Numerische Mathematik (2021) 149:87–137
https://doi.org/10.1007/s00211-021-01220-9

Numerische
Mathematik

A roadmap for Generalized Plane Waves and their interpolation properties

Lise-Marie Imbert-Gérard[1] · Guillaume Sylvand[2]

Received: 19 July 2019 / Revised: 14 January 2021 / Accepted: 14 January 2021 /
Published online: 23 August 2021

Abstract

This work focuses on the study of partial differential equation (PDE) based basis function for Discontinuous Galerkin methods to solve numerically wave-related boundary value problems with variable coefficients. To tackle problems with constant coefficients, wave-based methods have been widely studied in the literature: they rely on the concept of Trefftz functions, i.e. local solutions to the governing PDE, using oscillating basis functions rather than polynomial functions to represent the numerical solution. Generalized Plane Waves (GPWs) are an alternative developed to tackle problems with variable coefficients, in which case Trefftz functions are not available. In a similar way, they incorporate information on the PDE, however they are only approximate Trefftz functions since they don't solve the governing PDE exactly, but only an approximated PDE. Considering a new set of PDEs beyond the Helmholtz equation, we propose to set a roadmap for the construction and study of local interpolation properties of GPWs. Identifying carefully the various steps of the process, we provide an algorithm to summarize the construction of these functions, and establish necessary conditions to obtain high order interpolation properties of the corresponding basis.

Mathematics Subject Classification 65N30 · 65N80 · 68W25

✉ Lise-Marie Imbert-Gérard
lmig@math.arizona.edu

Guillaume Sylvand
guillaume.sylvand@airbus.com

[1] Department of Mathematics, The University of Arizona, Tucson, AZ, USA

[2] Central R&T, Airbus, AIRBUS Central R&T / XRV, 22 Rue du Gouverneur Général EBOUÉ,
92130 Issy Les Moulineaux, France

1 Introduction

Trefftz methods are Galerkin type of methods that rely on function spaces of local solutions to the governing partial differential equations (PDEs). They were initially introduced in [27,35], and the original idea was to use trial functions which satisfy the governing PDE to derive error bounds. They have been widely used in the engineering community [20] since the 60s, for instance for Laplace's equation [30], to the biharmonic equation [33] and to elasticity [23]. Later the general idea of taking advantage of analytical knowledge about the problem to build a good approximation space was used to develop numerical methods: in the presence of corner and interface singularities [10,34], boundary layers, rough coefficients, elastic interactions [2,3,28,29], wave propagation [2,9]. In the context of boundary value problems (BVPs) for time-harmonic wave propagation, several methods have been proposed following the idea of functions that solves the governing PDE [13], relying on incorporating oscillating functions in the function spaces to derive and discretize a weak formulation. Wave-based numerical methods have received attention from several research groups around the world, from the theoretical [13] and computational [14] point of view, and the pollution effect of plane wave Discontinuous Galerkin (DG) methods was studied in [11]. Such methods have also been implemented in industry codes,[1] for acoustic applications. The use of Plane Wave (PW) basis functions has been the most popular choice, while an attempt to use Bessel functions was reported in [25]. In [24], the authors present an interesting comparison of performance between high order polynomial and wave-based methods. More recently, application to space-time problems have been studied in [4,21,22,31,32].

In this context, numerical methods rely on discretizing a weak formulation via a set of exact solutions of the governing PDE. When no exact solutions to the governing PDE are available, there is no natural choice of basis functions to discretize the weak formulation. This is in particular the case for variable coefficient problems. In order to take advantage of Trefftz-type methods for problems with variable coefficients, Generalized Plane Waves (GPWs) were introduced in [17], as basis functions that are local approximate solutions—rather than exact solutions—to the governing PDE. GPWs were designed adding higher order terms in the phase of classical PWs, choosing these higher order terms to ensure the desired approximation of the governing PDE. In [15], the construction and interpolation properties of GPWs were studied for the Helmholtz equation

$$-\Delta u + \beta(x, y))u = 0, \tag{1}$$

with a particular interest for the case of a sign-changing coefficient β, including propagating solutions ($\beta < 0$), evanescent solutions ($\beta > 0$), smooth transition between them ($\beta = 0$) called cut-offs in the field of plasma waves. The interpolation properties of a set $\mathbb{V}$ spanned by resulting basis functions, namely $\|(I - P_{\mathbb{V}})u\|$ where $P_{\mathbb{V}}$ is the orthogonal projector on $\mathbb{V}$ while u is the solution to the original problem, play a crucial role in the error estimation of the corresponding numerical method [5]. For this

[1] http://www.waveller.com/Waveller_Acoustics/waveller_acoustics.shtml.

same equation the error analysis of a modified Trefftz method discretized with GPWs was presented in [18]. In [19], Generalized Plane Waves (GPWs) were used for the numerical simulation of mode conversion modeled by the following equation:

$$\left(\partial_x^2 + (d+\overline{d})\partial_x\partial_y + |d|^2\partial_y^2\right) F + (d-\overline{d})x\partial_y F \\ - \left(1 + \frac{1}{\mu} + x(x+y)\right) F = 0. \tag{2}$$

In the present work, we answer questions related to extending the work on GPW developed in [15]—the construction of GPWs on the one hand, and their interpolation properties on the other hand—from the Helmholtz operator $-\Delta + \beta$ to a wide range of partial differential operators. A construction process valid for some operators of order two or higher is presented, while a proof of interpolation properties is limited to some operators of order two. We propose a road map to identify crucial steps in our work:

1. Construction of GPWs φ such that $\mathcal{L}\varphi \approx 0$
   (a) Choose an ansatz for φ (Sect. 2).
   (b) Identify the corresponding N_{dof} degrees of freedom and N_{eqn} constraints (Sect. 2.1).
   (c) Choose the number of degrees of freedom adequately $N_{dof} \geq N_{eqn}$ (Sect. 2.1).
   (d) Study the structure of the resulting system and identify $N_{dof} - N_{eqn}$ additional constraints (Sects. 2.2, 2.3).
   (e) Compute the remaining N_{eqn} degrees of freedom at minimal computational cost (Sect. 2.4).
2. Interpolation properties
   (a) Study the properties of the remaining N_{eqn} degrees of freedom with respect to the $N_{dof} - N_{eqn}$ additional constraints
   (b) Identify a simple reference case depending only on the $N_{dof} - N_{eqn}$ additional constraints (Sect. 3).
   (c) Study the interpolation properties of this reference case (Sect. 4.1).
   (d) Relate the general case to the reference case (Sects. 3.1, 3.2).
   (e) Prove the interpolation properties of the GPWs from those of the reference case (Sect. 4.2).

We will consider linear partial differential operators with variable coefficients, defined as follows.

Definition 1 A linear partial differential operator of order $M \geq 2$, in two dimensions, with a given set of complex-valued coefficients $\alpha = \{\alpha_{k,\ell-k}, (k,\ell) \in \mathbb{N}^2, 0 \leq k \leq \ell \leq M\}$ will be denoted hereafter as

$$\mathcal{L}_{M,\alpha} := \sum_{\ell=0}^{M} \sum_{k=0}^{\ell} \alpha_{k,\ell-k}\,(x,y)\,\partial_x^k \partial_y^{\ell-k}.$$

Our goal is to build a basis of functions well suited to approximate locally any solution u to a given homogeneous variable-coefficient partial differential equation

$$\mathcal{L}_{M,\alpha} u = 0 \text{ on a domain } \Omega \subset \mathbb{R}^2,$$

where by locally we mean piecewise on a mesh $\mathcal{T}_h$ of Ω. Such interpolation properties are a building block for the convergence proof of Galerkin methods. For a constant coefficient operator, it is natural to use the same basis on each element $K \in \mathcal{T}_h$. However, with variable coefficients, it cannot be optimal to expect a single basis to have good approximation properties on the whole domain $\Omega \subset \mathbb{R}^2$. For instance, for the Helmholtz equation with a sign-changing coefficient, it can not be optimal to look for a single basis that would give a good approximation of solutions both in the propagating region and in the evanescent region. Therefore it is natural to think of local bases defined on each $K \in \mathcal{T}_h$: with GPWs we focus on local properties around a given point $(x_0, y_0) \in \mathbb{R}^2$ rather than on a given domain Ω. A simple idea would then be freezing the coefficients of the operator, that is to say studying, instead of $\mathcal{L}_{M,\alpha}$, the constant coefficient operator $\mathcal{L}_{M,\bar{\alpha}}$ with constant coefficients $\bar{\alpha} = \{\alpha_{k,l}(x_0, y_0), 0 \leq k+l \leq M\}$. However, as observed in [15,16], this leads to low order approximation properties, while we are interested in high order approximation properties. This is why new functions are needed to handle variable coefficients. This work will focus on two aspects: the construction and the interpolation properties of GPWs.

We follow the GPW design proposed in [15,17]. Retaining the oscillating feature while aiming for higher order approximation, GPW were designed with Higher Order Terms (HOT) in the phase function of a plane wave. These higher order terms are to be defined to ensure that a GPW function φ is an approximate solution to the PDE:

$$\begin{cases} \phi(x,y) = \exp i\kappa(\cos\theta x + \sin\theta y) \\ \left[-\Delta - \kappa^2\right]\phi = 0 \end{cases} \text{versus} \begin{cases} \varphi(x,y) = \exp\left(i\kappa(\cos\theta x + \sin\theta y) + HOT\right) \\ \mathcal{L}_{M,\alpha}\varphi \approx 0 \end{cases} \quad (3)$$

In Sect. 2, the construction of a GPW $\varphi(x,y) = e^{P(x,y)}$ will be described in detail, then a precise definition of GPW will be provided under the following hypothesis:

Hypothesis 1 Consider a given point $(x_0, y_0) \in \mathbb{R}^2$, a given approximation parameter $q \in \mathbb{N}, q \geq 1$, a given $M \in \mathbb{N}$, $M \geq 2$, and a partial differential operator $\mathcal{L}_{M,\alpha}$ defined by a given set of complex-valued coefficients $\alpha = \{\alpha_{k,l}, 0 \leq k+l \leq M\}$, defined in a neighborhood of (x_0, y_0), satisfying

- $\alpha_{k,l}$ is $\mathcal{C}^{q-1}$ at (x_0, y_0) for all (k, l) such that $0 \leq k+l \leq M$,
- $\alpha_{M,0}(x_0, y_0) \neq 0$.

This construction is equivalent to the construction of the bi-variate polynomial

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

and is performed by choosing the degree dP, and providing an explicit formula for the set of complex coefficients $\{\lambda_{ij}\}_{\{(i,j)\in\mathbb{N}^2, 0\leq i+j\leq dP\}}$, in order for φ to satisfy

$\mathcal{L}_{M,\alpha}\varphi(x,y) = O\left(\|(x,y)-(x_0,y_0)\|^q\right)$. An algorithm to construct a GPW is provided. In Sect. 3 properties of the λ_{ij}s are studied, while the interpolation properties of the corresponding set of basis functions are studied for the case $M=2$ in Sect. 4, under the following hypothesis:

Hypothesis 2 Under Hypothesis 1 we consider only operators $\mathcal{L}_{M,\alpha}$ such that M is even and the terms of order M satisfy

$$\sum_{k=0}^{M} \alpha_{k,M-k}(x_0,y_0)X^kY^{M-k} = (\gamma_1X^2+\gamma_2XY+\gamma_3Y^2)^{\frac{M}{2}}$$

for some complex numbers $(\gamma_1,\gamma_2,\gamma_3)$ such that there exists $(\mu_1,\mu_2)\in\mathbb{C}^2$, $\mu_1\mu_2\neq 0$, a non-singular matrix $A\in\mathbb{C}^{2\times 2}$ satisfying $\Gamma = A^tDA$ where $\Gamma = \begin{pmatrix}\gamma_1 & \gamma_2/2\\ \gamma_2/2 & \gamma_3\end{pmatrix}$ and $D=\begin{pmatrix}\mu_1 & 0\\ 0 & \mu_2\end{pmatrix}$, and therefore

$$\sum_{k=0}^{M} \alpha_{k,M-k}(x_0,y_0)X^kY^{M-k} = \Big(\mu_1(A_{11}X+A_{12}Y)^2+\mu_2(A_{21}X+A_{22}Y)^2\Big)^{\frac{M}{2}}.$$

For instance, these matrices are $\Gamma = D = Id$ for $\mathcal{L}_H := -\Delta-\kappa^2(x,y)$ or $\mathcal{L}_B := \Delta\mathcal{L}_H$, and $\Gamma = D = c(x_0,y_0)Id$ for $\mathcal{L}_C := -\nabla\cdot(c(x,y)\nabla)-\kappa^2(x,y)$. Note that if Γ is real, this is simply saying that its eigenvalues are non-zero. Finally, corresponding numerical results are then provided, for various operators $\mathcal{L}_{M,\alpha}$ of order $M=2$ in Sect. 5.

Our previous work was limited to the Helmholtz equation (1) for propagating and evanescent regions, transition between the two, absorbing regions, as well as caustics. The interpolation properties presented here cover more general second order equations, in particular equations that can be written as

$$\nabla\cdot(A\nabla u)+\mathbf{d}\cdot\nabla u+k^2mu=0, \tag{4}$$

with variable coefficients A matrix-valued, real and symmetric with non-zero eigenvalues, $\mathbf{d}$ vector-valued and m scalar-valued. It includes for instance

- Helmholtz equation with absorption corresponding to $A=I$ with $\Re(m)>0$ and $\Im(m)\neq 0$;
- the mild-slop equation [8] modeling the amplitude of the free-surface water waves corresponding to $m=c_pc_g$ being the product of c_p the phase speed of the waves and c_g the group speed of the waves with $A=mId$;
- if μ is the permeability and ϵ the permittivity, then the transverse-magnetic mode of Maxwell's equations for $A=\frac{1}{\mu}I$ and $m=\epsilon$, while the transverse-electric mode of Maxwell's equations for $A=\frac{1}{\epsilon}I$ and $m=\mu$.

Throughout this article, we will denote by $\mathbb{N}$ the set of non-negative integers, by $\mathbb{N}^*$ the set of positive integers, by $\mathbb{R}^+=[0;+\infty)$ the set of non-negative real numbers, and by $\mathbb{C}[z_1, z_2]$ the space of complex polynomials with respect to the two variables z_1 and z_2. As the first part of this work is dedicated to finding the coefficients λ_{ij}, we will reserve the word unknown to refer to the $\lambda_{i,j}$s. The length of the multi-index $(i, j) \in \mathbb{N}^2$ of an unknown λ_{ij}, $|(i, j)| = i + j$, will play a crucial role in what follows.

2 Construction of a GPW

The task of constructing a GPW is attached to a homogeneous PDE, it is not global on $\mathbb{R}^2$ but it is local as it is expressed in terms of a Taylor expansion. It consists in finding a polynomial $P \in \mathbb{C}[x, y]$ such that the corresponding GPW, namely $\varphi := e^P$, is locally an approximate solution to the PDE.

Consider $M = 2$, $\beta = \{\beta_{0,0}, \beta_{0,1} = \beta_{1,0} = \beta_{1,1} = 0, \beta_{0,2} \equiv -1, \beta_{2,0} \equiv -1\}$, and the corresponding the operator $\mathcal{L}_{2,\beta} = -\partial_x^2 - \partial_y^2 + \beta_{0,0}(x)$. Then for any polynomial $P \in \mathbb{C}[x, y]$:

$$\mathcal{L}_{2,\beta} e^{P(x,y)} = \left(-\partial_x^2 P - (\partial_x P)^2 - \partial_y^2 P - (\partial_y P)^2 + \beta_{0,0}(x, y)\right) e^{P(x,y)},$$

so the construction of an exact solution to the PDE would be equivalent to the following problem:

$$\begin{aligned} &\text{Find } P \in \mathbb{C}[x, y] \text{ such that} \\ &\quad \partial_x^2 P(x, y) + (\partial_x P)^2(x, y) + \partial_y^2 P(x, y) + (\partial_y P)^2(x, y) = \beta_{0,0}(x, y). \end{aligned} \tag{5}$$

Consider then the following examples.

- If $\beta_{0,0}(x, y)$ is constant, then it is straightforward to find a polynomial of degree one satisfying Problem (5); $\beta_{0,0}$ being negative this would correspond to a classical plane wave.
- If $\beta_{0,0}(x, y) = x$, then there is no solution to (5), since the total degree of the polynomial $\partial_x^2 P + (\partial_x P)^2 + \partial_y^2 P + (\partial_y P)^2$ is always even.
- If $\beta_{0,0}(x, y)$ is not a polynomial function, it is also straightforward to see that no polynomial P can satisfy Problem (5).

From these trivial examples we see that in general there is no such function, $\varphi(x, y) = e^{P(x,y)}$, P being a complex polynomial, solution to a variable coefficient partial differential equation exactly. It could seem that the restriction for P to be a polynomial is very strong. However since we are interested in approximation and smooth coefficients, rather than looking for a more general phase function we restrict the identity $\mathcal{L}\varphi = 0$ on Ω into an approximation on a neighborhood of $(x_0, y_0) \in \mathbb{R}^2$ in the following sense. We replace the too restrictive cancellation of $\mathcal{L}_{M,\alpha} e^{P(x,y)}$ by the cancellation of the lowest terms of its Taylor expansion around (x_0, y_0). So this section is dedicated to the construction of a polynomial $P \in \mathbb{C}[x, y]$, under Hypothesis 1, to ensure that the following local approximation property

$$\mathcal{L}_{M,\alpha} e^{P(x,y)} = O(\|(x - x_0, y - y_0)\|^q) \tag{6}$$

is satisfied. The parameter q will denote throughout this work the order of approximation of the equation to which the GPW is designed. In summary, the construction is performed:

- for a partial differential operator $\mathcal{L}_{M,\alpha}$ of order M defined by a set of smooth coefficients α,
- at a point $(x_0, y_0) \in \mathbb{R}^2$,
- at order $q \in \mathbb{N}^*$,
- to ensure that $\mathcal{L}_{M,\alpha} e^{P(x,y)} = O(|(x - x_0, y - y_0)|^q)$.

Even though the construction of a GPW will involve a non-linear system we propose to take advantage of the structure of this system to construct a solution via an explicit formula. In this way, even though a GPW $\varphi := e^P$ is a PDE-based function, the polynomial P can be constructed in practice from this formula, and therefore the function can be constructed without solving numerically any non-linear—or even linear—system. This remark is of great interest with respect to the use of such functions in a Discontinuous Galerkin method to solve numerically boundary value problems.

In order to illustrate the general formulas that will appear in this section, we will use the specific case $\mathcal{L}_{2,\gamma}$ where $\gamma = \{\gamma_{0,0}, \gamma_{1,0}, \gamma_{0,1}, \gamma_{2,0} \equiv -1, \gamma_{1,1}, \gamma_{0,2}\}$, for which we can write explicitly many formulas is a compact form. In order to simplify certain expressions that will follow we propose the following definition.

Definition 2 Assume $(i, j) \in \mathbb{N}^2$ and $(x_0, y_0) \in \mathbb{R}^2$. We define the linear partial differential operator $D^{(i,j)}$ by

$$\mathcal{D}^{(i,j)}: f \in \mathcal{C}^{i+j} \mapsto \frac{1}{i!j!}\partial_x^i \partial_y^j f.$$

A precise definition of GPW will be provided at the end of this section.

2.1 From the Taylor expansion to a non-linear system

We are seeking a polynomial $P(x, y) = \sum_{0\leq i+j\leq dP} \lambda_{ij}(x - x_0)^i (y - y_0)^j$ satisfying the Taylor expansion (6). Defining such a polynomial is equivalent to defining the set $\{\lambda_{ij}; (i, j) \in \mathbb{N}^2, 0 \leq i + j \leq dP\}$, and therefore we will refer to the λ_{ij}s as the unknowns throughout this construction process. The goal of this subsection is to identify the set of equations to be satisfied by these unknowns to ensure that P satisfies the Taylor expansion (6), and in particular to choose the degree of P so as to guarantee the presence of linear terms in each equation of the system.

According to the Faa di Bruno formula, the action of the partial differential operator $\mathcal{L}_{M,\alpha}$ on a function $\varphi(x, y) = e^{P(x,y)}$ is given by

$$\begin{aligned}\mathcal{L}_{M,\alpha} e^{P(x,y)} =& e^{P(x,y)} \Bigg(\alpha_{0,0}(x, y) \\ &+ \sum_{\ell=1}^{M} \sum_{k=0}^{\ell} \alpha_{k,\ell-k}(x, y)\, k!(\ell - k)! \sum_{1\leq\mu\leq\ell} \sum_{s=1}^{\ell} \sum_{p_s((k,\ell-k),\mu)}\end{aligned}$$

$$\prod_{m=1}^{s} \frac{1}{k_m!} \left(\mathcal{D}^{(i_m, j_m)} P(x, y)\right)^{k_m} \Bigg),$$

where the linear order $\prec$ on $\mathbb{N}^2$ is defined by

$$\forall(\mu, \nu) \in (\mathbb{N}^2)^2, \mu \prec \nu \Leftrightarrow \begin{array}{l} 1.\ \mu_1 + \mu_2 < \nu_1 + \nu_2;\ \text{or} \\ 2.\ \mu_1 + \mu_2 = \nu_1 + \nu_2 \text{ and } \mu_1 < \mu_2, \end{array}$$

and where $p_s((i, j), \mu)$ is equal to

$$\Bigg\{(k_1, \ldots, k_s; (i_1, j_1), \ldots, (i_s, j_s)) \colon k_i > 0, 0 \prec (i_1, j_1) \prec \cdots \prec (i_s, j_s),$$
$$\sum_{l=1}^{s} k_l = \mu, \sum_{l=1}^{s} k_l i_l = i, \sum_{l=1}^{s} k_l j_l = j \Bigg\}.$$

For the operator $\mathcal{L}_{2,\gamma}$ the Faa di Bruno formula becomes

$$\begin{aligned} \mathcal{L}_{2,\gamma} e^P = e^P \Big(& -\partial_x^2 P + \gamma_{1,1} \partial_x \partial_y P + \gamma_{0,2} \partial_y^2 P - (\partial_x P)^2 \\ & + \gamma_{1,1} \partial_x P \partial_y P + \gamma_{0,2} (\partial_y P)^2 \\ & + \gamma_{1,0} \partial_x P + \gamma_{0,1} \partial_y P + \gamma_{0,0} \Big). \end{aligned}$$

In order to single out the terms depending on P in the right hand side, this leads to the following definition.

Definition 3 Consider a given $M \in \mathbb{N}$, $M \geq 2$, a given set of complex-valued functions $\alpha = \{\alpha_{k,l}, 0 \leq k + l \leq M\}$, and the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$. We define the partial differential operator $\mathcal{L}_{M,\alpha}^A$ associated to $\mathcal{L}_{M,\alpha}$ as

$$\mathcal{L}_{M,\alpha}^A = \sum_{\ell=1}^{M} \sum_{k=0}^{\ell} k!(\ell - k)! \alpha_{k,\ell-k} \sum_{1 \leq \mu \leq \ell} \sum_{s=1}^{\ell} \sum_{p_s((k,\ell-k),\mu)} \prod_{m=1}^{s} \frac{1}{k_m!} \left(\mathcal{D}^{(i_m, j_m)}(\cdot)\right)^{k_m},$$

or equivalently, since the exponential of a bounded quantity is bounded away from zero:

$$\mathcal{L}_{M,\alpha}^A \colon f \in \mathcal{C}^M \mapsto \frac{\mathcal{L}_{M,\alpha} e^f}{e^f} - \alpha_{0,0}.$$

For the operator $\mathcal{L}_{2,\gamma}$ this gives

$$\begin{aligned} \mathcal{L}_{2,\gamma}^A P = & -\partial_x^2 P + \gamma_{1,1} \partial_x \partial_y P + \gamma_{0,2} \partial_y^2 P - (\partial_x P)^2 + \gamma_{1,1} \partial_x P \partial_y P \\ & + \gamma_{0,2} (\partial_y P)^2 + \gamma_{1,0} \partial_x P + \gamma_{0,1} \partial_y P. \end{aligned}$$

Since, for any polynomial P, the function e^P is locally bounded, and since $\mathcal{L}_{M,\alpha}[e^P] = \left(\mathcal{L}^A_{M,\alpha} e^P + \alpha_{0,0}\right)e^P$, then for a polynomial P to satisfy the approximation property (6), it is sufficient to satisfy

$$\mathcal{L}^A_{M,\alpha} P(x,y) = -\alpha_{0,0}(x,y) + O(|(x-x_0, y-y_0)|^q). \tag{7}$$

Therefore, the problem to be solved is now:

$$\begin{aligned} &\text{Find } P \in \mathbb{C}[x,y], \text{ s.t. } \forall (I,J) \in \mathbb{N}^2, 0 \le I+J < q,\\ &\mathcal{D}^{(I,J)}\mathcal{L}^A_{M,\alpha} P(x_0,y_0) = -\mathcal{D}^{(I,J)}\alpha_{0,0}(x_0,y_0). \end{aligned} \tag{8}$$

In order to define a polynomial $P(x,y) = \sum_{0\le i+j\le dP} \lambda_{ij}(x-x_0)^i(y-y_0)^j$, the degree dP of the polynomial determines the number of unknowns: there are $N_{dof} = \frac{(dP+1)(dP+2)}{2}$ unknowns to be defined, namely the $\{\lambda_{i,j}\}_{(i,j)\in\mathbb{N}, 0\le i+j\le dP}$. In order to design a polynomial P satisfying Eq. (7), the parameter q determines the number of equations to be solved: there are $N_{eqn} = \frac{q(q+1)}{2}$ terms to be canceled from the Taylor expansion. The first step toward the construction of a GPW is to define the value of dP for a given value of q.

At this point it is clear that if $dP \le q-1$, then the resulting system is overdetermined. Our choice for the polynomial degree dP relies on a careful examination of the linear terms in $\mathcal{L}^A_{M,\alpha}P$. We can already notice that, under Hypothesis 1, in $\mathcal{L}^A_{M,\alpha}P$ there is at least one non-zero linear term, namely $\alpha_{M,0}(x_0,y_0)\partial_x^M P$, and there is at least one non-zero non-linear term, namely $\alpha_{M,0}(x_0,y_0)(\partial_x P)^M$. This non-linear term corresponds to the following parameters from the Faa di Bruno formula: $\mu = M$, $s=1$, $(k_1,(i_1,j_1)) = (M,(1,0))$. The linear terms can only correspond to $s=1$, $\mu=1$ and $p_1((k,\ell-k),1) = \{(1,(k,\ell-k))\}$, see Definition 3. We can then split $\mathcal{L}^A_{M,\alpha}$ into its linear and non-linear parts.

Definition 4 Consider a given $M\in\mathbb{N}$, $M\ge 2$, a given set of complex-valued functions $\alpha = \{\alpha_{k,l}, 0\le k+l\le M\}$, and the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$. The linear part of the partial differential operator $\mathcal{L}^A_{M,\alpha}$ is defined by $\mathcal{L}^L_{M,\alpha} := \mathcal{L}_{M,\alpha} - \alpha_{0,0}\partial_x^0\partial_y^0$, or equivalently

$$\mathcal{L}^L_{M,\alpha} = \sum_{\ell=1}^{M}\sum_{k=0}^{\ell} \alpha_{k,\ell-k}\partial_x^k\partial_y^{\ell-k},$$

and its non-linear part $\mathcal{L}^N_{M,\alpha} := \mathcal{L}^A_{M,\alpha} - \mathcal{L}^L_{M,\alpha}$ can equivalently be defined by

$$\mathcal{L}^N_{M,\alpha} = \sum_{\ell=1}^{M}\sum_{k=0}^{\ell} k!(\ell-k)!\alpha_{k,\ell-k} \sum_{2\le\mu\le\ell}\sum_{s=1}^{\ell}\sum_{p_s((k,\ell-k),\mu)}\prod_{m=1}^{s}\frac{1}{k_m!}\left(\mathcal{D}^{(i_m,j_m)}(\cdot)\right)^{k_m}.$$

For the operator $\mathfrak{L}_{2,\gamma}$ this gives respectively

$$\begin{cases} \mathfrak{L}_{2,\gamma}^L P = -\partial_x^2 P + \gamma_{1,1}\partial_x\partial_y P + \gamma_{0,2}\partial_y^2 P + \gamma_{1,0}\partial_x P + \gamma_{0,1}\partial_y P, \\ \mathfrak{L}_{2,\gamma}^N P = -(\partial_x P)^2 + \gamma_{1,1}\partial_x P\partial_y P + \gamma_{0,2}(\partial_y P)^2. \end{cases}$$

Consider the (I, J) coefficients of the Taylor expansion of $\mathcal{L}_{M,\alpha}^L P$ for $(I, J) \in \mathbb{N}^2$ and $0 \le I + J < q$:

$$\mathcal{D}^{(I,J)}\left[\mathcal{L}_{M,\alpha}^L P\right](x_0, y_0) = \sum_{\ell=1}^{M}\sum_{k=0}^{\ell} \mathcal{D}^{(I,J)}\left[\alpha_{k,\ell-k}\partial_x^k\partial_y^{\ell-k} P\right](x_0, y_0),$$

so that in order to isolate the derivatives of highest order, i.e. of order $M + I + J$, we can write

$$\begin{aligned} &\mathcal{D}^{(I,J)}\left[\mathcal{L}_{M,\alpha}^L P\right](x_0, y_0) \\ &= \frac{1}{I!J!}\sum_{k=0}^{M}\alpha_{k,M-k}(x_0, y_0)\partial_x^{k+I}\partial_y^{M-k+J} P(x_0, y_0) \\ &\quad + \sum_{k=0}^{M}\sum_{\tilde{i}=0}^{I-1}\sum_{\tilde{j}=0}^{J-1}\frac{1}{\tilde{i}!\tilde{j}!}\mathcal{D}^{(I-\tilde{i},J-\tilde{j})}\alpha_{k,M-k}(x_0, y_0)\partial_x^{k+\tilde{i}}\partial_y^{M-k+\tilde{j}} P(x_0, y_0) \\ &\quad + \sum_{\ell=1}^{M-1}\sum_{k=0}^{\ell}\sum_{\tilde{i}=0}^{I}\sum_{\tilde{j}=0}^{J}\frac{1}{\tilde{i}!\tilde{j}!}\mathcal{D}^{(I-\tilde{i},J-\tilde{j})}\alpha_{k,\ell-k}(x_0, y_0)\partial_x^{k+\tilde{i}}\partial_y^{\ell-k+\tilde{j}} P(x_0, y_0). \quad (9) \end{aligned}$$

Back to Problem (8), the (I, J) terms (9) a priori depend on the unknowns $\{\lambda_{i,j}, (i, j) \in \mathbb{N}^2, 0 \le i + j \le dP\}$. Since

$$\forall (i, j) \in \mathbb{N}^2, \mathcal{D}^{(i,j)} P(x_0, y_0) = \begin{cases} \lambda_{i,j} & \text{if } i + j \le dP, \\ 0 & \text{otherwise,} \end{cases}$$

then under Hypothesis 1 any (I, J) term in System (8) has at least one non-zero linear term, as long as $I + J \le dP - M$, namely $\frac{(M+I)!}{I!}\alpha_{M,0}(x_0, y_0)\lambda_{M+I,J}$, while it does not necessarily have any linear term as soon as $I + J > dP - M$. Avoiding equations with no linear terms is natural, and it will be crucial for the construction process described hereafter.

Choosing the polynomial degree to be $dP = M + q - 1$ therefore guarantees the existence of at least one linear term in every equation of System (8). Therefore, from now on the polynomial P will be of degree $dP = M + q - 1$ and the new problem to be solved is

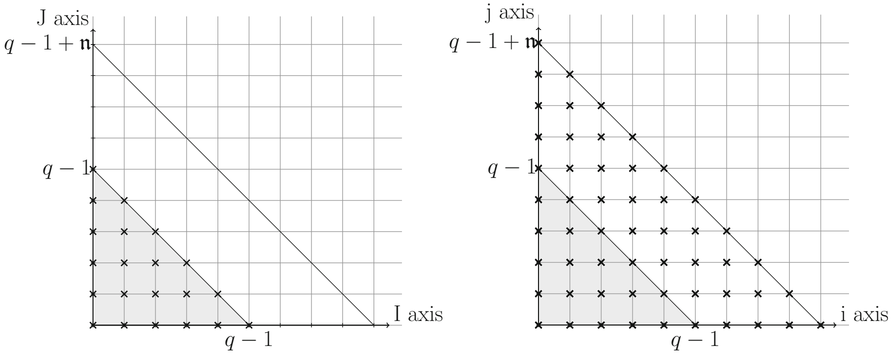

Fig. 1 Representation of the indices involved in the nonlinear system (10), for $q=6$ and $\mathfrak{n}=4$. Each cross in the (I,J) plane corresponds to the equation (I,J) in System (10) (Left panel), while each cross in the (i,j) plane corresponds to the unknown λ_{ij} (Right panel)

$$\begin{aligned}
&\text{Find } \{\lambda_{i,j}, (i,j)\in\mathbb{N}^2, 0\le i+j\le M+q-1\} \text{ such that}\\
&P(x,y):=\sum_{i=0}^{M+q-1}\sum_{j=0}^{M+q-1-i}\lambda_{i,j}(x-x_0)^i(y-y_0)^j\in\mathbb{C}[x,y], \text{ satisfies} \qquad (10)\\
&\forall(I,J)\in\mathbb{N}^2, 0\le I+J<q, \mathcal{D}^{(I,J)}\mathcal{L}_{M,\alpha}^{A}P(x_0,y_0)=-\mathcal{D}^{(I,J)}\alpha_{0,0}(x_0,y_0).
\end{aligned}$$

As a consequence the number of unknowns is $N_{dof}=\frac{(M+q)(M+q+1)}{2}$, and the system is under-determined : $N_{dof}-N_{eqn}=Mq+\frac{M(M+1)}{2}$. See Fig. 1 for an illustration of the equation and unknown count.

Note that this system is always non-linear. Indeed, under Hypothesis 1, the $(0,0)$ equation of the system always includes the non-zero non-linear term $\alpha_{M,0}(x_0,y_0)(\lambda_{1,0})^M$, corresponding to the following parameters from the Faa di Bruno formula: $\mu=M$, $s=1$, $(k_1,(i_1,j_1))=(M,(1,0))$.

The key to the construction procedure proposed next is a meticulous gathering of unknowns $\lambda_{i,j}$ with respect the length of their multi-index $i+j$. As we will now see, this will lead to splitting the system into a hierarchy of simple linear sub-systems.

2.2 From a non-linear system to linear sub-systems

The different unknowns appearing in each equation of System (10) can now be studied. A careful inspection of the linear and non-linear terms will reveal the underlying structure of the system, and will lead to identify a hierarchy of simple linear subsystems.

The inspection of the linear terms is very straightforward thanks to Eq. (9). The description of the unknowns in the linear terms is summarized here.

Lemma 1 *Consider a point* $(x_0,y_0)\in\mathbb{R}^2$, *a given* $q\in\mathbb{N}^*$, *a given* $M\in\mathbb{N}$, $M\ge 2$, *a given set of complex-valued functions* $\alpha=\{\alpha_{k,l}\in\mathcal{C}^{q-1}$ *at* $(x_0,y_0), 0\le k+l\le M\}$, *and the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$. In each equation (I, J) of System (10), the linear terms can be split as follows:*

- *a set of unknowns with length of the multi-index equal to $M+I+J$, corresponding to $\ell = M$ and $(\tilde{i}, \tilde{j}) = (I, J)$,*
- *a set of unknowns with length of the multi-index at most equal to $M+I+J-1$.*

Under Hypothesis 1, both sets are never empty.

Proof In terms of unknowns $\{\lambda_{i,j}, (i,j) \in \mathbb{N}^2, 0 \le i+j \le M+q-1\}$, Eq. (9) reads:

$$\begin{aligned}
&\partial_x^0 \partial_y^0 \left[\mathcal{L}_{M,\alpha}^L P\right](x_0, y_0)\\
&\quad = \sum_{k=0}^{M} (k)!(M-k)!\alpha_{k,M-k}(x_0,y_0)\lambda_{k,M-k}\\
&\qquad + \sum_{\ell=1}^{M-1}\sum_{k=0}^{\ell} (k)!\,(\ell-k)!\alpha_{k,\ell-k}(x_0,y_0)\lambda_{k,\ell-k};
\end{aligned} \tag{11}$$

$$\begin{aligned}
&\forall J > 0,\ \mathcal{D}^{(0,J)}\left[\mathcal{L}_{M,\alpha}^L P\right](x_0, y_0)\\
&\quad = \frac{1}{J!}\sum_{k=0}^{M} k!(M-k+J)!\alpha_{k,M-k}(x_0,y_0)\lambda_{k,M-k+J}\\
&\qquad + \sum_{k=0}^{M}\sum_{\tilde{j}=0}^{J-1} k!\frac{\left(M-k+\tilde{j}\right)!}{\tilde{j}!}\mathcal{D}^{(0,J-\tilde{j})}\alpha_{k,M-k}(x_0,y_0)\lambda_{k,M-k+\tilde{j}}\\
&\qquad + \sum_{\ell=1}^{M-1}\sum_{k=0}^{\ell}\sum_{\tilde{j}=0}^{J} k!\frac{\left(\ell-k+\tilde{j}\right)!}{\tilde{j}!}\mathcal{D}^{(0,J-\tilde{j})}\alpha_{k,\ell-k}(x_0,y_0)\lambda_{k,\ell-k+\tilde{j}};
\end{aligned} \tag{12}$$

$$\begin{aligned}
&\forall I > 0,\ \mathcal{D}^{(I,0)}\left[\mathcal{L}_{M,\alpha}^L P\right](x_0, y_0)\\
&\quad = \frac{1}{I!}\sum_{k=0}^{M} (k+I)!(M-k)!\alpha_{k,M-k}(x_0,y_0)\lambda_{k+I,M-k}\\
&\qquad + \sum_{k=0}^{M}\sum_{\tilde{i}=0}^{I-1} \frac{\left(k+\tilde{i}\right)!}{\tilde{i}!}(M-k)!\mathcal{D}^{(I-\tilde{i},0)}\alpha_{k,M-k}(x_0,y_0)\lambda_{k+\tilde{i},M-k}\\
&\qquad + \sum_{\ell=1}^{M-1}\sum_{k=0}^{\ell}\sum_{\tilde{i}=0}^{I} \frac{\left(k+\tilde{i}\right)!}{\tilde{i}!}(\ell-k)!\mathcal{D}^{(I-\tilde{i},0)}\alpha_{k,\ell-k}(x_0,y_0)\lambda_{k+\tilde{i},\ell-k};
\end{aligned} \tag{13}$$

$$\begin{aligned}
&\forall (I,J), IJ \neq 0,\ \mathcal{D}^{(I,J)}\left[\mathcal{L}_{M,\alpha}^L P\right](x_0, y_0)\\
&\quad = \frac{1}{I!J!}\sum_{k=0}^{M} (k+I)!(M-k+J)!\alpha_{k,M-k}(x_0,y_0)\lambda_{k+I,M-k+J}
\end{aligned}$$

$$+\sum_{k=0}^{M}\sum_{\tilde{i}=0}^{I-1}\sum_{\tilde{j}=0}^{J-1}\frac{\left(k+\tilde{i}\right)!}{\tilde{i}!}\frac{\left(M-k+\tilde{j}\right)!}{\tilde{j}!}\mathcal{D}^{(I-\tilde{i},J-\tilde{j})}\alpha_{k,M-k}(x_0,y_0)\lambda_{k+\tilde{i},M-k+\tilde{j}}$$
$$+\sum_{\ell=1}^{M-1}\sum_{k=0}^{\ell}\sum_{\tilde{i}=0}^{I}\sum_{\tilde{j}=0}^{J}\frac{\left(k+\tilde{i}\right)!}{\tilde{i}!}\frac{\left(\ell-k+\tilde{j}\right)!}{\tilde{j}!}\mathcal{D}^{(I-\tilde{i},J-\tilde{j})}\alpha_{k,\ell-k}(x_0,y_0)\lambda_{k+\tilde{i},\ell-k+\tilde{j}}. \tag{14}$$

The result is immediate for $I=J=0$ from (11). The following comments are valid for the right hand sides of (12)–(14): the third term only contains unknowns with a length of the multi-index equal to $\ell+\tilde{i}+\tilde{j}\leq M-1+I+J$, while the second term only contains unknowns with a length of the multi-index equal to $M+\tilde{i}+\tilde{j}\leq M+I+J-2$; as to the first term, it only contains unknowns with a length of the multi-index equal to $M+I+J$. This proves the claim. □

We then focus on the inspection of the non-linear terms. Each non-linear term in $\mathcal{L}_{M,\alpha}^{A}P$ reads from the definition of $\mathcal{L}_{M,\alpha}^{N}$

$$\alpha_{k,\ell-k}\prod_{m=1}^{s}\left(\partial_x^{i_m}\partial_y^{j_m}P\right)^{k_m}\ \text{with}\ \sum_{m=1}^{s}k_m>1 \tag{15}$$

and yields a sum of non-linear terms with respect to the unknowns $\{\lambda_{ij}\}_{\{(i,j),0\leq i+j\leq M+q-1\}}$, implicitly given by the following formula:

$$\begin{aligned}&\mathcal{D}^{(I,J)}\left[\alpha_{k,\ell-k}\prod_{m=1}^{s}\left(\partial_x^{i_m}\partial_y^{j_m}P\right)^{k_m}\right](x_0,y_0)\\&=\sum_{\tilde{i}=0}^{I}\sum_{\tilde{j}=0}^{J}\mathcal{D}^{(I-\tilde{i},J-\tilde{j})}\alpha_{k,\ell-k}(x_0,y_0)\mathcal{D}^{(\tilde{i},\tilde{j})}\left[\prod_{m=1}^{s}\left(\partial_x^{i_m}\partial_y^{j_m}P\right)^{k_m}\right](x_0,y_0).\end{aligned} \tag{16}$$

Therefore coming from the term (15), only a restricted number of unknowns contribute to the (I,J) equation of Problem (10).

In order to identify the unknowns contributing to (16), here are two simple yet important reminders are provided in “Appendix C”.

It is now straightforward to describe the unknowns $\lambda_{i,j}$ appearing in the non-linear terms of the equation (I,J) of System (10), unwinding formula (16).

Lemma 2 *Consider a point $(x_0,y_0)\in\mathbb{R}^2$, a given $q\in\mathbb{N}^*$, a given $M\in\mathbb{N}$, $M\geq 2$, a given set of complex-valued functions $\alpha=\{\alpha_{k,l}\in\mathcal{C}^{q-1}$ at $(x_0,y_0), 0\leq k+l\leq M\}$, and the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$. In each equation (I,J) of System* (10), *the unknowns $\lambda_{i,j}$ appearing in the non-linear terms have a length of the multi-index $i+j<M+I+J$.*

Proof Each term $\partial_x^{\tilde{i}}\partial_y^{\tilde{j}}\left[\prod_{m=1}^{s}\left(\partial_x^{i_m}\partial_y^{j_m}P\right)^{k_m}\right]$ in $\mathcal{L}_{M,\alpha}^{A}P$ is a polynomial, and its constant coefficient contains coefficients of the polynomial $\prod_{m=1}^{s}\left(\partial_x^{i_m}\partial_y^{j_m}P\right)^{k_m}$ with a length of the multi-index length of the multi-index at most equal to $\tilde{i}+\tilde{j}$, that is to say coefficients of the polynomials $\partial_x^{i_m}\partial_y^{j_m}P$ with a length of the multi-index length of the multi-index at most equal to $\tilde{i}+\tilde{j}$ for every (i_m, j_m) from the Faa di Bruno's formula, so coefficients $\lambda_{i,j}$ of the polynomial P with a length of the multi-index at most equal to $\tilde{i}+\tilde{j}+i_m+j_m$. Since the indices are such that $\tilde{i}\leq I$, $\tilde{j}\leq J$, and $i_m+j_m\leq\ell<M$, the unknowns $\lambda_{i,j}$ appearing in each term $\partial_x^{\tilde{i}}\partial_y^{\tilde{j}}\left[\prod_{m=1}^{s}\left(\partial_x^{i_m}\partial_y^{j_m}P\right)^{k_m}\right](x_0, y_0)$ have a length of the multi-index at most equal to $M+I+J-1$. It is therefore true for any linear combination such as (16). □

From the two previous Lemmas, we see that, in each equation (I, J) of System (10), unknowns with a length of the multi-index equal to $M+I+J$ appear only in linear terms, namely in

$$\sum_{k=0}^{n}\frac{(k+I)!}{I!}\frac{(M-k+J)!}{J!}\alpha_{k,M-k}(x_0, y_0)\lambda_{k+I,M-k+J},$$

whereas all the remaining unknowns have a length of the multi-index at most equal to $M+I+J-1$. It is consequently natural to subdivide the set of unknowns with respect to the length of their multi-index $M+\mathfrak{L}$, for $\mathfrak{L}$ between 0 and $q-1$ in order to take advantage of this linear structure.

2.3 Hierarchy of triangular linear systems

Our goal is now to construct a solution to the non-linear system (10), and our understanding of its linear part will lead to an explicit construction of such a solution without any need for any approximation.

The crucial point of our construction process is to take advantage of the underlying layer structure with respect to the length of the multi-index: it is only natural now to gather into subsystems all equations $(I, \mathfrak{L}-I)$ for I between 0 and $\mathfrak{L}$, while gathering similarly all unknowns with length of the multi-index equal to $M+\mathfrak{L}$. In the subsystem of layer $\mathfrak{L}$, we know that the unknowns with a length of the multi-index equal to $M+I+J$ only appear in linear terms, and we rewrite each equation (I, J) as

$$\begin{aligned}\sum_{k=0}^{n}&\frac{(k+I)!}{I!}\frac{(M-k+\mathfrak{L}-I)!}{(\mathfrak{L}-I)!}\alpha_{k,M-k}(x_0, y_0)\lambda_{k+I,M-k+\mathfrak{L}-I}\\ &=-\mathfrak{D}^{(I,J)}\alpha_{0,0}(x_0, y_0)-\mathfrak{D}^{(I,J)}\mathcal{L}_{M,\alpha}^{A}P(x_0, y_0)\\ &\quad+\sum_{k=0}^{n}\frac{(k+I)!}{I!}\frac{(M-k+\mathfrak{L}-I)!}{(\mathfrak{L}-I)!}\alpha_{k,M-k}(x_0, y_0)\lambda_{k+I,M-k+\mathfrak{L}-I}.\end{aligned}$$

For the sake of clarity, the resulting right-hand side terms can defined as follows.

Definition 5 Consider a point $(x_0, y_0) \in \mathbb{R}^2$, a given $q \in \mathbb{N}^*$, a given $M \in \mathbb{N}$, $M \geq 2$, a given set of complex-valued functions $\alpha = \{\alpha_{k,l} \in \mathcal{C}^{q-1} \text{ at } (x_0, y_0), 0 \leq k + l \leq M\}$, and the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$.

We define the quantity $N_{I,J}$ from Equation (I, J) from (10) as

$$\begin{aligned} N_{0,0} := &- \sum_{\ell=1}^{M-1} \sum_{k=0}^{\ell} (k)!\, (\ell-k)! \alpha_{k,\ell-k}(x_0, y_0) \lambda_{k,\ell-k} \\ &- \mathcal{L}_{M,\alpha}^{N} P(x_0, y_0) - \alpha_{0,0}(x_0, y_0); \end{aligned} \tag{17}$$

$$\begin{aligned} \forall J > 0,\ N_{0,J} := &\sum_{k=0}^{M} \sum_{\tilde{j}=0}^{J-1} \left(k+\tilde{i}\right)! \frac{\left(M-k+\tilde{j}\right)!}{\tilde{j}!} \mathcal{D}^{(0,J-\tilde{j})} \alpha_{k,M-k}(x_0, y_0) \lambda_{k,M-k+\tilde{j}} \\ &+ \sum_{\ell=1}^{M-1} \sum_{k=0}^{\ell} \sum_{\tilde{j}=0}^{J} (k)! \frac{\left(\ell-k+\tilde{j}\right)!}{\tilde{j}!} \mathcal{D}^{(0,J-\tilde{j})} \alpha_{k,\ell-k}(x_0, y_0) \lambda_{k,\ell-k+\tilde{j}} \\ &- \mathcal{D}^{(0,J)} \left[\mathcal{L}_{M,\alpha}^{N} P\right](x_0, y_0) - \mathcal{D}^{(0,J)} \alpha_{0,0}(x_0, y_0); \end{aligned} \tag{18}$$

$$\begin{aligned} \forall I > 0,\ N_{I,0} := &\sum_{k=0}^{M} \sum_{\tilde{i}=0}^{I-1} \frac{\left(k+\tilde{i}\right)!}{\tilde{i}!} (M-k)! \mathcal{D}_{k,M-k}^{(I-\tilde{i},\alpha)}(x_0, y_0) \lambda_{k+\tilde{i},M-k} \\ &+ \sum_{\ell=1}^{M-1} \sum_{k=0}^{\ell} \sum_{\tilde{i}=0}^{I} \frac{\left(k+\tilde{i}\right)!}{\tilde{i}!} (\ell-k)! \mathcal{D}^{(I-\tilde{i},0)} \alpha_{k,\ell-k}(x_0, y_0) \lambda_{k+\tilde{i},\ell-k} \\ &- \mathcal{D}^{(I,0)} \left[\mathcal{L}_{M,\alpha}^{N} P\right](x_0, y_0) - \mathcal{D}^{(I,0)} \alpha_{0,0}(x_0, y_0); \end{aligned} \tag{19}$$

$$\begin{aligned} &\forall (I, J),\ IJ \neq 0,\ N_{I,J} \\ &:= -\sum_{k=0}^{M} \sum_{\tilde{i}=0}^{I-1} \sum_{\tilde{j}=0}^{J-1} \frac{\left(k+\tilde{i}\right)! \left(M-k+\tilde{j}\right)!}{\tilde{i}!\tilde{j}!} \mathcal{D}^{(I-\tilde{i},J-\tilde{j})} \alpha_{k,M-k}(x_0, y_0) \lambda_{k+\tilde{i},M-k+\tilde{j}} \\ &\quad - \sum_{\ell=1}^{M-1} \sum_{k=0}^{\ell} \sum_{\tilde{i}=0}^{I} \sum_{\tilde{j}=0}^{J} \frac{\left(k+\tilde{i}\right)! \left(\ell-k+\tilde{j}\right)!}{\tilde{i}!\tilde{j}!} \mathcal{D}^{(I-\tilde{i},J-\tilde{j})} \alpha_{k,\ell-k}(x_0, y_0) \lambda_{k+\tilde{i},\ell-k+\tilde{j}} \\ &\quad - \mathcal{D}^{(I,J)} \left[\mathcal{L}_{M,\alpha}^{N} P\right](x_0, y_0) - \mathcal{D}^{(I,J)} \alpha_{0,0}(x_0, y_0). \end{aligned} \tag{20}$$

[EX] For the operator $\mathfrak{L}_{2,\gamma}$ the non-linear terms in $N_{0,0}$, $N_{1,0}$ and $N_{0,1}$ are respectively

$$\begin{aligned}
\mathfrak{L}^N_{2,\gamma} P(x_0,y_0) &= -\lambda_{1,0}^2 + \gamma_{1,1}(x_0,y_0)\lambda_{1,0}\lambda_{0,1} + \gamma_{0,2}(x_0,y_0)\lambda_{0,1}^2,\\
\partial_x[\mathfrak{L}^N_{2,\gamma} P](x_0,y_0) &= -2\lambda_{2,0}\lambda_{1,0} + \gamma_{1,1}(x_0,y_0)\left(2\lambda_{2,0}\lambda_{0,1} + \lambda_{1,0}\lambda_{1,1}\right)\\
&\quad + 2\gamma_{0,2}(x_0,y_0)\lambda_{1,1}\lambda_{0,1}\\
&\quad + \partial_x\gamma_{1,1}(x_0,y_0)\lambda_{1,0}\lambda_{0,1} + \partial_x\gamma_{0,2}(x_0,y_0)\lambda_{0,1}^2,\\
\partial_y[\mathfrak{L}^N_{2,\gamma} P](x_0,y_0) &= -2\lambda_{1,1}\lambda_{1,0} + \gamma_{1,1}(x_0,y_0)\left(\lambda_{1,1}\lambda_{0,1} + 2\lambda_{1,0}\lambda_{2,0}\right)\\
&\quad + 2\gamma_{0,2}(x_0,y_0)\lambda_{0,2}\lambda_{0,1}\\
&\quad + \partial_y\gamma_{1,1}(x_0,y_0)\lambda_{1,0}\lambda_{0,1} + \partial_y\gamma_{0,2}(x_0,y_0)\lambda_{0,1}^2.
\end{aligned}$$

We now consider the following subsystems for given $\mathfrak{L}$ between 0 and $q-1$:

$$\begin{cases}
\text{Find } \{\lambda_{i,j}, (i,j)\in\mathbb{N}^2, i+j = M+\mathfrak{L}\} \text{ such that}\\
\forall (I,J)\in\mathbb{N}^2, I+J=\mathfrak{L},\\
\displaystyle\sum_{k=0}^{M} \frac{(k+I)!(M-k+J)!}{I!J!}\alpha_{k,M-k}(x_0,y_0)\lambda_{k+I,M-k+J} = N_{I,J}.
\end{cases} \tag{21}$$

The layer structure follows from our understanding of the non-linearity of the original system:

Corollary 1 *Consider a point $(x_0,y_0)\in\mathbb{R}^2$, a given $q\in\mathbb{N}^*$, a given $M\in\mathbb{N}$, $M\geq 2$, a given set of complex-valued functions $\alpha = \{\alpha_{k,l}\in\mathcal{C}^{q-1}$ at (x_0,y_0), $0\leq k+l\leq M\}$, and the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$. For any $(I,J)\in\mathbb{N}^2$ such that $I+J<q$, the quantity $N_{I,J}$ only depends on unknowns $\lambda_{i,j}$ with length of the multi-index at most equal to $M+I+J-1$.*

Proof The result is straightforward from Lemmas 1 and 2 . □

Assuming that all unknowns $\lambda_{i,j}$ with length of the multi-index at most equal to $M+I+J-1$ are known, then (21) is a well-defined linear under-determined system with

- $\mathfrak{L}$ linear equations, namely the $(I,J) = (I,\mathfrak{L}-I)$ equations from System (10) for I between 0 and $\mathfrak{L}$;
- $M+\mathfrak{L}+1$ unknowns, namely the λ_{ij} for $i+j = M+\mathfrak{L}$.

Therefore, if all unknowns $\lambda_{i,j}$ with length of the multi-index at most equal to $M+I+J-1$ are known, we expect to be able to compute a solution to the subsystem $\mathfrak{L}$; this is the layer structure of our original problem (10). Figure 2 highlights the link between the layers of unknowns and equations of the initial nonlinear system on the one hand, and the layers unknowns and equations of the linear subsystems on the other hand.

At this stage, we have identified a hierarchy of under-determined linear subsystems, for increasing values of $\mathfrak{L}$ from 0 to $q-1$, and we are now going to propose one procedure to build a solution to each subsystem. There is no unique way to do so,

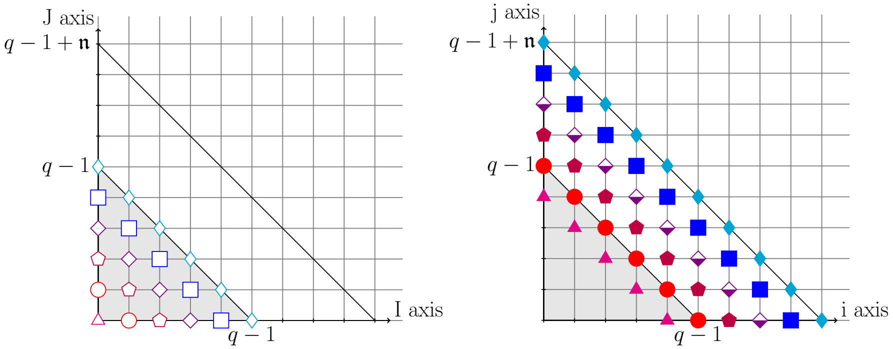

Fig. 2 Representation of the indices of equations and unknowns from the initial nonlinear system (10) divided up into linear subsystems (21). For $q = 6$ and $M = 4$, each shape of marker corresponds to one value of $\mathfrak{L}$: the indices (I, J) satisfying $I + J = \mathfrak{L}$ correspond to the subsystem's equations (Left panel), while the indices (i, j) satisfying $i + j = \mathfrak{L} + M$ correspond to the subsystem's unknowns (Right panel)

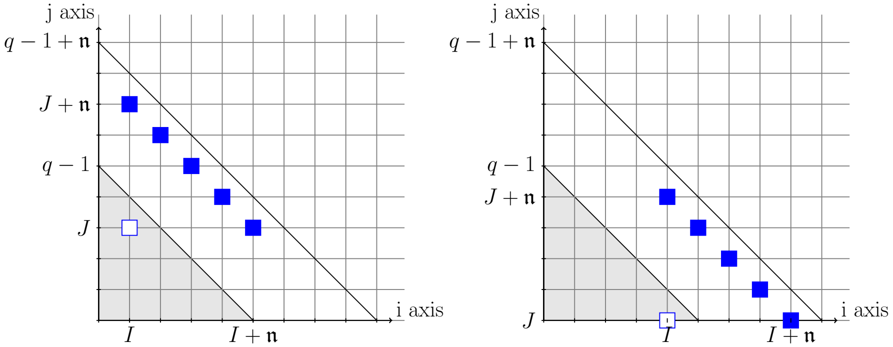

Fig. 3 Representation of the indices of unknowns involved in two equations (I, J) of the subsystem (21). For $q = 6$, for $M = 4$, and $\mathfrak{L} = 4$, each filled blue square marker corresponds in the (i, j) plane to an unknown λ_{ij}, involved in the $(I, J) = (1, 3)$ equation (Left panel), or in the $(I, J) = (4, 0)$ equation (Right panel) (colour figure online)

however if either $\alpha_{M,0}(x_0, y_0) \neq 0$ or $\alpha_{0,M}(x_0, y_0) \neq 0$ it provides a natural way to proceed. Indeed, the unknowns involved in an equation $(I, J) = (I, \mathfrak{L} - I)$ are $\{\lambda_{i,M+\mathfrak{L}-i}; i \in \mathbb{N}, I \leq i \leq I + M\}$; and the coefficient of the unknown $\lambda_{I+M,\mathfrak{L}-I}$ is proportional to $\alpha_{M,0}(x_0, y_0)$, which is non-zero under Hypothesis 1. Figure 3 provides two examples, in the (i, j) plane, of the indices of one equation's unknowns: for each equation, the coefficient of the term corresponding to the rightmost marker is non-zero. By adding M constraints corresponding to fixing the values of $\lambda_{i,M+\mathfrak{L}-i}$ for $0 \leq i < M$, that is the unknowns corresponding in the (i, j) plane to first M markers on the left at level $M + \mathfrak{L}$, we therefore guarantee that for increasing values of I from 0 to $\mathfrak{L}$ we can compute successively $\lambda_{I+M,\mathfrak{L}-I}$.

We can easily recast this in terms of matrices. At each level $\mathfrak{L}$, numbering the equations with increasing values of I and the unknowns with increasing values of i highlights the band-limited structure of each subsystem, while the entries of the

Mth super diagonal are all proportional to $\alpha_{M,0}(x_0, y_0)$, and therefore non-zero under Hypothesis 1. The matrix of the square linear system at level $\mathfrak{L}$ is then constructed from the first M lines of the identity, corresponding to the additional M constraints, placed on top of the matrix of the subsystem.

Definition 6 Consider a point $(x_0, y_0) \in \mathbb{R}^2$, a given $q \in \mathbb{N}^*$, a given $M \in \mathbb{N}$, $M \geq 2$, a given set of complex-valued functions $\alpha = \{\alpha_{k,l} \in \mathcal{C}^{q-1}$ at $(x_0, y_0), 0 \leq k + l \leq M\}$, and the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$. For a given level $\mathfrak{L} \in \mathbb{N}$ with $\mathfrak{L} < q$, we define the matrix of the square system of level $\mathfrak{L}$, $\mathsf{T}^{\mathfrak{L}} \in \mathbb{C}^{(M+\mathfrak{L}+1)\times(M+\mathfrak{L}+1)}$, as

$$\begin{cases} \mathsf{T}^{\mathfrak{L}}_{k+1,k+1} & = 1, & \forall k \ s.t.\ 0 \leq k \leq M-1, \\ \mathsf{T}^{\mathfrak{L}}_{I+M+1,I+k+1} & = \dfrac{(I+k)!(M-k+\mathfrak{L}-I)!}{I!(\mathfrak{L}-I)!}\alpha_{k,M-k}(x_0,y_0), & \forall (k,I) \ s.t.\ 0 \leq k \leq M,\ 0 \leq I \leq \mathfrak{L}, \\ \mathsf{T}^{\mathfrak{L}}_{k,k'} & = 0, & \text{otherwise,} \end{cases}$$

or equivalently

$$\mathsf{T}^{\mathfrak{L}} := \left[\begin{array}{cccccc} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ \hline \Pi_0^{0,\mathfrak{L}} A_0 & \cdots & \cdots & \Pi_M^{0,\mathfrak{L}} A_M & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & \Pi_0^{\mathfrak{L},\mathfrak{L}} A_0 & \cdots & \cdots & \Pi_M^{\mathfrak{L},\mathfrak{L}} A_M \end{array}\right] \text{ with } \begin{cases} \Pi_k^{i,\mathfrak{L}} := \frac{(k+i)!(M-k+\mathfrak{L}-i)!}{i!(\mathfrak{L}-i)!}, \\ A_k := \alpha_{k,M-k}(x_0,y_0). \end{cases}$$

Assuming that all unknowns $\lambda_{i,j}$ with length of the multi-index at most equal to $M + I + J - 1$ are known, then, as expected, a solution to the linear under-determined system (21) can be computed as follows.

Proposition 1 *Consider a point $(x_0, y_0) \in \mathbb{R}^2$, a given $q \in \mathbb{N}^*$, a given $M \in \mathbb{N}$, $M \geq 2$, a given set of complex-valued functions $\alpha = \{\alpha_{k,l} \in \mathcal{C}^{q-1}$ at $(x_0, y_0), 0 \leq k + l \leq M\}$, and the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$. For a given level $\mathfrak{L} \in \mathbb{N}$ with $\mathfrak{L} < q$, under Hypothesis 1, the matrix $\mathsf{T}^{\mathfrak{L}} \in \mathbb{C}^{(M+\mathfrak{L}+1)\times(M+\mathfrak{L}+1)}$ is non-singular.*

We now assume that the unknowns $\{\lambda_{i,j}, (i, j) \in \mathbb{N}^2, i + j < M + \mathfrak{L}\}$ are known, so that the terms $N_{I,\mathfrak{L}-I}$ for I from 0 to $\mathfrak{L}$ can be computed. Consider any vector $\mathsf{B}^{\mathfrak{L}} \in \mathbb{C}^{M+\mathfrak{L}+1}$ satisfying

$$\mathsf{B}^{\mathfrak{L}}_{M+1+I} = N_{I,\mathfrak{L}-I}, \forall I \ s.t.\ 0 \leq I \leq \mathfrak{L}.$$

Then independently of the first M components of $\mathsf{B}^{\mathfrak{L}}$, solving the linear system

$$\mathsf{T}^{\mathfrak{L}}\mathsf{X}^{\mathfrak{L}} = \mathsf{B}^{\mathfrak{L}} \tag{22}$$

by forward substitution provides a solution to (21) for

$$\lambda_{i,M+\mathfrak{L}-i} = \mathsf{X}^{\mathfrak{L}}_{i+1}, \ \forall i \in \mathbb{N} \text{ such that } 0 \leq i \leq M + \mathfrak{L}.$$

Proof The matrix $\mathsf{T}^{\mathfrak{L}}$ is lower triangular, therefore its determinant is

$$\det \mathsf{T}^{\mathfrak{L}} = \prod_{I=0}^{\mathfrak{L}} \left(\frac{(I+M)!(\mathfrak{L}-I)!}{I!(\mathfrak{L}-I)!} \alpha_{M,0}(x_0,y_0) \right) = \left(\prod_{I=0}^{\mathfrak{L}} \frac{(I+M)!}{I!} \right) \left(\alpha_{M,0}(x_0,y_0)\right)^{\mathfrak{L}+1},$$

which can not be zero under Hypothesis 1. The second part of the claim derives directly from the definition of $\mathsf{T}^{\mathfrak{L}}$ and $\mathsf{B}^{\mathfrak{L}}$ and the fact that the system is lower triangular, and can be illustrated as follows:

$$\underbrace{\left[\begin{array}{cccccc} 1 & & & & & \\ & \ddots & & & & \\ & & & 1 & & \\ \hline \Pi_0^{0,\mathfrak{L}} A_0 & \cdots & \cdots & \Pi_M^{0,\mathfrak{L}} A_M & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & \Pi_0^{\mathfrak{L},\mathfrak{L}} A_0 & \cdots & \cdots & \Pi_M^{\mathfrak{L},\mathfrak{L}} A_M \end{array}\right]}_{\mathsf{T}^{\mathfrak{L}}} \underbrace{\left[\begin{array}{c} \lambda_{0,\mathfrak{L}+M} \\ \vdots \\ \lambda_{M-1,\mathfrak{L}+1} \\ \hline \lambda_{M,\mathfrak{L}} \\ \vdots \\ \lambda_{\mathfrak{L}+M,0} \end{array}\right]}_{\mathsf{X}^{\mathfrak{L}}} = \underbrace{\left[\begin{array}{c} * \\ \vdots \\ * \\ \hline N_{0,\mathfrak{L}} \\ \vdots \\ N_{\mathfrak{L},0} \end{array}\right]}_{\mathsf{B}^{\mathfrak{L}}}$$

□

To summarize, we have defined for increasing values of $\mathfrak{L}$ a hierarchy of linear systems, each of which has the following characteristics:

- its unknowns are $\{\lambda_{i,M+\mathfrak{L}-i};\ \forall i \in \mathbb{N} \text{ such that } 0 \le i \le M+\mathfrak{L}\}$;
- its matrix $\mathsf{T}^{\mathfrak{L}} \in \mathbb{C}^{(M+\mathfrak{L}+1)\times(M+\mathfrak{L}+1)}$ is a square, non-singular, and triangular;
- its right-hand side depends both on $\{\lambda_{i,j};\ \forall (i,j) \in \mathbb{N}^2 \text{ such that } 0 \le i+j < M+\mathfrak{L}\}$ and on M additional parameters.

At each level $\mathfrak{L}$, assuming that the unknowns of inferior levels are known and provided M given values for $\lambda_{i,M+\mathfrak{L}-i}$ for $0 \le i < M$, Proposition 1 provides an explicit formula to compute $\lambda_{i,M+\mathfrak{L}-i}$ for $M \le i \le M+\mathfrak{L}$.

2.4 Algorithm

The non-linear system (10) had $N_{dof}^{(10)} = \frac{(M+q)(M+q+1)}{2}$ unknowns and $N_{eqn}^{(10)} = \frac{q(q+1)}{2}$ equations, whereas each linear triangular system introduced in the previous subsection has $N_{dof}^{T} = M+\mathfrak{L}+1$ unknowns and $N_{eqn}^{T} = M+\mathfrak{L}+1$ equations for each level $\mathfrak{L}$ such that $0 \le \mathfrak{L} \le q-1$. Therefore the hierarchy of triangular systems has a total of $N_{dof}^{H} = (M+1)q + \frac{q(q-1)}{2}$ unknowns and $N_{eqn}^{H} = N_{eqn}^{(10)} + Mq = Mq + \frac{q(q+1)}{2}$ equations, including the $\frac{q(q+1)}{2}$ equations of the initial non-linear system (10).

The remaining $N_{dof}^{(10)} - N_{dof}^{T} = \frac{M(M+1)}{2}$ unknowns, which are unknowns of none of the triangular systems but appear only on the right hand side of these systems, are the $\{\lambda_{i,j}, (i,j) \in \mathbb{N}^2, 0 \le i+j < M\}$. These are the unknowns with length of the multi-index at most equal to $M-1$, and the corresponding indices (i,j) are the only ones that are not marked on the right panel of Fig. 2. It is therefore natural to add

$\frac{M(M+1)}{2}$ constraints corresponding to fixing the values of the remaining unknowns $\{\lambda_{i,j}, (i,j) \in \mathbb{N}^2, 0 \leq i+j < M\}$. The final system we consider consists of these $\frac{M(M+1)}{2}$ constraints, guaranteeing that the unknowns $\{\lambda_{i,j}, (i,j) \in \mathbb{N}^2, 0 \leq i+j < M\}$ are known, together with the hierarchy of triangular systems (22) for increasing values of $\mathfrak{L}$ from 0 to $q-1$; it has $N^F_{dof} = \frac{(M+q)(M+q+1)}{2}$ unknowns, namely the unknowns of the original system (10), and $N^F_{eqn} = \frac{(M+q)(M+q+1)}{2}$ equations, namely the equations of the original system split into linear subsytems together with a total of $\frac{M(M+1)}{2} + qM$ additional constraints. A counting summary is presented here: Thanks to the $\frac{M(M+1)}{2}$ constraints, for increasing values of $\mathfrak{L}$ from 0 to $q-1$, the hypothesis of Proposition 1 is satisfied, the right hand side $\mathsf{B}^{\mathfrak{L}}$ can be evaluated and the triangular system (22) can be solved. So the unknowns $\{\lambda_{i,M+\mathfrak{L}-i};\ \forall i \in \mathbb{N} \text{ such that } 0 \leq i \leq M+\mathfrak{L}\}$ can be computed by induction on $\mathfrak{L}$, constructing a solution to the initial non-linear system (10) by induction on $\mathfrak{L}$.

	Number of unknowns	Number of equations
Original non-linear system (10)	$N^{(10)}_{dof} = \frac{(M+q)(M+q+1)}{2}$	$N^{(10)}_{eqn} = \frac{q(q+1)}{2}$
Subsystem at level $\mathfrak{L}$ (21)	$N^{\mathfrak{L}}_{dof} = M + \mathfrak{L} + 1$	$N^{\mathfrak{L}}_{eqn} = \mathfrak{L} + 1$
Triangular system at level $\mathfrak{L}$ (22)	$N^{T}_{dof} = M + \mathfrak{L} + 1$	$N^{T}_{eqn} = M + \mathfrak{L} + 1$
Hierarchy of triangular systems for $\mathfrak{L}$ from 0 to $q-1$	$N^{H}_{dof} = (M+1)q + \frac{q(q-1)}{2}$	$N^{H}_{eqn} = Mq + \frac{q(q+1)}{2}$
Final system (initial constraints + triangular systems)	$N^{F}_{dof} = \frac{(M+q)(M+q+1)}{2}$	$N^{F}_{eqn} = \frac{(M+q)(M+q+1)}{2}$

The following algorithm requires the value of $\frac{M(M+1)}{2} + qM$ parameters, to fix initially the set of unknowns $\{\lambda_{i,j}, (i,j) \in \mathbb{N}^2, 0 \leq i+j < M\}$ and then at each level $\mathfrak{L}$ the set of unknowns $\{\lambda_{i,M+\mathfrak{L}-i}, i \in \mathbb{N}, 0 \leq i < M\}$. Under Hypothesis 1, the algorithm presents a sequence of steps to construct explicitly a solution to Problem (10) and requires no approximation process.

Algorithm 1 Constructing a solution to Problem (10)

```
1: Fix $\{\lambda_{i,j}, (i,j) \in \mathbb{N}^2, 0 \leq i+j < M\}$    ▷ $\frac{M(M+1)}{2}$ unknowns
2: for $\mathfrak{L}$ from 0 to $q-1$ do    ▷ $q$ times
3:   Fix $\{\lambda_{i,M+\mathfrak{L}-i}, i \in \mathbb{N}, 0 \leq i < M\}$    ▷ $M$ unknowns
4:   for $I$ from 0 to $\mathfrak{L}$ do    ▷ $\mathfrak{L}+1$ times
5:     $\lambda_{I+M,\mathfrak{L}-I} := \frac{1}{\mathsf{T}^{\mathfrak{L}}_{I+M+1,I+M+1}} \left( \mathsf{B}^{\mathfrak{L}}_{I+M+1} - \sum_{k=0}^{M-1} \mathsf{T}^{\mathfrak{L}}_{I+M+1,I+k+1} \lambda_{I+k,M+\mathfrak{L}-I-k} \right)$    ▷ 1 unknown
```

From the definitions of $\mathsf{T}^{\mathfrak{L}}$ and $\mathsf{B}^{\mathfrak{L}}$ we immediately see that the step 5 boils down to

$$\lambda_{I+M,\mathfrak{L}-I} = \frac{I!}{(I+M)!\alpha_{M,0}(x_0,y_0)} \times \left(N_{I,\mathfrak{L}-I} - \sum_{k=0}^{M-1} \frac{(I+k)!(M-k+\mathfrak{L}-I)!}{I!(\mathfrak{L}-I)!} \alpha_{k,M-k}(x_0,y_0)\lambda_{I+k,M+\mathfrak{L}-I-k} \right) \tag{23}$$

If the set of unknowns $\{\lambda_{i,j}, (i,j) \in \mathbb{N}^2, 0 \leq i+j < M+q-1\}$ is computed from Algorithm 1, then the polynomial $P(x,y) := \sum_{0\leq i+j\leq q+M-1} \lambda_{i,j}(x-x_0)^i(y-y_0)^j$ is a solution to Problem (10), and therefore the function $\varphi(\mathbf{x}) := \exp P(\mathbf{x})$ satisfies (6). This is true independently of the values fixed in lines 1.1 and 1.3 of the algorithm.

Remark 1 It is interesting to notice that the algorithm applies to a wide range of partial differential operators, including type changing operators such as Keldysh operators, $L_K = \partial_x^2 + y^{2m+1}\partial_y^2 +$ lower order terms, or Tricomi operators, $L_T = \partial_x^2 + x^{2m+1}\partial_y^2 +$ lower order terms, that change from elliptic to hyperbolic type along a smooth parabolic curve.

To conclude this section, we provide a formal definition of a GPW associated to an partial differential operator at a given point.

Definition 7 Consider a point $(x_0,y_0) \in \mathbb{R}^2$, a given $q \in \mathbb{N}^*$, a given $M \in \mathbb{N}$, $M \geq 2$, a given set of complex-valued functions $\alpha = \{\alpha_{k,l} \in \mathcal{C}^{q-1}$ at $(x_0,y_0), 0 \leq k+l \leq M\}$, and the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$. A Generalized Plane Wave (GPW) associated to the differential operator $\mathcal{L}_{M,\alpha}$ at the point (x_0,y_0) is a function φ satisfying

$$\mathcal{L}_{M,\alpha}\varphi(x,y) = O(\|(x-x_0,y-y_0)\|^q).$$

Under Hypothesis 1, a Generalized Plane Wave (GPW) can be constructed as function $\varphi(x,y) = \exp P(x,y)$, where the coefficients of the polynomial P are computed by Algorithm 1, independently of the values fixed in the algorithm.

The crucial feature of the construction process is the exact solution provided in the algorithm: in practice, a solution to the initial non-linear rectangular system is computed without numerical resolution of any system, *with an explicit formula*.

The choice of the fixed values in Algorithm 1 will be discussed in the next paragraph. Even though these values does not affect the construction process, and the fact that the corresponding $\varphi(x,y) = \exp P(x,y)$ is a GPW, it will be key to prove the interpolation properties of the corresponding set of GPWs.

Remark 2 Under the hypothesis $\alpha_{0,M}(x_0,y_0) \neq 0$ it would be natural to fix the values of $\{\lambda_{i,j}, 0 \leq j \leq M-1, 0 \leq i \leq q+M-1-j\}$ instead of those of $\{\lambda_{i,j}, 0 \leq i \leq M-1, 0 \leq j \leq q+M-1-i\}$, and an algorithm very similar to Algorithm 1, exchanging the roles of i and j would construct the polynomial coefficients of a GPW.

3 Normalization

We will refer to normalization as the choice of imposed values in Algorithm 1. The discussion presented in this section will be summarized in Definition 8.

Within the construction process presented in the previous section, only the design of the function φ as the exponential of a polynomial is related to wave propagation, while Algorithm 1 works for partial differential operators not necessarily related to wave propagation. In particular, the property $\mathfrak{L}_{M,\alpha}\varphi(x,y) = O\left(\|(x,y)-(x_0,y_0)\|^q\right)$ of GPWs is independent of the choice of $(\lambda_{1,0},\lambda_{0,1})$. However, the normalization process described here carries on the idea of adding higher order terms to the phase function of a plane wave, see (3), as was proposed in [15].

We will now restrict our attention to a smaller set of partial differential operators that include several interesting operators related to wave propagation, thanks to an additional hypothesis on the highest order derivatives in $\mathfrak{L}_{M,\alpha}$, namely Hypothesis 2. Under this hypothesis we will be able to study the interpolation properties of associated GPWs in a unified framework. As we will see in this section, choosing only two non-zero fixed values in Algorithm 1 is sufficient to generate a set of linearly independent GPWs. It is then natural to study how the rest of the λ_{ij}s depend on those two values, and the related consequences of Hypothesis 2. These rely on Hypothesis 2 extending the fact that for classical PWs $(i\kappa\cos\theta)^2+(i\kappa\sin\theta)^2=-\kappa^2$ is independent of θ.

3.1 For every GPWs

In Algorithm 1, the number of prescribed coefficients is $\frac{M(M+1)}{2}+Mq$, and the set of coefficients to be prescribed is the set $\{\lambda_{i,j}, 0\leq i\leq M-1, 0\leq j\leq q+M-1-i\}$.

For the sake of simplicity, it is natural to choose most of these values to be zero. Since the unknown $\lambda_{0,0}$ never appears in the non-linear system, there is nothing more natural than setting it to zero: this ensures that any GPW φ will satisfy $\varphi(x_0,y_0)=1$. Concerning the subset of Mq unknowns corresponding to step 1.3 in Algorithm 1, setting these values to zero simply reduces the amount of computation involved in step 1.5 in the algorithm: indeed for $I=0$ then $\sum_{k=0}^{M-1}\mathsf{T}^{\mathfrak{L}}_{I+M+1,I+k+1}\lambda_{I+k,M+\mathfrak{L}-I-k}=0$, while for $0<I<M$ then

$$\sum_{k=0}^{M-1}\mathsf{T}^{\mathfrak{L}}_{I+M+1,I+k+1}\lambda_{I+k,M+\mathfrak{L}-I-k}=\sum_{k=M-\mathfrak{L}}^{M-1}\mathsf{T}^{\mathfrak{L}}_{I+M+1,I+k+1}\lambda_{I+k,M+\mathfrak{L}-I-k}.$$

As for the unknowns $\lambda_{1,0}$ and $\lambda_{0,1}$, they will be non-zero to mimic the classical plane wave case, and their precise choice will be discussed in the next subsection. For the remaining unknowns to be fixed, that is to say the set $\{\lambda_{i,j}, 2\leq i+j\leq M-1\}$, their values are set to zero, here again in order to reduce the amount of computation in computing the right hand side entries $\mathsf{B}^{\mathfrak{L}}_{M+1+I}$ and in applying 1.5.

For the operator $\mathfrak{L}_{2,\gamma}$ the non-linear terms in $N_{1,0}$ and $N_{0,1}$ respectively become with this normalization

$$\begin{aligned}\partial_x[\mathcal{L}^N_{2,\gamma}P](x_0,y_0) &= -2\lambda_{2,0}\lambda_{1,0}+\gamma_{1,1}(x_0,y_0)2\lambda_{2,0}\lambda_{0,1}\\ &\qquad +\partial_x\gamma_{1,1}(x_0,y_0)\lambda_{1,0}\lambda_{0,1}+\partial_x\gamma_{0,2}(x_0,y_0)\lambda_{0,1}^2,\\ \partial_y[\mathcal{L}^N_{2,\gamma}P](x_0,y_0) &= \gamma_{1,1}(x_0,y_0)2\lambda_{1,0}\lambda_{2,0}+\partial_y\gamma_{1,1}(x_0,y_0)\lambda_{1,0}\lambda_{0,1}.\end{aligned}$$

Since all but two of the unknowns to be fixed in Algorithm 1 are set to zero, it is now natural to express the $\lambda_{i,j}$ unknowns computed from 1.5 in the algorithm as functions of the two non-zero prescribed unknowns, $\lambda_{1,0}$ and $\lambda_{0,1}$.

Lemma 3 *Consider a point $(x_0,y_0)\in\mathbb{R}^2$, a given $q\in\mathbb{N}^*$, a given $M\in\mathbb{N}$, $M\geq 2$, a given set of complex-valued functions $\alpha=\{\alpha_{k,l}\in\mathcal{C}^{q-1}$ at $(x_0,y_0), 0\leq k+l\leq M\}$, and the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$. Under Hypothesis 1 consider a solution to Problem (10) constructed thanks to Algorithm 1 with all the prescribed values $\lambda_{i,j}$ such that $i<M$ and $i+j\neq 1$ set to zero. Each $\lambda_{i+M,j}$ can be expressed as an element of $\mathbb{C}[\lambda_{1,0},\lambda_{0,1}]$.*

Proof The fact that $\lambda_{i+M,j}$ can be expressed as a polynomial in two variables with respect to $\lambda_{1,0}$ and $\lambda_{0,1}$ is a direct consequence from the explicit formula in step 1.5 in Algorithm 1 combining with setting $\lambda_{i,j}$ such that $i<M$ and $i+j\neq 1$ to zero. □

Since unknowns are expressed as elements of $\mathbb{C}[\lambda_{1,0},\lambda_{0,1}]$, we will now study the degree of various terms from Algorithm 1 as polynomials with respect to $\lambda_{1,0}$ and $\lambda_{0,1}$. To do so, we will start by inspecting the product terms appearing in Faa di Bruno's formula.

Lemma 4 *Consider a point $(x_0,y_0)\in\mathbb{R}^2$, a given $q\in\mathbb{N}^*$, a given $M\in\mathbb{N}$, $M\geq 2$, a given set of complex-valued functions $\alpha=\{\alpha_{k,l}\in\mathcal{C}^{q-1}$ at $(x_0,y_0), 0\leq k+l\leq M\}$, and the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$. Consider a given polynomial $P\in\mathbb{C}[x,y]$. The non-linear terms $\mathcal{L}^N_{M,\alpha}P$, expressed as linear combinations of products of derivatives of P, namely $\prod_{m=1}^s\left(\partial_x^{i_m}\partial_y^{j_m}P\right)^{k_m}$, contain products of up to M derivatives of P, namely $\partial_x^{i_m}\partial_y^{j_m}P$, counting repetitions. The only products that have exactly M terms are $(\partial_xP)^k(\partial_yP)^{M-k}$ for $0\leq k\leq M$, whereas all the other products have less than M terms.*

Proof Since the operator $\mathcal{L}^N_{M,\alpha}$ is defined via Faa di Bruno's formula, we will proceed by careful examination of the summation and product indices in the latter.

The number of terms in the product term is s, with possible repetitions counted thanks to the k_ms, and the total number of terms counting repetitions is $\mu=\sum_{m=1}^s k_m$. Since in $\mathcal{L}^N_{M,\alpha}$ the indices are such that $1\leq\mu\leq\ell\leq M$, there cannot be more than M terms counting repetitions in any of the $\prod_{m=1}^s\left(\partial_x^{i_m}\partial_y^{j_m}P\right)^{k_m}$.

For $s=1$, in the set $p_1((k,\ell-k),\mu)$, $(i_1,j_1)\in\mathbb{N}^2$ are such that $i_1+j_1\geq 1$ and $k_1\in\mathbb{N}$ is such that $k_1(i_1+j_1)=\ell$. Since $\ell\leq M$, such a term appears in Faa di Bruno's formula as a product of $\mu=M$ terms if and only if $\ell=M$, $k_1=M$, and therefore $i_1+j_1=1$. There are then only two possibilities: either $(i_1,j_1)=(1,0)$ corresponding to the term $(\partial_xP)^M$, or $(i_1,j_1)=(0,1)$ corresponding to the term $(\partial_yP)^M$.

For $s = 2$, in the set $p_2((k, \ell - k), \mu)$, $(i_1, j_1, i_2, j_2) \in \mathbb{N}^4$ are such that $i_1 + j_1 \geq 1$, $i_2 + j_2 \geq 1$, $(i_1, j_1) \prec (i_2, j_2)$, and $(k_1, k_2) \in \mathbb{N}^2$ is such that $\mu = k_1 + k_2$ and $k_1(i_1 + j_1) + k_2(i_2 + j_2) = \ell$. Since $\ell = k_1(i_1 + j_1) + k_2(i_2 + j_2) \geq k_1 + k_2 = \mu$ and $\ell \leq M$ such a term appears in Faa di Bruno's formula as a product of $\mu = M$ terms if and only if $\ell = M$ and $k_1 + k_2 = M$. There are then two possible cases: either $i_2 + j_2 > 1$, then $M = k_1(i_1 + j_1) + k_2(i_2 + j_2) > k_1 + k_2 = M$, so there is no such term in the sum, or $i_2 + j_2 = 1$, then necessarily $(i_1, j_1) = (0, 1)$ and $(i_2, j_2) = (1, 0)$, corresponding to the terms $(\partial_x P)^k (\partial_y P)^{M-k}$ for any k from 0 to M.

For $s \geq 3$, in the set $p_s((k, \ell - k), \mu)$, for all $m \in \mathbb{N}$ such that $1 \leq m \leq s$, $(i_m, j_m) \in \mathbb{N}^2$ and $k_m \in \mathbb{N}$ are such that $i_m + j_m \geq 1$, $\sum_{m=1}^{s} k_m(i_m + j_m) = \ell$, $\mu = \sum_{m=1}^{s} k_m$ and $(i_1, j_1) \prec (i_2, j_2) \prec (i_3, j_3)$. Because of this last condition, it is clear that $i_3 + j_3 > 1$. Since $\ell \leq M$ and $\ell = \sum_{m=1}^{s} k_m(i_m + j_m) \geq \sum_{m=1}^{s} k_m = \mu$, such a term appears in Faa di Bruno's formula as a product of $\mu = M$ terms if and only if $\ell = M$ and $\sum_{m=1}^{s} k_m = M$. But then $M = \sum_{m=1}^{s} k_m(i_m + j_m) > \sum_{m=1}^{s} k_m = M$, so there is no such term in the sum.

The claim is proved. □

Lemma 5 *Consider a point $(x_0, y_0) \in \mathbb{R}^2$, a given $q \in \mathbb{N}^*$, a given $M \in \mathbb{N}$, $M \geq 2$, a given set of complex-valued functions $\alpha = \{\alpha_{k,l} \in \mathcal{C}^{q-1}$ at $(x_0, y_0), 0 \leq k + l \leq M\}$, and the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$. Consider a given polynomial $P \in \mathbb{C}[x, y]$. The quantity $\partial_x^{I_0} \partial_y^{J_0} \mathcal{L}_{M,\alpha}^N P$ is a linear combination of terms $\partial_x^{I_0} \partial_y^{J_0} \left(\prod_{m=1}^{s} \left(\partial_x^{i_m} \partial_y^{j_m} P \right)^{k_m} \right)$, where the indices come from Faa di Bruno's formula. Each of these $\partial_x^{I_0} \partial_y^{J_0} \left(\prod_{m=1}^{s} \left(\partial_x^{i_m} \partial_y^{j_m} P \right)^{k_m} \right)$ can be expressed as a linear combination of products $\prod_{m=1}^{t} (\partial_x^{a_m} \partial_y^{b_m} P)^{c_m}$ where the indices satisfy $\sum_{m=1}^{t} c_m(a_m + b_m) \leq I_0 + J_0 + M$.*

Proof Thanks to the product rule, the derivative $\partial_x^{I_0} \partial_y^{J_0} \left(\prod_{m=1}^{s} \left(\partial_x^{i_m} \partial_y^{j_m} P \right)^{k_m} \right)$ can be expressed as a linear combination of several terms $\prod_{m=1}^{s} \partial_x^{I_m} \partial_y^{J_m} \left[\left(\partial_x^{i_m} \partial_y^{j_m} P \right)^{k_m} \right]$, where $\sum_{m=1}^{t} I_m = I_0$ and $\sum_{m=1}^{t} J_m = J_0$.

We can prove by induction on k that $\partial_x^{I} \partial_y^{J} \left[\left(\partial_x^{i} \partial_y^{j} P \right)^{k} \right]$ can be expressed, for all $(i, j, I, J) \in \mathbb{N}^4$, as a linear combination of products $\prod_{m=1}^{M} (\partial_x^{a_m} \partial_y^{b_m} P)^{c_m}$ where the indices satisfy $\sum_{m=1}^{M} c_m(a_m + b_m) \leq I + J + k(i + j)$:

1. it is evidently true for $k = 1$;

2. suppose that it is true for $k_0 \geq 1$, then for any $(i, j, I, J) \in \mathbb{N}^4$ the product rule applied to $\partial_x^i \partial_y^j P \times \left(\partial_x^i \partial_y^j P\right)^{k_0}$ yields

$$\partial_x^I \partial_y^J \left[\left(\partial_x^i \partial_y^j P\right)^{k_0+1}\right] = \sum_{\tilde{i}=0}^{I} \sum_{\tilde{j}=0}^{J} \binom{I}{\tilde{i}} \binom{J}{\tilde{j}} \partial_x^{i+I-\tilde{i}} \partial_y^{j+J-\tilde{j}} P \partial_x^{\tilde{i}} \partial_y^{\tilde{j}} \left[\left(\partial_x^i \partial_y^j P\right)^{k_0}\right],$$

where by hypothesis each $\partial_x^{\tilde{i}} \partial_y^{\tilde{j}} \left[\left(\partial_x^i \partial_y^j P\right)^{k_0}\right]$ can be expressed as a linear combination of products $\prod_{m=1}^{M} (\partial_x^{a_m} \partial_y^{b_m} P)^{c_m}$ with $\sum_{m=1}^{M} c_m(a_m+b_m) \leq \tilde{i}+\tilde{j}+k_0(i+j)$, so that each term in the double sum can be expressed as a linear combination of products $\prod_{m=1}^{M+1} (\partial_x^{a_m} \partial_y^{b_m} P)^{c_m}$ where $a_{M+1} := i + I - \tilde{i}$, $b_{M+1} := j + J - \tilde{j}$ and $c_{M+1} := 1$, which yields $\sum_{m=1}^{M+1} c_m(a_m + b_m) = \sum_{m=1}^{M} c_m(a_m + b_m) + (i + I - \tilde{i} + j + J - \tilde{j})$ and therefore $\sum_{m=1}^{M+1} c_m(a_m + b_m) \leq k_0(i + j) + (i + I + j + J)$. This concludes the proof by induction.

Finally the derivative $\partial_x^{I_0} \partial_y^{J_0} \left(\prod_{m=1}^{s} \left(\partial_x^{i_m} \partial_y^{j_m} P\right)^{k_m}\right)$ can be expressed as a linear combination of several terms $\prod_{m=1}^{s} \prod_{\tilde{m}=1}^{M} (\partial_x^{a_{\tilde{m}}} \partial_y^{b_{\tilde{m}}} P)^{c_{\tilde{m}}}$, with $\sum_{\tilde{m}=1}^{M} c_{\tilde{m}}(a_{\tilde{m}} + b_{\tilde{m}}) \leq I_m + J_m + k_m(i_m + j_m)$, in other words it can be expressed as a linear combination of several terms $\prod_{m=1}^{Ms} (\partial_x^{a_m} \partial_y^{b_m} P)^{c_m}$, with $\sum_{m=1}^{Ms} c_m(a_m + b_m) \leq \sum_{m=1}^{s} I_m + J_m + k_m(i_m + j_m) = I_0 + J_0 + \sum_{m=1}^{s} k_m(i_m + j_m)$. For any $\partial_x^{I_0} \partial_y^{J_0} \left(\prod_{m=1}^{s} \left(\partial_x^{i_m} \partial_y^{j_m} P\right)^{k_m}\right)$ coming from $\partial_x^{I_0} \partial_y^{J_0} \mathcal{L}_{M,\alpha}^{N} P$, the summation indices from Faa di Bruno's formula satisfy $\sum_{m=1}^{s} k_m(i_m + j_m) = \ell$, so the products $\prod_{m=1}^{Ms} (\partial_x^{a_m} \partial_y^{b_m} P)^{c_m}$ are such that $\sum_{m=1}^{Ms} c_m(a_m + b_m) \leq I_0 + J_0 + M$. □

The two following results now turn to $\lambda_{i+M,j}$ computed in Algorithm 1.

Proposition 2 *Consider a point $(x_0, y_0) \in \mathbb{R}^2$, given $q \in \mathbb{N}^*$ and $M \in \mathbb{N}$, with $M \geq 2$, a set of complex-valued functions $\alpha = \{\alpha_{k,l} \in \mathcal{C}^{q-1}$ at $(x_0, y_0), 0 \leq k + l \leq M\}$, and the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$. Under Hypothesis 1 consider a solution to Problem (10) constructed thanks to Algorithm 1 with all the fixed values $\lambda_{i,j}$ such that $i < M$ and $i + j \neq 1$ set to zero. As an element of $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$, $\lambda_{M,0}$ is of degree equal to M.*

Proof The formula to compute $\lambda_{M,0}$ in Algorithm 1 comes from the $(I, J) = (0, 0)$ equation in System (10), that is to say $\mathcal{L}_{M,\alpha}^{A} P(x_0, y_0) = -\alpha_{0,0}(x_0, y_0)$. It reads

$$\lambda_{M,0} = \frac{1}{\mathsf{T}_{M+1,M+1}^{0}} \left(\mathsf{B}_{M+1}^{0} - \sum_{k=0}^{M-1} \mathsf{T}_{M+1,k+1}^{0} \lambda_{k,M-k}\right),$$

and the sum is actually zero since the $\lambda_{k,M-k}$ unknowns are prescribed to zero for $k < M$. The definitions of B^0 and L^0 then give

$$\lambda_{M,0} = \frac{1}{M!\alpha_{M,0}(x_0, y_0)}\left(-\sum_{\ell=0}^{M-1}\sum_{k=0}^{\ell} k!(\ell-k)!\alpha_{k,\ell-k}(x_0, y_0)\lambda_{k,\ell-k}\right.$$
$$\left. -\mathfrak{L}^N_{M,\alpha}P(x_0, y_0) - \alpha_{0,0}(x_0, y_0)\right).$$

Since the $\lambda_{k,\ell-k}$ unknowns are prescribed to zero for all $1 < \ell < M-1$ and all k, the double sum term reduces to $\alpha_{0,1}(x_0, y_0)\lambda_{0,1} + \alpha_{1,0}(x_0, y_0)\lambda_{1,0}$. The non-linear terms from $\mathfrak{L}^N_{M,\alpha}P$, namely $\prod_{m=1}^{s}(\partial_x^{i_m}\partial_y^{j_m}P)^{k_m}$, are products of at most M terms, counting repetitions, according to Lemma 4. So $\mathfrak{L}^N_{M,\alpha}P(x_0, y_0)$ is a linear combination of product terms reading $\prod_{m=1}^{s}(\lambda_{i_m,j_m})^{k_m}$ with at most M factors. Moreover, since P is constructed thanks to Algorithm 1, from Corollary 1 we know that these λ_{i_m,j_m}s have a length of the multi-index at most equal to $M-1$, so they are either $\lambda_{1,0}$ or $\lambda_{0,1}$ or prescribed to zero. This means that in $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ each one of these λ_{i_m,j_m} is at most of degree one. So in $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ each $\prod_{m=1}^{s}(\lambda_{i_m,j_m})^{k_m}$ is a product of at most M factors each of them of degree at most one, the product is therefore of degree at most M. As a result

$$\lambda_{M,0} = \frac{1}{M!\alpha_{M,0}(x_0, y_0)}\big(-\alpha_{0,1}(x_0, y_0)\lambda_{0,1} - \alpha_{1,0}(x_0, y_0)\lambda_{1,0}$$
$$-\mathfrak{L}^N_{M,\alpha}P(x_0, y_0) - \alpha_{0,0}(x_0, y_0)\big)$$

as an element of $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ is of degree at most M.

Finally, the term $(\partial_x P)^M$ from $\mathfrak{L}^N_{M,\alpha}P$ identified in Lemma 4 corresponds to a term $\alpha_{M,0}(x_0, y_0)(\lambda_{1,0})^M$ in the expression of $\lambda_{M,0}$, and this term is non-zero under Hypothesis 1. As a conclusion $\lambda_{M,0}$ as an element of $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ is of degree equal to M. □

Proposition 3 *Consider a point $(x_0, y_0) \in \mathbb{R}^2$, given $q \in \mathbb{N}^*$ and $M \in \mathbb{N}$, with $M \geq 2$, a given set of complex-valued functions $\alpha = \{\alpha_{k,l} \in \mathcal{C}^{q-1}$ at $(x_0, y_0), 0 \leq k+l \leq M\}$, and the corresponding partial differential operator $\mathfrak{L}_{M,\alpha}$. Under Hypothesis 1 consider a solution to Problem (10) constructed thanks to Algorithm 1 with all the fixed values $\lambda_{i,j}$ such that $i < M$ and $i + j \neq 1$ set to zero. As an element of $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$, each $\lambda_{i+M,j}$ has a total degree at most equal to the length of its multi-index $i+j+M$.*

Proof The formula to compute $\lambda_{I+M,\mathfrak{L}-I}$ in Algorithm 1 comes from the $(I, J) = (I, \mathfrak{L} - I)$ equation in System (10), that is to say $\partial_x^I\partial_y^{\mathfrak{L}-I}\mathfrak{L}^A_{M,\alpha}P(x_0, y_0) = -\partial_x^I\partial_y^{\mathfrak{L}-I}\alpha_{0,0}(x_0, y_0)$. It reads

$$\lambda_{I+M,\mathfrak{L}-I} = \frac{1}{\mathsf{T}^{\mathfrak{L}}_{I+M+1,I+M+1}}\left(\mathsf{B}^{\mathfrak{L}}_{I+M+1} - \sum_{k=0}^{M-1}\mathsf{T}^{\mathfrak{L}}_{I+M+1,I+k+1}\lambda_{I+k,M+\mathfrak{L}-I-k}\right)$$

$$= \frac{I!}{(M+I)!\alpha_{M,0}(x_0, y_0)} \left(\mathsf{N}_{I,\mathfrak{L}-I} - \sum_{k=0}^{M-1} \frac{(I+k)!(M-k+\mathfrak{L}-1)!}{I!(\mathfrak{L}-I)!} \alpha_{k,M-k}(x_0, y_0)\lambda_{I+k,M+\mathfrak{L}-I-k} \right). \tag{24}$$

We will proceed by induction on $\mathfrak{L}$:

1. the result has been proved to be true for $\mathfrak{L} = 0$ in Proposition 2;
2. suppose the result is true for $\mathfrak{L} \in \mathbb{N}$ as well as for all $\tilde{\mathfrak{L}} \in \mathbb{N}$ such that $\tilde{\mathfrak{L}} \leq \mathfrak{L}$, then all the linear terms in $N_{I,\mathfrak{L}+1-I}$ have a length of the multi-index at most equal to $M + \mathfrak{L}$ so by hypothesis their degree as elements of $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ is at most equal to $M + \mathfrak{L}$, and thanks to Lemma 5 all the non-linear terms in $N_{I,\mathfrak{L}+1-I}$ can be expressed as a linear combination of products $\prod_{m=1}^{t}(\lambda_{a_m,b_m})^{c_m}$ where the indices satisfy $\sum_{m=1}^{t} c_m(a_m + b_m) \leq \mathfrak{L} + 1 + M$ so by hypothesis their degree as elements of $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ is at most equal to $M + \mathfrak{L} + 1$; the last step is to prove that the $\lambda_{I+k,M+\mathfrak{L}+1-I-k}$ are also of degree at most equal to $M + \mathfrak{L} + 1$, and we will proceed by induction on I:
    (a) for $I = 0$, all $\lambda_{I+k,M+\mathfrak{L}+1-I-k}$ for $0 \leq k \leq M-1$ satisfy the two conditions $I + k < M$ and $I + k + M + \mathfrak{L} + 1 - I - k = M + \mathfrak{L} + 1 \neq 1$ so they are all prescribed to zero and their degree as element of $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ is at most equal to $M + \mathfrak{L} + 1$ that;
    (b) suppose that, for a given $I \in \mathbb{N}$, the $\lambda_{\tilde{I}+k,M+\mathfrak{L}+1-\tilde{I}-k}$ for all $\tilde{I} \in \mathbb{N}$ such that $\tilde{I} \leq I$ are also of degree at most equal to $M + \mathfrak{L} + 1$ then it is clear from Eq. (24) that $\lambda_{I+1+M,\mathfrak{L}-I-1}$ is also of degree at most equal to $M + \mathfrak{L} + 1$.

   This concludes the proof. □

As explained from an algebraic viewpoint in Sect. 3.2 in [15], the degree of $\lambda_{i+M,j}$ as an element of $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ will be affected by the choice of the last two prescribed values, namely $\lambda_{1,0}$ and $\lambda_{0,1}$. Indeed if $\lambda_{1,0}$ and $\lambda_{0,1}$ satisfy a polynomial identity $P_l(\lambda_{1,0}, \lambda_{0,1}) = 0$, then we can consider the quotient ring $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]/(P_l)$.

Note that choosing to set $\{\lambda_{i,j}, 1 < i + j \leq M - 1\}$ to values different from zero may be useful to treat operators that do not satisfy Hypothesis 2 but this is not our goal here.

3.2 For each GPW

In order to obtain a set of linearly independent GPWs, the values of $\lambda_{1,0}$ and $\lambda_{0,1}$ will be chosen different for each GPW. However the values of $\lambda_{1,0}$ and $\lambda_{0,1}$ will satisfy a common property for every GPWs. Very much as the coefficients of any plane wave of wavenumber κ satisfy $(\lambda_{1,0})^2 + (\lambda_{0,1})^2 = -\kappa^2$, independently of the direction of propagation θ since $\lambda_{1,0} = \imath\kappa\cos\theta$ and $\lambda_{0,1} = \imath\kappa\sin\theta$, under Hypothesis 2 the coefficients of each GPW will be chosen for the quantity

$$\sum_{k=0}^{M} \alpha_{k,M-k}(x_0, y_0)(\lambda_{1,0})^k(\lambda_{0,1})^{M-k} = \left(\begin{pmatrix}\lambda_{1,0}\\ \lambda_{0,1}\end{pmatrix}^t \Gamma \begin{pmatrix}\lambda_{1,0}\\ \lambda_{0,1}\end{pmatrix}\right)^{\frac{M}{2}}$$

to be identical for every GPWs, as we will see in the following proposition and theorem.

This will be crucial to prove interpolation properties of the corresponding set of functions, which will result from the consequence of this common property on the degree of each $\lambda_{i+M,j}$ as an element of $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$. As the plane wave case suggests, we will see that $\lambda_{i+M,j}$ can be expressed as a polynomial of lower degree thanks to a judicious choice for $\lambda_{1,0}$ and $\lambda_{0,1}$.

We first need an intermediate result concerning the polynomial $\mathcal{L}^N_{M,\alpha}P$.

Lemma 6 *Consider a point* $(x_0, y_0) \in \mathbb{R}^2$*, a given* $q \in \mathbb{N}^*$*, a given* $M \in \mathbb{N}$*,* $M \geq 2$*, a given set of complex-valued functions* $\alpha = \{\alpha_{k,l} \in \mathcal{C}^{q-1}$ *at* $(x_0, y_0), 0 \leq k + l \leq M\}$*, and the corresponding partial differential operator* $\mathcal{L}_{M,\alpha}$*. Consider a given polynomial* $P \in \mathbb{C}[x, y]$*. For any* $\mathfrak{L} \in \mathbb{N}$ *and any* $I \in \mathbb{N}$ *such that* $I \leq \mathfrak{L} + 1$*, the quantity* $\partial_x^I \partial_y^{\mathfrak{L}+1-I}\left[\mathcal{L}^N_{M,\alpha}P\right]$ *can be expressed as a linear combination of products* $\prod_{t=1}^{\mu} \partial_x^{i_t+I_t}\partial_y^{j_t+J_t}P$*, with* $\sum_{t=1}^{\mu} I_t = I$*,* $\sum_{t=1}^{\mu} J_t = \mathfrak{L} + 1 - I$*,* $\sum_{t=1}^{\mu} i_t = k$*, and* $\sum_{t=1}^{\mu} j_t = \ell - k$*. Moreover, for each product term, there exists* $t_0 \in \mathbb{N}$*,* $1 \leq t_0 \leq \mu$ *such that* $I_{t_0} \neq 0$ *or* $J_{t_0} \neq 0$*.*

Proof The quantity $\mathcal{L}^N_{M,\alpha}P$ can be expressed, from Faa di Bruno's formula, as a linear combination of products $\prod_{m=1}^{s}\left(\partial_x^{i_m}\partial_y^{j_m}P\right)^{k_m}$, with $(i_{m_1}, j_{m_1}) \neq (i_{m_2}, j_{m_2})$ for all $m_1 \neq m_2$, $\sum_{m=1}^{s} k_m = \mu$, $\sum_{m=1}^{s} k_m i_m = k$, and $\sum_{m=1}^{s} k_m j_m = \ell - k$. Therefore $\mathcal{L}^N_{M,\alpha}P$ can also be expressed, repeating terms, as a linear combination of products $\prod_{t=1}^{\mu} \partial_x^{i_t}\partial_y^{j_t}P$, with possibly $(i_{m_1}, j_{m_1}) = (i_{m_2}, j_{m_2})$ for $m_1 \neq m_2$, $\sum_{t=1}^{\mu} i_t = k$, and $\sum_{t=1}^{\mu} j_t = \ell - k$. So the quantity $\partial_x^I \partial_y^{\mathfrak{L}+1-I}\left[\mathcal{L}^N_{M,\alpha}P\right]$ can be expressed, from Leibniz's rule, as a linear combination of products $\prod_{t=1}^{\mu} \partial_x^{i_t+I_t}\partial_y^{j_t+J_t}P$, with $\sum_{t=1}^{\mu} I_t = I$ and $\sum_{t=1}^{\mu} J_t = \mathfrak{L} + 1 - I$.

Consider such a given product term $\prod_{t=1}^{\mu} \partial_x^{i_t+I_t}\partial_y^{j_t+J_t}P$, and suppose that for all t $I_t = J_t = 0$. Then $I = \sum_{t=1}^{\mu} I_t = 0$ and $\mathfrak{L} + 1 - I = \sum_{t=1}^{\mu} J_t = 0$, which is impossible since $\mathfrak{L} + 1 > 0$. □

The two following results gather the consequences of this choice on $\lambda_{i+M,j}$s computed in Algorithm 1.

Proposition 4 *Consider a point* $(x_0, y_0) \in \mathbb{R}^2$*, given* $q \in \mathbb{N}^*$ *and* $M \in \mathbb{N}$*, with* $M \geq 2$*, a given set of complex-valued functions* $\alpha = \{\alpha_{k,l} \in \mathcal{C}^{q-1}$ *at* $(x_0, y_0), 0 \leq k + l \leq M\}$*, and the corresponding partial differential operator* $\mathcal{L}_{M,\alpha}$*. Under Hypotheses 1 and 2 consider a solution to Problem (10) constructed thanks to Algorithm 1 with all the prescribed values* $\lambda_{i,j}$ *such that* $i < M$ *and* $i + j \neq 1$ *set to zero, and*

$$\begin{pmatrix}\lambda_{1,0}\\ \lambda_{0,1}\end{pmatrix} = \mathrm{i}\kappa A^{-1} D^{-1/2} \begin{pmatrix}\cos\theta\\ \sin\theta\end{pmatrix} \tag{25}$$

for some $\theta \in \mathbb{R}$ and $\kappa \in \mathbb{C}^*$. As an element of $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$, $\lambda_{M,0}$ can be expressed as a polynomial of degree at most equal to $M-1$, and its coefficients are independent of θ.

Note that once we impose this condition on $\lambda_{1,0}, \lambda_{0,1}$ any element of $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ can be expressed by different polynomials, possibly with different degrees, simply because under Hypothesis 2 and (25) we have

$$\sum_{k=0}^{M} \alpha_{k,M-k}(x_0, y_0)\lambda_{1,0}^{k}\lambda_{0,1}^{M-k} = \left(-\kappa^2\right)^{\frac{M}{2}}.$$

See paragraph 3.2 in [15] for an algebraic view point on this comment.

Proof Since

$$\lambda_{M,0} = \frac{1}{M!\alpha_{M,0}(x_0, y_0)}\Big(-\alpha_{0,1}(x_0, y_0)\lambda_{0,1} - \alpha_{1,0}(x_0, y_0)\lambda_{1,0} \\ - \mathcal{L}_{M,\alpha}^{N} P(x_0, y_0) - \alpha_{0,0}(x_0, y_0)\Big), \tag{26}$$

again the term to investigate is $\mathcal{L}_{M,\alpha}^{N} P(x_0, y_0)$. Lemma 4 identifies products of M terms in $\mathcal{L}_{M,\alpha}^{N} P$, and from the definition of $\mathcal{L}_{M,\alpha}^{N}$ they appear in the following linear combination

$$\sum_{k=0}^{M} k!(M-k)!\alpha_{k,M-k}\frac{(\partial_x P)^k}{k!}\frac{(\partial_y P)^{M-k}}{(M-k)!} = \sum_{k=0}^{M} \alpha_{k,M-k}(\partial_x P)^k(\partial_y P)^{M-k}.$$

Back to the expression of $\lambda_{M,0}$, and thanks to Hypothesis 2, the only possible terms of degree M therefore appear in the following linear combination:

$$\begin{aligned}
&\sum_{k=0}^{M} \alpha_{k,M-k}(x_0, y_0)(\lambda_{1,0})^k(\lambda_{0,1})^{M-k} \\
&= \left((\lambda_{1,0}\ \lambda_{0,1})\Gamma\begin{pmatrix}\lambda_{1,0}\\ \lambda_{0,1}\end{pmatrix}\right)^{\frac{M}{2}} = \left((\mathrm{i}\kappa)^2(\lambda_{1,0}\ \lambda_{0,1})A^t DA\begin{pmatrix}\lambda_{1,0}\\ \lambda_{0,1}\end{pmatrix}\right)^{\frac{M}{2}} \\
&= \left(-\kappa^2(\cos\theta\ \ \sin\theta)\begin{pmatrix}\cos\theta\\ \sin\theta\end{pmatrix}\right)^{\frac{M}{2}} = (-\kappa)^M
\end{aligned}$$

Finally thanks to (25), the only terms of degree M in (26) can be expressed as a polynomial of degree at most equal $M-1$. □

Proposition 5 *Consider a point $(x_0, y_0) \in \mathbb{R}^2$, given $q \in \mathbb{N}^*$ and $M \in \mathbb{N}$, with $M \geq 2$, a given set of complex-valued functions $\alpha = \{\alpha_{k,l} \in \mathcal{C}^{q-1}$ at $(x_0, y_0), 0 \leq k+l \leq M\}$, and the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$. Under Hypotheses 1 and*

2 *consider a solution to Problem* (10) *constructed thanks to Algorithm* 1 *with all the fixed values* $\lambda_{i,j}$ *such that* $i < M$ *and* $i + j \neq 1$ *set to zero, and*

$$\begin{pmatrix} \lambda_{1,0} \\ \lambda_{0,1} \end{pmatrix} = \mathrm{i}\kappa A^{-1} D^{-1/2} \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$$

for some $\theta \in \mathbb{R}$ *and* $\kappa \in \mathbb{C}^*$. *As an element of* $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$, *each* $\lambda_{i+M,j}$ *can be expressed as a polynomial of degree at most equal to* $i + j + M - 1$, *and its coefficients are independent of* θ.

Proof From Algorithm 1 the expression of $\lambda_{I+M,\mathfrak{L}-I}$ reads

$$\begin{aligned}
\lambda_{I+M,\mathfrak{L}-I} &= \frac{1}{\mathsf{T}^{\mathfrak{L}}_{I+M+1,I+M+1}} \left(\mathsf{B}^{\mathfrak{L}}_{I+M+1} - \sum_{k=0}^{M-1} \mathsf{T}^{\mathfrak{L}}_{I+M+1,I+k+1} \lambda_{I+k,M+\mathfrak{L}-I-k} \right) \\
&= \frac{I!}{(M+I)!\alpha_{M,0}(x_0,y_0)} \Big(\mathsf{N}_{I,\mathfrak{L}-I} \\
&\qquad - \sum_{k=0}^{M-1} \frac{(I+k)!(M-k+\mathfrak{L}-1)!}{I!(\mathfrak{L}-I)!} \alpha_{k,M-k}(x_0,y_0) \lambda_{I+k,M+\mathfrak{L}-I-k} \Big).
\end{aligned} \tag{27}$$

We will proceed again by induction on $\mathfrak{L}$:

1. the result has been proved to be true for $\mathfrak{L} = 0$ in Proposition 4;
2. suppose the result is true for $\mathfrak{L} \in \mathbb{N}$ as well as for all $\tilde{\mathfrak{L}} \in \mathbb{N}$ such that $\tilde{\mathfrak{L}} \leq \mathfrak{L}$, then we focus on $N_{I,\mathfrak{L}+1-I}$, given by

$$\begin{aligned}
N_{0,\mathfrak{L}+1} = &\sum_{k=0}^{M} \sum_{\tilde{j}=0}^{\mathfrak{L}} \left(k+\tilde{i}\right)! \frac{\left(M-k+\tilde{j}\right)!}{\tilde{j}!} \mathcal{D}^{(0,\mathfrak{L}+1-\tilde{j})} \alpha_{k,M-k}(x_0,y_0) \lambda_{k,M-k+\tilde{j}} \\
&+ \sum_{\ell=1}^{M-1} \sum_{k=0}^{\ell} \sum_{\tilde{j}=0}^{\mathfrak{L}+1} (k)! \frac{\left(\ell-k+\tilde{j}\right)!}{\tilde{j}!} \mathcal{D}^{(0,\mathfrak{L}+1-\tilde{j})} \alpha_{k,\ell-k}(x_0,y_0) \lambda_{k,\ell-k+\tilde{j}} \\
&- \mathcal{D}^{(0,\mathfrak{L}+1)} \left[\mathcal{L}^{N}_{M,\alpha} P \right](x_0,y_0) - \mathcal{D}^{(0,\mathfrak{L}+1)} \alpha_{0,0}(x_0,y_0) \text{ for } I = 0; \text{ and}
\end{aligned}$$

$$\begin{aligned}
&N_{I,\mathfrak{L}+1-I} \\
&= -\sum_{k=0}^{M} \sum_{\tilde{i}=0}^{I-1} \sum_{\tilde{j}=0}^{\mathfrak{L}-I} \frac{\left(k+\tilde{i}\right)! \left(M-k+\tilde{j}\right)!}{\tilde{i}!\tilde{j}!} \mathcal{D}^{(I-\tilde{i},\mathfrak{L}+1-I-\tilde{j})} \alpha_{k,M-k}(x_0,y_0) \lambda_{k+\tilde{i},M-k+\tilde{j}} \\
&- \sum_{\ell=1}^{M-1} \sum_{k=0}^{\ell} \sum_{\tilde{i}=0}^{I} \sum_{\tilde{j}=0}^{\mathfrak{L}+1-I} \frac{\left(k+\tilde{i}\right)! \left(\ell-k+\tilde{j}\right)!}{\tilde{i}!\tilde{j}!} \mathcal{D}^{(I-\tilde{i},\mathfrak{L}+1-I-\tilde{j})} \alpha_{k,\ell-k}(x_0,y_0) \lambda_{k+\tilde{i},\ell-k+\tilde{j}} \\
&- \mathcal{D}^{(I,\mathfrak{L}+1-I)} \left[\mathcal{L}^{N}_{M,\alpha} P \right](x_0,y_0) - \mathcal{D}^{(I,\mathfrak{L}+1-I)} \alpha_{0,0}(x_0,y_0) \text{ otherwise;}
\end{aligned}$$

all the linear terms in $N_{I,\mathfrak{L}+1-I}$, as elements of $\mathbb{C}[\lambda_{1,0},\lambda_{0,1}]$, by hypothesis have degree at most equal to $(I+M)+(\mathfrak{L}+1-I)-1 = M+\mathfrak{L}$, and thanks to Lemma 6 all the non-linear terms in $N_{I,\mathfrak{L}+1-I}$ can be expressed as a linear combination of products $\prod_{t=1}^{\mu}\lambda_{a_t,b_t}$ where the indices satisfy $\sum_{t=1}^{\mu}(a_t+b_t) \leq \mathfrak{L}+1+M$; in each such product, as element of $\mathbb{C}[\lambda_{1,0},\lambda_{0,1}]$, each λ_{a_t,b_t} is either of degree $a_t+b_t = 1$ if $(a_t,b_t)\in\{(0,1),(1,0)\}$, or of degree at most equal to a_t+b_t-1 otherwise by hypothesis; from Lemma 6 there is at least one t_0 such that $(a_{t_0},b_{t_0})\notin\{(0,1),(1,0)\}$, therefore each product $\prod_{t=1}^{\mu}\lambda_{a_t,b_t}$, as element of $\mathbb{C}[\lambda_{1,0},\lambda_{0,1}]$, can be expressed as a polynomial of degree at most $\left(\sum_{t=1}^{\mu}(a_t+b_t)\right)-1\leq\mathfrak{L}+M$; so all terms in $N_{I,\mathfrak{L}+1-I}$, as elements of $\mathbb{C}[\lambda_{1,0},\lambda_{0,1}]$, have degree at most equal to $M+\mathfrak{L}$; the last step is to prove that the $\lambda_{I+k,M+\mathfrak{L}+1-I-k}$ are also of degree at most equal to $M+\mathfrak{L}$, and we will proceed by induction on I:

(a) for $I=0$, all $\lambda_{I+k,M+\mathfrak{L}+1-I-k}$ for $0\leq k\leq M-1$ satisfy the two conditions $I+k<M$ and $I+k+M+\mathfrak{L}+1-I-k = M+\mathfrak{L}+1\neq 1$ so they are all prescribed to zero and their degree as element of $\mathbb{C}[\lambda_{1,0},\lambda_{0,1}]$ is at most equal to $M+\mathfrak{L}$ that;

(b) suppose that, for a given $I\in\mathbb{N}$, the $\lambda_{\tilde{I}+k,M+\mathfrak{L}+1-\tilde{I}-k}$ for all $\tilde{I}\in\mathbb{N}$ such that $\tilde{I}\leq I$ are also of degree at most equal to $M+\mathfrak{L}$ then it is clear from Eq. (27) that $\lambda_{I+1+M,\mathfrak{L}-I-1}$ is also of degree at most equal to $M+\mathfrak{L}$.

This concludes the proof. □

Finally, since we are interested in the local approximation properties of GPWs, it is natural to study their Taylor expansion coefficients, and how they can be expressed as elements of $\mathbb{C}[\lambda_{1,0},\lambda_{0,1}]$. In particular we will find what is the link between the Taylor expansion coefficients of a GPW, $\partial_x^i\partial_y^j\varphi(x_0,y_0)/(i!j!)$, and that of the corresponding PW, $(\lambda_{0,1})^j(\lambda_{1,0})^i/(i!j!)$.

Proposition 6 *Consider a point $(x_0,y_0)\in\mathbb{R}^2$, given $q\in\mathbb{N}^*$ and $M\in\mathbb{N}$, with $M\geq 2$, a given set of complex-valued functions $\alpha=\{\alpha_{k,l}\in\mathcal{C}^{q-1}$ at $(x_0,y_0), 0\leq k+l\leq M\}$, and the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$. Under Hypotheses 1 and 2 consider a solution to Problem (10) constructed thanks to Algorithm 1 with all the fixed values $\lambda_{i,j}$ such that $i<M$ and $i+j\neq 1$ set to zero, and*

$$\begin{pmatrix}\lambda_{1,0}\\ \lambda_{0,1}\end{pmatrix} = \mathrm{i}\kappa A^{-1}D^{-1/2}\begin{pmatrix}\cos\theta\\ \sin\theta\end{pmatrix},$$

for some $\theta\in\mathbb{R}$ and $\kappa\in\mathbb{C}^*$, and the corresponding $\varphi(x,y)=\exp\sum_{0\leq i+j\leq q+1}\lambda_{ij}(x-x_0)^i(y-y_0)^j$. Then for all $(i,j)\in\mathbb{N}^2$ such that $i+j\leq q+1$ the difference

$$R_{i,j} := \partial_x^i\partial_y^j\varphi(x_0,y_0)-(\lambda_{0,1})^j(\lambda_{1,0})^i \tag{28}$$

can be expressed as an element of $\mathbb{C}[\lambda_{1,0},\lambda_{0,1}]$ such that

- *its total degree satisfies $\mathrm{d}R_{i,j}\leq i+j-1$,*

- *its coefficients only depend on i, j, and on the derivatives of the PDE coefficients α evaluated at (x_0, y_0) but do not depend on θ.*

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

$$\partial_x^i \partial_y^j \varphi\,(x_0, y_0) = i!j! \sum_{\mu=1}^{i+j} \sum_{s=1}^{i+j} \sum_{p_s((i,j),\mu)} \prod_{l=1}^{s} \frac{(\lambda_{i_l, j_l})^{k_l}}{k_l!},$$

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

$$\Bigg\{ (k_l, (i_l, j_l))_{l \in [\![1,s]\!]} : k_l \in \mathbb{N}^*, 0 \prec (i_1, j_1) \\ \prec \cdots \prec (i_l, j_l), \sum_{l=1}^{s} k_l = \mu, \sum_{l=1}^{s} k_l (i_l, j_l) = (i, j) \Bigg\}.$$

For each partition $(k_l, (i_l, j_l))_{l \in [\![1,s]\!]}$ of (i, j), the corresponding product term, considered as an element of $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$, has degree $Deg\ \prod_{l=1}^{s} (\lambda_{i_l, j_l})^{k_l} = \sum_{l=1}^{s} k_l Deg\ \lambda_{i_l, j_l}$. Combining Proposition 5 and the fact that $\lambda_{i,j} = 0$ for all (i, j) such that $1 < i + j < M$, we can conclude that this degree is at most equal to

$$\sum_{i_l=0, j_l=1} k_l j_l + \sum_{i_l=1, j_l=0} k_l i_l + \sum_{1<i_l+j_l<M} k_l \cdot 0 + \sum_{i_l+j_l \geq M} k_l (i_l + j_l - 1). \quad (29)$$

The partition with two terms $(i, j) = j(0, 1) + i(1, 0)$ corresponds to the term $(\lambda_{0,1})^j (\lambda_{1,0})^i$, which is the leading term in $\partial_x^i \partial_y^j \varphi\,(x_0, y_0)$. Indeed, any other partition will include at least one term such that $i_l + j_l > 1$, and the degree corresponding to this term within the product is either $k_l \cdot 0$ or $k_l (i_l + j_l - 1)$, and in both case it is at most equal to $k_l (i_l + j_l) - 1$. As a result, the degree of the product term in (29) is necessarily less than $\sum_{l=1}^{s} k_l (i_l + j_l) = i + j$. So $R_{i,j}$, which is defined as the difference between $\partial_x^i \partial_y^j \varphi\,(x_0, y_0)$ and its leading term $(\lambda_{0,1})^j (\lambda_{1,0})^i$, is as expected of degree less than $i + j$.

Finally, the coefficients of $R_{i,j}$ share the same property as the coefficients of λ_{ij}s from Propositions 5. □

Remark 3 As mentioned in Remark 2, under the hypothesis $\alpha_{0,M}(x_0, y_0) \neq 0$, an algorithm very similar to Algorithm 1 would construct the polynomial coefficients of a GPW, fixing the values of $\{\lambda_{i,j}, 0 \leq j \leq M - 1, 0 \leq i \leq q + M - 1 - j\}$. The corresponding version of Proposition 6 could then be proved essentially by exchanging the roles of i and j in all the proofs.

3.3 Local set of GPWs

At this point for a given value of $\theta \in \mathbb{R}$ we can construct a GPW as a function $\varphi = \exp P$ where the polynomial P is a solution to Problem (10) constructed thanks to Algorithm 1 with all the fixed values $\lambda_{i,j}$ such that $i < M$ and $i + j \neq 1$ set to zero, and

$$\begin{pmatrix}\lambda_{1,0}\\ \lambda_{0,1}\end{pmatrix} = \mathrm{i}\kappa A^{-1} D^{-1/2} \begin{pmatrix}\cos\theta\\ \sin\theta\end{pmatrix}.$$

This parameter θ is then equivalent to the direction a classical plane wave, while $|\kappa|$ is equivalent to the wave number of a classical plane wave, and θ will now be used to construct a set of GPWs. Under Hypotheses 1 and 2 , by choosing p different angles $\{\theta_l, l \in \mathbb{N}^*, l \leq p\} \in \mathbb{R}^p$, we can consider p solutions to Problem (10) to construct p GPWs.

Definition 8 Consider a point $(x_0, y_0) \in \mathbb{R}^2$, a given $q \in \mathbb{N}^*$, a given $M \in \mathbb{N}, M \geq 2$, a given set of complex-valued functions $\alpha = \{\alpha_{k,l} \in \mathcal{C}^{q-1}$ at $(x_0, y_0), 0 \leq k+l \leq M\}$, and the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$. Let $p \in \mathbb{N}^*$ be the number of desired basis functions. Under Hypotheses 1 and 2 , consider the normalization $\lambda_{i,j}$ such that $i < M$ and $i + j \neq 1$ set to zero, and

$$\begin{pmatrix}\lambda^l_{1,0}\\ \lambda^l_{0,1}\end{pmatrix} = \kappa A^{-1} D^{-1/2} \begin{pmatrix}\cos\theta_l\\ \sin\theta_l\end{pmatrix}, \text{ for } \{\theta_l \in [0, 2\pi),$$
$$\forall l \in \mathbb{N}^*, l \leq p, \theta_{l_1} \neq \theta_{l_2}\ \forall l_1 \neq l_2, \kappa \in \mathbb{C}^*\}.$$

The set of corresponding GPWs constructed from Algorithm 1 will be denoted hereafter by

$$\mathbb{V}^0_{\alpha,p,q} = \{\varphi_l := \exp P_l, \forall l \in \mathbb{N}^*, l \leq p\}.$$

4 Interpolation properties

This section is restricted to operators of order $M = 2$.

We now have built tools to turn to the interpolation properties of GPWs. In particular, since the GPWs are constructed locally, and will be defined separately on each mesh element, we focus on local interpolation properties. Given a partial differential operator $\mathcal{L}$, a point $(x_0, y_0) \in \mathbb{R}^2$ and an integer $n \in \mathbb{N}$, the question is whether we can find a finite dimensional space $\mathbb{V}_h \subset \mathcal{C}^\infty$, with the following property:

$$\begin{aligned}&\forall u \text{ satisfying } \mathcal{L}u = 0, \exists u_a \in \mathbb{V}_h \text{ s. t.}\\ &\quad \forall (x, y) \in \mathbb{R}^2, |u(x, y) - u_a(x, y)| \leq C\|(x, y) - (x_0, y_0)\|^{n+1},\end{aligned} \tag{30}$$

that is to say there exists an element of $\mathbb{V}_h$ whose Taylor expansion at (x_0, y_0) matches the Taylor expansion of u at (x_0, y_0) up to order n, for any solution u of the PDE

$\mathcal{L}u = 0$. If $\{f_i, i \in \mathbb{N}^*, i \leq p\}$ is a basis of $\mathbb{V}_h$, this can be expressed in terms of linear algebra. Consider the vector space $\mathbb{F}$ and the matrix $\mathsf{M} \in \mathbb{C}^{(n+1)(n+2)/2\times p}$ defined as follows:

$$\begin{aligned}\mathbb{F} := \Big\{ &\mathsf{F} \in \mathbb{C}^{(n+1)(n+2)/2}, \exists u \text{ satisfying } \mathcal{L}u = 0 \text{ s.t.}\\ &\mathsf{F}_{\frac{(k_1+k_2)(k_1+k_2+1)}{2}+k+2+1} = \partial_x^{k_1}\partial_y^{k_2} u(x_0, y_0)/(k_1!k_2!)\Big\},\\ &\mathsf{M}_{\frac{(k_1+k_2)(k_1+k_2+1)}{2}+k_2+1,i} := \partial_x^{k_1}\partial_y^{k_2} f_i(x_0, y_0)/(k_1!k_2!).\end{aligned} \tag{31}$$

Then (30) is equivalent to

$$\forall \mathsf{F} \in \mathbb{F}, \exists \mathsf{X} \in \mathbb{C}^p \text{ s.t. } \mathsf{M}\mathsf{X} = \mathsf{F}, \tag{32}$$

and the choice of p, the number of basis functions, will be crucial to our study.

Our previous work on GPWs was focused on the Helmholtz equation, i.e. corresponding to the operator $\mathcal{L} = -\Delta + \beta(x, y)$, and in that case the classical PWs are exact solutions to the PDE if the coefficient is constant $\beta(x, y) = -\kappa^2$. However, even though the proof of the interpolation properties of GPWs relies strongly on that of classical PWs, it is not required, in order to obtain the GPW result, for classical PW to be solutions of the constant coefficient equation [15]. Indeed, what will be central to the proof that follows is the rank of the matrix M associated to a set of reference functions—not necessarily classical PWs—that are not required to satisfy any PDE. For the Helmholtz equation, the reference functions used in [15] were classical PWs if $\beta(x_0, y_0) < 0$ and real exponentials if $\beta(x_0, y_0) > 0$, and the structure of the proof provides useful guidelines for what follows.

4.1 Comments on a standard reference case

Interpolation properties of classical plane waves were already presented for instance in [15], and in [5], however the link between desired order of approximation n and number p of basis functions was simply provided as $p = 2n + 1$. We present here a new perspective, focusing on properties of trigonometric functions, to justify this choice. The corresponding set of trigonometric functions will constitute the reference case at the heart of the GPWs interpolation properties.

Definition 9 Consider a given $n \in \mathbb{N}^*$ and a given $p \in \mathbb{N}^*$. Considering for some $\kappa \in \mathbb{R}^*$ a space $\mathbb{V}_h^\kappa = Span\{\exp \mathrm{i}\kappa(\cos\theta_l(x - x_0) + \sin\theta_l(y - y_0)), 1 \leq l \leq p, \theta_l \in [0, 2\pi), \theta_{l_1} \neq \theta_{l_2}\ \forall l_1 \neq l_2\}$ of classical PWs, we define the corresponding matrix (31) for the plane wave functions spanning $\mathbb{V}_h^\kappa$, denoted M^C, as well as the reference matrix M^R, by

$$\forall (k_1, k_2) \in \mathbb{N}^2, k_1 + k_2 \leq n, \begin{cases} \left(\mathsf{M}_n^C\right)_{\frac{(k_1+k_2)(k_1+k_2+1)}{2}+k_2+1,l} := (\mathrm{i}\kappa)^{k_1+k_2}(\cos\theta_l)^{k_1}(\sin\theta_l)^{k_2}/(k_1!k_2!),\\ \left(\mathsf{M}_n^R\right)_{\frac{(k_1+k_2)(k_1+k_2+1)}{2}+k_2+1,l} := (\cos\theta_l)^{k_1}(\sin\theta_l)^{k_2}/(k_1!k_2!).\end{cases}$$

If we denote by $\mathsf{D}_n^{RC} = diag(d_k^{RC}, k \text{ from } 1 \text{ to } n+1)$ the block diagonal matrix with blocks of increasing size $d_k^{RC} = (\mathrm{i}\kappa)^{k-1} I_k \in \mathbb{C}^{k\times k}$, it is evident that $\mathsf{M}_n^C = \mathsf{D}_n^{RC}\mathsf{M}_n^R$, therefore trigonometric functions are closely related to interpolation properties of PWs.

Consider the two sets of functions

$$\begin{aligned} \mathcal{F}_n &= \{\theta \mapsto \cos^k\theta \sin^{K-k}\theta/(k!(K-k)!), 0 \leq k \leq K \leq n\}, \\ \text{and} \quad \mathcal{G}_n &= \{\theta \mapsto \exp ik\theta, -n \leq k \leq n\}. \end{aligned}$$

The first one, $\mathcal{F}_n$, is a set of $(n+1)(n+2)/2$ functions. The second one, $\mathcal{G}_n$, is a set of $2n+1$ linearly independent functions: indeed, any null linear combination of these functions $\sum_{-n\leq k\leq n} \nu_k \exp(ik\theta)$ would define a function $f(x) = \sum_{-n\leq k\leq n} \nu_k x^k$ that would be uniformly null on the circle $|x| = 1$, implying that the polynomial $x^n.f(x)$ has an infinite number of roots; hence all its coefficients ν_k are null. Moreover since

$$\begin{cases} \cos(\theta)^k \sin(\theta)^{K-k} = \left(\dfrac{e^{i\theta}+e^{-i\theta}}{2}\right)^k \left(\dfrac{e^{i\theta}-e^{-i\theta}}{2i}\right)^{K-k} = \dfrac{1}{2^K i^{K-k}} \displaystyle\sum_{l=0}^{k}\sum_{L=0}^{K-k} \binom{k}{l}\binom{K-k}{L} e^{i(2l+2L-K)\theta}, \\ \qquad\qquad \text{with } -K \leq 2l+2L-K \leq K \Rightarrow \mathcal{F}_n \subset Span\ \mathcal{G}_n, \\ \exp \pm ik\theta = \displaystyle\sum_{s=0}^{k} \binom{k}{s} (\pm i)^s (\cos\theta)^{k-s} (\sin\theta)^s \Rightarrow \mathcal{G}_n \subset Span\ \mathcal{F}_n, \end{cases}$$

we then have that $Span\ \mathcal{F}_n = Span\ \mathcal{G}_n$, and in particular the space spanned by $\mathcal{F}_n$ is of dimension $2n+1$.

Consider any matrix $\mathsf{A}^{\mathcal{F}} \in \mathbb{C}^{(n+1)(n+2)/2\times N_p}$ defined for some $\{\theta_l\}_{1\leq l\leq N_p} \in (\mathbb{R})^{N_p}$, with $N_p > 2n+1$, by

$$\mathsf{A}_{il}^{\mathcal{F}} = f_i(\theta_l), \text{ where } f_i \text{ denotes the elements of } \mathcal{F}_n \text{ (independently of their numbering).}$$

Its rank is at most $2n+1$. This is a simple consequence of the fact that the dimension of $Span\ \mathcal{F}_n$ is $2n+1 < (n+1)(n+2)/2$: indeed, this implies that there exists a matrix $\mathsf{C} \in \mathbb{C}^{((n+1)(n+2)/2-2n-1)\times(n+1)(n+2)/2}$ of rank $(n+1)(n+2)/2-2n-1$ such that

$$\forall i \in \mathbb{N}, 1 \leq i \leq (n+1)(n+2)/2 - 2n - 1, \quad \sum_{j=1}^{(n+1)(n+2)/2} \mathsf{C}_{ij} f_j = 0,$$

and therefore $\mathsf{C}\mathsf{A}^{\mathcal{F}} = 0_{((n+1)(n+2)/2-2n-1)\times N_p}$; as a result the N_p columns of $\mathsf{A}^{\mathcal{F}}$ belong to the kernel of C, which is of dimension $2n+1$; so the rank of $\mathsf{A}^{\mathcal{F}}$ is at most $2n+1$. In particular the matrix M_n^R introduced in Definition 9 is such a matrix $\mathsf{A}^{\mathcal{F}}$, and is therefore of rank at most $2n+1$.

We know that $\mathsf{M}_n^C = \mathsf{D}_n^{RC}\mathsf{M}_n^R$ and D_n^{RC} is non-singular, so $rk(\mathsf{M}_n^C) = rk(\mathsf{M}_n^R)$. The rank of M_n^C is at most equal to $2n+1$ for any choice of angles $\{\theta_l \in \mathbb{R}, 1 \leq l \leq p\}$. It was previously proved in Lemma 2 from [15] that for $p = 2n+1$ and directions such that $\{\theta_l \in [0, 2\pi), 1 \leq l \leq p, l_1 \neq l_2 \Rightarrow \theta_{l_1} \neq \theta_{l_2}\}$ the matrix M_n^C has rank $2n+1$. A trivial corollary of this proof is that, for any choice of p distinct angles in $[0, 2\pi)$,

$$rk(\mathsf{M}_n^C) = 2n+1 = rk(\mathsf{M}_n^R) \Leftrightarrow p \geq 2n+1. \tag{33}$$

In [15] we also proved that the space $\mathbb{F}$ for the constant coefficient Helmholtz operator is equal to the range of M_n^C for the corresponding wave number κ. As a direct consequence, a space $\mathbb{V}_h^\kappa = Span\{\exp \mathrm{i}\kappa(\cos\theta_l(x-x_0)+\sin\theta_l(y-y_0)), 1 \leq l \leq p\}$ for any choice of distinct angles in $[0, 2\pi)$ satisfies the interpolation property (30) for the Helmholtz equation if and only if $p \geq 2n+1$.

4.2 Generalized Plane Wave case

In order to prove that a GPW space $Span\ \mathbb{V}^0_{\alpha,p,q}$ (introduced in Definition 8) satisfies the interpolation property (30), we will rely on Proposition 6 to study the rank of the matrix (31) built from GPWs. As in the Helmholtz case, the proof relates the GPW matrix to the reference matrix, but here via an intermediate transition matrix.

Definition 10 Consider a point $(x_0, y_0) \in \mathbb{R}^2$, a given $q \in \mathbb{N}^*$, a given $M \in \mathbb{N}, M \geq 2$, a given set of complex-valued functions $\alpha = \{\alpha_{k,l} \in \mathcal{C}^{q-1}$ at $(x_0, y_0), 0 \leq k+l \leq M\}$, and the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$. For the corresponding set of GPWs, $\mathbb{V}^0_{\alpha,p,q} = \{\varphi_l := \exp P_l, \forall l \in \mathbb{N}^*, l \leq p, \theta_l \in [0, 2\pi), \theta_{l_1} \neq \theta_{l_2}\ \forall l_1 \neq l_2, \kappa \in \mathbb{C}^*\}$, we define the corresponding matrix (31), denoted M_n, as well as the transition matrix M_n^{Tr}, by

$$\begin{cases} \left(\mathsf{M}_n^{Tr}\right)_{\frac{(k_1+k_2)(k_1+k_2+1)}{2}+k_2+1,l} := (\lambda^l_{1,0})^{k_1}(\lambda^l_{0,1})^{k_2}/(k_1!k_2!), \\ (\mathsf{M}_n)_{\frac{(k_1+k_2)(k_1+k_2+1)}{2}+k_2+1,l} := \partial_x^{k_1}\partial_y^{k_2}\varphi_l(x_0, y_0)/(k_1!k_2!). \end{cases}$$

We first relate the transition matrix M_n^{Tr} to the reference matrix M_n^R.

Lemma 7 *Consider an open set $\Omega \subset \mathbb{R}^2$, $(x_0, y_0) \in \Omega$, a given $(M, n, p, q) \in (\mathbb{N}^*)^4$, $M \geq 2$, and a given set of complex-valued functions $\alpha = \{\alpha_{k_1,k_2} \in \mathcal{C}^{q-1}(\Omega), 0 \leq k_1+k_2 \leq M\}$, the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$ and set of GPWs $\mathbb{V}^0_{\alpha,p,q}$. There exists a block diagonal non-singular matrix D_n^{RT} such that $\mathsf{M}_n^{Tr} = \mathsf{D}_n^{RT}\mathsf{M}_n^R$, independently of the number p of GPWs in $\mathbb{V}^0_{\alpha,p,q}$.*

Proof As long as there are four complex numbers a, b, c, d such that

$$\forall p \in \mathbb{N}, 1 \leq l \leq p,\ \begin{pmatrix}\lambda^l_{1,0}\\ \lambda^l_{0,1}\end{pmatrix} = \begin{pmatrix} a & b \\ c & d\end{pmatrix}\begin{pmatrix}\cos\theta_l \\ \sin\theta_l\end{pmatrix},$$

then the diagonal blocks of $\mathsf{D}_n^{RT} = diag(d_K^{RT}, K \text{ from } 0 \text{ to } n)$ of increasing size $d_K^{RT} \in \mathbb{C}^{(K+1)\times(K+1)}$ can be built thanks to the following binomial formula

$$(\lambda^l_{1,0})^{K-k}(\lambda^l_{0,1})^k$$

$$= \sum_{i=0}^{K-k} \sum_{j=0}^{k} \binom{K-k}{i} \binom{k}{j} a^i c^j b^{K-k-i} d^{K-k-j} (\cos\theta_l)^{i+j} (\sin\theta_l)^{K-i-j}$$

since the coefficient of this linear combination of trigonometric functions are independent on l. □

The following step is naturally to relate the GPW matrix M_n to the reference matrix $\left(\mathbb{R}^2\right)$.

Proposition 7 *Consider an open set $\Omega \subset \mathbb{R}^2$, a point $(x_0, y_0) \in \Omega$, a given $(M, n, p, q) \in (\mathbb{N}^*)^4$, $M \geq 2$, $q \geq n-1$, and a given set of complex-valued functions $\alpha = \{\alpha_{k_1,k_2} \in \mathcal{C}^{\max(n,q-1)}(\Omega), 0 \leq k_1 + k_2 \leq M\}$, the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$ and set of GPWs $\mathbb{V}^0_{\alpha,p,q}$. There exists a lower triangular matrix L_n^R, whose diagonal coefficients are equal all non-zero and whose other non-zero coefficients depend only on derivatives of the PDE coefficients α evaluated at (x_0, y_0), such that*

$$\mathsf{M}_n = \mathsf{L}_n^R \cdot \mathsf{M}_n^R.$$

As a consequence $rk(\mathsf{M}_n) = rk(\mathsf{M}_n^R)$ independently of the number p of GPWs in $\mathbb{V}^0_{\alpha,p,q}$, and both $\|\mathsf{L}_n^R\|$ and $\|(\mathsf{L}_n^R)^{-1}\|$ are bounded by a constant depending only on the PDE coefficients α.

Remark 4 If $n = 1$, then the various matrices M belong to $\mathbb{C}^{3\times3}$, and we have the identity $\mathsf{M}_1 = \mathsf{M}_1^{Tr}$ independently of the value of q.

Proof Let's first relate M_n to M_n^{Tr}. The polynomials $R_{i,j} \in \mathbb{C}[X, Y]$ obtained in Proposition 6 have degree $\mathrm{d}R_{i,j} \leq i + j - 1$ and satisfy

$$\begin{aligned} \forall (i,j) \in \mathbb{N}^2, i + j \leq q + 1, \forall \varphi_l \in \mathbb{V}^0_{\alpha,p,q}, \ &\partial_x^i \partial_y^j \varphi_l(x_0, y_0) \\ &= (\lambda^l_{1,0})^i (\lambda^l_{1,0})^j + R_{i,j}(\lambda^l_{1,0}, \lambda^l_{1,0}). \end{aligned} \tag{34}$$

In order to apply this to all entries in the matrix M_n, it is sufficient for q to satisfy $n \leq q + 1$, which explains the assumption on q. Therefore each entry (i, j) of the matrix M_n can be written as the sum of the (i, j) entry of M_n^{Tr} and a linear combination of entries (k, j) of M_n^{Tr} for $k < i$. In other words, the existence of a lower triangular matrix L_n^T, whose diagonal coefficients are 1 and whose other non-zero coefficients depend only on the derivatives of the coefficients α evaluated at (x_0, y_0), such that $\mathsf{M}_n = \mathsf{L}_n^T \cdot \mathsf{M}_n^{Tr}$ is guaranteed since the coefficients of $R_{i,j}$ are independent of l and any monomial in $R_{i,j}(\lambda_{1,0}, \lambda_{1,0})$ has a degree lower than $i + j$.

As a consequence, the existence of L_n^R is guaranteed by Lemma 7 since the matrix $\mathsf{L}_n^R := \mathsf{L}_n^T \cdot \mathsf{D}_n^{RT}$ satisfies the desired properties. □

Everything is now in place to state and finally prove the necessary and sufficient condition on the number p of GPWs for the space $\mathbb{V}^0_{\alpha,p,q}$ to satisfy the interpolation property (30). We here turn to the specific case of second order operators.

Theorem 1 *Consider an open set $\Omega \subset \mathbb{R}^2$, a point $(x_0, y_0) \in \Omega$, $M = 2$, a given $(n, p, q) \in (\mathbb{N}^*)^3$, $n \geq M$, $q \geq n - 1$ and a given set of complex-valued functions $\alpha = \{\alpha_{k_1,k_2} \in \mathcal{C}^n(\Omega), 0 \leq k_1 + k_2 \leq M\}$, the corresponding partial differential operator $\mathcal{L}_{M,\alpha}$ and set of GPWs $\mathbb{V}^0_{\alpha,p,q}$. The space $\mathbb{V}^G_h := span \mathbb{V}^0_{\alpha,p,q}$ satisfies the property*

$$\begin{array}{l} \forall u \in \mathcal{C}^{n+2}(\Omega) \text{ satisfying } \mathcal{L}_{2,\alpha} u = 0, \exists u_a \in \mathbb{V}^G_h, \exists \text{ a constant } C(\Omega, n) \text{ s. t.} \\ \forall (x, y) \in \Omega, |u(x, y) - u_a(x, y)| \leq C(\Omega, n) \|(x, y) - (x_0, y_0)\|^{n+1}, \end{array} \tag{35}$$

if and only if $p \geq 2n + 1$.

Proof According to the discussion displayed in the introduction of Sect. 4, the proof focuses on the linear system (32) for the linear differential operator $\mathcal{L}_{2,\alpha}$. Indeed, defining the vector space

$$\begin{aligned} \mathbb{F}_\alpha := \Big\{ \mathsf{F} \in \mathbb{C}^{(n+1)(n+2)/2}, \exists v \in \mathcal{C}^{n+2}(\Omega) \text{ satisfying } \mathcal{L}_{2,\alpha} v = 0 \\ \text{s.t. } \mathsf{F}_{\frac{(k_1+k_2)(k_1+k_2+1)}{2}+k_2+1} = \partial_x^{k_1} \partial_y^{k_2} v(x_0, y_0)/(k_1! k_2!) \Big\} \end{aligned}$$

and considering $\mathsf{M}_n \in \mathbb{C}^{\frac{(n+1)(n+2)}{2} \times p}$ defined in (31) for a GPW basis $\mathbb{V}^G_h$, then (35) is equivalent to

$$\forall \mathsf{F} \in \mathbb{F}_\alpha, \exists \mathsf{X} \in \mathbb{C}^p \text{ s.t. } \mathsf{M}_n \mathsf{X} = \mathsf{F}. \tag{36}$$

Naturally, the two aspects of this proof are then associated to (1) the rank of M_n with respect to the choice of p, and (2) the relation between the right hand side and the range of the matrix.

Combining the fact that $rk(\mathsf{M}^R_n) = 2n + 1 \Leftrightarrow p \geq 2n + 1$ from (33) with the fact that $rk(\mathsf{M}_n) = rk(\mathsf{M}^R_n)$ for $q \geq n - 1$ from Proposition 7, we see immediately that, as long as $q \geq n - 1$, $rk(\mathsf{M}_n) = 2n + 1$ if and only if $p \geq 2n + 1$.

It is then sufficient to prove that the space $\mathbb{F}_\alpha$ belongs to the range of M_n, $\mathcal{R}(\mathsf{M}_n)$. To this end, we now define the space

$$\begin{aligned} \mathfrak{K} := \Big\{ \mathsf{K} \in \mathbb{C}^{(n+1)(n+2)/2}, \exists f \in \mathcal{C}^n(\Omega) \text{ satisfying } \mathcal{L}_{2,\alpha} f(x, y) = O(\|(x, y) - (x_0, y_0)\|^{n-1}) \\ \text{s.t. } \mathsf{K}_{\frac{(k_1+k_2)(k_1+k_2+1)}{2}+k_2+1} = \partial_x^{k_1} \partial_y^{k_2} f(x_0, y_0)/(k_1! k_2!) \Big\}. \end{aligned}$$

We can now see that

- $\mathcal{R}(\mathsf{M}_n) \subset \mathfrak{K}$ independently of the value of p, since by construction of GPWs, as long as $q \geq n - 1$, each column of M_n belongs to $\mathfrak{K}$;
- $\mathbb{F}_\alpha \subset \mathfrak{K}$, by definition of $\mathbb{F}_\alpha$;
- $\dim \mathfrak{K} = 2n + 1$, since - from the condition involving the Taylor expansion coefficients of $\mathcal{L}_{2,\alpha} f$ of order up to $n - 2$ at (x_0, y_0) set to zero - $\mathfrak{K} \subset \mathbb{C}^{(n+1)(n+2)/2}$ is the kernel of a matrix $\mathsf{A} \in \mathbb{C}^{n(n-1)/2 \times (n+1)(n+2)/2}$ with

$$\forall (i, j) \in \mathbb{N}^2, i + j < n - 1, \mathsf{A}_{\frac{(i+j)(i+j+1)}{2}+j+1, \frac{(i+j+2)(i+j+3)}{2}+j+1}$$

$$= \alpha_{2,0}(x_0, y_0) \neq 0 \text{ from Hypothesis 1},$$

$$\forall(\tilde{i}, \tilde{j}) \in \mathbb{N}^2, \tilde{i} + \tilde{j} < n - 1, \text{ if } \tilde{i} + \tilde{j} > i + j \text{ or if } \tilde{i} + \tilde{j} = i + j, \tilde{j} > j$$

$$\mathsf{A}_{\frac{(i+j)(i+j+1)}{2}+j+1, \frac{(\tilde{i}+\tilde{j}+2)(\tilde{i}+\tilde{j}+3)}{2}+\tilde{j}+1} = 0,$$

so that A is of maximal rank while its kernel has dimension $\frac{(n+1)(n+2)}{2} - \frac{n(n-1)}{2} = 2n + 1$.

Therefore, if $p \geq 2n+1$, we obtain that $\mathcal{R}(\mathsf{M}_n) = \mathfrak{K}$ and as a consequence $\mathbb{F}_\alpha \subset \mathcal{R}(\mathsf{M}_n)$ as expected. This concludes the proof. □

The necessary and sufficient condition on the number p of GPWs for the space $\mathbb{V}^0_{\alpha,p,q}$ to satisfy the interpolation property (30) when $M > 2$ are still unknown.

Remark 5 As in [15], the theorem holds in particular for the Helmholtz equation with sign changing.

5 Numerical experiments

In [15], GPWs where constructed and studied for the Helmholtz equation (1) with a variable and sign-changing coefficient β. The numerical experiments presented there were restricted to the Helmholtz equation at one point $(x_0, y_0) \in \mathbb{R}^2$, but considered a propagative case i.e. $\beta(x_0, y_0) < 0$, an evanescent case i.e. $\beta(x_0, y_0) > 0$, a cut-off case i.e. $\beta(x_0, y_0) = 0$. They also considered a case not covered by the convergence theorem, but important for future applications: considering GPWs centered at points (x_0, y_0) at a distance h from the cut-off.

Here, we are interested in illustrating the results presented in Theorem 1. Since the well known case of classical PW for the constant-coefficient Helmholtz equation is included by the hypotheses of the theorem, we cannot expect any improvement on the required number of basis functions p. However, we are interested in exploring the impact of the order of approximation q on the convergence of (35), in particular for anisotropic problems.

5.1 Test cases

We propose here four different test cases. Each test case consists of a partial differential operator of second order $\mathcal{L}$, an exact solution u of the equation $\mathcal{L}u = 0$, as well as a computational domain $\Omega \subset \mathbb{R}^2$, such that Hypotheses 1 and 2 hold at all $(x_0, y_0) \in \Omega$. The characteristics of the partial differential operators that we consider here are:

- polynomial coefficients α,
- non-polynomial coefficients α,
- anisotropy in the first order terms as $\overrightarrow{a}(x, y) \cdot \nabla$ for a vector-valued function $\overrightarrow{a}$;
- anisotropy in the second order terms as $\nabla \cdot (A(x, y)\nabla)$ for a matrix-valued function A.

The *Ad* est case We consider an isotropic partial differential operator with polynomial coefficients:

$$\begin{cases} \mathcal{L}_{Ad} := -\Delta + 2(x+y), \\ u_{Ad} : (x,y) \mapsto Ai(x+y), \\ \Omega_{Ad} := (-2,2)^2. \end{cases}$$

We have $\mathcal{L}_{Ad} u_{Ad} = 0$ on $\mathbb{R}^2$, all the coefficients of $\mathcal{L}_{Ad}$ belong to $\mathcal{C}^\infty\left(\mathbb{R}^2\right)$ and the coefficients $\{\alpha^{Ad}_{k,2-k}; k = 0,1,2\}$ satisfy

$$\sum_{k=0}^{2} \alpha^{Ad}_{k,2-k}(x_0,y_0) X^k Y^{2-k} = X^2 + Y^2 \quad \forall (x_0,y_0) \in \mathbb{R}^2,$$

so $\mathcal{L}_{Ad}$ satisfies Hypotheses 1 and 2 on $\mathbb{R}^2$. Note that the sign of the coefficient $\alpha^{Ad}_{0,0}(x,y) = 2(x+y)$ changes in the computational domain along the curve $x+y=0$.

The *Jc* test case We consider a partial differential operator with non-polynomial coefficients of the terms of order 1 and 0, and anisotropy in the first order term:

$$\begin{cases} \mathcal{L}_{Jc} := \nabla \cdot (x^2 \nabla) + \begin{pmatrix} -x \\ \cos y \end{pmatrix} \cdot \nabla + (\nu^2 - 2x^2 - \sin y), \\ u_{Jc} : (x,y) \mapsto J_1(x) \cos y, \\ \Omega_{Jc} := (1,4) \times (0, 2\pi). \end{cases}$$

We have $\mathcal{L}_{Jc} u_{Jc} = 0$ on $(0,\infty) \times \mathbb{R}$, all the coefficients of $\mathcal{L}_{Jc}$ belong to $\left(\mathbb{R}^+ \times \mathbb{R}\right)$ and the coefficients $\{\alpha^{Jc}_{k,2-k}; k = 0,1,2\}$ satisfy

$$\sum_{k=0}^{2} \alpha^{Jc}_{k,2-k}(x_0,y_0) X^k Y^{2-k} = x_0^2 (X^2 + Y^2) \quad \forall (x_0,y_0) \in \mathbb{R}^2,$$

so $\mathcal{L}_{Jc}$ satisfies Hypotheses 1 and 2 as long as $x > 0$.

The *JJ* test case We consider a partial differential operator with polynomial coefficients and anisotropy in the first and second order terms:

$$\begin{cases} \mathcal{L}_{JJ} := \nabla \cdot \begin{pmatrix} x^2 & 0 \\ 0 & y^2 \end{pmatrix} \nabla - \begin{pmatrix} x \\ y \end{pmatrix} \cdot \nabla + (x^2 + y^2 - 1), \\ u_{JJ} : (x,y) \mapsto J_0(x) J_1(y), \\ \Omega_{JJ} := (1,3) \times (1,3). \end{cases}$$

We have $\mathcal{L}_{JJ} u_{JJ} = 0$ on $(\mathbb{R}^+)^2$, all the coefficients of $\mathcal{L}_{JJ}$ belong to $\mathcal{C}^\infty\left((\mathbb{R}^+)^2\right)$ and the coefficients $\{\alpha^{JJ}_{k,2-k}; k = 0,1,2\}$ satisfy

$$\sum_{k=0}^{2} \alpha^{JJ}_{k,2-k}(x_0,y_0) X^k Y^{2-k} = x_0^2 X^2 + y_0^2 Y^2 \quad \forall (x_0,y_0) \in \mathbb{R}^2,$$

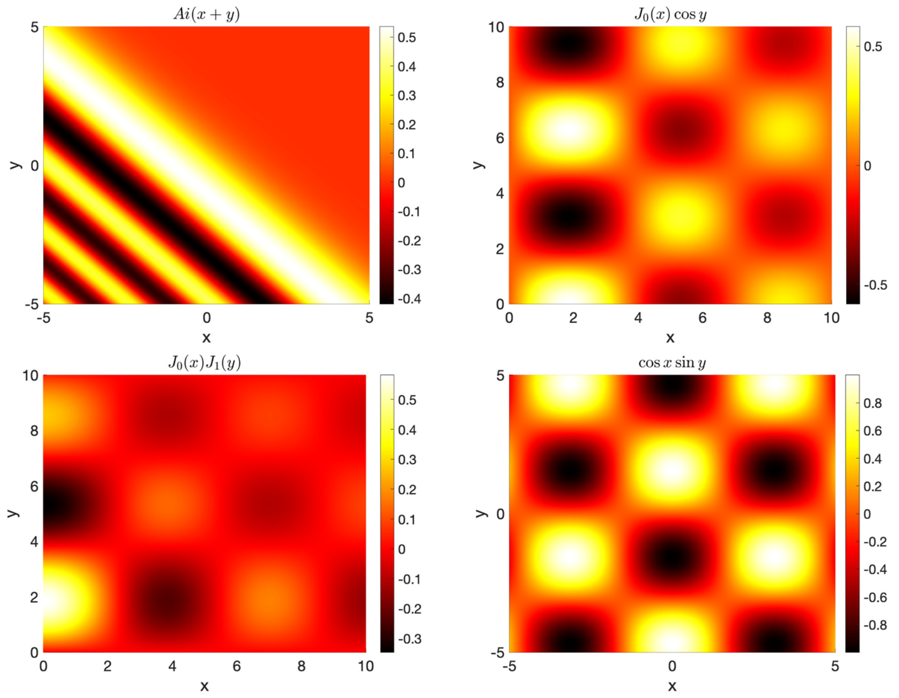

Fig. 4 Exact solutions of the four test cases

so $\mathcal{L}_{JJ}$ satisfies Hypotheses 1 and 2 as long as $xy \neq 0$.

The *cs* test case Finally we consider a partial differential operator with non-polynomial coefficients and anisotropy in the second order term:

$$\begin{cases} \mathcal{L}_{cs} := \nabla \cdot \begin{pmatrix} 1 & 0.1\cos x \sin y \\ 0.1\cos x \sin y & -2 \end{pmatrix} \nabla - 0.1 \begin{pmatrix} \cos x(\quad \cos y) \\ siny(- \sin x) \end{pmatrix} \cdot \nabla + (0.2 \sin x \cos y - 1), \\ \mathcal{L}_{cs} : = \partial_x^2 + 0.2 \cos x \sin y \; \partial_x \partial_y - 2\partial_y^2 + (0.2 \sin x \cos y - 1), \\ u_{cs} : (x, y) \mapsto \cos x \sin y, \\ \Omega_{cs} := (-1, 1)^2. \end{cases}$$

We have $\mathcal{L}_{cs} u_{cs} = 0$ on $\mathbb{R}^2$, all the coefficients of $\mathcal{L}_{cs}$ belong to $\mathcal{C}^\infty\left(\mathbb{R}^2\right)$ and the coefficients $\{\alpha_{k,2-k}^{cs}; k = 0, 1, 2\}$ satisfy

$$\sum_{k=0}^{2} \alpha_{k,2-k}^{cs}(x_0, y_0) X^k Y^{2-k} = \left(1 - \frac{(0.1)^2}{2} \cos^2 x_0 \sin^2 y_0\right) X^2 \\ - 2\left(Y - \frac{0.1}{2} \cos x_0 \sin y_0 X\right)^2 \quad \forall (x_0, y_0) \in \mathbb{R}^2,$$

so $\mathcal{L}_{cs}$ satisfies Hypotheses 1 and 2 on $\mathbb{R}^2$.

For reference, Fig. 4 displays the four solutions to the test cases.

5.2 Implementation of the construction algorithm

For a linear second order operator

$$\mathfrak{L}_{2,\alpha} = \alpha_{2,0}\partial_x^2 + \alpha_{1,1}\partial_x\partial_y + \alpha_{0,2}\partial_y^2 + \alpha_{1,0}\partial_x + \alpha_{0,1}\partial_y + \alpha_{0,0}$$

the associated operator $\mathfrak{L}^A_{2,\alpha}$ is defined by

$$\begin{aligned}\mathfrak{L}^A_{2,\alpha} P = &\underbrace{\alpha_{2,0}\partial_x^2 P + \alpha_{1,1}\partial_x\partial_y P + \alpha_{0,2}\partial_y^2 P}_{T_1}\\ &+ \underbrace{\alpha_{2,0}(\partial_x P)^2 + \alpha_{1,1}\partial_x P\partial_y P + \alpha_{0,2}(\partial_y P)^2}_{T_2} + \underbrace{\alpha_{1,0}\partial_x P + \alpha_{0,1}\partial_y P}_{T_3}.\end{aligned}$$

The implementation of Algorithm 1 simply requires, at each level $\mathfrak{L}$, the evaluation of $\{N_{I,\mathfrak{L}-I}, 0 \leq I \leq \mathfrak{L}\}$ to apply formula (23). At each level $\mathfrak{L}$ the coefficients $\{\mu_{ij}, (i,j) \in \mathbb{N}^2, i+j \leq q+1\}$ of $Q_{\mathfrak{L}} := \sum_{0\leq i+j\leq M+\mathfrak{L}-1} \lambda_{i,j}(x-x_0)^i(y-y_0)^j$ are computed as

$$\mu_{i,j} := \begin{cases} \lambda_{i,j} & \text{if } i+j \leq \mathfrak{L}+1 \\ 0 & \text{otherwise,} \end{cases}$$

and for $0 \leq I \leq \mathfrak{L}$ the different contributions to $N_{I,\mathfrak{L}-I}$ can be described as follows:

- the linear contributions from first order terms T_3

$$\begin{aligned}-\sum_{i=0}^{I}\sum_{j=0}^{\mathfrak{L}-I}\Big(&\mathcal{D}^{(I-i,\mathfrak{L}-I-j)}\alpha_{1,0}(x_0,y_0)(i+1)\mu_{i+1,j}\\ &+\mathcal{D}^{(I-i,\mathfrak{L}-I-j)}\alpha_{0,1}(x_0,y_0)(j+1)\mu_{i,j+1}\Big)\end{aligned}$$

- the non-linear contributions from the terms T_2

$$\begin{aligned}-&\sum_{i_1=0}^{I}\sum_{j_1=0}^{\mathfrak{L}-I}\sum_{i_2=0}^{i_1}\sum_{j_2=0}^{j_1}\\ &\Big(\mathcal{D}^{(I-i_1,\mathfrak{L}-I-j_1)}\alpha_{2,0}(x_0,y_0)(i_1-i_2+1)(i_2+1)\ \mu_{i_1-i_2+1,j_1-j_2}\mu_{i_2+1,j_2}\\ &+\mathcal{D}^{(I-i_1,\mathfrak{L}-I-j_1)}\alpha_{1,1}(x_0,y_0)(i_1-i_2+1)(j_2+1)\mu_{i_1-i_2+1,j_1-j_2}\mu_{i_2,j_2+1}\\ &+\mathcal{D}^{(I-i_1,\mathfrak{L}-I-j_1)}\alpha_{0,2}(x_0,y_0)(j_1-j_2+1)(j_2+1)\mu_{i_1-i_2,j_1-j_2+1}\mu_{i_2,j_2+1}\Big),\end{aligned}$$

- the linear contributions from the second order terms T_1

$$-\sum_{i=0}^{I}\sum_{j=0}^{\mathfrak{L}-I}\Big(\mathcal{D}^{(I-i,\mathfrak{L}-I-j)}\alpha_{2,0}(x_0,y_0)(i+2)(i+1)\mu_{i+2,j} \\ +\mathcal{D}^{(I-i,\mathfrak{L}-I-j)}\alpha_{1,1}(x_0,y_0)(j+1)(i+1)\mu_{i+1,j+1} \\ +\mathcal{D}^{(I-i,\mathfrak{L}-I-j)}\alpha_{0,2}(x_0,y_0)(j+2)(j+1)\mu_{i,j+2}\Big),$$

- the contribution from the zeroth order term $\alpha_{0,0}$

$$-\mathcal{D}^{(I,\mathfrak{L}-I)}\alpha_{0,0}(x_0,y_0).$$

Moreover, all experiments are conducted with the following choice of angles θ_l and κ parameters to build the GPW space $\mathbb{V}^0_{\alpha,p,q}$:

$$\begin{cases}\theta_l := \frac{\pi}{6}+\frac{2(l-1)\pi}{p}, \ \forall l\in\mathbb{N}, 1\le l\le p,\\ \kappa=\sqrt{-\alpha_{0,0}(x_0,y_0)}.\end{cases}$$

All exact solutions of the test cases are either products of a function of x by a function of y, or a function of $x+y$. Our particular choice of angles for the basis functions is made to avoid the unrealistically favorable case of having a basis function propagating in a direction aligned with the x direction, the y direction or the $x+y$ direction.

5.3 Construction of a solution to system (36)

In order to construct of a solution to System (36), we follow [15] in defining a matrix

$$(\mathsf{P}_n)_{n\pm k+1,\frac{k(k+1)}{2}+s+1}=(\pm i)^s,$$

and actually solving the square system

$$(\mathsf{P}_n\mathsf{M}_n)\mathsf{X}=\mathsf{P}_n\mathsf{F}.$$

5.4 Numerical results

The h-convergence results presented in Theorem 1 are stated as local properties at a given point. In order to illustrate them, for each test case, we consider the following procedure.

- At each of 50 random points (x_0,y_0) in the computational domain Ω
  1. Construct the set of GPWs from Algorithm 1 with the normalization proposed in Sect. 3.
  2. Compute u_a the linear combination of GPWs studied in the theorem's proof, matching its Taylor expansion to that of the exact solution.

- Estimate as a function of h the maximum L^∞ error on a disk of radius h centered at the random point: $\max_{(x_0,y_0)\in\Omega}\|u-u_a\|_{L^\infty(\{(x,y)\in\mathbb{R}^2,|(x,y)-(x_0,y_0)|<h\})}$.

We always consider a space $\mathbb{V}^0_{\alpha,p,q}$ of $p=2n+1$ GPWs. According to the theorem, we expect to observe convergence of order $n+1$ if the approximation parameter q in the construction of the basis functions is at least equal to $n-1$. For each of the four test cases proposed, we present: on the one hand results for n from 1 to 5 with $q=\max(1,n-1)$ (Left panel); on the other hand results for q from 1 to 4 with $n=4$ (Right panel). Hence with the first choice of parameters the theorem predicts convergence of order $n+1$, while with the second choice the theorem does not cover these cases.

The results are presented in Fig. 5 for the approximation of u_{Ad}, Fig. 6 for the approximation of u_{Jc}, Fig. 7 for the approximation of u_{JJ}, and Fig. 8 for the approximation of u_{cs}. We observe on Figs. 5 and 8 the effect of the large condition number of the matrix $\mathsf{P}_n\mathsf{M}_n$ on the accuracy of the approximation of u by u_a: even though the expected orders of convergence are observed for large values of h, when n increases the error stagnates at an increasing threshold for smaller values of h. Approximate condition number of the matrix $\mathsf{P}_n\mathsf{M}_n$ for the corresponding Ad and cs cases are provide in the following table.

	$n=1$ $q=1$	$n=2$ $q=1$	$n=3$ $q=2$	$n=4$ $q=3$	$n=5$ $q=4$	$n=4$ $q=1$	$n=4$ $q=2$	$n=4$ $q=3$	$n=4$ $q=4$
cond $\mathsf{P}_n\mathsf{M}_n^{Ad}$	$4.8\cdot10^0$	$4.8\cdot10^0$	$4.5\cdot10^1$	$3.2\cdot10^4$	$6.9\cdot10^5$	$4.8\cdot10^0$	$6.4\cdot10^2$	$3.2\cdot10^4$	$5.9\cdot10^5$
cond $\mathsf{P}_n\mathsf{M}_n^{cs}$	$1.5\cdot10^0$	$1.5\cdot10^0$	$2.0\cdot10^1$	$7.8\cdot10^4$	$1.6\cdot10^5$	$1.5\cdot10^0$	$2.0\cdot10^1$	$7.8\cdot10^4$	$1.4\cdot10^4$

Such problems of conditioning are inherent to wave-like bases, and for larger values of n, the condition number may become a limitation to compute accurate solutions. Techniques similar to the QR factorization proposed in [1] could be investigated in the future to improve the accuracy of this computation.

We also observe, on the left panels of Figs. 5, 6 and 7 , that these three test cases the constant $C(\Omega,n)$ from (35) in Theorem 1 does not seem to depend on n, even though the Theorem predicts that it does. The situation seems different on the left panel of Fig. 8.

We summarize in the following table the orders of convergence observed, always using $\mathbb{V}^0_{\alpha,p,q}$ with $p=2n+1$. The bold entries correspond to cases covered by Theorem 1 i.e. $n+1$ for $q\le n-1$, and the red entries correspond to cases with order of convergence observed higher than the theorem predicts.

$q\backslash n$	1	2	3	4	5
1	**2**	**3**	3	3/4	3
2	**2**	**3**	**4**	$\geq$ 4	$\geq$ 4
3	**2**	**3**	**4**	$\geq$ **5**	5
4	**2**	**3**	**4**	$\geq$ **5**	**6**

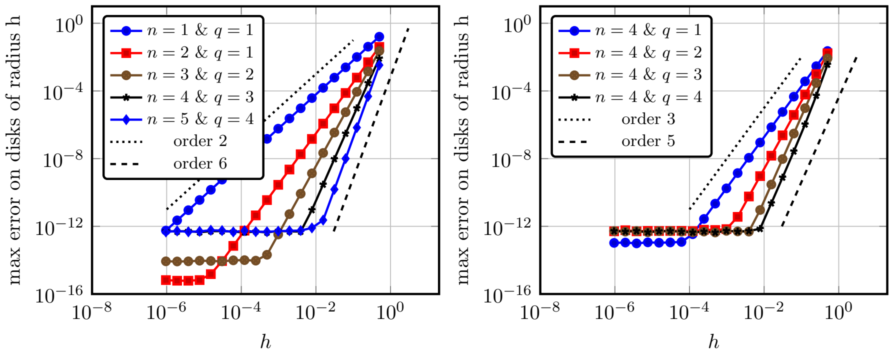

Fig. 5 GPW approximation of u_{Ad} by $u_a \in \mathbb{V}^0_{\alpha,p,q}$ with $p = 2n+1$. We represent the L^∞ error $\max_{(x_0,y_0)\in\Omega} \|u_{Ad} - u_a\|_{L^\infty(\{(x,y)\in\mathbb{R}^2, |(x,y)-(x_0,y_0)|<h\})}$, for 50 random points $(x_0, y_0) \in \Omega_{Ad}$. We compare results for parameters satisfying Theorem 1 hypotheses i.e. $q = \max(1, n-1)$ (Left panel), and for varying q with fixed n (Right panel)

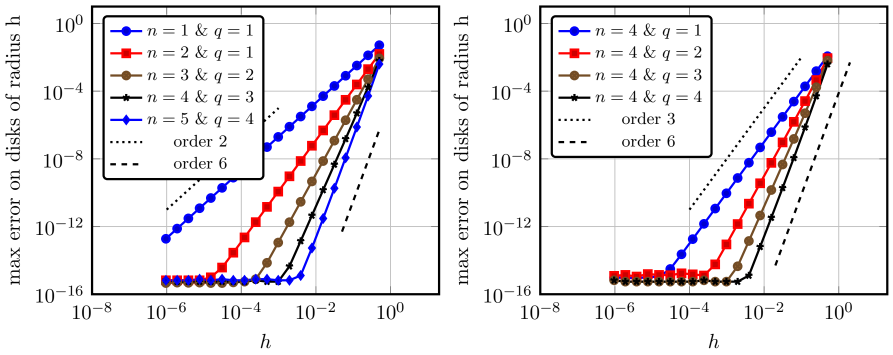

Fig. 6 GPW approximation of u_{Jc} by $u_a \in \mathbb{V}^0_{\alpha,p,q}$ with $p = 2n+1$. We represent the L^∞ error $\max_{(x_0,y_0)\in\Omega} \|u_{Jc} - u_a\|_{L^\infty(\{(x,y)\in\mathbb{R}^2, |(x,y)-(x_0,y_0)|<h\})}$, for 50 random points $(x_0, y_0) \in \Omega_{Jc}$. We compare results for parameters satisfying Theorem 1 hypotheses i.e. $q = \max(1, n-1)$ (Left panel), and for varying q with fixed n (Right panel)

We can see from this table that in all cases covered by the theorem, we observe a convergence rate equal or slightly better than predicted. But it would seem that the hypotheses of the theorem are sharp.

6 Conclusion

In this work we have considered local properties in the neighborhood of a point $(x_0, y_0) \in \mathbb{R}^2$, for an operator $\mathcal{L}_{M,\alpha}$. To summarize, we followed the steps announced in the introduction:

1. construction of GPWs φ such that $\mathcal{L}_{M,\alpha}\varphi(x, y) = O\left(\|(x, y) - (x_0, y_0)\|^q\right)$
   (a) choose an ansatz for $\varphi(x, y) = \exp \sum_{0\le i+j\le dP} \lambda_{ij}(x - x_0)^i (y - y_0)^j$

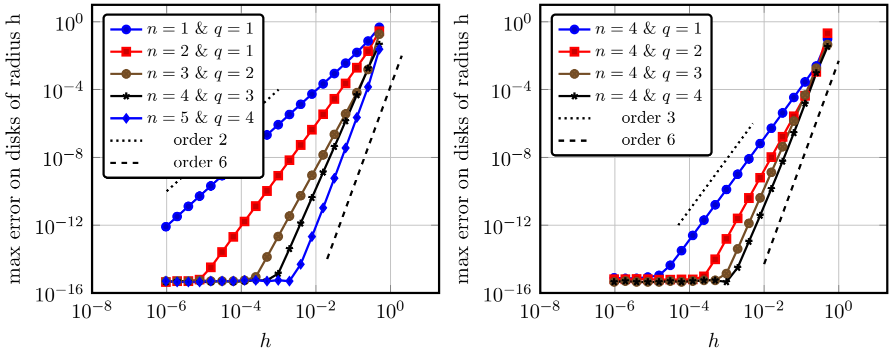

Fig. 7 GPW approximation of u_{JJ} by $u_a \in \mathbb{V}^0_{\alpha,p,q}$ with $p = 2n + 1$. We represent the L^∞ error $\max_{(x_0,y_0)\in\Omega} \|u_{JJ} - u_a\|_{L^\infty(\{(x,y)\in\mathbb{R}^2, |(x,y)-(x_0,y_0)|<h\})}$, for 50 random points $(x_0, y_0) \in \Omega_{JJ}$. We compare results for parameters satisfying Theorem 1 hypotheses i.e. $q = \max(1, n - 1)$ (Left panel), and for varying q with fixed n (Right panel)

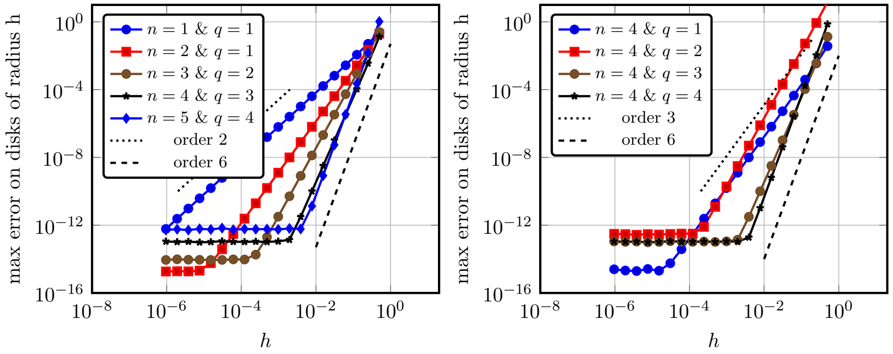

Fig. 8 GPW approximation of u_{cs} by $u_a \in \mathbb{V}^0_{\alpha,p,q}$ with $p = 2n + 1$. We represent the L^∞ error $\max_{(x_0,y_0)\in\Omega} \|u_{cs} - u_a\|_{L^\infty(\{(x,y)\in\mathbb{R}^2, |(x,y)-(x_0,y_0)|<h\})}$, for 50 random points $(x_0, y_0) \in \Omega_{cs}$. We compare results for parameters satisfying Theorem 1 hypotheses i.e. $q = \max(1, n - 1)$ (Left panel), and for varying q with fixed n (Right panel)

(b) identify the corresponding $N_{dof} = \frac{(dP+1)(dP+2)}{2}$ degrees of freedom, and $N_{eqn} = \frac{q(q+1)}{2}$ constraints, namely respectively

$$\{\lambda_{ij}; (i, j) \in \mathbb{N}^2, 0 \leq i + j \leq dP\},$$
$$\{\mathcal{D}^{(I,J)}\mathcal{L}_{M,\alpha}\varphi(x_0, y_0) = 0; (I, J) \in \mathbb{N}^2, 0 \leq I + J < q\}.$$

(c) for $dP = q + M - 1$, the number of degrees of freedom is $N_{dof} = \frac{(M+q)(M+q+1)}{2} > N_{eqn}$ and this ensures that there are linear terms in all the constraints

(d) identify $N_{dof} - N_{eqn} = Mq + \frac{M(M+1)}{2}$ additional constraints, namely

$$\text{Fixing } \{\lambda_{i,j}, (i, j) \in \mathbb{N}^2, i + j < q + M \text{ and } i < M\}$$

to obtain a global system that can be split into a hierarchy of linear triangular subsystems

(e) compute the remaining N_{eqn} degrees of freedom by forward substitution for each triangular subsystem, therefore at minimal computational cost

2. interpolation properties

(a) thanks to the normalization, in particular $\{\lambda_{i,j} = 0, (i,j) \in \mathbb{N}^2, i+j < M+q \text{ and } i < M, i+j \neq 1\}$, study the properties of the remaining N_{eqn} degrees of freedom, that is $\{\lambda_{i,j}, (i,j) \in \mathbb{N}^2, i+j < M+q \text{ and } i \geq M\}$, with respect to $(\lambda_{1,0}, \lambda_{0,1})$

(b) identify a simple reference case depending only on two parameters, that is basis functions $\phi(x,y) = \exp \lambda_{1,0}(x-x_0) + \lambda_{0,1}(y-y_0)$ depending only on the choice of $(\lambda_{1,0}, \lambda_{0,1})$, independently of ϕ being an exact solution to the constant coefficient equation

(c) study the interpolation properties of this reference case with classical PW techniques

(d) relate the general case to the reference case thanks to 2a

(e) prove the interpolation properties of the GPWs from those of the reference case

This construction process guarantees that the GPW function φ satisfies the approximate Trefftz property $\mathcal{L}_{M,\alpha}\varphi(x,y) = O\left(\|(x,y)-(x_0,y_0)\|^q\right)$ independently of the normalization, that is the values chosen for $\{\lambda_{i,j}, (i,j) \in \mathbb{N}^2, i+j < M\}$, while the proof of interpolation properties heavily rely on the normalization.

This work focuses on interpolation of solutions of a PDE, and is limited to local results, in the neighborhood of a given point. In order to address the convergence of a numerical method for a boundary value problem on a domain Ω with a GPW-discretized Trefftz method, on a mesh $\mathcal{T}_h$ of Ω, we will consider a space $\mathbb{V}_h$ of GPWs built element-wise, at the centroid $(x_0,y_0) = (x_K, y_K)$ of each element $K \in \mathcal{T}_h$, to study interpolation properties on Ω. In particular, meshing the domain Ω to respect any discontinuity in the coefficients, the interpolation error on Ω, $\|(I - P_{\mathbb{V}_h})\|$, will converge at the same order as the local interpolation error on each element, and the crucial point will be to investigate the behavior of the constant $C(\Omega, n)$ from Theorem 1. Related computational aspects of the construction of GPWs proposed in this work are currently under study.

We are also currently considering extensions to 3D problems. We expect to be able to follow a similar roadmap to construct GPWs and study their interpolation properties. However, even if we expect a similar layer structure for the system obtained to construct GPWs, the subsystems won't have a natural numbering making obvious their triangularity. We will therefore need new tools to construct solutions to the subsystems. Moreover, in 3D, choosing appropriate directions for the normalization of GPWs is challenging, and we anticipate that the study of interpolation properties will be more intricate.

Acknowledgements This material is based upon work supported by the National Science Foundation under Grants No. DMS-1818747 and DMS-2105487.

A 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 [26]. 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 mth 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 [7].

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 $(b_k)_{k\in[\![1,m]\!]}$ of $\sum_k kb_k = m$. 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 [6] applied to dimension 2. A linear order on $\mathbb{N}^2$ is defined by: $\forall(\mu, \nu) \in \left(\mathbb{N}^2\right)^2$, the relation $\mu \prec \nu$ 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

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

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

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

See [12] for a proof of the formula interpreted in terms of collapsing partitions.

B Faa di Bruno

The multivariate formula has been widely studied, the version described here is the one from [6] applied to dimension 2. A linear order on $\mathbb{N}^2$ is defined by: $\forall(\mu,\nu)\in\left(\mathbb{N}^2\right)^2$, the relation $\mu\prec\nu$ 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

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

$$\partial_x^k\partial_y^{\ell-k}e^{P(x,y)} = k!(\ell-k)!\sum_{1\leq\mu\leq\ell} e^{P(x,y)}\sum_{s=1}^{\ell}\sum_{p_s((k,\ell-k),\mu)}\prod_{m=1}^{s}\frac{1}{k_m!}\left(\frac{1}{i_m!j_m!}\partial_x^{i_m}\partial_y^{j_m}P(x,y)\right)^{k_m},$$

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

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

Note that s is the number of different terms appearing in the product, while μ is the number of terms in the product counting multiplicity, k_m is the multiplicity of the mth term in the product, while p_s represents the possible partitions of (i,j).

Note that since $k_m>0$, the condition $\sum_{m=1}^{s}k_m=\mu$ implies that $\mu=\sum_{m=1}^{s}k_m\geq\sum_{m=1}^{s}1=s$.

C Polynomial formulas

Here are two important comments. The first one concerns the product of polynomials. Assume that $\min(D_1,D_2)\geq q$. Then the product of two polynomials, respectively of degree D_1 and D_2, satisfies:

$$\left(\sum_{i_1=0}^{D_1}\sum_{j_1=0}^{D_1-i_1} p_{i_1,j_1}x^{i_1}y^{j_1}\right)\left(\sum_{i_2=0}^{D_2}\sum_{j_2=0}^{D_2-i_2} q_{i_2,j_2}x^{i_2}y^{j_2}\right)$$
$$=\sum_{i=0}^{q-1}\sum_{j=0}^{q-1-i}\left(\sum_{\tilde{i}=0}^{i}\sum_{\tilde{j}=0}^{j} p_{i-\tilde{i},j-\tilde{j}}q_{\tilde{i},\tilde{j}}\right)x^i y^j + O(h^q).$$

Since in particular the summation indices are such that $0 \le \tilde{i} \le i$, $0 \le i - \tilde{i} \le i$, $0 \le \tilde{j} \le j$, and $0 \le j - \tilde{j} \le j$, the only coefficients $p_{i,j}$ and $q_{i,j}$ appearing in the (I_0, J_0) coefficient of the product have a length of the multi-index $i + j \le I_0 + J_0$. As a consequence, the only coefficients of several polynomials appearing in the (I_0, J_0) coefficient of the product these several polynomials have a length of the multi-index $i + j \le I_0 + J_0$. The second comment turns to the derivative of a polynomial:

$$\partial_x^I \partial_y^J \left(\sum_{i=0}^{D}\sum_{j=0}^{D-i} p_{i,j}x^i y^j\right) = \sum_{i=0}^{D-I-J}\sum_{j=0}^{D-I-J-i} \frac{(i+I)!}{i!}\frac{(j+J)!}{j!} p_{i+I,j+J}x^i y^j.$$

In particular the only coefficients $p_{i,j}$ appearing in the (I_0, J_0) coefficient of the derivative has a length of the multi-index $i + j = I + J + I_0 + J_0$.

References

1. Antunes, P.: A numerical algorithm to reduce ill-conditioning in meshless methods for the Helmholtz equation. Numer. Algorithms **79**(3), 879–897 (2018)
2. Babuska, I., Melenk, J.M.: The partition of unity method. Int. J. Numer. Methods Eng. **40**(4), 727–758 (1997)
3. Babuska, I., Zhang, Z.: The partition of unity method for the elastically supported beam. In: Symposium on Advances in Computational Mechanics, Computer Methods in Applied Mechanics and Engineering, vol. 5. 152(1–2), pp. 1–18 (1998)
4. Buet, C., Despres, B., Morel, G.: Trefftz Discontinuous Galerkin basis functions for a class of Friedrichs systems coming from linear transport, hal-01964528
5. Cessenat, O.: In: Application d'une nouvelle formulation variationnelle aux équations d'ondes harmoniques. Problèemes de Helmholtz 2D et de Maxwell 3D, Université Paris 9 Dauphine (1996)
6. Constantine, G.M., Savits, T.H.: A multivariate Faà di Bruno formula with applications. Trans. Am. Math. Soc. **348**(2), 503–520 (1996)
7. Craik, A.D.D.: Prehistory of Faa di Bruno's formula. Am. Math. Mon. **112**(2), 119–130 (2005)
8. Eckart, C.: The propagation of gravity waves from deep to shallow water, Circular 20. National Bureau of Standards, pp. 165–173 (1952)
9. Farhat, C., Harari, I., Franca, L.P.: The discontinuous enrichment method. Comput. Methods Appl. Mech. Eng. **190**(48), 6455–6479 (2001)
10. Fix, G.J., Gulati, S., Wakoff, G.I.: On the use of singular functions with finite element approximations. J. Comput. Phys. **13**, 209–228 (1973)
11. Gittelson, C.J., Hiptmair, R.: Dispersion analysis of plane wave discontinuous Galerkin methods. Int. J. Numer. Methods Eng. **98**(5), 313–323 (2014)
12. Hardy, M.: Combinatorics of partial derivatives. Electron. J. Comb. **13**(1), 13 (2006)
13. Hiptmair, R., Moiola, A., Perugia, I.: A survey of Trefftz methods for the Helmholtz equation. In: Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations, Lecture Notes in Computational Science and Engineering, vol. 114, pp. 237–278. Springer, Berlin (2016)

14. Huttunen, T., Monk, P., Kaipio, J.P.: Computational aspects of the ultra-weak variational formulation. J. Comput. Phys. **182**(1), 27–46 (2002)
15. Imbert-Gérard, L.-M.: Interpolation properties of generalized plane waves. Numer. Math. **131**, 683–711 (2015)
16. Imbert-Gérard, L.-M.: Generalized plane waves for varying coefficients. In: Proceedings of Waves, Karslruhe (2015)
17. Imbert-Gerard, L.-M., Despres, B.: A generalized plane-wave numerical method for smooth nonconstant coefficients. IMA J. Numer. Anal. (2013). https://doi.org/10.1093/imanum/drt030
18. Imbert-Gérard, L.-M., Monk, P.: Numerical simulation of wave propagation in inhomogeneous media using Generalized Plane Waves. ESAIM: M2AN **51**(4), 1387–1406 (2017)
19. Imbert-Gérard, L.-M.: Well-posedness and generalized plane waves simulations of a 2D mode conversion model. J. Comput. Phys. **303**, 105–124 (2015)
20. Kita, E., Kamiya, N.: Trefftz method: an overview. Adv. Eng. Softw. **24**, 3–12 (1995)
21. Kretzschmar, F., Moiola, A., Perugia, I., Schnepp, S.M.: A priori error analysis of space-time Trefftz discontinuous Galerkin methods for wave problems. IMA J. Numer. Anal. **36**, 1599 (2016)
22. Kretzschmar, F., Schnepp, S.M., Tsukerman, I., Weiland, T.: Discontinuous Galerkin methods with Trefftz approximations. J. Comput. Appl. Math. **270**, 211–222 (2014)
23. Kupradze, V.D., Gegelia, T.G., Basheleishvili, M.O., Burchuladze, T.V.: Three-Dimensional Problems of the Mathematical Theory of Elasticity and Thermoelasticity. North-Holland, New York (1979)
24. Lieu, A., Gabard, G., Bériot, H.: A comparison of high-order polynomial and wave-based methods for Helmholtz problems. J. Comput. Phys. **321**, 105–125 (2016)
25. Luostari, T., Huttunen, T., Monk, P.: The ultra weak variational formulation using Bessel basis functions. Commun. Comput. Phys. **11**(2), 400–414 (2012)
26. Ma, T.-W.: Higher chain formula proved by combinatorics. Electron. J. Combin. **16**(1), 7 (2009)
27. Maunder, E.A.W.: Trefftz in translation. Comput. Assist. Mech. Eng. Sci. **10** (2003)
28. Melenk, J.M.: On Generalized Finite Element Methods, Ph.D. thesis. The University of Maryland (1995)
29. Melenk, J.M., Babuska, I.: The partition of unity finite element method: basic theory and applications. Comput. Methods Appl. Mech. Eng. **139**(1–4), 289–314 (1996)
30. Mikhlin, S.G.: Variational Methods in Mathematical Physics. Pergamon Press; distributed by Macmillan, New York (1964)
31. Morel, G., Buet, C., Despres, B.: Trefftz discontinuous Galerkin method for Friedrichs systems with linear relaxation: application to the P 1 Model. Comput. Methods Appl. Math. **18**(3), 521–557 (2018)
32. Moiola, A., Perugia, I.: A space-time Trefftz discontinuous Galerkin method for the acoustic wave equation in first-order formulation. Numer. Math. **138**(2), 389–435 (2018)
33. Rektorys, K.: Variational Methods in Mathematics, Science and Engineering. Springer, Berlin (2012)
34. Strang, G., Fix, G.J.: An Analysis of the Finite Element Method, Prentice-Hall Series in Automatic Computation. Prentice-Hall Inc, Englewood Cliffs (1973)
35. Trefftz, E.: In: Ein gegenstuck zum ritzschen verfahren, pp. 131–137. Orell Fussli Verlag, Zurich (1926)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.