.jussieu.fr

(27) yields the already simplified dielectric tensor is

$$\varepsilon(\mathbf{x}) = \begin{pmatrix} 1-a(\mathbf{x}) & iba(\mathbf{x}) & 0 \\ -iba(\mathbf{x}) & 1-a(\mathbf{x}) & 0 \\ 0 & 0 & 1-ca(\mathbf{x}) \end{pmatrix}, \qquad i^2 = -1,$$

where $b < 1$ and $c = 1 - b^2$. Typically the coefficient a satisfies $a = x/x_0 + \tilde{a}$, where the perturbation is described by $\tilde{a} = \tilde{a}_0 \exp\left(-\frac{(x-x_f)^2}{\omega_x^2}\right) \exp\left(-\frac{(y-y_f)^2}{\omega_y^2}\right) \cos(q(y-y_f))$, see Gusakov et al. (12). This is completed with boundary conditions of metallic or absorbing type. We refer to Monk (23) for the general theory of Maxwell's equations and to Cessenat & Després (2); Hiptmair et al. (14); Huttunen et al. (16) for the use of specific plane wave methods for the numerical approximation of the solutions of such problems. Two models for different propagation modes are often considered. Both are obtained from equation (1.2) under convenient assumptions on the direction and polarization of the electric field. The two dimensional equation for what is called the O-mode reduces to

$$-\Delta E_z - \frac{\omega^2}{c^2}\varepsilon_z(x,y)E_z = 0, \quad \Delta = \partial_{xx} + \partial_{yy}, \tag{1.3}$$

on the domain Ω and can be completed by the following boundary condition

$$(\partial_\nu + i\gamma)E_z = Q(-\partial_\nu + i\gamma)E_z + g$$

on the boundary domain Γ. Here ∂_ν denotes the normal derivative, $\gamma > 0$ is a smooth positive function and g is for instance an L^2-function on the boundary. Q is a smooth function allowing to fit the condition : if $Q = -1$ it gives a Dirichlet condition, if $Q = 1$ a Neumann condition or if $Q = 0$ a Robin condition. This O-mode (named for **O**rdinary mode) presents one cutoff : when ε_z is negative or positive the nature of the equation (1.3) is either elliptic coercive or elliptic propagative. This coefficient $\varepsilon_z \in \mathbb{R}$ is a real continuous function. It depends on the local density of electrons and on the exterior frozen magnetic field. Since the electron density is continuous, it explains why the coefficient of the equation is also a continuous function. A more general setting of the physical problem could be to introduce some dissipation with a complex valued coefficient. This case is not considered hereafter. A further simplified one dimensional model reads

$$-\frac{d^2}{dx^2}E_z + xE_z = 0. \tag{1.4}$$

Equations (1.3) and (1.4) are particular cases of our model problem (1.1). The fundamental solutions are the two Airy functions Ai and Bi. The first Airy function Ai displays important properties which are fundamentally related to the physics of the problem. It will be used for validation in our numerical tests.

More generally a challenge is to adapt advanced numerical methods which are at the frontier of what is used in classical engineering so as to obtain efficient algorithms which can be used in the context of the numerical modeling of Fusion plasmas. We are in particular interested by reflectometry which is a diagnostic method to measure density in fusion plasmas, based on the reflection of the probing wave at a plasma cut-off. The local density at the cut-off layer can be deduced from the reflected signal. Therefore the mathematical model of the cut-off is crucial for the application. Our motivation is that the equation (1.3) models the cut-off as the zero of the coefficient ε_z, it is then natural to take into account the smoothness feature of the coefficient in the design of the numerical method. In fact, classical numerical methods described in paragraph 1.2 consider only piecewise coefficients which would damage the description of the cut-off. We hope this work can be considered as a first step in this direction.

1.2 *Plane wave methods*

The numerical method that we propose is an extension of plane waves methods, such as the ultra weak variational formulation (UWVF) Després (4); Cessenat & Després (2, 3); Gittelson et al. (10); Huttunen et al. (17), to problems with smooth non constant coefficients. Indeed the standard UWVF uses constant coefficients per cell. This is optimal when the physical domain can be split into sub-domains in which the coefficients are constant. But if the coefficients of the problem to solve are non constant and smooth, such a procedure introduces a priori an important error. Our aim is to propose and analyze an extension of UWVF which uses original basis functions based on the generalized plane waves Melenk (20).

We think that the approach proposed in this work is not restricted to UWVF, and can be generalized to different plane wave methods that we describe here. PUFEM Melenk (19); Melenk & Babuska (21) falls in the same class of method Perrey-Debain et al. (24); Pluymers et al. (25). It has also been shown that UWVF can be interpreted as a special Discontinuous Galerkin procedure Gittelson et al. (10); Hiptmair et al. (14); Farhat et al. (6, 8). The analysis of the classical ultra weak variational formulation method described as a discontinuous Galerkin method is performed in Huttunen et al. (16), as well as the corresponding h and p convergence theory. It has been proved that the analysis of h-convergence takes great advantage of this fact in Buffa & Monk (1); Gittelson et al. (10). The analysis of p convergence is treated in Hiptmair et al. (13). Comparisons between these methods is investigated in Gabard et al. (9); Huttunen et al. (15); Strouboulis et al. (26); Wang et al. (28). Analysis with respect to the wave-number k is performed in Melenk & Sauter (22). The new family of generalized plane waves described in this work generates a high order method with respect to the basis functions and the coefficients of the problem: in this direction we refer also to the Enrichment method Kalashnikova et al. (18); Farhat et al. (7) which could provide an alternative to our method in order to enrich a more conventional polynomial basis with our generalized plane wave basis. Ultimately there is no opposition between all these approaches in the sense that it is possible at the level of principles to mix polynomial basis functions and generalized plane waves to obtain a method with improved approximation properties. However it is far beyond the scope of this paper and will not be considered.

The performance of the new method relies on a further investigation of the ultra weak variational formulations method presented in Després (4). In fact, the ultra weak variational formulation uses basis function adapted to the original problem, in the sense that they are exact solution to the problem. It makes the approximation of the exact solution more relevant for a given number of elements in the mesh and a given number of basis functions per element than the approximation obtained with classical finite elements methods. Then considering smooth non constant instead of piecewise constant coefficients, it's coherent to look for more general basis functions since classical plane waves can no more be exact solutions of the problem. So the idea is to construct, in the vicinity V_0 of a point $x_0 \in \Omega$ basis functions

- that are generalized plane waves, say $\varphi = e^{P(x)}$ for $x \in V_0$ and $P \in \mathbb{C}[X]$,

- that are solution to a modified problem : $-\Delta\varphi + \tilde{\alpha}\varphi = 0$, such that $\tilde{\alpha}$ satisfies the approximation property $\|\alpha - \tilde{\alpha}\|_{L^\infty(V_0)} \leqslant Ch^q$, where h denotes the size of V_0, C is a constant and $q \geqslant 1$ is a given entire number.

The aim of this paper is first to design the new adapted basis functions, second to construct the adapted tools to fit with the frame of non confirming finite element methods in order to follow the steps of the second Strang lemma for the estimation of the convergence rate and third to illustrate with basic numerical tests for the Airy equation.

1.3 *Plan*

This work is organized as follows. We present a family of generalized plane waves in the section 2 : these functions have been designed to be the standard plane waves in case the coefficients of the problem are constant in space and non positive. In section 3 we present the general principle of UWVF and adapt it to smooth coefficients. The next section 4 is devoted to the numerical analysis of the method. Our main theoretical result is a proof of convergence in dimension one, using the second Strang's lemma and some uniform coercivity estimates. This is probably the most original theoretical result in our work. To our knowledge it is the first time that it is introduced and analyzed in the context of generalized plane wave methods. Numerical results are provided in section 5 to illustrate the theoretical results. In particular we display experimental convergence estimates in dimension two. The numerical results suggest that a different normalization of the generalized plane waves may increase the accuracy, which is indeed what is observed. Additional technical material is provided in the appendix.

2. Generalized plane waves

Unlike the classical variational formulation used for instance by finite element methods, here the variational formulation requires meshing the domain as a preliminary task. This feature is shared by Ultra Weak Variational Formulations, Discontinuous Galerkin Methods, Enriched methods and other plane wave methods. The coupling strategy between the cells differs of course. But up to this fact the generalized plane waves can be used in principle for all such algorithms. We begin with some notations. The mesh of the domain Ω is denoted $\mathcal{T}_h = \{\Omega_k\}_{k\in[\![1,N_h]\!]}$, such that :

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

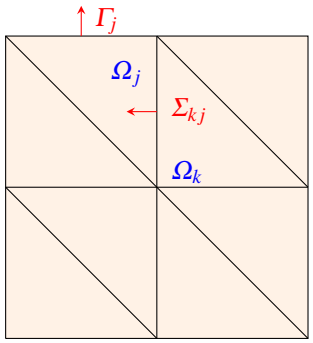


FIG. 1: Example of a meshed square domain Ω, with elements Ω_k, edges Σ_{kj} and Γ_j respectively oriented toward Ω_j and the exterior of the domain.

2.1 *A general method*

Shape or basis functions φ are sought as solutions of the homogeneous equation

$$(-\Delta+\alpha)\varphi = 0. \tag{2.1}$$

If the coefficient α is constant in the cell and negative, it is sufficient to use plane waves, that is in dimension two $x=(x_1,x_2)$

$$\varphi(x,y) = e^{\sqrt{\alpha}(d_1x+d_2y)} \text{ with } d=(d_1,d_2) \text{ and } (d,d)=1. \tag{2.2}$$

If the vector d is real, it is simply the direction of the plane wave. This is the basic idea of all plane wave methods. However if α is non constant in the cell, then we do not know of any simple and general analytical formula for φ. For example if $\alpha = x$ is linear with respect to the first variable, it is possible to construct φ from the Airy functions Ai and Bi. But the Airy functions are highly transcendental, they are not that evident to manipulate.

Our main goal is to describe a method of approximation which can be used for any function α. By comparison with (2.2) it is natural to to generalize the plane formula and to consider generalized plane waves such as

$$\varphi(x,y) = e^{P(x,y)} \qquad P \text{ a polynomial.}$$

One gets that φ is solution of (2.1) if and only if

$$\frac{\partial^2}{\partial x^2}P + \left(\frac{\partial}{\partial x}P\right)^2 + \frac{\partial^2}{\partial y^2}P + \left(\frac{\partial}{\partial y}P\right)^2 = \alpha(x,y). \tag{2.3}$$

However many tries showed that such a representation is not sufficient. The explanation is simple: as P can be expanded as a finite series of monomial $x^n y^m$, the result is a finite series of term

$$\begin{aligned}&\frac{\partial^2}{\partial x^2}x^n y^m + \left(\frac{\partial}{\partial x}x^n y^m\right)^2 + \frac{\partial^2}{\partial y^2}x^n y^m + \left(\frac{\partial}{\partial y}x^n y^m\right)^2\\ &= \left(n(n-1)x^{n-2}+nx^{2n-2}\right)y^m + x^n\left(m(m-1)y^{m-2}+m^2y^{2m-2}\right).\end{aligned} \tag{2.4}$$

For example let us consider the case $\alpha(x,y)=x$ and let us look for a polynomial

$$P = \sum_{n\leqslant K}\sum_{m\leqslant L} a_{nm}x^n y^m, \qquad a_{KL}\neq 0,$$

solution of (2.3). If $K\geqslant 2$ of $L\geqslant 2$, the maximal degree of (2.4) cannot decrease which is contradictory with the fact that $\alpha = x$ is a polynomial of degree one. So $K\leqslant 1$ and $L\leqslant 1$. In this case the degree of (2.4) with respect to x is 0, and the same for the degree with respect to y. In summary solutions of the functional equation (2.3) cannot be polynomial in the general case.

Therefore a modification is needed. Instead of considering that α is given and looking for solution of (2.3), we look for approximate solutions. That is we consider the approximate equation

$$\left(-\Delta+\alpha_k^l\right)\varphi_k^l = 0 \text{ in } \Omega_k$$

where α_k^l is an approximation of α in Ω_k.

Definition 2.1. Generalized plane waves *A generalized plane wave will be understood as any function (with support in cell Ω_k) of the form*

$$\varphi_k^l = e^{P_k^l(x,y)}$$

where P_k^l is a polynomial solution of

$$\frac{\partial^2}{\partial x^2}P_k^l + \left(\frac{\partial}{\partial x}P_k^l\right)^2 + \frac{\partial^2}{\partial y^2}P_k^l + \left(\frac{\partial}{\partial y}P_k^l\right)^2 = \alpha_k^l(x,y) \tag{2.5}$$

and α_k^l is an approximation of α.

2.2 *Design of the basis functions in dimension one*

The one dimensional case is a first step to construct the coefficients α_k^l and the generalized plane wave functions φ_k^l. Therefore we will suppose in this section that $\Omega =]a,b[\subset \mathbf{R}$ and that $\overline{\Omega} = \cup_{k\in[\![1,N_h]\!]}[x_k, x_{k+1}]$, with $x_k < x_{k+1}$. The middle of the open interval $\Omega_h =]x_k, x_{k+1}[$ is denoted by $x_{k+1/2} = \frac{x_k + x_{k+1}}{2}$. Apart from providing the technical details of the construction of the basis functions, the central result of this section is an explanation why it is necessary to use different approximations α_k^l of the function α in the same cell $[x_k, x_{k+1}]$ in order to avoid a singularity in the construction.

2.2.1 *Design principle.*

We want here to set our choice of basis functions : in order to generalize plane wave methods, we will consider exponential of polynomials

$$\varphi(x) = e^{P(x)}.$$

Notice that we only need two basis functions per element of the mesh in dimension one: this is a common property of plane wave methods in dimension one; the reason is the number of elementary solutions of a second order differential equation which is two. Plugging the previous representation formula into the homogeneous equation $-\varphi'' + \alpha\varphi = 0$ we find the functional equation

$$P''(x) + P'(x)^2 = \alpha(x), \qquad x \in [x_k, x_{k+1}].$$

This equation is non linear and no simple solution is available for general right hand side α. However if α is locally constant, that is

$$\alpha(x) = \alpha(x_{k+1/2}) \in \mathbb{R}, \qquad x \in [x_k, x_{k+1}],$$

then

$$P_k^{\pm}(x) = \pm\sqrt{\alpha(x_{k+1/2})}x$$

are two natural solutions which correspond to the two local plane waves $\varphi_k^{\pm}(x) = e^{P_k^{\pm}(x)}$ in the case $\alpha(x_{k+1/2}) < 0$.

2.2.2 *Local approximation.*

To ensure the local approximation of the α, one has to fit the polynomials' coefficients to approximate the Taylor expansion of the equation's coefficient α, which is performed with respect to the parameter h which represents the length of the mesh

$$h = \max_k(x_{k+1} - x_k).$$

A first idea is to approximate the Taylor expansion of α

$$\alpha = \alpha_{\pm} + O(h^q) \tag{2.6}$$

and to look for polynomials solutions of

$$P_{\pm}'' + (P_{\pm}')^2 = \alpha_{\pm}, \qquad x \in [x_k, x_{k+1}].$$

Without restriction we assume that α admits a local infinite expansion

$$\alpha = \sum_{i=0}^{\infty} \frac{d^i \alpha}{dx^i}(x_{k+1/2}) \left(x - x_{k+\frac{1}{2}}\right)^i, \qquad x \in [x_k, x_{k+1}].$$

Using the finite expansion $P_{\pm} = \sum_{i \leqslant I} \beta_i^{\pm} y^i$ where $y = x - x_{k+\frac{1}{2}}$, one obtains

$$\alpha_{\pm} = P_{\pm}'' + (P_{\pm}')^2 = \left(\sum_{i \leqslant I} \beta_i^{\pm} y^i\right)'' + \left(\left(\sum_{i \leqslant I} \beta_i^{\pm} y^i\right)'\right)^2.$$

In order to satisfy (2.6) we have to chose $I \in \mathbb{N}$ and $(\beta_i)_{0 \leqslant i \leqslant I}$ such that

$$\left(\sum_{i \leqslant I} \beta_i^{\pm} y^i\right)'' + \left(\left(\sum_{i \leqslant I} \beta_i^{\pm} y^i\right)'\right)^2 = \sum_{i=0}^{q-1} \frac{d^i \alpha}{dx^i}(x_{k+1/2})\, y^i + O(h^q). \tag{2.7}$$

Identifying the coefficients in the polynomial part of the previous equation leads to a system of q equations with I unknowns. Then choosing I high enough ensures that the system is easy to solve. At the same time it is reasonable to choose I as small as possible to minimize the amount of computations. The main question is therefore to determine the optimal value of the degree of the polynomials, parameter I, with respect to the order of approximation, parameter q. Some remarks and examples follow.

- Normalization : $\beta_0 = 0$. It is always possible to take $\beta_0 = 0$ since β_0 does not show up in (2.7). It implies that the amplitude of the corresponding basis function is normalized in the cell since

$$e^{P_{\pm}\left(x_{k+\frac{1}{2}}\right)} = e^0 = 1.$$

- Trivial case : $q = I = 1$. From (2.7) one obtains the equation $\beta_1^2 = \alpha\left(x_{k+\frac{1}{2}}\right)$. One recovers from this procedure $\beta_1 = \pm\sqrt{\alpha\left(x_{k+\frac{1}{2}}\right)}$ so

$$P_{\pm}(x) = \pm\sqrt{\alpha\left(x_{k+\frac{1}{2}}\right)}\left(x - x_{k+\frac{1}{2}}\right).$$

  In the case where $\alpha\left(x_{k+\frac{1}{2}}\right) < 0$, it yields two plane waves with opposite directions. This case is the trivial one.

• Counter-example : $q = I = 2$. The discrete equations are obtained from the first two terms in (2.7)

$$\left\{ \begin{array}{l} 2\beta_2 + \beta_1^2 = \alpha\left(x_{k+\frac{1}{2}}\right) \equiv a, \\ 4\beta_1\beta_2 = \alpha'\left(x_{k+\frac{1}{2}}\right) \equiv b. \end{array} \right. \tag{2.8}$$

Elimination of β_2 yields

$$-2\beta_1^3 + 2a\beta_1 = b. \tag{2.9}$$

It is of course possible in principle to compute β_1 as any root of this polynomial, β_2 will then be computed as a ratio, i.e. $\beta_2 = \frac{b}{4\beta_1}$. So in principle this method has the ability to generate at least two different polynomials $P_\pm$. However there is a possibility for β_1 to vanish for some value of a and b. In such a case β_2 would be singular. It must be noticed that we have used such a method in our first numerical tests: indeed it revealed a singularity near $\alpha(x) \approx 0$. Another problem is the generalization to high order : indeed this procedure requires to compute exactly the roots of a high order polynomial which generalizes (2.9); this is not possible for orders $\geqslant 5$. This is why we do not use this method to compute the coefficients β_1 and β_2.

• Example : $q = 2$ and $I = 3$. Since one needs at least one more degree of freedom in the system to be solved we modify (2.8) and take into account β_3. The system becomes

$$\left\{ \begin{array}{l} 2\beta_2 + \beta_1^2 = a, \\ 6\beta_3 + 4\beta_1\beta_2 = b. \end{array} \right. \tag{2.10}$$

This system has 3 unknowns and 2 equations. So it has a priori an infinite number of solutions. Very fortunately a natural normalization condition arises, by considering that the two basis function should be linearly independent. To insure this we impose that $\frac{d}{dx} e^{P_+\left(x_{k+\frac{1}{2}}\right)} = 0 \Longleftrightarrow P'_+\left(x_{k+\frac{1}{2}}\right) = 0$ and $\frac{d}{dx} e^{P_+\left(x_{k+\frac{1}{2}}\right)} = 1 \Longleftrightarrow P'_+\left(x_{k+\frac{1}{2}}\right) = 1$. The first case corresponds to $\beta_1 = 0$ and the second one to $\beta_1 = 1$. With this second normalization it is evident that β_2 and β_3 can be computed explicitly from (2.10) and that the resulting formulas are just polynomial expressions with respect to all coefficients. One obtains two sets of coefficients which are $\beta_1^+ = 1$, $\beta_2^+ = \frac{1}{2}(a-1)$, $\beta_3^+ = \frac{1}{3}(b-2a+2)$ and $\beta_1^- = 0$, $\beta_2^- = \frac{a}{2}$, $\beta_3^- = \frac{1}{3}(b-2a)$. Notice that $\alpha_+ \neq \alpha_-$ since $\beta_1^+ \neq \beta_1^-$.

We use the method described in the last example at any order. The first thing is to chose a convenient degree I for any order q. In order to obtain an invertible system, one has to consider the first q terms in the left hand side of (2.7). As long as one considers the terms of degree less than or equal to $I-2$, the index of the coefficient from P'' is higher than the indexes of all the coefficients arising from $(P')^2$ terms. So that - as long as $I-2 \leqslant q-1$ - the computation of the $I-1$ coefficients $\{\beta_j\}_{2\leqslant j\leqslant I}$ is straightforward, given the coefficient β_1. Moreover if $I < q+1$ the terms of degree higher than $I-2$ in (2.7) will give an overdetermined system. For this reason the choice of Ps degree is set to be $I = q+1$.

In fact, in this case we solve the system of q equations with $q+1$ unknowns obtained identifying the first q coefficients in both parts of the expansion (2.7) with the normalization

$$\beta_1^+ = 0 \text{ which corresponds to } P'_+\left(x_{k+\frac{1}{2}}\right) = 0$$

and

$$\beta_1^- = 1 \text{ which corresponds to } P'_-\left(x_{k+\frac{1}{2}}\right) = 1.$$

The coefficients β_2^+, β_3^+, β_4^+, ..., and β_2^-, β_3^-, β_4^-, ..., are calculated one after the other using the formula deduced from the definition (2.7) for all $n \leqslant q-1$

$$(n+2)(n+1)\beta_{n+2}^{\pm} = \frac{d^n}{dx^n}\alpha\left(x_{k+\frac{1}{2}}\right) - \sum_{0\leqslant j,j'\leqslant I-1}^{j+j'=n}(j+1)(j'+1)\beta_{j+1}^{\pm}\beta_{j'+1}^{\pm}.$$

By construction $\varphi_+ = e^{P_+}$ and $\varphi_- = e^{P_-}$ are linearly independent functions. Once the polynomials P_+ and P_- have been constructed up to order q, we set

$$\alpha_+ = P''_+ + (P'_+)^2 \text{ and } \alpha_- = P''_- + (P'_-)^2. \tag{2.11}$$

By construction the first q coefficients of these polynomials coincide. But of course all other coefficients have no reason to be equal, so

$$\alpha_+ \neq \alpha_- \text{ in the general case.}$$

One can summarize as follows.

Lemma 2.1. *The functions $\alpha^{\pm}$ defined in (2.11) satisfy the following statements.*

1. *They are bounded independently from the cell number k, as well as all there derivatives.*
2. *If α is constant in the cell, then $\alpha^{\pm} = \alpha$, and the basis functions are classical plane waves.*
3. *By construction there exists a constant C_q such that*

$$\|\alpha^{\pm} - \alpha\|_{L^\infty(\Omega_k)} \leqslant C_q h^q$$

   where $h = |\Omega_k|$ and q is the order of the approximation.

4. *This construction is valid even if the sign of α changes.*

The last point is critical to be able to address the numerical approximation of the Airy equation.

Remark 2.1. *It is also possible to choose another normalization such as $\beta_1^{\pm} = \pm\sqrt{x_{k+1/2}}$. This choice will be illustrated as a numerical example in section 5.*

Remark 2.2. *Property 3 of lemma (2.1) establishes a property of approximation with respect to h. One of our numerical tests shows that a similar property of convergence holds with respect to the order parameter q. In practice q-convergence is a highly desirable property since it allows to use big cells.*

2.3 *Design of the basis functions in dimension two*

The fundamental equation that defines a generalized plane wave is (2.5). Description of all solutions of this equation and of effective procedures for the computation of such solutions seems difficult. Nevertheless one can rely on a simple linearization procedure in order to define a set of generalized plane waves with an order of approximation $q > 1$.

2.3.1 *Linear coefficients+rotation.* A first remark is that a simple procedure exists in the case of a linear coefficient

$$\alpha = a + bx + cy.$$

Up to a local rotation it is always possible to assume that $c = 0$. Assuming the local form

$$P(x,y) = p(x) + \lambda y$$

one ends up with the equation

$$p''(x) + p'(x)^2 = \alpha(x) - \lambda^2$$

for which the procedure described in the previous section is well adapted for the construction of a discrete space of approximation. Some details about the choice of λ will be provided in the numerical section 5.3. Our numerical tests (in the numerical section) show this procedure yields a high order method at any order.

2.3.2 *Linearization.* For a smooth coefficient α which is not necessarily linear, it is always possible to approximate it by a linear function, that is

$$\alpha(x,y) = a + b(x - x_G) + c(y - y_G) + O(h^2)$$

where $a = \alpha(x_G, y_G)$, $b = \partial_x \alpha(x_G, y_G)$, $c = \partial_y \alpha(x_G, y_G)$ and $|x - x_G| + |y - y_G| = h$. We can write

$$\alpha(x,y) = \alpha_G(x,y) + O(h^2), \qquad \alpha_G(x,y) = a + b(x - x_G) + c(y - y_G).$$

The approximation of α_G with the method described above yields a procedure a approximation of α by generalized plane waves.

3. UWVF and generalized plane waves

We now described the introduction of generalized plane waves in the UWVF method.

3.1 *Notation*

The function space for the UWV formulation is denoted V as

$$V = \prod_{k \in [\![1, N_h]\!]} L^2(\partial \Omega_k),$$

equipped with the Hermitian product $(x,y) = \sum_k \int_{\partial \Omega_k} \frac{1}{\gamma} x_k \overline{y_k}$. It defines a norm: $\|x\| = \sqrt{(x,x)}$. In particular for any operator $A \in \mathscr{L}(V)$, the norm is

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

Remark 3.1. *The space V depends on the mesh. Moreover: if $\Omega \subset \mathbb{R}$ the dimension of V is finite; if $\Omega \subset \mathbb{R}^d$ with $d \geqslant 2$, the dimension of V is infinite.*

3.2 *A standard ultra weak variational formulation*

The ultra weak variational formulation is a convenient reformulation of the initial problem. We need to define

$$H_k(\alpha) = \left\{ v_k \in H^1(\Omega_k), \left| \begin{array}{l} (-\Delta+\alpha)v_k = 0, (\Omega_k), \\ ((-\partial_\nu + i\gamma)v_k)_{|\partial\Omega_k} \in L^2(\partial\Omega_k) \end{array} \right. \right\} \tag{3.1}$$

and

$$H = \prod_{k=1}^{N_h} H_k(\alpha).$$

Theorem 3.1. *Let $u \in H^1(\Omega)$ be a solution of problem* (1.1) *such that $\partial_{\nu_k} u \in L^2(\partial\Omega_k)$ for any k. Let $\gamma > 0$ be a given real number. Then $x \in V$ defined by $x_{|\partial\Omega_k} = x_k$ with $x_k = ((-\partial_\nu + i\gamma)u_{|\Omega_k})_{|\partial\Omega_k}$ satisfies*

$$\begin{array}{l} \displaystyle\sum_k \left(\int_{\partial\Omega_k} \frac{1}{\gamma} x_k \overline{(-\partial_\nu + i\gamma)e_k} - \sum_{j,j\neq k} \int_{\Sigma_{kj}} \frac{1}{\gamma} x_j \overline{(\partial_\nu + i\gamma)e_k} \right) \\ \displaystyle - \sum_{k,\overline{\Gamma_k}\neq\emptyset} \int_{\Gamma_k} \frac{Q}{\gamma} x_k \overline{(\partial_\nu + i\gamma)e_k} = -2i \sum_k \int_{\partial\Omega_k} f\overline{e} + \sum_k \int_{\Gamma_k} \frac{1}{\gamma} g \overline{(\partial_\nu + i\gamma)e_k}, \end{array} \tag{3.2}$$

for any $e = (e_k)_{k\in[\![1,N_h]\!]} \in H$. Conversely, if $x \in V$ is solution of (3.2) *then the function u defined locally by*

$$\left\{ \begin{array}{l} u_{|\Omega_k} = u_k \in H^1(\Omega_k), \\ (-\Delta+\alpha)u_k = f_{|\Omega_k}, \\ (-\partial_{\nu_k} + i\gamma)u_k = x_k, \end{array} \right. \tag{3.3}$$

is the unique solution of the problem (1.1)*.*

This result is classical in the context of UWVF. We refer to Cessenat & Després (3); Buffa & Monk (1); Hiptmair et al. (13); Huttunen et al. (16, 17); **(author?)** (Imbert-Gérard & Després). Our main task is to adapt this formulation to the generalized plane waves developed previously. In order to give a more compact formulation useful for further developments, some definitions are required.

Definition 3.1. *For any $f \in L^2(\Omega)$, let E_f be the extension mapping defined by :*

$$E_f : \left\{ \begin{array}{ccc} V & \to & H, \\ z & \mapsto & e = (e_k)_{k\in[\![1,N_h]\!]}, \end{array} \right.$$

where e is defined $\forall k \in [\![1,N_h]\!]$ by the unique solution of the following problem :

$$\left\{ \begin{array}{ll} (-\Delta+\alpha)e_k & = f \quad (\Omega_k), \\ (-\partial_{\nu_k} + i\gamma)e_k & = z_k \quad (\partial\Omega_k). \end{array} \right.$$

Also define E which is the homogeneous extension operator with vanishing right hand side, namely $E = E_0$.

Notice that E_f is well defined thanks to theorem A.1.

Definition 3.2. *Let F be the mapping defined by*

$$F : \left\{ \begin{array}{ccc} V & \to & V, \\ z & \mapsto & \left((\partial_\nu + i\gamma)E(z)_{|\partial\Omega_k}\right)_{k\in[\![1,N_h]\!]}. \end{array} \right.$$

This operator relates the outgoing and ingoing traces on the boundaries $\partial\Omega_k$.

Definition 3.3. *Let Π be the mapping defined by*

$$\Pi : \left\{ \begin{array}{ccc} V & \to & V, \\ z_{|\Sigma_{kj}} & \mapsto & z_{|\Sigma_{jk}}, \\ z_{|\Gamma_k} & \mapsto & Q z_{|\Gamma_k}. \end{array} \right.$$

Definition 3.4. *If F^* denotes the adjoint operator of the operator F, let A be the operator $F^*\Pi$.*

With these notations the problem (3.2) is equivalent Cessenat & Després (3) to

$$\left\{ \begin{array}{l} \text{Find } x \in V \text{ such that } \forall y \in V \\ (x,y) - (\Pi x, Fy) = (b,y), \end{array} \right. \tag{3.4}$$

where the right hand side $b \in V$ is given by the Riesz theorem

$$(b,y) = -2i \int_\Omega f\overline{E(y)} + \int_\Gamma \frac{1}{\gamma} g \overline{F(y)}, \qquad \forall y \in V.$$

More precisely

- If u is solution of the initial problem (1.1) such that $\big((-\partial_\nu + i\gamma)u_{|\partial\Omega_k}\big)_{k\in[\![1,N_h]\!]} \in V$, then $x = \big((-\partial_\nu + i\gamma)u_{|\partial\Omega_k}\big)_{k\in[\![1,N_h]\!]}$ is solution in V of (3.4).
- Conversely if x is solution of (3.4) then $u = E_f(x)$ is the unique solution of (3.2). The problem (3.4) is equivalent to

$$\left\{ \begin{array}{l} \text{For } b \in V, \text{ find } x \in V \\ (I - A)x = b. \end{array} \right. \tag{3.5}$$

We now give some properties of the operators defined previously. They will be useful for the theoretical study of the method.

Lemma 3.1. *The operator Π obviously satisfies $\|\Pi\| \leqslant 1$ for any complex function Q such that $|Q| \leqslant 1$.*

Lemma 3.2. *The operator F is an isometry.*

Proof. For any $y \in V$, let $e \in H$ be $E(y)$. Then

$$\begin{aligned} \|Fy\|^2 &= \sum_{k\in[\![1,N_h]\!]} \int_{\partial\Omega_k} \frac{1}{\gamma} |(\partial_\nu + i\gamma)e_k|^2 = \sum_{k\in[\![1,N_h]\!]} \int_{\partial\Omega_k} \frac{1}{\gamma} |\partial_\nu e_k|^2 + \gamma |e_k|^2 + 2\Im(\partial_\nu e_k \cdot \overline{e}_k), \\ \|y\|^2 &= \sum_{k\in[\![1,N_h]\!]} \int_{\partial\Omega_k} \frac{1}{\gamma} |(-\partial_\nu + i\gamma)e_k|^2 = \sum_{k\in[\![1,N_h]\!]} \int_{\partial\Omega_k} \frac{1}{\gamma} |\partial_\nu e_k|^2 + \gamma |e_k|^2 - 2\Im(\partial_\nu e_k \cdot \overline{e}_k). \end{aligned}$$

Since $\int_{\partial\Omega_k} \partial_\nu e_k \cdot \overline{e}_k = \int_{\Omega_k} |\nabla e_k|^2 + \alpha |e_k|^2 \in \mathbb{R}$, one gets that $\|Fy\|^2 = \|y\|^2$. This implies the result. □

Proposition 3.1. *The operator A satisfies $\|A\| \leqslant 1$.*

This operator also satisfies the following property.

Proposition 3.2. *The operator $I-A$ is injective.*

Proof. Let $x \in V$ such that $(I-A)x = 0$, which means $x = F^*\Pi x$. Define $z \in V$ such that $z = \Pi x$, then $F^*z = x$ so that $\Pi F^*z = z$. Then define $u \in H$ such that for all $k \in [\![1, N_h]\!]$

$$\left\{ \begin{array}{ll} -\Delta u + \alpha u &= 0, \qquad (\Omega_k), \\ (\partial_\nu + i\gamma)u &= z_{|\partial\Omega_k}, \quad (\partial\Omega_k). \end{array} \right. \tag{3.6}$$

In order to identify F^*z, define $y \in V$ such that $\forall k \in [\![1, N_h]\!]$, $y_k = (-\partial_\nu + i\gamma)u_{|\Omega_k}$. It is known that $\forall v \in V$, there exists $w \in H$ such that $w = E(v)$, which means w satisfies

$$\left\{ \begin{array}{ll} -\Delta w + \alpha w &= 0, \qquad (\Omega_k), \\ (-\partial_\nu + i\gamma)w &= v_{|\partial\Omega_k}, \quad (\partial\Omega_k). \end{array} \right. \tag{3.7}$$

Then

$$\begin{aligned} (y,v) &= \sum_{k\in[\![1,N_h]\!]} \int_{\partial\Omega_k} \frac{1}{\gamma}(-\partial_\nu + i\gamma)u_{|\Omega_k} \cdot \overline{(-\partial_\nu + i\gamma)w_{|\partial\Omega_k}}, \\ &= \sum_{k\in[\![1,N_h]\!]} \int_{\partial\Omega_k} \frac{1}{\gamma}\partial_\nu u \cdot \partial_\nu \overline{w} + \gamma u \cdot \overline{w} + i\partial_\nu u \cdot \overline{w} - iu \cdot \partial_\nu \overline{w}, \\ (z,Fv) &= \sum_{k\in[\![1,N_h]\!]} \int_{\partial\Omega_k} \frac{1}{\gamma}(\partial_\nu + i\gamma)u_{|\Omega_k} \cdot \overline{(\partial_\nu + i\gamma)w_{|\partial\Omega_k}}, \\ &= \sum_{k\in[\![1,N_h]\!]} \int_{\partial\Omega_k} \frac{1}{\gamma}\partial_\nu u \cdot \partial_\nu \overline{w} + \gamma u \cdot \overline{w} - i\partial_\nu u \cdot \overline{w} + iu \cdot \partial_\nu \overline{w}. \end{aligned}$$

On the other hand, from (3.6) and (3.7) for all $k \in [\![1, N_h]\!]$

$$\left\{ \begin{array}{ll} \int_{\partial\Omega_k} \partial_\nu u \cdot \overline{w} &= \int_{\Omega_k} \nabla u \cdot \nabla \overline{w} + \int_{\Omega_k} \alpha u \cdot \overline{w}, \\ \int_{\partial\Omega_k} u \cdot \partial_\nu \overline{w} &= \int_{\Omega_k} \nabla u \cdot \nabla \overline{w} + \int_{\Omega_k} \alpha u \cdot \overline{w}, \end{array} \right.$$

so that $\int_{\partial\Omega_k} -\partial_\nu u \cdot \overline{w} + u \cdot \partial_\nu \overline{w} = 0$. As a consequence

$$\forall v \in V, (y,v) = (z,Fv),$$

which exactly means that $y = F^*z$. Since $\Pi F^*z = z$, it leads to $\Pi y = z$.

To conclude let's read this last equation in terms of the function u defined in (3.6).

$$\forall (k,j) \in [\![1, N_h]\!]^2, \left\{ \begin{array}{ll} (-\partial_\nu + i\gamma)u_{|\Sigma_{jk}} &= (\partial_\nu + i\gamma)u_{|\Sigma_{kj}}, \\ Q(-\partial_\nu + i\gamma)u_{|\Gamma_k} &= (\partial_\nu + i\gamma)u_{|\Gamma_k}, \end{array} \right.$$

so that both u and $\partial_\nu u$ are continuous along every interface Σ_{kj}, and now

$$\left\{ \begin{array}{ll} -\Delta u + \alpha u &= 0, \qquad (\Omega), \\ (\partial_\nu + i\gamma)u &= Q(-\partial_\nu + i\gamma)u, \quad (\partial\Omega). \end{array} \right.$$

Thanks to the preliminary result, u is the unique solution of the corresponding (1.1) problem : it is the 0 solution. Then $z = 0$, and so $x = 0$. The proof is ended. □

3.3 *An abstract discretization procedure*

The next step consists in the discretization of equation (3.4). This could be treated thanks to a standard Galerkin method which is presented below. That is we consider a subspace $V_h \subset V$ with finite dimension. We seek the discrete solution $x_h \in V_h$ such that

$$\forall y_h \in V_h, (x_h, y_h) - (\Pi x_h, F y_h) = (b, y_h). \tag{3.8}$$

The definition of the operator F, through the operator E, is linked to the functional space H ; this fact means that solutions of the homogeneous equation are needed then. In other words this abstract Galerkin procedure needs a companion constructive procedure to design the basis functions to generate V_h.

Before describing in the next section what is our proposition to make such a Galerkin method effective, we explain below why such a Galerkin approach (3.8) yields a well posed discrete problem. We provide here an analysis of this well known fact which is slightly different from what can be found in the literature Després (4); Cessenat & Després (3); Gittelson et al. (10); Buffa & Monk (1); Hiptmair et al. (13, 14).

Definition 3.5. *Let us define the norm* $|||v||| = \|(I-A)v\|$ *for all* $v \in V$*, and the bilinear form of the formulation (3.4):* $a(x,y) = (x,y) - (\Pi x, Fy)$.

Since $I-A$ is injective, $|||\cdot|||$ is indeed a norm. In the rest of this paper, $\mathscr{R}(z)$ (resp. I(z))stands for the real (imaginary) part of $z \in \mathbb{C}$. A fundamental property is

Lemma 3.3. *The bilinear form is coercive with respect to the norm* $|||\cdot|||$

$$|||x|||^2 \leqslant 2\mathscr{R}\,(a(x,x)) \quad \forall x \in V,$$

and is bicontinuous in the sense

$$|a(x,y)| \leqslant |||x||| \times \|y\| \quad \forall x, y \in V.$$

Proof. One has by definition $|||x|||^2 = \|x\|^2 + \|Ax\|^2 - 2\mathscr{R}(x,Ax)$. Since $\|A\| \leqslant 1$ then

$$|||x|||^2 \leqslant 2\left(\|x\|^2 - \mathscr{R}(x,Ax)_V\right) = 2\mathscr{R}\left((I-A)x,x\right)_V = 2\mathscr{R}a(x,x).$$

The coercivity is proved. The skewed bicontinuity is evident from Cauchy-Schwarz inequality applied to $a(x,y) = ((I-A)x,y)$. □

Proposition 3.3. *Assume there exists x solution of the problem (3.5). Then any discrete solution* x_h *satisfies the inequality*

$$|||x - x_h||| \leqslant 2 \inf_{z_h \in V_h} \|x - z_h\|. \tag{3.9}$$

Proof. By construction $a(x-x_h, y_h) = 0\ \forall y_h \in V_h$. So

$$a(x-x_h, x-x_h) = a(x-x_h, x-z_h) \quad \text{with } z_h = y_y - x_h.$$

It ends the proof with the coercivity and skewed bicontinuity of lemma 3.3. □

Lemma 3.4. *For all* $b \in V$*, the discrete solution* x_h *exists and is unique.*

Proof. If x_h exists, it is solution of a linear system, the dimension of the system being the dimension of the discrete subspace V_h. Therefore it is sufficient to check that if $a(x_h, y_h) = 0$ for all $y_h \in V_h$, then $x_h = 0$. Apply the inequality (3.9) with the choice $x = b = 0$. It yields $\|x_h\| \leqslant 2\inf_{z_h \in V_h} \|z_h\| = 0$. □

3.4 *The new method*

We consider that the generalized plane waves φ_k^l have been constructed with the procedure described previously. The local discrete space is

$$W_k = Span\left\{(-\partial_\nu + i\gamma)\varphi_k^l\right\}_{1\leqslant l\leqslant p(k)} \subset L^2(\partial\Omega_k).$$

The global discrete space $V^q \subset V$ is defined by : $V^q = \prod_{1\leqslant k\leqslant N_h} W_k$. Regarding these definitions, one sees that the basis functions are defined on the boundaries of the mesh, and that they have compact support. That is the shape function defined from φ_k^l has support in $L^2(\partial\Omega_k)$ and vanishes in $L^2(\partial\Omega_{k'})$ for $k' \neq k$. It is therefore convenient to define the trace $v_k^l \in V$ by

$$v_k^l = (-\partial_\nu + i\gamma)\varphi_k^l \text{ on } L^2(\partial\Omega_k), \quad \text{ and } v_k^l = 0 \text{ on } L^2(\partial\Omega_{k'}) \;\; k' \neq k.$$

An equivalent way to define W_k and V^q could be

$$W_k = Span(v_k^l)_{1\leqslant l\leqslant p(k)} \text{ and } V^q = Span(v_k^l)_{1\leqslant k\leqslant p(k),\ 1\leqslant p\leqslant N_h}.$$

Next define what are the generalizations of operators E and F in this context. Let $E^q \in \mathscr{L}(V^q, H)$ be the discrete mapping defined $\forall k \in [\![1, N_h]\!]$ and $\forall l \in [\![1, p(k)]\!]$ by

$$E^q(v_k^l) = \varphi_k^l \text{ on } H^1(\Omega_k), \quad \text{ and } v_k^l = 0 \text{ on } H^1(\Omega_{k'}) \;\; k' \neq k. \tag{3.10}$$

Similarly define $F^q \in \mathscr{L}(V^q, V)$, $\forall k \in [\![1, N_h]\!]$ and $\forall l \in [\![1, p(k)]\!]$, by

$$F^q(v_k^l) = (\partial_\nu + i\gamma)(\varphi_k^l) \text{ on } L^2(\partial\Omega_k), \quad \text{ and } v_k^l = 0 \text{ on } L^2(\partial\Omega_{k'}) \;\; k' \neq k.$$

With these notations and definitions, the abstract UWVF with generalized plane waves is defined as follows.

Definition 3.6. (UWVF method with generalized plane waves) *Find $x_h \in V^q$ such that*

$$\forall y_h \in V^q, (x_h, y_h)_V - (\Pi x_h, F^q y_h)_V = (b^q, y_h)_V \tag{3.11}$$

with the right hand side given by

$$(b^q, y_h)_V = -2i \int_\Omega f\overline{E^q(y_h)} + \int_\Gamma \frac{1}{\gamma} g\overline{F^q(y_h)}, \qquad \forall y_h \in V^q. \tag{3.12}$$

4. Numerical analysis of the method

In this section we desire to provide tools for the proof of the convergence of the discrete solution defined by (3.11) to the exact solution. Since the discrete method (3.11) can be viewed as a convenient modification of the bilinear form (3.8), it is not surprising that that the convergence analysis strongly relies on the second Strang's lemma as it is the case for non conforming finite element methods Farhat et al. (6). However the technicalities attached to ultra weak formulations are such that the convergence proof will be completed only in dimension one. This is due to the fact that some uniform coercivity properties which are part of the second Strang's lemma are easy to prove in dimension one, see proposition 4.2, but are open problems in higher dimension.

4.1 *Simplified notations in dimension one*

Let the order of approximation q be a given number. Assume that one has two polynomials $P_{k,1}$ and $P_{k,2}$ for all $k \in [\![1,N_h]\!]$. The corresponding basis functions and coefficients are denoted $\varphi_{k,1}$, $\alpha_{k,1}$ and $\varphi_{k,2}$, $\alpha_{k,2}$. For the sake of simplicity, the basis functions space will be now denoted by $\{\varphi_j\}_{j\in[\![1,2N_h]\!]}$ and the corresponding coefficients $\mathscr{D} = \{\alpha_j\}_{j\in[\![1,2N_h]\!]}$; $\{z_j\}_{j\in[\![1,2N_h]\!]}$ will denote the corresponding traces, i.e.

$$\forall j \in [\![1,2N_h]\!], z_j = \{(-\partial_\nu + i\gamma)\varphi_{j|\partial\Omega_k}\}_{k\in[\![1,N_h]\!]}.$$

The family $\{z_j\}_{j\in[\![1,2N_h]\!]}$ is a basis of the functional space V^q. A fundamental property is that

$$V^q = V \text{ only in dimension one}.$$

This will greatly reduce the technicalities of the proof. In fact, the lemmas 4.1 and 4.3 rely on the fact that in dimension one $dim V^q = 2$

4.2 *Preliminary results*

For the sake of completeness, here are classical results useful for the study of this new method. The proofs are postponed to the appendix.

Theorem 4.1. *Let $\mathscr{O}$ be a one-dimensional open interval with length h. Let w be the unique solution of*

$$\left\{\begin{array}{rl} -\Delta w + \alpha w &= 0, \quad (\mathscr{O}), \\ (-\partial_\nu + i\gamma)\, w &= g, \quad (\partial\mathscr{O}). \end{array}\right. \tag{4.1}$$

Then there exists two constants h_0 and C which depend of $\|\alpha\|_{L^\infty(\mathscr{O})}$ and γ such that $\forall h < h_0$

$$\|w\ \|_{L^2(\mathscr{O})} \leqslant C\sqrt{h}\,\|g\|_{L^2(\partial\mathscr{O})}\,, \tag{4.2}$$

Remark that the existence and uniqueness of the solution is given by theorem A.1.

We will also need a result on the approximation error between the problem

$$\left\{\begin{array}{rcl} -\Delta w + \alpha w &=& f, \quad (\mathscr{O}), \\ (-\partial_\nu + i\gamma)\, w &=& g, \quad (\partial\mathscr{O}), \end{array}\right. \tag{4.3}$$

and the modified problem

$$\left\{\begin{array}{rcl} -\Delta w + \alpha_h w &=& f, \quad (\mathscr{O}), \\ (-\partial_\nu + i\gamma)\, w &=& g, \quad (\partial\mathscr{O}), \end{array}\right. \tag{4.4}$$

where $\mathscr{O}$ represents any open set with length h included in Ω.

Theorem 4.2. *Let $\mathscr{O}$ be a one-dimensional open interval with length h. If u is solution of the problem* (4.3) *and u_h is solution of the problem* (4.4)*, then for small h there exists a constant C such that*

$$\|u - u_h\|_{L^2(\mathscr{O})} \leqslant C\left(h^{\frac{3}{2}}\|g\|_{L^2(\partial\mathscr{O})} + h^2\|f\|_{L^2(\mathscr{O})}\right)\|\alpha - \alpha_h\|_{L^\infty(\mathscr{O})}. \tag{4.5}$$

4.3 *The discrete problem*

This paragraph is devoted to showing that the operator F^q described in section 3.4 is an approximation of the operator F up to the order $q+1$ in h. Consider the following problem

$$\begin{cases} \text{Find } x_h \in V_q \text{ such that} \\ (I - A^q)x_h = b, \end{cases} \tag{4.6}$$

where $A^q = (F^q)^* \Pi$. Here h and q are given. This result relies on a preliminary lemma.

Lemma 4.1. *Let $q \geqslant 2$. Suppose h is small enough and basis functions are constructed as described in paragraph 2.2.2. For all $k \in [\![1, N_h]\!]$, there exists a constant C independent k such that $\forall z \in V^q$ and $\forall k \in [\![1, N_h]\!]$*

$$\sum_{j\in\{1,2\}} |x_j| \|z_j\|_{L^2(\partial\Omega_k)} \leqslant C \left\| \sum_{j\in\{1,2\}} x_j z_j \right\|_{L^2(\partial\Omega_k)}.$$

Proof. Set $k \in [\![1, N_h]\!]$ and $z = x_1 z_1 + x_2 z_2$. First x_j can be written as a function of z. This is a priori possible using $\{w_j\}_{j\in\{1,2\}}$ which is the dual basis of $\{z_j\}_{j\in\{1,2\}}$. For all $(j,l) \in \{1,2\}^2$, the dual function w_j is defined by

$$(w_j, z_l)_V = \delta_{jl}, \tag{4.7}$$

where δ denotes the Kronecker symbol. The proof proceeds in several steps.

First step. One has that $x_j = (z, w_j)_V$, therefore

$$\sum_{j\in\{1,2\}} |x_j| \|z_j\| \leqslant \left(\sum_{j\in\{1,2\}} \|z_j\| \|w_j\| \right) \|z\|.$$

So the claim is proved provided the term between parentheses can be estimated.

Second step: estimation of $\| \sum_{j\in\{1,2\}} \|z_j\| \|w_j\| \|$. From (4.7) it turns out that

$$\begin{aligned} w_1 &= \frac{-\|z_2\|^2}{|(z_1,z_2)|^2 - \|z_1\|^2\|z_2\|^2} z_1 + \frac{(z_1,z_2)}{|(z_1,z_2)|^2 - \|z_1\|^2\|z_2\|^2} z_2, \\ w_2 &= \frac{(z_1,z_2)}{|(z_1,z_2)|^2 - \|z_1\|^2\|z_2\|^2} z_1 - \frac{\|z_1\|^2}{|(z_1,z_2)|^2 - \|z_1\|^2\|z_2\|^2} z_2, \end{aligned}$$

so that

$$\sum_{j\in\{1,2\}} \|z_j\| \|w_j\| \leqslant 2 \frac{\|z_1\|^2 \|z_2\|^2}{\|z_1\|^2\|z_2\|^2 - |(z_1,z_2)|^2}.$$

Let us set for convenience $A = \frac{|(z_1,z_2)|}{\|z_1\|\|z_2\|}$ so that $\sum_{j\in\{1,2\}} \|z_j\| \|w_j\| \leqslant 2\frac{1}{1-A^2}$. It means that the whole proof relies on an upper bound for A.

Third step: end of the proof. By definition $(z_j)_{|\partial\Omega_k} = \left((-\partial_\nu + i\gamma)e^{P_j}\right)_{|\partial\Omega_k}$. By construction $P_j(x_{k+1/2}) = 0$ for $j = 1,2$, $P_1'(x_{k+1/2}) = 0$ and $P_2'(x_{k+1/2}) = 1$. Since by construction all derivatives of P_1 and P_2 are uniformly bounded, one has $P_j(x) = O(h)$ for $j = 1,2$, $P_1'(x) = O(h)$ and $P_2'(x) = 1 + O(h)$ when h goes to 0 and for all $x \in [x_k, x_{k+1}]$.

So one can estimate

$$||z_1||^2 = \frac{1}{\gamma}\left|-P_1'(x_{k+1}) + i\gamma P_1(x_{k+1})\right|^2 |\exp(P_1(x_{k+1}))|^2 + \frac{1}{\gamma}\left|P_1'(x_k) + i\gamma P_1(x_k)\right|^2 |\exp(P_1(x_k))|^2$$

$$= \frac{1}{\gamma}\left|-P_1'(x_{k+\frac{1}{2}}) + i\gamma P_1(x_{k+\frac{1}{2}})\right|^2 + \frac{1}{\gamma}\left|P_1'(x_{k+\frac{1}{2}}) + i\gamma P_1(x_{k+\frac{1}{2}})\right|^2 + O(h),$$

that is $||z_1||^2 = 2\gamma + O(h)$. With the same method one obtains

$$||z_2||^2 = 2\frac{1+\gamma^2}{\gamma} + O(h) = \frac{1+\gamma^2}{\gamma^2}2\gamma + O(h),$$

and

$$(z_1, z_2) = \frac{1}{\gamma}(-P_1'(x_{k+1/2}) + i\gamma)\overline{(-P_2'(x_{k+1/2}) + i\gamma)}$$

$$+\frac{1}{\gamma}(P_1'(x_{k+1/2}) + i\gamma)\overline{(P_2'(x_{k+1/2}) + i\gamma)} + O(h)$$

that is $(z_1, z_2) = 2\gamma + O(h)$. Therefore $A^2 = \frac{\gamma^2}{1+\gamma^2} + O(h)$. It proves the claim for h sufficiently small.

Final comment. By construction the polynomials designed in dimension one in section 2.2.2 by the approximation of the Taylor expansion (2.7) are such that all their coefficients are uniformly bounded up to order q for all cells in the domain. This is why the error $O(h)$ in the above analysis is uniform with respect to the cell index k, which is therefore not indicated. This is not true if one constructs the polynomials with the method constructed in the counter example (2.8).

□

Lemma 4.2. *For small h and considering the basis functions constructed as described in paragraph 2.2.2, there exists a constant C*

$$\|F^q - F\| \leqslant Ch^{q+1} \tag{4.8}$$

Proof.

For all $j \in [\![1, 2N_h]\!]$, the function φ_j is by construction $\varphi_j = E^q(z_j)$ such that

$$\varphi_j \in \{\varphi_l\}_{l\in[\![1,2N_h]\!]} \text{ satisfies } \forall k \in [\![1, N_h]\!] \begin{cases} z_j = (-\partial_v + i\gamma)\varphi_j, & (\partial\Omega_k), \\ \left(-\frac{d^2}{dx^2} + \alpha_j\right)\varphi_j = 0, & (\Omega_k). \end{cases}$$

We also define $\psi_j = E(z_j)$ such that and the equation with the exact coefficient α

$$\psi_j \in H \text{ satisfies } \forall k \in [\![1, N_h]\!] \begin{cases} z_j = (-\partial_v + i\gamma)\psi_j, & (\partial\Omega_k), \\ \left(-\frac{d^2}{dx^2} + \alpha\right)\psi_j = 0, & (\Omega_k). \end{cases}$$

Then

$$\begin{aligned} |(F^q - F)z_j|^2 &= |(\partial_v + i\gamma)(\varphi_j - \psi_j)|^2, \\ &= |(-\partial_v + i\gamma)(\varphi_j - \psi_j)|^2 + 2\Re(i\gamma(\varphi_j - \psi_j)\partial_v\overline{(\varphi_j - \psi_j)}), \\ &= -2\gamma\Im((\varphi_j - \psi_j)\partial_v\overline{(\varphi_j - \psi_j)}), \end{aligned}$$

since φ_j and ψ_j satisfy the same boundary condition: $(-\partial_\nu + i\gamma)(\varphi_j - \psi_j) = 0$. Then on the only one element where z_j is non zero numbered $k = k(j)$

$$\begin{aligned}\int_{\partial\Omega_k} \frac{1}{\gamma}|(F^q - F)z_j|^2 &= -2\Im \int_{\partial\Omega_k} (\varphi_j - \psi_j)\partial_\nu \overline{(\varphi_j - \psi_j)},\\ &= -2\Im \int_{\Omega_k} (\varphi_j - \psi_j)\frac{d^2}{dx^2}\overline{(\varphi_j - \psi_j)} - 2\Im \int_{\Omega_k} \left|\frac{d}{dx}(\varphi_j - \psi_j)\right|^2,\\ &\leqslant -2\Im \int_{\Omega_k} (\varphi_j - \psi_j)(\alpha_j\overline{\varphi_j} - \alpha\overline{\psi_j}),\end{aligned}$$

since both φ_j and ψ_j satisfy homogeneous equations. Then

$$\begin{aligned}\int_{\partial\Omega_k} \frac{1}{\gamma}|(F^q - F)z_j|^2 &\leqslant -\Im\left(\int_{\Omega_k} (\alpha_j + \alpha)|\varphi_j - \psi_j|^2 + \int_{\Omega_k} (\alpha_j - \alpha)(\varphi_j - \psi_j)\overline{(\varphi_j + \psi_j)}\right),\\ &\leqslant \|\alpha_j + \alpha\|_{L^\infty(\Omega_k)}\|\varphi_j - \psi_j\|^2_{L^2(\Omega_k)}\\ &\quad + \|\alpha_j - \alpha\|_{L^\infty(\Omega_k)}\|\varphi_j - \psi_j\|_{L^2(\Omega_k)}\left(\|\varphi_j\|_{L^2(\Omega_k)} + \|\psi_j\|_{L^2(\Omega_k)}\right),\end{aligned}$$

thanks to Cauchy-Schwarz inequality. On the other hand, from (4.2) and (4.5) for small hs

$$\begin{aligned}\|\varphi_j - \psi_j\|_{L^2(\Omega_k)} &\leqslant Ch^{\frac{3}{2}}\|z_j\|_{L^2(\partial\Omega_k)}\|\alpha - \alpha_j\|_{L^\infty(\Omega_k)},\\ \|\varphi_j\|_{L^2(\Omega_k)} &\leqslant C\sqrt{h}\|z_j\|_{L^2(\partial\Omega_k)},\\ \|\psi_j\|_{L^2(\Omega_k)} &\leqslant C\sqrt{h}\|z_j\|_{L^2(\partial\Omega_k)},\end{aligned}$$

and $\|\alpha_j + \alpha\|_{L^\infty(\Omega_k)}$ is bounded as noticed in remark 2.1. So for small h

$$\|(F^q - F)z_j\|^2_{L^2(\partial\Omega_k)} \leqslant C'h^2\|\alpha_j - \alpha\|^2_{L^\infty(\Omega_k)}\|z_j\|^2_{L^2(\partial\Omega_k)},$$

where still k denotes $k(j)$. Now for all $k \in [\![1, N_h]\!]$ let $L(k)$ be the set of indexes $j \in [\![1, 2N_h]\!]$ such that Ω_k is the support of z_j. Hence, for all $z \in V^q$ then $z_{|\partial\Omega_k} = \sum_{l\in L(k)} x_l z_l$ where both z_ls vanish on $\partial\Omega_j$ for all $j \neq k$, it yields

$$\begin{aligned}\|(F^q - F)z\|_{L^2(\partial\Omega_k)} &\leqslant \sum_{l\in L(k)} |x_l|\|(F^q - F)z_l\|_{L^2(\partial\Omega_k)}\\ &\leqslant Ch \max_{l\in L(k)} \|\alpha^l_k - \alpha\|_{L^\infty(\Omega_k)}\left(\sum_{l\in\{1,2\}} |x_l|\|z_l\|_{L^2(\partial\Omega_k)}\right).\end{aligned}$$

Thanks to lemma 4.1 it means that

$$\|(F^q - F)z\|_{L^2(\partial\Omega_k)} \leqslant \sqrt{C'}h \max_{l\in L(k)} \|\alpha^l_k - \alpha\|_{L^\infty(\Omega_k)}\|z\|_{L^2(\partial\Omega_k)}.$$

Going back to the definition of the V norm for all $z \in V$

$$\|(F^q - F)z\| \leqslant Ch \max_{j\in[\![1,2N_h]\!]} \|\alpha_j - \alpha\|_{L^\infty(\Omega_k)}\|z\|,$$

which exactly means $\|F^q - F\| \leqslant Ch \max_{j\in[\![1,2N_h]\!]} \|\alpha - \alpha_j\|_{L^\infty(\Omega_k)}$. The result then comes from equation (2.6) ensured by the construction of approximated coefficients α_js. □

4.4 *Some norms*

The whole point of this paragraph is to define a useful norm to adapt the second Strang lemma.

Lemma 4.3. *There exists a constant C such that for all $x \in V$: $Ch^{3/2}\|x\| \leqslant \|(I-A)x\|$.*

Remark that, in dimension one, the dimension of the space V is finite, so all the norms are equivalent; but the constants in the continuity inequalities depend on h, and this lemma specifies the dependence in this mesh parameter. *Proof.*

First step Take $x \in V$, and define $b = (I-A)x$. In order to interpret this equality in V, define $u = E(x)$ and $w = E(b)$, so that $(u,w) \in H \times H$ and

$$\forall k \in [\![1,N_h]\!] \left\{ \begin{array}{rll} \left(-\frac{d^2}{dx^2}+\alpha\right)u &= 0, & (\Omega_k), \\ (-\partial_\nu + i\gamma)u &= x_k, & (\partial\Omega_k), \end{array} \right.$$

$$\forall k \in [\![1,N_h]\!] \left\{ \begin{array}{rll} \left(-\frac{d^2}{dx^2}+\alpha\right)w &= 0, & (\Omega_k), \\ (-\partial_\nu + i\gamma)w &= b_k, & (\partial\Omega_k). \end{array} \right.$$

Since F is an isometry one has

$$Fx - \Pi x = Fb.$$

It means on every interface

$$\forall k \in [\![1,N_h]\!], \left\{ \begin{array}{ll} (-\partial_\nu + i\gamma)u_{|\Omega_k}(x_k) - \mathbf{1}_{k\neq 1}(-\partial_\nu + i\gamma)u_{|\Omega_{k-1}}(x_k) &= (-\partial_\nu + i\gamma)w_{|\Omega_k}(x_k), \\ (\partial_\nu + i\gamma)u_{|\Omega_k}(x_{k+1}) - \mathbf{1}_{k\neq N_h}(\partial_\nu + i\gamma)u_{|\Omega_{k+1}}(x_{k+1}) &= (\partial_\nu + i\gamma)w_{|\Omega_k}(x_{k+1}). \end{array} \right.$$

This leads to a system of jump conditions on the interfaces

$$\left\{ \begin{array}{l} (-\partial_\nu + i\gamma)u_{|\Omega_1}(x_1) = (-\partial_\nu + i\gamma)w_{|\Omega_1}(x_1), \\ \forall k \in [\![2,N_h]\!], \left| \begin{array}{l} \left(\frac{d}{dx}u_{|\Omega_{k-1}} - \frac{d}{dx}u_{|\Omega_k}\right)(x_k) = \frac{1}{2}\left((-\partial_\nu + i\gamma)w_{|\Omega_k} - (\partial_\nu + i\gamma)w_{|\Omega_{k-1}}\right)(x_k), \\ \left(u_{|\Omega_k} - u_{|\Omega_{k-1}}\right)(x_k) = \frac{1}{2i\gamma}\left((-\partial_\nu + i\gamma)w_{|\Omega_k} - (\partial_\nu + i\gamma)w_{|\Omega_{k-1}}\right)(x_k), \end{array} \right. \\ (\partial_\nu + i\gamma)u_{|\Omega_{N_h}}(x_{N_h+1}) = (\partial_\nu + i\gamma)w_{|\Omega_{N_h}}(x_{N_h+1}). \end{array} \right. \tag{4.9}$$

Considering U_0 and U_1 the two fundamental solutions of the homogeneous equation such that $\left(-\frac{d^2}{dx^2}+\alpha\right)u = 0$ on Ω, then u satisfies

$$\forall k \in [\![1,N_h]\!], u_{|\Omega_k} = \delta_0^k U_0 + \delta_1^k U_1, \tag{4.10}$$

where $(\delta_0^k, \delta_1^k)_{k\in[\![1,N_h]\!]}$ completely determine $u \in H$. Plugging (4.10) in (4.9), and defining

$$\left\{ \begin{array}{l} \lambda_0 = (-\partial_\nu + i\gamma)w_{|\Omega_1}(x_1), \\ \forall k \in [\![2,N_h]\!], \left| \begin{array}{l} \lambda_{k-1} = \frac{1}{2}\left((-\partial_\nu + i\gamma)w_{|\Omega_k} - (\partial_\nu + i\gamma)w_{|\Omega_{k-1}}\right)(x_k), \\ \mu_{k-1} = \frac{1}{2i\gamma}\left((-\partial_\nu + i\gamma)w_{|\Omega_k} - (\partial_\nu + i\gamma)w_{|\Omega_{k-1}}\right)(x_k), \end{array} \right. \\ \mu_{N_h} = (\partial_\nu + i\gamma)w_{|\Omega_{N_h}}(x_{N_h+1}), \end{array} \right.$$

then

$$\left\{\begin{array}{l|l} (-\partial_\nu+i\gamma)U_0(x_1)\delta_0^1+(-\partial_\nu+i\gamma)U_1(x_1)\delta_1^1=\lambda_0, & \\ \forall k\in[\![2,N_h]\!], & \frac{d}{dx}U_0(x_k)(\delta_0^{k-1}-\delta_0^k)+\frac{d}{dx}U_1(x_k)(\delta_1^{k-1}-\delta_1^k)=\lambda_{k-1}, \\ & U_0(x_k)(\delta_0^{k-1}-\delta_0^k)+\quad U_1(x_k)(\delta_1^{k-1}-\delta_1^k)=\mu_{k-1}, \\ (\partial_\nu+i\gamma)U_0(x_{N_h+1})\delta_0^{N_h}+(\partial_\nu+i\gamma)U_1(x_{N_h+1})\delta_1^{N_h}=\mu_{N_h}. & \end{array}\right. \tag{4.11}$$

Given the change of variable

$$\forall k\in[\![1,N_h-1]\!]\left\{\begin{array}{l} D_0^k=\delta_0^k-\delta_0^{k+1}, \\ D_1^k=\delta_1^k-\delta_1^{k+1}, \end{array}\right. \tag{4.12}$$

the system (4.11) gives a linear system with unknowns $(D_0^k,D_1^k)_{k\in[\![1,N_h-1]\!]}$. Defining the Wronskian $W_0=U_1\frac{d}{dx}U_0-U_0\frac{d}{dx}U_1$ - which is non zero - the solution is

$$\forall k\in[\![1,N_h-1]\!]\left\{\begin{array}{l} D_0^k=\dfrac{1}{W_0}\left(\lambda_kU_1(x_{k+1})-\mu_k\dfrac{d}{dx}U_1(x_{k+1})\right), \\ D_1^k=\dfrac{1}{W_0}\left(\mu_k\dfrac{d}{dx}U_0(x_{k+1})-\lambda_kU_0(x_{k+1})\right). \end{array}\right.$$

Then the structure of the system (4.11) is

$$\left\{\begin{array}{l} \alpha\delta_0^1+\beta\delta_1^1=\lambda_0, \\ \delta_0^k-\delta_0^{k+1}=D_0^k,\forall k\in[\![1,N_h-1]\!], \\ \delta_1^k-\delta_1^{k+1}=D_1^k,\forall k\in[\![1,N_h-1]\!], \\ \gamma\delta_0^{N_h}+\eta\delta_1^{N_h}=\mu_{N_h}. \end{array}\right.$$

Eliminating $(\delta_0^k,\delta_1^k)_{k\in[\![1,N_h-1]\!]}$ it yields

$$\left\{\begin{array}{l} \delta_0^1=\sum_{k=1}^{N_h-1}D_0^k+\delta_0^{N_h}, \\ \delta_1^1=\sum_{k=1}^{N_h-1}D_1^k+\delta_1^{N_h}, \end{array}\right.$$

and

$$\left\{\begin{array}{l} \alpha\delta_0^{N_h}+\beta\delta_1^{N_h}=L, \\ \gamma\delta_0^{N_h}+\eta\delta_1^{N_h}=\mu_{N_h}, \end{array}\right. \tag{4.13}$$

with

$$L=\lambda_0-(-\partial_\nu+i\gamma)U_0(a)\sum_{k=1}^{N_h-1}D_0^k-(-\partial_\nu+i\gamma)U_1(a)\sum_{k=1}^{N_h-1}D_1^k. \tag{4.14}$$

The determinant of the system (4.13) is $W_1=(-\partial_\nu+i\gamma)U_0(a)(\partial_\nu+i\gamma)U_1(b)-(\partial_\nu+i\gamma)U_0(b)(-\partial_\nu+i\gamma)U_1(a)$. If it were zero, then its columns would be linearly dependent, say $a_0C_1+a_1C_2=0$; this would mean $(\partial_\nu+i\gamma)(a_0U_0+a_1U_1)(x_1)=0$ and $(\partial_\nu+i\gamma)(a_0U_0+a_1U_1)(x_{N_h})=0$ so that $u=a_0U_0+a_1U_0$ would satisfy

$$\left\{\begin{array}{l} -u''+\alpha u=0, \\ (\partial_\nu+i\gamma)u=0. \end{array}\right.$$

Then u would be the unique solution (zero) of this last system, which is not possible since U_0 and U_1 are independent. Then W_1 is non zero. One finally obtains that

$$\left\{\begin{array}{l|l} \delta_0^{N_h} = \dfrac{1}{W_1}\Big(L(\partial_\nu + i\gamma)U_1(b) - \mu_{N_h}(-\partial_\nu + i\gamma)U_1(a)\Big), & \\ \delta_1^{N_h} = \dfrac{1}{W_1}\Big(\mu_{N_h}(-\partial_\nu + i\gamma)U_0(a) - L(-\partial_\nu + i\gamma)U_1(a)\Big), & \\ \forall k \in [\![1, N_h - 1]\!] & \begin{array}{l} \delta_0^k = \delta_0^{N_h} + \displaystyle\sum_{j=k}^{N_h-1} D_0^j, \\ \delta_1^k = \delta_1^{N_h} + \displaystyle\sum_{j=k}^{N_h-1} D_1^j. \end{array} \end{array}\right. \tag{4.15}$$

Now u is completely known.

Second step The next step is the estimation of the coefficients $(\delta_0^k, \delta_1^k)_{k\in[\![1,N_h]\!]}$ using (4.15). Since F is an isometry, and λ_k and μ_k are linear combinations of the components of Fb

$$\left\{\begin{array}{l} \forall k \in [\![0, N_h - 1]\!], |\lambda_k| \leqslant \sqrt{\gamma}\|b\|, \\ \forall k \in [\![1, N_h]\!], |\mu_k| \leqslant \dfrac{1}{\sqrt{\gamma}}\|b\|. \end{array}\right. \tag{4.16}$$

Thus from (4.12) and (4.16), with C depending on U_0, U_1, γ and W_0,

$$\left|\sum_{k=1}^{N_h-1} D_0^k\right| \leqslant CN_h\|b\|,, \quad \left|\sum_{k=1}^{N_h-1} D_1^k\right| \leqslant CN_h\|b\|.$$

From (4.14), $|L| \leqslant CN_h\|b\|$, and since $|\mu_{N_h}| \leqslant C\|b\|$ one has from (4.15)

$$\left|\delta_i^{N_h}\right| \leqslant CN_h\|b\|, \forall i \in \{0,1\},$$

and next for $k \in [\![1, N_h - 1]\!]$: $\left|\delta_i^k\right| \leqslant \left|\delta_i^{N_h}\right| + \displaystyle\sum_{k=1}^{N_h-1}\left|D_i^k\right| \leqslant CN_h\|b\|$. Then all δ terms satisfy $\left|\delta_i^k\right| \leqslant CN_h\|b\|$ for $i \in \{0,1\}$ and $k \in [\![1, N_h]\!]$.

End of the proof A last calculus leads to the following inequalities

$$\begin{aligned} \|x\|^2 &= \sum_{k\in[\![1,N_h]\!]} \left\|\delta_0^k(-\partial_\nu + i\gamma)U_0 + \delta_1^k(-\partial_\nu + i\gamma)U_1\right\|^2_{L^2(\partial\Omega_k)} \\ &\leqslant \sum_{k\in[\![1,N_h]\!]} \Big(2C(|\delta_0^k| + |\delta_1^k|)\Big)^2 \leqslant C \sum_{k\in[\![1,N_h]\!]} N_h^2\|b\|^2 \leqslant C\|b\|^2 N_h^3, \end{aligned}$$

so that $\|x\| \leqslant Ch^{-3/2}\|b\|$.

□

Definition 4.1. *Let us define $|||x|||_q = \|(I - A^q)x\|$ for all $x \in V$. This is a norm under the condition of the next proposition.*

Proposition 4.1. *Let $q \geqslant 2$ be given and let h be small enough. There exists a constant $C > 0$ such that*

$$Ch^{3/2}\|x\| \leqslant |||x|||_q, \forall x \in V. \tag{4.17}$$

Proof. One has

$$\begin{aligned} \forall x \in V, \quad \|(I-A)x\| &\leqslant \|(I-A^q)x\| + \|(A^q-A)x\| \\ &\leqslant \|(I-A^q)x\| + Ch^{q+1}\|x\|. \end{aligned}$$

So $\|(I-A)x\|_V - Ch^{q+1}\|x\| \leqslant \|(I-A^q)x\|, \forall x \in V$. Then lemma 4.3 concludes the proof since $h^{q+1} < h^{3/2}$ for h small enough. □

Proposition 4.2. *There exists a constant $h_1 > 0$ such that the bilinear form $a_q(x,y) = ((I-A^q)x,y)$ is uniformly coercive, i.e. $\forall h \leqslant h_1$*

$$|||x|||_q^2 \leqslant 3\mathscr{R}\left(a_q(x,x)\right), \forall x \in V.$$

Proof.

One has $|||x|||_q^2 \leqslant \|x\|^2 - 2\mathscr{R}\left(A_q x, x\right) + \|A_q x\|^2$. Since

$$\|A_q x\| \leqslant \|Ax\| + \|(A_q - A)x\| \leqslant (1 + Ch^{q+1})\|x\| \tag{4.18}$$

there exists another constant denoted as $C' > 0$ such that $\|A_q x\|^2 \leqslant (1 + C'h^{q+1})\|x\|^2$. Therefore

$$|||x|||_q^2 \leqslant 2\|x\|^2 + C'h^{q+1}\|x\|^2 - 2\mathscr{R}\left(A_q x, x\right),$$

that is $|||x|||_q^2 - C'h^{q+1}\|x\|^2 \leqslant 2\mathscr{R}\left(a_q(x,x)\right)$. For small h since $q > 3/2$ and due to the proposition 4.1 one has

$$C'h^q\|x\|^2 \leqslant C'h^{1/3}(h^{3/2} - h^q)\|x\|^2 \leqslant h^{1/3}|||x|||_q^2,$$

then

$$\frac{2}{3}|||x|||_q^2 \leqslant |||x|||_q^2 - C'h^q\|x\|^2.$$

Combined with the previous inequality it proves the claim. □

4.5 *Convergence*

The main convergence result is an adapted version of Strang second lemma with the $|||\cdot|||_q$ norm.

Theorem 4.3. *Suppose that $q \geqslant 2$ and $h \leqslant \min(h_0, h_1)$. Denote $x \in V$ the solution of the exact problem* (3.5) *in dimension one and $x_h \in V$ the solution of the discrete problem* (4.6)*. Then there exists a constant $C > 0$ such that*

$$|||x - x_h|||_q \leqslant Ch^{-3/2}\left(\inf_{y_h \in V} |||x - y_h|||_q + \sup_{w_h \in V - \{0\}} \frac{|a_q(x, w_h) - f_q(w_h)|}{\|w_h\|}\right), \tag{4.19}$$

where $f_q(y) = (b^q, y)_V$.

The proof relies on the following intermediate result already proved in (4.18).

Lemma 4.4. *The operator A^q satisfies $\|A^q\| \leqslant 1 + Ch^{q+1}$.*

Proof. Of theorem 4.3

- The first remark is the uniform coercivity with respect to $|||.|||_q$ needed in the second Strang lemma. It is proved in proposition 4.2.
- The second step consists in characterizing the uniform continuity of a_q. For all $(x,y) \in V^2$

$$\begin{aligned} |a_q(x,y)| &= |((I-A^q)x,y)|, \\ &\leqslant |||x|||_q \, \|y\|. \end{aligned}$$

Using (4.17) one has $\|w_h\| \leqslant Ch^{-3/2}|||w_h|||_q$ for some constant C, so that for small h

$$\forall (x,y) \in V^2, \ |a_q(x,y)| \leqslant Ch^{-3/2}|||x|||_q|||y|||_q.$$

- The last step is the inequality itself. The triangular inequality yields

$$|||x-x_h|||_q \leqslant |||x-y_h|||_q + |||x_h-y_h|||_q, \forall y_h \in V.$$

On the other hand proposition 3.3 shows that

$$\begin{aligned} \frac{1}{3}|||x_h-y_h|||_q^2 &\leqslant \left|a_q(x_h-y_h,x_h-y_h)\right|, \\ &\leqslant \left|a_q(x-y_h,x_h-y_h)\right| + \left|a_q(x-x_h,x_h-y_h)\right|, \\ &\leqslant Ch^{-3/2}|||x-y_h|||_q|||x_h-y_h|||_q + \left|a_q(x,x_h-y_h)-b_q(x_h-y_h)\right|. \end{aligned}$$

As $w_h = x_h - y_h \in V$, then

$$\frac{1}{3}|||x_h-y_h|||_q \leqslant Ch^{-3/2}|||x-y_h|||_q + \frac{|a_q(x,w_h)-b_q(w_h)|}{\|w_h\|}\frac{\|w_h\|}{|||w_h|||_q}.$$

Using one more time $\|w_h\| \leqslant Ch^{-3/2}|||w_h|||_q$, it yields the desired result.

□

We now have to estimate the error defined by

$$D_h(x,w_h) = |a_q(x,w_h) - b_q(w_h)|, \forall w_h \in V.$$

In order to simplify the proof and to match to the physical meaning of the problem, we will assume that the right hand side (3.12) is characterized by $f = 0$ and $g \in L^2(\Gamma) \subset V$.

Lemma 4.5. *There exists a constant $C > 0$ such that*

$$\forall w_h \in V - \{0\}, \ \frac{D_h(x,w_h)}{\|w_h\|} \leqslant Ch^{q+1}\left(\|x\| + \|g\|\right). \tag{4.20}$$

Proof.

$$\begin{aligned} \forall w_h \in V-\{0\}, \ D_h(x,w_h) &= |((I-A^q)x,w_h)_V - (b,w_h)_V|, \\ &\leqslant |((A-A^q)x,w_h)_V| + |((I-A)x,w_h)_V - (b,w_h)_V| + \left|(b-b_q,w_h)\right|, \\ &\leqslant Ch^{q+1}\|x\| \, \|w_h\| + Ch^{q+1}\|g\| \, \|w_h\| \end{aligned}$$

since the second term vanishes $(I-A)x = b$. The third term is bounded using (3.12) like $\left|(b-b_q,w_h)\right| \leqslant C\|F-F_q\| \, \|g\| \, \|w_h\| \leqslant Ch^{q+1}\|g\| \, \|w_h\|$. This gives exactly (4.20). □

It is now easy to prove the theoretical convergence of the method in dimension one.

Theorem 4.4. *One has the estimation*

$$|||x-x_h|||_q = O(h^{q-1/2}). \tag{4.21}$$

Proof. In dimension one the discrete space of approximation is equal to V whatever the method of construction of basis functions is. This is why one can choose $y_h = x$ in (4.19). So $\inf_{y_h\in V} |||x-y_h|||_q = 0$. The remaining term is bounded with (4.20). □

It is useful to rewrite this inequality using a norm with the usual scaling

$$\overline{\|z\|} = \sqrt{\sum_{k\in[\![1,N_h]\!]} h|z_k|^2}.$$

By construction $\overline{\|z\|} = h^{\frac{1}{2}}\|z\|$. Using (4.17) one gets $\overline{\|z\|} \leqslant Ch^{-1}|||z|||_q$. Therefore a corollary of the theorem is the estimate of convergence

$$\overline{\|x-x_h\|} = O(h^{q-3/2}). \tag{4.22}$$

5. Numerical examples

All the following examples are linked with Airy functions since it is the physical problem (1.3)-(1.4) we are interested in. We only consider here coefficients $\beta(x) = x$ and $\beta(x,y) = x$, so that in dimension one as in dimension two that Airy functions are exact solutions of the equation. All the linear systems are assembled and solved with Matlab.

The parameter γ is set to be constant equal to 1 everywhere.

5.1 *One dimensional test case*

The test problem considered here is the following : on an interval $\Omega =]a,b[\subset\mathbb{R}$

$$\begin{cases} -u''(x) + x\,u(x) &= 0, & (]a,b[),\\ (\partial_\nu + i\gamma)u(x) &= (\partial_\nu + i\gamma)Ai(x), & (\{a,b\}), \end{cases}$$

The points of the uniform mesh are denoted $\{x_k\}_{k\in[\![1,N_h+1]\!]}$, where N_h stands for the number of elements defining the mesh. For a given value of q the basis functions are designed as in paragraph 2.2.2. The solution computed corresponds to an element $x_h \in V$. A simple formula to expresses the traces of u_h in funcrion of x_h

$$\begin{cases} 2i\gamma u_h = (I+\Pi)x_h + g & (\{a,b\}),\\ 2i\gamma u_h = (I+\Pi)x_h & (\{x_k\}_{k\in[\![2,N_h]\!]}). \end{cases}$$

In all simulations, the accuracy is reported using a discrete l^2 norm so that the relative error is computed as

$$\frac{\sqrt{\sum_{k\in[\![1,N_h+1]\!]} |u_{ex}(x_k) - u_h(x_k)|^2}}{\sqrt{\sum_{k\in[\![1,N_h+1]\!]} |u_{ex}(x_k)|^2}}.$$

Considering the domain $\Omega =]-5,5[$, one gets the the typical result of figure 2 where we plot the exact analytical Airy function and the numerical solution computed with our method. The rates of convergence are described in figures 1 and 3. The numerical rates of convergence are better than the theoretical estimates.

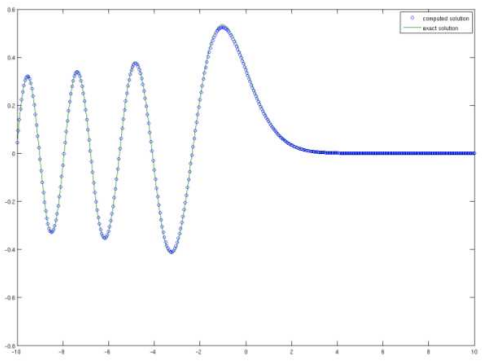


FIG. 2: Plot of the analytical Airy function, and comparison with the numerical solution. Here we used a large (≈ 200) number of cells and two high order generaalized plane waves per cell. One clearly distinguishes between the propagative medium $x < 0$ and the non propagative medium $x > 0$.

TABLE 1. *Errors and orders of convergence for different orders of approximation q depending on the number of unknowns N.*

	q=2		q=3		q=4		q=5		q=6	
l	Error	Rate	Error	Rate	Error	Rate	Error	Rate	Error	Rate
4	9.5e-01	-	9.9e-01	-	8.6e-01	-	8.6e-01	-	NaN	-
8	9.2e-01	-0.05	9.7e-01	-0.03	9.7e-01	0.18	9.9e-01	0.20	9.9e-01	NaN
16	7.8e-01	-0.23	9.5e-01	-0.03	9.2e-01	-0.09	9.6e-01	-0.04	9.4e-01	-0.04
32	6.0e-01	-0.39	3.3e-01	-1.51	2.5e-01	-1.89	1.5e-01	-2.65	1.1e-01	-3.14
64	2.0e-01	-1.59	3.2e-02	-3.4	2.0e-02	-3.61	3.2e-03	-5.6	2.0e-03	-5.75
128	5.4e-02	-1.89	2.1e-03	-3.91	1.3e-03	-3.93	5.2e-05	-5.94	3.2e-05	-5.96
256	1.4e-02	-1.97	1.3e-04	-3.98	8.4e-05	-3.98	8.2e-07	-5.99	5.0e-07	-5.99
512	3.4e-03	-1.99	8.3e-06	-4.00	5.3e-06	-4.00	1.3e-08	-6.00	7.9e-09	-6.00
1024	8.6e-04	-2.00	5.2e-07	-4.00	3.3e-07	-4.00	2.0e-10	-6.00	1.2e-10	-6.00
2048	2.2e-04	-2.00	3.3e-08	-4.00	2.1e-08	-4.00	3.1e-12	-5.99	1.9e-12	-6.00
4096	5.4e-05	-2.00	2.0e-09	-4.00	1.3e-09	-4.00	7.3e-14	-5.43	7.5e-14	-4.69
8192	1.3e-05	-2.00	1.3e-10	-4.00	8.1e-11	-4.00	1.6e-14	-2.21	5.8e-14	-0.37
16384	3.4e-06	-2.00	7.9e-12	-4.01	5.0e-12	-4.01	5.0e-14	1.67	5.0e-14	-0.20

One can also notice that we observe better convergence rates for odd values of q compared to even values. We have no explanation for the moment.

One can see that on the finest meshes the solution is accurate to machine precision for the highest values of parameter q.

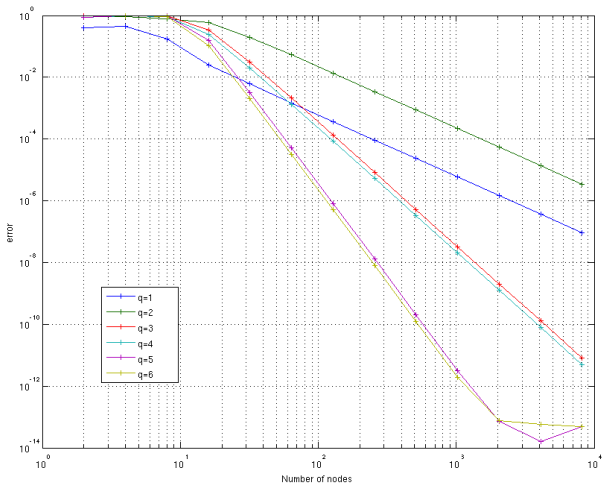


FIG. 3: Convergence of the method increasing the parameter q, relative discrete L^2 error as a function of the number of elements defining the mesh.

5.2 *About q convergence*

In figure 3, when the number of nodes is fixed, the error decreases when the parameter $q \geqslant 2$ increases.To obtain better understanding of this phenomenon, we plot in figure 4 for different values of q and around two points x_0 the Airy function and its approximations thanks to the two basis functions φ constructed in section 2.2.2.

We observe that the approximation is uniform in $]x_0-\varepsilon, x_0+\varepsilon[$, with ε independent of q.

5.3 *Two dimensional test case*

A first test case in dimension two is presented here. Consider an open set $\Omega \subset \mathbb{R}^2$ and the following simple problem

$$\left\{\begin{array}{rll} -\Delta u(x,y)+x\,u(x,y) & = 0, & (\Omega), \\ (\partial_\nu + i\gamma)u(x,y) & = (\partial_\nu + i\gamma)Ai(x), & (\partial\Omega), \end{array}\right.$$

so that the exact solution is again the Airy function Ai. The domain considered here is square and meshed with regular triangles. A comparison between the exact solution and the numerical solution is diplayed in figure 5.

As explained in section 2.3.1, the design of basis functions is easy in the case of a coefficient depending on only one coordinate, performing a one dimension reduction. The basis function φ is defined by $\varphi(x,y) = e^{P(x,y)}$ with $P(x,y) = p(x) + \lambda y$ where λ still has to be defined. Here we chose

$$\lambda \in \left\{ i \sin\left(\frac{2\pi k}{3}\right), k \in [\![1,3]\!] \right\}.$$

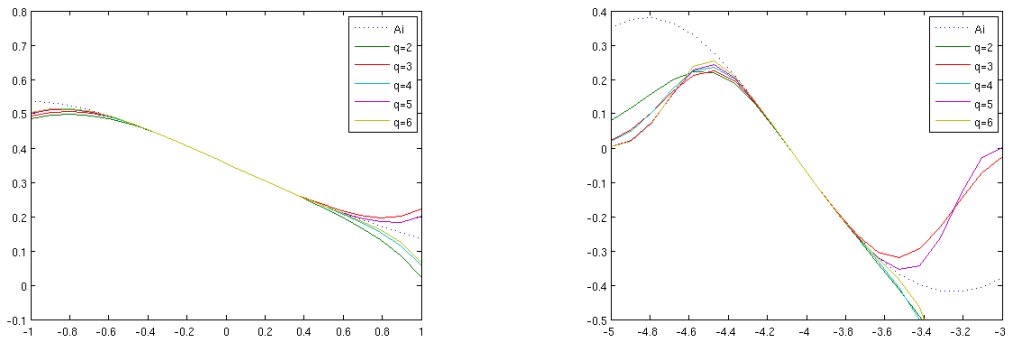


FIG. 4: Approximation of Airy function by corresponding basis functions for different values of q, in the vicinity of $x_0 = 0$ and $x_0 = -4$.

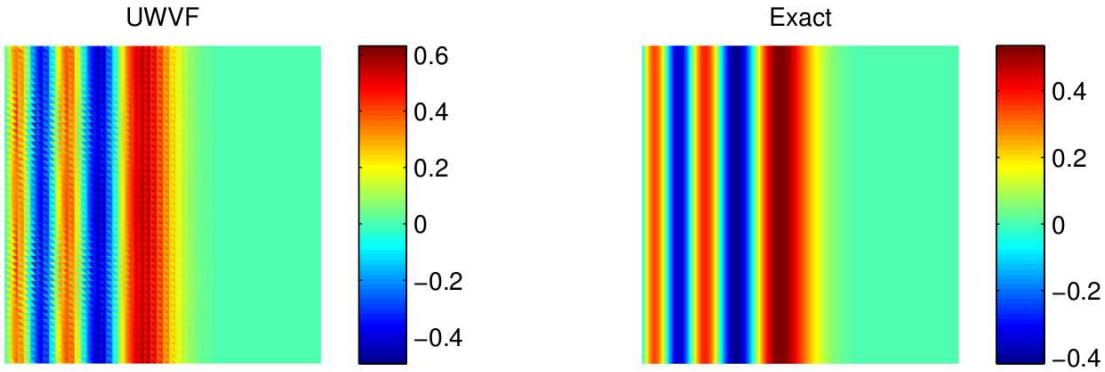
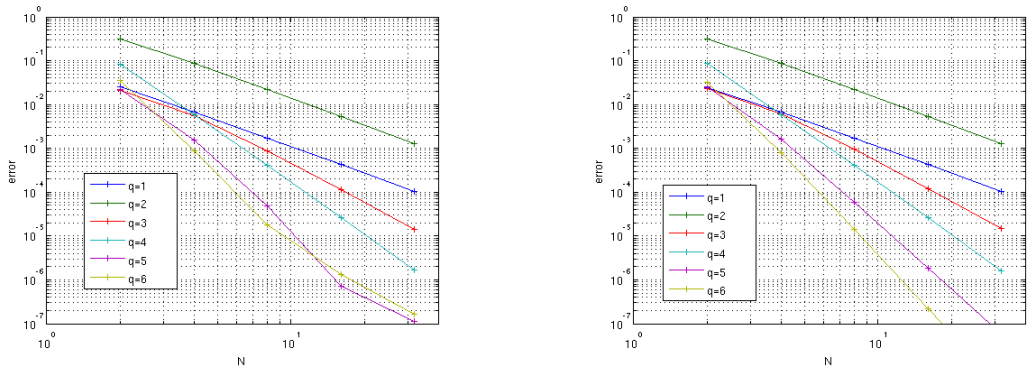


FIG. 5: Comparison between the exact Airy function on the right and the numerical solution computed with 6 basis function per element on the left. The numerical solution is interpolated on a finer mesh. Here the error is 0.0256%. The tables of errors 2 and 3 show high order convergence.



(a) With Simpson quadrature formulas.

(b) With Boole quadrature formulas.

FIG. 6: First test case in dimension two, with triangular mesh, on the square $]-1,1[\times]-1,1[$, with N nodes on each edge of the square.

TABLE 2. *Errors and orders of convergence depending on the number of unknowns N for the two dimensional case with Simpson quadrature formulas.*

	q=2		q=3		q=4		q=5		q=6	
l	Error	Rate	Error	Rate	Error	Rate	Error	Rate	Error	Rate
48	3.1e-01	-	2.1e-02	-	8.4e-02	-	2.2e-02	-	3.5e-02	-
192	8.6e-02	-1.84	5.6e-03	-1.92	5.9e-03	-3.83	1.6e-03	-3.85	8.9e-04	-5.3
768	2.2e-02	-1.98	8.9e-04	-2.67	4.1e-04	-3.85	5.0e-05	-4.97	1.8e-05	-5.62
3072	5.3e-03	-2.04	1.1e-04	-2.96	2.6e-05	-3.96	7.1e-07	-6.12	1.3e-06	-3.74
12288	1.3e-03	-2.04	1.4e-05	-3.01	1.6e-06	-4.00	1.1e-07	-2.67	1.7e-07	-3.03

For each λ the corresponding functions p_+ and p_- are constructed as in the one dimensional case since φ being a solution of the homogeneous equation means that $-\left(e^{p(x)}\right)'' + (x-\lambda^2)e^{p(x)} = 0$.

The other difference with the one dimensional case is the numerical estimation of boundary integrals. It requires numerical quadrature. The quadrature is performed with a given number of points with either Simpson or Boole method. The corresponding results are given in figures 6, 2 and 3. One can observe a clear improvement in the results obtained using Boole formulas compared to the results obtained using Simpson formulas.

5.4 *Other basis functions*

Figures 7 and 4 present the numerical convergence results obtained with basis functions designed with the normalization $\beta_{l,\pm} = \pm\sqrt{\alpha(x_{k+1/2})}$. Comparing to figures 1 and 3, one can see that the convergence rate is not modified by this new choice, however for a given number of mesh elements the error is smaller when the method is constructed with these new basis functions than with the basis functions described in section 2.2.2. In fact, for a given order q, the numerical results show that the constant underlying

TABLE 3. *Errors and orders of convergence depending on the number of unknowns N for the two dimensional case with Simpson quadrature formulas.*

	q=2		q=3		q=4		q=5		q=6	
l	Error	Rate	Error	Rate	Error	Rate	Error	Rate	Error	Rate
48	3.2e-01	-	2.3e-02	-	8.5e-02	-	2.5e-02	-	3.2e-02	-
192	8.7e-02	-1.86	6.2e-03	-1.92	5.9e-03	-3.86	1.6e-03	-3.94	8.0e-04	-5.32
768	2.2e-02	-1.98	9.5e-04	-2.69	4.0e-04	-3.86	5.9e-05	-4.78	1.4e-05	-5.85
3072	5.3e-03	-2.04	1.2e-04	-2.97	2.6e-05	-3.96	1.9e-06	-4.99	2.2e-07	-5.98
12288	1.3e-03	-2.04	1.5e-05	-3.01	1.6e-06	-4.00	5.7e-08	-5.02	3.4e-09	-6.00

TABLE 4. *Errors and orders of convergence depending on the number of unknowns N for the two dimensional case with Simpson quadrature formulas.*

	q=2		q=3		q=4		q=5		q=6	
l	Error	Rate	Error	Rate	Error	Rate	Error	Rate	Error	Rate
16	1.9e-01	-1.92	3.9e-02	-3.69	4.7e-02	-5.65	5.4e-03	-7.07	2.0e-02	-5.19
32	6.2e-02	-1.64	2.9e-03	-3.75	4.2e-03	-3.48	1.4e-04	-5.28	4.2e-04	-5.54
64	1.6e-02	-1.93	1.9e-04	-3.95	2.8e-04	-3.92	2.4e-06	-5.86	6.9e-06	-5.93
128	4.2e-03	-1.98	1.2e-05	-3.99	1.8e-05	-3.98	3.8e-08	-5.97	1.1e-07	-5.98
256	1.0e-03	-1.99	7.4e-07	-4.00	1.1e-06	-3.99	6.0e-10	-5.99	1.7e-09	-6.00
512	2.6e-04	-2.00	4.6e-08	-4.00	7.0e-08	-4.00	9.4e-12	-6.00	2.7e-11	-6.00
1024	6.5e-05	-2.00	2.9e-09	-4.00	4.4e-09	-4.00	1.6e-13	-5.92	4.3e-13	-5.98
2048	1.6e-05	-2.00	1.8e-10	-4.00	2.7e-10	-4.00	9.8e-15	-3.99	1.4e-14	-4.95
4096	4.1e-06	-2.00	1.1e-11	-4.00	1.7e-11	-4.00	2.4e-14	1.28	1.9e-14	0.42
8192	1.0e-06	-2.00	7.5e-13	-3.90	1.0e-12	-4.06	1.3e-13	2.43	1.4e-13	2.88
16384	2.6e-07	-2.00	2.1e-13	-1.84	2.0e-13	-2.32	2.1e-13	0.73	2.1e-13	0.61

in estimation (4.22) is much better : for a given number of mesh elements the numerical error can be improved by a factor $\approx 10^2$. Once again the only difference between these two different choices of basis functions relies on the fact that the leading coefficient in $P_{\pm}$ does depend or not on the coefficient α. The theoretical tools that developed previously can be adapted without difficulty to this new family of basis functions but the vertical shift visible in figures 3 to 7 will require more research to be fully understood.

As in dimension one, one can see that with Simpson method on the finest meshes the errors increase for $q = 5$ and 6. It is also supposedly linked with machine precision.

6. Perspectives

The method proposed in this work has been designed in any dimension : its goal is to increase the accuracy of any plane wave methods in the case the coefficients of the equation are smooth. Preliminary tests in dimensions one and two for h-convergence assess effective gain in accuracy. The evolution of this method may be pursued in at least two directions, numerical and theoretical.

Our first interest is to validate the method on challenging test problems inspired by the physics of reflectometry : in particular we have in mind to use non uniform meshes to reduce the computational burden for problems such that $|\Omega|^{1/d} >> \lambda$ where $|\Omega|$ is the size of the domain and λ the characteristic

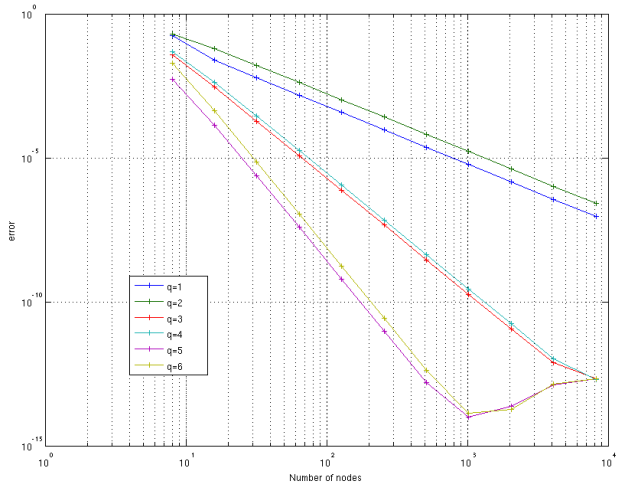


FIG. 7: Relative discrete L^2 error as a function of the number of elements defining the mesh, using the normalization $\beta_{1,\pm} = \pm\sqrt{\alpha(x_{k+1/2})}$. Different curves correspond to increasing order parameter q.

wave length of the wave. The design of new families of generalized plane waves with increased or even better optimal accuracy is of course of major interest for practical applications.

The numerical analysis of the method is also rich of new theoretical questions in dimension higher than two. We distinguish two particular problems. A first problem is to determine the best choice for $\frac{\partial P}{\partial x}(G_j)$ and $\frac{\partial P}{\partial y}(G_j)$ where G_j is the center of mass of the cell and $\varphi = e^P$ is a basis function; this problem corresponds to the choice of β_1 described in section 2.2.2. A second problem, more fundamental from a theoretical perspective, is the generalization in two dimension of the inverse inequalities like those of lemma 4.3 and proposition 4.1. The main difficulty stems from the fact that V is finite dimensional in dimension one and has infinite dimension in higher dimension : as a consequence $V^q \neq V$ in dimension two and more. Up to this difference we think nevertheless that the functional setting that we have developed, which is based on the second Strang's lemma, is still convenient in dimension higher than two. In this context it could be worthwhile to make the connection with the Discontinuous Galerkin formalism developed for example in Huttunen et al. (16).

A. Appendix

For the sake of completeness of this work, we review some very classical results needed for the proof of the convergence of our algorithm.

A.1 *On the initial problem*

It concerns the solution of the system in bounded domains

$$\left\{ \begin{array}{rcll} -\Delta u + \alpha u & = & f, & x \in \Omega, \\ (\partial_\nu + i\gamma) u & = & Q(-\partial_\nu + i\gamma) u + g, & x \in \Gamma. \end{array} \right.$$

It is necessary to assume that the regularity of the boundary of Ω is sufficient so that a unique continuation principle holds. We do not want to discuss it because it is not in the scope of this work. We refer the reader to Monk (23) p.92. Moreover, for classical results on polygonal domains we refer to Grisvard (11). To simplify here Q is constant. A proof can be found in **(author?)** (Imbert-Gérard & Després).

Theorem A.1. *Let Ω be a bounded domain in $\mathbb{R}^2$ with a Lipschitz and piecewise $\mathscr{C}^2$ boundary Γ. Let $f \in L^2(\Omega)$, $g \in L^2(\Gamma)$ and $\zeta \in \mathbb{C}$ such that $\Re(\zeta) \neq 0$. Then there exists a unique solution $u \in H^1(\Omega)$ to the variational formulation*

$$\int_\Omega \nabla u \cdot \overline{\nabla v} + \int_\Omega \alpha u \overline{v} + i\zeta \int_\Gamma u\overline{v} = \int_\Omega f\overline{v} + \int_\Gamma g\overline{v}, \quad \forall v \in H^1(\Omega).$$

Using the notations of (A.1) *with $|Q| < 1$, then $\Re(\frac{1-Q}{1+Q}) \neq 0$ so there exists a unique solution $u \in H^1$ to* (1.1)*, i.e. such that*

$$\int_\Omega \nabla u \cdot \overline{\nabla v} + \int_\Omega \alpha u \overline{v} + i\frac{1-Q}{1+Q}\gamma \int_\Gamma u\overline{v} = \int_\Omega f\overline{v} + \frac{1}{1+Q}\int_\Gamma g\overline{v}, \quad \forall v \in H^1(\Omega).$$

Remark A.1. *This result can be generalized to the case where $|Q| \leqslant 1$ almost everywhere on Γ and $|Q| < 1$ on a smooth part of Γ which length is non zero.*

A.2 *Proof of inequality* (4.2)

We need a very classical Poincaré inequality in one dimension.

Proposition A.1. *There exists a constant C such that for all $h > 0$, all open interval $\mathscr{O} \subset \mathbb{R}$, for all $u \in L(\mathscr{O})$*

$$\|u\|_{L^2(\mathscr{O})} \leqslant C\left(\sqrt{h}\|u\|_{L^2(\partial\mathscr{O})} + h\|u'\|_{L^2(\mathscr{O})}\right) \tag{A.1}$$

Proof. There exists $a \in \mathbb{R}$ such that $\mathscr{O} =]a, a+h[$. From $u(x) = u(a) + \int_a^x u'(t)dt$ it yields $\int_a^{a+h} |u(x)|^2 dx \leqslant 2h|u(a)|^2 + 2\int_a^{a+h} \left(\int_a^x |u'(t)|dt\right)^2 dx$, so that $\|u\|_{L^2(\mathscr{O})} \leqslant \sqrt{2h}\|u\|_{L^2(\partial\mathscr{O})} + \sqrt{2}h\|u'\|_{L^2(\mathscr{O})}$. It gives the result for $C = \sqrt{2}$. □

Proof. We will show a more general inequality than (4.2) . We use u as test function in the variational formulation (A.1) corresponding to the following problem

$$\left\{ \begin{array}{ll} -u'' + \beta u = f, & (\mathscr{O}) \\ (-\partial_\nu + i\gamma)u = g, & (\partial\mathscr{O}). \end{array} \right. \tag{A.2}$$

One gets

$$\int_{\mathscr{O}} |u'|^2 + i\gamma \int_{\partial\mathscr{O}} |u|^2 = \int_{\mathscr{O}} f\overline{u} - \int_{\mathscr{O}} \beta|u|^2 + \int_{\partial\mathscr{O}} g\overline{u}.$$

We obtain

$$\begin{cases} \|u\|^2_{L^2(\partial\mathscr{O})} \leqslant \dfrac{1}{\gamma}\|f\|_{L^2(\mathscr{O})}\|u\|_{L^2(\mathscr{O})} + \dfrac{1}{\gamma}\|g\|_{L^2(\partial\mathscr{O})}\|u\|_{L^2(\partial\mathscr{O})}, \\ \|u'\|^2_{L^2(\mathscr{O})} \leqslant \|g\|_{L^2(\partial\mathscr{O})}\|u\|_{L^2(\partial\mathscr{O})} + \|\beta\|_{L^\infty(\mathscr{O})}\|u\|^2_{L^2(\mathscr{O})} + \|f\|_{L^2(\mathscr{O})}\|u\|_{L^2(\mathscr{O})}. \end{cases}$$

The first inequality yields

$$\|u\|^2_{L^2(\partial\mathscr{O})} \leqslant \frac{2}{\gamma}\|f\|_{L^2(\mathscr{O})}\|u\|_{L^2(\mathscr{O})} + \frac{1}{\gamma^2}\|g\|^2_{L^2(\partial\mathscr{O})}.$$

A standard inequality yields

$$\|g\|_{L^2(\partial\mathscr{O})}\|u\|_{L^2(\partial\mathscr{O})} \leqslant \frac{1}{2\gamma}\|g\|^2_{L^2(\partial\mathscr{O})} + \frac{\gamma}{2}\|u\|^2_{L^2(\partial\mathscr{O})}$$

$$\leqslant \frac{1}{2\gamma}\|g\|^2_{L^2(\partial\mathscr{O})} + \|f\|_{L^2(\mathscr{O})}\|u\|_{L^2(\mathscr{O})} + \frac{1}{2\gamma}\|g\|^2_{L^2(\partial\mathscr{O})}.$$

Inserting in the second inequality we obtain

$$\|u'\|^2_{L^2(\mathscr{O})} \leqslant \frac{1}{\gamma}\|g\|^2_{L^2(\partial\mathscr{O})} + 2\|f\|_{L^2(\mathscr{O})}\|u\|_{L^2(\mathscr{O})} + \|\beta\|_{L^\infty(\mathscr{O})}\|u\|^2_{L^2(\mathscr{O})}.$$

Then from (A.1)

$$\|u\|^2_{L^2(\mathscr{O})} \leqslant C\left(h\left(\frac{2}{\gamma}\|f\|_{L^2(\mathscr{O})}\|u\|_{L^2(\mathscr{O})} + \frac{1}{\gamma^2}\|g\|^2_{L^2(\partial\mathscr{O})}\right)\right.$$
$$\left. + h^2\left(\frac{1}{2\gamma}\|g\|^2_{L^2(\partial\mathscr{O})} + 2\|f\|_{L^2(\mathscr{O})}\|u\|_{L^2(\mathscr{O})} + \|\beta\|_{L^\infty(\mathscr{O})}\|u\|^2_{L^2(\mathscr{O})}\right)\right).$$

For h small enough we obtain

$$\|u\|^2_{L^2(\mathscr{O})} \leqslant C\left(\frac{h}{\gamma^2}\|g\|^2_{L^2(\partial\mathscr{O})} + \frac{h^2}{\gamma^2}\|f\|^2_{L^2(\mathscr{O})}\right). \tag{A.3}$$

One can notice that the scaling of this estimate is optimal. Indeed considering that γ is the dimension of the inverse of a length which is evident from the boundary condition, all quantities have the same dimension at inspection of (A.2). Inequality (4.2) is obtained by taking $f = 0$ in the previous inequality. □

A.3 *Proof of theorem 4.2*

Proof. Suppose that u and u_h are the solutions of the two following problems

$$\begin{cases} -u'' + \beta u = f, & (\mathscr{O}) \\ (-\partial_\nu + i\gamma)u = g, & (\partial\mathscr{O}). \end{cases}$$

and

$$\begin{cases} -u_h'' + \beta_h u_h = f, & (\mathscr{O}) \\ (-\partial_\nu + i\gamma)u_h = g, & (\partial\mathscr{O}). \end{cases}$$

Then $e_h := u - u_h$ satisfies

$$\begin{cases} -e_h'' + \beta_h e_h = (\beta_h - \beta)u, & (\mathscr{O}) \\ (-\partial_\nu + i\gamma)e_h = 0, & (\partial\mathscr{O}). \end{cases}$$

Inequality (A.3) yields

$$\|e_h\|_{L^2(\mathscr{O})} \leqslant C\frac{h}{\gamma}\|(\beta_h - \beta)u\|_{L^2(\mathscr{O})} \leqslant C\frac{h}{\gamma}\|\beta_h - \beta\|_{L^\infty(\mathscr{O})}\|u\|_{L^2(\mathscr{O})}.$$

Using one more time (A.3) to estimate u and regarding γ which is a positive number, we get

$$\|e_h\|_{L^2(\mathscr{O})} \leqslant C\left(h^{\frac{3}{2}}\|g\|_{L^2(\partial\mathscr{O})} + h^2\|f\|_{L^2(\mathscr{O})}\right)\|\beta_h - \beta\|_{L^\infty(\mathscr{O})}.$$

□

REFERENCES

BUFFA, A. & MONK, P. (2008) Error estimates for the Ultra Weak Variational Formulation of the Helmholtz equation, *ESAIM: Mathematical Modelling and Numerical Analysis*, November, **42**, 925–940.

CESSENAT, O. & DESPRÉS, B. (2003) Using plane waves as base functions for solving time harmonic equations with the ultra weak variational formulation, *Journal of Computational Acoustics*, **11** no. 2, 227–238.

CESSENAT, O. & DESPRÉS, B. (1998) Application of an ultra weak variational formulation of elliptic PDEs to the two dimensional Helmholtz problem, *SIAM J. Numer. Anal.*, **vol. 55**, no1, 255–299.

DESPRÉS, B. (1994) Sur une formulation variationnelle de type ultra-faible, *C. R. Acad. Sci. Paris Sr. I Math.* **318** no. 10, 939–944.

DAUTRAY, R. & LIONS, J.-L. (1984) *Mathematical Analysis and Numerical Methods for Science and Technology*. Masson 1984, volume 3, chapitre 8.

FARHAT, C., HARARI, I. & FRANCA, L. (2001) The discontinuous enrichment method, *Computer Methods in Applied Mechanics and Engineering*, **190**, 6455–6479.

FARHAT, C., TEZAUR, R. & WIEDEMANN-GOIRAN, P. (2004) Higher-order extensions of a discontinuous Galerkin method for mid-frequency Helmholtz problems, *International Journal for Numerical Methods in Engineering*, **61** 1938–1956.

FARHAT, C., TEZAUR, R. & TOIVANEN, J. (2009) A domain decomposition method for discontinuous Galerkin discretizations of Helmholtz problems with plane waves and Lagrange multipliers, *International Journal for Numerical Methods in Engineering*, **78**, 1513–1531.

GABARD, G., GAMALLO, P. & HUTTUNEN, T. (2011) A comparison of wave-based discontinuous Galerkin, ultra-weak and least-square methods for wave problems, *International Journal for Numerical Methods in Engineering*, **85**, 380–402.

GITTELSON, C. J., HIPTMAIR, R. & PERUGIA, I. (2009) Plane wave discontinuous Galerkin methods: Analysis of the h-version, *ESAIM: Mathematical Modelling and Numerical Analysis*, **43**, 297–331.

GRISVARD, P. (1986) Problèmes aux limites dans les polygones. Mode d'emploi. *EDF Bull. Direction tudes Rech. Sr. C Math. Inform.*, no. 1, 3, 21–59.

GUSAKOV, E. Z., LECLERT, G., BOUCHER, I., HEURAUX, S., HACQUIN, S., COLIN, M., BULANIN, V. V., PETROV, A. V., YAKOVLEV, B. O., CLAIRET, F. & ZOU, X. L. (2002) Small-angle scattering and spatial resolution of fluctuation reflectometry: comparison of 2D analytical theory with numerical calculations. *Plasma physics and controlled fusion*, **vol. 44**, no8, 1565–1579.

HIPTMAIR, R., MOIOLA, A. & PERUGIA, I. (2009) Plane wave discontinuous Galerkin methods for the 2D Helmholtz equation: analysis of the p-version, Preprint 2009-20, SAM Report, ETH Zürich, Switzerland.

HIPTMAIR, R., MOIOLA, A. & PERUGIA, I. (2011) Error analysis of Trefftz-discontinuous Galerkin methods for the time-harmonic Maxwell equations, ETHZ, Research Report No. 2011-09.

HUTTUNEN, T., GAMMALO, P. & ASTLEY, R. J. (2009) Comparison of two wave element methods for the Helmholtz problem. *Communications in Numerical Methods in Engineering*, **25**, 35–52.

HUTTUNEN, T., MALINEN, M. & MONK, P. (2007) Solving Maxwells equations using the ultra weak variational formulation, *Journal of Computational Physics*, **223**, Issue 2, 731–758.

HUTTUNEN, T., MONK, P. & KAIPIO, J. P. (2002) Computational Aspects of the Ultra-Weak Variational Formulation, *Journal of Computational Physics*, **182**, Issue 1, 27–46.

IMBERT-GÉRARD, L.-M. & DESPRÉS, B. (2011) A generalized plane wave numerical method for smooth non constant coefficients, Tech. report R11034, LJLL-UPMC.

KALASHNIKOVA, I., TEZAUR, R. & FARHAT, C. (2010) A discontinuous enrichment method for variable coefficient advection-diffusion at high Péclet number. *International Journal for Numerical Methods in Engineering*, **87**, 309–335.

MELENK, J. (1995) On Generalized Finite Element Methods, PhD thesis, University of Maryland, USA.

MELENK, J. (1999) Operator adapted spectral element methods I: harmonic and generalized harmonic polynomials, *Numerische Mathematik*, **84**, 35–69.

MELENK, J.M. & BABUSKA, I. (1996) The partition of unity method finite element method: basic theory and applications, *Computer Methods in Applied Mechanics and Engineering*, **139**, 289–314.

MELENK, J.M. & SAUTER, S. (2011) Wavenumber explicit convergence analysis for Galerkin discretizations of the Helmholtz equation *SIAM J. Numer. Anal.*, **49**, 1210–1243.

MONK, P. (2003) *Finite element methods for Maxwell's equations*, Calderon press Oxford.

PERREY-DEBAIN, E., LAGHROUCHE, O., BETTESS, P. & TREVELYAN, J. (2004) Plane-wave basis finite elements and boundary elements for three-dimensional wave scattering, *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, **362**, 561–577.

PLUYMERS, B., DESMET, W., VANDEPITTE, D. & SAS, P. (2006) Wave based modelling methods for steady-state interior acoustics: an overview, *ISMA*.

STROUBOULIS, T., BABUSKA, I. & HIDAJAT, R. (2006) The generalized finite element method for Helmholtz equation: theory, computation, and open problems, *Computational Methods in Applied Mechanical Engineering*, **195**, 4711–4731.

SWANSON, D. G. (2003) *Plasma Waves, 2nd Edition*, Series in Plasma Physics.

WANG, D., TEZAUR, R., TOIVANEN, J. & FARHAT, C. (2012) Overview of the discontinuous enrichment method, the ultra-weak variational formulation, and the partition of unity method for acoustic scattering in medium frequency regime and performance comparisons *Int. J. Numer. Meth. Engng.* **89(4)**, 403–417.