

Interpolation properties of generalized plane waves

Lise-Marie Imbert-Gérard

Received: 30 October 2013 / Accepted: 6 October 2014 / Published online: 15 January 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract This paper is aimed at developing new shape functions adapted to the scalar wave equation with smooth (possibly vanishing) coefficients and investigates the numerical analysis of their interpolation properties. The interpolation is local, but high order convergence is shown with respect to the size of the domain considered. The new basis functions are then implemented in a numerical method to solve a scalar wave equation problem with a mixed boundary condition. The main theoretical result states that any given order of approximation can be achieved by an appropriate choice of parameters for the design of the shape functions. The convergence is studied with respect to the size of the domain, which is referred to in the literature as h -convergence.

Mathematics Subject Classification 65D05 · 65N99

1 Introduction

This paper focuses on the design of Generalized Plane Waves (GPW) to approximate smooth solutions $u \in C^\infty(\Omega)$ of the model problem

$$-\Delta u + \beta u = 0, \quad \text{in } \Omega \subset \mathbb{R}^2, \quad (1)$$

where β is in $C^\infty(\Omega)$. This time-harmonic equation, generally called the scalar wave equation, models for instance the acoustic pressure describing the behavior of sound

This work was supported by the Applied Mathematical Sciences Program of the U.S. Department of Energy under contract DEFG0288ER25053.

L.-M. Imbert-Gérard (✉)
Courant Institute of Mathematical Sciences, NYU, 251 Mercer Street, New York, NY, USA
e-mail: imbertgerard@cims.nyu.edu

in matter or a polarized electromagnetic wave propagating in an isotropic medium. If $\beta = -\omega^2$, $\omega \in \mathbb{R}$, the equation is the classical Helmholtz equation and is still the subject of recent research, see for instance [27]. If $\beta < 0$ is non constant, this is a simple model of wave propagation in an inhomogeneous medium. If $\beta > 0$, it models an absorbing medium and the partial differential equation is coercive. The applications considered here include both propagative and absorbing media, as well as smooth transitions in between them, i.e. respectively $\beta > 0$, $\beta < 0$ and $\beta = 0$.

Several types of numerical methods are used for the simulation of wave propagation. Classical finite element methods applied to such problems are known to be polluted by dispersion, see [1]. An alternative is to consider approximation methods based on shape functions that are local solutions of the homogeneous equation: this justifies the development of Trefftz-based methods, first introduced in [32], that rely on solutions of the homogeneous governing domain equation: information about the problem is embedded in the finite basis. The present work originated from the idea of applying such a method to a problem modeled by (1) in which the coefficient is likely to vanish: shape functions adapted to this problem are here designed and studied. See the previous work [22] for the physical motivation of the problem. We refer to [29] and references therein for more recent developments of these Trefftz-based methods, and to [12, 17] for applications linked to one specific method, the so-called Ultra Weak Variational Formulation (UWVF). The method coupling the latter to the adapted shape functions is the topic of [22], and the present work includes numerical results for the h convergence of the coupled method.

The novelty in the present lies in the fact that the shape functions, called Generalized Plane Waves (GPWs), are selected from a smooth coefficient β of the governing domain Eq. (1). This work can be compared to recent progress that focus on non polynomial methods for smooth varying coefficients, see for instance [3, 31]. The design of shape functions adapted to smooth and possibly vanishing coefficients, that is the core of this work, starts from mimicking the equation

$$(-\Delta + \beta)e^{i\omega \vec{k} \cdot \vec{x}} = \left(-(\iota\omega \|\vec{k}\|)^2 - \omega^2\right)e^{i\omega \vec{k} \cdot \vec{x}} = 0,$$

where $\iota = \sqrt{-1}$. It shows that classical plane waves functions $e^{i\omega \vec{k} \cdot \vec{x}}$ are exact solutions of (1) when $\beta = -\omega^2$ is constant and negative.

The case of a piecewise constant coefficient is addressed for example in [5, 11], and the more general case of a smooth coefficient is generally approximated by a piecewise constant coefficient. A very simple extension of classical plane waves for a positive or negative constant coefficient would be to consider at a point $\vec{g} = (x_0, y_0) \in \mathbb{R}^2$ the shape function

$$\varphi(x, y) = \exp\left(\sqrt{\operatorname{sgn}(\beta(\vec{g}))} \sqrt{|\beta(\vec{g})|} ((x - x_0) \cos \theta + (y - y_0) \sin \theta)\right), \quad (2)$$

where the parameter θ represents the direction of the plane wave. Indeed, the case $\beta(\vec{g}) < 0$ corresponds to the classical plane wave whereas the case $\beta(\vec{g}) > 0$ corresponds to a purely imaginary wavenumber. This choice will provide a tool to extend the interpolation results cited previously. Note that if the coefficient $\beta(\vec{g}) < 0$ goes to zero then the corresponding classical plane waves functions, generated by equispaced directions θ , tend not to be independent anymore.

To address the general case of a smooth coefficient, the idea developed in the present work is to design approximate solutions of the governing equation, in the form of exponential of polynomials: the GPWs. Typically, in the case $\beta(x, y) = x$ the Airy functions A_i and B_i are exact solutions; however, in the general case there is no exact analytic solution known. Indeed, as was explained in Section 2.1 of [22], no exponential of a polynomial can solve a generic scalar wave equation. It is then natural to generalize the classical plane waves by approximate solutions. A parameter q will describe the order of this approximation. The explicit design of those GPWs is the first concern of this paper. It is key to prove the interpolation property.

The desired interpolation property aims at approximating any smooth solution of Eq. (1) using a basis of GPWs. It is the second concern of this paper. The goal of the theoretical part of this paper is to prove high order approximation properties on such sets of basis functions, provided that a sufficient number of basis functions is used with respect to the approximation parameter q . It will be stated with more precision in Theorem 1, and can be announced/summarized as follows.

Claim Assume $n \in \mathbb{N}$, standing for an interpolation parameter and $\vec{g} = (x_0, y_0)$, standing for a point in a bounded domain Ω . Suppose that the coefficients of the Taylor expansion

$$\begin{aligned} \forall \vec{m} = (x, y) \in \Omega, \beta(\vec{m}) &= \sum_{0 \leq i+j < n+1} \frac{\partial_x^i \partial_y^j \beta(\vec{g})}{i!j!} (x - x_0)^i (y - y_0)^j + O(|\vec{m} - \vec{g}|^{n+1}) \end{aligned}$$

are given, where β is the coefficient of the scalar wave equation (1). Consider a smooth solution $u \in C^{n+1}(\Omega)$ of Eq. (1). There exists a linear combination of GPWs u_a which is an approximation of order $n + 1$ of u in the following sense: there is a constant C_Ω such that for all $\vec{m} \in \mathbb{R}^2$

$$\begin{cases} |u(\vec{m}) - u_a(\vec{m})| \leq C_\Omega |\vec{m} - \vec{g}|^{n+1} \|u\|_{C^{n+1}}, \\ \|\nabla u(\vec{m}) - \nabla u_a(\vec{m})\| \leq C_\Omega |\vec{m} - \vec{g}|^n \|u\|_{C^{n+1}}. \end{cases}$$

Note that this result is local; it is stated at a given point $\vec{g} \in \mathbb{R}^2$. Theorem 1 specifies a sufficient number of GPWs to achieve an arbitrarily high order of convergence. It can be related to what is called in the literature Hermite-Birkhoff interpolation thanks to GPWs, see [30].

Also note that in the literature of finite element or plane wave numerical methods, such a result is referred to as h -convergence as it provides a rate of convergence with respect to h if $|\vec{m} - \vec{g}| \leq h$. It is opposed to the p -convergence which focuses on the convergence with respect to the number of basis functions. A first glance at the p -version is proposed following the proof of the Theorem.

There are two main approaches in proving such interpolation results for Helmholtz equation that have been developed in the literature. One of them is based on Vekua theory, which was first translated into English in [14] for functions in \mathbb{R}^2 . A more recent introduction to the topic can be found in [2]. Theoretical studies based on this technical tool can be found in [25], and more recently in [28]. In the latter, the case of the Helmholtz equation with constant coefficient is explicitly studied and

interpolation properties are obtained with explicit dependence with respect to the parameters. However, even if this theory is powerful, in the case of a smooth coefficient it gives no explicit estimates with respect to the different parameters. The second main stream method relies on Taylor expansions, and was proposed in [5]. Since the design of solutions developed in this paper is based on Taylor expansions as well, this second approach will be the one followed here.

Section 2 describes precisely the design process of the GPWs. The approximation property is obtained subject to constraints on the function's coefficients. Then a set of approximated functions is obtained by specifying these constraints, mimicking the classical plane wave functions. The notion of a GPW is made precise in Definition 1. Section 3 focuses on the main approximation result of the paper, namely Theorem 1. This theorem states a theoretical framework to obtain the approximation properties presented in the previous claim. Its proof relies on fundamental properties of the GPWs. Section 4 presents a series of numerical validations of Theorem 1. It considers two different normalizations of the GPWs. The numerical test cases are chosen to consider problems linked with reflectometry, a radar diagnostic technique for fusion plasma, see [22] for more details.

Notation. The complex number $\sqrt{-1}$ is denoted by i . The symbol $l \in \llbracket n, m \rrbracket$ stands for the statement: $l \in \mathbb{N}$ such that $n \leq l \leq m$. The symbol \mathcal{C}^n represents the set of functions whose derivatives up to the order n exist and are continuous.

2 Design of the generalized plane waves

A GPW is designed as a function $\varphi = e^P$, where P is a complex polynomial in two variables, which satisfies *locally* an approximated version of the governing Eq. (1): $(-\Delta + \beta)e^P \approx 0$. This section is split into two parts: a first one dedicated to the design of a GPW as an approximate solution to the scalar wave Eq. (1), and a second one dedicated to the construction of a basis of such approximate solutions. While the design process depends on a single approximation order denoted q , the construction of a basis depends on two additional parameters: the number p of basis functions as well as a normalization parameter denoted N .

Definition 1 states a precise definition of a GPW.

2.1 An approximate solution

Since the design process is local, denote by $\vec{g} = (x_0, y_0)$ a point in \mathbb{R}^2 . Consider a complex polynomial

$$P(x, y) = \sum_{i=0}^{dP} \sum_{j=0}^{dP-i} \lambda_{i,j} (x - x_0)^i (y - y_0)^j, \quad (3)$$

together with the associated function $\varphi = e^P$. In order for φ to be a GPW, its image by the operator $-\Delta + \beta$ has to be small. Since β is not constant but smooth, P will be designed to ensure that

$(-\Delta + \beta)e^P = (- (\partial_{x^2}P + (\partial_x P)^2 + \partial_{y^2}P + (\partial_y P)^2) + \beta)e^P$ is *locally* small.

To simplify the notation, define the polynomial $P_\Delta = \partial_{x^2}P + (\partial_x P)^2 + \partial_{y^2}P + (\partial_y P)^2$, so that $(-\Delta + \beta)e^P = (-P_\Delta + \beta)e^P$. Since e^P is *locally* close to $e^{\lambda_{0,0}} \neq 0$, the idea is to set $-P_\Delta + \beta$ to be small.

The design of the GPW is then performed by setting the coefficients of P to match the Taylor expansions of P_Δ and β : introducing an approximation parameter $q \in \mathbb{N}$, $q \geq 1$, the design process described in this paper is based on the *local* equation

$$- P_\Delta(x, y) + \beta(x, y) = O (\|(x, y) - (x_0, y_0)\|^q). \tag{4}$$

The specific case $q = 1$ corresponds to the simplest generalization of plane waves described previously in (2). Indeed, in that case (4) is reduced to only one equation, reading for $dP \geq 2$

$$-2\lambda_{2,0} - \lambda_{1,0}^2 - 2\lambda_{0,2} - \lambda_{0,1}^2 + \beta(\vec{g}) = 0.$$

So setting

$$(\lambda_{1,0}, \lambda_{0,1}) = \sqrt{\text{sgn}\beta(\vec{g})} \sqrt{|\beta(\vec{g})|} (\cos \theta, \sin \theta)$$

together with $\lambda_{0,2} = \lambda_{2,0} = 0$ corresponds to the shape function described in (2) and is sufficient to satisfy the approximation property (4) for $q = 1$. In the general case $q \geq 1$, the procedure includes choosing the degree of the polynomial and giving an explicit expression to compute the coefficients of the polynomial. These two choices are not independent. A precise analysis of Eq. (4) leads to choosing the degree of P such that the computation of the coefficients appears to be straightforward.

Consider now the coefficients of P as unknowns in the general case $q > 1$. In order to solve to ensure that Eq. (4) holds, consider a system described by:

- $N_{un} = \frac{(dP+1)(dP+2)}{2}$ unknowns, namely the coefficients of P ,
- $N_{eq} = \frac{q(q+1)}{2}$ equations, corresponding to the cancellation of the terms of degree lower than q in the Taylor expansion of $\beta - P_\Delta$.

As a result the system is overdetermined if $dP + 1 < q$, and in such a case the existence of a solution is not guaranteed. The idea is then to find the smallest value of $dP \geq q - 1$ that would provide an invertible system.

The case $dP = q - 1$ provides a square, but generally not invertible system. For instance consider the case $q = 2$ and $dP = 1$. Then (4) is reduced to three equations reading

$$\begin{cases} -\lambda_{1,0}^2 - \lambda_{0,1}^2 + \beta(\vec{g}) = 0, \\ \partial_x \beta(\vec{g}) = 0, \\ \partial_y \beta(\vec{g}) = 0. \end{cases}$$

Since no unknowns appear in the second and third equations this system obviously has no solution in the general case.

The case $dP = q$ is more intricate than the next one, since the q equation stemming from the terms of degree $q - 1$ have no linear term. In a such case, the system is underdetermined however there is no straightforward way to obtain an invertible system, because of the nonlinearity. For instance consider the case $dP = q = 2$. Then (4) is reduced to three equations reading

$$\begin{cases} -2\lambda_{2,0} - \lambda_{1,0}^2 - 2\lambda_{0,2} - \lambda_{0,1}^2 + \beta(\vec{g}) = 0, \\ -4\lambda_{2,0}\lambda_{1,0} + \partial_x\beta(\vec{g}) = 0, \\ -4\lambda_{0,2}\lambda_{0,1} + \partial_y\beta(\vec{g}) = 0. \end{cases}$$

This system is underdetermined, however for a general coefficient β it has no obvious solution because of the nonlinearity: finding a solution corresponds to computing non vanishing roots of a polynomial of degree 4 in two variables $(\lambda_{1,0}, \lambda_{0,1})$. So the case $dP = q$ does not—in general—lead to a convenient invertible system.

As for the case $dP = q + 1$, the system is underdetermined and the number of additional equations to be imposed to get a square system is $N_{un} - N_{eq} = 2q + 3$. Moreover, since

$$\beta(x, y) = \sum_{(i,j)/0 \leq i+j \leq q-1} \frac{\partial_x^i \partial_y^j \beta(x_0, y_0)}{i!j!} (x - x_0)^i (y - y_0)^j + O(\|(x, y) - (x_0, y_0)\|^q),$$

then the N_{eq} equations of the system that come from (4) actually reads

$$\begin{aligned} &\forall (i, j) \text{ s.t. } 0 \leq i + j \leq q - 1, \\ &\frac{\partial_x^i \partial_y^j \beta(\vec{g})}{i!j!} = (i + 2)(i + 1)\lambda_{i+2,j} + (j + 2)(j + 1)\lambda_{i,j+2} \\ &\quad + \sum_{k=0}^i \sum_{l=0}^j (i - k + 1)(k + 1)\lambda_{i-k+1,j-l}\lambda_{k+1,l} \\ &\quad + \sum_{k=0}^j \sum_{l=0}^i (j - k + 1)(k + 1)\lambda_{i-l,j-k+1}\lambda_{l,k+1}. \end{aligned} \tag{5}$$

As a consequence, to obtain an invertible system the choice proposed in this paper is to fix the set of coefficients $\{\lambda_{i,j}, i \in \{0, 1\}, j \in \llbracket 0, q + 1 - i \rrbracket\}$. Thus this choice corresponds to the $2q + 2$ additional constraints that, together with Eq. (5), form a square system. The next result states the existence and uniqueness of a solution to this square system.

Proposition 1 *Assume $\vec{g} = (x_0, y_0)$ is a point in \mathbb{R}^2 and the approximation parameter $q \in \mathbb{N}$ is such that $q \geq 1$. Finally suppose $\beta \in C^{q-1}$ at \vec{g} and consider the complex*

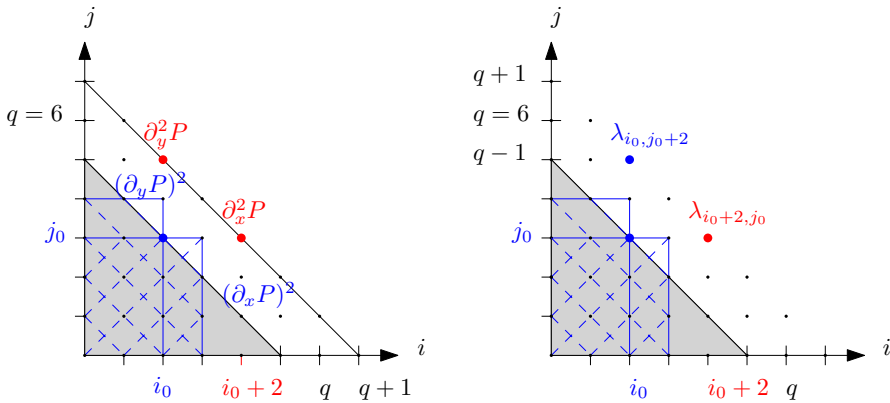


Fig. 1 For a given (i_0, j_0) , the left part of the figure shows the contributions from P_Δ to the $x^{i_0} y^{j_0}$ term in $\beta - P_\Delta$. The right part shows that λ_{i_0+2, j_0} can be explicitly expressed as long as $\lambda_{k,l}$ are known for all $k \leq i_0 + 1$ and $l \leq dP - 2 - k$

unknowns $\{\lambda_{i,j}, 0 \leq i + j \leq q + 1\}$. The system described by (5) together with the additional constraints of fixing the elements of $\{\lambda_{i,j}, i \in \{0, 1\}, j \in \llbracket 0, q + 1 - i \rrbracket\}$ has a unique solution, given by

$$\begin{aligned} \forall (i, j) \text{ s.t. } 0 \leq i + j \leq q - 1, \\ \lambda_{i+2,j} = \frac{1}{(i + 2)(i + 1)} \left(\frac{\partial_x^i \partial_y^j \beta(\vec{g})}{i!j!} - (j + 2)(j + 1)\lambda_{i,j+2} \right. \\ \left. - \sum_{k=0}^i \sum_{l=0}^j (i - k + 1)(k + 1)\lambda_{i-k+1,j-l}\lambda_{k+1,l} \right. \\ \left. - \sum_{k=0}^j \sum_{l=0}^i (j - k + 1)(k + 1)\lambda_{i-l,j-k+1}\lambda_{l,k+1} \right). \end{aligned} \tag{6}$$

Proof For any given set of coefficients $\{\lambda_{i,j}, i \in \{0, 1\}, j \in \llbracket 0, q + 1 - i \rrbracket\}$, the existence and uniqueness of a solution to (5) stems directly from the induction relation (6). See Fig. 1. □

As a direct consequence of Proposition 1, the choice $dP = q + 1$ provides an invertible system to compute the coefficients of P . Thanks to the preceding study, this choice $dP = q + 1$ is actually the smallest value of dP providing such an invertible system.

Note that for $dP > q + 1$, the system’s equations are still described by (5). However there are more unknowns than in the previous case, namely $dP = q + 1$. More precisely there are $N_{un} - N_{eq} = \frac{(dP+1)(dP+2)-q(q+1)}{2}$ extra unknowns: again fixing $\{\lambda_{i,j}, i \in \{0, 1\}, j \in \llbracket 0, q + 1 - i \rrbracket\}$ uniquely defines the set of coefficients $\{\lambda_{i,j}, 0 \leq i + j \leq q + 1\}$ since the induction formula (6) still holds. But the remain-

ing coefficients $\{\lambda_{i,j}, q + 1 < i + j \leq dP\}$ do not appear in the equations: they need to be computed even if they are not involved in the approximation Eq. (4).

Consequently from now on the polynomial P will be of degree $q + 1$. The following result summarizes how to design a GPW, providing $2q + 1$ degree of freedom to do so.

Corollary 1 Assume $\vec{g} = (x_0, y_0)$ is a point in \mathbb{R}^2 and the approximation parameter satisfies $q \in \mathbb{N}$ such that $q \geq 1$. Finally suppose $\beta \in C^{q-1}$ at \vec{g} . Consider a given set of complex numbers $\{\lambda_{i,j}, i \in \{0, 1\}, j \in \llbracket 0, q + 1 - i \rrbracket\}$, and the corresponding coefficients $\{\lambda_{i,j}, 0 \leq i + j \leq q + 1\}$ computed by the induction formula (6). The function $\varphi(x, y) = \exp \sum_{0 \leq i+j \leq q+1} \lambda_{i,j} (x - x_0)^i (y - y_0)^j$ satisfies $(-\Delta + \beta(x, y))\varphi(x, y) = O(\|(x, y) - (x_0, y_0)\|^q)$.

The explicit choice for the set of coefficients $\{\lambda_{i,j}, i \in \{0, 1\}, j \in \llbracket 0, q + 1 - i \rrbracket\}$ that will lead to the explicit definition of a GPW will include a parameter meant to generate not only one, but a set of independent GPWs at a single point \vec{g} .

2.2 A set of approximate solutions

The construction of a set of approximate solutions will take advantage of the degrees of freedom available to design an approximate solution, provided by Corollary 1. Inspired by the construction of classical plane waves, the general design procedure proposed in this paper involves two parameters:

- a parameter θ corresponding to the direction of the wave,
- a parameter $N \neq 0$, where $-iN$ can be interpreted as the local wave number of the wave.

These parameters are used to set $(\lambda_{1,0}, \lambda_{0,1}) = N(\cos \theta, \sin \theta)$. It justifies the name given to the new shape functions: generalized plane waves. Varying θ then provides different functions φ , only as long as $N \neq 0$.

The additional constraints that correspond the coefficients $\{\lambda_{i,j}, i \in \{0, 1\}, j \in \llbracket 0, q + 1 - i \rrbracket\}$ will be fixed in the following way to give a precise definition of a GPW.

Definition 1 Assume $\vec{g} = (x_0, y_0)$ is a point in \mathbb{R}^2 , the approximation parameter satisfies $q \in \mathbb{N}$ such that $q \geq 1$, $\theta \in \mathbb{R}$ and $N \in \mathbb{C}$ such that $N \neq 0$. Finally suppose $\beta \in C^{q-1}$ at \vec{g} . A generalized plane wave adapted to the operator $-\Delta + \beta$ is a function $\varphi = \exp \sum_{0 \leq i+j \leq q+1} \lambda_{i,j} (x - x_0)^i (y - y_0)^j$ whose coefficients satisfy the induction formula (6) together with the additional constraints

- $\lambda_{0,0} = 0$,
- $(\lambda_{1,0}, \lambda_{0,1}) = N(\cos \theta, \sin \theta)$,
- $\lambda_{i,j} = 0$ for $i \in \{0, 1\}$ and $1 < i + j \leq q + 1$.

The first item prevents any blow up of the shape function linked to the exponential, since then $\varphi(\vec{g}) = e^{\lambda_{0,0}}$ is independent of \vec{g} and β , while the second item mimics classical plane waves. The last item is the simplest possible choice and is meant to simplify both the numerical computations—by a substantial decrease of basic operations necessary to evaluate a shape function—and the analysis of the method.

Remark 1 (Other possible choices) Other choices to obtain an invertible system would give the same theoretical results. For instance choosing to fix $\{\lambda_{i,j}, j \in \{0, 1\}, i \in \llbracket 0, q + 1 - j \rrbracket\}$ is possible as well. But numerically, as will be seen later on, there is no evidence of the lack of symmetry with respect to the two space variables.

The condition $N \neq 0$ is mandatory to define a set of linearly independent shape functions: they then form a basis of an approximation space $\mathcal{E}_{\vec{g}}(N, p, q)$, where p is the dimension of the space and q the order of approximation in (4).

Definition 2 Assume $\vec{g} = (x_0, y_0)$ is a point in \mathbb{R}^2 , the approximation parameter satisfies $q \in \mathbb{N}$ such that $q \geq 1$, $p \in \mathbb{N}$ such that $p \leq 3$ and $N \in \mathbb{C}$ such that $N \neq 0$. Finally suppose $\beta \in \mathcal{C}^{q-1}$ at \vec{g} . Consider then for all $l \in \llbracket 1, p \rrbracket$

- $\theta_l = 2\pi(l - 1)/p$ a direction, all directions being equi-spaced,
- $(\lambda_{1,0}^l, \lambda_{0,1}^l) = N(\cos \theta_l, \sin \theta_l)$ the corresponding coefficients of the degree one terms,
- φ_l the corresponding generalized plane wave as introduced in Definition 1.

The set of p shape functions adapted to the operator $-\Delta + \beta$, denoted $\mathcal{E}_{\vec{g}}(N, p, q)$, is defined by $\mathcal{E}_{\vec{g}}(N, p, q) = \{\varphi_l\}_{l \in \llbracket 1, p \rrbracket}$.

The parameter N is then the main degree of freedom to be fixed to compute explicitly the approximation space $\mathcal{E}_{\vec{g}}(N, p, q)$.

Finally, the explicit design of a set of GPWs at a point $\vec{g} = (x_0, y_0)$ can be summarized by the following steps:

1. fix $p \in \mathbb{N}$, $p \geq 3$, the number of functions, and compute for all $l \in \llbracket 1, p \rrbracket$: $\theta_l = 2\pi(l - 1)/p$ the direction of each function,
2. fix $N \in \mathbb{C}$, $N \neq 0$, and compute for all $l \in \llbracket 1, p \rrbracket$: $(\lambda_{1,0}^l, \lambda_{0,1}^l) = N(\cos \theta_l, \sin \theta_l)$,
3. fix $q \in \mathbb{N}$, $q \geq 1$, the order of approximation, and compute for all $l \in \llbracket 1, p \rrbracket$ the set of coefficients of each function $\{\lambda_{i,j}^l, 0 \leq i + j \leq q + 1\}$ according to Corollary 1,
4. for all $l \in \llbracket 1, p \rrbracket$ define $\varphi_l = \exp \sum_{0 \leq i+j \leq q+1} \lambda_{i,j}^l (x - x_0)^i (y - y_0)^j$;

the desired set of GPWs is $\{\varphi_l\}_{l \in \llbracket 1, p \rrbracket}$.

3 Interpolation

The interpolation properties of the set $\mathcal{E}_{\vec{g}}(N, p, q)$ are proved in Theorem 1. This section is devoted to the proof of this result, which states that, in order to approximate to a given order $n + 1$ the solution of the scalar wave Eq. (1) around a point \vec{g} , a number $p = 2n + 1$ of basis functions together with a approximation parameter $q = n + 1$ are sufficient. The gradient of the solution is then approximated to the order n .

The proof of this main theorem relies on a fundamental property of the GPWs, which is first proved and commented.

3.1 A fundamental property of a generalized pane wave

Since the design and the interpolation study are based on different Taylor expansions, the derivatives of the shape function φ are important quantities. Moreover, only two coefficients among the set $\{\lambda_{i,j}, i \in \{0, 1\}, j \in \llbracket 0, q + 1 - i \rrbracket\}$ are non-zero to define a GPW. So it is natural to express the other coefficients with respect to those two. Both

- the coefficients $\lambda_{i,j}$ s defining a shape function φ
- the derivatives of φ

can actually be expressed as polynomials in two variables with respect to $(\lambda_{1,0}, \lambda_{0,1})$. The coefficients of these polynomials depend on the indices i, j , the GPW parameter N and the derivatives of β evaluated at \vec{g} . They do not depend on θ , and this is crucial to prove Theorem 1.

Proposition 2 states this fundamental property of the derivatives of a generalized pane wave, and its proof relies on the intermediate Lemma 1. These result strongly use the fact that for a GPW: $\lambda_{1,0}^2 + \lambda_{0,1}^2 = N^2$.

Lemma 1 *Assume $\vec{g} = (x_0, y_0)$ is a point in \mathbb{R}^2 , the approximation parameter satisfies $q \in \mathbb{N}$ such that $q \geq 1$, $\theta \in \mathbb{R}$ and $N \in \mathbb{C}$ such that $N \neq 0$. Finally suppose $\beta \in C^{q-1}$ at \vec{g} . Consider a GPW adapted to the operator $-\Delta + \beta$: $\varphi = \exp \sum_{0 \leq i+j \leq q+1} \lambda_{i,j} (x - x_0)^i (y - y_0)^j$.*

Then each coefficient of the set $\{\lambda_{i,j}, 0 \leq i + j \leq q + 1\}$ can be described as polynomials in two variables in $(\lambda_{1,0}, \lambda_{0,1})$ as follows.

$$\begin{cases} \forall i \geq 2 \\ \lambda_{i,j} \text{ is a polynomial of total degree at most } i - 2. \end{cases} \tag{7}$$

The coefficients of these polynomials depend on the indices i, j , the GPW parameter N and the derivatives of β evaluated at \vec{g} . They do not depend on θ .

The following proof relies on a close examination of the induction formula (5), considered as polynomial in two variables, namely $(\lambda_{1,0}, \lambda_{0,1})$. The idea is to track the terms with higher degree.

Proof Because of the null coefficients, the induction formula (6) for $i = 0$ and $i = 1$ reads

$$\begin{cases} \lambda_{2,0} = \frac{1}{2} \left(\beta(\vec{g}) - (\lambda_{1,0})^2 - (\lambda_{0,1})^2 \right), \\ \lambda_{2,j} = \frac{1}{2} \frac{\partial_y^j \beta(\vec{g})}{j!} & \forall j > 0, \\ \lambda_{3,0} = \frac{1}{6} \left(\partial_x \beta(\vec{g}) - 4\lambda_{2,0}\lambda_{1,0} \right), \\ \lambda_{3,j} = \frac{1}{6} \left(\frac{\partial_x \partial_y^j \beta(\vec{g})}{j!} - 4\lambda_{2,j}\lambda_{1,0} \right) & \forall j > 0. \end{cases} \tag{8}$$

Then (7) for $i = 2$ stems from the choice of $(\lambda_{1,0}, \lambda_{0,1})$ in the GPW definition. Indeed for $j = 0$ the sum $(\lambda_{1,0})^2 + (\lambda_{0,1})^2$ does not depend on $(\lambda_{1,0}, \lambda_{0,1})$ themselves but only on N . Afterwards (7) for $i = 3$ is clear from (8).

Now set $i \geq 2$ and suppose that the statement (7) holds true for all $\tilde{i} \in \llbracket 3, i + 1 \rrbracket$. Then, isolating $\lambda_{i+2,j}$ in (5), the higher possible degree of each term is

- $i - 2$ for the term in $\lambda_{i,j+2}$,
- $(i - 1) + 1$ for the term in $\lambda_{i+1,j}\lambda_{1,0}$,
- $(i - k - 1) + (k - 1)$ for the terms in $\lambda_{i-k+1,j-l}\lambda_{k+1,l}$ with $k \neq 0$ and $k \neq i$,
- $(i - 2) + 1$ for the term in $\lambda_{i,j+1}\lambda_{0,1}$,
- $(i - l - 2) + (l - 2)$ for the term in $\lambda_{i-l,j-k+1}\lambda_{l,k+1}$ with $l \neq 0$ and $l \neq i$, note that $\lambda_{i-l,j-k+1}\lambda_{l,k+1} = 0$ with $l \neq 1$ and $l \neq i - 1$ because of null coefficients.

As a consequence the terms with higher degree appearing in the expression of $\lambda_{i+2,j}$ have degree at most equal to i . It completes the proof of (7) for $i > 2$ by induction.

The dependence of the polynomial expressions with respect to the different parameters is clear through the previous induction proof: the i, j and β terms appear explicitly in the induction formula (6), while the N appears in the expression of $\lambda_{2,0}$ of (8) since $N^2 = \lambda_{1,0}^2 + \lambda_{0,1}^2$. The parameter θ does not appear anywhere. \square

Proposition 2 Assume $\vec{g} = (x_0, y_0)$ is a point in \mathbb{R}^2 , the approximation parameter satisfies $q \in \mathbb{N}$ such that $q \geq 1$, $\theta \in \mathbb{R}$ and $N \in \mathbb{C}$ such that $N \neq 0$. Finally suppose $\beta \in C^{q-1}$ at \vec{g} . Consider a GPW adapted to the operator $-\Delta + \beta$: $\varphi = \exp \sum_{0 \leq i+j \leq q+1} \lambda_{i,j}(x-x_0)^i(y-y_0)^j$. Then for all $(i, j) \in \mathbb{N}^2$ such that $i + j \leq q + 1$ the difference

$$R_{i,j} := \partial_x^i \partial_y^j \varphi(\vec{g}) - (\lambda_{0,1})^j (\lambda_{1,0})^i \tag{9}$$

can be expressed as a complex polynomial with respect to the two variables $\lambda_{1,0}$ and $\lambda_{0,1}$. such that

- its total degree satisfies $dR_{i,j} \leq i - 2$,
- its coefficients only depend on i, j, N and on the derivatives of β but do not depend on θ .

Remark 2 Since $(\lambda_{1,0})^2 + (\lambda_{0,1})^2$ is fixed, none of the polynomial expressions that occur in Proposition 2 can be unique. For instance, any occurrence of $(\lambda_{1,0})^2$ could be replaced by $N^2 - (\lambda_{0,1})^2$ which would change the term of higher degree. This is the reason why $R_{i,j}$ is not unique: see Sect. 3.2 for a different point of view. However, formula (6) from Proposition 1 gives an explicit procedure for the computation of all $\lambda_{i,j}$: this is the crucial point that will be used for practical implementation.

One could have expected the degree of $R_{i,j}$ to be smaller than $i + j - 1$. The fact that it does actually not depend on j is due to the choice of the coefficients $\{\lambda_{i,j}, i \in \{0, 1\}, i + j > 1\}$ to be zero. The fact that it is smaller than $i - 2$ is due to the fact that the degree of $\lambda_{2,j}$ is 0, since $(\lambda_{1,0})^2 + (\lambda_{0,1})^2 = N^2$ is constant with respect to $\lambda_{0,1}$ and $\lambda_{1,0}$. See Definition 1.

Proof Applying the chain rule introduced in ‘Bivariate version’ in Appendix to the GPW φ one gets for all $(i, j) \in \mathbb{N}^2$,

$$\partial_x^i \partial_y^j \varphi(\vec{g}) = i!j! \sum_{\mu=1}^{i+j} \sum_{s=1}^{i+j} \sum_{p_s((i,j),\mu)} \prod_{l=1}^s \frac{(\lambda_{i_l,j_l})^{k_l}}{k_l!},$$

where $p_s((i, j), \mu)$ is the set of partitions of (i, j) with length μ :

$$\left\{ (k_l, (i_l, j_l))_{l \in \llbracket 1, s \rrbracket} : k_l \in \mathbb{N}^*, 0 < (i_1, j_1) < \dots < (i_l, j_l), \sum_{l=1}^s k_l = \mu, \sum_{l=1}^s k_l(i_l, j_l) = (i, j) \right\}.$$

See ‘Bivariate version’ in Appendix for a definition of this ordering relation. Now consider such a partition to be given and focus on the degree of the corresponding product term, namely $\prod_{l=1}^s (\lambda_{i_l,j_l})^{k_l}$. Thanks to Lemma 1 one can split this product into different terms regarding their degree as polynomials with respect to $(\lambda_{1,0}, \lambda_{0,1})$. As a result, since $Deg \prod_{l=1}^s (\lambda_{i_l,j_l})^{k_l} = \sum_{l=1}^s k_l Deg \lambda_{i_l,j_l}$, this quantity is also at most equal to

$$\sum_{i_l=0, j_l=1} k_l j_l + \sum_{i_l=1, j_l=0} k_l i_l + \sum_{i_l=2} k_l \cdot 0 + \sum_{i_l \geq 3} k_l(i_l - 2), \tag{10}$$

where the two first sums contain at most one term each.

Obviously the leading term in $\partial_x^i \partial_y^j \varphi(\vec{g})$ is $(\lambda_{0,1})^j (\lambda_{1,0})^i$, it corresponds to the partition $(i, j) = j(0, 1) + i(1, 0)$. Indeed, as long as a partition contains at least one term such that $i_l \geq 2$, the resulting degree computed from (10) will contain at least one term $k_l \cdot 0$ or $k_l(i_l - 2)$, and any of them is at most $k_l(i_l + j_l) - 2$; as a consequence the degree computed in (10) is then strictly lower than $\sum_{l=1}^s k_l(i_l + j_l) - 2 = i + j - 2$.

Since the product term corresponding to the partition $j(0, 1) + i(1, 0)$ is $(\lambda_{0,1})^j (\lambda_{1,0})^i / (j!i!)$ it proves both the polynomial nature of $R_{i,j}$ and its degree.

The claim concerning the coefficients of $R_{i,j}$ directly stems from the same property of the coefficients of $\lambda_{i,j}$ s from Lemma 1. □

3.2 A more algebraic viewpoint

This paragraph presents a more algebraic point of view on Remark 2.

Suppose $N \in \mathbb{C}$ is such that $N \neq 0$, and consider $\{\lambda_{i,j}\}$ the coefficient of any given GPW defined with the parameters N and a fixed θ . The value of $(\lambda_{1,0}, \lambda_{0,1})$ gives that the polynomial P_N defined as $P_N = (\lambda_{1,0})^2 + (\lambda_{0,1})^2 - N^2$ satisfies $P_N = 0$ independently of θ . From then on, considering other quantities as polynomials in two variables in $(\lambda_{0,1}, \lambda_{1,0})$ is in fact computing in the quotient ring $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]/(P_N)$

of $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ modulo the ideal generated by P_N . For instance, the system (8) reads

$$\left\{ \begin{array}{l} \lambda_{2,0} = \frac{\beta(\vec{g}) - N^2}{2} \quad (P_N), \\ \lambda_{2,j} = \frac{\partial_y^j \beta(\vec{g})}{2(j!)} \quad (P_N), \quad \forall j > 0, \\ \lambda_{3,0} = \frac{\partial_x \beta(\vec{g}) - 2\lambda_{1,0}(\beta(\vec{g}) - N^2)}{6} \quad (P_N), \\ \lambda_{3,j} = \frac{\partial_x \partial_y^j \beta(\vec{g})}{6(j!)} + 2 \frac{\partial_y^j \beta(\vec{g})}{j!} \lambda_{1,0} \quad (P_N), \quad \forall j > 0. \end{array} \right.$$

Of course in this quotient ring, each equivalence class has an infinite number of elements, and all the computations of the previous subsection are performed on elements of these classes. Thus any equality applies to all the elements of the same class. Note that since the ring considered here is the ring of polynomials in two variables, there is no such thing as the Euclidean division. As a result there is nothing like a canonical element of a class used for computations. One can easily see that for $q \geq 4$

$$\begin{aligned} \partial_x^4 \partial_y \varphi(\vec{g}) &= (\lambda_{1,0})^4 (\lambda_{1,0}) + 2 \partial_y \beta(\vec{g}) \left((\lambda_{1,0})^2 - (\lambda_{0,1})^2 \right) + 2 \partial_x \beta(\vec{g}) \lambda_{0,1} \lambda_{1,0} \\ &\quad + 2 \partial_x \partial_y \beta(\vec{g}) \lambda_{1,0} + (-3 \partial_y^2 \beta(\vec{g}) + \partial_x \beta(\vec{g})) \lambda_{0,1} - \partial_y^3 \beta(\vec{g}) \\ &\quad + \partial_x^2 \partial_y \beta(\vec{g}), = (\lambda_{1,0})^4 (\lambda_{1,0}) + 2 \partial_y \beta(\vec{g}) \left((\lambda_{1,0})^2 + (\lambda_{0,1})^2 \right) \\ &\quad + 2 \partial_x \beta(\vec{g}) \lambda_{0,1} \lambda_{1,0} + 2 \partial_x \partial_y \beta(\vec{g}) \lambda_{1,0} + (-3 \partial_y^2 \beta(\vec{g}) \\ &\quad + \partial_x \beta(\vec{g})) \lambda_{0,1} - \partial_y^3 \beta(\vec{g}) + \partial_x^2 \partial_y \beta(\vec{g}) - 2N^2 \partial_y \beta(\vec{g}), \end{aligned}$$

which gives two possible $R_{4,1} \in \mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ satisfying (9) in Proposition 2.

3.3 Theoretical result

This subsection focuses on the interpolation property of the set of basis functions $\mathcal{E}_{\vec{g}}(N, p, q)$. The sketch of the proof is inspired by the one developed by Cessenat and Després [5], but it is adapted to the generalized plane wave basis functions.

This proof requires the definition of two matrices, M_n and M_n^C , containing the derivatives of classical and generalized plane waves.

Definition 3 Assume $N \in \mathbb{C}$ is such that $N \neq 0$, $n \in \mathbb{N}$ is such that $n > 0$, $q \in \mathbb{N}$ such that $q \geq n + 1$ and $\vec{g} = (x_0, y_0) \in \mathbb{R}^2$. Finally suppose $\beta \in \mathcal{C}^{q-1}$ at \vec{g} . For all $l \in \mathbb{N}$ such that $1 \leq l \leq 2n + 1$ consider the direction $\theta_l = 2\pi(l - 1)/(2n + 1)$, the corresponding GPW $\varphi_l, \kappa = -lN \in \mathbb{C}^*$ and the function

$$e_l(x, y) = e^{i\kappa((x-x_0) \cos \theta_l + (y-y_0) \sin \theta_l)},$$

which is a classical plane wave if $N \in i\mathbb{R}$. The $(n + 1)(n + 2)/2 \times (2n + 1)$ matrices M_n^C and M_n are defined as follows: for all $(k_1, k_2) \in \mathbb{N}^2$, such that $k_1 + k_2 \leq n$

$$\begin{cases} (M_n^C)_{\frac{(k_1+k_2)(k_1+k_2+1)}{2}+k_2+1,l} = \frac{\partial_x^{k_1} \partial_y^{k_2} e_l(\vec{g})}{k_1!k_2!}, \\ (M_n)_{\frac{(k_1+k_2)(k_1+k_2+1)}{2}+k_2+1,l} = \frac{\partial_x^{k_1} \partial_y^{k_2} \varphi_l(\vec{g})}{k_1!k_2!}. \end{cases}$$

Their l th columns contain respectively the Taylor expansion coefficients of the functions e_l and φ_l .

For instance, one has $M_1 = \begin{pmatrix} \varphi_1(\vec{g}) & \varphi_2(\vec{g}) & \varphi_3(\vec{g}) \\ \partial_x \varphi_1(\vec{g}) & \partial_x \varphi_2(\vec{g}) & \partial_x \varphi_3(\vec{g}) \\ \partial_y \varphi_1(\vec{g}) & \partial_y \varphi_2(\vec{g}) & \partial_y \varphi_3(\vec{g}) \end{pmatrix}$,

$M_1^C = \begin{pmatrix} 1 & 1 & 1 \\ i\kappa \cos \theta_1 & i\kappa \cos \theta_2 & i\kappa \cos \theta_3 \\ i\kappa \sin \theta_1 & i\kappa \sin \theta_2 & i\kappa \sin \theta_3 \end{pmatrix}$ and

$$M_2 = \begin{pmatrix} \varphi_1(\vec{g}) & \varphi_2(\vec{g}) & \varphi_3(\vec{g}) & \varphi_4(\vec{g}) & \varphi_5(\vec{g}) \\ \partial_x \varphi_1(\vec{g}) & \partial_x \varphi_2(\vec{g}) & \partial_x \varphi_3(\vec{g}) & \partial_x \varphi_4(\vec{g}) & \partial_x \varphi_5(\vec{g}) \\ \partial_y \varphi_1(\vec{g}) & \partial_y \varphi_2(\vec{g}) & \partial_y \varphi_3(\vec{g}) & \partial_y \varphi_4(\vec{g}) & \partial_y \varphi_5(\vec{g}) \\ \partial_x^2 \varphi_1(\vec{g})/2 & \partial_x^2 \varphi_2(\vec{g})/2 & \partial_x^2 \varphi_3(\vec{g})/2 & \partial_x^2 \varphi_4(\vec{g})/2 & \partial_x^2 \varphi_5(\vec{g})/2 \\ \partial_x \partial_y \varphi_1(\vec{g}) & \partial_x \partial_y \varphi_2(\vec{g}) & \partial_x \partial_y \varphi_3(\vec{g}) & \partial_x \partial_y \varphi_4(\vec{g}) & \partial_x \partial_y \varphi_5(\vec{g}) \\ \partial_y^2 \varphi_1(\vec{g})/2 & \partial_y^2 \varphi_2(\vec{g})/2 & \partial_y^2 \varphi_3(\vec{g})/2 & \partial_y^2 \varphi_4(\vec{g})/2 & \partial_y^2 \varphi_5(\vec{g})/2 \end{pmatrix}.$$

The rank of the matrix M_n^C is computed in Lemma 2, which profits from the fact that the result proved by Cessenat and Després [5] for $\kappa > 0$ is actually still valid for $\kappa \in \mathbb{C}^*$. The proof of Theorem 1 relies on Lemma 3 that explicits the link between the matrix M_n^C and the corresponding matrix M_n built with the generalized plane waves.

Lemma 2 Assume $N \in \mathbb{C}$ is such that $N \neq 0$, $n \in \mathbb{N}$ is such that $n > 0$ and $\vec{g} = (x_0, y_0) \in \mathbb{R}^2$. Finally suppose $\beta \in C^0$ at \vec{g} . There are two matrices: a rectangle matrix $P_n \in \mathbb{C}^{(2n+1) \times (n+1)(n+2)/2}$ only depending on $\beta(\vec{g})$ and a square invertible matrix $S_n \in \mathbb{C}^{(2n+1) \times (2n+1)}$ only depending on the directions θ_l such that $S_n = P_n \cdot M_n^C$ and $rk(S_n) = 2n + 1$. Moreover $rk(M_n^C) = 2n + 1$.

Proof Consider M_n^C to be the matrix introduced in Definition 3 so that for all $(k_1, k_2) \in \mathbb{N}^2$, such that $k_1 + k_2 \leq n$ since $e_l(\vec{g}) = 1$

$$(M_n^C)_{\frac{(k_1+k_2)(k_1+k_2+1)}{2}+k_2+1,l} = \frac{\partial_x^{k_1} \partial_y^{k_2} e_l(\vec{g})}{k_1!k_2!} = \frac{(i\kappa)^{k_1+k_2}}{k_1!k_2!} \cos^{k_1} \theta_l \sin^{k_2} \theta_l.$$

Define for all $k \in \llbracket 0, n \rrbracket$

$$(S_n)_{n\pm k+1,l} = \frac{1}{(i\kappa)^k} (\partial_x \pm i\partial_y)^k e_l(\vec{g}) = \frac{k!}{(i\kappa)^k} \sum_{s=0}^k \frac{(\pm i)^s \partial_x^{(k-s)} \partial_y^s e_l(\vec{g})}{(k-s)!s!}.$$

Thanks to the definition of M_n^C one can check that

$$(S_n)_{n\pm k+1,l} = \frac{k!}{(\iota\kappa)^k} \sum_{s=0}^k (\pm\iota)^s (M_n^C)_{\frac{((k-s)+s)((k-s)+s+1)}{2}+s+1,l},$$

so that S_n is a $(2n + 1) \times (2n + 1)$ matrix that is a linear transform of M_n^C . More precisely, define P_n as a $(2n + 1) \times \frac{(n+1)(n+2)}{2}$ matrix such that

$$(P_n)_{n\pm k+1, \frac{k(k+1)}{2}+s+1} = k!(\pm\iota)^s/(\iota\kappa)^k.$$

Then $S_n = P_n \cdot M_n^C$. As a consequence, $rk(M_n^C) \geq rk(S_n)$.

The rank of S_n is now to be evaluated thanks to the definition of the plane waves e_l . Since $e_l(x, y) = e^{(\iota\kappa)((x-x_0) \cos \theta_l + (y-y_0) \sin \theta_l)}$ then

$$(\partial_x \pm \iota \partial_y)^k e_l = (\iota\kappa)^k (\cos \theta_l \pm \iota \sin \theta_l)^k e_l.$$

Consider that $z_l = \cos \theta_l + \iota \sin \theta_l = (\cos \theta_l - \iota \sin \theta_l)^{-1}$ because $|z_l| = 1$, and since $e_l(\vec{g}) = 1$ it yields

$$(\partial_x \pm \iota \partial_y)^k e_l(\vec{g}) = (\iota\kappa)^k (z_l)^{\pm k} e_l(\vec{g}) = (\iota\kappa)^k (z_l)^{\pm k} \Rightarrow (S_n)_{n\pm k+1,l} = (z_l)^{\pm k}.$$

Thus S_n 's columns are proportional to the one of a Vandermonde matrix and

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

From the choice of θ_l s, for all $i \neq j$: $z_i \neq z_j$ so that S_n is invertible and $rk(M_n^C) \geq rk(S_n) = 2n + 1$. Since

$$rk(M_n^C) \leq \min\left(2n + 1, \frac{(n + 1)(n + 2)}{2}\right) = 2n + 1$$

the proof is then completed. □

Lemma 3 Assume $N \in \mathbb{C}$ is such that $N \neq 0$, $n \in \mathbb{N}$ is such that $n > 0$, $q \in \mathbb{N}$ is such that $q \geq n + 1$ and $\vec{g} = (x_0, y_0) \in \mathbb{R}^2$. Finally suppose $\beta \in C^{q-1}$ at \vec{g} . Consider $\mathcal{E}_{\vec{g}}(N, p, q)$ introduced in Definition 2, together with M_n and M_n^C introduced in Definition 3. Then there is a lower triangular matrix L_n , whose diagonal coefficients are all equal to 1 and whose other coefficients are linear combinations of the derivatives of β evaluated at \vec{g} , such that

$$M_n = L_n \cdot M_n^C. \tag{11}$$

As a consequence $rk(M_n) = rk(M_n^C)$ and both $\|L_n\|$ and $\|(L_n)^{-1}\|$ are bounded by a constant only depending on N and β .

The following proof is straightforward considering the feature of the derivatives of φ_l described in Proposition 2.

Proof From (9) there exists a polynomial $R_{i,j} \in \mathbb{C}[X, Y]$ with $Deg R_{i,j} \leq i - 2$ such that

$$\forall (i, j) \in \mathbb{N}^2, \partial_x^i \partial_y^j \varphi_l(\vec{g}) = \partial_x^i \partial_y^j e_l(\vec{g}) + R_{i,j}(\partial_x e_l(\vec{g}), \partial_y e_l(\vec{g})). \tag{12}$$

The coefficients of $R_{i,j}$ do not depend on the basis function considered, but only depends on β and its derivatives evaluated at \vec{g} . By construction of the classical plane wave e_l , one has

$$\begin{cases} \partial_x^k \partial_y^m e_l(\vec{g}) = (\partial_x e_l(\vec{g}))^k (\partial_y e_l(\vec{g}))^m, \\ = (i\kappa)^{k+m} \cos(\theta)^k (i \sin(\theta))^m. \end{cases}$$

The numbering of the rows in matrices M_n^C and M_n is set up such that the derivatives of smaller order appear higher in the matrix, which proves (11). Indeed (12) shows that any coefficient of M_n is the sum of the corresponding coefficient in M_n^C plus a linear combination—whose coefficients do not depend on the column that is considered but only on β and its derivatives evaluated at \vec{g} —of terms that appear higher in the corresponding column of M_n .

The rank of M_n is then equal to the rank of M_n^C , and $\|L_n\|$ and $\|(L_n)^{-1}\|$ do only depend on the coefficients of $R_{i,j}$. As a result they do not depend on the basis functions but only on the parameter N , the coefficient β and its derivatives at \vec{g} . \square

The hypothesis of the following theorem give a sufficient condition on the relation between the parameters p and q to achieve a high order interpolation.

Theorem 1 *Assume $\Omega \subset \mathbb{R}^2$ is a bounded domain, $N \in \mathbb{C}$ is such that $N \neq 0$, $n \in \mathbb{N}$ is such that $n > 0$, $q \geq n + 1$, $p = 2n + 1$ and $\vec{g} = (x_0, y_0) \in \Omega$. Finally suppose $\beta \in C^{q-1}$ at \vec{g} . Consider that u is a solution of scalar wave Eq. (1) which belongs to C^{n+1} . Consider then $\mathcal{E}_{\vec{g}}(N, p, q)$ introduced in Definition 2. Then there are a function $u_a \in Span \mathcal{E}_{\vec{g}}(N, p, q)$ depending on β and n , and a constant $C(N, \Omega, n)$ depending on N , β and n such that for all $\vec{m} \in \mathbb{R}^2$*

$$\begin{cases} |u(\vec{m}) - u_a(\vec{m})| \leq C(N, \Omega, n) |\vec{m} - \vec{g}|^{n+1} \|u\|_{C^{n+1}(\Omega)}, \\ \|\nabla u(\vec{m}) - \nabla u_a(\vec{m})\| \leq C(N, \Omega, n) |\vec{m} - \vec{g}|^n \|u\|_{C^{n+1}(\Omega)}. \end{cases} \tag{13}$$

Note that it provides an upper bound for the best approximation property of the function space $\mathcal{E}_{\vec{g}}(N, p, q)$:

$$\begin{aligned} & \max \left\{ \|u - \varphi\|_{L^\infty(|\vec{m} - \vec{g}| \leq h)}, \varphi \in Span \mathcal{E}_{\vec{g}}(N, p, q) \right\} \\ & \leq C(N, \Omega, n) \|u\|_{C^{n+1}(\Omega)} h^{n+1}. \end{aligned}$$

The behavior of the constant $C(N, \Omega, n)$ as N goes to zero is commented in Sect. 3.4. It suggests the need for a parameter N that is bounded away from zero, see Sect. 4.1.

Proof The idea of the proof is to look for $u_a = \sum_{l=1}^{2n+1} X_l \varphi_l$ by fitting its Taylor expansion to the one of u . This will be done by solving a linear system concerning the unknowns $(X_l)_{l \in \llbracket 1, 2n+1 \rrbracket}$.

Since u belongs to C^{n+1} and for all $l \in \llbracket 1, 2n + 1 \rrbracket$ the basis function φ_l belongs to C^∞ , their Taylor expansions read: there is a constant C such that for all $\vec{m} = (x, y) \in \Omega$

$$\left| u(x, y) - \sum_{m=0}^n \sum_{k_1+k_2=m} B_{k_1 k_2} x^{k_1} y^{k_2} \right| \leq C |\vec{m} - \vec{g}|^{n+1} \|u\|_{C^{n+1}},$$

$$\left| \varphi_l(x, y) - \sum_{m=0}^n \sum_{k_1+k_2=m} M_{k_1 k_2}^l x^{k_1} y^{k_2} \right| \leq C |\vec{m} - \vec{g}|^{n+1} \|\varphi_l\|_{C^{n+1}},$$

where for the sake of simplicity $M_{k_1 k_2}^l$ stands for the term $\partial_x^{k_1} \partial_y^{k_2} \varphi_l / (k_1! k_2!)$, which is the definition of $(M_n)_{\binom{(k_2+k_1)(k_2+k_1+1)}{2} + k_2 + 1, l}$. In the same way, define a vector B_n : B_{k_1, k_2} stands for the Taylor coefficient of u , which is the definition of the coefficient $(B_n)_{\binom{(k_2+k_1)(k_2+k_1+1)}{2} + k_2 + 1}$. Note that at every point \vec{g} in Ω , there is a constant $C = C_{\vec{g}}$ such that the two previous estimates hold. This constant $C_{\vec{g}}$ depends continuously on \vec{g} , through the derivatives of β evaluated at \vec{g} . Since Ω is compact, the constant $C = \max_{\vec{g} \in \Omega} C_{\vec{g}}$ is finite. The system to be solved is then

$$\begin{cases} \text{Find } (X_l)_{l \in \llbracket 1, 2n+1 \rrbracket} \in \mathbb{C}^{2n+1} \text{ s.t.} \\ \sum_{l=1}^{2n+1} M_{k_1, k_2}^l X_l = B_{k_1, k_2}, \quad \forall m \in \llbracket 0, n \rrbracket, \quad \forall (k_1, k_2) \in \llbracket 0, n \rrbracket^2 \text{ s.t. } k_1 + k_2 = m. \end{cases}$$

In order to study the system’s matrix, the equations depending on (k_1, k_2) have to be numbered: they will be considered with increasing $m = k_1 + k_2$, and with decreasing k_1 for a fixed value of m . Defining the corresponding vector $B_n \in \mathbb{C}^{\frac{(n+1)(n+2)}{2}}$, together with the unknown $X^n = (X_1, X_2, \dots, X_{2n+1}) \in \mathbb{C}^{2n+1}$, the system now reads

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

where $M_n \in \mathbb{C}^{\frac{(n+1)(n+2)}{2} \times (2n+1)}$ is the matrix from Definition 3.

Since the system is not square, there is a solution if and only if $B_n \in \text{Im}(M_n)$. The following points are steps toward the proof of existence of such a solution.

(i) The technical point is to prove that $\text{rk}(M_n) = 2n + 1$. It is straightforward from Lemmas 3 and 2.

(ii) Build subset $\mathfrak{K} \subset \mathbb{C}^{\frac{(n+1)(n+2)}{2}}$ such that $\text{Im}(M_n) \subset \mathfrak{K}$ and $B_n \in \mathfrak{K}$. Such a subspace \mathfrak{K} can be built from the fact that the basis functions are designed to fit the Taylor expansion of the scalar wave equation: for all $l \in \llbracket 1, 2n + 1 \rrbracket$, for all $(k_1, k_2) \in$

$[[0, n - 2]]^2$ such that $k_1 + k_2 \leq n - 2$, the element of the l th column of M_n satisfy

$$\begin{aligned} & (k_1 + 1)(k_1 + 2)(M_n)_{k_1+2,k_2} + (k_2 + 1)(k_2 + 2)(M_n)_{k_1,k_2+2} \\ &= \sum_{i=0}^{k_1} \sum_{j=0}^{k_2} \frac{\partial_x^i \partial_y^j \beta(\vec{g})}{i!j!} (M_n)_{k_1-i,k_2-j}. \end{aligned}$$

This is clear from the design of the GPWs, since for all the basis functions φ_l , all the terms of order smaller than q in the Taylor expansion of $(-\Delta + \beta)\varphi_l = (-P_{\Delta,l} + \beta)\varphi_l$ vanish. So define

$$\begin{aligned} \mathfrak{K} := & \left\{ (C_{k_1,k_2}) \in \mathbb{C}^{\frac{(n+1)(n+2)}{2}}, \forall (k_1, k_2) \in \mathbb{N}^2, k_1 + k_2 \leq n - 2, \right. \\ & (k_1 + 1)(k_1 + 2)C_{k_1+2,k_2} + (k_2 + 1)(k_2 + 2)C_{k_1,k_2+2} \\ & \left. = \sum_{i=0}^{k_1} \sum_{j=0}^{k_2} \frac{\partial_x^i \partial_y^j \beta(\vec{g})}{i!j!} C_{k_1-i,k_2-j} \right\} \tag{14} \end{aligned}$$

With $q \geq n + 1$, it is then straightforward to see that $Im(M_n) \subset \mathfrak{K}$. The fact that $B_n \in \mathfrak{K}$ simply stems from plugging the Taylor expansions of u and β into scalar wave equation.

(iii) The dimension of \mathfrak{K} defined by (14) is $\dim \mathfrak{K} = 2n + 1$. Indeed, one can check—using the same numbering as previously for the equations—that \mathfrak{K} is defined by $n(n - 1)/2$ linearly independent relations on $\mathbb{C}^{\frac{(n+1)(n+2)}{2}}$, so that its dimension is $(n + 1)(n + 2)/2 - n(n - 1)/2$.

As a result of points (i), (ii), and (iii), $Im(M_n)$ there is a solution to the system $M_n \cdot X^n = B_n$.

As a consequence, consider a solution X^n to the system $M_n \cdot X^n = B_n$, and define $u_a = \sum_{l=1}^{2n+1} X_l \varphi_l$. Thanks to that definition and to the Taylor expansions of u and the φ_l s it yields

$$|u(\vec{m}) - u_a(\vec{m})| \leq C |\vec{m} - \vec{g}|^{n+1} (\|u\|_{\mathcal{C}^{n+1}} + \|u_a\|_{\mathcal{C}^{n+1}}).$$

Moreover one has the identity $X^n = (S_n)^{-1} P_n (L_n)^{-1} B_n$, where $(S_n)^{-1} P_n$ is bounded from above by $\sup_{l \in [[1, 2n+1]]} \|e_l\|_{\mathcal{C}^{n+1}}$, see Lemma 2, $(L_n)^{-1}$ is bounded from above by a constant depending only on β and its derivatives from Lemma 3, and B_n is bounded by $\|u\|_{\mathcal{C}^{n+1}}$. Since for all $l \in [[1, 2n + 1]]$ it yields $|X_l| \leq C(N, \Omega, n)\|u\|_{\mathcal{C}^{n+1}}$, it turns out to be the first part of (13):

$$|u(\vec{m}) - u_a(\vec{m})| \leq C(N, \Omega, n)(2n + 2) |\vec{m} - \vec{g}|^{n+1} \|u\|_{\mathcal{C}^{n+1}}.$$

At last, the second part of (13) stems from taking the Taylor Lagrange formula of the gradient of $u - u_a$, up to the order n , since

$$\sum_{m=0}^n \sum_{k_1+k_2=m} \left(\mathbf{B}_{k_1 k_2} (x-x_0)^{k_1} (y-y_0)^{k_2} - \sum_{l=1}^{2n+1} \left(\mathbf{X}_l \mathbf{M}_{k_1 k_2}^l (x-x_0)^{k_1} (y-y_0)^{k_2} \right) \right) = 0.$$

That is: for all $\vec{m} = (x, y) \in \Omega$ there are ζ_1, ζ_2 in \mathbb{R}^2 on the segment line between \vec{m} and \vec{g} such that

$$\begin{cases} \partial_x(u - u_a)(x, y) = \sum_{l=0}^n \frac{\partial_x^{l+1} \partial_y^{n-l}(u - u_a)(\zeta_1)}{l!(n-l)!} (x - x_0)^l (y - y_0)^{n-l}, \\ \partial_y(u - u_a)(x, y) = \sum_{l=0}^n \frac{\partial_x^l \partial_y^{n-l+1}(u - u_a)(\zeta_2)}{l!(n-l)!} (x - x_0)^l (y - y_0)^{n-l} \end{cases}$$

which indeed leads to the desired inequality. □

3.4 Comments on the asymptotic behavior of $C(N, \Omega, n)$

Consider first the behavior of this constant with respect to N . Because the Taylor expansion actually reads

$$(u - u_a)(x, y) = \sum_{j=0}^q \left(\frac{\partial_x^j \partial_y^{q-j} u(\vec{g})}{j!(q-j)!} - \sum_{l=1}^p \mathbf{X}_l \frac{\partial_x^j \partial_y^{q-j} \varphi_l(\vec{g})}{j!(q-j)!} \right) (x - x_0)^j (y - y_0)^{q-j} + O(|\vec{m} - \vec{g}|^{q+1}),$$

one can see that if $c_{N, \vec{g}} = \sum_{l=1}^p \mathbf{X}_l \sum_{j=0}^q \frac{\partial_x^j \partial_y^{q-j} \varphi_l(\vec{g})}{j!(q-j)!}$ blows up when N goes to zero, then so does $C(N, \Omega, n)$.

Since \mathbf{X}^n satisfies $\mathbf{S}_n \cdot \mathbf{X}^n = \mathbf{P}_n (\mathbf{L}_n)^{-1} \mathbf{B}_n$, one can describe the asymptotic behavior of the \mathbf{X}_l s with respect to N as follows. The matrix of this system doesn't depend on N , but only on p . Now consider the right hand side of this system:

- the vector \mathbf{B}_n only depends on u and n ,
- thanks to Proposition 2, the matrix \mathbf{L}_n has a lower triangular block structure: particularly the n diagonal square blocs of increasing size $k \times k$ for k from 1 to $n + 1$ are identity blocs, while the sub-diagonal rectangle blocs of size $(k + 1) \times k$ for k from 1 to n are zero blocs, and the rest of the sub-diagonal coefficients are polynomials with respect to N^2 ,
- the inverse $(\mathbf{L}_n)^{-1}$ shares this sparse structure with \mathbf{L}_n and the same dependence with respect to N ,
- by construction the only non zero blocs of the matrix \mathbf{P}_n are row blocs, each element of such a bloc having the same modulus $1/N^k$, particularly for the first

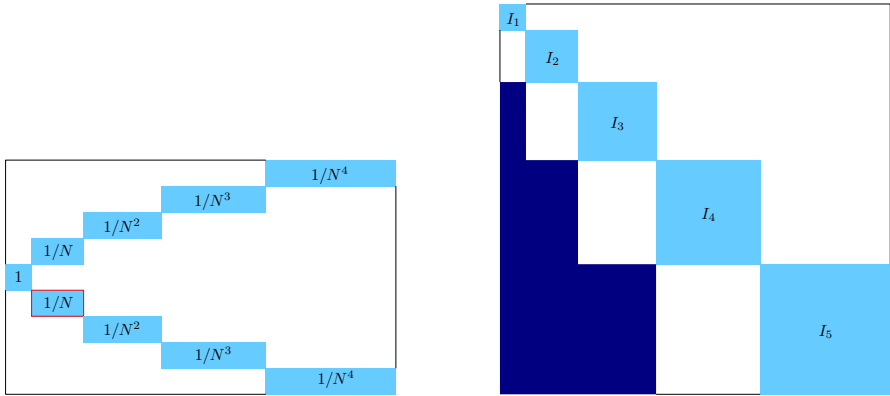


Fig. 2 Profile of matrices P_n and L_n for $n = 4$. All white blocs are zeros. Left row bloc structure of the P_n matrix. For k from 0 to n , the rows $n \pm k + 1$ contain a $1 \times (k + 1)$ non zero bloc. These non zero blocs are represented in sky blue. As stated in the definition of P_n , every non zero element on the row $n \pm k + 1$ behaves asymptotically as $1/N^k$ as N goes to zero. Right bloc structure of the L_n matrix. The diagonal blocs are identity blocs of increasing size k , they are represented in sky blue. The sub-diagonal blocs of size $k \times (k + 1)$ are zeros, so they are represented in white. The rest of the sub-diagonal coefficients, represented on midnight blue, are polynomials with respect to N^2 (color figure online)

and last rows of P_n which first $\frac{n(n+1)}{2}$ terms are zero while the last $n + 1$ terms are respectively $n!(\pm i)^s/N^n$ for s from 0 to n .

It yields that the right hand side behaves as $1/N^n$ as N goes to zero, see Fig. 2 for an illustration of these different structures. As a result the coefficients $X_{i,s}$ behave as $1/N^n$ as N goes to zero as well. As opposed to Proposition 2, a result concerning the lowest degree term of $\partial_x^j \partial_y^{q-j} \varphi$ as polynomials with respect to $(\lambda_{1,0}, \lambda_{0,1})$ would provide an explicit estimate for $C(N, \Omega, n)$ as N goes to zero.

Consider now the behavior of the constant $C(N, \Omega, n)$ with respect to the local number of basis functions, $p = 2n + 1$. The only direct dependence with respect to p stems from the matrix S_n , more precisely from its determinant. Indeed, it is clear to see that

$$|\det S_n| = \prod_{j=2}^p \prod_{i=1}^{j-1} \left(\exp \frac{2i\pi(i-1)}{p} - \exp \frac{2i\pi(j-1)}{p} \right) = \prod_{l=1}^{\frac{p-1}{2}} \left(\sin \frac{2\pi l}{p} \right)^p.$$

In order to get a first estimate, one can consider that the convergence will be driven by the terms such that $l/p \leq 100$ and simply bound the other terms by 1. Then from Stirling’s approximation

$$\prod_{l=1}^{p/100} \sin \frac{2\pi l}{p} \approx 2\pi \prod_{l=1}^{p/100} \frac{l}{p} \approx 2\pi \sqrt{2\pi \frac{p}{100}} \frac{1}{(100e)^{\frac{p}{100}}},$$

so that $|\det \mathbf{S}_n| \leq C \left(\frac{p}{100}\right)^{p/2} \frac{1}{(100e)^{\frac{p^2}{100}}}$. This first estimate is encouraging since in the plane wave literature, and especially for applications linked to the UWVF, the hp -convergence seems more attractive than the h -convergence, see for instance [18, 20].

4 Numerical validation

In order to validate the Theorem 1, each of the numerical validation case is computed, for a given value of n , setting $q = n + 1$ and $p = 2n + 1$. The test case considered is $\beta(x, y) = x - 1$, to approximate the exact solution $u_e(x, y) = \text{Airy}(x)e^{ly}$. See [22] for the physical motivation of this test case: its main interest is that the coefficient vanishes along the line $x = 1$, which represents a plasma cut-off that reflects incoming waves.

First a discussion addresses the choice of the normalization parameter N to design the basis functions, introduced in Definition 1.

Of course since the theoretical results give local approximation properties, the validation procedure itself will be local as well. As stated in the theorem u_e can be approximated by a function u_a that belongs to the approximation space $\mathcal{E}_{\vec{g}}(N, p, q)$, space that is built with the previously discussed normalizations.

The idea is to follow the error $\max |u_e - u_a|$ on disks with decreasing radius h in order to observe the order of convergence with respect to h . Several different cases are proposed to validate the theoretical order of convergence, and additional cases concern the behavior of the basis functions with respect to the cut-off.

4.1 Two families of generalized plane waves

Two different choices of the normalization parameter N will be considered in this paper. A first choice is comprised of setting $N = \sqrt{\text{sgn}\beta(\vec{g})} \sqrt{|\beta(\vec{g})|}$, see Definition 4. It gives a direct generalization of a classical plane waves, since in this case $\beta(\vec{g}) < 0$, so that $\sqrt{-\beta(\vec{g})}$ is the local wave number. However, this choice is local since N does depend on $\vec{g} \in \Omega$, and cannot be used if $\beta(\vec{g}) = 0$: it is a classical problem in low frequency regime, see [15, 21]. To overcome this limitation and consider the stationary limit case, a second possibility is to choose one constant and non zero value for N : it will not depend on \vec{g} anymore.

The first type of shape functions corresponds to a *local* choice since it does depend on $\vec{g} \in \mathbb{R}^2$.

Definition 4 The β -normalization for GPWs is defined by choosing $N = \sqrt{\text{sgn}\beta(\vec{g})} \sqrt{|\beta(\vec{g})|}$ in Definition 1, which means setting

1. $(\lambda_{1,0}, \lambda_{0,1}) = \sqrt{\text{sgn}\beta(\vec{g})} \sqrt{|\beta(\vec{g})|} (\cos \theta, \sin \theta)$.
2. $\{\lambda_{i,j}, 0 \leq i + j \leq q + 1, i + j \neq 1\}$ are set to zero.

Remark 3 (Back to classical Plane Waves from the β -normalization) The fact that the quantity $(\lambda_{1,0})^2 + (\lambda_{0,1})^2$ is equal to $\beta(\vec{g})$ however gives that the value of $\beta(\vec{g})$ does actually never appear in the expression of the other coefficients explicitly, but only in product terms involving $\lambda_{1,0}$ or $\lambda_{0,1}$. One can easily check by induction that each one of the terms that are summed in formula (6) contains at least one derivative of β . Indeed (8) proves it for $\lambda_{i,j}$, $i = 2$ and 3 , and the induction is then clear for $i \geq 3$ thanks to (6).

As a consequence, for $\beta = -\omega^2 < 0$ and for any $q \geq 1$, all the coefficients $\lambda_{i,j}$ such that $i > 1$ are actually zero, which means that the corresponding function $\varphi = \exp \sum_{0 \leq i+j \leq q+1} \lambda_{i,j}(x-x_0)^i(y-y_0)^j$ is nothing more than a classical plane wave.

As already remarked, it is also obvious that for $q = 1$ this new shape function is again nothing more than a classical plane wave as long as $\beta < 0$. This case $q = 1$ corresponds to the classical fact of approximating a smooth coefficient by its piecewise constant value at the center of the cells.

The fact is that since the terms $(\lambda_{1,0}, \lambda_{0,1})$ of the β -normalization are proportional to the square root of β , they will vanish when $\beta = 0$. Note that the approximation properties of classical plane waves do not deteriorate as their wave number goes to zero, see [26, Section 3.4]. However, numerical results show that it causes severe damaging to the conditioning of the discrete UWVF problem if β tends to zero. Moreover the theoretical estimate displayed in Sect. 3.4 justifies the need for a second normalization to ensure the desired interpolation property. As a consequence, a second normalization is considered, with a *global* choice of N independent from $\vec{g} \in \mathbb{R}^2$.

Definition 5 The constant-normalization for GPW is defined by choosing $N = \iota$ in Definition 1, which means setting

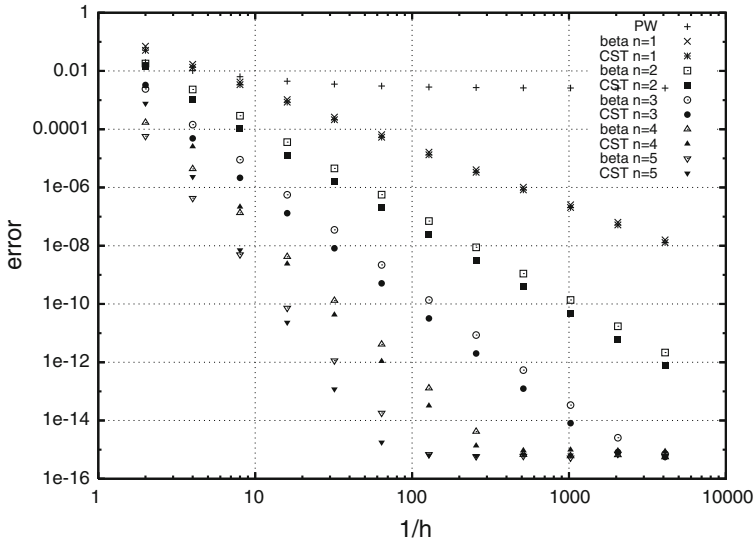
1. $(\lambda_{1,0}, \lambda_{0,1}) = \iota(\cos \theta, \sin \theta)$.
2. $\{\lambda_{i,j}, 0 \leq i+j \leq q+1, i+j \neq 1\}$ are set to zero.

Remark 4 (Classical Plane Waves and the constant-normalization) In order to balance Remark 3, note that since the constant-normalization does not depend on β it arises that for $\beta = -\omega^2 \neq -1$ the term $\beta(\vec{g})$ appears in higher order terms. For instance it is clear that $\lambda_{2,0} = (1 + \beta(\vec{g}))/2$. As a consequence, neither when $\beta(\neq -1) < 0$ is constant nor when $q = 1$ the shape function designed from the constant-normalization can be a classical plane wave.

4.2 In the propagative zone

The point $\vec{g} = (-3, 1)$ is in the propagative zone. Then concentric disks are centered on \vec{g} with radius $h = 1/2^k$, increasing the value of k . Following the theorem, the expected order of convergence is $n + 1$.

Figure 3 displays computed convergence results that fit perfectly the theoretical result. A set of $p = 11$ classical Plane Waves is used as a control case, since $p = 11$ is the higher number of basis functions used with both normalizations of the GPWs. As expected, since both normalizations satisfy $N \neq 0$ away from the cut-off, the orders of



	$p = 11$		$n = 1$		$n = 2$		$n = 3$		$n = 4$		$n = 5$	
h	PW	β	CST	β	CST	β	CST	β	CST	β	CST	
$1/2^2$	0.92	2.07	1.94	2.96	3.81	4.07	6.08	5.29	6.88	7.06	8.37	
$1/2^3$	0.69	2.02	1.98	3.00	3.27	3.99	4.50	5.02	6.88	6.44	8.35	
$1/2^4$	0.51	2.00	2.00	3.00	3.06	4.00	4.04	5.00	6.49	6.09	8.26	
$1/2^5$	0.34	2.00	2.00	3.00	3.01	4.00	4.00	5.00	5.82	6.00	7.61	
$1/2^6$	0.21	2.00	2.00	3.00	3.00	4.00	4.00	5.00	5.30	5.97	6.07	

Fig. 3 Convergence results in the propagative zone, computed at $(-3, 1) \in \mathbb{R}^2$ with different basis functions. Comparison between classical Plane Waves and Generalized Plane Waves for both β and constant normalizations. Some of the associated orders of convergence are also provided. In the legend, *beta* and *CST* respectively refer to the β and constant normalizations, while *n* is the parameter introduced in Theorem 1: in each case the expected order of convergence is $n + 1$. *PW* refers to the classical Plane Waves

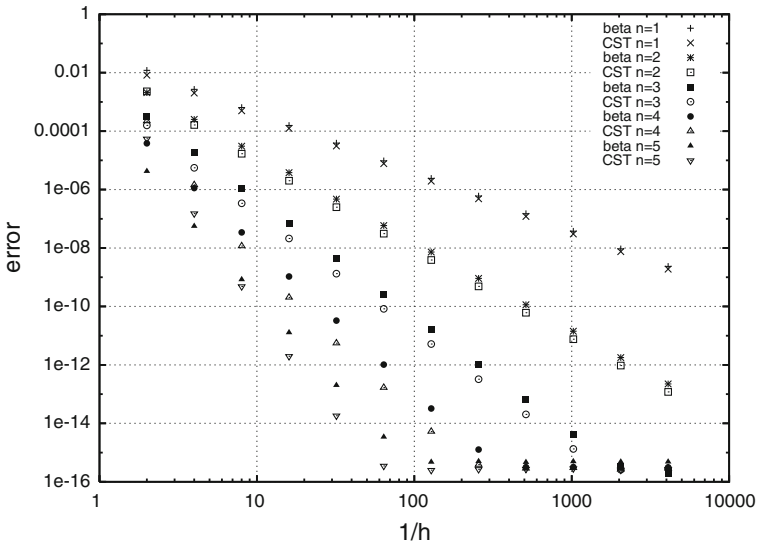
convergence observed are close to $n + 1$. No significant difference is observed between the two types of normalization.

Note that machine precision is reached in the higher order cases.

4.3 In the non propagative zone

The point $\vec{g} = (2, 1)$ is in the non propagative zone. Again concentric disks are centered on \vec{g} with radius $h = 1/2^k$, increasing the value of k , and the expected order of convergence is $n + 1$. There is no classical Plane Wave that can be computed here since $\beta(\vec{g}) > 0$.

Figure 4 displays computed convergence results that fit perfectly the theoretical result as well. As was observed in the propagative case, since both normalizations satisfy $N \neq 0$, the orders of convergence observed are equal or slightly higher than



	n = 1		n = 2		n = 3		n = 4		n = 5	
h	β	CST	β	CST	β	CST	β	CST	β	CST
1/2 ²	2.16	2.03	3.05	3.82	4.14	4.82	5.09	7.26	6.24	8.50
1/2 ³	2.07	2.01	3.03	3.27	4.05	4.03	5.04	6.96	6.07	8.28
1/2 ⁴	2.03	2.00	3.02	3.07	4.02	4.00	5.02	5.83	6.02	7.93
1/2 ⁵	2.02	2.00	3.01	3.01	4.01	4.00	5.01	5.21	6.00	6.76
1/2 ⁶	2.01	2.00	3.00	3.00	4.00	4.00	5.00	5.05	5.87	5.70

Fig. 4 Convergence results in the non-propagative zone, computed at $(2, 1) \in \mathbb{R}^2$ with different basis functions. Comparison between Generalized Plane Waves for β and constant normalizations. Some of the associated orders of convergence are also provided. In the legend, *beta* and *CST* respectively refer to the β and constant normalizations, while *n* is the parameter introduced in Theorem 1: in each case the expected order of convergence is $n + 1$

the theoretical order $n + 1$. Still no significant difference appears between the two types of normalization.

Again machine precision is reached in the higher order cases.

4.4 Along the cut-off: $\beta = 0$

The point $\vec{g} = (1, 1)$ lies exactly on the vanishing line of β . Then again concentric disks are centered on \vec{g} with radius $h = 1/2^k$, increasing the value of k . Both classical Plane Waves and GPWs with β -normalization would provide only one function since they correspond to $N = \beta(\vec{g}) = 0$. As to the GPWs with constant-normalization, the theoretical results show that their interpolation property holds along the cut-off as well as anywhere else in the domain.

As Figs. 3, 4 and 5 displays results that fit perfectly the theoretical result. The orders of convergence observed are again slightly better than $n + 1$. It is an example of efficient approximation of the exact solution u_e along the cut-off.

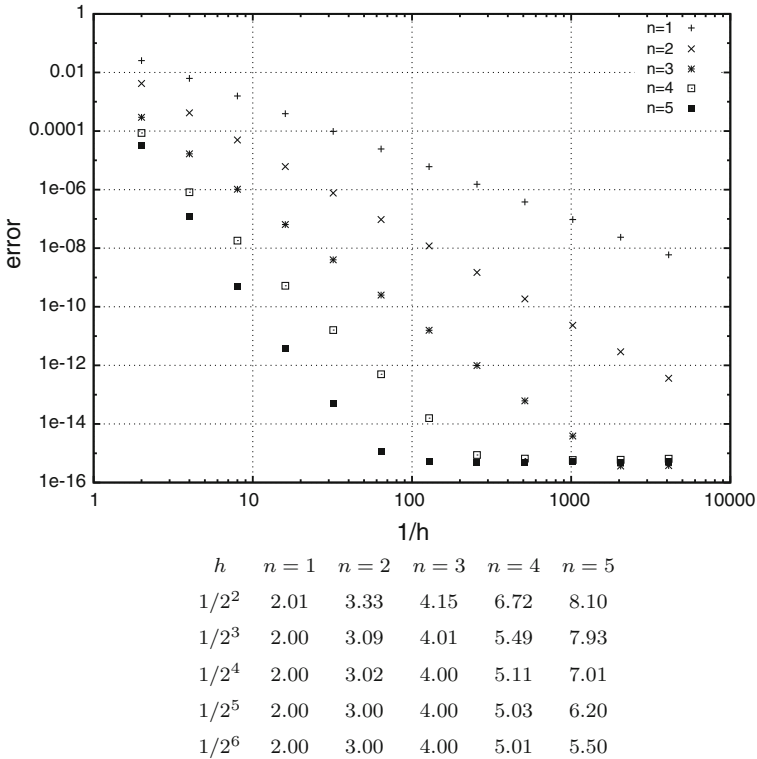
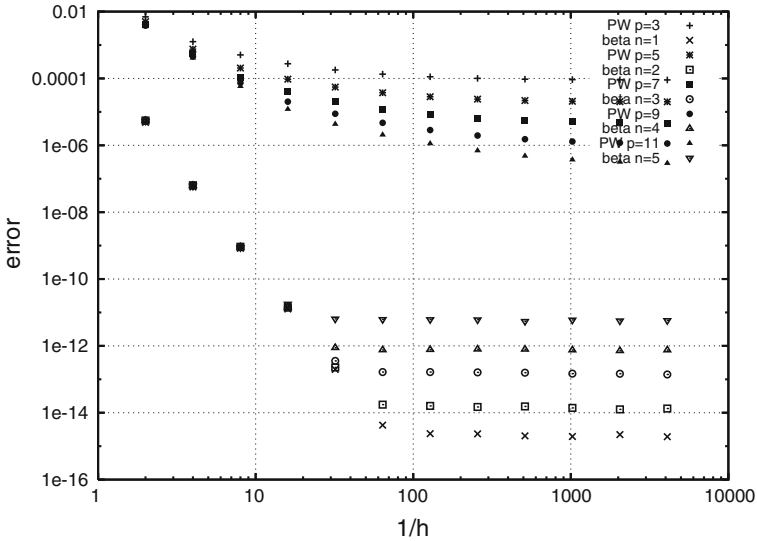


Fig. 5 Convergence results computed at $(1, 1) \in \mathbb{R}^2$ where $\beta(1, 1) = 0$ using Generalized Plane Waves with the constant-normalization. Some of the associated orders of convergence are also provided. In the legend, n is the parameter introduced in Theorem 1: in each case the expected order of convergence is $n + 1$

4.5 Toward the cut-off: $\beta \rightarrow 0$

Since $\beta(x = 1, y) = 0$, it is interesting to look at what happens with the β -normalization along this line. Again the value of h is $h = 1/2^k$, increasing the value of k . The point $\vec{g}_h = [1 - h, 1]$ remains in the propagative zone. Then disks are here centered on a point \vec{g}_h that stands at a distance h from the line $x = 1$, still with radius h . As a result all the disks are tangent to the cut-off line defined by $x = 1$. Classical Plane Waves are compared to the β -normalization with the same number of basis functions. Even if this test case does not correspond to the theoretical result proved in the present work, it is of significant interest for further inclusion of the GPWs to plane wave methods.

Figure 6 show that the β normalized Generalized plane waves give a high order approximation of u even getting closer to the vanishing line $x = 1$, as long as h is not too small. Note first that there seem to be a significant difference between the classical and generalized plane waves, as the error obtained with the classical plane waves rapidly reaches a threshold which is seven orders of magnitude bigger than with the



	n = 1		n = 2		n = 3		n = 4		n = 5	
h	PW	β	PW	β	PW	β	PW	β	PW	β
1/2 ²	2.35	2.37	3.27	3.24	3.17	4.09	3.21	5.32	3.21	6.25
1/2 ³	2.23	2.24	3.19	3.19	2.77	4.13	2.82	5.14	2.82	6.11
1/2 ⁴	2.14	2.15	3.12	3.11	2.47	4.09	2.40	5.08	2.40	6.03
1/2 ⁵	2.08	2.09	3.08	3.07	2.27	4.06	2.24	5.05	2.24	1.13
1/2 ⁶	2.05	2.04	3.05	3.04	2.15	4.04	2.13	4.00	2.13	-2.24

Fig. 6 Convergence results toward $\beta = 0$, computed at $(1 - h, 1) \in \mathbb{R}^2$. Comparison between Classical Plane Waves and Generalized Plane Waves with the β -normalization. Some of the associated orders of convergence are also provided. In the legend, *beta* and *PW* respectively refer to the β normalization and classical Plane Waves, while *n* is the parameter introduced in Theorem 1: in each case the expected order of convergence is $n + 1$, and *p* is the number of classical Plane Waves

GPWs. However it is obvious that, as the parameter n increases, the minimum error obtained with the β -normalized GPWs increases. It justifies the use of the constant-normalization for further applications.

Another possibility is to compare the influence of two parameters: the size of the disk h and the distance d between \vec{g} and the line $x = 1$. In this case, the error $e = \max |u_e - u_a|$ depends on both parameters, so one can write $e(h, d)$. Figure 7 displays the error computed for h and d convergence with the β -normalization. The h convergence is clearly damaged for decreasing values of d . This is linked to the low frequency limit when β goes to zero. However, looking at the h convergence with $d = h$, one can see that the error $e(h, h)$ converges as the error $e(h, 1/2)$ until $h = 1/2^5$.

$h \setminus d$	$1/2^1$	$1/2^2$	$1/2^3$	$1/2^4$	$1/2^5$	$1/2^6$	$1/2^7$	$1/2^8$	$1/2^9$	$1/2^{10}$
$1/2^1$	4.8e-06	5.5e-06	5.5e-06	5.4e-06	5.4e-06	5.3e-06	5.2e-06	5.2e-06	5.2e-06	5.2e-06
$1/2^2$	5.7e-08	6.4e-08	6.4e-08	6.2e-08	6.1e-08	6.0e-08	5.9e-08	5.8e-08	5.8e-08	6.9e-08
$1/2^3$	8.3e-10	9.2e-10	9.2e-10	9.0e-10	8.8e-10	8.7e-10	9.2e-10	1.2e-09	3.5e-09	2.4e-08
$1/2^4$	1.3e-11	1.4e-11	1.4e-11	1.4e-11	1.8e-11	3.6e-11	1.0e-10	5.4e-10	3.2e-09	2.2e-08
$1/2^5$	2.0e-13	2.3e-13	3.5e-13	8.8e-13	6.4e-12	2.8e-11	1.2e-10	5.4e-10	3.8e-09	2.2e-08
$1/2^6$	4.3e-15	1.7e-14	1.6e-13	7.6e-13	6.2e-12	3.0e-11	1.0e-10	6.0e-10	3.1e-09	2.0e-08
$1/2^7$	2.4e-15	1.6e-14	1.6e-13	7.7e-13	6.2e-12	2.8e-11	9.8e-11	5.1e-10	2.9e-09	2.3e-08
$1/2^8$	2.3e-15	1.5e-14	1.6e-13	7.9e-13	6.1e-12	2.7e-11	1.0e-10	5.0e-10	2.5e-09	1.6e-08
$1/2^9$	2.0e-15	1.5e-14	1.6e-13	7.9e-13	5.4e-12	2.5e-11	9.7e-11	4.9e-10	2.5e-09	1.9e-08
$1/2^{10}$	1.9e-15	1.4e-14	1.5e-13	7.5e-13	6.0e-12	2.5e-11	8.7e-11	5.0e-10	2.5e-09	1.8e-08

Fig. 7 Error computed on a disk of radius h centered at $(1-d; 1) \in \mathbb{R}^2$. The approximation is computed with β -normalized basis functions and with $n = 5$

5 Conclusion

A procedure to design a set of generalized plane waves that are locally approximate solution of the scalar wave equation has been successfully developed, the novelty being to consider smooth and non constant coefficients. It is to be noted that the design procedure is still valid as the coefficient vanishes. Both theoretical and numerical results evidence the high order approximation property of the GPWs, corresponding to h -convergence.

The design procedure could easily be generalized to many differential operators, as described in [23]. Moreover a natural idea would be to extend the generalization process from the phase to the amplitude of plane waves, by considering a looking for a shape function as $\varphi = Qe^P$ where P and Q are two polynomials.

Acknowledgments I would like to thank Peter Monk for bringing to my attention the importance of such interpolation properties and for his hospitality during my visit to the University of Delaware This visit was funded by the Fondation Pierre Ledoux. I would also like to thank Bruno Després for his help.

Appendix: Chain rule in dimension 1 and 2

For the sake of completeness, this section is dedicated to describing the formula to derive a composition of two functions, in dimensions one and two. A wide bibliography about this formula is to be found in [24]. It is linked to the notion of partition of an integer or the one of a set. The 1D version is not actually used in this work but is displayed here as a comparison with a 2D version, mainly concerning this notion of partition.

A.1 Faa Di Bruno formula

Faa Di Bruno formula gives the m th derivative of a composite function with a single variable. It is named after Francesco Faa Di Bruno, but was stated in earlier work of Louis F.A. Arbogast around 1800, see [8].

If f and g are functions with sufficient derivatives, then

$$\frac{d^m}{dx^m} f(g(x)) = m! \sum f^{(\sum_k b_k)}(g(x)) \prod_{k=1}^m \frac{1}{b_k!} \left(\frac{g^{(k)}(x)}{k!} \right)^{b_k},$$

where the sum is over all different solutions in nonnegative integers $(b_k)_{k \in \llbracket 1, m \rrbracket}$ of $\sum_k k b_k = m$. These solutions are actually the partitions of m .

A.2 Bivariate version

The multivariate formula has been widely studied, the version described here is the one from [7] applied to dimension 2. A linear order on \mathbb{N}^2 is defined by: $\forall (\mu, \nu) \in (\mathbb{N}^2)^2$, the relation $\mu < \nu$ holds provided that

1. $\mu_1 + \mu_2 < \nu_1 + \nu_2$; or
2. $\mu_1 + \mu_2 = \nu_1 + \nu_2$ and $\mu_1 < \nu_1$.

If f and g are functions with sufficient derivatives, then

$$\begin{aligned} \partial_x^i \partial_y^j f(g(x, y)) &= i! j! \sum_{1 \leq \mu \leq i+j} f^\mu(g(x, y)) \\ &\times \sum_{s=1}^{i+j} \sum_{p_s((i, j), \mu)} \prod_{l=1}^s \frac{1}{k_l!} \left(\frac{1}{i_l! j_l!} \partial_x^{i_l} \partial_y^{j_l} (g(x, y)) \right)^{k_l}, \end{aligned}$$

where the partitions of (i, j) are defined by the following sets: $\forall \mu \in \llbracket 1, i + j \rrbracket$, $\forall s \in \llbracket 1, i + j \rrbracket$, $p_s((i, j), \mu)$ is equal to

$$\left\{ (k_1, \dots, k_s; (i_1, j_1), \dots, (i_s, j_s)) : k_i > 0, 0 < (i_1, j_1) < \dots < (i_s, j_s), \sum_{l=1}^s k_l = \mu, \sum_{l=1}^s k_l i_l = i, \sum_{l=1}^s k_l j_l = j \right\}.$$

See [13] for a proof of the formula interpreted in terms of collapsing partitions.

References

1. Babuka, I., Sauter, S.A.: Is the pollution effect of the FEM avoidable for the Helmholtz’s equation considering high wave numbers? *SIAM Rev.* **42**, 451–484 (2000)
2. Betcke, T.: Numerical computation of eigenfunctions of planar regions, Thesis (2005)
3. Betcke, T., Phillips, J.: Approximation by dominant wave directions in plane wave methods. (unpublished)
4. Buffa, A., Monk, P.: Error estimates for the ultra weak variational formulation of the Helmholtz equation. *ESAIM Math. Model. Numer. Anal.* **42**, 925–940 (2008)
5. Cessenat, O., Després, B.: Application of an ultra weak variational formulation of elliptic PDEs to the two dimensional Helmholtz problem. *SIAM J. Numer. Anal.* **55**(1), 255–299 (1998)

6. Ciarlet, P.G.: The finite element method for elliptic problems. In: Studies in Mathematics and its Applications, vol. 4. North-Holland Publishing Co., Amsterdam (1978)
7. Constantine, G.M., Savits, T.H.: A multivariate Faà di Bruno formula with applications. *Trans. Am. Math. Soc.* **348**(2), 503–520 (1996)
8. Craik, A.D.D.: Prehistory of Faà di Bruno's formula. *Am. Math. Mon.* **112**(2), 119–130 (2005)
9. Demailly, J.P.: *Analyse numérique et équations différentielles* (Nelle édition) Grenoble Sciences, EDP Sciences, ISBN 9782759801121 (2012)
10. Després, B.: Sur une formulation variationnelle de type ultra-faible. *C. R. Acad. Sci. Paris Sér. I Math.* **318**(10), 939–944 (1994)
11. Farhat, C., Harari, I., Franca, L.: The discontinuous enrichment method. *Comput. Methods Appl. Mech. Eng.* **190**, 6455–6479 (2001)
12. Gittelson, C.J., Hiptmair, R., Perugia, I.: Plane wave discontinuous Galerkin methods: analysis of the h-version. *ESAIM Math. Model. Numer. Anal.* **43**, 297–331 (2009)
13. Hardy, M.: Combinatorics of partial derivatives. *Electron. J. Comb.* **13**(1), 13 (2006)
14. Henrici, P.: A survey of I. N. Vekua's theory of elliptic partial differential equations with analytic coefficients. *Z. Angew. Math. Phys.* **8**, 169–203 (1957)
15. Hiptmair, R.: Low frequency stable Maxwell formulations. Oberwolfach reports (2008)
16. Hiptmair, R., Moiola, A., Perugia, I.: Plane wave discontinuous Galerkin methods for the 2D Helmholtz equation: analysis of the p-version. *SIAM J. Numer. Anal.* **49**(1), 264–284 (2011)
17. Hiptmair, R., Moiola, A., Perugia, I.: Error analysis of Trefftz-discontinuous Galerkin methods for the time-harmonic Maxwell equations. *Math. Comput.* **82**(281), 247–268 (2013)
18. Hiptmair, R., Moiola, A., Perugia, I.: Plane wave discontinuous Galerkin methods: exponential convergence of the hp-version. SAM, ETH Zürich, Switzerland, Report 2013–31 (Submitted to *Found. Comput. Math*) (2013)
19. Huttunen, T., Malinen, M., Monk, P.: Solving Maxwell's equations using the ultra weak variational formulation. *J. Comput. Phys.* **223**(2), 731–758 (2007)
20. Huttunen, T., Monk, P., Kaipio, J.P.: Computational aspects of the ultra-weak variational formulation. *J. Comput. Phys.* **182**(1), 27–46 (2002)
21. Imbert-Gérard, L.-M., Després, B.: A generalized plane wave numerical method for smooth non constant coefficients, Tech. report R11034, LJLL-UPMC (2011)
22. Imbert-Gérard, L.-M., Despres, B.: A generalized plane-wave numerical method for smooth nonconstant coefficients. *IMA J. Numer. Anal.* (2013). doi:[10.1093/imanum/drt030](https://doi.org/10.1093/imanum/drt030)
23. Imbert-Gérard, L.-M.: Mathematical and numerical problems of some wave phenomena appearing in magnetic plasmas. Ph.D. thesis, oai:tel.archives-ouvertes.fr:tel-00870184 (2013)
24. Ma, T.W.: Higher chain formula proved by combinatorics. *Electron. J. Comb.* **16**(1), 7 (2009). (Note 21)
25. Melenk, J.: Operator adapted spectral element methods I: harmonic and generalized harmonic polynomials. *Numerische Mathematik* **84**, 35–69 (1999)
26. Moiola, A.: Trefftz-discontinuous Galerkin methods for time-harmonic wave problems. Diss. no. 19957, ETH Zürich, 2011 (2012)
27. Moiola, A., Spence, E.A.: Is the Helmholtz equation really sign-indefinite. *Siam Review* **56**(2):274–312 (2014)
28. Moiola, A., Hiptmair, R., Perugia, I.: Vekua theory for the Helmholtz operator. *Z. Angew. Math. Phys.* **62**(5), 779–807 (2011)
29. Pluymers, B., Hal, B., Vandepitte, D., Desmet, W.: Trefftz-based methods for time-harmonic acoustics. *Arch. Comput. Methods Eng.* **14**(4), 343–381 (2007)
30. Schoenberg, I.J.: On Hermite-Birkhoff interpolation. *J. Math. Anal. Appl.* **16**(3), 538–543 (1966)
31. Tezaur, R.: Discontinuous enrichment method for smoothly variable wavenumber medium-frequency Helmholtz problems. In: Proceedings of the 11th International Conference on Mathematical and Numerical Aspects of Waves, Tunis, Tunisie, pp. 353–354 (2013)
32. Trefftz, E.: Ein Gegenstück zum Ritzschen Verfahren. In: Proceedings of the 2nd International Congress on Applied Mechanics, Zürich, Switzerland, pp. 131–137 (1926)