SIAM J. NUMER. ANAL.
Vol. 59, No. 3, pp. 1663–1686

AMPLITUDE-BASED GENERALIZED PLANE WAVES: NEW QUASI-TREFFTZ FUNCTIONS FOR SCALAR EQUATIONS IN TWO DIMENSIONS*

LISE-MARIE IMBERT-GERARD†

Abstract. Generalized plane waves (GPWs) were introduced to take advantage of Trefftz methods for problems modeled by variable coefficient equations. Despite the fact that GPWs do not satisfy the Trefftz property, i.e., they are not exact solutions to the governing equation, they instead satisfy a quasi-Trefftz property: They are only approximate solutions. They lead to high-order numerical methods, and this quasi-Trefftz property is critical for their numerical analysis. The present work introduces a new family of GPWs: amplitude-based. The motivation lies in the poor behavior of the phase-based GPW approximations in the preasymptotic regime, which will be tamed by avoiding high-degree polynomials within an exponential. The new ansatz introduces higher-order terms in the amplitude rather than in the phase of a plane wave as was initially proposed. The new functions' construction and the study of their approximation properties are guided by the road map proposed in [L.-M. Imbert-Gérard and G. Sylvand, *Numer. Math.*, to appear]. For the sake of clarity, the first focus is on the two-dimensional Helmholtz equation with spatially varying wave number. The extension to a range of operators allowing for anisotropy in the first- and second-order terms follows. Numerical simulations illustrate the theoretical study of the new quasi-Trefftz functions.

Key words. quasi-Trefftz methods, generalized plane waves, best approximation properties

AMS subject classifications. 65N30, 65N80, 68W25

DOI. 10.1137/20M136791X

1. Introduction.

Our interest lies in the numerical simulation of time-harmonic wave propagation in inhomogeneous media for boundary value problems with a scalar-valued governing equation. While homogeneous media lead to governing PDEs with constant coefficients, inhomogeneous media lead to variable-coefficient PDEs. In this article we introduce a new family of basis functions to be used as a basis in a quasi-Trefftz numerical method, show how to compute them, and study their approximation properties. Trefftz methods, a particular type of Galerkin method, have the specificity to rely on function spaces of solutions to the governing PDE as opposed to standard function spaces of sufficiently smooth functions. They take advantage of this specificity via the derivation of a weak formulation of the boundary value problem of interest, leading to a weak formulation with no volume term. This of course results in a considerable reduction of the discretization's computational cost. We will hereafter refer to solutions of the governing PDE as Trefftz functions and to spaces of Trefftz functions as Trefftz spaces.

The initial idea of using solutions to the governing PDE was introduced by Trefftz in 1926 [37], more recently translated from German [34], to obtain estimates on the solution of boundary value problems. It was then developed as early as the 1930s under the name of Trefftz method to study, for instance, problems of elasticity [35, 29] and torsion [18, 11]. Later the method was fruitfully extended to numerical

*Received by the editors September 18, 2020; accepted for publication (in revised form) March 10, 2021; published electronically June 16, 2021.
https://doi.org/10.1137/20M136791X
Funding: The work of the author was supported by National Science Foundation grant DMS-1818747.
†University of Arizona, Tucson, AZ 85721 USA (lmig@math.arizona.edu, https://www.math.arizona.edu/~lmig/).

approximation. In [9], the author classifies the method as a boundary method, as the solution to a boundary value problem is approximated by a linear combination of Trefftz functions, while this linear combination is constructed to take into account the boundary condition; see also [17]. Trefftz functions were used as a local basis on a region and coupled to a finite element region in [44]. In [26], they were used as a local basis on individual mesh elements instead of polynomial basis functions; an overview of such early Trefftz-type Galerkin methods can be found in [27].

Several versions of Trefftz methods have been actively developed to the numerical simulation of time-harmonic wave propagation problems for more than 20 years [16], such as the Trefftz and discontinuous Galerkin method [15, 14] and the ultraweak variational formulation [12, 6, 7]. More recent development include the Trefftz virtual element method [32, 33]. Trefftz spaces as well as Trefftz functions are absolutely fundamental to all Trefftz methods, throughout the weak formulation derivation as well as throughout the analysis of the resulting numerical method.

Ideally the discretization of a Trefftz weak formulation is performed via a finite-dimensional vector subspace of the Trefftz space. Trefftz functions are available for several problems of time-harmonic wave propagation through homogeneous media, i.e., when the governing PDE has constant coefficients. Typical examples include the most obvious plane, circular or spherical waves, but also functions constructed from Bessel functions [8, 31]. By contrast, for wave propagation through inhomogeneous media, i.e., when the governing PDE has either smooth or piecewise smooth variable coefficients, in general Trefftz functions are not available. Nevertheless, the idea of replacing Trefftz functions by approximate solutions to the governing PDE to discretize a Trefftz formulation when the PDE has variable coefficients was introduced in [22], and various aspects of these functions, called generalized plane waves (GPWs), as well as the resulting method, were studied in [20, 21, 23, 24].

For reference, ideas related to Taylor-based constructions of exact *polynomial* solutions can be found in [28, 43], while equation-based finite-difference approaches can be found in [3, 38, 39].

In general we will refer to approximate solutions to the PDE as quasi-Trefftz functions and to associated methods as quasi-Trefftz methods. As the PDE coefficients are variable, there is no hope to guarantee global approximation properties, and therefore the construction of quasi-Trefftz functions should emphasize local properties. For instance, GPWs are constructed locally and satisfy a local approximation of the PDE in the sense of a Taylor expansion. In general, quasi-Trefftz functions are defined piecewise, element per element on a mesh of the computational domain, and the approximation of the PDE is expected to be accurate locally on each element. We will therefore focus on the neighborhood of a generic point $(x_c, y_c) \in \mathbb{R}^2$ that could, for instance, be the centroid of each mesh element.

The GPWs introduced in [22] are phase-based, as they rely on the addition of higher-order terms in the phase of classical plane waves (PWs). Because of the presence of high-degree polynomials within an exponential, they exhibited a poor behavior in the preasymptotic regime, i.e., at a distance $h \approx 1$ from (x_c, y_c). This motivates the introduction of a new family of GPWs, avoiding the presence of high-degree polynomials within an exponential. In this article we introduce a family of amplitude-based GPWs, with a phase identical to that of a PW but with higher-degree polynomials in the amplitude. The first challenge is to design of a new ansatz. Thanks to an appropriate choice of ansatz, the problem of constructing a GPW will be reformulated into a problem that has solutions, but, most importantly, a problem for which a particular solution can be constructed at a limited computational cost. The second challenge is

to decipher the relation between the new functions and PWs in order to study the local approximation properties of these new GPWs.

The construction of amplitude-based GPWs and the study of their approximation properties will be addressed in sections 2 and 3, respectively, for the two-dimensional Helmholtz equation, namely:

$$\Big[-\Delta - \kappa^2(x,y)\Big]u = 0, \tag{1.1}$$

following the road map proposed in [24]. The extension of these results to a set of anisotropic second-order PDEs will be the focus of section 4. Then section 5 presents numerical experiments illustrating approximation properties and emphasizing the improvement obtained with respect to phase-based GPWs in terms of approximation properties.

To describe various useful sets of indices, we will write $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.

2. Construction of amplitude-based GPWs. The construction of this new family of GPWs relies on the choice of a new ansatz, shifting the focus from the phase to the amplitude of PWs. Here we emphasize the motivation behind the new ansatz and the reformulation of the problem of defining an amplitude-based GPW into a well-posed problem. The final goal of this section is to provide an algorithm to construct such GPWs, guaranteeing that they satisfy an approximation of the PDE (1.1); see Algorithm 2.1 and Proposition 2.3.

Since GPWs satisfy a local approximation of the PDE in the sense of a Taylor expansion, assumptions on the smoothness of the variable coefficient are stated in terms of continuous differentiability at the point $(x_c, y_c) \in \mathbb{R}^2$ and will of course depend on the desired order of approximation of the Taylor expansion.

2.1. Amplitude-based versus phase-based GPWs. The original idea behind GPW was to start by considering a classical PW, and its relation to the constant coefficient Helmholtz equation:

$$\begin{cases} W(x,y) = \exp\Big(\lambda_{10}(x-x_c) + \lambda_{01}(y-y_c)\Big), \quad \text{where} \quad (\lambda_{10},\lambda_{01}) \in \mathbb{C}^2, \\ (-\Delta - \kappa^2)W = 0 \Leftrightarrow \lambda_{10}^2 + \lambda_{01}^2 + \kappa^2 = 0. \end{cases}$$

In this case the ansatz chosen for W has two degrees of freedom, namely, $(\lambda_{10}, \lambda_{01}) \in \mathbb{C}^2$. The single constraint $\lambda_{10}^2 + \lambda_{01}^2 + \kappa^2 = 0$ on these two degrees of freedom is sufficient to ensure that the ansatz W solves the governing PDE. Moreover, by choosing $(\lambda_{10}, \lambda_{01}) = \mathrm{i}\kappa(\cos\theta, \sin\theta)$ for some real parameter θ, the constraint is satisfied, and any family of such functions, associated to any distinct values of $\theta \in [0, 2\pi)$, is linearly independent, so it is a basis for a finite-dimensional subspace of the Trefftz space for the Helmholtz equation. This subspace has been the most widely used to discretize Trefftz weak formulations.

In the case of a variable coefficient $\kappa(x,y)$, however, there is no general closed formula for exact solutions of the PDE. The original idea behind GPW might be summarized in two points:

- relaxing the Trefftz property, $(-\Delta - \kappa^2)W = 0$, into an approximation $(-\Delta - \kappa^2)G \approx 0$;
- choosing an ansatz with more degrees of freedom by adding for higher-order terms (HOT) to the phase of the classical PW: $G(x,y) = \exp(\lambda_{10}(x-x_c) + \lambda_{01}(y-y_c) + HOT)$.

The image of G by the differential operator, i.e., the function $(-\Delta - \kappa^2)G$, will not be zero but instead will locally approximate zero: It is the Taylor polynomial of this function that will be equal to zero. So the parameter q will refer to the order of the Taylor expansion approximation. Throughout the construction process, q's value remains unconstrained; it only comes into play to guarantee high-order approximation properties. A GPW was then initially defined in the vicinity of a point $(x_c, y_c) \in \mathbb{R}^2$ as

$$\begin{cases} G(x,y) := \exp P(x,y) \text{ with } P \in \mathbb{C}[X,Y] \text{ such that} \\ (-\Delta - \kappa^2(x,y))G(x,y) = O(|(x,y)-(x_c,y_c)|^q). \end{cases}$$

As a consequence, the construction of a GPW was equivalent to the following problem:

$$\begin{cases} \text{Find a polynomial } P \in \mathbb{C}[X,Y] \text{ such that} \\ G(x,y) := \exp P(x,y) \text{ satisfies } (-\Delta - \kappa^2(x,y))G(x,y) = O(|(x,y)-(x_c,y_c)|^q). \end{cases}$$

Moreover, since

$$(-\Delta - \kappa^2(x,y)) \exp P(x,y) = \left(-\Delta P(x,y) - \nabla P(x,y)\cdot\nabla P(x,y) - \kappa^2(x,y)\right) \exp P(x,y)$$

while $\exp P$ is bounded in the vicinity of (x_c, y_c), the construction of a GPW was also equivalent to the following problem:

$$\text{(2.1)} \qquad \begin{cases} \text{Find a polynomial } P \in \mathbb{C}[X,Y] \text{ such that} \\ -\Delta P(x,y) - \nabla P(x,y)\cdot\nabla P(x,y) - \kappa^2(x,y) = O(|(x,y)-(x_c,y_c)|^q) \\ G(x,y) := \exp P(x,y). \end{cases}$$

We are now interested in exploring a new type of quasi-Trefftz functions, this time choosing an ansatz with more degrees of freedom as higher-order terms (HOT) added to the amplitude rather than the phase of a classical PW: $G(x,y) = (1 + HOT)\exp(\lambda_{10}(x-x_c) + \lambda_{01}(y-y_c))$. An amplitude-based GPW is then defined in the vicinity of a point $(x_c, y_c) \in \mathbb{R}^2$ as

$$\begin{cases} G(x,y) := Q(x,y)\exp\left(\lambda_{10}(x-x_c) + \lambda_{01}(y-y_c)\right) \\ \text{ with } Q \in \mathbb{C}[X,Y], (\lambda_{10},\lambda_{01}) \in \mathbb{C}^2 \text{ such that} \\ (-\Delta - \kappa^2(x,y))G(x,y) = O(|(x,y)-(x_c,y_c)|^q). \end{cases}$$

As a consequence, the construction of an amplitude-based GPW is equivalent to the following problem:

$$\begin{cases} \text{Find } (Q,(\lambda_{10},\lambda_{01})) \in \mathbb{C}[X,Y]\times\mathbb{C}^2 \text{ such that} \\ G(x,y) := Q(x,y)\exp\left(\lambda_{10}(x-x_c) + \lambda_{01}(y-y_c)\right) \text{ satisfies} \\ (-\Delta - \kappa^2(x,y))G(x,y) = O(|(x,y)-(x_c,y_c)|^q). \end{cases}$$

Moreover, defining $\vec{d} := (\lambda_{10}, \lambda_{01})$ and $\widetilde{\kappa}^2 := \vec{d}\cdot\vec{d} + \kappa^2$, since

$$\begin{aligned} &(-\Delta - \kappa^2(x,y))Q(x,y)\exp(\lambda_{10}(x-x_c) + \lambda_{01}(y-y_c)) \\ &\quad = \left(-\Delta Q(x,y) - 2\nabla Q(x,y)\cdot\vec{d} - \widetilde{\kappa}^2(x,y)Q(x,y)\right)\exp\left(\lambda_{10}(x-x_c) + \lambda_{01}(y-y_c)\right) \end{aligned}$$

while $\exp(\lambda_{10}(x-x_c) + \lambda_{01}(y-y_c))$ is bounded in the vicinity of (x_c, y_c), the construction of an amplitude-based GPW is also equivalent to the following problem:

$$\text{(2.2)} \qquad \begin{cases} \text{Find } \left(Q,\vec{d}\right) \in \mathbb{C}[X,Y]\times\mathbb{C}^2 \text{ such that} \\ -\Delta Q(x,y) - 2\nabla Q(x,y)\cdot\vec{d} - \widetilde{\kappa}^2(x,y)Q(x,y) = O(|(x,y)-(x_c,y_c)|^q) \\ G(x,y) := Q(x,y)\exp\vec{d}\cdot\begin{pmatrix} x-x_c \\ y-y_c \end{pmatrix}. \end{cases}$$

Remark 2.1. Comparing (2.2) to (2.1), we observe that the new problem is not linear anymore.

However, there is no guarantee that this problem is well-posed; hence, subsections 2.2 and 2.3 turn to the study of the well-posedness of this problem. Next, given both the order q and the center (x_c, y_c) of the Taylor expansion, we will focus on developing an algorithm to construct such a GPW, which is equivalent to computing the coefficients of the polynomial Q so that the coefficients of the Taylor expansion of $-\Delta Q - 2\nabla Q \cdot \vec{d} - \widetilde{\kappa}^2 Q$ are zero; see subsection 2.4. Beyond the construction of one GPW, the goal is evidently to construct a family of linearly independent GPWs with good approximation properties, and these aspects will be addressed in section 3.

2.2. A linear system. In order to underline the structure of the problem at stake, we will now consider $\vec{d} := (\lambda_{10}, \lambda_{01}) \in \mathbb{C}^2$ to be fixed while we identify any polynomial $Q \in \mathbb{C}[X, Y]$ to the set of its coefficients $\{\mu_{i_x,i_y}, (i_x, i_y) \in \mathbb{N}_0^2, i_x + i_y \leq \deg Q\}$ via $Q(x, y) = \sum_{0 \leq i_x + i_y \leq \deg Q} \mu_{i_x,i_y} (x - x_c)^{i_x} (y - y_c)^{i_y}$. Later $\vec{d}$ will be used to define a family of GPWs. In this case, (2.2) boils down to a linear system:

- the unknowns are $\{\mu_{i_x,i_y}, (i_x, i_y) \in \mathbb{N}_0^2, i_x + i_y \leq \deg Q\}$;
- the equations are

$$\left\{ \partial_x^{j_x} \partial_y^{j_y} \left[-\Delta Q - 2\nabla Q \cdot \vec{d} - \widetilde{\kappa}^2 Q \right](x_c, y_c) = 0, (j_x, j_y) \in \mathbb{N}_0^2, j_x + j_y \leq q - 1 \right\}.$$

This linear system therefore has $N_{dof} := \frac{(\deg Q+1)(\deg Q+2)}{2}$ unknowns and $N_{eqn} := \frac{q(q+1)}{2}$ equations, and in order to build a solution to problem (2.2), for clarity we will consistently use indices (i_x, i_y) to refer to unknowns, while indices (j_x, j_y) will refer to equations.

Obviously, choosing the degree of Q, $\deg Q$, will affect the well-posedness of this system, as depending on $\deg Q + 1$ being smaller than, larger than, or equal to q, the system is, respectively, overdetermined, underdetermined, or square. Similarly to the choice of $\deg P$ for phase-based GPWs, we will choose $\deg Q = q + 1$ to exploit the structure provided by the Laplacian; see Remark 2.2. According to this choice, as $\partial_x^{j_x} \partial_y^{j_y} Q(x_c, y_c) = j_x! j_y! \mu_{j_x,j_y}$ for $j_x + j_y \leq q+1$ and zero otherwise, the linear system can then be written as

(2.3)
$$\begin{cases} \forall (j_x, j_y) \in \mathbb{N}_0^2, j_x + j_y \leq q - 1, \\ (j_x + 2)(j_x + 1)\mu_{j_x+2,j_y} + (j_y + 2)(j_y + 1)\mu_{j_x,j_y+2} \\ \qquad + 2(j_x + 1)\lambda_{10}\mu_{j_x+1,j_y} + 2(j_y + 1)\lambda_{01}\mu_{j_x,j_y+1} \\ \qquad + \displaystyle\sum_{k_x=0}^{j_x} \sum_{k_y=0}^{j_y} \frac{1}{(j_x - k_x)!(j_y - k_y)!} \partial_x^{j_x-k_x} \partial_y^{j_y-k_y} \widetilde{\kappa}^2(x_c, y_c) \mu_{k_x,k_y} = 0. \end{cases}$$

The goal in this section is then to leverage the structure of this linear system in order to build nontrivial solutions and hence build corresponding GPWs $Q(x, y) = \sum_{0 \leq i_x + i_y \leq q+1} \mu_{i_x,i_y} (x - x_c)^{i_x} (y - y_c)^{i_y}$.

Remark 2.2. Note that for any value of $\deg Q \geq q + 1$, the corresponding linear system could be written exactly as (2.3). Therefore, picking $\deg Q > q + 1$ would would increase the number of degrees of freedom without affecting the properties of the system.

On the other hand, if we had chosen $\deg Q < q+1$, at least the equations (j_x, j_y) such that $j_x + j_y = \deg Q - 1$ would read

(2.4)
$$\begin{cases} \forall (j_x, j_y) \in \mathbb{N}_0^2, j_x + j_y = \deg Q - 1, \\ 2(j_x+1)\lambda_{10}\mu_{j_x+1,j_y} + 2(j_y+1)\lambda_{01}\mu_{j_x,j_y+1} \\ \qquad + \displaystyle\sum_{k_x=0}^{j_x}\sum_{k_y=0}^{j_y} \frac{1}{(j_x-k_x)!(j_y-k_y)!}\partial_x^{j_x-k_x}\partial_y^{j_y-k_y}\widetilde{\kappa}^2(x_c,y_c)\mu_{k_x,k_y} = 0, \end{cases}$$

and the upcoming study of the system would not hold.

2.3. Layer structure and well-posedness. Even though there are no nonlinear terms in the system for amplitude-based GPWs, the structure of the system is analogous to that of the equivalent system for phase-based GPWs. We can gather unknowns μ_{i_x,i_y} according to the total degree of their monomial in Q, that is, according to $i_x + i_y$. Likewise, for each value of ℓ from 0 to $q-1$, we can gather equations according to $\ell = j_x + j_y$ into subsystems of $\ell + 1$ equations for the unknowns $\{\mu_{i_x,i_y}, (i_x, i_y) \in \mathbb{N}_0^2, i_x + i_y = \ell + 2\}$. Therefore, each subsystem can be rewritten as

(2.5)
$$\begin{cases} \forall (j_x, j_y) \in \mathbb{N}_0^2, j_x + j_y = \ell, \\ (j_x+2)(j_x+1)\mu_{j_x+2,j_y} + (j_y+2)(j_y+1)\mu_{j_x,j_y+2} \\ = -2(j_x+1)\lambda_{10}\mu_{j_x+1,j_y} - 2(j_y+1)\lambda_{01}\mu_{j_x,j_y+1} \\ \qquad - \displaystyle\sum_{k_x=0}^{j_x}\sum_{k_y=0}^{j_y} \frac{1}{(j_x-k_x)!(j_y-k_y)!}\partial_x^{j_x-k_x}\partial_y^{j_y-k_y}\widetilde{\kappa}^2(x_c,y_c)\mu_{k_x,k_y}. \end{cases}$$

The right-hand side of each equation in (2.5) only involves unknowns $\{\mu_{i_x,i_y}, (i_x, i_y) \in \mathbb{N}_0^2, i_x + i_y < \ell + 2\}$. The subsystems can then be considered sequentially for increasing values of ℓ, as the right-hand side would then be known from the previous layers $\widetilde{\ell} < \ell$. Each subsystem consists of $\ell + 1$ equations and has $\ell + 3$ unknowns, namely, $\{\mu_{i_x,i_y}, (i_x, i_y) \in \mathbb{N}_0^2, i_x + i_y = \ell + 2\}$; furthermore, the three unknowns $\{\mu_{0,0}, \mu_{1,0}, \mu_{0,1}\}$ only appear in the right-hand sides of these subsystems.

We can now justify why each subsystem has a solution thanks to a reformulation of each subsystem in terms of the partial operator Δ defined on the space of homogeneous polynomials. Denote by $\mathbb{A} = \mathbb{C}[X, Y]$ the space of complex polynomials in two variables and by $\mathbb{A}_d \subset \mathbb{A}$ the space of homogeneous polynomials of degree d. Since $\mathbb{A}_d = \mathrm{Span}\{X^i Y^{d-i}, 0 \le i \le d\}$, it is clear that $\dim \mathbb{A}_d = d + 1$. For a given level $\ell \in \mathbb{N}_0$, consider the Laplacian of a homogeneous polynomial of degree $\ell + 2$:

$$\Delta\left[\sum_{i=0}^{\ell+2} p_i X^i Y^{\ell+2-i}\right] = \sum_{i=0}^{\ell}\Big((i+2)(i+1)p_{i+2} + (\ell+2-i)(\ell+1-i)p_i\Big)X^i Y^{\ell-i}. \tag{2.6}$$

Then the restriction Δ_ℓ of the Laplacian operator to a space of homogeneous polynomials $\mathbb{A}_{\ell+2}$ is defined as

$$\begin{array}{rccc} \Delta_\ell : & \mathbb{A}_{\ell+2} & \to & \mathbb{A}_\ell \\ & P & \mapsto & \Delta P, \end{array}$$

and we are interested in the range of this linear operator. Indeed, the subsystem (2.5) has a solution if and only if the operator Δ_ℓ is surjective.

Let us focus first on the kernel of Δ_ℓ. It is clear from (2.6) that

$$\ker \Delta_\ell = \left\{ \sum_{i=0}^{\ell+2} p_i X^i Y^{\ell+2-i}, \{p_i\}_{0\le i\le \ell+2} \in \mathbb{C}^{\ell+3}, \right. \\ \left. (i+2)(i+1)p_{i+2} + (\ell+2-i)(\ell+1-i)p_i = 0 \text{ for } 0 \le i \le \ell \right\},$$

which is equivalent to

$$\ker \Delta_\ell = \operatorname{Span}\left\{ \sum_{i=0}^{\lfloor \frac{\ell+2}{2} \rfloor} (-1)^i \frac{(\ell+2)!}{(2i)!(\ell+2-2i)!} X^{2i} Y^{\ell+2-2i}, \right. \\ \left. \sum_{i=1}^{\lfloor \frac{\ell+3}{2} \rfloor} (-1)^{i-1} \frac{(\ell+1)!}{(2i-1)!(\ell+3-2i)!} X^{2i-1} Y^{\ell+3-2i} \right\}.$$

So $\dim(\ker \Delta_\ell) = 2$. As a consequence, since $\dim \mathbb{A}_{\ell+2} = \ell+3$ while $\dim \mathbb{A}_\ell = \ell+1$, the rank-nullity theorem shows that the operator Δ_ℓ is full-rank.

This then shows that each subsystem (2.5) has a solution.

In turn, thanks to the layer structure of system (2.3), we have then proved that the system has a solution. With $N_{dof} = \frac{(q+2)(q+3)}{2}$ unknowns and $N_{eqn} = \frac{q(q+1)}{2}$ equations, system (2.3) can be solved using $N_{dof} - N_{eqn} = 2q+3$ appropriate additional constraints. First, the three unknowns $\{\mu_{0,0}, \mu_{1,0}, \mu_{0,1}\}$ can be fixed, as we have already noted that they appear only on right-hand sides of the subsystems. Next, for increasing values of ℓ, each of the q subsystems is triangular. This can be evidenced via numbering of the equations with increasing values of j_x and numbering of the unknowns with increasing values of i_x. It is then clear that by adding the two additional constraints corresponding to fixing $\mu_{0,\ell+2}$ and $\mu_{1,\ell+1}$, we obtain a well-posed problem for each layer ℓ. The three initial constraints plus the two constraints per subsystem altogether form $2q+3$ additional constraints, and there is a unique solution to (2.3) augmented by these constraints.

In summary, the problem of constructing a GPW can then be reformulated into a well-posed problem:

$$\left\{ \begin{array}{ll} \text{Fix } \{\mu_{0,0}, \mu_{1,0}, \mu_{0,1}\}, & \\ \forall \ell \in [\![0, q-1]\!] & \text{Fix } \mu_{0,\ell+2}, \mu_{1,\ell+1}, \\ & \text{Solve (2.5).} \end{array} \right. \tag{2.7}$$

2.4. Construction of solutions. The well-posedness of (2.7)—and therefore the existence of a solution to (2.2)—is independent of the values chosen to fix the additional constraints. However, in order for the resulting GPW to have useful approximation properties, we make the following choices. To obtain a GPW of the form $G(x,y) = (1+HOT)\exp(\lambda_{10}(x-x_c) + \lambda_{01}(y-y_c))$ as announced, we choose to fix $\mu_{0,0} = 1$, as any choice of a nonzero value independent of $(\lambda_{10}, \lambda_{01})$ would not affect any of the results that follow. Next we choose for simplicity to fix $\mu_{1,0} = 0$ and $\mu_{0,1} = 0$ and, for increasing values of ℓ, $\mu_{0,\ell+2} = 0$ and $\mu_{1,\ell+1} = 0$. The unique solution to the reformulated problem (2.7) can be constructed thanks to the following algorithm.

Algorithm 2.1 Construction of an amplitude-based GPW for the Helmholtz equation.

Given $\vec{d} = (\lambda_{1,0}, \lambda_{0,1}) \in \mathbb{C}^2$ and $(x_c, y_c) \in \mathbb{R}^2$
Fix $\mu_{0,0} = 1$ as well as $(\mu_{1,0}, \mu_{0,1}) = (0, 0)$
for $\ell \leftarrow 0,\ q-1$ **do**
  Fix $\mu_{0,\ell+2} = 0$ and $\mu_{1,\ell+1} = 0$
  for $j_x \leftarrow 0,\ \ell$ **do**
    $\mathsf{RHS} := -2(j_x+1)\lambda_{10}\mu_{j_x+1,\ell-j_x} - 2(\ell-j_x+1)\lambda_{01}\mu_{j_x,\ell-j_x+1}$
    $$- \sum_{k_x=0}^{j_x} \sum_{k_y=0}^{\ell-j_x} \frac{1}{(j_x-k_x)!(\ell-j_x-k_y)!} \partial_x^{j_x-k_x} \partial_y^{\ell-j_x-k_y} \widetilde{\kappa}^2(x_c, y_c) \mu_{k_x,k_y}$$
    $$\mu_{j_x+2,\ell-j_x} := \frac{1}{(j_x+2)(j_x+1)} \Big(\mathsf{RHS} - (\ell-j_x+2)(\ell-j_x+1)\mu_{j_x,\ell-j_x+2} \Big)$$
$$Q(x,y) \leftarrow \sum_{0 \le i_x+i_y \le q+1} \mu_{i_x,i_y} (x-x_c)^{i_x} (y-y_c)^{i_y}$$
$$G_{\vec{d}}(x,y) \leftarrow Q(x,y) \exp\Big(\lambda_{10}(x-x_c) + \lambda_{01}(y-y_c) \Big)$$

Thanks to the triangular structure of subsystems (2.5), which hence are solved by substitution, the computational cost of Algorithm 2.1 is simply linear with respect to the number of polynomial coefficients of Q. Moreover, even though it does not affect the linear rate, the choice to fix to zero $\mu_{1,0}$, $\mu_{0,1}$, $\mu_{0,\ell+2}$, and $\mu_{1,\ell+1}$ further reduces the computational cost.

From the derivation of this algorithm, we immediately obtain the following result.

PROPOSITION 2.3. *Assume that the parameter $q \in \mathbb{N}$ is given and that κ^2 is a function of class $\mathcal{C}^{q-1}$ in a neighborhood of a given point $(x_c, y_c) \in \mathbb{R}^2$. An amplitude-based GPW $G_{\vec{d}}$ constructed from Algorithm 2.1 for any $\vec{d} = (\lambda_{1,0}, \lambda_{0,1}) \in \mathbb{C}^2$ satisfies the local approximation property in the vicinity of (x_c, y_c)*

$$(-\Delta - \kappa^2(x,y)) G_{\vec{d}}(x,y) = O(|(x,y) - (x_c,y_c)|^q). \tag{2.8}$$

This property actually holds not only independently of the choice of $\vec{d} = (\lambda_{1,0}, \lambda_{0,1}) \in \mathbb{C}^2$ but also independently of the other values fixed in the algorithm since (2.8) holds simply by construction. The particular choice of $\vec{d}$ that will now be proposed is, however, the cornerstone of the proofs leading to approximation properties of the new GPWs.

The choice of the $\vec{d}$, related to the direction of propagation of PWs, is referred to as normalization. In order to build a family of amplitude-based GPWs, we simply mimic classical PW by choosing $(\lambda_{1,0}, \lambda_{0,1}) = \sqrt{-\kappa^2(x_c, y_c)}(\cos\theta, \sin\theta)$ and construct the corresponding GPW for any value of θ.

DEFINITION 2.4. *Assume that the parameter $q \in \mathbb{N}$ is given and that κ^2 is a function of class $\mathcal{C}^{q-1}$ in a neighborhood of a given point $(x_c, y_c) \in \mathbb{R}^2$ with $\kappa^2(x_c, y_c) \neq 0$. For any set of p angles $\mathrm{A} := \{\theta_k \in [0, 2\pi); 1 \le k \le p\}$, for each angle we define the associated $\vec{d}_k := \sqrt{-\kappa^2(x_c, y_c)}(\cos\theta_k, \sin\theta_k)$ and GPW $G_k := G_{\vec{d}_k}$. We will denote the corresponding set of GPW functions $\mathcal{B}_{\mathrm{A},q} := \{G_k; \theta_k \in \mathrm{A}, 1 \le k \le p\}$.*

As a by-product of the proof of approximation properties, we will prove the linear independence of this set of GPWs under the mere condition that A is a set of distinct angles.

Obviously, if $\kappa^2(x_c, y_c) = 0$, then all G_k are identical because then all $\vec{d}_k$ are the same. However, in this case, the normalization could be chosen differently, and a family of linearly independent GPWs would still be constructed. We will not focus on this aspect here, but a study of approximation properties with a normalization independent of κ was presented in [24] for phase-based GPWs, and the same approach is applicable to the amplitude-based GPWs. Approximation properties of these new functions indeed hold even with a κ-independent normalization, so the new GPWs could be used (like the phase-based ones) in domains including points at which κ vanishes. For instance, for applications to plasma physics where a cutoff be defined as a line along which κ vanishes, it would be possible to change the normalization in a neighborhood of a cutoff.

3. Best approximation properties. We will now follow the road map proposed in [24] to study approximation properties of GPWs, adapting each step to this new framework of amplitude-based GPWs. At each step before the last one, we will emphasize interpretation of the central idea in this new framework and state the desired properties in lemmas, finally resulting in the approximation properties summarized in Theorem 3.7. The theorem's proof relies on constructing a GPW approximation to any smooth solution of the PDE by matching their Taylor expansions. A natural element coming into play when matching the Taylor expansion of a linear combination of p functions $\{f_i, i \in \mathbb{N}, i \leq p\}$ is the matrix $\mathsf{M}_n \in \mathbb{C}^{(n+1)(n+2)/2 \times p}$ built columnwise from the Taylor expansion coefficients of each function, and we will use the notation

(3.1)
$$\forall (j_x, j_y) \in (\mathbb{N}_0)^2, j_x + j_y \leq n,\ (\mathsf{M}_n)_{\frac{(j_x+j_y)(j_x+j_y+1)}{2}+j_y+1,k} := \partial_x^{j_x}\partial_y^{j_y} f_k(x_c, y_c)/(j_x! j_y!).$$

Step 1. In this first step, we seek common properties of the unknowns that are computed inside the nested loop in Algorithm 2.1. In the phase-based context, this amounts to studying the coefficients of highest-degree terms in the polynomial P with respect to the only nonzero fixed unknowns in the construction algorithm, namely, $(\lambda_{1,0}, \lambda_{0,1})$. So it involves exclusively the phase polynomial. By contrast, here it will couple the amplitude polynomial with the phase term. The coefficients of highest-degree terms in the polynomial Q cannot possibly be expressed exclusively in terms of the only nonzero fixed unknown in the construction algorithm, which would be $\mu_{0,0}$. Instead, as appears clearly from the computation of RHS in the algorithm, they are intrinsically coupled to the fixed terms from the phase, namely, $(\lambda_{1,0}, \lambda_{0,1})$. Despite this practical difference, the fundamental idea remains the same independently of the choice of ansatz for the GPW: Both ansatz are designed starting from a classical PW, so the new unknowns introduced in each ansatz are studied with respect to the parameters defining a classical PW, namely, $(\lambda_{1,0}, \lambda_{0,1})$.

The relation between $\{\mu_{i_x,i_y}, i_x \geq 2\}$ and $(\lambda_{0,1}, \lambda_{1,0})$ is polynomial, and each of these μ_{i_x,i_y} has a particular degree as a polynomial in $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$.

LEMMA 3.1. *Assume that the parameter $q \in \mathbb{N}$ is given and that κ^2 is a function of class $\mathcal{C}^{q-1}$ in a neighborhood of a given point $(x_c, y_c) \in \mathbb{R}^2$ with $\kappa^2(x_c, y_c) \neq 0$. Consider the set of unknowns $\{\mu_{i_x,i_y}, (i_x, i_y) \in \mathbb{N}_0^2, i_x + i_y \leq q + 1\}$ constructed in Algorithm* 2.1*, under the assumption (inspired by classical PWs) that the quantity $(\lambda_{1,0})^2 + (\lambda_{0,1})^2$ is equal to $-\kappa^2(x_c, y_c)$. While $\mu_{0,0} = 1$, each μ_{i_x,i_y} for $i_x + i_y \geq 1$ can be expressed as a polynomial in $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ of degree at most equal to $i_x + i_y - 2$.*

In order for the following proof to hold, it is sufficient for the quantity $(\lambda_{1,0})^2 + (\lambda_{0,1})^2$ to have a fixed value in $v \in \mathbb{C}$, as explained in [20]. In other words, instead

of considering elements of the polynomial ring $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$, we consider the quotient ring $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]/((\lambda_{1,0})^2 + (\lambda_{0,1})^2 - v)$. This explains the phrasing of the lemma: μ_{i_x,i_y} *can be expressed as* a polynomial with a certain degree—as opposed to *has* a certain degree.

Proof. In view of Algorithm 2.1's formula 2, the result is clear for $(\mu_{1,0}, \mu_{0,1})$. In view of Algorithm 2.1's formula 8, we will further proceed by nested induction on ℓ and j_x. Following the layer structure of our problem, we start by the induction with respect to ℓ.

For $\ell = 0$, $\mu_{0,2}$ and $\mu_{1,1}$ are both fixed to zero, so they clearly are polynomials in $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ of degree at most equal to 0. Then, since $\mu_{0,0} = 1$, the last μ_{i_x,i_y} with $i_x + i_y = 2$ can be written as

$$\mu_{2,0} = \frac{1}{2}\left(-2\lambda_{1,0}\mu_{1,0} - 2\lambda_{0,1}\mu_{0,1} - \widetilde{\kappa}^2(x_c, y_c)\right). \tag{3.2}$$

From the definition of $\widetilde{\kappa}^2$ and the assumption on $(\lambda_{1,0}, \lambda_{0,1})$, it is clear that $\widetilde{\kappa}^2(x_c, y_c) = 0$. Since moreover $\mu_{1,0} = \mu_{0,1} = 0$, we obtain that $\mu_{2,0} = 0$, and the result is proved for $\ell = 0$.

Given $\ell \in \mathbb{N}_0$, $\ell < q - 1$, assume that each μ_{i_x,i_y} for $1 \leq i_x + i_y \leq \ell + 2$ is a polynomial in $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ of degree at most equal to $i_x + i_y - 2$. Since $\mu_{0,\ell+3}$ and $\mu_{1,\ell+2}$ are both fixed to zero, they clearly are polynomials in $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ of degree at most equal to $\ell + 1$. We then want to prove the result for all $\mu_{j_x+2,\ell+1-j_x}$ with $0 \leq j_x \leq \ell + 1$, and we will naturally proceed by induction on j_x. For $j_x = 0$, since $\mu_{0,\ell+3} = 0$, Algorithm 2.1's formula 8 shows that $\mu_{2,\ell+1}$ as a polynomial in $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ will simply be a multiple of RHS, while, thanks to the definition of $\widetilde{\kappa}^2$ with $\mu_{0,\ell+1} = \mu_{1,\ell+1} = \mu_{0,\ell+2} = 0$, we have

$$\mathsf{RHS} = -\sum_{k_y=0}^{\ell+1} \frac{1}{(\ell + 1 - k_y)!} \partial_y^{\ell+1-k_y} \kappa^2(x_c, y_c) \mu_{0,k_y}.$$

So the induction hypothesis for ℓ together with $\mu_{0,0} = 1$ show that the result holds for $\mu_{2,\ell+1}$. Given $j_x \in \mathbb{N}_0$, $j_x < \ell+1$, we now assume that each μ_{i_x,i_y} for $i_x+i_y = \ell+3$ as well as $i_x \leq j_x+2$ is a polynomial in $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ of degree at most equal to i_x+i_y-2. The degree of $\mu_{j_x+3,\ell-j_x}$ at most equal to the maximum of

- the degree of $\mu_{j_x+1,\ell-j_x+2}$, at most equal to $\ell+1$ from the induction hypothesis for j_x;
- the degree of $\lambda_{10}\mu_{j_x+2,\ell-j_x}$, at most equal to $\ell + 1$ from the induction hypothesis for ℓ;
- the degree of $\lambda_{01}\mu_{j_x+1,\ell-j_x+1}$, at most equal to $\ell + 1$ from the induction hypothesis for ℓ;
- the degree of $((\lambda_{1,0})^2 + (\lambda_{0,1})^2)\mu_{j_x+1,\ell-j_x}$, at most equal to $\ell - 1$ from the induction hypothesis for ℓ;
- the degree of $\partial_x^{j_x+1}\partial_y^{\ell-j_x}\kappa^2(x_c, y_c)\mu_{0,0}$, at most equal to 0 since $\mu_{0,0} = 1$;
- the degree of $\partial_x^{j_x+1-k_x}\partial_y^{\ell-j_x-k_y}\kappa^2(x_c, y_c)\mu_{k_x,k_y}$ for $0 \leq k_x \leq j_x + 1$ and $0 \leq k_y \leq \ell - j_x$, at most equal to $\ell - 1$ from the induction hypothesis for ℓ.

So the degree of $\mu_{j_x+3,\ell-j_x}$ as a polynomial in $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ is as expected at most equal to $\ell + 1$, which concludes the proof. ∎

Step 2. This step focuses on identifying a reference set of functions, simpler to study than the GPWs. From the choice of ansatz for amplitude-based GPWs, it is again clear that the reference case of classical PWs will play a fundamental role in the study of approximation properties, similarly to the phase-based GPW case [20].

The normalization introduced before Definition 2.4 is used to define the reference space of classical PWs corresponding to the GPWs as follows.

DEFINITION 3.2. *Assume that the point* $(x_c, y_c) \in \mathbb{R}^2$ *as well as the parameter* p *are given. For any set of* p *angles* $\mathrm{A} := \{\theta_k \in [0, 2\pi); 1 \leq k \leq p\}$*, for each angle we define the associated* $\vec{d_k} := \sqrt{-\kappa^2(x_c, y_c)}(\cos\theta_k, \sin\theta_k)$ *and classical PW* $H_k := \exp(\vec{d_k} \cdot (\cdot - (x_c, y_c)))$*. We will denote the corresponding set of PW functions* $\mathcal{B}_{\mathrm{A}}^{ref} := \{H_k; \theta_k \in \mathrm{A}, 1 \leq k \leq p\}$.

Even though as discussed earlier we will assume that $\kappa^2(x_c, y_c) \neq 0$, the following study will hold independently of the sign of $\kappa^2(x_c, y_c)$; however, the H_k functions are oscillating PWs only if $\kappa^2(x_c, y_c) > 0$, while they are exponentially decaying (or increasing) functions if $\kappa^2(x_c, y_c) < 0$.

Step 3. Let us turn to the study of approximation properties of this reference case. Here the reference case is the same for the amplitude-based GPWs as it was for the phase-based GPWs. The approximation properties of the reference space $\mathcal{B}_{\mathrm{A}}^{ref}$ were proved in [7], even though they were not stated as a stand-alone result. The precise result that we will use in the following is only a part of that proof. It is simply presented here as a reminder.

LEMMA 3.3. *Assume that the point* $(x_c, y_c) \in \mathbb{R}^2$ *as well as the parameters* $n \in \mathbb{N}$ *are given. For any set of* $p = 2n + 1$ *distinct angles* $\mathrm{A} := \{\theta_k \in [0, 2\pi); 1 \leq k \leq p\}$*, we consider the matrix* (3.1) *for the reference set of PW* $\mathcal{B}_{\mathrm{A}}^{ref}$*, denoted* M_n^C *as in* [20]*. Then, as long as* $\kappa^2(x_c, y_c) \neq 0$*, the rank of this matrix is* $rk(\mathsf{M}_n^C) = 2n + 1$.

See [24, section 4.1] for a comment on the need for p to be at least equal to $2n+1$ to guarantee a rank equal to $2n+1$ in relation to properties of trigonometric functions.

Step 4. In order to relate the new GPW case to the reference case, we introduce the matrix (3.1) for the new GPW set $\mathcal{B}_{\mathrm{A},q}$, denoted M_n^G. To study its relation to the reference matrix M_n^C, we will first express each of its entries: the new basis functions' derivatives evaluated at (x_c, y_c) in terms of the reference basis functions' derivatives evaluated at (x_c, y_c).

LEMMA 3.4. *Assume that the point* $(x_c, y_c) \in \mathbb{R}^2$ *as well as the parameters* $(p, q) \in \mathbb{N}^2$ *are given and that* κ^2 *is a function of class* $\mathcal{C}^{q-1}$ *in a neighborhood of* (x_c, y_c)*. For any set of* p *distinct angles* $\mathrm{A} := \{\theta_k \in [0, 2\pi); 1 \leq k \leq p\}$*, ranked in a given order, we consider the reference and GPW bases* $\mathcal{B}_{\mathrm{A}}^{ref}$ *and* $\mathcal{B}_{\mathrm{A},q}$*, with their elements ranked in the same order. For* $(j_x, j_y) \in \mathbb{N}_0^2$ *such that* $j_x + j_y \leq q + 1$*, there exists a polynomial* $R_{(j_x,j_y)} \in \mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ *with* $\deg R_{(j_x,j_y)} < j_x + j_y$ *such that for all* $k \in \mathbb{N}$ *with* $k \leq p$*, we have*

$$\frac{1}{j_x! j_y!}\partial_x^{j_x}\partial_y^{j_y} G_k(x_c, y_c) = \frac{1}{j_x! j_y!}\partial_x^{j_x}\partial_y^{j_y} H_k(x_c, y_c) + R_{(j_x,j_y)}(\lambda_{1,0}, \lambda_{0,1}).$$

Proof. For the vector $\vec{d_k} := \sqrt{-\kappa^2(x_c, y_c)}(\cos\theta_k, \sin\theta_k)$ associated to any angle $\theta_k \in \mathrm{A}$, the reference and new basis functions are, respectively, H_k and G_k. We first notice that $G_k = Q \cdot H_k$, where Q is the polynomial constructed via Algorithm 2.1. Then, since $\partial_x^{i_x}\partial_y^{i_y} Q(x_c, y_c) = i_x! i_y! \mu_{i_x,i_y}$ for any $(i_x, i_y) \in \mathbb{N}_0^2$ such that $i_x + i_y \leq q+1$, the product rule shows that as long as $j_x + j_y \leq q + 1$, we have

$$\frac{1}{j_x! j_y!}\partial_x^{j_x}\partial_y^{j_y} G_k(x_c, y_c) = \sum_{i_x=0}^{j_x}\sum_{i_y=0}^{j_y} \frac{1}{(j_x - i_x)!(j_y - i_y)!}\mu_{i_x,i_y} (\lambda_{1,0})^{j_x - i_x} (\lambda_{0,1})^{j_y - i_y}.$$

In view of the normalization chosen in Algorithm 2.1, it gives, for $(j_x, j_y) \in \mathbb{N}_0^2$ such that $j_x + j_y \leq q+1$,

$$\begin{cases} \dfrac{1}{j_y!}\partial_y^{j_y} G_k(x_c,y_c) = \dfrac{1}{j_y!}\left(\lambda_{0,1}\right)^{j_y} \text{ if } j_x = 0, \\ \dfrac{1}{j_y!}\partial_x\partial_y^{j_y} G_k(x_c,y_c) = \dfrac{1}{j_y!}\left(\lambda_{1,0}\right)^{1}\left(\lambda_{0,1}\right)^{j_y} \text{ if } j_x = 1, \\ \dfrac{1}{j_x!j_y!}\partial_x^{j_x}\partial_y^{j_y} G_k(x_c,y_c) \\ \quad = \dfrac{1}{j_x!j_y!}\left(\lambda_{1,0}\right)^{j_x}\left(\lambda_{0,1}\right)^{j_y} + \displaystyle\sum_{i_x=2}^{j_x}\sum_{i_y=0}^{j_y} \frac{\left(\lambda_{0,1}\right)^{j_y-i_y}\left(\lambda_{1,0}\right)^{j_x-i_x}}{(j_x-i_x)!(j_y-i_y)!}\mu_{i_x,i_y} \text{ if } j_x \geq 2. \end{cases}$$

On the other hand, $\partial_x^{j_x}\partial_y^{j_y} H_k(x_c,y_c) = \left(\lambda_{0,1}\right)^{j_y}\left(\lambda_{1,0}\right)^{j_x}$ for any $(j_x,j_y) \in \mathbb{N}_0^2$. So the result is proved for $j_x = 0$ and $j_x = 1$, while for $j_x \geq 2$ the result is a direct consequence of Lemma 3.1. ∎

LEMMA 3.5. *Assume that the point $(x_c,y_c) \in \mathbb{R}^2$ as well as the parameters $n \in \mathbb{N}$ are given and that κ^2 is a function of class $\mathcal{C}^{q-1}$ in a neighborhood of (x_c,y_c). For any set of $p = 2n+1$ distinct angles $\mathrm{A} := \{\theta_k \in [0,2\pi); 1 \leq k \leq p\}$, ranked in a given order, we consider the reference and GPW matrices, M_n^C and M_n^G, respectively, defined for the reference and the GPW bases $\mathcal{B}_{\mathrm{A}}^{ref}$ and $\mathcal{B}_{\mathrm{A},q}$ for any $q \geq \max(n-1,1)$, with their elements ranked in the same order. There exists a nonsingular matrix $\mathsf{L}_n \in \mathbb{C}^{(n+1)(n+2)/2 \times (n+1)(n+2)/2}$ such that*

$$\mathsf{M}_n^G = \mathsf{L}_n \mathsf{M}_n^C.$$

As a consequence, as long as $\kappa^2(x_c,y_c) \neq 0$, $rk(\mathsf{M}_n^G) = 2n+1$.

Proof. The choice $q \geq \max(n-1,1)$ guarantees that $n \leq q+1$; therefore, the result of Lemma 3.4 holds in particular for all entries of the matrices, and it can be restated as follows: There exists a complex polynomial in two variables $R_{(j_x,j_y)}$ with $\deg R_{(j_x,j_y)} < j_x + j_y$ such that for all $k \in \mathbb{N}$ with $k \leq p$, we have

$$\begin{aligned} &\left(\mathsf{M}_n^G\right)_{\frac{(j_x+j_y)(j_x+j_y+1)}{2}+j_y+1,k} \\ &\quad = \left(\mathsf{M}_n^C\right)_{\frac{(j_x+j_y)(j_x+j_y+1)}{2}+j_y+1,k} + R_{(j_x,j_y)}\left(\left(\mathsf{M}_n^C\right)_{2,k}, \left(\mathsf{M}_n^C\right)_{3,k}\right). \end{aligned} \tag{3.3}$$

Moreover, from the definition of the reference basis functions H_k, we have

$$\partial_x^{j_x}\partial_y^{j_y} H_k(x_c,y_c) = \left(\lambda_{1,0}\right)^{j_x}\left(\lambda_{0,1}\right)^{j_y} = \partial_x H_k(x_c,y_c)^{j_x}\partial_y H_k(x_c,y_c)^{j_y},$$

where $(\lambda_{1,0},\lambda_{0,1}) = \sqrt{-\kappa^2(x_c,y_c)}(\cos\theta_k, \sin\theta_k)$. Hence, from the definition of the reference matrix M_n^C, we can identify

$$\left(\left(\mathsf{M}_n^C\right)_{2,k}\right)^{j_x}\left(\left(\mathsf{M}_n^C\right)_{3,k}\right)^{j_y} = \left(\mathsf{M}_n^C\right)_{\frac{(j_x+j_y)(j_x+j_y+1)}{2}+j_y+1,k}.$$

It is now clear that (3.3) is precisely stating that M_n^G's row number $\frac{(j_x+j_y)(j_x+j_y+1)}{2} + j_y + 1$ can be written as a linear combination of M_n^C's rows; more precisely, it can be written as the sum of

- once M_n^C's row number $\frac{(j_x+j_y)(j_x+j_y+1)}{2}+j_y+1$;
- a linear combination of M_n^C's row of index at most equal to $\frac{(j_x+j_y)(j_x+j_y+1)}{2}$.

Finally, this is equivalent to the existence of a lower unitriangular matrix L_n such that $\mathsf{M}_n^G = \mathsf{L}_n\mathsf{M}_n^C$.

As a consequence, $rk(\mathsf{M}_n^G) = rk(\mathsf{M}_n^C)$. Hence, from Lemma 3.3, if $\kappa^2(x_c, y_c) \neq 0$, then $rk(\mathsf{M}_n^G) = 2n+1$. ∎

Step 5. Finally, we can pull the pieces together to study the approximation properties of the space spanned by the new GPWs.

DEFINITION 3.6. *Assume that the point* $(x_c, y_c) \in \mathbb{R}^2$ *as well as the parameters* $(p, q) \in \mathbb{N}^2$ *are given and that* κ^2 *is a function of class* $\mathcal{C}^{q-1}$ *in a neighborhood of* (x_c, y_c). *For any set of* p *distinct angles* $\mathrm{A} := \{\theta_k \in [0, 2\pi); 1 \leq k \leq p\}$, *ranked in a given order, we consider the GPW basis* $\mathcal{B}_{\mathrm{A},q}$ *for any* $q \geq \max(n-1, 1)$. *The complex vector space spanned by* $\mathcal{B}_{\mathrm{A},q}$ *will be denoted* $\mathbb{V}^{\mathrm{A},q}_{(x_c,y_c)}$.

In order to obtain the desired approximation properties, it is therefore sufficient to pick the approximation parameter q to be equal to $\max(n-1, 1)$. Picking a higher value would guarantee the same properties but would result in an unnecessary increase of the GPW construction's computational cost.

THEOREM 3.7. *Assume that the point* $(x_c, y_c) \in \mathbb{R}^2$ *as well as the parameters* $n \in \mathbb{N}$ *are given and that* κ^2 *is a function of class* $\mathcal{C}^n$ *in a neighborhood of* (x_c, y_c), *with* $\kappa^2(x_c, y_c) \neq 0$. *For any set of* $p = 2n+1$ *distinct angles* $\mathrm{A} := \{\theta_k \in [0, 2\pi); 1 \leq k \leq p\}$, *we consider the GPW space* $\mathbb{V}^{\mathrm{A},\max(n-1,1)}_{(x_c,y_c)}$.

For any solution u *of the PDE* (1.1) *which is of class* $\mathcal{C}^n$ *at* (x_c, y_c), *there exists a GPW function* $u_a \in \mathbb{V}^{\mathrm{A},\max(n-1,1)}_{(x_c,y_c)}$ *and a constant* C *such that in a neighborhood of* (x_c, y_c),

$$\begin{cases} |(u-u_a)(x,y)| \leq C|(x,y)-(x_c,y_c)|^{n+1}, \\ |(\nabla u - \nabla u_a)(x,y)| \leq C|(x,y)-(x_c,y_c)|^{n}. \end{cases} \tag{3.4}$$

Thanks to our preliminary lemmas, the proof is identical to the one from [20] for phase-based GPWs. We repeat the proof here for the sake of completeness. Even though the preliminary steps were studied under the assumption that $\kappa^2(x_c, y_c) \neq 0$, as it was mentioned in section 2, amplitude-based GPWs with an appropriate normalization enjoy the same approximation properties even if $\kappa^2(x_c, y_c) = 0$.

Proof. It is sufficient for the Taylor expansion of a function $u_a \in \mathbb{V}^{\mathrm{A},\max(n-1,1)}_{(x_c,y_c)}$ to match that of the solution u to prove the theorem.

Any element of $\mathbb{V}^{\mathrm{A},\max(n-1,1)}_{(x_c,y_c)}$ can be written as $\sum_{k=1}^{2n+1} X_k G_k$, and its Taylor expansion matches that of the solution u if and only if

$$\mathsf{M}_n^G \mathsf{X} = \mathsf{U}, \tag{3.5}$$

where the kth entry of $\mathsf{X} \in \mathbb{C}^{2n+1}$ is the coefficient X_k, while for all $(j_x, j_y) \in (\mathbb{N}_0)^2$, the $\frac{(j_x+j_y)(j_x+j_y+1)}{2}+j_y+1$th entry of $\mathsf{U} \in \mathbb{C}^{\frac{(n+1)(n+2)}{2}}$ is the Taylor expansion coefficient $\partial_x^{j_x}\partial_y^{j_y}u(x_c, y_c)/(j_x!j_y!)$. From Lemma 3.5 the matrix $\mathsf{M}_n^G \in \mathbb{C}^{\frac{(n+1)(n+2)}{2}\times(2n+1)}$

has maximal rank $2n+1$. Moreover, the range of M_n^G can be identified as

$$\mathfrak{K} := \left\{ (C_{j_x,j_y}) \in \mathbb{C}^{\frac{(n+1)(n+2)}{2}}, \forall (j_x,j_y) \in \mathbb{N}^2, j_x+j_y \le n-2, \right.$$
$$(j_x+1)(j_x+2)C_{j_x+2,j_y} + (j_y+1)(j_y+2)C_{j_x,j_y+2}$$
$$\left. = -\sum_{i_x=0}^{j_x}\sum_{i_y=0}^{j_y} \frac{\partial_x^{i_x}\partial_y^{i_y}\kappa^2(x_c,y_c)}{i_x!i_y!} C_{j_x-i_x,j_y-i_y} \right\},$$

and the right-hand side U of (3.5) clearly belongs to $\mathfrak{K}$ as the function u solves the PDE (1.1).

As a result, there exists a solution X to the linear system (3.5), and the corresponding GPW function $u_a := \sum_{k=1}^{2n+1} X_k G_k \in \mathbb{V}_{(x_c,y_c)}^{\mathrm{A},\max(n-1,1)}$ is guaranteed to have the same Taylor expansion as u up to order n. In other words, we have $(u-u_a)(x,y) = O(|(x,y)-(x_c,y_c)|^{n+1})$, and this clearly implies the result (3.4). ∎

4. Extension beyond the Helmholtz equation. It is a natural question to consider how can this work, developed in the previous sections for the Helmholtz operator $-\Delta - \kappa^2(x,y)$, be extended to other linear partial differential operators. This was the goal of [24] for phase-based GPWs, which considered operators of order $M \ge 2$, in two dimensions, of the form

$$\mathcal{L}_{M,\alpha} := \sum_{\ell=0}^{M}\sum_{k=0}^{\ell} \alpha_{k,\ell-k}(x,y)\,\partial_x^k\partial_y^{\ell-k}, \tag{4.1}$$

where $\alpha = \{\alpha_{k,\ell-k}, (k,\ell) \in \mathbb{N}^2, 0 \le k \le \ell \le M\}$ represents the set of complex-valued coefficients. The situation is similar for amplitude-based GPWs. On the one hand, the construction process of amplitude-based GPWs proposed earlier can be extended as-is to a large family of linear PDEs of the form (4.1) under a simple assumption $\alpha_{M,0}(x_c,y_c) \neq 0$. This is simply because the key to the construction algorithm does not lie in any of the normalization choices (as the values chosen for the normalization obviously do not affect the structure of the algorithm) but rather in the explicit formula 8 in Algorithm 2.1. On the other hand, the approximation properties strongly rely on the normalization, as we have seen that the fact for $(\lambda_{1,0})^2 + (\lambda_{0,1})^2$ to be constant was fundamental as early as in Step 1.

In this section, instead of considering general operators (4.1) as in [24], we will focus on operators of second order allowing for anisotropy in the first- and second-order terms, written under the form

$$\mathcal{L} := \nabla \cdot A(x,y)\nabla + V(x,y)\cdot\nabla + s(x,y), \tag{4.2}$$

where A a matrix-valued function, V a vector-valued function, and s a scalar-valued function, under the assumption that the matrix $A(x_c,y_c)$ is real symmetric with nonzero eigenvalues. Our work is motivated by one such equation, the convected Helmholtz equation, of interest to the aeroacoustics industry [4, 5]. Some recent work on anisotropic wave propagation includes asymptotic techniques [25], finite-difference aspects [40], from the scattering point of view both finite element or integral equation aspects [10, 13, 1, 42] boundary element aspects [2], and combined spectral and wave aspects [41]. However, there is little work on anisotropic media in the field of Trefftz methods, but anisotropic Maxwell problems have been investigated in [30] in

the constant coefficient case, with anisotropy restricted to diagonal permittivity and permeability matrices, as well as in [19] in the variable coefficient case, with a real permeability and a complex permittivity not assumed to be diagonal but under some assumptions of symmetry.

We will point out how the construction process, and the proof of approximation properties can both be adapted to operators (4.2); see, respectively, subsections 4.1 and 4.2. However, similarly to the work presented in [24], this work extends to certain operators of higher order like (4.1)—under the same hypothesis.

4.1. The construction process. In the construction process, the crucial point lies in the identification of linear subsystems thanks to the layer structure. Defining

$$\begin{aligned}\widetilde{V} &:= \begin{pmatrix} \partial_x A_{11} + 2\lambda_{1,0}A_{11} + A_{12}\lambda_{0,1} + \partial_y A_{21} + A_{21}\lambda_{0,1} + V_1 \\ \partial_y A_{22} + 2\lambda_{0,1}A_{22} + A_{21}\lambda_{1,0} + \partial_x A_{12} + A_{12}\lambda_{1,0} + V_2 \end{pmatrix}, \\ \widetilde{s} &:= s + V_1\lambda_{1,0} + V_2\lambda_{0,1} + (\partial_x A_{11} + \partial_y A_{21})\lambda_{1,0} + (\partial_x A_{12} + \partial_y A_{22})\lambda_{0,1} \\ &\quad + A_{11}\lambda_{1,0}^2 + (A_{12} + A_{21})\lambda_{1,0}\lambda_{0,1} + A_{22}\lambda_{0,1}^2,\end{aligned}$$

one can easily verify that the equivalent to problem (2.2) reads

$$\text{(4.3)} \quad \begin{cases} \text{Find } (Q, (\lambda_{10}, \lambda_{01})) \in \mathbb{C}[X,Y] \times \mathbb{C}^2 \text{ such that} \\ A_{11}\partial_x^2 Q(x,y) + (A_{12} + A_{21})\partial_x\partial_y Q(x,y) + A_{22}\partial_y^2 Q(x,y) \\ \qquad + \widetilde{V}(x,y)\cdot\nabla Q(x,y) + \widetilde{s}(x,y)Q(x,y) = O(|(x,y) - (x_c,y_c)|^q), \\ G(x,y) := Q(x,y)\exp\Big(\lambda_{10}(x - x_c) + \lambda_{01}(y - y_c)\Big), \end{cases}$$

where it is important to keep in mind that the second-order term coefficients are not constant here. For a given $(\lambda_{10}, \lambda_{01}) \in \mathbb{C}^2$, as in subsection 2.2, the linear system's unknowns are still the polynomial coefficients of Q, while the linear system's equations are still the Taylor expansion coefficients of (4.3). The degree of Q, for the same reason as earlier, is set to $\deg Q = q + 1$. For a more compact notation, we will write $A^c := A(x_c, y_c)$. Instead of the Laplacian defined on spaces of homogeneous polynomials, it is the operator $\nabla \cdot A^c \nabla$ defined on spaces of homogeneous polynomials which is here key to the layer structure. As a consequence, the well-posedness relies on the size of the kernel of this operator, which is again equal to 2 according to the identity

$$\begin{aligned}&\nabla \cdot A^c \nabla \left[\sum_{i=0}^{\ell+2} p_i X^i Y^{\ell+2-i}\right] \\ &\quad = \sum_{i=0}^{\ell} \Big((i+2)(i+1)A_{11}^c p_{i+2} + (i+1)(\ell+1-i)(A_{12}^c + A_{21}^c)p_{i+1} \\ &\qquad\qquad + (\ell+2-i)(\ell+1-i)A_{22}^c p_i\Big) X^i Y^{\ell-i}.\end{aligned}$$

Therefore, the linear system also has a solution, and an algorithm following the same stages as Algorithm 2.1, with updated formulas for RHS and $\mu_{j_x+2,\ell-j_x}$, constructs an amplitude-based GPW for any $\vec{d} = (\lambda_{1,0}, \lambda_{0,1}) \in \mathbb{C}^2$ which satisfies the local approximation property in the vicinity of (x_c, y_c), namely,

$$\Big[\nabla \cdot A(x,y)\nabla + V(x,y)\cdot\nabla + s(x,y)\Big] G_{\vec{d}}(x,y) = O\left(|(x,y) - (x_c,y_c)|^q\right),$$

independently of the normalization.

4.2. Best approximation properties. The impact of the normalization on the approximation properties appears in section 3 when we express the coefficients $\{\mu_{i_x,i_y}, 0 \leq i_x + i_y \leq q+1\}$ of Q in terms of the phase coefficients $(\lambda_{1,0}, \lambda_{0,1})$. In Lemma 3.1's proof, the assumption that $(\lambda_{1,0})^2 + (\lambda_{0,1})^2$ is fixed is crucial to show that $\mu_{2,0}$ can be expressed as a constant polynomial in $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$. In the case of the generalized operator (4.2), the expression for the unknown μ_{20} corresponding to (3.2) reads

$$\mu_{2,0} = -\frac{1}{2A_{11}(x_c, y_c)} \left(\widetilde{V}_1(x_c, y_c)\mu_{1,0} + \widetilde{V}_2(x_c, y_c)\mu_{0,1} + \widetilde{s}(x_c, y_c)\mu_{0,0} \right).$$

Since $\mu_{1,0} = \mu_{0,1} = 0$ while $\mu_{0,0} = 1$, we can focus our attention on the last term. From the definition of $\widetilde{s}$, we can consider $\widetilde{s}(x_c, y_c)$ as an element $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$:

$$\begin{aligned}\widetilde{s}(x_c, y_c) = {} & s(x_c, y_c) + V_1\lambda_{1,0} + V_2\lambda_{0,1} \\ & + (\partial_x A_{11} + \partial_y A_{21})(x_c, y_c)\lambda_{1,0} + (\partial_x A_{12} + \partial_y A_{22})(x_c, y_c)\lambda_{0,1} \\ & + A_{11}(x_c, y_c)\lambda_{1,0}^2 + (A_{12} + A_{21})(x_c, y_c)\lambda_{1,0}\lambda_{0,1} + A_{22}(x_c, y_c)\lambda_{0,1}^2.\end{aligned}$$

The choice of normalization for $(\lambda_{1,0}, \lambda_{0,1})$ then determines how, in turn, $\mu_{2,0}$ can be expressed as an element of $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$. For the Helmholtz case, since $A_{12} = A_{21} = 0$ and $A_{11} = A_{22} = 1$, the second-degree terms were reduced to $(\lambda_{1,0})^2 + (\lambda_{0,1})^2$, and there were no first-degree terms. The normalization assumption that $(\lambda_{1,0})^2 + (\lambda_{0,1})^2$ was fixed therefore resulted in the expression of $\mu_{2,0}$ as a constant polynomial in $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$. But here the first-degree terms are nonzero unless the matrix $A(x_c, y_c)$ is constant, and there are three second-degree terms. However, since the matrix $A(x_c, y_c)$ is assumed to be real symmetric with nonzero eigenvalues, denoted here γ_1, γ_2, it can be diagonalized by an orthogonal matrix: There exists an orthogonal matrix P such that $\left(\begin{smallmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{smallmatrix}\right) = PA(x_c, y_c)P^T$, and the second-degree terms can then be rewritten as

$$\begin{aligned}& A_{11}(x_c, y_c)\lambda_{1,0}^2 + (A_{12} + A_{21})(x_c, y_c)\lambda_{1,0}\lambda_{0,1} + A_{22}(x_c, y_c)\lambda_{0,1}^2 \\ & \quad = \gamma_1 \left(P_{11}\lambda_{1,0} + P_{12}\lambda_{0,1}\right)^2 + \gamma_2 \left(P_{21}\lambda_{1,0} + P_{22}\lambda_{0,1}\right)^2.\end{aligned}$$

The normalization assumption that

$$\begin{pmatrix} \lambda_{1,0} \\ \lambda_{0,1} \end{pmatrix} \propto P^T \begin{pmatrix} 1/\sqrt{\gamma_1} & 0 \\ 0 & 1/\sqrt{\gamma_2} \end{pmatrix} \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \tag{4.4}$$

then guarantees that $\mu_{2,0}$ can be written as a polynomial of degree at most equal to 1 in $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$. So the normalization is different from that of the Helmholtz case, although if A is the identity matrix and V is the zero vector, then (4.4) reduces to the Helmholtz PW normalization.

Starting from this point and following a similar reasoning as that of Lemma 3.1's proof, we can prove that each μ_{i_x,i_y} for $i_x + i_y \geq 1$ can be expressed as a polynomial in $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ of degree at most equal to $i_x + i_y - 1$.

The road map's second and third steps are independent of the GPW normalization choice, as the reference case is still the classical PW case, with propagation direction $(\cos\theta, \sin\theta)$ but without any restriction on the wave number, that is, functions $H_k := \exp(\vec{d}_k \cdot (\cdot - (x_c, y_c)))$ with some $\vec{d}_k \propto (\cos\theta_k, \sin\theta_k)$.

There is an important consequence of the normalization to comment on concerning the road map's fourth step. The functions $F_k := \exp(\vec{e}_k \cdot (\cdot - (x_c, y_c)))$ with $\vec{e}_k :=$

$(\lambda_{1,0}, \lambda_{0,1})$ with the normalization (4.4) for $\theta = \theta_k$ are a natural intermediate between the GPWs and the classical PWs, useful to write the GPWs as $G_k = Q \cdot F_k$. While in the Helmholtz case it is clear that $H_k = F_k$ for the appropriate choice of wave number, this is not the case as soon as the matrix A is distinct from the identity. The pending result to Lemma 3.4 would then relate the derivatives of G_ks to derivatives of F_k: Any derivative of order $j_x + j_y$ of the difference $G_k - F_k$ could be expressed as a polynomial of degree smaller than $j_x + j_y$ in $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$. The only impact of the normalization on the proof is through Lemma 3.1, and μ_{i_x,i_y} being expressed as a polynomial of degree at most equal to i_x+i_y-1 is sufficient to conclude. The fact that $F_k \neq H_k$ then requires an additional step to relate the matrices M_n^G and M_n^C thanks to the introduction of the matrix M_n^F corresponding to the intermediate set of functions $\{F_k, k \in \mathbb{N}, k \leq p\}$. Relating M_n^G to M_n^F follows from the previous comment, just as in Lemma 3.5. Relating M_n^F to M_n^C is straightforward, as can be seen from [24, Lemma 7]. This provides again all the pieces to prove that the rank of M_n^G is $2n+1$ as expected.

Finally, one more time, the fifth step is a direct consequence of the preliminary work. So the proof of Theorem 3.7 does not need to be adapted, and the same approximation properties hold for the operator (4.2) as for the Helmholtz operator under the assumptions that

- the matrix $A(x_c, y_c)$ is real symmetric with nonzero eigenvalues;
- the normalization of $(\lambda_{1,0}, \lambda_{0,1})$ is chosen according to (4.4).

5. Numerical results. This section is dedicated to illustrating the approximation properties of amplitude-based GPWs for the Helmholtz equation and beyond. The most natural aspect to discuss is the expected high-order convergence, but we will also emphasize aspects of conditioning as well as a comparison with phase-based GPWs in terms of preasymptotic behavior.

The testing procedure follows the structure of the theoretical part of this article. The parameters are set according to the theorem: $p = 2n + 1$, $q = \max(n - 1, 1)$, for increasing values of n. A set of GPWs with appropriate normalization is constructed following Algorithm 2.1, and the GPW approximation u_a's coefficients in the GPW basis are computed following the proof of Theorem 3.7. However, rather than considering a single point (x_c, y_c), we will consider 50 random points distributed in a given domain Ω, as we keep in mind that the GPWs are meant to be a local basis for a problem set on the full domain.

In order to illustrate precisely the first estimate from Theorem 3.7,

$$|(u - u_a)(x, y)| \leq C|(x, y) - (x_c, y_c)|^{n+1},$$

we investigate the h convergence, i.e., in the regime $h := |(x, y)-(x_c, y_c)|$ approaching 0, of the difference between the exact solution u to be approximated and the constructed GPW approximation u_a. The approximate function u_a itself is independent of h but depends on n. The error reported here is an estimate of the L^∞ norm of the difference $u - u_a$ on the circle centered at (x_c, y_c) of radius h. This is different from the norm reported in previous work where we reported the L^∞ norm of the difference $u - u_a$ on the disk centered at (x_c, y_c) of radius h. We then report the largest error among errors obtained at each of the 50 random points (x_c, y_c). According to the theorem, for each value of n, we expect to observe convergence of order $n + 1$. The constant C depends both on n and on (x_c, y_c).

All the numerical experiments presented in this article were computed with the same set of normalization angles, $\mathrm{A} := \{\theta_k = \frac{2(k-1)\pi}{p} + \frac{\pi}{6}; 1 \leq k \leq p\}$, in order to avoid any particular alignment of the basis functions with the coordinate axes.

5.1. List of test cases. Each test case consists of a variable-coefficient PDE (1.1) or (4.2), a given domain, and an exact solution u to this PDE. It is straightforward to verify that these operators satisfy the hypothesis described in section 4.

Ref.	Operator	Domain	Exact solution
Ae	$\Delta - (x-1)$	$[-2,2]\times[-2,2]$	$u(x,y) = Ai(x)\exp iy$
Ac	$\Delta - (x-1)$	$[-2,2]\times[-2,2]$	$u(x,y) = Ai(x)\cos y$
A+	$\Delta - 2(x+y)$	$[-2,2]\times[-2,2]$	$u(x,y) = Ai(x+y)$
cs	$\partial_x^2 + .2\cos x \sin y \partial_x\partial_y - 2\partial_y^2$ $+(.2\sin x\cos y - 1)$	$[-1,1]\times[-1,1]$	$u(x,y) = \cos x \sin y$
ey	$\Delta + 1$	$[-1,1]\times[0,2\pi]$	$u(x,y) = \exp(iy)$
Jc	$x^2\Delta - x\partial_x + \cos y\partial_y$ $-(1-2x^2-\sin y)$	$[1,5]\times[0,2\pi]$	$u(x,y) = J_1(x)\cos y$
JJ	$x^2\partial_x^2 + y^2\partial_y^2 + x\partial_x + y\partial_y$ $+(x^2+y^2-1)$	$[1,3]\times[0,3]$	$u(x,y) = J_0(x)J_1(y)$

5.2. Higher-order convergence. Figure 1 displays the results obtained for the ey test case. In this case, the operator is the Helmholtz operator with a constant wave number. The GPWs constructed by Algorithm 2.1 are then exactly the classical PWs, as is the case with phase-based GPWs. For each value of n from 1 to 20, the order of convergence observed is the order obtained in the theorem. The error decreases until it reaches machine precision, and as h keeps decreasing beyond this point, the error remains of the same magnitude.

Figures 2 and 3 display the results obtained for four different test cases: Ae, Ac, A+, and cs. The first three correspond to the Helmholtz equation with a variable wave number, the variable wave number changes sign within each of the associated domains, and the exact solutions oscillate for $\kappa^2 > 0$ and decay exponentially for $\kappa^2 < 0$. The last case corresponds to a more general operator (4.2), with anisotropy in both the second- and the first-order terms. For each value of n from 1 to 20, the order of convergence observed is the order obtained in the theorem. However, the matrix M_n^G's condition number increases with the value of n, so the coefficients X_k of the linear combination $u_a \sum_{k=1}^{2n+1} X_k G_k$ are computed with decreasing accuracy

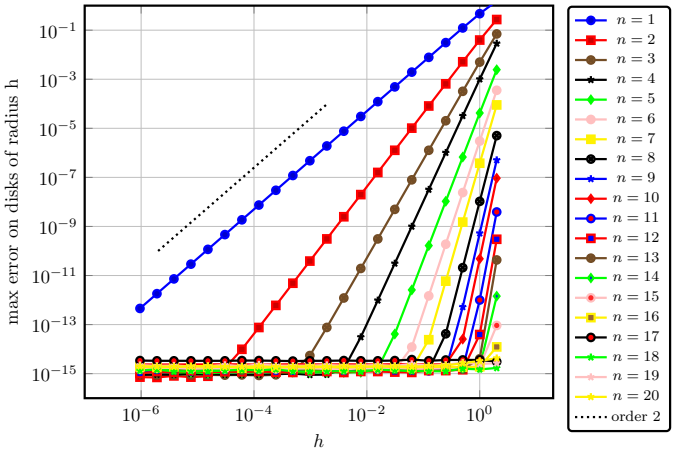

FIG. 1. *Convergence results for the ey test case for n from 1 to 20. For each value of n, the expected order of convergence, namely, $n+1$, is observed, and the error decreases until it reaches machine precision.*

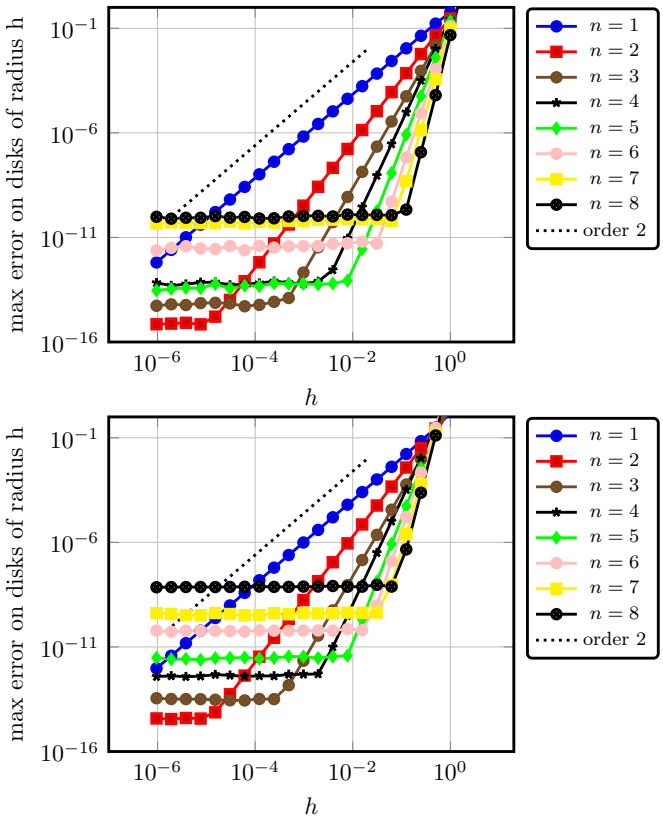

FIG. 2. *Convergence results for the A+ (top) and cs (bottom) test cases for n from* 1 *to* 8. *While the expected order of convergence is observed in each case, the matrix* M_n^G*'s ill-conditioning increases together with the value of n, which limits the accuracy of the corresponding GPW approximation.*

as n increases. This is illustrated by the behavior of the errors in Figures 2 and 3: the error's decrease is limited by the accuracy of the coefficient X_k, and as h keeps decreasing beyond this point, the error remains of the same magnitude.

Preliminary results suggest that the normalization chosen to compute the GPWs has a strong impact on M_n^G's condition number, as illustrated in Figure 4. These results were produc with a different normalization of the GPWs, namely, $\vec{d} \propto (\cos\theta, \sin\theta)$, instead of the normalization (4.4). Note, for instance, that for $n = 8$, the threshold shifts from approximately 10^{-8} in Figure 2 (bottom) to less than 10^{-12} in Figure 4. However, we would like to emphasize that this simpler normalization provides better results for the cs test case, while it does not impact significantly the other three test cases presented in Figures 2 and 3. Further investigation in this direction is the topic of ongoing research.

5.3. Comparison with phase-based GPWs. Figures 5 and 6 evidence the benefit of amplitude-based GPWs over phase-based GPWs in the preasymptotic regime. The construction of increasingly high-order GPWs involves polynomials of increasingly high degree. While phased-based GPWs are designed as

$$G(x,y) := \exp P(x,y) \text{ with } P(x,y) = \sum_{0 \le i_x + i_y \le \deg P} \lambda_{i_x,i_y} (x - x_c)^{i_x} (y - y_c)^{i_y},$$

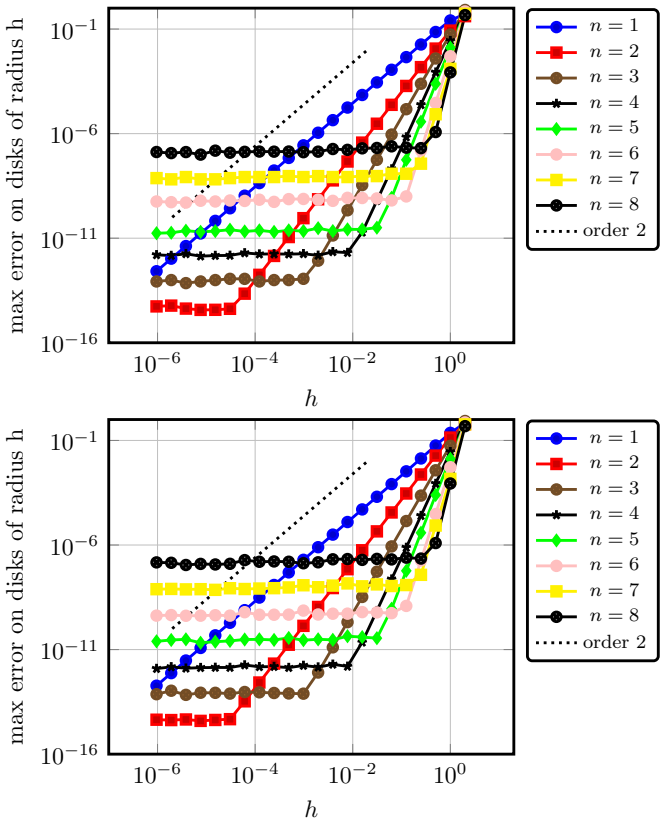

FIG. 3. *Convergence results for the Ac (top) and Ae (bottom) test cases for n from* 1 *to* 8. *While the expected order of convergence is observed in each case, the matrix* M_n^G*'s ill-conditioning increases together with the value of n, which limits the accuracy of the corresponding GPW approximation.*

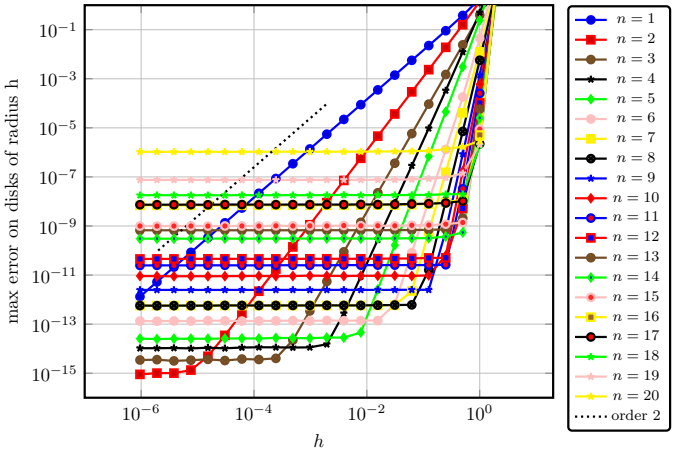

FIG. 4. *Convergence results for the cs test case for n from* 1 *to* 20 *with the classical PW normalization. For each value of n, the expected order of convergence, namely,* $n + 1$*, is observed, and when compared to Figure* 2 *(top right), we observe a decrease of the thresholds by several orders of magnitude.*

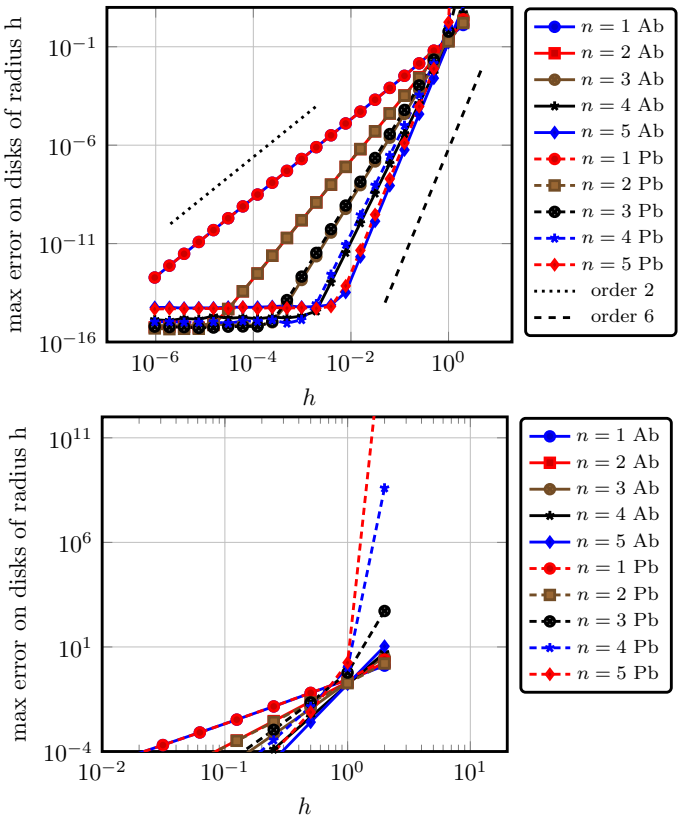

FIG. 5. *Convergence results for the JJ test case for n from* 1 *to* 5. *Results obtained from amplitude-based and phase-based GPWs are, respectively, represented with solid and dashed lines for comparison. While the expected order of convergence is observed in each case, the zoom displayed at the bottom evidences the advantage of the amplitude-based GPWs in the preasymptotic regime.*

amplitude-based GPWs are designed as

$$G(x,y) := Q(x,y)W(x,y) \text{ with } Q(x,y) = \sum_{0 \le i_x + i_y \le \deg Q} \mu_{i_x,i_y} (x-x_c)^{i_x} (y-y_c)^{i_y}$$

$$\text{and } W(x,y) := \exp \begin{pmatrix} \lambda_{1,0} \\ \lambda_{0,1} \end{pmatrix} \cdot \begin{pmatrix} x-x_c \\ y-y_c \end{pmatrix}.$$

Therefore, in the first case, evaluating a GPW implies evaluating $\exp \lambda_{i_x,i_y}(x - x_c)^{i_x}(y - y_c)^{i_y}$ with $0 \le i_x + i_y \le \deg P$, while in the second case, the exponential terms are limited to $\exp \lambda_{1,0}(x - x_c)$ and $\exp \lambda_{0,1}(y - y_c)$. This explains the large values of phase-based GPWs evaluated at (x, y) such that $|(x, y) - (x_c, y_c)| \ge 1$.

6. Conclusion. Amplitude-based GPWs constructed thanks to Algorithm 2.1 are quasi-Trefftz functions in the sense that the Taylor expansion of their image by the differential operator is zero up to a desired order. These new GPWs enjoy the same asymptotic approximation properties as their phase-based counterparts, and the numerical results presented in the previous section agree with the theoretical results. Moreover, we illustrate the fact that the best approximation error of the amplitude-

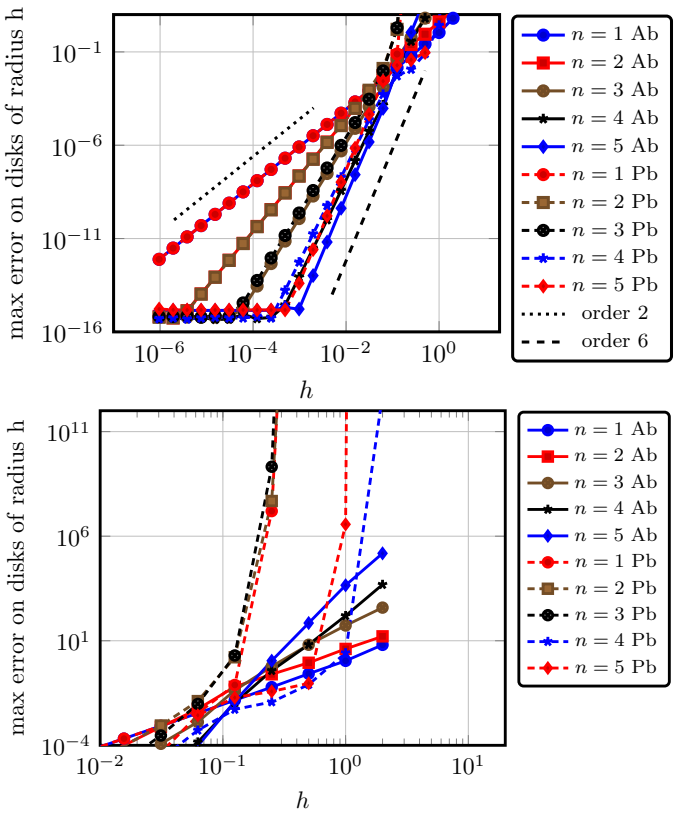

FIG. 6. *Convergence results for the Jc test case for n from* 1 *to* 5. *Results obtained from amplitude-based and phase-based GPWs are, respectively, represented with solid and dashed lines for comparison. While the expected order of convergence is observed in each case, the zoom displayed at the bottom evidences the advantage of the amplitude-based GPWs in the preasymptotic regime.*

based GPW approximation exhibits as expected a better preasymptotic behavior than the best approximation error of the phase-based GPW approximation.

Future directions of research include studying the impact of the normalization on the conditioning of the GPW basis, optimizing the choice of propagation directions, and investigating the behavior of GPWs in the high-frequency regime $\kappa \gg 1$.

REFERENCES

[1] S. AMBIKASARAN, C. BORGES, L.-M. IMBERT-GERARD, AND L. GREENGARD, *Fast, adaptive, high-order accurate discretization of the Lippmann-Schwinger equation in two dimensions*, SIAM J. Sci. Comput., 38 (2016), pp. A1770–A1787.

[2] I. M. AZIS, *Numerical solutions for the Helmholtz boundary value problems of anisotropic homogeneous media*, J. Comput. Phys., 381 (2019), pp. 42–51.

[3] S. BRITT, S. TSYNKOV, AND E. TURKEL, *Numerical simulation of time-harmonic waves in inhomogeneous media using compact high order schemes*, Commun. Comput. Phys., 9 (2011), pp. 520–541.

[4] N. BALIN, F. CASENAVE, F. DUBOIS, E. DUCEAU, S. DUPREY, AND I. TERRASSE, *Boundary element and finite element coupling for aeroacoustics simulations*, J. Comput. Phys., 294 (2015), pp. 274–296.

[5] F. CASENAVE, A. ERN, AND G. SYLVAND, *Coupled BEM-FEM for the convected Helmholtz equation with non-uniform flow in a bounded domain*, J. Comput. Phys., 257 (2014), pp. 627–644

[6] O. Cessenat, *Application d'une nouvelle formulation variationnelle aux équations d'ondes harmoniques. Problèmes de Helmholtz* 2*D et de Maxwell* 3*D*, Thesis, Université Paris 9 Dauphine, 1996.
[7] O. Cessenat and B. Després, *Application of an ultra weak variational formulation of elliptic PDEs to the two dimensional Helmholtz problem*, SIAM J. Numer. Anal., 55 (1998), pp. 255–299.
[8] Y. K. Cheung, W. G. Jin, and O. C. Zienkiewicz, *Solution of Helmholtz equation by Trefftz method*, Internat. J. Numer. Methods Engrg., 32 (1991), pp. 63–78.
[9] L. Collatz, *The Numerical Treatment of Differential Equations*, 3rd ed., Springer-Verlag, Berlin, 1966.
[10] J. Coyle and P. Monk, *Scattering of time-harmonic electromagnetic waves by anisotropic inhomogeneous scatterers or impenetrable obstacles*, SIAM J. Numer. Anal., 37 (2000), pp. 1590–1617.
[11] J. B. Diaz and A. Weinstein, *The torsional rigidity and variational methods*, Amer. J. Math., 70 (1948), pp. 107–116.
[12] B. Després, *Sur une formulation variationnelle de type ultra-faible. [An ultra-weak-type variational formulation]*, C. R. Acad. Sci. Paris Sér. I Math., 318 (1994), pp. 939–944.
[13] A. Gillman, A. H. Barnett, and P.-G. Martinsson, *A spectrally accurate direct solution technique for frequency-domain scattering problems with variable media*, BIT, 55 (2015), pp. 141–170.
[14] C. J. Gittelson and R. Hiptmair, *Dispersion analysis of plane wave discontinuous Galerkin methods*, Internat. J. Numer. Methods Engrg., 98 (2014), pp. 313–323.
[15] C. J. Gittelson, R. Hiptmair, and I. Perugia, *Plane wave discontinuous Galerkin methods: Analysis of the h-version*, Math. Model. Numer. Anal., 43 (2009), pp. 297–332.
[16] R. Hiptmair, A. Moiola, and I. Perugia, *A survey of Trefftz Methods for the Helmholtz equation*, in Building Bridges: Connections and Challenges in Modern Approaches to Numerical PDEs, Lect. Notes Comput. Sci. Eng. 114, Springer-Verlag, Berlin, 2016, pp. 237–278.
[17] I. Herrera, *Trefftz method*, in Topics in Boundary Element Research, Vol. 1: Basic Principles and Applications, Springer-Verlag, 1984, pp. 225–253.
[18] D. L Holl and E. W. Anderson, *Limits of approximate solutions of a torsion problem*, Bull. Amer. Math. Soc., 37 (1931), pp. 580–584.
[19] T. Huttunen and P. Monk, *The use of plane waves to approximate wave propagation in anisotropic media*, J. Comput. Math., 25 (2007), pp. 350–367.
[20] L.-M. Imbert-Gérard, *Interpolation properties of generalized plane waves*, Numer. Math., 131 (2015), pp. 683–711.
[21] L.-M. Imbert-Gérard, *Well-posedness and generalized plane waves simulations of a* 2*D mode conversion model*, J. Comput. Phys., 303 (2015), pp. 105–124.
[22] L.-M. Imbert-Gerard and B. Despres, *A generalized plane-wave numerical method for smooth nonconstant coefficients*, IMA J. Numer. Anal., (2013), https://doi.org/10.1093/imanum/drt030.
[23] L.-M. Imbert-Gérard and P. Monk, *Numerical simulation of wave propagation in inhomogeneous media using Generalized Plane Waves*, ESAIM: Math. Models. Numer. Anal., 51 (2017), pp. 1387–1406.
[24] L.-M. Imbert-Gérard and G. Sylvand, *A roadmap for Generalized Plane Waves and their interpolation properties*, Numer. Math., to appear.
[25] M. Jacobs and S. Luo, *Asymptotic solutions for high frequency Helmholtz equations in anisotropic media with Hankel functions*, J. Sci. Comput., 80 (2019), pp. 808–833.
[26] J. Jirousek and N. Leon, *A powerful finite element for plate bending*, Comput. Methods Appl. Mech. Engrg., 12 (1977), pp. 77–96.
[27] J. Jirousek and A. Wróblewski, *T-elements: State of the art and future trends*, Arch. Comput. Methods Eng., 3 (1996), pp. 323–434.
[28] V. Kompis, *Finite elements satisfying all governing equations inside the element*, Comput. Struct., 50 (1994), pp. 273–278.
[29] A. Kuhelj, *Energy criterion of elastic stability for thin shells*, Publ. Inst. Math. (Beograd) (N.S.), (1953) , pp. 77–102.
[30] Y. Long and H. Qiya, *Error analysis of the plane wave discontinuous Galerkin method for Maxwell's equations in anisotropic media*, Commun. Comput. Phys., 25 (2019), pp. 1496–1522.
[31] T. Luostari, T. Huttunen, and P. Monk, *The ultra weak variational formulation using Bessel basis functions*, Commun. Comput. Phys., 11 (2011), pp. 400–414.
[32] L. Mascotto, I. Perugia, and A. Pichler, *A nonconforming Trefftz virtual element method for the Helmholtz problem*, Math. Models Methods Appl. Sci., 29 (2019), pp. 1619–1656.

[33] L. Mascotto and A. Pichler, *Extension of the nonconforming Trefftz virtual element method to the Helmholtz problem with piecewise constant wave number*, Appl. Numer. Math., 155 (2020), pp. 160–180.
[34] E. A. W. Maunder, *Trefftz in translation*, Comput. Assist. Mech. Eng. Sci., 10 (2003), pp. 545–563.
[35] W. Prager and J. L. Synge, *Approximations in elasticity based on the concept of function space*, Quart. Appl. Math., 5 (1943), pp. 241–269.
[36] M. Stojek, *Least-squares Trefftz-type elements for the Helmholtz equation*, Internat. J. Numer. Methods Engrg., 41 (1998), pp. 831–849.
[37] E. Trefftz, *Ein gegenstuck zum ritzschen verfahren*, in Proceedings of the 2nd International Congress of Applied Mechanics, Orell Fussli Verlag, Zurich, Switzerland, 1926, pp. 131–137.
[38] I. Tsukerman, *A class of difference schemes with flexible local approximation*, J. Comput. Phys., 211 (2006), pp. 659–699.
[39] I. Tsukerman, *Electromagnetic applications of a new finite-difference calculus*, IEEE Trans. Magn., 41 (2005), pp. 2206–2225.
[40] Z. Wu and T. Alkhalifah, *A highly accurate finite-difference method with minimum dispersion error for solving the Helmholtz equation*, J. Comput. Phys., 365 (2018), pp. 350–361.
[41] L. Yuan, *A combined scheme of the local spectral element method and the generalized plane wave discontinuous Galerkin method for the anisotropic Helmholtz equation*, Appl. Numer. Math., 150 (2020), pp. 341–360.
[42] L. Zepeda-Núñez and H. Zhao, *Fast alternating bidirectional preconditioner for the 2D high-frequency Lippmann-Schwinger equation*, SIAM J. Sci. Comput., 38 (2016), pp. B866–B888.
[43] A. P. Zielinski, *On trial functions applied in the generalized Trefftz method*, Adv. Eng. Softw., 24 (1995), pp. 147–155.
[44] O. C. Zienkiewicz, D. W. Kelly, and P. Bettess, *The coupling of the finite element method and boundary solution procedures*, Internat. J. Numer. Methods Engrg., 11 (1977), pp. 355–375.