ESAIM: M2AN 42 (2008) 925–940
DOI: 10.1051/m2an:2008033

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

ERROR ESTIMATES FOR THE ULTRA WEAK VARIATIONAL FORMULATION OF THE HELMHOLTZ EQUATION

Annalisa Buffa[1] and Peter Monk[2]

Abstract. The Ultra Weak Variational Formulation (UWVF) of the Helmholtz equation provides a variational framework suitable for discretization using plane wave solutions of an appropriate adjoint equation. Currently convergence of the method is only proved on the boundary of the domain. However substantial computational evidence exists showing that the method also converges throughout the domain of the Helmholtz equation. In this paper we exploit the fact that the UWVF is essentially an upwind discontinuous Galerkin method to prove convergence of the solution in the special case where there is no absorbing medium present. We also provide some other estimates in the case when absorption is present, and give some simple numerical results to test the estimates. We expect that similar techniques can be used to prove error estimates for the UWVF applied to Maxwell's equations and elasticity.

Mathematics Subject Classification. 65N15, 65N30, 35J05.

Received July 20, 2007.
Published online August 12, 2008.

1. Introduction

The idea of using a complete family of solutions of a linear partial differential equation to approximate the solution of a boundary value problem has a long history. Trefftz [21] is usually credited with this idea which was further developed by Bergman and Vekua in the 1940's (see [11] for a review up to the mid 1980s). Recent work in this area includes [3,4]. More recently, in an attempt to avoid ill-conditioning and slow convergence in some situations, methods have been developed that use complete families locally on small sub-regions of the domain. These local solutions are then patched together to form an approximate global solution. Possible techniques include the partition of unity finite element method [16,17], a Lagrange multiplier technique [20], least squares methods [18,19] or the Ultra Weak Variational Formulation (UWVF) [5–7]. It is the last of these techniques that will be the focus of this paper.

The UWVF, which we shall describe precisely in Section 2, particularly equation (2.17), is a variational formulation of the Helmholtz equation due to Cessenat and Després [5,6]. It is based on a mesh of the domain where the Helmholtz equation is to be solved and computes the trace of the solution of the Helmholtz equation and its normal derivative on the skeleton of the mesh (*i.e.* the faces (in 3D) or edges (in 2D) of the mesh). The discrete UWVF uses solutions of the Helmholtz equation (usually plane waves, although Fourier-Bessel functions or other complete families could be used) in a key step of the algorithm, and it is traces of these

Keywords and phrases. Helmholtz equation, UWVF, plane waves, error estimate.

[1] Istituto di Matematica Applicata e Tecnologie Informatiche, via Ferrata 1, 27100 Pavia, Italy. annalisa@imati.cnr.it

[2] Department of Mathematical Sciences, University of Delaware, Newark, DE 19716, USA. monk@math.udel.edu

Article published by EDP Sciences

© EDP Sciences, SMAI 2008

functions that are used to approximate the desired solution. Once the solution of the Helmholtz equation is approximated on the skeleton of the mesh, the full solution can be approximated element by element by solving local problems, again using plane waves for example.

Cessenat and Després [5,6] prove an error estimate for the method showing that the solution of the UWVF converges to an appropriate impedance trace of the true solution on the boundary of the domain and in particular they prove that the boundary error is bounded by a suitable best approximation error in the entire domain. They then derive explicit error estimates by analyzing the convergence of the best approximation error. We shall give more details of their results later in this paper. Extensive numerical experiments by Cessenat and Després, as well as by others [12–14], shows that the method converges throughout the domain of computation. The main purpose of this paper is to prove global convergence of the UWVF applied to the Helmholtz equation in the case where the medium is not absorbing as is often the case for scattering calculations. We shall also provide an alternative error estimate for the case of an absorbing medium.

The method of proof uses the connection between upwind discontinuous Galerkin methods and the UWVF. This was noted in [15] where it was shown that the UWVF for Maxwell's equations can be derived using discontinuous Galerkin (DG) techniques and a special choice of degrees of freedom. This observation also holds for the Helmholtz equation (see also [9]). Using techniques of analysis appropriate for the discontinuous Galerkin method we can prove bounds on the jump of the error across element boundaries, and *via* duality techniques from [18] we obtain global convergence.

We now describe the problem to be approximated in this paper. Let Ω be a bounded Lipschitz polyhedral (alternatively smooth) domain in $\mathbb{R}^3$ with boundary Γ and consider the problem of finding an acoustic field u such that

$$\Delta u + k^2 u = 0 \text{ in } \Omega, \tag{1.1}$$

$$\frac{\partial u}{\partial \mathbf{n}} + \mathrm{i}k\eta u = -\mathrm{i}kg \text{ on } \Gamma, \tag{1.2}$$

where $\boldsymbol{n}$ is the unit outward normal, $k > 0$ is a real parameter and $\eta \in L^\infty(\Gamma)$ is a strictly positive and bounded function on Γ. The function $g \in L^2(\Gamma)$ is given data. Here we have adopted the sign conventions used in [6] so that the resulting solution of the wave equation is $\Re(u(\boldsymbol{x})\exp(\mathrm{i}\omega t))$ where $\omega = k/c$ is the temporal frequency of the wave and c is the speed of sound.

The standard UWVF uses a more complicated boundary condition

$$\frac{\partial u}{\partial \mathbf{n}} + \mathrm{i}k\eta u = Q\left(\frac{\partial u}{\partial \mathbf{n}} - \mathrm{i}k\eta u\right) - \mathrm{i}kg \text{ on } \Gamma \tag{1.3}$$

where Q is a real parameter with $|Q| \le 1$. This is useful for implementing Dirichlet ($Q = 1$) and Neumann ($Q = -1$) boundary conditions. However we are currently unable to analyze these two important cases (numerical tests show that the method does converge even when $Q = \pm 1$). When $|Q| < 1$ we may rewrite (1.3) as

$$\frac{\partial u}{\partial \boldsymbol{n}} + \mathrm{i}k\frac{(1+Q)}{(1-Q)}\eta u = -\frac{\mathrm{i}k}{(1-Q)}g$$

so it suffices to consider only (1.2) here. Our estimates will be proved under the assumption that there are constants $\eta_{\min}$ and $\eta_{\max}$ such that

$$0 < \eta_{\min} \le \eta(\boldsymbol{x}) \le \eta_{\max} < \infty \tag{1.4}$$

for all $\boldsymbol{x} \in \Gamma$ corresponding to the restriction $|Q| < 1$.

The UWVF can be derived in a variety of ways. In the original work of Cessenat [5] (see also [6,7]) an identity termed the "isometry lemma" was proved using integration by parts on an element. This leads directly and elegantly to the UWVF. In [15] we showed that, in the case of Maxwell's equations, but obviously also for the Helmholtz equation written as a first order system, the UWVF results from a standard upwind discontinuous

Galerkin method with a suitable choice of degrees of freedom. Here we give a third, equivalent, derivation using the techniques introduced to unify the analysis of DG methods in [2] because we wish to use methods from the analysis of DG methods to analyze the UWVF. It is well known (see for example [1]) that the upwind discontinuous Galerkin method can be written in this framework using a suitable choice of fluxes. This derivation is given in Section 2. Very recently a general class of plane wave discontinuous Galerkin methods have been analyzed in [10]. This analysis does not cover the UWVF, but several interesting numerical comparisons of DG and UWVF results are provided.

As we have mentioned previously, the only known error estimates for the UWVF show that the method converges on the boundary Γ and an extra condition is imposed: that $\eta = 1$. We wish to consider more general choices of η to approximate scattering by imperfect conducting obstacles and because it is possible that more general choices of η can be used to help the convergence of the scheme. However the main focus of this paper is to prove global convergence of the method (always seen in our experience of practical calculations). In Section 3 we derive a basic estimate on fluxes given by the UWVF method. We also extend some of the estimates from [5–7] to the case of general η. This theory gives an explicit error estimate in a mesh dependent norm. Under restrictive conditions, we show, in Section 4, that the mesh dependent norm can be used to estimate the standard $L^2(\Omega)$ norm of the error. In Section 5 we then obtain some results for the case of an absorbing medium using the techniques of [5]. In Section 6 we then provide some explicit estimates in 2D and test the results using a simple numerical test case. Finally, in Section 7, we summarize our results and discuss further directions.

2. Derivation of the UWVF

To facilitate the derivation of the UWVF *via* DG techniques, we start by writing (1.1)–(1.2) as a first order system by introducing a field $\boldsymbol{v}$ such that $\mathrm{i}k\boldsymbol{v} = -\nabla u$. Then the problem consists of finding $\boldsymbol{v}$ and u such that

$$-\mathrm{i}k\boldsymbol{v} = \nabla u \text{ in } \Omega, \tag{2.1}$$

$$-\mathrm{i}ku = \nabla \cdot \boldsymbol{v} \text{ in } \Omega, \tag{2.2}$$

$$\eta u - \boldsymbol{v} \cdot \boldsymbol{n} = -g \text{ on } \Gamma. \tag{2.3}$$

Note that this system is a normalized version of the first-order system of linear acoustics and hence more physically relevant than the original Helmholtz equation.

To derive a DG scheme corresponding to this system we suppose Ω is covered by a regular finite element mesh of elements of maximum diameter h denoted $\mathcal{T}_h$. Tetrahedral or hexahedral meshes could be used in principle (or triangle/rectangle elements in 2D).

Following the DG strategy of [2], we now multiply (2.1) and (2.2) by the complex conjugate (denoted by an overbar) of smooth test functions $\boldsymbol{\phi}$ (a vector) and ξ (a scalar) and integrate over an element K in the mesh as follows

$$\int_K (-\mathrm{i}k\boldsymbol{v} \cdot \overline{\boldsymbol{\phi}} + u \nabla \cdot \overline{\boldsymbol{\phi}})\, \mathrm{d}V = \int_{\partial K} u\, \boldsymbol{n}_K \cdot \overline{\boldsymbol{\phi}}\, \mathrm{d}A, \tag{2.4}$$

$$\int_K (-\mathrm{i}ku\, \overline{\xi} + \boldsymbol{v} \cdot \nabla \overline{\xi})\, \mathrm{d}V = \int_{\partial K} \boldsymbol{v} \cdot \boldsymbol{n}_K \overline{\xi}\, \mathrm{d}A, \tag{2.5}$$

where $\boldsymbol{n}_K$ is the unit outward normal to K. For the DG method the fluxes u and $\boldsymbol{v}$ on ∂K are replaced by quantities computed from averages and jumps of appropriate quantities on each element meeting at the appropriate face (*i.e.* we have to allow discontinuous fields, since the numerical scheme will be based on discontinuous expansions). In particular if K and K' are two elements meeting at a face f then we define, on f,

$$\{\!\{u\}\!\} = \frac{u|_K + u|_{K'}}{2}, \qquad \{\!\{\boldsymbol{v}\}\!\} = \frac{\boldsymbol{v}|_K + \boldsymbol{v}|_{K'}}{2},$$

$$[\![u]\!] = u|_K \boldsymbol{n}_K + u|_{K'} \boldsymbol{n}_{K'}, \qquad [\![\boldsymbol{v}]\!] = \boldsymbol{v}|_K \cdot \boldsymbol{n}_K + \boldsymbol{v}|_{K'} \cdot \boldsymbol{n}_{K'}.$$

Using these definitions we choose the flux functions on interior faces f of the grid to be

$$\hat{u} = \{\!\{u\}\!\} + \frac{1}{2\eta}[\![\boldsymbol{v}]\!], \tag{2.6}$$

$$\hat{\boldsymbol{v}} = \{\!\{\boldsymbol{v}\}\!\} + \frac{\eta}{2}[\![u]\!], \tag{2.7}$$

where we have taken a special form of these quantities appropriate for the UWVF compared to standard generalized DG schemes. Here $\eta > 0$ is a bounded strictly positive and piecewise smooth real function on the faces or skeleton of the mesh (a reasonable choice would be $\eta = 1$ on interior faces, and this is the choice made in the original UWVF). Of course on Γ this function (*i.e.* η) is data for the original problem. In particular, this choice corresponds to an upwind DG scheme. On a boundary face, we choose $\hat{u} = u$ and $\hat{\boldsymbol{v}} = \boldsymbol{v}$.

Replacing u on the right hand side of (2.4) with $\hat{u}$ from (2.6) and $\boldsymbol{v}$ on the right hand side of (2.5) with $\hat{\boldsymbol{v}}$ from (2.7) and then adding the result we obtain

$$\int_K \left[\boldsymbol{v}\cdot(\overline{\mathrm{i}k\boldsymbol{\phi}+\nabla\xi}) + u(\overline{\mathrm{i}k\xi+\nabla\cdot\boldsymbol{\phi}})\right]\,\mathrm{d}V = \int_{\partial K}\left(\hat{u}\,\boldsymbol{n}_K\cdot\overline{\boldsymbol{\phi}} + \hat{\boldsymbol{v}}\cdot\boldsymbol{n}_K\overline{\xi}\right)\,\mathrm{d}A.$$

Now we choose $\boldsymbol{\phi}$ and ξ to be smooth solutions of the first order system

$$\mathrm{i}k\boldsymbol{\phi}+\nabla\xi = 0 \text{ and } \mathrm{i}k\xi+\nabla\cdot\boldsymbol{\phi} = 0 \text{ on } K \tag{2.8}$$

(equivalently $\Delta\xi + k^2\xi = 0$ in K) so that we obtain the identity

$$\int_{\partial K}\left(\hat{u}\boldsymbol{n}_K\cdot\overline{\boldsymbol{\phi}} + \hat{\boldsymbol{v}}\cdot\boldsymbol{n}_K\overline{\xi}\right)\,\mathrm{d}A = 0. \tag{2.9}$$

We shall show that adding (2.9) over all elements in the mesh gives rise to the Ultra Weak Variational Formulation of the Helmholtz equation after a suitable choice of degrees of freedom.

Suppose we index the elements K_j, $j = 1,\ldots,N_h$, and suppose $\Sigma_{j,\ell}$ is the face shared between elements K_j and K_ℓ (or the empty set if the two elements do not share a face). In addition let $\boldsymbol{n}_j = \boldsymbol{n}_{K_j}$. If we apply (2.9) on K_j (assuming this is an interior element of the mesh so that it is entirely surrounded by other elements) and use the notation that $u_j = u|_{K_j}$, $\xi_j = \xi|_{K_j}$ (with similar notation for other indexed quantities) we have

$$\int_{\partial K_j}\left(\hat{u}\boldsymbol{n}_j\cdot\overline{\boldsymbol{\phi}}_j + \hat{\boldsymbol{v}}\cdot\boldsymbol{n}_j\overline{\xi}_j\right)\,\mathrm{d}A = \sum_\ell\int_{\Sigma_{j,\ell}}\left(\hat{u}\boldsymbol{n}_j\cdot\overline{\boldsymbol{\phi}}_j + \hat{\boldsymbol{v}}\cdot\boldsymbol{n}_j\overline{\xi}_j\right)\,\mathrm{d}A. \tag{2.10}$$

On each face

$$\begin{aligned}\int_{\Sigma_{j,\ell}}\left(\hat{u}\boldsymbol{n}_j\cdot\overline{\boldsymbol{\phi}}_j + \hat{\boldsymbol{v}}\cdot\boldsymbol{n}_j\overline{\xi}_j\right)\,\mathrm{d}A &= \int_{\Sigma_{j,\ell}}\left(\frac{u_j+u_\ell}{2} + \frac{1}{2\eta}(\boldsymbol{v}_j\cdot\boldsymbol{n}_j + \boldsymbol{v}_\ell\cdot\boldsymbol{n}_\ell)\right)\boldsymbol{n}_j\cdot\overline{\boldsymbol{\phi}}_j\,\mathrm{d}A\\ &\quad+\int_{\Sigma_{j,\ell}}\left(\frac{\boldsymbol{v}_j+\boldsymbol{v}_\ell}{2} + \frac{\eta}{2}(u_j\boldsymbol{n}_j + u_\ell\boldsymbol{n}_\ell)\right)\cdot\boldsymbol{n}_j\overline{\xi}_j\,\mathrm{d}A\\ &= \int_{\Sigma_{j,\ell}}\left(\frac{u_j}{2} + \frac{1}{2\eta}\boldsymbol{v}_j\cdot\boldsymbol{n}_j\right)\boldsymbol{n}_j\cdot\overline{\boldsymbol{\phi}}_j + \left(\frac{\boldsymbol{v}_j}{2} + \frac{\eta}{2}u_j\boldsymbol{n}_j\right)\cdot\boldsymbol{n}_j\overline{\xi}_j\,\mathrm{d}A\\ &\quad+\int_{\Sigma_{j,\ell}}\left(\frac{u_\ell}{2} + \frac{1}{2\eta}\boldsymbol{v}_\ell\cdot\boldsymbol{n}_\ell\right)\boldsymbol{n}_j\cdot\overline{\boldsymbol{\phi}}_j + \left(\frac{\boldsymbol{v}_\ell}{2} + \frac{\eta}{2}u_\ell\boldsymbol{n}_\ell\right)\cdot\boldsymbol{n}_j\overline{\xi}_j\,\mathrm{d}A.\end{aligned}$$

Using the fact that on $\Sigma_{j,\ell}$ the normals are related by $\boldsymbol{n}_\ell = -\boldsymbol{n}_j$ we have

$$\int_{\Sigma_{j,\ell}} \left(\hat{u}\boldsymbol{n}_j \cdot \overline{\boldsymbol{\phi}} + \hat{\boldsymbol{v}} \cdot \boldsymbol{n}_j \overline{\xi}\right) \mathrm{d}A = \int_{\Sigma_{j,\ell}} \left(\frac{u_j}{2} + \frac{1}{2\eta}\boldsymbol{v}_j \cdot \boldsymbol{n}_j\right) \boldsymbol{n}_j \cdot \overline{\boldsymbol{\phi}}_j + \left(\frac{\boldsymbol{v}_j \cdot \boldsymbol{n}_j}{2} + \frac{\eta}{2} u_j\right) \overline{\xi}_j \,\mathrm{d}A$$
$$+ \int_{\Sigma_{j,\ell}} \left(\frac{u_\ell}{2} + \frac{1}{2\eta}\boldsymbol{v}_\ell \cdot \boldsymbol{n}_\ell\right) \boldsymbol{n}_j \cdot \overline{\boldsymbol{\phi}}_j + \left(-\frac{\boldsymbol{v}_\ell \cdot \boldsymbol{n}_\ell}{2} - \frac{\eta}{2} u_\ell\right) \overline{\xi}_j \,\mathrm{d}A.$$

Hence, rearranging once more,

$$\int_{\Sigma_{j,\ell}} \left(\hat{u}\boldsymbol{n}_j \cdot \overline{\boldsymbol{\phi}}_j + \hat{\boldsymbol{v}} \cdot \boldsymbol{n}_j \overline{\xi}_j\right) \mathrm{d}A = \int_{\Sigma_{j,\ell}} \frac{1}{2\eta} \left(\eta u_j + \boldsymbol{v}_j \cdot \boldsymbol{n}_j\right) \left(\overline{\eta\xi_j + \boldsymbol{\phi}_j \cdot \boldsymbol{n}_j}\right) \mathrm{d}A$$
$$- \int_{\Sigma_{j,\ell}} \frac{1}{2\eta} \left(\eta u_\ell + \boldsymbol{v}_\ell \cdot \boldsymbol{n}_\ell\right) \left(\overline{\eta\xi_j - \boldsymbol{\phi}_j \cdot \boldsymbol{n}_j}\right) \mathrm{d}A.$$

Defining $\mathcal{X}_j = (\eta u_j + \boldsymbol{v}_j \cdot \boldsymbol{n}_j)$, $\mathcal{Y}_j = (\eta\xi_j + \boldsymbol{\phi}_j \cdot \boldsymbol{n}_j)$ and $F_j(\mathcal{Y}_j) = (\eta\xi_j - \boldsymbol{\phi}_j \cdot \boldsymbol{n}_j)$ we obtain

$$\int_{\Sigma_{j,\ell}} \left(\hat{u}\boldsymbol{n}_j \cdot \overline{\boldsymbol{\phi}}_j + \hat{\boldsymbol{v}} \cdot \boldsymbol{n}_j \overline{\xi}_j\right) \mathrm{d}A = \int_{\Sigma_{j,\ell}} \frac{1}{2\eta} \mathcal{X}_j \overline{\mathcal{Y}_j} \,\mathrm{d}A - \int_{\Sigma_{j,\ell}} \frac{1}{2\eta} \mathcal{X}_\ell \overline{F_j(\mathcal{Y}_j)} \,\mathrm{d}A. \tag{2.11}$$

Using (2.11) in equation (2.10) we have, for an interior element K_j,

$$\int_{\partial K_j} \frac{1}{2\eta} \mathcal{X}_j \overline{\mathcal{Y}}_j \,\mathrm{d}A - \sum_{\ell=1,\, \ell\neq j}^{N_h} \int_{\Sigma_{j,\ell}} \frac{1}{2\eta} \mathcal{X}_\ell \overline{F_j(\mathcal{Y}_j)} \,\mathrm{d}A = 0. \tag{2.12}$$

On a boundary face (or perhaps a union of boundary faces) we must proceed slightly differently. Suppose $\Gamma_j = \partial K_j \cap \Gamma$ and recall that we define the flux functions to be $\hat{u} = u$ and $\hat{\boldsymbol{v}} = \boldsymbol{v}$ on such faces so the relevant contribution to (2.9) on Γ_j is

$$\begin{aligned} \int_{\Gamma_j} (u_j\, \boldsymbol{n}_j \cdot \overline{\boldsymbol{\phi}}_j + \boldsymbol{v}_j \cdot \boldsymbol{n}_j\, \overline{\xi}_j) \,\mathrm{d}A \quad &= \quad \int_{\Gamma_j} \frac{1}{2\eta} (\eta u_j + \boldsymbol{v}_j \cdot \boldsymbol{n}_j)(\overline{\eta\xi_j + \boldsymbol{\phi}_j \cdot \boldsymbol{n}_j}) \,\mathrm{d}A \\ &\quad - \int_{\Gamma_j} \frac{1}{2\eta} (\eta u_j - \boldsymbol{v}_j \cdot \boldsymbol{n}_j)(\overline{\eta\xi_j - \boldsymbol{\phi}_j \cdot \boldsymbol{n}_j}) \,\mathrm{d}A \\ &= \quad \int_{\Gamma_j} \frac{1}{2\eta} \mathcal{X}_j\, \overline{\mathcal{Y}_j} \,\mathrm{d}A - \int_{\Gamma_j} \frac{1}{2\eta} F_j(\mathcal{X}_j)\, \overline{F_j(\mathcal{Y}_j)} \,\mathrm{d}A. \end{aligned} \tag{2.13}$$

For the solution of the original boundary value problem, we want u and $\boldsymbol{v}$ to satisfy the boundary condition (1.2) and so replace $F_j(\mathcal{X}_j)$ by g on the right hand side above to obtain the following contribution for a boundary face:

$$\int_{\Gamma_j} (u_j\, \boldsymbol{n}_j \cdot \overline{\boldsymbol{\phi}}_j + \boldsymbol{v}_j \cdot \boldsymbol{n}_j\, \overline{\xi}_j) \,\mathrm{d}A \quad = \quad \int_{\Gamma_j} \frac{1}{2\eta} \mathcal{X}_j\, \overline{\mathcal{Y}_j} \,\mathrm{d}A - \int_{\Gamma_j} \frac{1}{2\eta} g\, \overline{F_j(\mathcal{Y}_j)} \,\mathrm{d}A. \tag{2.14}$$

Let $L^2_\eta(\partial K)$ denote the set of functions in $L^2(\partial K)$ with the weighted norm

$$\|u\|^2_{L^2_\eta(\partial K)} = \int_{\partial K} \frac{1}{\eta} |u|^2 \,\mathrm{d}s.$$

Adding expressions (2.11) and (2.14) to obtain the contributions for each face in (2.9) and defining $\Gamma_j = \emptyset$ if $\Gamma \cap \partial\Omega_j = \emptyset$ we obtain the problem of finding $\mathcal{X}_j \in L^2_\eta(\partial K_j)$, $1 \le j \le N_h$ such that

$$\int_{\partial K_j} \frac{1}{2\eta}\mathcal{X}_j\overline{\mathcal{Y}}_j \,\mathrm{d}A - \sum_{\ell=1,\,\ell\neq j}^{N_h} \int_{\Sigma_{j,\ell}} \frac{1}{2\eta}\mathcal{X}_\ell \overline{F_j(\mathcal{Y}_j)} \,\mathrm{d}A = \int_{\Gamma_j} \frac{1}{2\eta} g\,\overline{F_j(\mathcal{Y}_j)} \,\mathrm{d}A \tag{2.15}$$

for all $\mathcal{Y}_j \in L^2_\eta(\partial K_j)$ and for $1 \le j \le N_h$. Note that the sum does not include faces on the boundary because $\Sigma_{j,\ell}$ is always a face between two elements.

Equation (2.17) is the Ultra Weak Variational Formulation (UWVF) of the Helmholtz equation before discretization generalized to the case when η is variable. We rewrite this further by defining $\vec{\mathcal{X}} = (\mathcal{X}_1, \ldots, \mathcal{X}_{N_h})$, $\mathcal{X}_j \in L^2_\eta(\partial K_j)$, $1 \le j \le N_h$ and similarly $\vec{\mathcal{Y}}$ we set

$$a(\vec{\mathcal{X}}, \vec{\mathcal{Y}}) = \sum_{j=1}^{N_h} \left[\int_{\partial K_j} \frac{1}{2\eta}\mathcal{X}_j\overline{\mathcal{Y}}_j \,\mathrm{d}A - \sum_{\ell=1,\,\ell\neq j}^{N_h} \int_{\Sigma_{j,\ell}} \frac{1}{2\eta}\mathcal{X}_\ell \overline{F_j(\mathcal{Y}_j)} \,\mathrm{d}A \right]. \tag{2.16}$$

In addition we define the data term

$$b(\vec{\mathcal{Y}}) = \sum_{j=1}^{N_h} \int_{\Gamma_j} \frac{1}{2\eta} g\,\overline{F_j(\mathcal{Y}_j)} \,\mathrm{d}A,$$

where we recall that $\Gamma_j = \Gamma \cap \partial K_j$ is empty if K is an interior triangle. Let $X = \Pi_{K\in\mathcal{T}_h} L^2_\eta(\partial K)$ so that X has the norm

$$\|\mathcal{X}\|^2_X = \sum_{j=1}^{N_h} \int_{\partial K_j} \frac{1}{\eta}|\mathcal{X}_j|^2 \,\mathrm{d}A$$

and inner product

$$(\vec{\mathcal{X}}, \vec{\mathcal{Y}})_X = \sum_{j=1}^{N_h} \int_{\partial K_j} \frac{1}{2\eta}\mathcal{X}_j\overline{\mathcal{Y}}_j \,\mathrm{d}A.$$

Then the UWVF can be written as the problem of finding $\vec{\mathcal{X}} \in X$ such that

$$a(\vec{\mathcal{X}}, \vec{\mathcal{Y}}) = b(\vec{\mathcal{Y}}), \quad \text{for all } \vec{\mathcal{Y}} \in X. \tag{2.17}$$

We now turn to the discrete system. Let $X^h_{K_j} \subset L^2_\eta(\partial K_j)$ for $1 \le j \le N_h$ denote a collection of finite dimensional spaces and set $X^h = \Pi_{j=1}^{N_h} X^h_{K_j}$. Then we seek $\vec{\mathcal{X}}^h \in X^h$ such that

$$a(\vec{\mathcal{X}}^h, \vec{\mathcal{Y}}^h) = b(\vec{\mathcal{Y}}^h), \quad \text{for all } \vec{\mathcal{Y}}^h \in X^h. \tag{2.18}$$

It is shown in [5,6] that this system has a unique solution.

Note that (2.18) is not necessarily easy to implement. In particular for any $\mathcal{Y}^h_j \in X^h_{K_j}$ we need to compute $F_j(\mathcal{Y}^h_j)$ which involves solving the boundary value problem

$$\begin{aligned} \Delta\xi_j + k^2\xi_j &= 0 \text{ on } K_j, \\ \eta\xi_j - \frac{1}{\mathrm{i}k}\frac{\partial\xi_j}{\partial \boldsymbol{n}_j} &= \mathcal{Y}^h_j \text{ on } \partial K_j, \end{aligned}$$

and setting

$$F_j(\mathcal{Y}^h_j) = \eta\xi_j + \frac{1}{\mathrm{i}k}\frac{\partial\xi_j}{\partial \boldsymbol{n}_j} \text{ on } \partial K_j. \tag{2.19}$$

Cessenat and Després point out that if we first choose an auxiliary finite dimensional space $M^h_{K_j}$ of solutions of the Helmholtz equation on K_j then for $\xi_j \in M^h_{K_j}$ we may set

$$\mathcal{Y}^h_j = \eta\xi_j - \frac{1}{\mathrm{i}k}\frac{\partial \xi_j}{\partial \boldsymbol{n}_j} \text{ on } \partial K_j$$

and to find $F_j(\mathcal{Y}^h_j)$ it suffices to use (2.19). In this case

$$X^h_{K_j} = \left\{ \eta\xi_j - \frac{1}{\mathrm{i}k}\frac{\partial \xi_j}{\partial \boldsymbol{n}_j} \;\middle|\; \xi_j \in M^h_{K_j} \right\}.$$

The price to pay for this indirect choice of $X^h_{K_j}$ is that the accuracy of the solution now depends on the properties of appropriate traces of the functions in $M^h_{K_j}$ which are less well known than for polynomial basis functions used in standard finite element methods.

Cessenat and Després suggest to choose p_j independent directions $\boldsymbol{d}^{K_j}_\ell$, $1 \le \ell \le p_j$ where $|\boldsymbol{d}^{K_j}_\ell| = 1$ and take $M^h_{K_j}$ to be a space of plane wave solutions of the Helmholtz equation in each direction

$$M^h_{K_j} = \operatorname{span}\left\{ \exp(\mathrm{i}k\boldsymbol{d}^{K_j}_\ell \cdot \boldsymbol{x}),\ 1 \le \ell \le p_j \right\}. \tag{2.20}$$

In fact Cessenat and Després choose $p_j = p$ for all j (*i.e.* the same number of plane waves on each element). We have found it necessary to choose p_j depending on the element K_j and wave number k in order to control ill-conditioning in certain matrices in the formulation [13].

3. Basic error estimates

In order to estimate the error we now set $\vec{e}^h = \vec{\mathcal{X}} - \vec{\mathcal{X}}^h$. We also define $P^h : X \to X^h$ to be the best approximation (projection) operator in the X norm. We shall use several results from [5,6]. These are proved in the case $\eta = 1$ but the proofs carry over directly to the current case. In particular Lemma 8 of Section I.3.3 of [5] (see also Lem. 3.1 of [6]) can be used to show the following estimate.

Lemma 3.1. *The following estimate holds where $P^h : X \to X^h$ is the X^h best approximation operator:*

$$|a(\vec{e}^h, \vec{e}^h)| \le 2\|(I - P^h)\vec{\mathcal{X}}\|^2_X. \tag{3.1}$$

Proof. Since $\vec{e}^h = \vec{\mathcal{X}} - \vec{\mathcal{X}}^h$ the following Galerkin property holds by subtracting (2.18) from (2.17):

$$a(\vec{e}^h, \vec{\mathcal{Y}}^h) = 0, \qquad \text{for all } \vec{\mathcal{Y}}^h \in X^h.$$

Hence

$$a(\vec{e}^h, \vec{e}^h) = a(\vec{e}^h, (I - P^h)\vec{\mathcal{X}}) + a(\vec{e}^h, P^h\vec{\mathcal{X}} - \vec{\mathcal{X}}^h) = a(\vec{e}^h, (I - P^h)\vec{\mathcal{X}}). \tag{3.2}$$

To complete the proof we can use Lemma 3.1 of [6] after introducing some additional notation. In particular, following [6], we introduce the operators $\Pi : X \to X$ defined such that if $\Sigma_{j,k} \ne \phi$ then

$$(\Pi\mathcal{X}_j)|_{\Sigma_{j,k}} = \mathcal{X}_k$$

and if $\Gamma_j \ne \phi$ then

$$(\Pi\mathcal{X}_j)|_{\Gamma_j} = 0.$$

This operator performs the task of selecting information from adjoining elements (by faces in 3D or edges in 2D).

In addition $\vec{F} : X \to X$ is defined such that $\vec{F} = (F_1, \ldots, F_{N_h})$ where $F_j : L^2(\partial T_j) \to L^2(\partial T_j)$ is given by (2.19). It is then easy to see that (2.16) is equal to

$$a(\vec{\mathcal{X}}, \vec{\mathcal{Y}}) = (\vec{\mathcal{X}}, \vec{\mathcal{Y}})_X - (\Pi\vec{\mathcal{X}}, \vec{F}(\vec{\mathcal{Y}}))_X$$

for all $\vec{\mathcal{X}}, \vec{\mathcal{Y}} \in X$. Hence

$$|a(\vec{\mathcal{X}}, \vec{\mathcal{Y}})| = |((I - A)\vec{\mathcal{X}}, \vec{\mathcal{Y}})_X| \leq \|(I - A)\vec{\mathcal{X}}\|_X \|\vec{\mathcal{Y}}\|_X, \tag{3.3}$$

where $A = \vec{F}^*\Pi$ and the superscript $*$ denotes the adjoint operator in the X inner-product. Cessenat and Després (Prop. 1.10 of [6]) show that the operator norm of $A : X \to X$ satisfies $\|A\|_{X\to X} \leq 1$ and in Lemma 3.1 of the same paper, they show that this implies that

$$\|(I - A)\vec{e}^h\|_X \leq 2\|(I - P^h)\vec{\mathcal{X}}\|_X.$$

These results hold for general η. Using (3.3) and this result in (3.2) proves the desired estimate. □

The next result is the main convergence estimate of [6], again extended directly to general η satisfying (1.4).

Lemma 3.2.

$$\|\vec{e}^h\|_{L^2_\eta(\Gamma)} \leq \sqrt{2}\|(I - P^h)\vec{\mathcal{X}}\|_X \tag{3.4}$$

where

$$\|\vec{e}^h\|^2_{L^2_\eta(\Gamma)} = \sum_{j=1}^{N_h} \int_{\Gamma_j} \frac{1}{2\eta} \left|\mathcal{X}_j - \mathcal{X}_j^h\right|^2 \mathrm{d}A.$$

Our goal for this section is to prove global convergence of the UWVF even away from the boundary. To analyze this problem, we show that $a(\cdot, \cdot)$ has sufficient coercivity to provide an error estimate. This is done by relating the method back to the original DG scheme *via* an auxiliary sesquilinear form that we now define. Given $\vec{\mathcal{X}}, \vec{\mathcal{Y}} \in X$ define $(u_j, \boldsymbol{v}_j)$ to satisfy

$$\begin{aligned} -\mathrm{i}k\boldsymbol{v}_j &= \nabla u_j \text{ in } K_j, & (3.5)\\ -\mathrm{i}ku &= \nabla \cdot \boldsymbol{v}_j \text{ in } K_j, & (3.6)\\ \eta u_j + \boldsymbol{v}_j \cdot \boldsymbol{n}_j &= \mathcal{X}_j \text{ on } \partial K_j, & (3.7) \end{aligned}$$

and $(\xi_j, \boldsymbol{\phi}_j)$ to satisfy

$$\begin{aligned} -\mathrm{i}k\boldsymbol{\phi}_j &= \nabla \xi_j \text{ in } K_j, & (3.8)\\ -\mathrm{i}k\xi_j &= \nabla \cdot \boldsymbol{\phi}_j \text{ in } K_j, & (3.9)\\ \eta\xi_j + \boldsymbol{\phi}_j \cdot \boldsymbol{n}_j &= \mathcal{Y}_j \text{ on } \partial K_j. & (3.10) \end{aligned}$$

Now we define the following auxiliary sesquilinear form:

$$a_0(\vec{\mathcal{X}}, \vec{\mathcal{Y}}) = \sum_{j=1}^{N_h} \int_{\partial K_j} \left(\hat{u}\boldsymbol{n}_j \cdot \overline{\boldsymbol{\phi}}_j + \hat{\boldsymbol{v}} \cdot \boldsymbol{n}_j \overline{\xi}_j\right) \mathrm{d}A \tag{3.11}$$

where $\hat{u}$ and $\hat{\boldsymbol{v}}$ are the numerical fluxes defined in (2.6) and (2.7). We can now rewrite this sesquilinear form as follows by rewriting the definition in terms of a sum over faces in the grid (using (3.5)–(3.10) to extend $\mathcal{X}_j$ and $\mathcal{Y}_j$ into each element K_j):

$$a_0(\vec{\mathcal{X}}, \vec{\mathcal{Y}}) = \sum_{f:\text{interior faces}} \int_f \left(\hat{u}_f \overline{[\![\boldsymbol{\phi}]\!]_f} + \hat{\boldsymbol{v}}_f \cdot \overline{[\![\xi]\!]_f}\right) \mathrm{d}A + \sum_{f:\text{boundary faces}} \int_f \left(u_f \boldsymbol{n} \cdot \overline{\boldsymbol{\phi}}_f + \boldsymbol{v}_f \cdot \boldsymbol{n}\overline{\xi}_f\right) \mathrm{d}A \tag{3.12}$$

where we have emphasized that dependence on the faces f by using the subscript f to denote the restriction of the appropriate quantity to f so $\hat{u}_f = \hat{u}|_f$.

We have the following result.

Lemma 3.3. *The bilinear forms a_0 (see (3.11)) and a (see (2.16)) are related by*

$$a(\vec{\mathcal{X}}, \vec{\mathcal{Y}}) = a_0(\vec{\mathcal{X}}, \vec{\mathcal{Y}}) + \sum_j \int_{\Gamma_j} \frac{1}{2\eta} F_j(\mathcal{X}_j)\overline{F_j(\mathcal{Y}_j)}\,\mathrm{d}A.$$

for all $\vec{\mathcal{X}}, \vec{\mathcal{Y}} \in X$.

Proof. Summing over interior and boundary faces in the mesh using (2.11) and (2.13) as appropriate shows that

$$\sum_{j=1}^{N_h} \int_{\partial K_j} \left(\hat{u}\boldsymbol{n}_K \cdot \overline{\boldsymbol{\phi}}_j + \hat{\boldsymbol{v}} \cdot \boldsymbol{n}_j \overline{\xi}_j\right) \mathrm{d}A = a(\vec{\mathcal{X}}, \vec{\mathcal{Y}}) - \sum_j \int_{\Gamma_j} \frac{1}{2\eta} F_j(\mathcal{X}_j)\overline{F_j(\mathcal{Y}_j)}\,\mathrm{d}A,$$

as required. □

In particular we estimate $a(\vec{\mathcal{X}}, \vec{\mathcal{X}})$ as summarized in the following lemma.

Lemma 3.4. *Suppose (1.4) holds at every point on the skeleton (faces) of the mesh. Then*

$$\Re(a(\vec{\mathcal{X}}, \vec{\mathcal{X}})) = \sum_{f:\text{interior faces}} \int_f \left(\frac{\eta}{2}|[\![u]\!]_f|^2 + \frac{1}{2\eta}|[\![\boldsymbol{v}]\!]_f|^2\right) \mathrm{d}A + \sum_j \int_{\Gamma_j} \frac{1}{2\eta}|F_j(\mathcal{X}_j)|^2\,\mathrm{d}A. \tag{3.13}$$

Proof. Lemma 3.3 shows that

$$a(\vec{\mathcal{X}}, \vec{\mathcal{X}}) = a_0(\vec{\mathcal{X}}, \vec{\mathcal{X}}) + \sum_j \int_{\Gamma_j} \frac{1}{2\eta}|F_j(\mathcal{X}_j)|^2\,\mathrm{d}A. \tag{3.14}$$

To prove coercivity we take $\xi = u_j$ and $\boldsymbol{\phi} = \boldsymbol{v}_j$ on each element K_j (of course these are not independent quantities) in (3.12) to obtain

$$a_0(\vec{\mathcal{X}}, \vec{\mathcal{X}}) = \sum_{f:\text{interior faces}} \int_f \hat{u}_f\overline{[\![\boldsymbol{v}]\!]_f} + \hat{\boldsymbol{v}}_f \cdot \overline{[\![u]\!]_f}\,\mathrm{d}A + \sum_{f:\text{boundary faces}} \int_f u_f\boldsymbol{n} \cdot \overline{\boldsymbol{v}}_f + \boldsymbol{v}_f \cdot \boldsymbol{n}\overline{u}_f\,\mathrm{d}A.$$

On each interior face, using the definition of $\hat{u}$ and $\hat{\boldsymbol{v}}$ we have

$$\hat{u}_f\overline{[\![\boldsymbol{v}]\!]_f} + \hat{\boldsymbol{v}}_f \cdot \overline{[\![u]\!]_f} = \{\!\{u\}\!\}_f\overline{[\![\boldsymbol{v}]\!]_f} + \{\!\{\boldsymbol{v}\}\!\}_f \cdot \overline{[\![u]\!]_f} + \frac{\eta}{2}|[\![u]\!]_f|^2 + \frac{1}{2\eta}|[\![\boldsymbol{v}]\!]_f|^2.$$

Then if two elements K_1 and K_2 meet at f and if we write $u_1 = u|_{K_1}$ and $u_2 = u|_{K_2}$ (similarly $\boldsymbol{v}_1$ and $\boldsymbol{v}_1$), and if $\boldsymbol{n}_1$ is the unit outward normal to K_1 and $\boldsymbol{n}_2$ is the unit outward normal to K_2

$$\begin{aligned}\{\!\{u\}\!\}_f\overline{[\![\boldsymbol{v}]\!]_f} + \{\!\{\boldsymbol{v}\}\!\} \cdot \overline{[\![u]\!]_f} \quad &= \quad (u_1\overline{\boldsymbol{v}_1} \cdot \boldsymbol{n}_1 + \boldsymbol{v}_1 \cdot \boldsymbol{n}_1\overline{u_1}) + (u_1\overline{\boldsymbol{v}_2} \cdot \boldsymbol{n}_2 + \boldsymbol{v}_2 \cdot \boldsymbol{n}_1\overline{u_1}) \\ &\quad + (u_2\overline{\boldsymbol{v}_1} \cdot \boldsymbol{n}_1 + \boldsymbol{v}_1 \cdot \boldsymbol{n}_2\overline{u_2}) + (u_2\overline{\boldsymbol{v}_2} \cdot \boldsymbol{n}_2 + \boldsymbol{v}_2 \cdot \boldsymbol{n}_2\overline{u_2}).\end{aligned}$$

Taking into account the change in sign of the normals, and denoting by $\Re(.)$ the real part of the corresponding expression we have

$$\Re[(u_1\overline{\boldsymbol{v}_2} \cdot \boldsymbol{n}_2 + \boldsymbol{v}_2 \cdot \boldsymbol{n}_1\overline{u_1}) + (u_2\overline{\boldsymbol{v}_1} \cdot \boldsymbol{n}_1 + \boldsymbol{v}_1 \cdot \boldsymbol{n}_2\overline{u_2})] = 0.$$

Now using this in the above expression we obtain

$$\Re(a_0(\vec{\mathcal{X}},\vec{\mathcal{X}})) \;=\; \sum_K \int_{\partial K} (u_K\overline{\boldsymbol{v}_K}\cdot\boldsymbol{n}_K + \boldsymbol{v}_K\cdot\boldsymbol{n}_K\overline{u_K})\,\mathrm{d}A \\ \sum_{f:\text{interior faces}} \int_f \left(\frac{\eta}{2}|[\![u]\!]_f|^2 + \frac{1}{2\eta}|[\![\boldsymbol{v}]\!]_f|^2\right)\mathrm{d}A.$$

Using the fact that $\boldsymbol{v} = -(\mathrm{i}k)^{-1}\nabla u$ on each element, the first term on the right hand side above can be written as

$$\int_{\partial K} (u_K\overline{\boldsymbol{v}_K}\cdot\boldsymbol{n}_K + \boldsymbol{v}_K\cdot\boldsymbol{n}_K\overline{u_K})\,\mathrm{d}A = (\mathrm{i}k)^{-1}\int_{\partial K}\left(u\overline{\frac{\partial u}{\partial n_K}} - \frac{\partial u}{\partial n_K}\overline{u}\right)\mathrm{d}A = 0$$

by Green's second identity and the Helmholtz equation. Thus, we have

$$\Re(a_0(\vec{\mathcal{X}},\vec{\mathcal{X}})) \;=\; \sum_{f:\text{interior faces}} \int_f \left(\frac{\eta}{2}|[\![u]\!]_f|^2 + \frac{1}{2\eta}|[\![\boldsymbol{v}]\!]_f|^2\right)\mathrm{d}A.$$

Using this estimate in (3.14) proves the result. □

We now can use (3.13) to derive the following global error estimate involving a mesh dependent norm:

Theorem 3.5. *Suppose η satisfies (1.4). Let $u_j^h, \boldsymbol{v}_j^h$ denote the solution of (3.5)–(3.7) when $\mathcal{X}_j$ is replaced by $\mathcal{X}_j^h$, and let $u^h \in L^2(\Omega)$ be such that $u^h|_{K_j} = u_j$, $1 \le j \le N_h$ (and similarly $\boldsymbol{v}^h|_{K_j} = \boldsymbol{v}_j$). Then we have the error estimate*

$$\sum_{f:\text{interior faces}} \int_f \left(\frac{\eta}{2}|[\![u-u^h]\!]_f|^2 + \frac{1}{2\eta}|[\![\boldsymbol{v}-\boldsymbol{v}^h]\!]_f|^2\right)\mathrm{d}A + \sum_j \int_{\Gamma_j} \frac{1}{2\eta}|F_j(\mathcal{X}_j - \mathcal{X}_j^h)|^2\,\mathrm{d}A \\ + \sum_j \int_{\Gamma_j} \frac{1}{2\eta}|\mathcal{X}_j - \mathcal{X}_j^h|^2\,\mathrm{d}A \le 4\|(I-P^h)\vec{\mathcal{X}}\|_X^2. \quad (3.15)$$

Remark 3.6.

(1) It is a little surprising to have a precise constant on the right hand side of the error estimate (usually, in error estimates for the Helmholtz equation, there appears a constant C depending on k in an unknown way). We might expect "pollution error" to appear in this estimate, but this type of error may be hidden in the mesh dependent norm on the left and right hand sides of the estimate.

(2) The convergence rate observed in practice will depend on the smoothness of the solution $\mathcal{X}$ and on the subspaces $X_{K_j}^h$, $1 \le j \le N_h$ (and hence on the auxiliary subspaces $M_{K_j}^h$) used for the calculation.

Proof. Using the definition of $(u_j^h, \boldsymbol{v}_j^h)$ in the theorem, and using the conclusion of Lemmas 3.1 and 3.2 we have

$$\sum_{f:\text{interior faces}} \int_f \left(\frac{\eta}{2}|[\![u-u^h]\!]_f|^2 + \frac{1}{2\eta}|[\![\boldsymbol{v}-\boldsymbol{v}^h]\!]_f|^2\right)\mathrm{d}A \quad + \sum_j \int_{\Gamma_j} \frac{1}{2\eta}|F_j(e_j^h)|^2\,\mathrm{d}A = \Re(a(\vec{e}^h,\vec{e}^h)) \\ \le |a(\vec{e}^h,\vec{e}^h)| \le 2\|(I-P^h)\vec{\mathcal{X}}\|_X^2. \quad (3.16)$$

We can then combine this estimate with (3.4) to proved the desired estimate. □

Although we know from Cessenat and Després' work that the UWVF and discrete UWVF both have unique solutions, the previous theorem provides a direct proof. We concentrate on the discrete problem.

Corollary 3.7. *The discrete problem* (2.18) *has exactly one solution for any choice of* $g \in L^2(\Gamma)$.

Proof. The discrete UWVF (2.18) results in a square linear system so it suffices to consider the case when $g = 0$ (in that case we know the only solution of the continuous problem is $u = 0$ and $\boldsymbol{v} = 0$). Suppose $\vec{\mathcal{X}}^h$ is a solution of (2.18). Defining u_j and $\boldsymbol{v}_j$, $j = 1, \ldots, N_h$ using (3.5)–(3.7) with $\mathcal{X}_j$ replaced by $\mathcal{X}_j^h$ we conclude from (3.15) that

$$\sum_{f:\text{interior faces}} \int_f \left(\frac{\eta}{2} |[\![u^h]\!]_f|^2 + \frac{1}{2\eta} |[\![\boldsymbol{v}^h]\!]_f|^2 \right) \mathrm{d}A + \sum_j \int_{\Gamma_j} \frac{1}{2\eta} |F_j(\mathcal{X}_j^h)|^2 \, \mathrm{d}A + \sum_j \int_{\Gamma_j} \frac{1}{2\eta} |\mathcal{X}_j^h|^2 \, \mathrm{d}A = 0.$$

The vanishing jumps imply that the composite solution on all the elements is a global solution of the Helmholtz equation, and the vanishing data $\mathcal{X}_j$ on Γ_j implies that this solution is identically zero. Hence we have verified the result. □

4. An error estimate in $L^2(\Omega)$

While the estimate in Theorem 3.5 certainly implies the convergence of the UWVF throughout the domain provided the spaces $X_{K_j}^h$, $1 \le j \le N_h$, have good approximation properties, it does not explicitly predict a convergence rate because the left and right hand sides of (3.15) are mesh dependent. To obtain an explicit global estimate we now assume Ω is convex and $\eta = 1$ (because we wish to use a duality estimate). We shall estimate the $L^2(\Omega)$ error of the solution computed from the approximate solution $\vec{\mathcal{X}}^h$ of the discrete UWVF. Obviously other norms of the solution would also be interesting. It is possible that the piecewise $H^1(\Omega)$ norm (or "broken" H^1 norm in DG terminology) could also be estimated by similar techniques, but we have not pursued that direction.

Theorem 4.1. *Suppose the mesh is regular and quasi-uniform, that* Ω *is a convex polyhedral domain in* $\mathbb{R}^3$, *and that* $\eta = 1$ *then*

$$\|u - u^h\|_{L^2(\Omega)} \le C h^{-1/2} \|(I - P^h)\vec{\mathcal{X}}\|_X$$

where C *depends on* k *and* Ω *but is independent of* h *and* u.

Remark 4.2. It is likely that the assumption on η can be relaxed. The assumption that Ω is convex is used to ensure that the solution of a certain dual problem has $H^2(\Omega)$ regularity. Unfortunately the k dependence of the constant C cannot derived from our proof since the proof is based on Theorem 3.1 of [18] which uses *a priori* estimates for the solution of the Helmholtz equation. The k dependence of these estimates is only know in special cases [8].

Proof. To derive the error estimate we need to use an estimate from Theorem 3.1 of [18] that is proved for a domain in $\mathbb{R}^2$ but which can be extended to $\mathbb{R}^3$. This shows that for any piecewise solution w of the Helmholtz equation on a regular and quasi-uniform mesh (*i.e.* w_j is a solution of the Helmholtz equation on each element K_j, $1 \le j \le N_h$ in the mesh)

$$\|w\|_{L^2(\Omega)}^2 \le \frac{C}{h} \left[\sum_{f:\text{interior faces}} \|[\![\nabla w]\!]\|_{L^2(f)}^2 + k^2 \|[\![w]\!]\|_{L^2(f)}^2 + \sum_{f:\text{boundary faces}} \left\| \frac{\partial w}{\partial \boldsymbol{n}} - \mathrm{i}kw \right\|_{L^2(f)}^2 \right].$$

Using $w = u - u^h$ (where u^h is the composite solution found from $\mathcal{X}_j^h$ on each element), taking into account the definition of the flux $\boldsymbol{v}$, and using C as a generic constant independent of h, u and u_h we then have

$$\|u-u^h\|^2_{L^2(\Omega)} \le \frac{C}{h}\left[\sum_{f:\text{interior faces}} \|[\![\frac{1}{\mathrm{i}k}\nabla(u-u_h)]\!]\|^2_{L^2(e)} + \|[\![u-u_h]\!]\|^2_{L^2(f)} \right.$$
$$\left. + \sum_{f:\text{boundary faces}} \left\| \frac{1}{\mathrm{i}k}\frac{\partial(u-u_h)}{\partial \boldsymbol{n}} - (u-u_h)\right\|^2_{L^2(f)}\right] \le \frac{C}{h}\left[|a(\vec{e}^h,\vec{e}^h)| + \|\vec{e}^h\|^2_{L^2(\Gamma)}\right].$$

Then using estimate (3.15) we have proved the error estimate. □

5. Estimates for an absorbing medium

The UWVF can also be applied to an absorbing medium, but the derivation of the UWVF given in Section 2 no longer holds. Thus the error estimates derived using this point of view also do not apply. In this case the unknown field u satisfies

$$\begin{aligned} \Delta u + k^2 n u &= 0 \text{ in } \Omega, \\ \frac{\partial u}{\partial \mathbf{n}} + \mathrm{i}k\eta u &= -\mathrm{i}kg \text{ on } \Gamma, \end{aligned}$$

where $n = n_1 - \mathrm{i}n_2$ and where n_1 and n_2 are real and $n_1 > 0$ and $n_2 \ge 0$. We assume that n_1 and n_2 are piecewise constant on the UWVF mesh. The choice of the sign for the imaginary part of n is dictated by the choice of the sign in the boundary condition and corresponds to a time variation of $\exp(\mathrm{i}\omega t)$ where $\omega = kc$ is the temporal frequency of the wave and c is the wave speed.

Following the derivation in [5], Section I.5, the UWVF (2.17) holds with one modification: the operator $\vec{F} : X \to X$ must be computed *via* the adjoint equation so that if $\mathcal{Y}_j \in L^2(\partial K_j)$, $1 \le j \le N_h$, then

$$F_j(\mathcal{Y}_j) = \eta\xi_j + \frac{1}{\mathrm{i}k}\frac{\partial \xi_j}{\partial \boldsymbol{n}_j} \text{ on } \partial K_j,$$

where ξ_j satisfies

$$\Delta\xi_j + k^2\overline{n}\xi_j = 0 \text{ in } K_j, \tag{5.1}$$
$$\eta\xi_j - \frac{1}{\mathrm{i}k}\frac{\partial \xi_j}{\partial \boldsymbol{n}_j} = \mathcal{Y}_j \text{ on } \partial K_j, \tag{5.2}$$

where $\overline{n} = n_1 + \mathrm{i}n_2$. This interior problem is well posed for any $k > 0$.

It is no longer true that $\vec{F}$ is an isometry. Instead we have the following estimate (see Lem. 11 of Sect. I.1.5.13 of [5], proved here for convenience).

Lemma 5.1. *The operators $F_j : L^2_\eta(\partial K_j) \to L^2_\eta(\partial K_j)$, $1 \le j \le N_h$, satisfy the estimate*

$$\|F_j(\mathcal{Y}_j)\|^2_{L^2_\eta(\partial K_j)} = \|\mathcal{Y}_j\|^2_{L^2_\eta(\partial K_j)} - 4k\int_{K_j} n_2|\xi_j|^2 \,\mathrm{d}V.$$

Proof. Proceeding directly using the definitions of $\mathcal{Y}_j$ and $F_j(\mathcal{Y}_j)$,

$$\|F_j(\mathcal{Y}_j)\|^2_{L^2_\eta(\partial K_j)} = \|\mathcal{Y}_j\|^2_{L^2_\eta(\partial K_j)} + \frac{2}{\mathrm{i}k}\int_{\partial K_j}\left(\frac{\partial \xi_j}{\partial \boldsymbol{n}_j}\overline{\xi_j} - \overline{\frac{\partial \xi_j}{\partial \boldsymbol{n}_j}}\xi_j\right)\mathrm{d}A.$$

Using Green's theorem and the adjoint Helmholtz equation

$$\begin{aligned}\|F_j(\mathcal{Y}_j)\|^2_{L^2_\eta(\partial K_j)} &= \|\mathcal{Y}_j\|^2_{L^2_\eta(\partial K_j)} + \frac{2}{\mathrm{i}k}\int_{K_j}\left((-k^2\overline{n}\xi_j)\overline{\xi_j} - \overline{(-k^2\overline{n}\xi_j)}\xi_j\right)\mathrm{d}V \\ &= \|\mathcal{Y}_j\|^2_{L^2(\partial K_j)} - 2\mathrm{i}k\int_{K_j}|\xi_j|^2(n-\overline{n})\,\mathrm{d}V,\end{aligned}$$

and the proof is complete. □

The discrete UWVF (2.18) also holds provided $\vec{F}$ is interpreted in the sense of this section. A practical implementation now uses the plane wave solutions of the adjoint problem on each element: again, on element K_j, we choose p_j directions $\boldsymbol{d}_\ell^{K_j}$, $1 \le \ell \le p_j$ where $|\boldsymbol{d}_\ell^{K_j}| = 1$ and take $M_h^{K_j}$ to be a space of plane wave solutions of equation (5.1)

$$M_h^{K_j} = \operatorname{span}\left\{\exp(\mathrm{i}k\sqrt{\overline{n}}\boldsymbol{d}_\ell^{K_j}\cdot\boldsymbol{x}),\ 1 \le \ell \le p_j\right\}. \tag{5.3}$$

No error estimates are available for the problem in this case. Here we prove the analogue of the fundamental error estimate Lemma 3.2.

Lemma 5.2. *Let ε_j denote the $H^1(K_j)$ solution of (5.1)–(5.2) when $\mathcal{Y}_j = \mathcal{X}_j - \mathcal{X}_j^h$. Then*

$$\|\vec{e}^h\|^2_{L^2_\eta(\Gamma)} + 8k\sum_{j=1}^{N_h}\int_{K_j} n_2|\varepsilon_j|^2\,\mathrm{d}A \le 8\|(I-P^h)\vec{\mathcal{X}}\|^2_X \tag{5.4}$$

where, as before,

$$\|\vec{e}^h\|^2_{L^2_\eta(\Gamma)} = \sum_{j=1}^{N_h}\int_{\Gamma_j}\frac{1}{2\eta}\left|\mathcal{X}_j - \mathcal{X}_j^h\right|^2\,\mathrm{d}A.$$

Remark 5.3. On the one hand this estimate proves the convergence of the method on the boundary. On the other hand, it also proves convergence on any element in which $n_2 > 0$, although in a weak and difficult to interpret mesh dependent norm (by the uniqueness of the interior problem on K_j, if $\varepsilon_j = 0$ in K_j then $\mathcal{X}_j - \mathcal{X}_j^h = 0$ on ∂K_j).

Proof. The proof follows that of Lemma 9 of Section I.1.3.3.3.1 of [5] with suitable modifications. Letting $\vec{e}^h = \vec{\mathcal{X}} - \vec{\mathcal{X}}^h$ we have, using the Cauchy-Schwarz and arithmetic geometric mean inequalities,

$$\begin{aligned}((I-A)\vec{e}^h, \vec{e}^h)_X &\ge \|\vec{e}^h\|^2_X - \|\Pi\vec{e}^h\|_X\|\vec{F}\vec{e}^h\|_X \\ &\ge \|\vec{e}^h\|^2_X - \frac{1}{2}\|\Pi\vec{e}^h\|^2_X - \frac{1}{2}\|\vec{F}\vec{e}^h\|_X.\end{aligned}$$

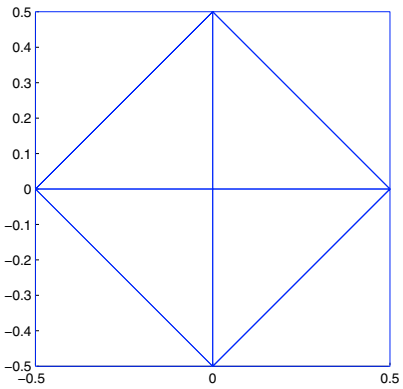
But Cessenat and Després [6] show that,

$$\|\Pi\vec{e}^h\|^2_X \le \|\vec{e}^h\|^2_X - \frac{1}{2}\|\vec{e}^h\|^2_{L^2_\eta(\Gamma)}.$$

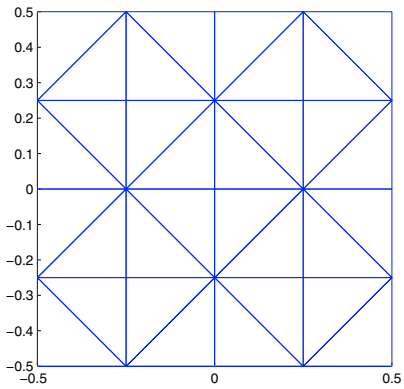
(Strictly this result is only proved when $\eta = 1$, but holds for general $\eta > 0$.) In addition Lemma 5.2 provides an estimate for $\|\vec{F}\vec{e}^h\|^2_X$ and hence

$$2((I-A)\vec{e}^h, \vec{e}^h)_X \ge \frac{1}{2}\|\vec{e}^h\|^2_{L^2_\eta(\Gamma)} + 4k\sum_{j=1}^{N_h}\int_{K_j} n_2|\varepsilon_j|^2\,\mathrm{d}A. \tag{5.5}$$

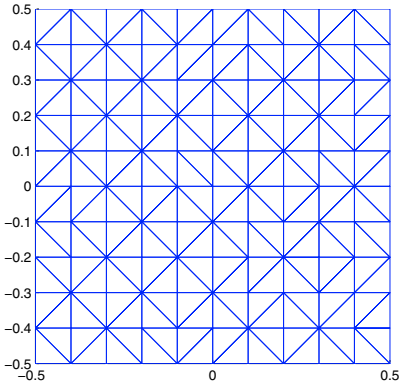
On the other hand, an extension of Lemma 3.1 to the case of absorbing media (proved using Lem. 5.1) can be used to estimate the right hand side of (5.5) and complete the proof. □



$h = 0.5$



$h = 0.25$



$h = 0.1$

FIGURE 1. Meshes used in this study. The mesh size h is the length of the vertical edge of a triangle.

6. Numerical results

In this section we assume that $n = 1$ (no absorption) and that the local plane wave basis given in (2.20) is used with the additional restriction that $p_j = p$ for each element K_j, and we consider a two dimensional problem for convenience of numerical experiments. Both Theorems 3.5 and 4.1 hold when $\Omega \subset \mathbb{R}^2$. In this case, we may use the remarkable error estimate of Theorem 3.7 of [6] that shows that if we take $p = 2\mu + 1$ and use a regular and quasi-uniform grid of triangles then

$$\|(I - P^h)\vec{\mathcal{X}}\|_X \leq Ch^{\mu-1/2}\|u\|_{C^{\mu+1}(\Omega)}.$$

We can conclude from Theorem 4.1 that in 2D

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^{\mu-1}\|u\|_{C^{\mu+1}(\Omega)}$$

where C is independent of h and u. Thus for smooth solutions of the Helmholtz equation, the UWVF can converge to high order if p is taken large enough (in particular we predict $O(h^4)$ if $\mu = 5$ or $p = 11$, $O(h^6)$ if $p = 15$ and $O(h^9)$ if $p = 21$).

TABLE 1. Results for $k = 20$ (wavelength 0.31, $p = 15$ ($\mu = 7$)). The measured convergence order for the error is roughly 7.8, whereas our theory predicts $O(h^6)$. The results of Cessenat and Després [6] predict that the condition number of D will increase $O(h^{-12})$ and this is born out by the results above.

Mesh size h	# DoF	N_{lam}	$L^2(\Omega)$ error (%)	Order	cond(D)	Order
0.50	120	4.6	4.38	-	0.64×10^2	-
0.25	480	9.2	0.01873	7.9	0.20×10^6	–11.6
0.10	3000	22.9	1.51×10^{-5}	7.8	0.22×10^{11}	–12.7

TABLE 2. Results for $k = 40$ (wavelength 0.16, $p = 21$ ($\mu = 10$)). The measured convergence order above is between 9.5 and 10.5, whereas out theory predicts $O(h^9)$. The results of Cessenat and Després [6] predict that the condition number of D will increase $O(h^{-18})$ and this is seen in practice.

Mesh size h	# DoF	N_{lam}	$L^2(\Omega)$ error (%)	Order	cond(D)	Order
0.50	168	2.0	25.2	-	8.7	-
0.25	672	4.1	0.0337	9.55	0.94×10^5	–13.4
0.1	4200	10.2	2.32×10^{-6}	10.5	0.14×10^{13}	–18

To test the predicted rate of convergence we report some numerical results due to Dr. Tomi Huttunen of the University of Kuopio in Finland. To simplify the numerical experiment as far as possible we consider the domain $\Omega = [-0.5, 0.5]^2$ and use the three grids shown in Figure 1. These grids are uniform, but tests in [6] on uniform and unstructured grids show the same convergence rate on both grids, so, at least in the $L^2(\Omega)$ norm there does not seem to be any "superconvergence" or other special convergence mechanisms. However previous tests, for example by Cessenat and Després [6] used a plane wave solution as the exact solution but this might not be representative of general solutions. So here we take the exact solution to be $u(\boldsymbol{x}) = \frac{\mathrm{i}}{4} H_0^{(1)}(k|\boldsymbol{x} - \boldsymbol{x}_0|)$ with $\boldsymbol{x}_0 = (-0.75, 0)^T$ where $H_0^{(1)}$ is the Hankel function of first kind and order zero. We choose this solution since it is singular near to the computational domain and also has curved solution contours.

In the first test case we choose $k = 20$ (so the domain is slightly over 3 wavelengths across) and $p = 15$ directions on each triangle. Results are given in Table 1. In this table, "DoF" records the total number of degrees of freedom for the problem, N_{lam} is an estimate of the number of degrees of freedom points per wavelength, and cond(D) is an estimate of the condition number of the matrix corresponding to $\langle ., . \rangle_X$. This latter quantity is often found to be a good estimate for the overall condition number of the linear system corresponding to the discrete UWVF and limits how small h can be chosen. The empirical orders computed in this table are the convergence rate of the relative $L^2(\Omega)$ error (left most "Order" column) and for the growth rate of the condition number (right most "Order" column).

For $p = 15$ we have $\mu = 7$ and so our predicted rate of convergence in the $L^2(\Omega)$ norm is $O(h^6)$. Clearly this is an over estimate of the error (or an under estimate of the convergence rate).

The second test case is for $k = 40$ and $p = 21$ and is summarized in Table 2. In this case we would expect 9th order convergence in the error, but we observe roughly 10th order.

7. CONCLUSION

We have derived error estimates that show that the UWVF converges globally. For a non-absorbing medium, the first estimate (Thm. 3.5) is relatively general but involves a mesh dependent norm. The second estimate (Thm. 4.1) requires more stringent assumptions but shows that the solution converges in the $L^2(\Omega)$ norm globally provided the best approximation error in X converges at a rate better than $O(h^{1/2})$. Numerical tests

of the $L^2(\Omega)$ error estimate show that it under-estimates the convergence rate. This maybe because we were unable to find a duality estimate for relating the estimates in Theorem 3.5 to those in Theorem 4.1. Clearly it would be highly desirable to fill this gap and we hope this paper will stimulate efforts to do this.

Two more obvious gaps exist. First no estimates are available for the error in the presence of a boundary singularity. Computational results suggest that convergence should be provable in this case [9] but serious conditioning problems can arise. Secondly there are no estimates when both h and p are refined (an hp-method).

Extensions to Maxwell's equations and higher global norms could now be considered.

Acknowledgements. We would like to thank Dr. Tomi Huttunen, University of Kuopio, Finland for supplying us with the numerical results in Section 6. The research of PM is partially supported by a grant from the US AFOSR under grant number F49620-02-1-0071.

References

[1] M. Ainsworth, P. Monk and W. Muniz, Dispersive and dissipative properties of discontinuous Galerkin methods for the wave equation. *J. Sci. Comput.* **27** (2006) 5–40.
[2] D. Arnold, F. Brezzi, B. Cockburn and L. Marini, Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.* **39** (2002) 1749–1779.
[3] A. Barnett and T. Betcke, Stability and convergence of the method of fundamental solutions for Helmholtz problems on analytic domains. *J. Comp. Phys.* (to appear).
[4] T. Betcke, A GSVD formulation of a domain decomposition method for planar eigenvalue problems. *IMA J. Numer. Anal.* **27** (2007) 451–478.
[5] O. Cessenat, *Application d'une nouvelle formulation variationnelle aux équations d'ondes harmoniques. Problèmes de Helmholtz 2D et de Maxwell 3D.* Ph.D. thesis, Université Paris IX Dauphine, France (1996).
[6] O. Cessenat and B. Després, Application of the ultra-weak variational formulation of elliptic PDEs to the 2-dimensional Helmholtz problem. *SIAM J. Numer. Anal.* **35** (1998) 255–299.
[7] O. Cessenat and B. Després, Using plane waves as base functions for solving time harmonic equations with the Ultra Weak Variational Formulation. *J. Comput. Acoustics* **11** (2003) 227–238.
[8] P. Cummings and X. Feng, Sharp regularity coefficient estimates for complex-valued acoustic and elastic Helmholtz equations. *Math. Models Methods Appl. Sci.* **16** (2006) 139–160.
[9] P. Gamallo and R. Astley, A comparison of two Trefftz-type methods: The ultraweak variational formulation and the least-squares method, for solving shortwave 2-D Helmholtz problems. *Int. J. Numer. Meth. Eng.* **71** (2007) 406–432.
[10] C. Gittelson, R. Hiptmair and I. Perugia, *Plane wave discontinuous Galerkin methods.* Preprint NI07088-HOP, Isaac Newton Institute Cambride, Cambridge, UK, December (2007) http://www.newton.cam.ac.uk/preprints/NI07088.pdf.
[11] I. Herrera, *Boundary Methods: an Algebraic Theory.* Pitman (1984).
[12] T. Huttunen, J. Kaipio and P. Monk, The perfectly matched layer for the ultra weak variational formulation of the 3D Helmholtz equation. *Int. J. Numer. Meth. Eng.* **61** (2004) 1072–1092.
[13] T. Huttunen, P. Monk and J. Kaipio, Computational aspects of the Ultra Weak Variational Formulation. *J. Comput. Phys.* **182** (2002) 27–46.
[14] T. Huttunen, P. Monk, F. Collino and J. Kaipio, The Ultra Weak Variational Formulation for elastic wave problems. *SIAM J. Sci. Comput.* **25** (2004) 1717–1742.
[15] T. Huttunen, M. Malinen and P. Monk, Solving Maxwell's equations using the Ultra Weak Variational Formulation. *J. Comput. Phys.* **223** (2007) 731–758.
[16] J. Melenk, *On generalized finite element methods.* Ph.D. thesis, University of Maryland, College Park, USA (1995).
[17] J. Melenk and I. Babuška, The partition of unity finite element method: Basic theory and applications. *Comput. Meth. Appl. Mech. Eng.* **139** (1996) 289–314.
[18] P. Monk and D. Wang, A least squares method for the Helmholtz equation. *Comput. Meth. Appl. Mech. Eng.* **175** (1999) 121–136.
[19] M. Stojek, Least-squares Trefftz-type elements for the Helmholtz equation. *Int. J. Numer. Meth. Eng.* **41** (1998) 831–849.
[20] R. Tezaur and C. Farhat, Three-dimensional discontinuous Galerkin elements with plane waves and Lagrange multipliers for the solution of mid-frequency Helmholtz problems. *Int. J. Numer. Meth. Eng.* **66** (2006) 796–815.
[21] E. Trefftz, Ein gegenstück zum Ritz'schen verfahren, in *Proc. 2nd Int. Congr. Appl. Mech.*, Zurich (1926) 131–137.