SIAM J. NUMER. ANAL.
Vol. 58, No. 5, pp. 3091–3123

A NOVEL LEAST SQUARES METHOD FOR HELMHOLTZ EQUATIONS WITH LARGE WAVE NUMBERS[*]

QIYA HU[†] AND RONGRONG SONG[†]

Abstract. In this paper we are concerned with numerical methods for Helmholtz equations with large wave numbers. We design a least squares method for discretization of the considered Helmholtz equations. In this method, an auxiliary unknown is introduced on the common interface of any two neighboring elements and a quadratic objective functional is defined by the jumps of the traces of the solutions of local Helmholtz equations across all the common interfaces, where the local Helmholtz equations are defined on elements and are imposed Robin-type boundary conditions given by the auxiliary unknowns. A minimization problem with the objective functional is proposed to determine the auxiliary unknowns. The resulting discrete system of the auxiliary unknowns is Hermitian positive definite and so it can be solved by the preconditioned conjugate gradient method. Under some assumptions we show that the generated approximate solutions possess almost the same L^2 convergence order as the plane wave methods (for the case of constant wave number). Moreover, we construct a substructuring preconditioner for the discrete system of the auxiliary unknowns. Numerical experiments show that the proposed methods are very effective and have little "wave number pollution" for the tested Helmholtz equations with large wave numbers.

Key words. Helmholtz equations, inhomogeneous media, large wave number, auxiliary unknowns, least squares, error estimates, preconditioner

AMS subject classifications. 65N30, 65N55

DOI. 10.1137/19M1294101

1. Introduction.

Let Ω be a bounded, connected Lipschitz domain in $\mathbb{R}^2$. Consider the Helmholtz equations

$$\begin{cases} -\Delta u-\kappa^2 u=f & \text{in } \Omega,\\ \dfrac{\partial u}{\partial \mathbf{n}}+i\kappa u=g & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where $\mathbf{n}$ denotes the unit outward normal on the boundary $\partial\Omega$ and κ is the wave number defined by $\kappa(\mathbf{x})=\frac{\omega}{c(\mathbf{x})}>0$, with $\omega>0$ being a constant and $c(\mathbf{x})$ being a bounded and positive function defined on Ω. In applications, ω denotes the angular frequency, which may be very large, and $c(\mathbf{x})$ denotes the wave speed (the acoustic velocity), which may not be a constant function on Ω, i.e., the involved media is inhomogeneous.

The Helmholtz equation is the basic model in sound propagation. It is a very important topic to design a high accuracy method for Helmholtz equations with large wave numbers, such that the so-called wave number pollution can be reduced. The "wave number pollution" says that, for a finite element method for the discretization of (1.1), the mesh size h must satisfy $h\omega^{1+\delta}=O(1)$ for some positive number δ to achieve

[*] Received by the editors October 18, 2019; accepted for publication (in revised form) August 13, 2020; published electronically October 29, 2020.
https://doi.org/10.1137/19M1294101
Funding: The work of the authors was supported by National Natural Science Foundation of China grant G12071469.

[†] LSEC, ICMSEC, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China, and School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China (hqy@lsec.cc.ac.cn, songrongrong@lsec.cc.ac.cn).

a given accuracy of the approximate solutions when the wave number ω increases, which means that the accuracies of the approximate solutions are obviously destroyed if fixing the value of $h\omega$ but increasing the wave number ω. For convenience, we call the parameter δ the "pollution index," which describes the degree of wave number pollution. For the standard linear finite element method, the "pollution index" $\delta = 1$ (see [32]). Of course, we hope to design a "good" finite element method for (1.1) such that the pollution index δ is sufficiently small.

In recent years, many interesting methods for the discretization of Helmholtz equations with large wave numbers have been proposed, for example, the higher-order finite element methods [9, 32], the ultra weak variational formulation [3, 29], the plane wave least squares (PWLS) methods [25, 26, 33], the plane wave discontinuous Galerkin (PWDG) methods [16, 23], the method of fundamental solutions [1, 6], the plane wave method with Lagrange multipliers [12], the variational theory of complex rays [37], the high-order element discontinuous Galerkin method [9, 13], the local discontinuous Galerkin method [14], the hybridizable discontinuous Galerkin method (HDG) [4, 5, 17, 22, 34, 35, 38] and the discontinuous Petrov–Galerkin (DPG) method [7, 18, 19], the ray-based finite element method [11], and the generalized plane wave method [30]. All these methods are superior to the standard linear finite element method in the sense that the pollution index $\delta < 1$.

It is known that the plane wave finite element methods have little "wave number pollution" (i.e., the pollution index δ is very small) and can generate higher accuracy approximations than the polynomial basis finite element methods for solving the Helmholtz equations with large (piecewise constant) wave numbers when finite element spaces have the same degrees of freedom. A comparison of finite element methods based on high-order polynomial basis functions and plane wave basis functions was given in [31]. The numerical results reported in [31] indicate that if only the degrees of freedom on element boundaries for a high-order polynomial method are calculated (the degrees of freedom in the interior of elements are eliminated), the high-order polynomial method can deliver comparably to the PWDG method. Unfortunately, the plane wave methods cannot be directly applied to the discretization of nonhomogeneous Helmholtz equations in inhomogeneous media. A plane wave method combined with a local spectral element for nonhomogeneous Helmholtz equations in homogeneous media was proposed in [27] (see also [26]). A generalized plane wave method for homogeneous Helmholtz equations in inhomogeneous media was introduced in [30].

The HDG-type methods (and the DPG method) have been studied in many works (see the references listed above). We would like to simply recall the ideas of the HDG methods. Let Ω be decomposed into a union of elements $\{\Omega_k\}$, and let γ denote the element interface, which is a union of all the common edges of two neighboring elements. For the HDG-type methods, (1.1) is first transformed into a first-order system of the original unknown u and an auxiliary unknown $\Phi = (i\omega)^{-1}\nabla u$, then the restrictions of the unknowns u and Φ on the elements $\{\Omega_k\}$ are eliminated by solving all the local first-order systems to obtain an interface equation of the trace $u|_\gamma$ (and the trace $(\nabla u \cdot \mathbf{n})|_\gamma$ in [34]) in some manner. For the DPG method, there are two interface unknowns that are defined by the traces $u|_\gamma$ and $(\nabla u \cdot \mathbf{n})|_\gamma$ and the interface equation becomes Hermitian positive definite by introducing nonstandard test space that is the image of the trial space under a suitable mapping. In both the HDG-type methods and the DPG method, the unknown needing to be globally solved was defined on the interface γ, so these methods have less cost of calculation than the standard hp finite element method proposed in [32]. The HDG-type methods and the DPG method have their respective merits: the HDG-type methods are easier to implement than the DPG method since the HDG-type methods use the standard polynomial basis functions; the interface equation needed to be solved globally is Hermitian positive definite for the DPG method, but it is still indefinite as the original equation (1.1) for the HDG-type methods.

In the present paper, we design a novel discretization method for Helmholtz equations with large wave numbers such that the method can absorb the merits of the HDG-type methods and the DPG method. The basic ideas of the new method can be roughly described as follows. We introduce an auxiliary unknown λ_h that is an edgewise q-order polynomial on γ and compute p-order ($p \geq q+2$) polynomial solutions $\{u_{h,k}\}$ of the discrete variational problems of all local Helmholtz equations, where each local Helmholtz equation is the restriction of (1.1) on some element Ω_k and is imposed a Robin-type boundary condition given by the auxiliary unknown λ_h. We define a minimization problem with a quadratic objective functional defined by the jumps of the traces of the solutions $\{u_{h,k}\}$ across the interface γ. This minimization problem results in a Hermitian positive definite algebraic system of the auxiliary unknown λ_h. After solving the algebraic system, we can easily obtain an approximate solution of the original Helmholtz equation by solving small local problems on the elements in parallel manner. This method has some similarity with the HDG method but it has essential differences from the HDG method: (a) each element subproblem is just the local variational problem of the original Helmholtz equation (1.1), so only one internal unknown $u_{h,k}$ needs to be computed for an element Ω_k; (b) the interface unknown λ_h, which may be discontinuous on the interface γ, is defined independently on every edge of elements; (c) the interface unknown λ_h is determined by a minimization problem, so the interface equation is Hermitian positive definite.

Since the resulting approximate solutions (u_h, λ_h) neither satisfy a mixed variational problem (comparing the Lagrange multiplier methods) nor satisfy a hybridizable variational problem (comparing the HDG methods), well-posedness and convergence of the proposed method cannot be proved by the techniques developed in existing works. By developing some new techniques, we show that the proposed discretization method is well-posed and the resulting approximate solution possesses almost the same L^2 error estimate as the plane wave methods under suitable assumptions, which indicate that the proposed method has little "wave number pollution."

In addition, we construct a domain decomposition preconditioner for the algebraic system of λ_h. The balancing domain decomposition by constraints (BDDC) is a popular substructuring domain decomposition method, which was first proposed in [8] and then was extended to various models by many researchers. The key idea of the BDDC method is to compute basis functions of the coarse space by solving local minimization problems. This method has some advantages over the traditional substructuring methods, but the minimization problems for computing coarse basis functions can be defined only for symmetric and positive definite systems. Thanks to the Hermitian positive definiteness of the algebraic system of λ_h, we can construct a substructuring preconditioner for the system by the BDDC method. However, we find that the coarse space defined by the BDDC method is unsatisfactory for the current situation. Because of this, we construct a variant of the BDDC preconditioner for the algebraic system of λ_h by changing the definition of coarse space.

Numerical results indicate that the proposed discretization method and preconditioner are very efficient for the tested Helmholtz equations with large wave numbers.

The new method possesses the following merits: (i) the method possesses simple construction since the subproblem for computing $u_{h,k}$ on an element Ω_k is directly defined by the original second-order Helmholtz equation and the basis functions on every element Ω_k and every element edge are standard polynomials; (ii) the resulting approximate solutions have nice error estimates with high accuracies; (iii) the method has good numerical performances that are comparable with the plane wave methods in middle and high frequency regimes; (iv) the method is practical to general nonhomogeneous Helmholtz equations in inhomogeneous media; (v) the algebraic system of λ_h is Hermitian positive definite, so it can be solved by the preconditioned conjugate gradient (PCG) method, which has stable convergence and less cost of calculation, and the construction of a preconditioner for this system has more choices; (vi) it is cheap to implement this method since only one unknown $u_{h,k}$ is introduced in each element Ω_k and only one unknown $\lambda_h|_{\gamma_{lj}}$ is involved on each local interface γ_{lj}.

The paper is organized as follows. In section 2, we describe the proposed least squares variational formulation for Helmholtz equations. In section 3, we construct a substructuring preconditioner for the discrete system. The main results about error estimates are presented in section 4. In section 5, we give proofs of the main results in detail. Finally, we report some numerical results to confirm the effectiveness of the new method in section 6.

2. A least squares variational formulation.

2.1. Notation. As usual we partition Ω into elements in the sense that

$$\overline{\Omega} = \bigcup_{k=1}^{N} \overline{\Omega}_k, \quad \Omega_k \bigcap \Omega_j = \emptyset, \quad \text{for } k \neq j.$$

Here each Ω_k may be curve polyhedron. We use h_k to denote the diameter of Ω_k and set $h = \max\{h_k\}$. Let $\mathcal{T}_h$ denote the partition comprised of elements $\{\Omega_k\}_{k=1}^N$. As usual we assume that the partition $\mathcal{T}_h$ is quasi-uniform and regular.

Let γ_{kj} denote the common edge of two neighboring elements Ω_k and Ω_j, and set $\gamma_k = \partial\Omega_k \cap \partial\Omega$ when the intersection is an edge of the element Ω_k. For convenience, define $\gamma = \cup_{k\neq j}\gamma_{kj}$.

Let $q \geq 1$ be an integer and choose $p \geq q+2$. Throughout this paper we use the following notation:

- $V_h^p(\Omega_k) = \{v \in H^1(\Omega_k) : v \text{ is a polynomial whose order does not exceed } p\}$.
- $V_h^p(\mathcal{T}_h) = \prod_{k=1}^N V_h^p(\Omega_k)$.
- $V_h^p(\partial\Omega_k) = \{v|_{\partial\Omega_k} : v \in V_h^p(\Omega_k)\}$.
- $W(\gamma) = \prod_{k\neq j}^N H^{-\frac{1}{2}}(\gamma_{kj})$.
- $W_h^q(\gamma_{kj}) = \{\mu \in H^1(\gamma_{kj}) : \mu \text{ is a polynomial whose order does not exceed } q\}$.
- $W_h^q(\gamma) = \prod_{k\neq j} W_h^q(\gamma_{kj})$.
- $W_h^q(\partial\Omega_k\backslash\partial\Omega) = \{\mu|_{\partial\Omega_k\backslash\partial\Omega} : \mu \in W_h^q(\gamma)\}$.
- The jump of v across γ_{kj}: $[v] = v_k - v_j$, where v is a piecewise smooth function on $\mathcal{T}_h$ and $v_k = v|_{\Omega_k}$.
- $(u,v)_{\Omega_k} = \int_{\Omega_k} u \cdot v \, dx, \quad \langle u,v\rangle_{\partial\Omega_k} = \int_{\partial\Omega_k} u \cdot v \, ds.$

2.2. A continuous variational formulation. Let $u \in H^1(\Omega)$. For each element Ω_k, set $u|_{\Omega_k} = u_k$. For $k > j$, define $\lambda \in W(\gamma)$ as

$$\lambda|_{\gamma_{kj}} = \left(\frac{\partial u_k}{\partial \mathbf{n}_k} + i\rho u_k\right)\Bigg|_{\gamma_{kj}} = \left(-\frac{\partial u_j}{\partial \mathbf{n}_j} + i\rho u_j\right)\Bigg|_{\gamma_{kj}},$$

where $\rho > 0$, $\mathbf{n}_k$, and $\mathbf{n}_j$ separately denote the unit outward normal on $\partial\Omega_k$ and $\partial\Omega_j$. It is clear that the solution u of (1.1) satisfies the local Helmholtz equation on each element Ω_k $(k = 1, \ldots, N)$

$$\begin{cases} -\Delta u_k - \kappa^2 u_k = f & \text{in } \Omega_k, \\ \dfrac{\partial u_k}{\partial \mathbf{n}_k} \pm i\rho u_k = \pm\lambda & \text{on } \partial\Omega_k \backslash \partial\Omega, \\ \dfrac{\partial u_k}{\partial \mathbf{n}_k} + i\kappa u_k = g & \text{on } \partial\Omega_k \cap \partial\Omega. \end{cases} \tag{2.1}$$

Here the sign " $\pm$ " means that two inverse signs are used on the two sides of each local interface $\gamma_{kj} = \partial\Omega_k \cap \partial\Omega_j$: it takes "+" on $\gamma_{kj} \subset \partial\Omega_k$, and it takes "−" on $\gamma_{kj} \subset \partial\Omega_j$.

For each element Ω_k, define the local sesquilinear form

$$a^{(k)}(v,w) = (\nabla v, \nabla \overline{w})_{\Omega_k} - \left(\kappa^2 v, \overline{w}\right)_{\Omega_k} \pm i\rho\langle v, \overline{w}\rangle_{\partial\Omega_k \backslash \partial\Omega} + i\langle \kappa v, \overline{w}\rangle_{\partial\Omega_k \cap \partial\Omega}, \quad v, w \in H^1(\Omega_k)$$

and the local functional

$$L^{(k)}(v) = (f, \overline{v})_{\Omega_k} + \langle g, \overline{v}\rangle_{\partial\Omega_k \cap \partial\Omega}, \quad v \in H^1(\Omega_k).$$

It is easy to see that the variational formulation of (2.1) is to find $u_k(\lambda) \in H^1(\Omega_k)$ such that

$$a^{(k)}(u_k(\lambda), v) = L^{(k)}(v) + \langle \pm\lambda, \overline{v}\rangle_{\partial\Omega_k \backslash \partial\Omega} \quad \forall\, v \in H^1(\Omega_k). \tag{2.2}$$

We define the quadratic functional

$$J(\mu) = \sum_{\gamma_{kj}} \int_{\gamma_{kj}} |u_k(\mu) - u_j(\mu)|^2 ds, \quad \mu \in W(\gamma), \tag{2.3}$$

and consider the following minimization problem: find $\lambda \in W(\gamma)$ such that

$$J(\lambda) = \min_{\mu \in W(\gamma)} J(\mu). \tag{2.4}$$

It is clear that u is the solution of (1.1) if and only if $J(\lambda) = 0$, which means that λ is the solution of the minimization problem (2.4).

In order to give a variational problem of (2.4), we write the solution of (2.1) as $u_k(\lambda) = u_k^{(1)}(\lambda) + u_k^{(2)}$, which respectively satisfy

$$a^{(k)}\left(u_k^{(1)}(\lambda), v\right) = \pm\langle \lambda, \overline{v}\rangle_{\partial\Omega_k \backslash \partial\Omega} \quad \forall\, v \in H^1(\Omega_k)$$

and

$$a^{(k)}\left(u_k^{(2)}, v\right) = L^{(k)}(v) \quad \forall\, v \in H^1(\Omega_k).$$

Then $J(\mu)$ can be written as

$$J(\mu) = \sum_{\gamma_{kj}} \int_{\gamma_{kj}} \left| \left(u_k^{(1)}(\mu) - u_j^{(1)}(\mu)\right) + \left(u_k^{(2)} - u_j^{(2)}\right) \right|^2 ds.$$

Define the sesquilinear form

$$s(\lambda,\mu)=\sum_{\gamma_{kj}}\int_{\gamma_{kj}}\left(u_k^{(1)}(\lambda)-u_j^{(1)}(\lambda)\right)\cdot\overline{\left(u_k^{(1)}(\mu)-u_j^{(1)}(\mu)\right)}ds,\quad \lambda,\mu\in W(\gamma)$$

and the functional

$$l(\mu)=-\sum_{\gamma_{kj}}\int_{\gamma_{kj}}\left(u_k^{(2)}-u_j^{(2)}\right)\cdot\overline{\left(u_k^{(1)}(\mu)-u_j^{(1)}(\mu)\right)}ds,\quad \mu\in W(\gamma).$$

Therefore the variational problem of the minimization problem (2.4) can be expressed as follows: find $\lambda\in W(\gamma)$ such that

$$s(\lambda,\mu)=l(\mu)\quad\forall\ \mu\in W(\gamma). \tag{2.5}$$

2.3. The discrete variational formulation. Let $\lambda_h\in W_h^q(\gamma)$. For each element Ω_k, define $u_{h,k}(\lambda_h)\in V_h^p(\Omega_k)$ by

$$a^{(k)}(u_{h,k}(\lambda_h),v_h)=L^{(k)}(v_h)+\langle\pm\lambda_h,\overline{v}_h\rangle_{\partial\Omega_k\backslash\partial\Omega}\quad\forall\, v_h\in V_h^p(\Omega_k). \tag{2.6}$$

It is easy to see that the above problem is uniquely solvable.

As in the continuous situation, we decompose $u_{h,k}$ into $u_{h,k}=u_{h,k}^{(1)}(\lambda_h)+u_{h,k}^{(2)}$, which are respectively defined by

$$a^{(k)}\left(u_{h,k}^{(1)}(\lambda_h),v_h\right)=\pm\langle\lambda_h,\overline{v}_h\rangle_{\partial\Omega_k\backslash\partial\Omega}\quad\forall\, v_h\in V_h^p(\Omega_k)$$

and

$$a^{(k)}\left(u_{h,k}^{(2)},v_h\right)=L^{(k)}(v_h)\quad\forall\, v_h\in V_h^p(\Omega_k).$$

From the computational point of view, the function $u_{h,k}^{(2)}$ can be preliminarily calculated, but the function $u_{h,k}^{(1)}$ cannot be calculated until the function λ_h is obtained.

Define the discrete sesquilinear form

$$s_h(\lambda_h,\mu_h)=\sum_{\gamma_{kj}}\int_{\gamma_{kj}}\left(u_{h,k}^{(1)}(\lambda_h)-u_{h,j}^{(1)}(\lambda_h)\right)\cdot\overline{\left(u_{h,k}^{(1)}(\mu_h)-u_{h,j}^{(1)}(\mu_h)\right)}ds,\quad \lambda_h,\mu_h\in W_h^q(\gamma),$$

and the functional

$$l_h(\mu_h)=-\sum_{\gamma_{kj}}\int_{\gamma_{kj}}\left(u_{h,k}^{(2)}-u_{h,j}^{(2)}\right)\cdot\overline{\left(u_{h,k}^{(1)}(\mu_h)-u_{h,j}^{(1)}(\mu_h)\right)}ds,\quad \mu_h\in W_h^q(\gamma).$$

Therefore the discrete variational problem of (2.5) can be written as follows: find $\lambda_h\in W_h^q(\gamma)$ such that

$$s_h(\lambda_h,\mu_h)=l_h(\mu_h)\quad\forall\ \mu_h\in W_h^q(\gamma). \tag{2.7}$$

After λ_h is solved from (2.7), we can easily compute $u_{h,k}$ in parallel by (2.6) for every Ω_k. Define $u_h\in V_h^p(\mathcal{T}_h)$ by $u_h|_{\Omega_k}=u_{h,k}(\lambda_h)$ $(k=1,\dots,N)$. Then u_h should be an approximate solution of u. We would like to emphasize the discrete system (2.7) has relatively fewer degrees of freedom, so it is cheaper to solve.

Let $\mathcal{S}$ be the stiffness matrix associated with the sesquilinear form $s_h(\cdot,\cdot)$, and let b denote the vector associated with $l_h(\cdot)$. Then the discretization problem (2.7) leads to the algebraic system

$$\mathcal{S}X = b. \tag{2.8}$$

From the definition of the sesquilinear form $s_h(\cdot,\cdot)$, we know that the matrix $\mathcal{S}$ is Hermitian positive definite, so the system (2.8) can be solved by the PCG method with a positive definite preconditioner. The construction of an efficient preconditioner for $\mathcal{S}$ is an important task (see the next section).

Remark 2.1. Since each local finite element space $V_h^p(\Omega_k)$ consists of the standard polynomials, instead of solutions of a homogeneous Helmholtz equation in the plane wave methods, from the viewpoint of the algorithm the proposed method is practical for general nonhomogeneous Helmholtz equations in inhomogeneous media.

Remark 2.2. As in the traditional Lagrange multiplier method, we can derive another discrete system of λ_h by the constraints (for all element interfaces γ_{kj})

$$\langle u_{h,k} - u_{h,j}, \mu\rangle_{\gamma_{kj}} = 0 \quad \forall \mu \in W_h^q(\gamma).$$

However, the coefficient matrix of the resulting system is still indefinite as (1.1) (comparing the system (2.8)), which makes the solution of the system more difficult.

3. A domain decomposition preconditioner.

This section is devoted to the construction of a preconditioner $\mathcal{K}$ for $\mathcal{S}$. Thanks to the Hermitian positive definiteness of the matrix $\mathcal{S}$, we can construct a (Hermitian positive definite) substructuring preconditioner absorbing some ideas in the BDDC method first introduced in [8] (see section 1 for simple descriptions of the BDDC method). As we will see, the preconditioner designed in this section has essential differences from the one defined in the standard BDDC method.

For convenience, we will define the preconditioner in operator form. To this end, let $S : W_h^q(\gamma) \to W_h^q(\gamma)$ denote the discrete operator corresponding to the stiffness matrix $\mathcal{S}$, i.e.,

$$\langle S\lambda_h, \mu_h\rangle = s_h(\lambda_h, \mu_h) \quad \forall \lambda_h, \mu_h \in W_h^q(\gamma).$$

As usual we coarsen the partition as follows: let Ω be decomposed into a union of $D_1, D_2, \dots, D_{n_0}$ such that D_r is just a union of several elements $\Omega_k \in \mathcal{T}_h$ and satisfies (refer to the left graph of Figure 1)

$$\overline{\Omega} = \bigcup_{r=1}^{n_0} \overline{D}_r, \quad D_r \bigcap D_l = \emptyset \quad \text{for } r \neq l.$$

Let d denote the size of the subdomains $D_1, D_2, \dots, D_{n_0}$, and let $\mathcal{T}_d$ denote the partition comprised of the subdomains $\{D_r\}_{r=1}^{n_0}$.

For the construction of a substructuring preconditioner, we need to define a suitable "interface" Γ such that the degrees of freedom in all the subdomain interiors (i.e., $\Omega\backslash\Gamma$) can be eliminated independently for different subdomains. We first explain that, for the current situation, an interface Γ cannot be defined in the standard manner, where Γ is just a union of all the intersections of two neighboring subdomains. To this end, we want to investigate basis functions associated with two neighboring subdomains D_r and D_l, which have the nonempty common part $\partial D_r \cap \partial D_l$. Let e and e' be two fine edges that satisfy $e \in \bar{D}_r\backslash(\partial D_r \cap \partial D_l)$ and $e' \in \bar{D}_l\backslash(\partial D_r \cap \partial D_l)$, and let μ_e

and $\mu_{e'}$ denote two basis functions on e and e', respectively. It can be checked that if e and e' are close to $\partial D_r \cap \partial D_l$, then μ_e and $\mu_{e'}$ still have coupling, i.e., $s_h(\mu_e, \mu_{e'}) \neq 0$. This means that if the interface is defined in the standard manner, namely, is defined as the union of all $\partial D_r \cap \partial D_l$, the degrees of freedom in subdomain interiors cannot be eliminated independently. According to this observation, in the current situation an interface should be defined as a union of some elements instead of a union of some edges.

For each D_r, let $D_r^b \subset D_r$ be a union of the elements that touch the right and the lower boundary of $\partial D_r \backslash \partial\Omega$ (refer to the right graph in Figure 1). We define an interface as

$$\Gamma = \bigcup_{r=1}^{n_0} D_r^b.$$

Of course, the definition of such an interface is not unique (see [28] and [36] for similar definitions of interfaces), for example, an interface Γ can be defined as a union of all the elements that touch the standard interface $\cup_{k\neq j}(\partial D_k \cap \partial D_j)$.

In the following we describe various subspaces of $W_h^q(\gamma)$ and the corresponding solvers, which are needed in the construction of the desired preconditioner.

At first we define a subspace associated with each D_r. Set $D_r^0 = D_r \backslash D_r^b$ (see the right graph in Figure 1), and define the subspace for each subdomain D_r^0

$$W_h^q(D_r^0) = \left\{\mu \in W_h^q(\gamma) : \operatorname{supp} \mu \subset D_r^0\right\}, \ r = 1, 2, \ldots, n_0.$$

The local solver on the local space $W_h^q(D_r^0)$ is defined in the standard manner. Let $S_r^0 : W_h^q(D_r^0) \to W_h^q(D_r^0)$ be the restriction of S on $W_h^q(D_r^0)$,

$$\langle S_r^0 \varphi, \psi \rangle = \langle S\varphi, \psi\rangle, \quad \varphi, \psi \in W_h^q(D_r^0).$$

For the definition of solvers associated with the interface, we need to give a decomposition of the interface Γ. Let $\mathcal{V}_d$ denote the set of all the nodes corresponding to the coarse partition $\mathcal{T}_d$. For a coarse node $V \in \mathcal{V}_d$, let D_V denote the top left corner element that touches the vertex V (see the left graph in Figure 2).

Let D_{rl} denote the union of the elements that touch the intersection $\partial D_r \cap \partial D_l$ from the left side (or the upper side) but do not touch the lower (or the right) endpoints of $\partial D_r \cap \partial D_l$ (see the right graph in Figure 2).

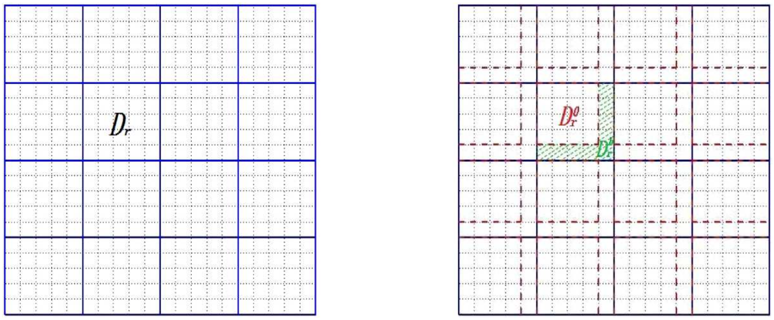

Fig. 1. *The left graph: each small square with dotted lines denotes an element Ω_k and each square with solid lines denotes a subdomain D_r. The right graph: each rectangle with red dotted lines denotes a subdomain D_r^0, each L-shape or reverse L-shape domain denotes a subdomain D_r^b (the green shade domain), where $D_r^b \cup D_r^0 = D_r$.*

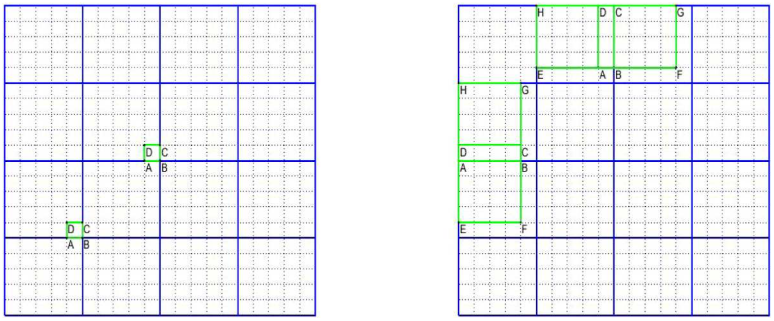

FIG. 2. *The left graph: the rectangle ABCD denotes a subdomain D_V. The right graph: the rectangle ABCD denotes a subdomain D_{rl} and the rectangle EFGH denotes a subdomain $\tilde{D}_{rl}$.*

It is easy to see that the interface can be decomposed into

$$\Gamma = \left(\bigcup_{rl} D_{rl} \right) \bigcup \left(\bigcup_{V \in \mathcal{V}_d} D_V \right).$$

Next we define local interface spaces. Set

$$\tilde{D}_{rl} = D_{rl} \cup D_r^0 \cup D_l^0$$

and define the discrete $s_h(\cdot,\cdot)$-harmonic extension spaces

$$W_H^q(\tilde{D}_{rl}) = \left\{ \mu \in W_h^q(\gamma) : supp\, \mu \subset \tilde{D}_{rl}; s_h(\mu, w) = 0, \forall w \in W_h^q\left(D_r^0\right) \cup W_h^q\left(D_l^0\right) \right\}.$$

Notice that the basis functions of these local spaces are not given explicitly, so the variational problems defined on these spaces cannot be solved in a direct manner. In order to overcome this difficulty, instead of computing such basis functions, as usual (see, for example, [8]) we transform the corresponding local interface problem into a residual equation, which is defined on the natural restriction space of the global space $W_h^q(\gamma)$ on the subdomain $\tilde{D}_{rl}$ (such a residual equation will be described exactly in Step 2 of Algorithm 3.1). However, solution of the residual equation is expensive since the restriction space contains many more basis functions than each local space $W_h^q(D_r^0)$, which is defined on a smaller subdomain D_r^0 than $\tilde{D}_{rl}$.

In order to decrease the cost of calculation, we choose to reduce the sizes of the subdomains $\tilde{D}_{rl}$ and define discrete $s_h(\cdot,\cdot)$-harmonic on the reduced subdomains. We reduce $\tilde{D}_{rl}$ to $\tilde{D}_{rl}^{half}$ such that the resulting subdomains have almost the same size d with D_r (see Figure 3).

Define the local spaces

$$W_h^q\left(\tilde{D}_{rl}^{half}\right) = \left\{ \mu \in W_h^q(\gamma) : \text{supp}\, \mu \subset \tilde{D}_{rl}^{half} \right\}.$$

For $\mu \in W_H^q(\tilde{D}_{rl})$, define $\mu_{rl}^{half} \in W_h^q(\tilde{D}_{rl}^{half})$ such that $\mu_{rl}^{half}|_{D_{rl}} = \mu|_{D_{rl}}$ and μ_{rl}^{half} is discrete $s_h(\cdot,\cdot)$-harmonic in the complement domain $\tilde{D}_{rl}^{half} \backslash D_{rl}$.

Define the discrete operator $K_{rl}^0 : W_H^q(\tilde{D}_{rl}) \to W_H^q(\tilde{D}_{rl})$ by

$$\langle K_{rl}^0 \mu, w \rangle = s_h\left(\mu_{rl}^{half}, w_{rl}^{half}\right),\ \mu \in W_H^q(\tilde{D}_{rl})\ \forall w \in W_H^q(\tilde{D}_{rl}).$$

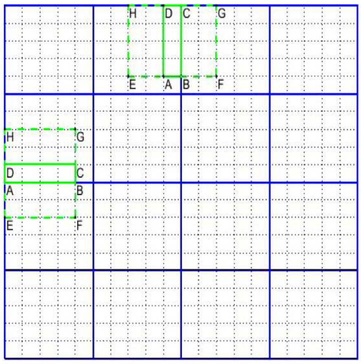

FIG. 3. *The rectangle ABCD denotes a subdomain D_{rl} and the rectangle EFGH denotes a subdomain $\tilde{D}_{rl}^{half}$.*

Notice that the action of $(K_{rl}^0)^{-1}$ is implemented by solving a residual equation defined on the "half" space $W_h^q(\tilde{D}_{rl}^{half})$ (see Algorithm 3.1), so K_{rl}^0 can be regarded as an "inexact" local interface solver based on the "compressed" harmonic extension μ_{rl}^{half} (refer to [28]). It is easy to see that the dimension of $W_h^q(\tilde{D}_{rl}^{half})$ is about half of the dimension of $W_h^q(\tilde{D}_{rl})$ and almost equals the dimension of $W_h^q(D_r^0)$. Then almost the same cost is needed for the solution of each subproblem in Step 2 and Step 1 (and Step 3) of Algorithm 3.1, which make the loading balance be guaranteed in parallel calculation (in applications, we choose $d \approx \sqrt{h}$).

Finally we construct a coarse space $W_d^q(\gamma)$ by some local energy minimizations.

For a coarse node $V \in \mathcal{V}_d$, let $\phi_V^{(m)}$ be a basis function in the subspace

$$W_h^q(D_V) = \{\mu|_{D_V} : \mu \in W_h^q(\gamma)\}.$$

Since the function $\phi_V^{(m)}$ is well defined only on the fine edges of D_V, we need to extend $\phi_V^{(m)}$ in a suitable manner such that $\phi_V^{(m)}$ has definitions on all the fine edges of $\mathcal{T}_h$. The desired coarse space will be spanned by the extensions of all $\phi_V^{(m)}$.

Let $\tilde{\phi}_V^{(m)}$ be the initial extension of $\phi_V^{(m)}$ such that $\tilde{\phi}_V^{(m)}$ is $s_h(\cdot,\cdot)$-harmonic on each subspace $W_h^q(D_r^0)$ and vanishes on all the fine edges in $\Gamma \backslash D_V$. In order to define further extension of $\tilde{\phi}_V^{(m)}$, let Γ_V denote a union of the coarse edges that touch the vertex V. For each $\Gamma_{rl} \in \Gamma_V$, let $\Phi_{V,rl}^{(m)} \in W_h^q(\tilde{D}_{rl}^{half})$ be the solution of the minimization problem

$$\min_{\Psi \in W_h^q(\tilde{D}_{rl}^{half})} \left\{ s_h^{(r)}\left(\tilde{\phi}_V^{(m)} + \Psi, \tilde{\phi}_V^{(m)} + \Psi\right) + s_h^{(l)}\left(\tilde{\phi}_V^{(m)} + \Psi, \tilde{\phi}_V^{(m)} + \Psi\right)\right\}, \tag{3.1}$$

where $s_h^{(r)}(\cdot,\cdot)$ denotes the restriction of $s_h(\cdot,\cdot)$ on the fine edges on D_r. Then $\Phi_{V,rl}^{(m)} \in W_h^q(\tilde{D}_{rl}^{half})$ can be obtained by solving the local equation

$$\sum_{k=r,l} s_h^{(k)}\left(\Phi_{V,rl}^{(m)}, v\right) = -\sum_{k=r,l} s_h^{(k)}\left(\tilde{\phi}_V^{(m)}, v\right) \quad \forall v \in W_h^q\left(\tilde{D}_{rl}^{half}\right). \tag{3.2}$$

Define

$$\Phi_V^{(m)} = \tilde{\phi}_V^{(m)} + \sum_{\Gamma_{rl} \in \Gamma_V} R_{rl}^t \Phi_{V,rl}^{(m)}, \tag{3.3}$$

where R_{rl}^t denotes the zero extension operators from $W_h^q(\tilde{D}_{rl})$ into $W_h^q(\gamma)$. The coarse space $W_d^q(\gamma)$ is spanned by all the basis functions $\Phi_V^{(m)}$, namely,

$$W_d^q(\gamma) = \text{span}\left\{\Phi_V^{(m)}\right\}.$$

Let the coarse solver $S_d : W_d^q(\gamma) \to W_d^q(\gamma)$ be the discrete operator which is the restriction of S on $W_d^q(\gamma)$ as usual.

Now we can define the preconditioner $K : W_h^q(\gamma) \to W_h^q(\gamma)$ as

$$K^{-1} = \sum_r (S_r^0)^{-1} Q_r + \sum_{\Gamma_{rl}} (K_{rl}^0)^{-1} Q_{rl} + S_d^{-1} Q_d,$$

where Q_r, Q_{rl}, and Q_d denote the L^2 projectors into $W_h^q(D_r^0), W_H^q(\tilde{D}_{rl})$, and $W_d^q(\gamma)$, respectively.

The action of the preconditioner K^{-1} can be described by the following algorithm.

Algorithm 3.1. *For $\xi \in W_h^q(\gamma)$, the solution $\lambda_\xi = K^{-1}\xi \in W_h^q(\gamma)$ can be obtained as follows:*

Step 1. *Computing $\lambda_r^0 \in W_h^q(D_r^0)$ in parallel by*

$$s_h^{(r)}\left(\lambda_r^0, \mu_h\right) = \langle \xi, \mu_h \rangle \quad \forall \mu_h \in W_h^q(D_r^0),\ r = 1, 2, \ldots, n_0.$$

Step 2. *Computing $\lambda_{rl} \in W_h^q(\tilde{D}_{rl}^{half})$ in parallel by*

$$\sum_{k=r,l} s_h^{(k)}(\lambda_{rl}, \mu_h) = \langle \xi, \mu_h \rangle - \sum_{k=r,l} s_h^{(k)}\left(\lambda_r^0, \mu_h\right) \quad \forall \mu_h \in W_h^q\left(\tilde{D}_{rl}^{half}\right).$$

Step 3. *Computing $\lambda_d \in W_d^q(\gamma)$ by*

$$s_h(\lambda_d, \mu_h) = \langle \xi, \mu_h \rangle - \sum_r s_h^{(r)}\left(\lambda_r^0, \mu_h\right) \quad \forall \mu_h \in W_d^q(\gamma).$$

Step 4. *Set $\phi = \sum \lambda_{rl} + \lambda_d$, and compute harmonic extensions $\lambda_r^H \in W_h^q(D_r)$ for all r in parallel, such that $\lambda_r^H = \phi$ on D_r^b and satisfies*

$$s_h^{(r)}\left(\lambda_r^H,\ \mu_h\right) = 0 \quad \forall \mu_h \in W_h^q\left(D_r^0\right),\ r = 1, 2, \ldots, n_0.$$

Step 5. *Computing*

$$\lambda_\xi = \sum_r \lambda_r^0 + \sum_r \lambda_r^H.$$

Remark 3.1. The minimization problem (3.1) is different from that in the BDDC method. In the BDDC method, each minimization problem which determines coarse basis functions is defined on one subdomain, so the solutions of the two minimization problems associated with two neighboring subdomains have different values on their common interface. In order to define coarse basis functions, in the BDDC method one has to compute some average of the values of the two solutions on the common interface. Since the minimization problem (3.1) is defined on the subdomain $\tilde{D}_{rl}^{half}$, the solution of this minimization problem has a unique value on the interface D_{rl} and the coarse basis functions can be directly obtained by (3.3). We found that if minimization problems are defined as in the BDDC method, then the resulting preconditioner is unstable.

Remark 3.2. Since the stiffness matrix of S_h^r has almost the same structure as the stiffness matrix $\mathcal{S}$ of the global system, the condition number of the stiffness matrix of S_h^r cannot be significantly decreased compared to the original stiffness matrix $\mathcal{S}$. However, such a local stiffness matrix has much lower order than $\mathcal{S}$, so each subproblem in Step 1 of Algorithm 3.1 can be solved in a direct manner (using LU decomposition), which is not sensitive to the condition number of this local stiffness matrix, where the global Step 1 is implemented in parallel. Notice that the variational problem (3.2) and the variational problem in Step 2 of Algorithm 3.1 correspond to the same stiffness matrix (with different right hands only). Thus the cost of computation for the coarse basis functions by solving every subproblem (3.2) in parallel only increases a little by using LU decomposition made in Step 2 for each local stiffness matrix (when Step 2 is implemented in the direct method).

Remark 3.3. When Ω is a general domain instead of a rectangle, we can first define a domain decomposition such that every subdomain D_r is a polygon, and then define a triangle partition on each subdomain D_r, all of which constitute a partition $\mathcal{T}_h$ of Ω. In this situation, the "interface" Γ and the reduced subdomain $\tilde{D}_{rl}^{half}$ can be defined in a similar manner, but their shapes may be more complicated.

4. Main results.

Throughout this paper, C denotes a generic positive constant that may have different values in different occurrences, where C is always independent of $\omega, h, p,$ and q but may depend on the shape of Ω and the maximal value and minimal value of $c(\mathbf{x})$ on Ω. Before presenting the main results, we give several assumptions.

Assumption 1. The domain Ω is a strictly star-shaped; the function $c(\mathbf{x})$ belongs to $W^{1,\infty}(\Omega)$.

The first condition in the above assumption appeared in many existing works to build error estimates with little wave number pollution (see, for example, [9] and [32]); the second condition was used in [2] to build stability results of an analytic solution.

Assumption 2. The mesh size h satisfies the condition: $\omega h \leq C_0$ with a possibly small constant C_0 independent of ω, h, p, and q (but may depend on the shape of Ω and the maximal value and minimal value of the function $c(\mathbf{x})$).

The above assumption is weaker than that required in analysis of the HDG-type methods. The following assumption has no restriction to the proposed method.

Assumption 3. The parameter ρ in the variational formula is not large: $\rho \leq C_0 \min\{1, \omega^2 h\}$ for a possibly small constant C_0 independent of ω, h, p, and q.

From the viewpoint of the algorithm, all the discretization methods based on polynomial basis functions are practical for the case with variable wave numbers (in inhomogeneous media). However, almost existing error estimates with little wave number pollution were established only for the case with constant wave numbers (see, for example, [9] and [32]). The main reason is that one does not know whether the result on "stable decomposition of solution," which was built in Theorem 4.10 of [32] and plays a key role in the derivations of good error estimates, still holds for the case with variable wave numbers. In this paper we try to investigate the possibility that the proposed method possesses error estimates with little wave number pollution even for the case of variable wave numbers. In order to cover the case of variable wave number, we introduce an additional assumption.

For $\tilde{f} \in L^2(\Omega)$, consider a dual problem with Robin-type boundary condition

$$\begin{cases} -\Delta\phi - \kappa^2\phi = \tilde{f} & \text{in } \Omega, \\ \dfrac{\partial \phi}{\partial n} - i\kappa\phi = 0 & \text{on } \partial\Omega. \end{cases} \tag{4.1}$$

Let $\tilde{V}_h^p(\mathcal{T}_h) \subset H^1(\Omega)$ denote the continuous piecewise p-order polynomial space associated with the partition $\mathcal{T}_h$.

Assumption 4. The finite element solution $\phi_h \in \tilde{V}_h^p(\mathcal{T}_h)$ of (4.1) possesses a weak convergence with respect to p for large p,

$$||\nabla(\phi - \phi_h)||_{0,\Omega} + \omega||(\phi - \phi_h)||_{0,\Omega} \lesssim p^{-\frac{1}{2}}||\tilde{f}||_{0,\Omega}. \tag{4.2}$$

This assumption can be met easily when $c(\mathbf{x})$ is a constant. In fact, for this case the following stronger result has been built in [32, Corollary 5.10] under the assumptions that Ω is a strictly star-shaped domain with an analytic boundary and the discretization parameters satisfy the mild conditions $\frac{\omega h}{p} \le C_0$ and $p \ge 1 + c_0 \log \omega$:

$$||\nabla(\phi - \phi_h)||_{0,\Omega} + \omega||(\phi - \phi_h)||_{0,\Omega} \lesssim hp^{-1}||\tilde{f}||_{0,\Omega}. \tag{4.3}$$

Therefore, when $c(\mathbf{x})$ is a constant, Assumption 4 should be changed into: Ω has an analytic boundary; $p \ge 1 + c_0 \log \omega$ (since Assumption 2 implies $\frac{\omega h}{p} \le C_0$).

Whether the error estimate (4.3) still holds for the case of variable $c(\mathbf{x})$ seems an open problem, but the weak error estimate (4.2) should be valid even for a variable $c(\mathbf{x})$ under the above assumptions. In this situation, Assumption 4 can be replaced by the conditions that Ω has an analytic boundary and $p \ge 1 + c_0 \log \omega$.

Now we list the main results, which will be proved in the next section. First, we give a result about the local inf-sup condition.

THEOREM 4.1. *Let $q \ge 1$ and $p \ge q + 2$. For any $\mu \in W_h^q(\partial\Omega_k \backslash \partial\Omega)$, there exists a nonzero function $v \in V_h^p(\partial\Omega_k)$ such that*

$$\langle \mu, v\rangle_{\partial\Omega_k\backslash\partial\Omega} \ge Cq^{-\frac{1}{2}}||\mu||_{0,\partial\Omega_k\backslash\partial\Omega}||v||_{0,\partial\Omega_k\backslash\partial\Omega}, \tag{4.4}$$

where C is a constant independent of ω, h, p, and q.

Next we give a result on the coerciveness of the sesquilinear form $s_h(\cdot,\cdot)$, which implies that the discrete problem (2.7) is well-posed.

THEOREM 4.2. *Let Assumptions 1–4 be satisfied. Suppose $q \ge 1$ and $p \ge q + 2$. Then, for any $\mu_h \in W_h^q(\gamma)$, we have*

$$s_h(\mu_h, \mu_h) \ge C\omega^{-2}h^2p^{-1}q^{-1}\sum_{\gamma_{kj}} \|\mu_h\|^2_{0,\gamma_{kj}}, \tag{4.5}$$

where C is a constant independent of ω, h, p, and q.

Finally, we give error estimates of the approximation u_h. Define the subspace $W^{r-\frac{1}{2}}(\gamma) = \prod_{k \ne j} H^{r-\frac{1}{2}}(\gamma_{kj})$, which is equipped with the norm

$$||\mu||_{r-\frac{1}{2},\gamma} = \left(\sum_{\gamma_{kj}} ||\mu||^2_{r-\frac{1}{2},\gamma_{kj}}\right)^{\frac{1}{2}} \quad \forall\ \mu \in W^{r-\frac{1}{2}}(\gamma).$$

For ease of notation, we also define the seminorms ($r \geq 1$)

$$|\mu|_{r-\frac{1}{2},\gamma} = \left(\sum_{\gamma_{kj}} |\mu|^2_{r-\frac{1}{2},\gamma_{kj}} \right)^{\frac{1}{2}}, \qquad \mu \in W^{r-\frac{1}{2}}(\gamma),$$

and

$$|v|_{r,\Omega} = \left(\sum_{k=1}^{N} |v|^2_{r,\Omega_k} \right)^{\frac{1}{2}}, \qquad v \in \prod_{k=1}^{N} H^r(\Omega_k).$$

Set

$$H^{r+1}(\mathcal{T}_h) = \left\{ v \in H^2(\Omega): \quad v|_{\Omega_k} \in H^{r+1}(\Omega_k),\ \frac{\partial v}{\partial \mathbf{n}}|_{\gamma_{kj}} \in H^{r-\frac{1}{2}}(\gamma_{kj}) \right\}.$$

THEOREM 4.3. *Suppose that $q \geq 1$ and $p \geq q+2$. Let Assumptions* 1–4 *be satisfied. Assume that the analytical solution u of the Helmholtz problem* (1.1) *belongs to $H^{r+1}(\mathcal{T}_h)$ with $1 \leq r \leq q$ ($r \in \mathbb{N}$). Then the approximate solution u_h defined in subsection* 2.3 *satisfies*

$$|u - u_h|_{1,\Omega} \leq Ch^{r-1} \left(p^{-r}|u|_{r+1,\Omega} + q^{-r}|\lambda|_{r-\frac{1}{2},\gamma} \right) \tag{4.6}$$

and

$$||u - u_h||_{0,\Omega} \leq C\omega^{-1}h^{r-1} \left(p^{-r}|u|_{r+1,\Omega} + q^{-r}|\lambda|_{r-\frac{1}{2},\gamma} \right), \tag{4.7}$$

where C is a constant independent of ω, h, p, and q.

Remark 4.1. Comparing Theorem 4.3 with Theorem 3.15 in [23] (and Theorem 3.4 in [33]), we can see that the proposed discretization method possesses almost the same L^2 convergence order as the plane wave methods (for the case of constant wave number), which have fast convergence and small "wave number pollution." As pointed out in section 1, the standard plane wave methods are not practical for the case with variable wave numbers, but there is not this problem for the proposed method (see Remark 2.1).

5. Proof of the main results. This section is devoted to the proofs of Theorems 4.1, 4.2, and 4.3. Since the approximate solutions (u_h, λ_h) neither satisfy a mixed variational problem (comparing the Lagrange multiplier methods) nor satisfy a hybridizable variational problem (comparing the HDG methods), so the results cannot be proved by the techniques developed in existing works. As we shall see, the proofs are very technical, so this section is divided into three subsections, in which many auxiliary results need to be established.

In order to shorten the length of this article, we only give detailed proofs of two key auxiliary results and three theorems themselves, but we omit the proofs of the other auxiliary results (the proofs can be found in the e-print [24]).

For ease of notation, we use the shorthand notation $x \lesssim y$ and $y \gtrsim x$ for the inequality $x \leq Cy$ and $y \geq Cx$, where C is a constant independent of ω, h, p, and q but may depend on the shape of Ω and the maximal value and minimal value of $c(\mathbf{x})$ on Ω. Throughout this section, we use p and q to denote two positive integers.

We first verify the local inf-sup condition given in Theorem 4.1 by using Jacobi polynomials.

5.1. Analysis on the local inf-sup condition. The main difficulty for the proof of Theorem 4.1 is the fact that the functions in $W_h^q(\partial\Omega_k\backslash\partial\Omega)$ are defined independently for different edges of $\partial\Omega_k$ and may be discontinuous at the vertices of $\partial\Omega_k$ but the functions in $V_h^p(\partial\Omega_k)$ are defined globally on $\partial\Omega_k$ and must be continuous at the vertices of $\partial\Omega_k$. Because of this, we have to split the q-order polynomial space into a sum of two polynomial subspaces, one of which consists of all the q-order polynomials vanishing at the vertices of $\partial\Omega_k$, so that the construction of a function v satisfying (4.4) for $\mu \in W_h^q(\partial\Omega_k\backslash\partial\Omega)$ becomes easier by using this splitting and Jacobi polynomial basis functions of this subspace (we can require that such function v vanishes at all the vertices of $\partial\Omega_k$).

Set $J = [0, 1]$ and let $\mathcal{P}_q$ stand for the space of all polynomials on J with orders $\leq q$. First, we give a space decomposition of $\mathcal{P}_q$ on J

$$\mathcal{P}_q = \mathcal{P}_1^* + \mathcal{P}_q' \quad \text{with } \mathcal{P}_1^* \perp \mathcal{P}_q'. \tag{5.1}$$

The specific definition of $\mathcal{P}_1^*$ and $\mathcal{P}_q'$ will be given next.

Let $\mathcal{P}_1$ and $\mathcal{P}_q'$ denote the linear part and high-order part of $\mathcal{P}_q$, respectively. Then the two basis functions of $\mathcal{P}_1$ are $\phi_1 = x$ and $\phi_2 = 1 - x$. If $q = 1$, then $\mathcal{P}_q' = \emptyset$ and set $P_1^* = \{\phi_1, \phi_2\}$. For a unified description below, we define $\phi_1^* = \phi_1$ and $\phi_2^* = \phi_2$ when $q = 1$ and write $P_1^* = \{\phi_1^*, \phi_2^*\}$.

In the following we assume that $q \geq 2$. Let $\{\psi_k\}_{k=1}^{q-1}$ denote the basis functions of the subspace $\mathcal{P}_q'$. Define

$$\phi_1^* = \phi_1 - \sum_{k=1}^{q-1} \alpha_k \psi_k \quad \text{and} \quad \phi_2^* = \phi_2 - \sum_{k=1}^{q-1} \beta_k \psi_k. \tag{5.2}$$

Here the numbers $\{\alpha_k\}$ and $\{\beta_k\}$ are determined by $\langle \phi_1^*, \psi_k\rangle_J = 0$ and $\langle \phi_2^*, \psi_k\rangle_J = 0$. Apparently we can get

$$(\alpha_1, \alpha_2, \ldots, \alpha_{q-1})^t = A^{-1}b_1 \text{ and } (\beta_1, \beta_2, \ldots, \beta_{q-1})^t = A^{-1}b_2,$$

where $A = (\langle \psi_k, \psi_j\rangle_J)_{(q-1)*(q-1)}$ and $b_1 = (\langle \phi_1, \psi_k\rangle_J)_{(q-1)*1}, b_2 = (\langle \phi_2, \psi_k\rangle_J)_{(q-1)*1}$, which means ϕ_1^*, ϕ_2^* are uniquely determined. Let $\mathcal{P}_1^* = \text{span}\{\phi_1^*, \phi_2^*\}$, which satisfies the space decomposition (5.1).

Next we give a set of orthogonal basis functions of $\mathcal{P}_q' = \text{span}\{\psi_1, \psi_2, \ldots, \psi_{q-1}\}$ $(q \geq 2)$. In order to explicitly write the orthogonal basis functions and conveniently compute the involved integrations, we use a set of Jacobi polynomials $\{G_k\}$ (see [39]). For convenience, we let the coefficient of the first Jacobi polynomial be 1. Then, for $p \geq q + 2$, the Jacobi polynomials are defined as

$$G_k = (-1)^{k-1} \frac{(k+3)!}{(2k+2)!} x^{-2}(1-x)^{-2} \frac{d^{k-1}}{dx^{k-1}} \left(x^{k+1}(1-x)^{k+1}\right), \quad k = 1, \ldots, p. \tag{5.3}$$

It is known that

$$\int_0^1 x^2(1-x)^2 G_k G_j dx = \begin{cases} 0, & k \neq j, \\ \frac{(k-1)!(k+1)!^2(k+3)!}{(2k+2)!(2k+3)!}, & k = j. \end{cases}$$

We also have the recursion relations

$$\begin{cases} G_1 = 1, \ G_2 = x - \dfrac{1}{2}, \\ G_k = \left(x - \dfrac{1}{2}\right) G_{k-1} - \dfrac{(k-2)(k+2)}{4(2k-1)(2k+1)} G_{k-2}, \quad 3 \leq k \leq p. \end{cases}$$

Define

$$\psi_k = \frac{(2k+2)!}{(k-1)!(k+1)!}x(1-x)G_k, \quad k = 1, \dots, q, \dots, p. \tag{5.4}$$

It is clear that $\psi_k(0) = \psi_k(1) = 0$ and

$$\int_0^1 \psi_k \psi_j dx = \begin{cases} 0, & k \neq j, \\ \frac{k(k+1)(k+2)(k+3)}{2k+3}, & k = j. \end{cases} \tag{5.5}$$

Furthermore $\{\psi_k\}_{k=1}^{p-1}$ satisfy the recursion relations

$$\begin{cases} \psi_1 = 12x(1-x), \ \psi_2 = 120x(1-x)\left(x - \dfrac{1}{2}\right), \\ \psi_k = \dfrac{2(2k+1)}{k-1}\left(x - \dfrac{1}{2}\right)\psi_{k-1} - \dfrac{k+2}{k-1}\psi_{k-2}, \quad 3 \leq k \leq p. \end{cases} \tag{5.6}$$

The functions $\{\psi_k\}_{k=1}^{q-1}$ constitute a set of orthogonal bases of $\mathcal{P}'_q$.

The following result can be directly obtained by mathematical induction and the recursion relations (5.6).

Lemma 5.1. *Let $q \geq 2$. For $\phi_1 = x$, $\phi_2 = 1 - x$ and $\{\psi_k\}_{k=1}^{q-1}$ defined by* (5.4), *we have*

$$\langle \phi_1, \psi_k \rangle_J = 1 \ \textit{and} \ \langle \phi_2, \psi_k \rangle_J = (-1)^{k-1}, \ k = 1, 2, \dots, q-1.$$

It is easy to see that the following two equalities hold for any positive integer m:

$$\sum_{k=1}^{m} \frac{2k+3}{k(k+1)(k+2)(k+3)} = \frac{1}{3} - \frac{1}{(m+1)(m+3)} \tag{5.7}$$

and

$$\sum_{k=1}^{m} \frac{(-1)^{k-1}(2k+3)}{k(k+1)(k+2)(k+3)} = \frac{1}{6} + \frac{(-1)^{m-1}}{(m+1)(m+2)(m+3)}. \tag{5.8}$$

The following result can be reduced by using (5.2), together with (5.5), (5.7), Lemma 5.1, and (5.8).

Lemma 5.2. *Let $\phi_1^* = x$ and $\phi_2^* = 1 - x$ for $q = 1$. For $q \geq 2$, let $\{\phi_1^*, \phi_2^*\}$ be defined as* (5.2) *with ψ_k given by* (5.4)*. Then we have*

$$\langle \phi_1^*, \phi_1^* \rangle_J = \langle \phi_2^*, \phi_2^* \rangle_J = \frac{1}{q(q+2)}, \quad \langle \phi_1^*, \phi_2^* \rangle_J = \frac{(-1)^{q-1}}{q(q+1)(q+2)}, \tag{5.9}$$

and

$$||a_1\phi_1^*||_{0,J}^2 + ||a_2\phi_2^*||_{0,J}^2 \leq \frac{q+1}{q}||a_1\phi_1^* + a_2\phi_2^*||_{0,J}^2 \quad \forall\, a_1, a_2 \in \mathbb{R}. \tag{5.10}$$

Proof of Theorem 4.1. For an element Ω_k, let n_k denote the number of the edges of Ω_k and write its boundary as $\partial\Omega_k = \bigcup_{j=1}^{n_k} J_j$, where J_j is the jth edge of Ω_k. If

$J_j \subset \partial\Omega_k \cap \partial\Omega$, we set $\mu|_{J_j} = 0$. Then we only need to prove that for any $\mu \in W_h^q(\partial\Omega_k)$, there exists a function $v \in V_h^p(\partial\Omega_k)$ such that

$$\langle \mu, v\rangle_{\partial\Omega_k} \geq C_{p,q}||\mu||_{0,\partial\Omega_k}||v||_{0,\partial\Omega_k},$$

where $C_{p,q}$ is a positive constant which may only depend on p and q. Since Ω_k is regular, we can simply set $J_j = [0, 1]$ by the scaling transformation.

By the space decomposition (5.1), the function $\mu \in W_h^q(\partial\Omega_k)$ can be written as

$$\mu|_{J_j} = \sum_{k=1}^{q-1} \xi_k\psi_k + a_1\phi_1^* + a_2\phi_2^*, \quad a_1, a_2, \xi_k \in \mathbb{R}, \tag{5.11}$$

where $\{\phi_1^*, \phi_2^*\}$ are two basis functions of $\mathcal{P}_1^*$, and $\{\psi_k\}_{k=1}^{q-1}$ denote the orthogonal basis functions of $\mathcal{P}_q'$; see (5.4). Then we choose

$$v|_{J_j} = \sum_{k=1}^{q-1} \xi_k\psi_k + \sum_{k=q}^{p-1} \frac{\langle a_1\phi_1^* + a_2\phi_2^*, \psi_k\rangle_{J_j}}{\langle \psi_k, \psi_k\rangle_{J_j}}\psi_k, \tag{5.12}$$

where $\{\psi_k\}_{k=1}^{p-1}$ are defined by (5.4). It is clear that $v|_{J_j}(0) = v|_{J_j}(1) = 0$. Then we have $v \in V_h^p(\partial\Omega_k)$.

Using the orthogonality condition (5.5) yields

$$\langle \mu, v\rangle_{J_j} = \sum_{k=1}^{q-1} \xi_k^2\langle \psi_k, \psi_k\rangle_{J_j} + \sum_{k=q}^{p-1} \frac{\langle a_1\phi_1^* + a_2\phi_2^*, \psi_k\rangle_{J_j}^2}{\langle \psi_k, \psi_k\rangle_{J_j}} = ||v||_{0,J_j}^2.$$

It follows that

$$\langle \mu, v\rangle_{\partial\Omega_k} = \sum_{j=1}^{n_k} \langle \mu, v\rangle_{J_j} = \sum_{j=1}^{n_k} ||v||_{0,J_j}^2 = ||v||_{0,\partial\Omega_k}^2.$$

Thus, we only need to prove there exists $C_{p,q}$, such that

$$||v||_{0,\partial\Omega_k} \geq C_{p,q}||\mu||_{0,\partial\Omega_k} \quad \text{or} \quad ||v||_{0,J_j} \geq C_{p,q}||\mu||_{0,J_j}.$$

To do this, we use (5.5), (5.9), and Lemma 5.1, which gives

$$\begin{aligned} ||v||_{0,J_j}^2 &= \sum_{k=1}^{q-1} \xi_k^2\langle \psi_k, \psi_k\rangle_{J_j} + \sum_{k=q}^{p-1} \frac{(2k+3)(a_1 + (-1)^{k-1}a_2)^2}{k(k+1)(k+2)(k+3)} \\ &= \sum_{k=1}^{q-1} \xi_k^2\langle \psi_k, \psi_k\rangle_{J_j} + \left(\frac{1}{q(q+2)} - \frac{1}{p(p+2)}\right)(a_1^2 + a_2^2) \\ &\quad + \left(\frac{(-1)^{q-1}}{q(q+1)(q+2)} - \frac{(-1)^{p-1}}{p(p+1)(p+2)}\right)2a_1a_2. \end{aligned}$$

Then, using (5.9) again, we have

$$\begin{aligned} ||\mu||_{0,J_j}^2 &= \sum_{k=1}^{q-1} \xi_k^2\langle \psi_k, \psi_k\rangle_{J_j} + a_1^2\langle \phi_1^*, \phi_1^*\rangle_{J_j} + a_2^2\langle \phi_2^*, \phi_2^*\rangle_{J_j} + 2a_1a_2\langle \phi_1^*, \phi_2^*\rangle_{J_j} \\ &= \sum_{k=1}^{q-1} \xi_k^2\langle \psi_k, \psi_k\rangle_{J_j} + \frac{1}{q(q+2)}(a_1^2 + a_2^2) + \frac{(-1)^{q-1}}{q(q+1)(q+2)}2a_1a_2. \end{aligned}$$

So we choose

$$C_{p,q}^2 = \begin{cases} 1 - \frac{(q+1)(q+2)}{(p+1)(p+2)}, & p+q = \text{even}, \\ 1 - \frac{(q+1)(q+2)}{p(p+1)}, & p+q = \text{odd}, \end{cases}$$

which satisfies

$$||v||_{0,J_j}^2 \geq C_{p,q}^2 ||\mu||_{0,J_j}^2 .$$

Since $q \geq 1$ and $p \geq q+2$, we have $C_{p,q}^2 \geq \frac{2}{q+3}$ and so $||v||_{0,J_j}^2 \gtrsim q^{-1}||\mu||_{0,J_j}^2$, namely,

$$\langle \mu, v\rangle_{J_j} \gtrsim q^{-\frac{1}{2}}||\mu||_{0,J_j}||v||_{0,J_j}.$$

It concludes the proof of the local inf-sup condition given by (4.4). □

Remark 5.1. If $q \geq 2$ and $p = 2q$, we have $C_{p,q}^2 \geq \frac{1}{2}$, which implies that

$$\langle \mu, v\rangle_{J_j} \gtrsim ||\mu||_{0,J_j}||v||_{0,J_j}.$$

Then, when $q \geq 2$ and $p = 2q$, the inequality (4.4) can be replaced by the optimal inf-sup condition

$$\langle \mu, v\rangle_{\partial\Omega_k\backslash\partial\Omega} \geq C||\mu||_{0,\partial\Omega_k\backslash\partial\Omega}||v||_{0,\partial\Omega_k}.$$

5.2. Analysis on the coerciveness. The proofs of Theorems 4.2 and 4.3 will depend on a *jump-controlled* stability estimate (which will be given by Proposition 5.1). A technical tool for the derivation of this stability estimate is a Poincaré-type inequality given by Lemma 5.4. In order to prove this Poincaré-type inequality, we have to develop a special technique: construct a globally continuous p-finite element function to “approximate” a piecewise continuous p-finite element function and derive a corresponding “approximate” result (Lemma 5.3). There seems to be no similar technique and result in the existing literature.

Let $\tilde{V}_h^p(\mathcal{T}_h) \subset H^1(\Omega)$ denote the continuous piecewise p-order finite element space associated with the partition $\mathcal{T}_h$. For a given function $v \in V_h^p(\mathcal{T}_h)$ ($\not\subseteq H^1(\Omega)$), we want to construct a correction function $\tilde{v} \in \tilde{V}_h^p(\mathcal{T}_h)$, which should satisfy the estimates stated in Lemma 5.3.

Let $v \in V_h^p(\mathcal{T}_h)$. For each element Ω_k, we set $v|_{\Omega_k} = v_k$, which denotes the restriction of v on the element Ω_k. We need only to define a suitable correction function $\tilde{v}_k$ of v_k for each Ω_k. After it is done, we then define the desired function $\tilde{v}$ such that $\tilde{v}|_{\Omega_k} = \tilde{v}_k$. For ease of understanding, we want to describe the basic idea for defining such function $\tilde{v}_k$. Consider the standard decomposition

$$v_k = v_k^0 + v_k^\partial,$$

where $v_k^\partial|_{\partial\Omega_k} = v_k|_{\partial\Omega_k}$ and $v_k^\partial \in V_h^p(\Omega_k)$ is the discrete harmonic extension of $v_k|_{\partial\Omega_k}$ into Ω_k. It is easy to see that $v_k^0|_{\partial\Omega_k} = v_k|_{\partial\Omega_k} - v_k^\partial|_{\partial\Omega_k} = 0$, which can be naturally extended into Ω. However, in general we have $v_k^\partial|_{\gamma_{kj}} \neq v_j^\partial|_{\gamma_{kj}}$, where $\gamma_{kj} = \partial\Omega_k \cap \partial\Omega_j$ is an element edge.

Since we require that the desired function $\tilde{v} \in H^1(\Omega)$, we need to define a correction $\tilde{v}_k^\partial$ of v_k^∂ in a special manner such that $\tilde{v}_k^\partial|_{\gamma_{kj}} = \tilde{v}_j^\partial|_{\gamma_{kj}}$. After it is done, we naturally define

$$\tilde{v}_k = v_k^0 + \tilde{v}_k^\partial,$$

where $\tilde{v}_k^\partial \in V_h^p(\Omega_k)$ is the discrete harmonic extension of $\tilde{v}_k^\partial|_{\partial\Omega_k}$ into Ω_k.

In the following we give a definition of $\tilde{v}_k^{\partial}|_{\partial\Omega_k}$. Let e denote an edge of $\partial\Omega_k$. When $e = \partial\Omega_k \cap \partial\Omega$, we simply define $\tilde{v}_k^{\partial}|_e = v_k^{\partial}|_e$. If $e = \partial\Omega_k \cap \partial\Omega_j$, we define $\tilde{v}_k^{\partial}|_e$ as follows.

As in the beginning of subsection 5.1, we can define the spaces $\mathcal{P}_1^*$ and $\mathcal{P}_p'$ on the edge e by the standard scaling technique. Then we have the decomposition

$$v_k^{\partial}|_e = v_{k1}^{\partial} + v_{k0}^{\partial},$$

where $v_{k1}^{\partial} \in \mathcal{P}_1^*$ and $v_{k0}^{\partial} \in \mathcal{P}_p'$. Let $\{\phi_1^{*e}, \phi_2^{*e}\}$ denote the two basis functions of $\mathcal{P}_1^*$, and let v_1 and v_2 denote the two endpoints of the edge e. It is easy to see that v_{k1}^{∂} can be written as

$$v_{k1}^{\partial} = v_k(\mathrm{v}_1)\phi_1^{*e} + v_k(\mathrm{v}_2)\phi_2^{*e}.$$

Set

$$\Lambda_{\mathrm{v}_i} = \{r,\ \Omega_r \text{ contains } \mathrm{v}_i \text{ as one of its vertices}\} \quad (i = 1, 2),$$

and let n_{v_i} denote the number of all the elements that contain v_i as their common vertex, namely, the dimension of set Λ_{v_i}. For $e = \partial\Omega_k \cap \partial\Omega_j$, define

$$\tilde{v}_{k1}^{\partial}|_e = \frac{1}{n_{\mathrm{v}_1}} \sum_{r\in\Lambda_{\mathrm{v}_1}} v_r(\mathrm{v}_1)\,\phi_1^{*e} + \frac{1}{n_{\mathrm{v}_2}} \sum_{r\in\Lambda_{\mathrm{v}_2}} v_r(\mathrm{v}_2)\,\phi_2^{*e} \tag{5.13}$$

and

$$\tilde{v}_{k0}^{\partial}|_e = \frac{1}{2}(v_{k0}^{\partial} + v_{j0}^{\partial}). \tag{5.14}$$

Now we define $\tilde{v}_k^{\partial}|_e = \tilde{v}_{k1}^{\partial}|_e + \tilde{v}_{k0}^{\partial}|_e$ for each $e \subset \partial\Omega_k$, and let $\tilde{v}_k^{\partial} \in V_h^p(\Omega_k)$ be the discrete harmonic extension of $\tilde{v}_k^{\partial}|_{\partial\Omega_k}$. From the definition of $\tilde{v}_{k1}^{\partial}$, we know that $\tilde{v}_k^{\partial}|_{\gamma_{kj}} = \tilde{v}_j^{\partial}|_{\gamma_{kj}}$. Thus we can define $\tilde{v}_k = \tilde{v}_k^{\partial} + v_k^0$. It is clear that $\tilde{v}_k|_{\gamma_{kj}} = \tilde{v}_j|_{\gamma_{kj}}$.

Finally we define $\tilde{v}$ by $\tilde{v}|_{\Omega_k} = \tilde{v}_k$ and we have $\tilde{v} \in \tilde{V}_h^p(\mathcal{T}_h)$.

The following result can be verified by Lemma 5.2 (replacing q by p).

LEMMA 5.3. *For $v \in V_h^p(\mathcal{T}_h)$, let $\tilde{v} \in \tilde{V}_h^p(\mathcal{T}_h)$ be defined above. Then we have*

$$\left(\sum_{k=1}^{N} |v - \tilde{v}|_{1,\Omega_k}^2 + h^{-2}||v - \tilde{v}||_{0,\Omega}^2\right)^{\frac{1}{2}} \lesssim h^{-\frac{1}{2}} p^{\frac{1}{2}} \left(\sum_{\gamma_{kj}} ||[v]||_{0,\gamma_{kj}}^2\right)^{\frac{1}{2}}. \tag{5.15}$$

In the following we want to build a Poincaré-type inequality for the functions in $V_h^p(\mathcal{T}_h)$ by Lemma 5.3.

LEMMA 5.4. *Let Assumptions 1–4 be satisfied. Assume that, for some $\lambda_h \in W_h^q(\gamma)$, the function $v \in V_h^p(\mathcal{T}_h)$ satisfies*

$$a^{(k)}(v, w) = \langle \pm\lambda_h, \overline{w}\rangle_{\partial\Omega_k \setminus \partial\Omega} \quad (k = 1, 2, \ldots, N) \quad \forall\, w \in V_h^p(\mathcal{T}_h). \tag{5.16}$$

Then

$$||v||_{0,\Omega}^2 \lesssim h^{-1} \sum_{\gamma_{kj}} ||[v]||_{0,\gamma_{kj}}^2. \tag{5.17}$$

Proof. A standard technique to estimate L^2 norm of a function is the introduction of a suitable dual problem (see, for example, [23]). Consider the dual problem

$$\begin{cases} -\Delta\phi-\kappa^2\phi=v & \text{in } \Omega,\\ \dfrac{\partial\phi}{\partial n}-i\kappa\phi=0 & \text{on } \partial\Omega. \end{cases} \tag{5.18}$$

Let $\phi\in H^1(\Omega)$ and $\phi_h\in\tilde{V}_h^p(\mathcal{T}_h)$ denote its weak solution and p-order finite element solution, which are defined respectively by

$$(\nabla\phi,\overline{\nabla\psi})_\Omega-(\kappa^2\phi,\overline{\psi})_\Omega-i\langle\kappa\phi,\overline{\psi}\rangle_{\partial\Omega}=(v,\overline{\psi})_\Omega\quad\forall\psi\in H^1(\Omega) \tag{5.19}$$

and

$$(\nabla\phi_h,\overline{\nabla\psi}_h)_\Omega-(\kappa^2\phi_h,\overline{\psi}_h)_\Omega-i\langle\kappa\phi_h,\overline{\psi}_h\rangle_{\partial\Omega}=(v,\overline{\psi}_h)_\Omega\quad\forall\psi_h\in\tilde{V}_h^p(\mathcal{T}_h). \tag{5.20}$$

Using (5.18) and Green's formula, we obtain

$$\begin{aligned} ||v||^2_{0,\Omega}&=\sum_{k=1}^N(v,\overline{v})_{\Omega_k}=\sum_{k=1}^N\big((v,-\overline{\Delta\phi})_{\Omega_k}-(v,\overline{\kappa^2\phi})_{\Omega_k}\big)\\ &=\sum_{k=1}^N\big((\nabla v,\overline{\nabla\phi})_{\Omega_k}-\langle v,\overline{\nabla\phi\cdot n}\rangle_{\partial\Omega_k}-(\kappa^2v,\overline{\phi})_{\Omega_k}\big)\\ &=\sum_{k=1}^N(\nabla v,\overline{\nabla\phi})_{\Omega_k}-\sum_{k=1}^N(\kappa^2v,\overline{\phi})_{\Omega_k}-\sum_{\gamma_{kj}}\langle[v],\overline{\nabla\phi\cdot n}\rangle_{\gamma_{kj}}-\langle v,\overline{i\kappa\phi}\rangle_{\partial\Omega}\\ &=\sum_{k=1}^N(\nabla v,\overline{\nabla\phi_h})_{\Omega_k}-\sum_{k=1}^N(\kappa^2v,\overline{\phi_h})_{\Omega_k}-\sum_{\gamma_{kj}}\langle[v],\overline{\nabla\phi\cdot n}\rangle_{\gamma_{kj}}+i\langle\kappa v,\overline{\phi}\rangle_{\partial\Omega}\\ &\quad+\sum_{k=1}^N(\nabla v,\overline{\nabla(\phi-\phi_h)})_{\Omega_k}-\sum_{k=1}^N\big(\kappa^2v,\overline{\phi-\phi_h}\big)_{\Omega_k}. \end{aligned} \tag{5.21}$$

In the last equality, we introduced the finite element function ϕ_h since (5.16) holds only for finite element function w (if $v\in H^1(\Omega_k)$ satisfies (5.16) for any $w\in H^1(\Omega_k)$, then the proof is trivial).

Letting $w=\phi_h$ in (5.16) and summing the resulting equality over k, and using the fact that ϕ_h is continuous across the inner edges, gives

$$\sum_{k=1}^N a^{(k)}(v,\phi_h)=\sum_{k=1}^N\langle\pm\lambda_h,\overline{\phi_h}\rangle_{\partial\Omega_k\backslash\partial\Omega}=0,$$

which implies that

$$\sum_{k=1}^N(\nabla v,\overline{\nabla\phi}_h)_{\Omega_k}-\sum_{k=1}^N(\kappa^2v,\overline{\phi}_h)_{\Omega_k}=-i\rho\sum_{\gamma_{kj}}\langle[v],\overline{\phi}_h\rangle_{\gamma_{kj}}-i\langle\kappa v,\overline{\phi}_h\rangle_{\partial\Omega}.$$

This, together with (5.21), leads to

$$\begin{aligned} ||v||_{0,\Omega}^2 &= -i\rho \sum_{\gamma_{kj}} \langle [v], \overline{\phi_h}\rangle_{\gamma_{kj}} - i\langle \kappa v, \overline{\phi}_h\rangle_{\partial\Omega} - \sum_{\gamma_{kj}} \langle [v], \overline{\nabla\phi\cdot n}\rangle_{\gamma_{kj}} + i\langle \kappa v, \overline{\phi}\rangle_{\partial\Omega} \\ &\quad + \sum_{k=1}^{N} (\nabla v, \overline{\nabla(\phi-\phi_h)})_{\Omega_k} - \sum_{k=1}^{N} (\kappa^2 v, \overline{\phi-\phi_h})_{\Omega_k} \\ &= \sum_{k=1}^{N} (\nabla v, \overline{\nabla(\phi-\phi_h)})_{\Omega_k} - \sum_{k=1}^{N} (\kappa^2 v, \overline{\phi-\phi_h})_{\Omega_k} + i\langle \kappa v, \overline{\phi-\phi_h}\rangle_{\partial\Omega} \\ &\quad - \sum_{\gamma_{kj}} \langle [v], \overline{\nabla\phi\cdot n}\rangle_{\gamma_{kj}} - i\rho \sum_{\gamma_{kj}} \langle [v], \overline{\phi}_h\rangle_{\gamma_{kj}}. \end{aligned} \tag{5.22}$$

If we directly estimate the terms containing the error $\phi-\phi_h$, we cannot build the inequality (5.17) unless a stronger assumption on the mesh size h is made. Because of this, we have to introduce a globally continuous finite element "approximation" of v such that the energy orthogonality of $\phi-\phi_h$ can be used.

For $v \in V_h^p(\mathcal{T}_h)$, we construct $\tilde{v} \in \tilde{V}_h^p(\mathcal{T}_h)$ as in Lemma 5.3. For ease of notation, set

$$R = \sum_{k=1}^{N} (\nabla(v-\tilde{v}), \overline{\nabla(\phi-\phi_h)})_{\Omega_k} - \sum_{k=1}^{N} \left(\kappa^2(v-\tilde{v}), \overline{\phi-\phi_h}\right)_{\Omega_k} + i\langle \kappa(v-\tilde{v}), \overline{\phi-\phi_h}\rangle_{\partial\Omega}.$$

Then (5.22) can be written as

$$\begin{aligned} ||v||_{0,\Omega}^2 &= R + \sum_{k=1}^{N} (\nabla\tilde{v}, \overline{\nabla(\phi-\phi_h)})_{\Omega_k} - \sum_{k=1}^{N} (\kappa^2\tilde{v}, \overline{\phi-\phi_h})_{\Omega_k} + i\langle \kappa\tilde{v}, \overline{\phi-\phi_h}\rangle_{\partial\Omega} \\ &\quad - \sum_{\gamma_{kj}} \langle [v], \overline{\nabla\phi\cdot n}\rangle_{\gamma_{kj}} - i\rho \sum_{\gamma_{kj}} \langle [v], \overline{\phi}_h\rangle_{\gamma_{kj}}. \end{aligned} \tag{5.23}$$

Choosing $\psi = \tilde{v}$ in (5.19) and $\psi_h = \tilde{v}$ in (5.20), we get the difference

$$(\nabla(\phi-\phi_h), \overline{\nabla\tilde{v}})_{\Omega} - \left(\kappa^2(\phi-\phi_h), \overline{\tilde{v}}\right)_{\Omega} - i\langle \kappa(\phi-\phi_h), \overline{\tilde{v}}\rangle_{\partial\Omega} = 0, \tag{5.24}$$

which is called the energy orthogonality of $\phi-\phi_h$. The complex conjugation of (5.24) becomes

$$(\nabla\tilde{v}, \overline{\nabla(\phi-\phi_h)})_{\Omega} - \left(\kappa^2\tilde{v}, \overline{\phi-\phi_h}\right)_{\Omega} + i\langle \kappa\tilde{v}, \overline{\phi-\phi_h}\rangle_{\partial\Omega} = 0. \tag{5.25}$$

Substituting (5.25) into (5.23), we obtain

$$||v||_{0,\Omega}^2 = R - \sum_{\gamma_{kj}} \langle [v], \overline{\nabla\phi\cdot n}\rangle_{\gamma_{kj}} - i\rho \sum_{\gamma_{kj}} \langle [v], \overline{\phi}_h\rangle_{\gamma_{kj}}. \tag{5.26}$$

Let $M = \sum\limits_{\gamma_{kj}} ||[v]||_{0,\gamma_{kj}}^2$. Using the Cauchy–Schwarz inequality for the sums on the right side of (5.26) yields

$$\begin{aligned} ||v||_{0,\Omega}^2 &\le |R| + \sum_{\gamma_{kj}} ||[v]||_{0,\gamma_{kj}} ||\nabla\phi\cdot n||_{0,\gamma_{kj}} + \rho \sum_{\gamma_{kj}} ||[v]||_{0,\gamma_{kj}} ||\phi_h||_{0,\gamma_{kj}} \\ &\le |R| + \left(\sum_{\gamma_{kj}} ||\nabla\phi\cdot n||_{0,\gamma_{kj}}^2\right)^{\frac{1}{2}} M^{\frac{1}{2}} + \rho \left(\sum_{\gamma_{kj}} ||\phi_h||_{0,\gamma_{kj}}^2\right)^{\frac{1}{2}} M^{\frac{1}{2}}. \end{aligned} \tag{5.27}$$

Using the Cauchy–Schwarz inequality, Assumption 4, and Lemma 5.3, we can deduce that (more details can be found in the e-print [24])

$$|R| \lesssim \left(1+\omega h + +\omega^{\frac{1}{2}}h^{\frac{1}{2}}\right)h^{-\frac{1}{2}}M^{\frac{1}{2}}||v||_{0,\Omega}. \tag{5.28}$$

On the other hand, by the stabilities (which are derived as in the proof of Lemma 3.3 of [9], together with Assumption 1 and Theorem 1 of [2]) we can verify that

$$\sum_{\gamma_{kj}} ||\nabla\phi\cdot n||^2_{0,\gamma_{kj}} \lesssim (h^2\omega^2+1)h^{-1}||v||^2_{0,\Omega} \tag{5.29}$$

and (using Assumption 4 again)

$$\sum_{\gamma_{kj}} ||\phi_h||^2_{0,\gamma_{kj}} \lesssim (h^2+\omega^{-2})h^{-1}||v||^2_{0,\Omega}. \tag{5.30}$$

Substituting the inequalities (5.28), (5.29), and (5.30) into (5.27) and using Assumptions 2 and 3 yields

$$||v||^2_{0,\Omega} \lesssim h^{-\frac{1}{2}}\cdot M^{\frac{1}{2}}||v||_{0,\Omega}.$$

Finally, we obtain the desired inequality (5.17). □

Remark 5.2. The inequality (5.17) can be viewed as an extension of the Poincaré inequality held for plane wave functions to the piecewise polynomial functions in $V_h^p(\mathcal{T}_h)$. Comparing the inequality (5.17) with the Poincaré-type inequality given by Lemma 3.7 of [23] for the plane wave functions, we find that the right sides of the two inequalities contain the same term $h^{-1}\sum_{\gamma_{kj}}||[v]||^2_{0,\gamma_{kj}}$, and (5.17) is more succinct thanks to the condition (5.16) (there are extra terms in the inequality in Lemma 3.7 of [23]). However, the proof of (5.17), which depends on the estimates (5.15) and (4.3), is much more technical than that of the inequality in Lemma 3.7 of [23] since the considered functions do not satisfy the homogeneous Helmholtz equation satisfied by the plane wave functions.

Remark 5.3. As pointed out in Remark 2.1, the proposed method is practical for the case with variable wave numbers, but we have to use Assumption 4 to give the theoretical analysis of (5.17). The main reason is that we do not know whether the estimate (4.3) proved in [32] for constant wave numbers is still valid for the case with variable wave numbers (the condition (4.2) that we used is much weaker than (4.3)). We failed to build a similar inequality with (5.17) without Assumption 4.

In the rest of this paper, we always use $a^{(k)}(\cdot,\cdot)$ to denote the local sesquilinear form defined in subsection 2.2. For $v\in\prod_{k=1}^N H^1(\Omega_k)$, define

$$|||v||| = \left(\sum_{k=1}^N \|\nabla v\|^2_{0,\Omega_k} + \omega^2\|v\|^2_{0,\Omega}\right)^{\frac{1}{2}}.$$

The following auxiliary result can be directly verified by the local inf-sup condition given in Theorem 4.1, together with Assumptions 3 and 2.

LEMMA 5.5. *Let Assumption* 3 *be satisfied. Suppose* $q\geq 1$ *and* $p\geq q+2$. *Assume that* $v\in V_h^p(\mathcal{T}_h)$ *and* $\lambda_h\in W_h^q(\gamma)$ *satisfy the relation*

$$a^{(k)}(v,w) = \langle \pm\lambda_h, \overline{w}\rangle_{\partial\Omega_k\backslash\partial\Omega} \qquad (k=1,2,\ldots,N), \quad \forall\, w\in V_h^p(\mathcal{T}_h). \tag{5.31}$$

Then the following estimate holds:

$$\sum_{\gamma_{kj}} ||\lambda_h||^2_{0,\gamma_{kj}} \lesssim h^{-1}pq|||v|||^2. \tag{5.32}$$

By Lemmas 5.4 and 5.5, we can prove a crucial auxiliary result given below, which can be viewed as a *jump-controlled* stability estimate. As we will see, this auxiliary result plays a key role in the proof of Theorems 4.2 and 4.3.

PROPOSITION 5.1. *Assume that $q \geq 1$ and $p \geq q+2$. Let Assumptions* 1–4 *be satisfied, and let $v \in V_h^p(\mathcal{T}_h)$ satisfy*

$$a^{(k)}(v,w) = \langle \pm\lambda_h, \overline{w}\rangle_{\partial\Omega_k\backslash\partial\Omega} \qquad (k=1,2,\dots,N), \ \forall\, w \in V_h^p(\mathcal{T}_h). \tag{5.33}$$

Then the following estimate holds:

$$|||v|||^2 \lesssim \omega^2 h^{-1} \sum_{\gamma_{kj}} ||[v]||^2_{0,\gamma_{kj}}. \tag{5.34}$$

Now we can easily prove Theorem 4.2 *by Lemma* 5.5 *and Proposition* 5.1.

Proof of Theorem 4.2. For $\lambda_h \in W_h^q(\gamma)$, let $u_{h,k}^{(1)}(\lambda_h)$ be the function defined in subsection 2.3. From the definition of $u_{h,k}^{(1)}(\lambda_h)$, we have

$$a^{(k)}\left(u_{h,k}^{(1)}(\lambda_h), \overline{w}_h\right) = \langle \pm\lambda_h, \overline{w}_h\rangle_{\partial\Omega_k\backslash\partial\Omega}, \quad k=1,\dots,N; \quad \forall\, w_h \in V_h^p(\mathcal{T}_h).$$

Namely, $u_h^{(1)}(\lambda_h)$ satisfies (5.31). It follows by Lemma 5.5 that

$$\sum_{\gamma_{kj}} ||\lambda_h||^2_{0,\gamma_{kj}} \lesssim h^{-1}pq\left|\left|\left|u_h^{(1)}(\lambda_h)\right|\right|\right|^2. \tag{5.35}$$

Obviously, $u_h^{(1)}(\lambda_h)$ satisfies (5.33) too. It follows by Proposition 5.1 that

$$\left|\left|\left|u_h^{(1)}(\lambda_h)\right|\right|\right|^2 \lesssim \omega^2 h^{-1} \sum_{\gamma_{kj}} \left\|\left[u_h^{(1)}(\lambda_h)\right]\right\|^2_{0,\gamma_{kj}}.$$

This, together with (5.35), leads to

$$\sum_{\gamma_{kj}} ||\lambda_h||^2_{0,\gamma_{kj}} \lesssim \omega^2 h^{-2}pq \sum_{\gamma_{kj}} \left\|\left[u_h^{(1)}(\lambda_h)\right]\right\|^2_{0,\gamma_{kj}}.$$

Thus

$$s_h(\lambda_h,\lambda_h) = \sum_{\gamma_{kj}} \left\|\left[u_h^{(1)}(\lambda_h)\right]\right\|^2_{0,\gamma_{kj}} \gtrsim \omega^{-2}h^2p^{-1}q^{-1} \sum_{\gamma_{kj}} ||\lambda_h||^2_{0,\gamma_{kj}}. \qquad \square$$

Remark 5.4. Remark 5.1 tells us that, when $q \geq 2$ and $p = 2q$, a slightly better result than (4.5) can be built:

$$s_h(\lambda_h,\lambda_h) = \sum_{\gamma_{kj}} \left\|\left[u_h^{(1)}(\lambda_h)\right]\right\|^2_{0,\gamma_{kj}} \gtrsim \omega^{-2}h^2p^{-1} \sum_{\gamma_{kj}} ||\lambda_h||^2_{0,\gamma_{kj}}.$$

5.3. Analysis on the error estimates. In order to prove Theorem 4.3, we need more auxiliary results. We will decompose the error $u - u_h$ into three parts, where the first part and the second part have some particular property and the third part is a p-order finite element function. The third part can be estimated by Proposition 5.1, but the estimates of the first part and the second part are more technical, which depend on a key auxiliary result (Lemma 5.6).

We first build the key auxiliary result mentioned above. For an element Ω_k and a function $v \in H^1(\Omega_k)$, we use the notation in this subsection,

$$F_k(v) = ||\nabla v||^2_{0,\Omega_k} - (\kappa^2 v, v)_{\Omega_k} \pm i\rho ||v||^2_{0,\partial\Omega_k\setminus\partial\Omega} + i\langle \kappa v, v\rangle_{\partial\Omega_k\cap\partial\Omega}. \tag{5.36}$$

It is clear that $F_k(v) = a^{(k)}(v, v)$.

Lemma 5.6. *Let Assumptions* 2 *and* 3 *be satisfied. For one element Ω_k, assume that $v \in H^1(\Omega_k)$ has the property*

$$-(\kappa^2 v, 1)_{\Omega_k} \pm i\rho\langle v, 1\rangle_{\partial\Omega_k\setminus\partial\Omega} + i\langle \kappa v, 1\rangle_{\partial\Omega_k\cap\partial\Omega} = 0. \tag{5.37}$$

Then

$$\|\nabla v\|^2_{0,\Omega_k} + \omega\|v\|^2_{0,\Omega_k} \leq C|F_k(v)|. \tag{5.38}$$

Proof. Taking the module to (5.36) leads to

$$\big|||\nabla v||^2_{0,\Omega_k} - (\kappa^2 v, v)_{\Omega_k}\big| \leq |F_k(v)| \tag{5.39}$$

and

$$\big| \pm \rho||v||^2_{0,\partial\Omega_k\setminus\partial\Omega} + \langle \kappa v, v\rangle_{\partial\Omega_k\cap\partial\Omega}\big| \leq |F_k(v)|. \tag{5.40}$$

We first assume that $\partial\Omega_k \cap \partial\Omega \neq \emptyset$. It follows, by (5.40) and the trace inequality, that

$$\begin{aligned} \omega||v||^2_{0,\partial\Omega_k\cap\partial\Omega} &\lesssim |F_k(v)| + \rho||v||^2_{0,\partial\Omega_k\setminus\partial\Omega} \\ &\lesssim |F_k(v)| + \rho h||\nabla v||^2_{0,\Omega_k} + \rho h^{-1}||v||^2_{0,\Omega_k}. \end{aligned} \tag{5.41}$$

Using the Poincaré inequality and (5.41) yields

$$\begin{aligned} \omega^2||v||^2_{0,\Omega_k} &\lesssim \omega^2 h^2||\nabla v||^2_{0,\Omega_k} + \omega^2 h||v||^2_{0,\partial\Omega_k\cap\partial\Omega} \\ &\lesssim \omega^2 h^2||\nabla v||^2_{0,\Omega_k} + \omega h|F_k(v)| + \rho\omega h^2||\nabla v||^2_{0,\Omega_k} + \rho\omega||v||^2_{0,\Omega_k}, \end{aligned}$$

which implies that

$$(1 - \rho C\omega^{-1})\omega^2||v||^2_{0,\Omega_k} \lesssim (\omega^2 h^2 + \rho\omega h^2)||\nabla v||^2_{0,\Omega_k} + \omega h|F_k(v)|.$$

Then, from Assumption 3, we have

$$\omega^2||v||^2_{0,\Omega_k} \lesssim \omega^2 h^2||\nabla v||^2_{0,\Omega_k} + \omega h|F_k(v)|. \tag{5.42}$$

On the other hand, by (5.39) and (5.42), we deduce that

$$||\nabla v||^2_{0,\Omega_k} \lesssim \omega^2||v||^2_{0,\Omega_k} + |F_k(v)| \lesssim \omega^2 h^2||\nabla v||^2_{0,\Omega_k} + (1 + \omega h)|F_k(v)|,$$

which gives

$$\left(1 - C\omega^2 h^2\right) ||\nabla v||^2_{0,\Omega_k} \lesssim |F_k(v)|.$$

This, together with (5.42), leads to

$$\left(1 - C\omega^2 h^2\right) \omega^2 ||v||^2_{0,\Omega_k} \lesssim \left(\omega^2 h^2 + \omega h(1 - C\omega^2 h^2)\right) |F_k(v)|.$$

Using Assumption 2, the above two inequalities give (5.38) when $\partial\Omega_k \cap \partial\Omega \neq \emptyset$.

In the following we assume that $\partial\Omega_k \cap \partial\Omega = \emptyset$. It follows by (5.37) that

$$\omega^2 |\Omega_k||\gamma_{\Omega_k}(v)| \lesssim \rho |\langle v, 1\rangle_{\partial\Omega_k \setminus \partial\Omega}| \lesssim \rho |\partial\Omega_k|^{\frac{1}{2}} ||v||_{0,\partial\Omega_k},$$

where $\gamma_{\Omega_k}(v) = \frac{1}{|\Omega_k|}\int_{\Omega_k} v\,dx$. Thus

$$\omega^4 |\gamma_{\Omega_k}(v)|^2 \lesssim \rho^2 |\partial\Omega_k| \cdot |\Omega_k|^{-2} ||v||^2_{0,\partial\Omega_k}.$$

This, together with the trace inequality (or ε-inequality), leads to

$$\begin{aligned} \omega^4 ||\gamma_{\Omega_k}(v)||^2_{0,\Omega_k} &\lesssim \rho^2 |\partial\Omega_k| \cdot |\Omega_k|^{-1} ||v||^2_{0,\partial\Omega_k} \\ &\lesssim \rho^2 h^{-1} \left(h ||\nabla v||^2_{0,\Omega_k} + h^{-1} ||v||^2_{0,\Omega_k}\right) \\ &= \rho^2 ||\nabla v||^2_{0,\Omega_k} + \rho^2 h^{-2} ||v||^2_{0,\Omega_k}. \end{aligned} \tag{5.43}$$

Using Friedrichs' inequality and (5.43), we deduce that

$$\begin{aligned} \omega^4 ||v||^2_{0,\Omega_k} &\leq \omega^4 ||v - \gamma_{\Omega_k}(v)||^2_{0,\Omega_k} + \omega^4 ||\gamma_{\Omega_k}(v)||^2_{0,\Omega_k} \\ &\lesssim \omega^4 h^2 ||\nabla v||^2_{0,\Omega_k} + \rho^2 ||\nabla v||^2_{0,\Omega_k} + \rho^2 h^{-2} ||v||^2_{0,\Omega_k}. \end{aligned}$$

So we get

$$\left(1 - \rho^2 \omega^{-4} h^{-2}\right) \omega^2 ||v||^2_{0,\Omega_k} \lesssim (\omega^2 h^2 + \rho^2 \omega^{-2}) ||\nabla v||^2_{0,\Omega_k}.$$

Thus, by Assumption 3, we have

$$\omega^2 ||v||^2_{0,\Omega_k} \lesssim \omega^2 h^2 ||\nabla v||^2_{0,\Omega_k}.$$

In addition, combining (5.39) with the above inequality, we get

$$||\nabla v||^2_{0,\Omega_k} \lesssim \omega^2 ||v||^2_{0,\Omega_k} + |F_k(v)| \lesssim \omega^2 h^2 ||\nabla v||^2_{0,\Omega_k} + |F_k(v)|.$$

Therefore we obtain (if $C\omega^2 h^2 < 1$)

$$\left(1 - C\omega^2 h^2\right) ||\nabla v||^2_{0,\Omega_k} \lesssim |F_k(v)|$$

and

$$\left(1 - C\omega^2 h^2\right) \omega^2 ||v||^2_{0,\Omega_k} \lesssim \omega^2 h^2 |F_k(v)|.$$

Using Assumption 2 again, the above two inequalities give (5.38) for the case that $\partial\Omega_k \cap \partial\Omega = \emptyset$. □

For each element Ω_k, let $\hat{u}_{h,k}(\lambda) \in V_h^p(\Omega_k)$ be determined by the variational problem

$$a^{(k)}(\hat{u}_{h,k}(\lambda), \overline{v}_h) = L^{(k)}(\overline{v}_h) + \langle \pm\lambda, \overline{v}_h \rangle_{\partial\Omega_k \setminus \partial\Omega} \quad \forall\, v_h \in V_h^p(\Omega_k). \tag{5.44}$$

Then define $\hat{u}_h(\lambda) \in V_h^p(\mathcal{T}_h)$ such that $\hat{u}_h(\lambda)|_{\Omega_k} = \hat{u}_{h,k}(\lambda)$ $(k = 1, \ldots, N)$.

The forthcoming two auxiliary results can be proved by Lemma 5.6 and the error estimates of the *hp* finite element (see [20]).

Lemma 5.7. *Assume that* $u \in H^{r+1}(\mathcal{T}_h)$ *with* $1 \le r \le p$. *Let Assumptions* 2 *and* 3 *be satisfied. Then*

$$|u(\lambda) - \hat{u}_h(\lambda)|_{1,\Omega} \le Ch^r p^{-r}|u|_{r+1,\Omega}$$

and

$$||u(\lambda) - \hat{u}_h(\lambda)||_{0,\Omega} \le C\omega^{-1}h^r p^{-r}|u|_{r+1,\Omega}.$$

Let $Q_h^q : L^2(\gamma) \to W_h^q(\gamma)$ denote the L^2 projector, and let $u_h(Q_h^q\lambda) \in V_h^p(\Omega_k)$ be defined as in (2.6), by choosing $\lambda_h = Q_h^q\lambda$.

Lemma 5.8. *Suppose that* $q \ge 1$, $p \ge q+2$, *and* $\lambda \in W^{r-\frac{1}{2}}(\gamma)$ *with* $1 \le r \le q+1$. *Let Assumptions* 1–4 *be satisfied. Then*

$$|\hat{u}_h(\lambda) - u_h(Q_h^q\lambda)|_{1,\Omega} \le Ch^r q^{-r}|\lambda|_{r-\frac{1}{2},\gamma}$$

and

$$||\hat{u}_h(\lambda) - u_h(Q_h^q\lambda)||_{0,\Omega} \le C\omega^{-1}h^r q^{-r}|\lambda|_{r-\frac{1}{2},\gamma}.$$

The following result can be verified by using Proposition 5.1, together with Lemmas 5.7 and 5.8.

Lemma 5.9. *Suppose that* $q \ge 1$, $p \ge q+2$, *and* $u \in H^{r+1}(\mathcal{T}_h)$ *with* $1 \le r \le p$. *Let Assumptions* 1–4 *be satisfied. Then*

$$|u_h(Q_h^q\lambda) - u_h(\lambda_h)|_{1,\Omega} \le Ch^{r-1}\left(p^{-r}|u|_{r+1,\Omega} + q^{-r}|\lambda|_{r-\frac{1}{2},\gamma}\right)$$

and

$$||u_h(Q_h^q\lambda) - u_h(\lambda_h)||_{0,\Omega} \le C\omega^{-1}h^{r-1}\left(p^{-r}|u|_{r+1,\Omega} + q^{-r}|\lambda|_{r-\frac{1}{2},\gamma}\right).$$

Now we can easily prove Theorem 4.3 by Lemmas 5.7–5.9.

Proof of Theorem 4.3. By the triangle inequality, we have

$$\begin{aligned}&||u(\lambda) - u_h(\lambda_h)||_{0,\Omega}\\&\quad \le ||u(\lambda) - \hat{u}_h(\lambda)||_{0,\Omega} + ||\hat{u}_h(\lambda) - u_h(Q_h^q\lambda)||_{0,\Omega} + ||u_h(Q_h^q\lambda) - u_h(\lambda_h)||_{0,\Omega}\end{aligned}$$

and

$$\begin{aligned}&|u(\lambda) - u_h(\lambda_h)|_{1,\Omega}\\&\quad \le |u(\lambda) - \hat{u}_h(\lambda)|_{1,\Omega} + |\hat{u}_h(\lambda) - u_h(Q_h^q\lambda)|_{1,\Omega} + |u_h(Q_h^q\lambda) - u_h(\lambda_h)|_{1,\Omega}.\end{aligned}$$

Then, by the estimates given in Lemmas 5.7, 5.8, and 5.9, we obtain

$$||u(\lambda) - u_h(\lambda_h)||_{0,\Omega} \lesssim \omega^{-1}h^{r-1}\left(p^{-r}|u|_{r+1,\Omega} + q^{-r}|\lambda|_{r-\frac{1}{2},\gamma}\right)$$

and

$$|u(\lambda) - u_h(\lambda_h)|_{1,\Omega} \lesssim h^{r-1}\left(p^{-r}|u|_{r+1,\Omega} + q^{-r}|\lambda|_{r-\frac{1}{2},\gamma}\right). \qquad \square$$

Remark 5.5. The proposed method can be extended directly to the case with other boundary conditions that can guarantee the well-posedness of the equations, provided that the subproblems defined on the elements touching the boundary $\partial\Omega$ impose the corresponding boundary conditions (the analysis is simpler for the variational problems with part Dirichlet boundary condition). The proposed discretization method can also be extended to three-dimensional problems, but the coarse subspace involved in the construction of the preconditioner needs to be modified and the analysis is more difficult (for example, the proofs of Theorem 4.1 and Lemma 5.3 need to be changed).

6. Numerical experiments. In this section we report some numerical results to illustrate that the proposed least squares method and domain decomposition preconditioner are efficient for Helmholtz equations with large wave numbers.

In the discretization method described in section 2, the parameter ρ can be relatively arbitrarily positive number. We find that the different choices of ρ do not affect the accuracy of the resulting approximations provided that the value of ρ is less than 1. In this section, we simply choose $\rho = 10^{-5}$ for numerical experiments.

For the considered example, the domain Ω is a rectangle so we adopt a uniform partition $\mathcal{T}_h$ for the domain Ω as follows: Ω is divided into some small rectangles with a same size h, where h denotes the length of the longest edge of the elements.

To measure the accuracy of the numerical solution u_h, we introduce the following relative L^2 error:

$$\text{Err.} = \frac{||u_{ex} - u_h||_{L^2(\Omega)}}{||u_{ex}||_{L^2(\Omega)}}.$$

For a discretization method, when the value of ωh is fixed but ω increases (h decreases), the relative L^2 error Err. may obviously increase (if the number of basis functions on each element does not increase). This phenomenon is called "wave number pollution." The efficiency of a discretization method for Helmholtz equations can be characterized by the degree of wave number pollution. For convenience, we define a positive parameter δ to measure the *degree of wave number pollution* as follows: the parameter δ is the minimal positive number such that, when ω increases and h decreases to keep the value of $\omega^{1+\delta}h$ being a constant, the relative L^2 error Err. does not increase. If $\delta = 0$, the discretization method has no "wave number pollution." For the standard linear finite element method, the existing results imply that $\delta = 1$ (see [32]). That a discretization method is ideal means that $\delta \ll 1$. For concrete examples, it is difficult to exactly calculate such parameter δ. Because of this, we want to give a similar definition of δ, which can be explicitly calculated.

When ω increases from ω_1 to ω_2, the mesh size h decreases from h_1 to h_2. We fix the value ωh, i.e., $\omega_2 h_2 = \omega_1 h_1$. Let Err_1 and Err_2 denote the relative L^2 errors with $\omega = \omega_1$ ($h = h_1$) and $\omega = \omega_2$ ($h = h_2$), respectively. Define $\delta > 0$ by

$$\frac{\omega_2^{1+\delta} h_2}{\omega_1^{1+\delta} h_1} = \frac{\text{Err}_2}{\text{Err}_1}.$$

It is easy to see that the parameter δ can be expressed as

$$\delta = \frac{\ln(\text{Err}_2/\text{Err}_1)}{\ln(\omega_2/\omega_1)} + \left(\frac{\ln(h_1/h_2)}{\ln(\omega_2/\omega_1)} - 1\right) = \frac{\ln(\text{Err}_2/\text{Err}_1)}{\ln(\omega_2/\omega_1)} \qquad (\text{since} \ \ \omega_2 h_2 = \omega_1 h_1).$$

For a given ω, we define the error order with respect to h in the standard manner, namely,

$$\text{order} = \frac{\ln(\text{Err}_2/\text{Err}_1)}{\ln(h_2/h_1)},$$

where Err_1 and Err_2 denote the relative L^2 errors corresponding to $h = h_1$ and $h = h_2$ (p, q are fixed), respectively.

Throughout this section we can simply choose $p = q + 2$ to avoid extra cost of calculation.

6.1. Wave propagation in a duct with rigid walls. In this subsection, we give some comparisons between the proposed method and the PWLS method for a homogeneous Helmholtz equation with constant wave number. For the comparisons, we recall the basic ideas of the PWLS method (see subsection 2.3 in [28]). In the PWLS method, the solution space consists of plane wave basis functions that exactly satisfy the considered homogeneous Helmholtz equation, and the variational formula is derived by a minimization problem with a quadratic subject functional defined by the jumps of function values and normal derivations across all the element interfaces. Since the basis functions satisfy the considered Helmholtz equation, one need not introduce auxiliary unknowns on the element interfaces and so does not solve local Helmholtz equations on elements.

We consider the following model Helmholtz equation for the acoustic pressure u (see [25])

$$\left\{ \begin{array}{ll} -\Delta u - \omega^2 u = 0 & \text{in} \quad \Omega, \\ \frac{\partial u}{\partial \mathbf{n}} + i\omega u = g & \text{on} \quad \partial\Omega, \end{array} \right. \tag{6.1}$$

where $\Omega = [0, 2] \times [0, 1]$, and $g = (\frac{\partial}{\partial \mathbf{n}} + i\omega)u_{ex}$. The analytic solution u_{ex} of the problem can be obtained in the closed form as

$$u_{ex}(x, y) = \cos(k\pi y) \left(A_1 e^{-i\omega_x x} + A_2 e^{i\omega_x x}\right)$$

with $\omega_x = \sqrt{\omega^2 - (k\pi)^2}$, and the coefficients A_1 and A_2 satisfying the equation

$$\begin{pmatrix} \omega_x & -\omega_x \\ (\omega - \omega_x)e^{-2i\omega_x} & (\omega + \omega_x)e^{2i\omega_x} \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} = \begin{pmatrix} -i \\ 0 \end{pmatrix}. \tag{6.2}$$

Let NLS denote the novel least squares method proposed in this paper. In addition, let $\hat{p}$ be the number of plane wave basis functions on every element. For convenience, we use “dof.” to denote the number of degrees of freedom in the resulting algebraic systems (which means the system (2.8) for the NLS method).

In Tables 1 and 2, we compare the required numbers of degrees of freedom to achieve almost the same accuracies of the approximate solutions generated by the two methods.

It can be seen from the above data that, for the new least squares method, fewer degrees of freedom in the solved algebraic system are enough to achieve almost the same accuracies (with the same choices of ω and h). For the proposed method, a little extra cost is needed when solving all the local problems defined on the elements (in parallel). In addition, the system (2.8) has more complex structure than the one in the PWLS method and so its preconditioner is more difficult to construct. In summary, the proposed method is at least comparable to the plane wave method even if the wave number is a constant (otherwise, the plane wave method may be unpractical).

Table 1
Approximate errors: fixing $\omega = 40\pi$ and decreasing h (setting $k = 19$).

	PWLS, $\hat{p} = 12$		NLS, $(q,p) = (4,6)$	
h	dof.	Err.	dof.	Err.
$\frac{1}{28}$	18816	1.24e-2	15260	4.57e-4
$\frac{1}{36}$	31104	2.19e-3	25380	7.26e-5
$\frac{1}{44}$	46464	2.67e-4	38060	1.80e-5
$\frac{1}{52}$	64896	5.99e-5	53300	6.02e-6

Table 2
Approximate errors: fixing $\omega h = 5\pi/8$ and increasing ω (setting $k = 12$).

		PWLS, $\hat{p} = 12$		NLS, $(q,p) = (3,5)$	
ω	h	dof.	Err.	dof.	Err.
30π	$\frac{1}{48}$	55296	1.29e-4	36288	3.24e-5
35π	$\frac{1}{56}$	75264	3.43e-4	49504	3.56e-5
40π	$\frac{1}{64}$	98304	5.73e-4	64768	3.81e-5
45π	$\frac{1}{72}$	124416	7.85e-4	82080	4.00e-5

6.2. An example with variable wave numbers. In this subsection, we consider the following Helmholtz equations with variable wave numbers:

$$\begin{cases} -\Delta u - \kappa^2 u = f & \text{in } \Omega, \\ \dfrac{\partial u}{\partial n} + i\kappa u = g & \text{on } \partial\Omega, \end{cases} \tag{6.3}$$

where $\Omega = [0,1] \times [0,1]$ and $\kappa = \frac{\omega}{c(x,y)}$. We define the velocity field $c(\mathbf{x})$ as a smooth converging lens with a Gaussian profile at the center $(r_1, r_2) = (1/2, 1/2)$ (refer to [10]):

$$c(x,y) = \frac{4}{3}\left(1 - \frac{1}{8}\exp\big(-32((x-r_1)^2 + (y-r_2)^2)\big)\right). \tag{6.4}$$

The analytic solution of the problem is given by

$$u_{ex}(x,y) = c(x,y)\exp(i\omega xy). \tag{6.5}$$

For this example, the standard plane wave methods are unpractical. In Tables 3 and 4, we list the accuracies of the approximate solutions generated by the proposed least squares method, where the algebraic systems are solved in the exact manner.

From the two tables, we can see that the approximate solutions generated by the proposed method indeed have high accuracies and have little "wave number pollution."

Since the resulting stiffness matrix is Hermitian positive definite, we can solve the system by the CG method and the PCG method with the preconditioner constructed in section 3. As usual we choose $d \approx \sqrt{h}$ as the subdomain size in this preconditioner

Table 3
Degrees of wave number pollution: fixing ωh to be a constant and increasing ω (and decreasing h).

	$\omega h = 1,\ (q,p) = (2,4)$			$\omega h = 2,\ (q,p) = (3,5)$			$\omega h = 2,\ (q,p) = (4,6)$		
ω	h	Err.	δ	h	Err.	δ	h	Err.	δ
64	$\frac{1}{64}$	3.079e-5		$\frac{1}{32}$	2.8643e-5		$\frac{1}{32}$	1.690e-6	
128	$\frac{1}{128}$	3.132e-5	0.024	$\frac{1}{64}$	2.913e-5	0.035	$\frac{1}{64}$	1.731e-6	0.034
256	$\frac{1}{256}$	3.241e-5	0.049	$\frac{1}{128}$	2.952e-5	0.019	$\frac{1}{128}$	1.753e-6	0.019
512	$\frac{1}{512}$	3.318e-5	0.034	$\frac{1}{256}$	2.981e-5	0.014	$\frac{1}{256}$	1.770e-6	0.014

Table 4
Convergence orders of the approximations with respect to h: fixing $\omega = 64$ and decreasing h.

	$(q,p) = (2,4)$			$(q,p) = (3,5)$			$(q,p) = (4,6)$		
ω	h	Err.	order	h	Err.	order	h	Err.	order
64	$\frac{1}{32}$	4.484e-4		$\frac{1}{16}$	1.186e-3		$\frac{1}{16}$	1.381e-4	
64	$\frac{1}{64}$	3.079e-5	3.864	$\frac{1}{32}$	2.843e-5	5.383	$\frac{1}{32}$	1.690e-6	6.352
64	$\frac{1}{128}$	2.016e-6	3.933	$\frac{1}{64}$	7.390e-7	5.266	$\frac{1}{64}$	2.639e-8	6.001
64	$\frac{1}{256}$	1.282e-7	3.975	$\frac{1}{128}$	2.174e-8	5.087	$\frac{1}{128}$	4.135e-10	5.996
64	$\frac{1}{512}$	8.068e-9	3.990	$\frac{1}{256}$	6.710e-10	5.018	$\frac{1}{256}$	6.731e-12	5.941

to guarantee the loading balance. The stopping criterion in the iterative algorithms is that the relative L^2 norm of the residual of the iterative approximation satisfies $\epsilon < 1.0e - 6$.

Moreover, let N_{iter}^{CG} represent the iteration count for solving the algebraic system by the CG method and let N_{iter}^{PCG} represent the iteration count for solving the algebraic system by the PCG method with the domain decomposition preconditioner. When the wave number ω increases (and the mesh size h decreases), the iteration count N_{iter} (representing N_{iter}^{CG} or N_{iter}^{PCG}) also increases. In order to describe the growth rate of the iteration count N_{iter} with respect to the wave number ω, we introduce a new notation ρ^{iter}. Let ω_1 and ω_2 be two wave numbers, and let $N_{iter}^{(1)}$ and $N_{iter}^{(2)}$ denote the corresponding iteration counts, respectively. Then we define the positive number ρ^{iter} by

$$\frac{N_{iter}^{(2)}}{N_{iter}^{(1)}} = \left(\frac{\omega_2}{\omega_1}\right)^{\rho^{iter}}.$$

For example, when $\rho^{iter} = 1$, the growth is linear; if $\rho^{iter} \to 0^+$, then the preconditioner possesses the optimal convergence. For a preconditioner, the positive number ρ^{iter} defined above is called the “relative growth rate” of the iteration count. Of course, we hope that the relative growth rate ρ^{iter} is small.

In Tables 5 and 6, we compare the iteration counts and the “relative growth rate” for the CG method and the PCG method with the domain decomposition preconditioner constructed in section 3.

The above data indicate that the proposed preconditioner is very efficient and the iteration counts of the corresponding PCG method have a small relative growth rate when the wave number increases.

TABLE 5
Effectiveness of the preconditioner: the case with $\omega h \approx 1$ and $(q, p) = (2, 4)$.

ω	h	d	N_{iter}^{CG}	ρ_{CG}^{iter}	N_{iter}^{PCG}	ρ_{PCG}^{iter}	Err.
20π	$\frac{1}{64}$	$\frac{1}{8}$	1556		105		2.8825e-5
40π	$\frac{1}{121}$	$\frac{1}{11}$	2504	0.6864	139	0.4047	3.8198e-5
80π	$\frac{1}{256}$	$\frac{1}{16}$	5158	1.0426	191	0.4585	3.1632e-5
160π	$\frac{1}{484}$	$\frac{1}{22}$	9643	0.9027	251	0.3941	4.2254e-5

TABLE 6
Effectiveness of the preconditioner: the case with $\omega h \approx 2$ and $(q, p) = (4, 6)$.

ω	h	d	N_{iter}^{CG}	ρ_{CG}^{iter}	N_{iter}^{PCG}	ρ_{PCG}^{iter}	Err.
20π	$\frac{1}{36}$	$\frac{1}{6}$	489		78		7.3846e-7
40π	$\frac{1}{64}$	$\frac{1}{8}$	646	0.4017	106	0.4425	1.5461e-6
80π	$\frac{1}{121}$	$\frac{1}{11}$	1002	0.6333	144	0.4420	2.2087e-6
160π	$\frac{1}{256}$	$\frac{1}{16}$	1758	0.8111	191	0.4075	1.5775e-6

Acknowledgment. The authors would like to thank the two anonymous reviewers, who give many insightful suggestions to improve the presentation of this paper.

REFERENCES

[1] C. Alves and S. Valtchev, *Numerical comparison of two meshfree methods for acoustic wave scattering*, Eng. Anal. Bound. Elem., 29 (2005), pp. 371–382.

[2] D. L. Brown, D. Gallist, and D. Peterseim, *Multiscale petrov-Galerkin method for high-frequency heterogeneous Helmholtz equations*, in Meshfree Methods for Partial Differential Equations VII, Lect. Notes Comput. Sci. Eng. 100, Springer, New York, 2016.

[3] O. Cessenat and B. Despres, *Application of an ultra weak variational formulation of elliptic PDEs to the two-dimensional Helmholtz problem*, SIAM J. Numer. Anal., 35 (1998), pp. 255–299.

[4] H. Chen, P. Lu, and X. Xu, *A hybridizable discontinuous Galerkin method for the Helmholtz equation with high wave number*, SIAM J. Numer. Anal., 51 (2013), pp. 2166–2188.

[5] J. Cui and W. Zhang, *An analysis of HDG methods for the Helmholtz equation*, IMA J. Numer. Anal., 33 (2013), pp. 1–17

[6] E. Deckers, O. Atak, L. Coox, R. Damico, H. Devriendt, S. Jonckheere, K. Koo, B. Pluymers, D. Vandepitte, and W. Desmet, *The wave based method: An overview of* 15 *years of research*, Wave Motion, 51 (2014), pp. 550–565.

[7] L. Demkowicz, J. Gopalakrishnan, I. Muga, and J. Zitelli, *Wavenumber explicit analysis of a DPG method for the multidimensional Helmholtz equation*, Comput. Methods Appl. Mech. Engrg., 213 (2012), pp. 126–138.

[8] C. Dohrmann, *A preconditioner for substructuring based on constrained energy minimization*, SIAM J. Sci. Comput., 25 (2003), pp. 246–258.

[9] Y. Du and H. Wu, *Preasymptotic error analysis of higher order FEM and CIP-FEM for Helmholtz equation with high wave number*, SIAM J. Numer. Anal., 53 (2015), pp. 782–804.

[10] B. Engquist and L. Ying, *Sweeping preconditioner for the Helmholtz equation: Moving perfectly matched layers*, Multiscale Model. Simul., 9 (2011), pp. 686–710.

[11] J. Fang, J. Qian, L. Zepeda-Núñez, and H. Zhao, *Learning dominant wave directions for plane wave methods for high-frequency Helmholtz equations*, Res. Math. Sci., 4 (2017).

[12] C. Farhat, I. Harari, and U. Hetmaniuk, *A discontinuous Galerkin method with Lagrange multipliers for the solution of Helmholtz problems in the mid-frequency regime*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 1389–1419.

[13] X. Feng and H. Wu, *hp-discontinuous Galerkin methods for the Helmholtz equation with large wave number*, Math. Comp., 80 (2011), pp. 1997–2024.

[14] X. Feng and Y. Xing, *Absolutely stable local discontinuous Galerkin methods for the Helmholtz equation with large wave number*, Math. Comp., 82 (2013), pp. 1269–1296.

[15] G. Gabard, *Discontinuous Galerkin methods with plane waves for time-harmonic problems*, J. Comput. Phys., 225 (2007), pp. 1961–1984.

[16] C. J. Gittelson, R. Hiptmair, and I. Perugia, *Plane wave discontinuous Galerkin methods: Analysis of the h-version*, ESAIM Math. Model. Numer. Anal., 43 (2009), pp. 297–331.

[17] J. Gopalakrishnan, S. Lanteri, N. Olivares, and R. Perrussel, *Stabilization in relation to wavenumber in HDG methods*, Adv. Model. Simul. Eng. Sci., 2 (2015).

[18] J. Gopalakrishnan, I. Muga, and N. Olivares, *Dispersive and dissipative error in the DPG method with scaled norms for Helmholtz equartion*, SIAM J. Sci. Comput., 36 (2014), pp. A20–A39.

[19] J. Gopalakrishnan, M. Solano, and F. Vargas, *Dispersion analysis of HDG methods*, J. Sci. Comput., 77 (2018), pp. 1703–1735.

[20] B. Guo and W. Sun, *The optimal convergence of the hp version of the finite element method with quasi-uniform meshes*, SIAM J. Numer. Anal., 45 (2007), pp. 698–730.

[21] U. Hetmaniuk, *Stability estimates for a class of Helmholtz problems*, Commun. Math. Sci., 5 (2007), pp. 665–678.

[22] R. Griesmair and P. Monk, *Error analysis for a hybridizable discontinuous Galerkin method for the Helmholtz equation*, J. Sci. Comput., 49 (2011), pp. 291–310.

[23] R. Hiptmair, A. Moiola, and I. Perugia, *Plane wave discontinuous Galerkin methods for the 2D Helmholtz equation: analysis of the p-version*, SIAM J. Numer. Anal., 49 (2011), pp. 264–284.

[24] Q. Hu and R. Song, *A Novel Least Squares Method for Helmholtz Equations with Large Wave Numbers*, arXiv:1902.01166 [math.NA], 2019.

[25] Q. Hu and L. Yuan, *A weighted variational formulation based on plane wave basis for discretization of Helmholtz equations*, Int. J. Numer. Anal. Model., 11 (2014), pp. 587–607.

[26] Q. Hu and L. Yuan, *A plane wave least-squares method for time-harmonic Maxwell's equations in absorbing media*, SIAM J. Sci. Comput., 36 (2014), pp. A1911–A1936.

[27] Q. Hu and L. Yuan, *A Plane wave method combined with local spectral elements for nonhomogeneous Helmholtz equation and time-harmonic Maxwell equations*, Adv. Comput. Math., 44 (2018), pp. 245–275.

[28] Q. Hu and H. Zhang, *Substructuring preconditioners for the systems arising from plane wave discretization of Helmholtz equations*, SIAM J. Sci. Comput., 38 (2016), pp. A2232–A2261.

[29] T. Huttunen, M. Malinen, and P. Monk, *Solving Maxwell's equations using the ultra weak variational formulation*, J. Comput. Phys., 223 (2007), pp. 731–758.

[30] L. Imbert-Gérard and P. Monk, *Numerical simulation of wave propagation in inhomogeneous media using Generalized Plane Waves*, ESAIM Math. Model. Numer. Anal., 51 (2017), pp. 1387–1406.

[31] A. Lieu, G. Gabard, and H. Bériot, *A comparison of high-order polynomial and wave-based methods for Helmholtz problems*, J. Comput. Phys., 321 (2016), pp. 105–125.

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

[33] P. Monk and D. Wang, *A least-squares method for the helmholtz equation*, Comput. Methods Appl. Mech. Engrg., 175 (1999), pp. 121–136.

[34] P. Monk, J. Schöberl, and A. Sinwel, *Hybridizing Raviart-Thomas elements for the Helmholtz equation*, Electromagnetics, 30 (2010), pp. 149–176.

[35] N. C. Nguyena, J. Peraire, F. Reitich, and B. Cockburn, *A phase-based hybridizable discontinuous Galerkin method for the numerical solution of the Helmholtz equation*, J. Comput. Phys., 290 (2015), pp. 318–335.

[36] J. Peng, J. Wang, and S. Shu, *Adaptive BDDC algorithms for the system arising from plane wave discretization of Helmholtz equations*, Internat. J. Numer. Methods Engrg., 116 (2018), pp. 683–707.

[37] H. Riou, P. Ladevèze, and B. Sourcis, *The multiscale VTCR approach applied to acoustics problems*, J. Comput. Acoust., 16 (2008), pp. 487–505.

[38] M. Stanglmeier, N. C. Nguyena, J. Peraire, and B. Cockburn, *An explicit hybridizable discontinuous Galerkin method for the acoustic wave equation*, Comput. Methods Appl. Mech. Engrg., 300 (2016), pp. 748–769

[39] G. Szegö, *Orthogonal Polynomials*, 4th ed., Amer. Math. Soc. Colloq. Publ. 23, AMS, Providence, RI, 1975.