A generalized plane wave numerical method for smooth non constant coefficients

Lise-Marie Imbert-Gerard*, Bruno Despres†

November 7, 2011

Abstract

Maxwell's equations with hermitian permittivity ε are used to model reflectometry in fusion plasma. Simplified models split them into two different propagation modes. Here we focus on the O-mode equation. We propose an original method based on generalized plane waves and approximated coefficients for the numerical approximation. This is justified in dimension one by a high order convergence estimate rate. Some numerical results are presented in dimension one and two.

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 real non constant coefficient

$$\left\{ \begin{array}{lcll} -\Delta u + \alpha u & = & f, & x \in \Omega, \\ (\partial_\nu + i\gamma)\, u & = & Q\,(-\partial_\nu + i\gamma)\, u + g, & x \in \Gamma. \end{array} \right. \tag{1}$$

The real smooth function is $\alpha \in \mathbb{R}$. The γ function can be a variable physical parameter satisfying $0 < \gamma_m \le \gamma \le \gamma_M$, but for the sake of simplicity we will consider it constant and positive. The unknown $u(\mathbf{x}) \in \mathbb{C}$ is sought in the space of complex valued functions.

1.1 Plane wave methods

The numerical method that we propose is an extension of plane waves methods, such as the ultra weak variational formulation (UWVF) [5, 3, 4, 11, 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 [19].

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 [18, 20] falls in the same class of method [23, 24]. It has also been shown that UWVF can be interpreted as a special Discontinuous Galerkin procedure [11, 13, 7, 9]. It has been proved that the analysis of h-convergence takes great advantage of this fact in [2, 11]. The analysis of p convergence is treated in [12]. Comparisons between these methods is investigated in [10, 14, 25]. Analysis with respect to the wave-number k is performed in [21].

In this work we recast the classical UWVF as a special Galerkin procedure with a bilinear form which is coercive and bicontinuous in appropriate spaces. It helps to develop the family of generalized plane wave methods needed to treat variable coefficients. This family of plane waves

*LJLL, UPMC, 4 place Jussieu, 75252 Paris (imbert@ann.jussieu.fr).
†LJLL, UPMC, 4 place Jussieu, 75252 Paris, (despres@ann.jussieu.fr).

generates a high order method with respect to the basis functions and the coefficients of the problem: in this direction we refer also to [8, 15]. Our most original theoretical result is probably the fact that the underlying non conformity of the new bilinear form can be treated with the second Strang's lemma. Classically non conformal methods are analyzed in the context of Finite Element Methods. To our knowledge it is the first time that it is introduced and analyzed in the context of generalized plane wave methods.

1.2 Physical motivations

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

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

where E denotes the electric field, ω is the pulsation, c the sound speed and ε the dielectric tensor. The hermitian dielectric tensor represents the electromagnetic behavior of the media. The cold plasma theory [26] 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 $c = 1 - b^2$. This is completed with boundary conditions of metallic or absorbing type. We refer to [22] for the general theory of Maxwell's equations and to [3, 13, 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 (2) under convenient assumptions on the direction and polarization of the electric field. The 2D equation for what is called the O-mode reduces to

$$-\Delta E_z - \frac{\omega^2}{c^2}\varepsilon_z(x, y)E_z = 0, \quad \Delta = \partial_{xx} + \partial_{yy}, \tag{3}$$

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 a 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 (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 further simplified 1D model writes

$$-\frac{d^2}{dx^2}E_z + xE_z = 0. \tag{4}$$

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. This equation will be used for numerical purposes. Equations (3) and (4) are particular cases of our model problem (1).

1.3 Plan

This work is organized as follows. In section 2 we present the general principle of UWVF and adapt it to smooth coefficients. It is made possible with new basis functions. The next section

3 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. Numerical results are provided in section 4 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 Description of the proposed numerical method

2.1 Notations

Unlike the classical variational formulation used for instance by finite element methods, here the variational formulation requires meshing the domain as a preliminary task. The mesh of the domain

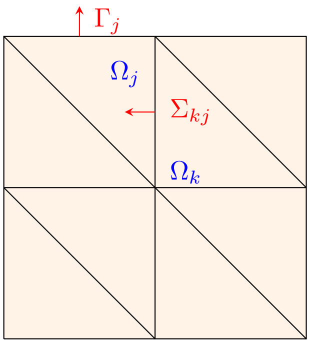


Figure 1: Example of a meshed square domain Ω, with elements Ω_k, edges Σ_{kj} and Γ_j respectively oriented toward Ω_j and the exterior of the domain.

Ω is denoted $\mathcal{T}_h = \{\Omega_k\}_{k\in[\![1,N_h]\!]}$, such that :

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

The functional 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

$$(x, y) = \sum_k \int_{\partial\Omega_k} x_k \overline{y_k}.$$

It defines a norm: $\|x\| = \sqrt{(x, x)}$. In particular for any operator $A \in \mathcal{L}(V)$, the norm is

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

Remark 2.1. *It is fundamental to notice that the space V already depends on the mesh. Moreover: if $\Omega \subset \mathbb{R}$ the dimension of V is finite; if $\Omega \subset \mathbb{R}^d$ with $d \geq 2$, the dimension of V is infinite.*

2.2 A standard ultra weak variational formulation

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

$$H_k(\alpha) = \left\{ v_k \in H^1(\Omega_k), \middle| \begin{array}{l} (-\Delta+\alpha)v_k = 0, (\Omega_k), \\ ((-\partial_\nu + i\gamma)v_k)_{|\partial\Omega_k} \in L^2(\partial\Omega_k) \end{array} \right\} \tag{5}$$

and

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

Theorem 2.1. *Let $u \in H^1(\Omega)$ be a solution of the problem (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*

$$\begin{aligned} &\sum_k \left(\int_{\partial\Omega_k} \frac{1}{\gamma} x_k \overline{(-\partial_\nu + i\gamma)e_k} - \sum_{j, j\neq k} \int_{\Sigma_{kj}} \frac{1}{\gamma} x_j \overline{(\partial_\nu + i\gamma)e_k} \right) \\ &- \sum_{k, \Gamma_k \neq \emptyset} \int_{\Gamma_k} \frac{Q}{\gamma} x_k \overline{(\partial_\nu + i\gamma)e_k} \\ &= -2i \sum_k \int_{\partial\Omega_k} f\overline{e} + \sum_k \int_{\Gamma_k} \frac{1}{\gamma} g \overline{(\partial_\nu + i\gamma)e_k}, \end{aligned} \tag{6}$$

for any $e = (e_k)_{k\in[\![1,N_h]\!]} \in H$.

Conversely, if $x \in V$ is solution of (6) then the function u defined locally by

$$\left\{ \begin{array}{l} u_{|\Omega_k} = u_k \in H^1(\Omega_k), \\ (-\Delta+\alpha)u_k = f_{|\Omega_k}, \\ (-\partial_{\nu_k} + i\gamma)u_k = x_k, \end{array} \right. \tag{7}$$

is the unique solution of the problem (1).

Proof. By hypothesis $u \in H^1$ and the normal derivatives $\partial_\nu u$ are square integrable. It allows us to write for a given $k \in [\![1, N_h]\!]$

$$\int_{\partial\Omega_k} \frac{1}{\gamma}(-\partial_\nu + i\gamma)u \cdot \overline{(-\partial_\nu + i\gamma)e_k} = \int_{\partial\Omega_k} \frac{1}{\gamma}(\partial_\nu + i\gamma)u \cdot \overline{(\partial_\nu + i\gamma)e_k} - 2i \int_{\partial\Omega_k} (u\overline{\partial_\nu e_k} - \partial_\nu u \overline{e}_k), \tag{8}$$

then definition (5) and problem (1) yields

$$\left\{ \begin{array}{ll} (-\Delta+\alpha)u = f, & (\Omega_k), \\ (-\Delta+\alpha)e_k = 0, & (\Omega_k). \end{array} \right. \tag{9}$$

Performing two integrations by part, the following holds $\forall k \in [\![1, N_h]\!]$

$$\left\{ \begin{array}{ll} \int_{\Omega_k} \nabla u \cdot \nabla\overline{e_k} + \int_{\Omega_k} \alpha u \cdot \overline{e_k} - \int_{\partial\Omega_k} \partial_\nu u \cdot \overline{e_k} & = \int_{\Omega_k} f \cdot \overline{e_k}, \\ \int_{\Omega_k} \nabla u \cdot \nabla\overline{e_k} + \int_{\Omega_k} \alpha u \cdot \overline{e_k} - \int_{\partial\Omega_k} u \cdot \overline{\partial_\nu e_k} & = 0. \end{array} \right.$$

So using the boundary conditions together with the smoothness of the solution u, namely $\forall k \in [\![1, N_h]\!]$

$$\left\{ \begin{array}{ll} (\partial_\nu + i\gamma)u_{|\Sigma_{kj}} & = (-\partial_\nu + i\gamma)u_{|\Sigma_{jk}}, \\ (\partial_\nu + i\gamma)u_{|\Gamma_k} & = Q(-\partial_\nu + i\gamma)u_{|\Gamma_k} + g, \end{array} \right. \tag{10}$$

the identity (8) yields $\forall k \in [\![1, N_h]\!]$

$$\begin{aligned} &\left(\int_{\partial\Omega_k} \frac{1}{\gamma} x_k \overline{(-\partial_\nu + i\gamma)e_k} - \sum_{j, j\neq k} \int_{\Sigma_{kj}} \frac{1}{\gamma} x_j \overline{(\partial_\nu + i\gamma)e_k} \right) \\ &-\mathbf{1}_{\Gamma_k \neq \emptyset} \int_{\Gamma_k} \frac{Q}{\gamma} x_k \overline{(\partial_\nu + i\gamma)e_k} \\ &= -2i \int_{\partial\Omega_k} f\overline{e} + \int_{\Gamma_k} \frac{1}{\gamma} g \overline{(\partial_\nu + i\gamma)e_k}. \end{aligned}$$

Summing over k then gives the UWVF (6).

Conversely, let x be a solution of (6) and let u satisfy (7) on every Ω_k. The hypothesis on u and e, gives (9) and then $\forall k \in [\![1, N_h]\!]$

$$\int_{\partial\Omega_k} \frac{1}{\gamma}(-\partial_\nu + i\gamma)u \cdot \overline{(-\partial_\nu + i\gamma)e_k} - \int_{\partial\Omega_k} \frac{1}{\gamma}(\partial_\nu + i\gamma)u \cdot \overline{(\partial_\nu + i\gamma)e_k} = -2i \int_{\Omega_k} f\overline{e_k}.$$

Summing over k and combining the result with (6) satisfied by x we obtain for all $e = (e_k) \in H$

$$\begin{aligned}
&\sum_{k,j\neq k} \int_{\Sigma_{kj}} \frac{1}{\gamma} x_k \cdot \overline{(\partial_\nu + i\gamma)e_k} + \sum_{k,\Gamma_k\neq\emptyset} \int_{\Gamma_k} \frac{1}{\gamma} x_k \cdot \overline{(\partial_\nu + i\gamma)e_k} \\
&= \sum_{k,j\neq k} \int_{\Sigma_{kj}} \frac{1}{\gamma} x_j \cdot \overline{(\partial_\nu + i\gamma)e_k} + \sum_{k,\Gamma_k\neq\emptyset} \int_{\Gamma_k} \frac{1}{\gamma} (Qx_k + g) \cdot \overline{(\partial_\nu + i\gamma)e_k}.
\end{aligned}$$

Therefore u satisfies (10). It shows that u is the unique smooth solution of (1) given by theorem A.1 in the appendix. □

In order to give a more compact formulation of this problem, some definitions are required.

Definition 2.1. *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 :

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

We 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 2.2. *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 2.3. *Let Π be the mapping defined by*

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

Definition 2.4. *If F^* denotes the adjoint operator of the operator F, let A be the operator $F^*\Pi$.*

The proof of the following result is to be found in [4].

Theorem 2.2. *The problem (6) is equivalent to*

$$\left\{ \begin{array}{l} \textit{Find } x \in V \textit{ such that } \forall y \in V \\ (x, y) - (\Pi x, Fy) = (b, y), \end{array} \right. \tag{11}$$

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) *such that* $\left((-\partial_\nu + i\gamma)u_{|\partial\Omega_k}\right)_{k\in[\![1,N_h]\!]} \in V$*, then* $x = \left((-\partial_\nu + i\gamma)u_{|\partial\Omega_k}\right)_{k\in[\![1,N_h]\!]}$ *is solution in V of* (11).
- *Conversely if x is solution of* (11) *then $u = E_f(x)$ is the unique solution of* (6). *The problem* (11) *is equivalent to*

$$\left\{\begin{array}{l} \text{For } b \in V, \text{ find } x \in V \\ (I - A)x = b. \end{array}\right. \tag{12}$$

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

Lemma 2.1. *The operator Π obviously satisfies $\|\Pi\| \leq 1$ for any complex function Q such that $|Q| \leq 1$.*

Lemma 2.2. *The operator F is an isometry.*

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

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

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

On the other hand, for all $k \in [\![1, N_h]\!]$

$$\int_{\Omega_k} |\nabla e_k|^2 + \alpha|e_k|^2 - \int_{\partial\Omega_k} \partial_\nu e_k \cdot \overline{e}_k = 0,$$

so that

$$\|Fy\|^2 = \|y\|^2.$$

This clearly implies the result. □

As a consequence, it yields

Proposition 2.1. *The operator A satisfies $\|A\| \leq 1$.*

This operator also satisfies the following property.

Proposition 2.2. *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 $k \in [\![1, N_h]\!]$

$$\left\{\begin{array}{rll} -\Delta u + \alpha u &= 0, & (\Omega_k), \\ (\partial_\nu + i\gamma)u &= z_{|\partial\Omega_k}, & (\partial\Omega_k). \end{array}\right. \tag{13}$$

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

$$\left\{\begin{array}{rll} -\Delta w + \alpha w &= 0, & (\Omega_k), \\ (-\partial_\nu + i\gamma)w &= v_{|\partial\Omega_k}, & (\partial\Omega_k). \end{array}\right. \tag{14}$$

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 (13) and (14) 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 (13).

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

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

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

Thanks to the preliminary result, u is the unique solution of the corresponding (1) problem : it is the 0 solution. Then $z=0$, and so $x=0$. The proof is ended. □

2.3 An abstract discretization procedure

The next step consists in the discretization of equation (11). 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

$$\forall y_h\in V_h, (x_h,y_h)-(\Pi x_h,Fy_h)=(b,y_h). \tag{15}$$

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 Galerkin procedure is only abstract until on provides a 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 (15) 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 [5, 4, 11, 2, 12, 13].

Definition 2.5. *Let us define the norm* $|||\cdot|||$

$$\forall v\in V,\quad |||v|||=\|(I-A)v\|$$

and the bilinear form of the formulation (11)

$$a(x,y)=(x,y)-(\Pi x,Fy)\,.$$

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

A fundamental property is

Lemma 2.3. *The bilinear form is coercive with respect to the norm* $|||\cdot|||$

$$|||x|||^2 \leq 2\mathcal{R}\left(a(x,x)\right) \quad \forall x \in V,$$

and is bicontinuous in the sense

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

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

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

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

Proposition 2.3. *Assume there exists x solution of the problem (12). Then any discrete solution x_h satisfies the inequality*

$$|||x - x_h||| \leq 2 \inf_{z_h \in V_h} \|x - z_h\|. \tag{16}$$

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 2.3. □

Lemma 2.4. *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 $x_h = 0$.

We apply the inequality (16) with the choice $x = b = 0$. It yields

$$\|x_h\| \leq 2 \inf_{z_h \in V_h} \|z_h\| = 0.$$

□

2.4 The new method

If one desires to implement the discrete ultra weak formulation (15), it is necessary to manipulate shape or basis functions φ which are based on solutions of the homogeneous equation

$$(-\Delta + \alpha)\varphi = 0.$$

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

$$\varphi(x) = e^{\sqrt{\alpha}(d,x)} \text{ with } d = (d_1, d_2) \text{ and } (d,d) = 1.$$

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 $\alpha = x_1$ is linear, it is possible to construct φ from the Airy functions Ai and Bi. But the Airy functions are highly transcendantal, 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. Instead of approximating α by a piecewise constant function on every element of the mesh, here the approximation of the coefficient is performed up to order q in h. This is the main novelty compared to the classical method. More precisely, for all cell $k \in [\![1, N_h]\!]$, define $p(k) \in \mathbb{N}^*$ functions α_k^l, null on $\Omega - \{\Omega_k\}$ and satisfying for all $l \in [\![1, p(k)]\!]$

$$\|\alpha - \alpha_k^l\|_{L^\infty(\Omega_k)} \le C(k)h^q, \tag{17}$$

where h denotes the size of the mesh and $C(k)$ denotes a constant independent of h but depending on k. The letter q indeed refers to the order of approximation of the initial equation's coefficient α. We will assume that there exists a constant C independent of h and k, such that

$$\max_{k\in[\![1,N_h]\!]} \max_{l\in[\![1,p(k)]\!]} \|\alpha - \alpha_k^l\|_{L^\infty(\Omega_k)} \le Ch^q. \tag{18}$$

We also assume that we are able to construct a corresponding smooth function φ_k^l such that

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

Here smooth means that

$$(-\partial_\nu + i\gamma)(\varphi_k^l) \in L^2(\partial\Omega_k).$$

Under these assumptions we are able to make the following general definitions.

Definition 2.6. *The local discrete space is*

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

The global discrete space $V^q \subset V$ is defined by : $V^q = \prod_{1\le k\le N_h} W_k$.

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

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

Next we define what are the generalizations of operators E and F in this context.

Definition 2.7. *Let $E^q \in \mathcal{L}(V^q, H)$ be the discrete mapping defined $\forall k \in [\![1, N_h]\!]$ and $\forall l \in [\![1, p(k)]\!]$ by*

$$E^q(v_k^l) = \varphi_k^l \text{ on } H^1(\Omega_k), \quad \text{ and } v_k^l = 0 \text{ on } H^1(\Omega_{k'}) \;\; k' \neq k. \tag{19}$$

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

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

The corresponding numerical method now writes : find $x_h \in V^q$ such that

$$\forall y_h \in V^q, (x_h, y_h)_V - (\Pi x_h, F^q y_h)_V = (b^q, y_h)_V \tag{20}$$

with the right hand side given by

$$(b^q, y_h)_V = -2i\int_\Omega f\overline{E^q(y_h)} + \int_\Gamma \frac{1}{\gamma} g\overline{F^q(y_h)}, \qquad \forall y_h \in V^q. \tag{21}$$

Before studying the method we desire to describe the exact construction of the basis functions.

2.5 Design of the basis functions in dimension one

The one dimensional case is enough to explain how we propose 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.

2.5.1 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. The reason is that $dim(H_k(\alpha)) = 2$ because the number of elementary solutions of a second order differential equation 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 $\alpha(x_{k+1/2}) < 0$.

2.5.2 Local approximation

To ensure the local approximation of the α coefficient (18) using exponential of polynomials, one has to fit the polynomials' coefficients to approximate the Taylor expansion of the equation's coefficient α. The Taylor expansion 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 look a priori for approximate functions $\alpha_{\pm}$ such that

$$\alpha = \alpha_{\pm} + O(h^q) \tag{22}$$

holds together with

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

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 $P_{\pm} = \sum_{i\le I} \beta_i \left(x - x_{k+\frac{1}{2}}\right)^i$

$$\alpha_{\pm} = P''_{\pm} + (P'_{\pm})^2 = \left(\sum_{i\le I} \beta_i \left(x - x_{k+\frac{1}{2}}\right)^i\right)'' + \left(\left(\sum_{i\le I} \beta_i \left(x - x_{k+\frac{1}{2}}\right)^i\right)'\right)^2.$$

In order to satisfy (22) we have to chose $I \in \mathbb{N}$ and $(\beta_i)_{0\le i\le I}$ such that

$$\left(\sum_{i\le I} \beta_i \left(x - x_{k+\frac{1}{2}}\right)^i\right)'' + \left(\left(\sum_{i\le I} \beta_i \left(x - x_{k+\frac{1}{2}}\right)^i\right)'\right)^2 = \sum_{i=0}^{q} \frac{d^i\alpha}{dx^i}(x_{k+1/2}) \left(x - x_{k+\frac{1}{2}}\right)^i + O(h^q). \tag{23}$$

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. 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 (23). 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 (23) 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 (23)

$$\begin{cases} 2\beta_2 + \beta_1^2 = \alpha\left(x_{k+\frac{1}{2}}\right) \equiv a, \\ 4\beta_1\beta_2 = \alpha'\left(x_{k+\frac{1}{2}}\right) \equiv b. \end{cases} \tag{24}$$

  Elimination of β_2 yields $-2\beta_1^3 + 2a\beta_1 = b$. 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. Ultimately the inequality (17) will not be true near a singularity. It must be noticed that we have used such a method in our first numerical tests: indeed it revealed a singularity near $\alpha(x) \approx 0$. 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 (24) and take into account β_3. The system becomes

$$\begin{cases} 2\beta_2 + \beta_1^2 = a, \\ 3\beta_3 + 4\beta_1\beta_2 = b. \end{cases} \tag{25}$$

  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

$$\frac{d}{dx}e^{P_+(x_{k+\frac{1}{2}})} = 0 \Longleftrightarrow P_+'(x_{k+\frac{1}{2}}) = 0$$

  and

$$\frac{d}{dx}e^{P_+(x_{k+\frac{1}{2}})} = 1 \Longleftrightarrow P_+'(x_{k+\frac{1}{2}}) = 1.$$

  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 (25) and that the resulting formulas are just polynomial expressions with respect to all coefficients. Notice that a priori $\alpha_+ \neq \alpha_-$.

- **General case :** $q > 2$ **and** $I = q + 1$**.** We use this method at any order. That is we solve the system of q equations with $q+1$ unknowns obtained identifying the first q coefficients in both parts of the expansion (23) with the normalization

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

  and

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

By construction $\varphi_+ = e^{P_+}$ and $\varphi_- = e^{P_-}$ are linearly independent functions. In practice we use an automatic procedure with Maple to compute the solutions, but it can easily be done by hand. The coefficients β_2, β_3, β_4, $\dots$, are calculated one after the other.

Once the polynomials P_+ and P_- have been constructed up to order q, we set

$$\alpha_+ = P_+'' + (P_+')^2 \text{ and } \alpha_- = P_-'' + (P_-')^2.$$

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

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

That is we use $I > q$ to get rid of the singularity described in the case $I = q = 2$. We observe then that when $q > 1$, α_+ and α_- are different since $P_+ \neq P_-$. This construction is the major motivation for the introduction of the general formalism (19)-(21) which permits to define and study such non conformal methods.

Remark 2.2. *Note that α and all the α_j functions constructed here, as well as all there derivatives, are bounded independently from k.*

Remark 2.3. *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 4.*

2.6 Design of the basis functions in dimension two

The generalization in dimension two corresponds to a basis function $\varphi(x, y) = e^{P(x,y)}$ solution to $-\Delta\varphi + \alpha\varphi = 0$. It is associated to the equation

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

In theory a local expansion with respect to the x and y variables is possible, as it was performed in dimension one.

2.6.1 Linear coefficients+rotation

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

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

one ends up with the equation

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

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 θ will be provided in the numerical section 4.3.

3 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 (20) to the exact solution. Since the discrete method (20) can be viewed as a convenient modification of the bilinear form (15), it is not surprising that that the convergence analysis strongly relies on the second Strang's lemma as it is the case for non conformal finite element methods [7]. 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 3.2, but are open problems in greater dimension.

3.1 Simplified notations in dimension one

Let the order of approximation q be a given number. We assume that we have 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 $\mathcal{D} = \{\alpha_j\}_{j\in[\![1,2N_h]\!]}$; $\{z_j\}_{j\in[\![1,2N_h]\!]}$ 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

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

This will greatly reduce the technicalities of the proof.

3.2 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 3.1. *Let $\mathcal{O}$ be a one-dimensional open interval with length h. Let w be the unique solution of*

$$\begin{cases} -\Delta w + \beta w &= 0, \quad (\mathcal{O}), \\ (-\partial_\nu + i\gamma)\, w &= g, \quad (\partial\mathcal{O}). \end{cases} \tag{26}$$

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

$$\|w\ \|_{L^2(\mathcal{O})} \leq C\sqrt{h}\, \|g\|_{L^2(\partial\mathcal{O})}\,, \tag{27}$$

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

$$\begin{cases} -\Delta w + \beta w &= f, \quad (\mathcal{O}), \\ (-\partial_\nu + i\gamma)\, w &= g, \quad (\partial\mathcal{O}), \end{cases} \tag{28}$$

and the modified problem

$$\begin{cases} -\Delta w + \beta_h w &= f, \quad (\mathcal{O}), \\ (-\partial_\nu + i\gamma)\, w &= g, \quad (\partial\mathcal{O}), \end{cases} \tag{29}$$

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

Theorem 3.2. *Let $\mathcal{O}$ be a one-dimensional open interval with length h. If u is solution of the problem* (28) *and u_h is solution of the problem* (29)*, then for small h there exists a constant C such that*

$$\|u - u_h\|_{L^2(\mathcal{O})} \leq C\left(h^{\frac{3}{2}}\|g\|_{L^2(\partial\mathcal{O})} + h^2\|f\|_{L^2(\mathcal{O})}\right)\|\beta - \beta_h\|_{L^\infty(\mathcal{O})}. \tag{30}$$

3.3 The discrete problem

This paragraph is devoted to showing that the operator F^q described in section 2.4 is an approximation of the operator F up to the order $q + 1$ in h. Consider the following problem

$$\begin{cases} \text{Find } x_h \in V_q \text{ such that} \\ (I - A^q)x_h = b, \end{cases} \tag{31}$$

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

Lemma 3.1. *Let $q \geq 2$. Suppose h is small enough and basis functions are constructed as described in paragraph 2.5.2. For all $k \in [\![1, N_h]\!]$, there exists a constant C independent k such that $\forall z \in V^q$ and $\forall k \in [\![1, N_h]\!]$*

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

Proof. Set $k \in [\![1, N_h]\!]$ and $z = x_1 z_1 + x_2 z_2$. First we desire to write x_j 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

$$(w_j, z_l)_V = \delta_{jl}, \tag{32}$$

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

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

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

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 (32) 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

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

Let us set for convenience

$$A = \frac{|(z_1,z_2)|}{\|z_1\|\|z_2\|},$$

so that

$$\sum_{j\in\{1,2\}} \|z_j\|\|w_j\| \le 2 \frac{1}{1-A^2}.$$

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

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

$$= \frac{1}{\gamma}\left|-P_1'(x_{k+\frac12}) + i\gamma P_1(x_{k+\frac12})\right|^2 + \frac{1}{\gamma}\left|P_1'(x_{k+\frac12}) + i\gamma P_1(x_{k+\frac12})\right|^2 + O(h),$$

that is

$$||z_1||^2 = 2\gamma + O(h).$$

With the same method we obtain

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

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 2.5.2 by the approximation of the Taylor expansion (23) 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 (24).

□

Lemma 3.2. *For small* h *and considering the basis functions constructed as described in paragraph 2.5.2, there exists a constant* C

$$\|F^q - F\| \le C h^{q+1} \tag{33}$$

Proof. Notation: for all $j \in [\![1, 2N_h]\!]$, k will denote the index of z_j's support. For all $j \in [\![1, 2N_h]\!]$, the function φ_j is by construction such that

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

We also define ψ_j which satisfies the same boundary condition and the equation with the exact coefficient α

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

Then

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

since φ_j and ψ_j satisfy the same boundary condition: $(-\partial_\nu + i\gamma)(\varphi_j - \psi_j) = 0$. Then

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

since both φ_j and ψ_j satisfy homogeneous equations. Then

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

thanks to Cauchy-Schwarz inequality. On the other hand, from (27) and (30) for small hs

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

and $\|\alpha_j + \alpha\|_{L^\infty(\Omega_k)}$ is bounded as noticed in remark 2.2. So for small h

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

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

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

Thanks to lemma 3.1 it means that

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

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

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

which exactly means

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

The result then comes from equation (22) ensured by the construction of approximated coefficients α_js. □

We now want to address the convergence problem.

3.4 Some norms

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

Lemma 3.3. *There exists a constant C such that for all $x \in V$*

$$Ch^{3/2}\|x\| \leq \|(I - A)x\|.$$

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 does 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 we define $u = E(x)$ and $w = E(b)$, so that $(u, w) \in H \times H$ and

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

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

Since F is an isometry one has

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

It means on every interface

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

This leads to a system of jump conditions on the interfaces

$$\left\{\begin{array}{l} (-\partial_\nu + i\gamma) u_{|\Omega_1}(x_1) = (-\partial_\nu + i\gamma) w_{|\Omega_1}(x_1), \\ \forall k \in [\![2, N_h]\!], \left| \begin{array}{l} \left(\frac{d}{dx} u_{|\Omega_{k-1}} - \frac{d}{dx} u_{|\Omega_k}\right)(x_k) = \frac{1}{2}\left((-\partial_\nu + i\gamma) w_{|\Omega_k} - (\partial_\nu + i\gamma) w_{|\Omega_{k-1}}\right)(x_k), \\ \left(u_{|\Omega_k} - u_{|\Omega_{k-1}}\right)(x_k) = \frac{1}{2i\gamma}\left((-\partial_\nu + i\gamma) w_{|\Omega_k} - (\partial_\nu + i\gamma) w_{|\Omega_{k-1}}\right)(x_k), \end{array}\right. \\ (\partial_\nu + i\gamma) u_{|\Omega_{N_h}}(x_{N_h+1}) = (\partial_\nu + i\gamma) w_{|\Omega_{N_h}}(x_{N_h+1}). \end{array}\right. \tag{34}$$

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

$$\forall k \in [\![1, N_h]\!], u_{|\Omega_k} = \delta_0^k U_0 + \delta_1^k U_1, \tag{35}$$

where $(\delta_0^k, \delta_1^k)_{k\in[\![1,N_h]\!]}$ completely determine $u \in H$. Plugging (35) in (34), 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

$$\left\{\begin{array}{l} (-\partial_\nu + i\gamma) U_0(x_1)\delta_0^1 + (-\partial_\nu + i\gamma) U_1(x_1)\delta_1^1 = \lambda_0, \\ \forall k \in [\![2, N_h]\!], \left| \begin{array}{l} \frac{d}{dx} U_0(x_k)(\delta_0^{k-1} - \delta_0^k) + \frac{d}{dx} U_1(x_k)(\delta_1^{k-1} - \delta_1^k) = \lambda_{k-1}, \\ U_0(x_k)(\delta_0^{k-1} - \delta_0^k) + U_1(x_k)(\delta_1^{k-1} - \delta_1^k) = \mu_{k-1}, \end{array}\right. \\ (\partial_\nu + i\gamma) U_0(x_{N_h+1})\delta_0^{N_h} + (\partial_\nu + i\gamma) U_1(x_{N_h+1})\delta_1^{N_h} = \mu_{N_h}. \end{array}\right. \tag{36}$$

Given the change of variable

$$\forall k \in [\![1, N_h - 1]\!] \left\{\begin{array}{l} D_0^k = \delta_0^k - \delta_0^{k+1}, \\ D_1^k = \delta_1^k - \delta_1^{k+1}, \end{array}\right. \tag{37}$$

the system (36) gives a linear system with unknowns $(D_0^k, D_1^k)_{k\in[\![1,N_h-1]\!]}$. Defining the Wronskien $W_0 = U_1 \frac{d}{dx} U_0 - U_0 \frac{d}{dx} U_1$ - which is non zero - the solution is

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

Then the structure of the system (36) 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

$$\left\{\begin{array}{l} \alpha\delta_0^{N_h} + \beta\delta_1^{N_h} = L, \\ \gamma\delta_0^{N_h} + \eta\delta_1^{N_h} = \mu_{N_h}, \end{array}\right. \tag{38}$$

with

$$L = \lambda_0 - (-\partial_\nu + i\gamma)U_0(a) \sum_{k=1}^{N_h-1} D_0^k - (-\partial_\nu + i\gamma)U_1(a) \sum_{k=1}^{N_h-1} D_1^k. \tag{39}$$

The determinant of the system (38) 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

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

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. We finally obtain that

$$\begin{cases} \delta_0^{N_h} = \dfrac{1}{W_1}\Big(L(\partial_\nu + i\gamma)U_1(b) - \mu_{N_h}(-\partial_\nu + i\gamma)U_1(a)\Big), \\ \delta_1^{N_h} = \dfrac{1}{W_1}\Big(\mu_{N_h}(-\partial_\nu + i\gamma)U_0(a) - L(-\partial_\nu + i\gamma)U_1(a)\Big), \\ \forall k \in [\![1, N_h - 1]\!] \left| \begin{array}{l} \delta_0^k = \delta_0^{N_h} + \displaystyle\sum_{j=k}^{N_h-1} D_0^j, \\ \delta_1^k = \delta_1^{N_h} + \displaystyle\sum_{j=k}^{N_h-1} D_1^j. \end{array} \right. \end{cases} \tag{40}$$

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 (40). Since F is an isometry, and λ_k and μ_k are linear combinations of the components of Fb

$$\begin{cases} \forall k \in [\![0, N_h - 1]\!], |\lambda_k| \leq \sqrt{\gamma}\|b\|, \\ \forall k \in [\![1, N_h]\!], |\mu_k| \leq \dfrac{1}{\sqrt{\gamma}}\|b\|. \end{cases} \tag{41}$$

Thus from (37) and (41), with C depending on U_0, U_1, γ and W_0,

$$\begin{cases} \left|\displaystyle\sum_{k=1}^{N_h-1} D_0^k\right| \leq CN_h\|b\|, \\ \left|\displaystyle\sum_{k=1}^{N_h-1} D_1^k\right| \leq CN_h\|b\|. \end{cases}$$

From (39), $|L| \leq CN_h\|b\|$, and since $|\mu_{N_h}| \leq C\|b\|$ one has from (40)

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

and next for $k \in [\![1, N_h - 1]\!]$

$$\begin{aligned} \left|\delta_i^k\right| &\leq \left|\delta_i^{N_h}\right| + \sum_{k=1}^{N_h-1} \left|D_i^k\right|, \\ &\leq CN_h\|b\|. \end{aligned}$$

Then all δ terms satisfy for $i \in \{0, 1\}$ and $k \in [\![1, N_h]\!]$

$$\left|\delta_i^k\right| \leq CN_h\|b\|.$$

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

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

so that

$$\|x\|\le Ch^{-3/2}\|b\|.$$

□

Definition 3.1. *Let us define the norm* $|||\cdot|||_q$

$$|||x|||_q=\|(I-A^q)x\|,\qquad \forall x\in V.$$

Proposition 3.1. *Let q be given. There exists a constant $C>0$ such that*

$$C(h^{3/2}-h^{q+1})\|x\|\le |||x|||_q, \forall x\in V. \tag{42}$$

Proof. One has

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

So

$$\|(I-A)x\|_V-Ch^{q+1}\|x\|\le\|(I-A^q)x\|, \forall x\in V.$$

Then lemma 3.3 concludes the proof. □

Proposition 3.2. *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. $\forall h\le h_1$*

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

Proof. One has

$$|||x|||_q^2\le\|x\|^2-2\mathcal{R}\left(A_qx,x\right)+\|A_qx\|^2.$$

Since

$$\|A_qx\|\le\|Ax\|+\|(A_q-A)x\|\le(1+Ch^{q+1})\|x\| \tag{43}$$

there exists another constant denoted as $C'>0$ such that

$$\|A_qx\|^2\le(1+C'h^{q+1})\|x\|^2.$$

Therefore

$$|||x|||_q^2\le 2\|x\|^2+C'h^{q+1}\|x\|^2-2\mathcal{R}\left(A_qx,x\right),$$

that is

$$|||x|||_q^2-C'h^{q+1}\|x\|^2\le 2\mathcal{R}\left(a_q(x,x)\right).$$

For small h since $q>3/2$ and due to the proposition 3.1 one has

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

then

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

Combined with the previous inequality it proves the claim. □

3.5 Convergence

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

Theorem 3.3. *Suppose that $q \geq 2$ and $h \leq \min(h_0, h_1)$. Denote $x \in V$ the solution of the exact problem (12) in dimension one and $x_h \in V$ the solution of the discrete problem (31). Then there exists a constant $C > 0$ such that*

$$|||x - x_h|||_q \leq Ch^{-3/2} \inf_{y_h \in V} |||x - y_h|||_q + 3 \sup_{w_h \in V - \{0\}} \frac{|a_q(x, w_h) - f_q(w_h)|}{\|w_h\|}, \tag{44}$$

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

The proof relies on the following intermediate result already proved in (43)

Lemma 3.4. *The operator F^q satisfies $\|A^q\| \leq 1 + Ch^{q+1}$.*

Proof. Of theorem 3.3

- The first remark is the uniform coercivity with respect to $|||.|||_q$ needed in the second Strang lemma. It is proved in proposition 3.2.
- The second step consists in characterizing the uniform continuity of a_q. For all $(x, y) \in V^2$

$$\begin{aligned} |a_q(x, y)| &= |((I - A^q)x, y)|, \\ &\leq |||x|||_q \; \|y\|, \\ &\leq \frac{1}{C(h^{3/2} - h^{q+1})} |||x|||_q |||y|||_q \end{aligned}$$

so that there exists a constant C such that for small h

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

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

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

On the other hand proposition 2.3 shows that

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

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

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

Finally we minimize the first error term with respect to y_h. It yields the desired result.

□

We now have to estimate the error defined by

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

Lemma 3.5. *There exists a constant $C > 0$ such that*

$$\forall w_h \in V - \{0\}, \; \frac{D_h(x, w_h)}{\|w_h\|} \leq Ch^{q+1} \|x\|. \tag{45}$$

Proof.

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

since $(I - A)x = b$. This gives exactly (45). □

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

Theorem 3.4. *One has the estimation*

$$|||x - x_h|||_q = O(h^{q+1}). \tag{46}$$

Proof. In dimension one the discrete space of approximation is equal to V whatever is the method of construction of basis functions. This is why we can choose $y_h = x$ in (44). So $\inf_{y_h \in V} |||x - y_h|||_q = 0$. The remaining term is bounded with (45). □

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 (42) we get $\overline{\|z\|} \leq Ch^{-1}|||z|||_q$. Therefore a corollary of the theorem is the estimate of convergence

$$\overline{\|x - x_h\|} = O(h^q). \tag{47}$$

Numerical experiments show that this estimate is optimal for q even, and under-optimal by a factor one for q odd.

4 Numerical examples

All the following examples are linked with Airy functions since it is the physical problem (3)-(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 the exact solutions.

All the linear systems are assembled and inversed with Matlab.

4.1 One dimensional test case

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

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

The discretization of the domain is

$$x_k = a + \frac{b - a}{N_h}(k - 1),\ \forall k \in [\![1, N_h + 1]\!],$$

where N_h stands for the number of elements defining the mesh and Ω_k denotes $]x_k, x_{k+1}[$, so that the mesh is uniform. For a given value of q the basis functions are designed as in paragraph 2.5.2. The solution computed corresponds to an element $x_h \in V$. A simple calculus permits to express the approximation u_h of the initial unknown u. In fact, since :

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

	q=2		q=3		q=4		q=5		q=6	
N	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

Figure 2: Errors and orders of convergence for different orders of approximation q depending on the number of unknowns N.

the discrete solution u_h satisfies :

$$\left\{\begin{array}{l} 2i\gamma u_{h|\Omega_1} = g + \displaystyle\sum_{j\in J(1)} (x_h)_j(-\partial_\nu + i\gamma)\varphi_j, \\ 2i\gamma u_{h|\Omega_k} = \displaystyle\sum_{j\in J(k)} (x_h)_j(-\partial_\nu + i\gamma)\varphi_j + \sum_{j\in J(k-1)} (x_h)_j(-\partial_\nu + i\gamma)\varphi_j, \forall k \in [\![2, N_h - 1]\!], \\ 2i\gamma u_{h|\Omega_{N_h}} = g + \displaystyle\sum_{j\in J(N_h)} (x_h)_j(-\partial_\nu + i\gamma)\varphi_j, \end{array}\right.$$

where, for all $k \in [\![1, N_h]\!]$, $J(k)$ denotes the set of indexes of basis functions supported in Ω_k. As a consequence, for all $k \in [\![1, N_h]\!]$

$$\left\{\begin{array}{rl} 2i\gamma u_h(x_1) = & (x_h)_{\Omega_1,1}.\left(\varphi'_{\Omega_1,1}(x_1) + i\gamma\varphi_{\Omega_1,1}(x_1)\right) \\ & +(x_h)_{\Omega_1,2}.\left(\varphi'_{\Omega_1,2}(x_1) + i\gamma\varphi_{\Omega_1,2}(x_1)\right) + g(x_1), \\ 2i\gamma u_h(x_{N_h+1}) = & (x_h)_{\Omega_{N_h},1}.\left(-\varphi'_{\Omega_{N_h},1}(x_k) + i\gamma\varphi_{\Omega_{N_h},1}(x_{N_h+1})\right) \\ & +(x_h)_{\Omega_{N_h},2}.\left(-\varphi'_{\Omega_{N_h},2}(x_{N_h+1}) + i\gamma\varphi_{\Omega_{N_h},2}(x_{N_h+1})\right) + g(x_{N_h+1}), \\ 2i\gamma u_h(x_k) = & \sum_{\delta\in\{1,2\}}(x_h)_{\delta,\Omega_{k-1}}\left(-\varphi'_{\delta,\Omega_{k-1}}(x_k) + i\gamma\varphi_{\delta,\Omega_{k-1}}(x_k)\right) \\ & +\sum_{\delta\in\{1,2\}}(x_h)_{\delta,\Omega_k}\left(-\varphi'_{\delta,\Omega_k}(x_k).(-1) + i\gamma\varphi_{\delta,\Omega_k}(x_k)\right), \end{array}\right.$$

this last line standing only for $k \in [\![2, N_h - 1]\!]$. In all simulations, the accuracy is reported using a discrete l^2 norm so that the relative error is computed as

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

Considering the domain $\Omega =]-5, 5[$, one gets the results described in figures 2 and 3. The rates of convergence are equal to the theoretical estimates for q even, and better (h^{q+1} instead of h^q) for q odd.

4.2 About q convergence

On figure 3, when the number of nodes is fixed, the error decreases when the parameter $q \geq 2$ increases.The classical method with plane waves corresponds to $q = 1$. To obtain better understanding of this phenomenon, we plot in figure 4 for different values of q and around two points

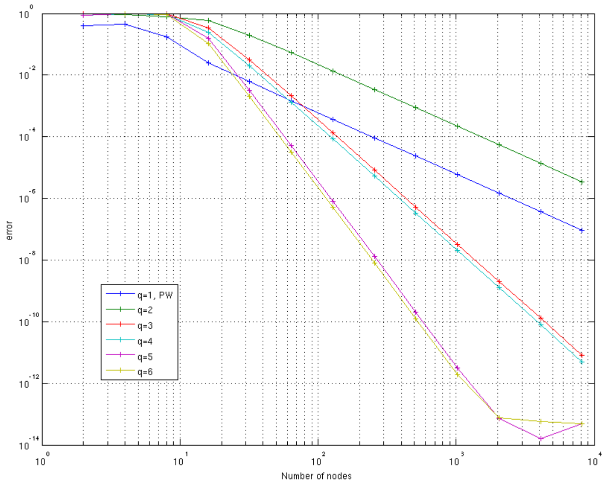


Figure 3: Convergence of the method increasing the parameter q, relative discrete L^2 error as a function of the number of elements defining the mesh.

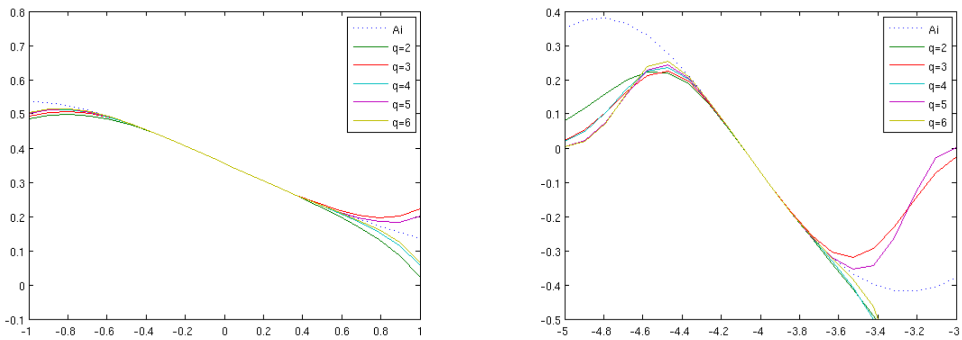


Figure 4: Approximation of Airy function by corresponding basis functions for different values of q, in the vicinity of $x_0 = 0$ and $x_0 = -4$.

x_0 the Airy function and its approximations thanks to the two basis functions φ constructed in section 2.5.2.

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

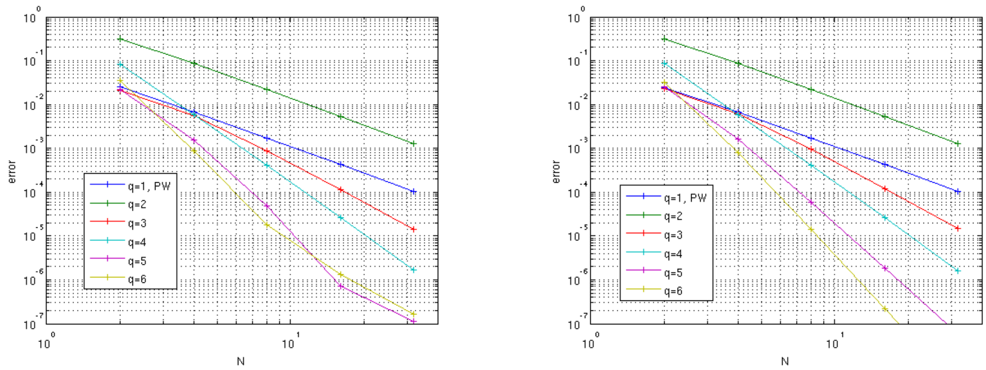
4.3 Two dimensional test case

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

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

so that the exact solution is again the Airy function Ai. The domain considered here is square and meshed with regular triangles.

As explained in section 2.6.1, 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 has the form $\varphi(x,y) = e^{P(x,y)}$ with $P(x,y) = p(x) + \theta y$ and θ still to be defined. In practise we



(a) With Simpson quadrature formulas. (b) With Boole quadrature formulas.

Figure 5: 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.

	q=2		q=3		q=4		q=5		q=6	
N	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

Figure 6: Errors and orders of convergence depending on the number of unknowns N for the two dimensional case with Simpson quadrature formulas.

chose

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

Then r has to be odd since for even values, $k = r/2$ and r would give the same value of θ so that the resulting family of basis functions would no more be independent. For each θ the corresponding functions P_+ and P_- are constructed as in the one dimensional case.

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, for $r = 3$ in figures 5, 6 and 7. One can observe a clear improvement in the results obtained using Boole formulas compared to the results obtained using Simpson formulas.

	q=2		q=3		q=4		q=5		q=6	
N	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

Figure 7: Errors and orders of convergence depending on the number of unknowns N for the two dimensional case with Boole quadrature formulas.

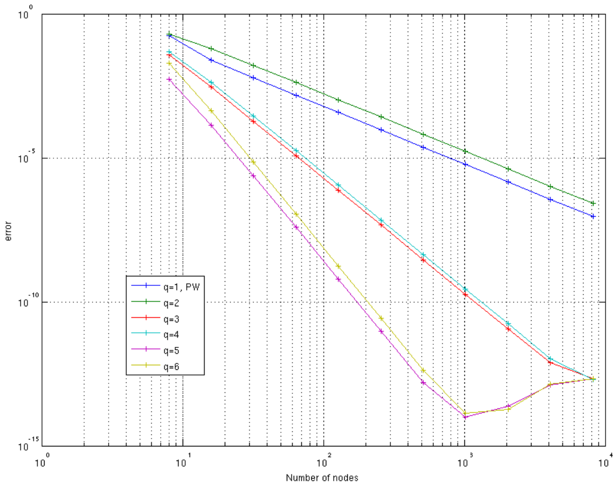


Figure 8: Relative discrete L^2 error as a function of the number of elements defining the mesh, using the normalization $\beta_{1,\pm} = \pm\sqrt{\alpha(x_{k+1/2})}$. Different curves correspond to increasing order parameter q.

4.4 Other basis functions

Figures 8 and 9 present the numerical convergence results obtained with basis functions designed with the normalization $\beta_{1,\pm} = \pm\sqrt{\alpha(x_{k+1/2})}$. Comparing to figures 2 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 2.5.2. In fact, for a given order q, the numerical results show that the constant underlying in estimation (47) 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 developped previously can be adapted without difficulty to this new family of basis functions but the vertical shift visible on figures 3 to 8 will require more research to be fully understood.

	q=2		q=3		q=4		q=5		q=6	
N	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

Figure 9: Errors and orders of convergence for different orders of approximation q depending on the number of unknowns N, using the normalization $\beta_{1,\pm} = \pm\sqrt{\alpha(x_{k+1/2})}$.

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

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

To prove the next result, 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 [22] p.92. To simplify here Q is constant.

Theorem A.1. *Let Ω be a bounded domain in $\mathbb{R}$ with a Lipschitz and piecewise $\mathcal{C}^2$ 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)*, i.e. such that*

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

Proof. This very classical result will be used in the following.
This proof relies on classical methods for variational formulations. Let us introduce an intermediate problem

$$\begin{cases} -\Delta w + w &= f_i, (\Omega), \\ (\partial_\nu + i\zeta\gamma)\, w &= g_i, (\Gamma). \end{cases}$$

Let a and l be the corresponding sesquilinear and antilinear forms, so that for any u and v in $H^1(\Omega)$

$$a(u,v) = \int_\Omega \nabla u \cdot \overline{\nabla v} + \int_\Omega u\overline{v} + i\zeta \int_\Gamma u\overline{v}, \qquad l(v) = \int_\Omega f_i\overline{v} + \int_\Gamma g_i\overline{v}.$$

As a is sesquilinear and continuous, b is antilinear continuous and $Re(a(v,v))$ is coercive, there exists a unique $u \in H^1$ such that

$$a(u,v) = l(v), \forall v \in H^1, \tag{48}$$

for any couple $(g_i, f_i) \in L^2(\Omega) \times L^2(\Gamma)$. See [6] for this version of Lax-Milgram theorem. Then let us define the linear operator A by

$$A : (f_i, g_i) \in L^2(\Omega) \times L^2(\Gamma) \mapsto u \in L^2(\Omega),$$

where u is the solution given by (48). Moreover, we can notice that from a classical a priori estimate we have $\|u\|_{H^1} \leq \|f_i\|_{L^2} + \|g_i\|_{L^2(\Gamma)}$. A is compact since the injection of H^1 in L^2 is compact as $\Omega \subset \mathbb{R}$. We remark that

$$\begin{aligned} u \text{ is solution of (1)} &\Leftrightarrow u = A\left((id-\alpha)u + f, \frac{1}{1+Q}g\right), \\ &\Leftrightarrow [I - A((id-\alpha)\cdot, 0)]\, u = A\left(f, \frac{1}{1+Q}g\right). \end{aligned}$$

Since α is bounded, the operator $K := A((id-\alpha)\cdot, 0))$ is also compact, and we are in the frame of Fredholm alternative, see [1]. So uniqueness is equivalent to existence of a solution for the problem (1). Then suppose $u \in L^2$, actually also in H^1, is such that $(I-K)u = 0$, which means

$$\int_\Omega \nabla u \cdot \overline{\nabla v} + \int_\Omega \alpha u \overline{v} + i\zeta \int_\Gamma u\overline{v} = 0, \forall v \in H^1. \tag{49}$$

Choosing $v = u$ as test function, and taking the imaginary part of (49) we get that $u = 0$ on Γ.The regularity of Γ is such that $\partial_\nu u = 0$. A unique continuation principle given in [22] shows that there exists a unique solution to (1). □

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

A.2 Proof of inequality (27)

We will 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 $\mathcal{O} \subset \mathbb{R}$, for all $u \in L^2(\mathcal{O})$*

$$\|u\|_{L^2(\mathcal{O})} \leq C\left(\sqrt{h}\|u\|_{L^2(\partial\mathcal{O})} + h\|u'\|_{L^2(\mathcal{O})}\right) \tag{50}$$

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

Proof. We will show a more general inequality than (27) . We use u as test function in the variational formulation (A.1) corresponding to the following problem

$$\begin{cases} -u'' + \beta u = f, & (\mathcal{O}) \\ (-\partial_\nu + i\gamma)u = g, & (\partial\mathcal{O}). \end{cases} \tag{51}$$

One gets

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

We obtain

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

The first inequality yields

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

A standard inequality yields

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

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

Inserting in the second inequality we obtain

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

Then from (50)

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

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

For h small enough we obtain

$$\|u\|^2_{L^2(\mathcal{O})}\le C\left(\frac{h}{\gamma^2}\|g\|^2_{L^2(\partial\mathcal{O})}+\frac{h^2}{\gamma^2}\|f\|^2_{L^2(\mathcal{O})}\right). \tag{52}$$

One can notice that the dimension of this estimate. Considering that γ is the dimension of the inverse of a length which is evident from the boundary conditon, all quantities have the same dimension at inspection of (51). Inequality (27) is obtained by taking $f = 0$ in the previous inequality. □

A.3 Proof of of theorem 3.2

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

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

and

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

Then $e_h := u - u_h$ satisfies

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

Inequality (52) yields

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

Using one more time (52) to estimate u and regarding γ which is a positive number, we get

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

□

References

[1] H. Brézis, *Analyse Fonctionnelle.* Dunod, 1983. (version anglaise p 299)

[2] A. Buffa and P. Monk, Error estimates for the Ultra Weak Variational Formulation of the Helmholtz equation, ESAIM: Mathematical Modelling and Numerical Analysis November 2008, 42: 925-940.

[3] O. Cessenat and B. Després, Using plane waves as base functions for solving time harmonic equations with the ultra weak variational formulation, Journal of Computational Acoustics, 2003

[4] O. Cessenat, B. Després, *Application of an ultra weak variational formulation of elliptic PDEs to the two dimensional Helmholtz problem.* SIAM J. Numer. Anal. 1998, vol. 55, no1, pp. 255-299.

[5] B. Després, *Sur une formulation variationnelle de type ultra-faible.* C. R. Acad. Sci. Paris Sér. I Math. 318 1994, no. 10, 939–944.

[6] R. Dautray, J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology.* Masson 1984, volume 3, chapitre 8.

[7] C. Farhat, I.Harari, and L. Franca, The discontinuous enrichment method, Computer Methods in Applied Mechanics and Engineering, 190 (2001), pp. 6455-6479.

[8] Farhat C, Tezaur R, Wiedemann-Goiran P, Higher-order extensions of a discontinuous Galerkin method for mid-frequency Helmholtz problems. International Journal for Numerical Methods in Engineering 2004; 61: 1938-1956.

[9] Farhat C, Tezaur R, Toivanen J. A domain decomposition method for discontinuous Galerkin discretizations of Helmholtz problems with plane waves and Lagrange multipliers. International Journal for Numerical Methods in Engineering 2009; 78:1513-1531.

[10] Gabard G, Gamallo P, Huttunen T, A comparison of wave-based discontinuous Galerkin, ultra-weak and least-square methods for wave problems. International Journal for Numerical Methods in Engineering 2011; 85:380-402.

[11] Gittelson, Claude J, Hiptmair, Ralf and Perugia, Ilaria, Plane wave discontinuous Galerkin methods: Analysis of the h-version, ESAIM: Mathematical Modelling and Numerical Analysis, 2009, 43: 297-331

[12] R. Hiptmair, A. Moiola, and I. Perugia, Plane wave discontinuous Galerkin methods for the 2D Helmholtz equation: analysis of the p-version, Preprint 2009-20, SAM Report, ETH Zürich, Switzerland, 2009.

[13] R. Hiptmair, A. Moiola and I. Perugia, Error analysis of Trefftz-discontinuous Galerkin methods for the time-harmonic Maxwell equations, ETHZ, Research Report No. 2011-09.

[14] Huttunen T, Gammalo P, Astley RJ, Comparison of two wave element methods for the Helmholtz problem. Communications in Numerical Methods in Engineering 2009; 25:35-52.

[15] Kalashnikova I, Tezaur R, Farhat C, A discontinuous enrichment method for variable coefficient advection-diffusion at high Péclet number. International Journal for Numerical Methods in Engineering 2010; 87:309-3356

[16] T. Huttunen, M. Malinen and P. Monk, Solving Maxwell's equations using the ultra weak variational formulation, Journal of Computational Physics Volume 223, Issue 2, 1 May 2007, Pages 731-758.

[17] Tomi Huttunen, Peter Monk and Jari P. Kaipio, Computational Aspects of the Ultra-Weak Variational Formulation, Journal of Computational Physics Volume 182, Issue 1, 10 October 2002, Pages 27-46.

[18] J. Melenk, On Generalized Finite Element Methods, PhD thesis, University of Maryland, USA, 1995.

[19] J. Melenk, Operator adapted spectral element methods I: harmonic and generalized harmonic polynomials, Numerische Mathematik, 84 (1999), pp. 35-69.

[20] Melenk JM, Babuska I. The partition of unity method finite element method: basic theory and applications. Computer Methods in Applied Mechanics and Engineering 1996; 139:289-314.

[21] J.M. Melenk and S. Sauter, Wavenumber explicit convergence analysis for Galerkin discretizations of the Helmholtz equation SIAM J. Numer. Anal. 49 (2011), pp. 1210–1243.

[22] P. Monk, *Finite element methods for Maxwell's equations.* Calderon press Oxford 2003.

[23] E. Perrey-Debain, O. Laghrouche, P. Bettess and J. Trevelyan, Plane-wave basis finite elements and boundary elements for three-dimensional wave scattering, Phil. Trans. R. Soc. Lond. A 2004 362, 561-577.

[24] B. Pluymers, W. Desmet, D. Vandepitte, P. Sas, Wave based modelling methods for steady-state interior acoustics: an overview, ISMA, 2006.

[25] Strouboulis T, Babuska I, Hidajat R. The generalized finite element method for Helmholtz equation: theory, computation, and open problems. Computational Methods in Applied Mechanical Engineering 2006; 195:4711-4731.

[26] D. G. Swanson, Plasma Waves, 2nd Edition, Series in Plasma Physics, 2003.