ASC Report No. 06/2012

# Generalized DG-Methods for Highly Indefinite Helmholtz Problems based on the Ultra-Weak Variational Formulation

J.M. Melenk, A. Parsania, and S. Sauter

Institute for Analysis and Scientific Computing
Vienna University of Technology — TU Wien
www.asc.tuwien.ac.at ISBN 978-3-902627-05-6

## Most recent ASC Reports

05/2012 *J.M. Melenk, H. Rezaijafari, B. Wohlmuth*
Quasi-optimal a priori estimates for fluxes in mixed finite element methods and applications to the Stokes-Darcy coupling

04/2012 *M. Langer, H. Woracek*
Indefinite Hamiltonian systems whose Titchmarsh-Weyl coefficients have no finite generalized poles of non-negativity type

03/2012 *M. Aurada, M. Feischl, J. Kemetmüller, M. Page, D. Praetorius*
Adaptive FEM with inhomogeneous dirichlet data: Convergence and quasi-optimality in Rd

02/2012 *P. Goldenits, G. Hrkac, D. Praetorius, D. Suess*
An Effective Integrator for the Landau-Lifshitz-Gilbert Equation

01/2012 *A. Arnold, L. Neumann, W. Hochhauser*
Stability of glued and embedded Glass Panes: Dunkerley straight Line as a conservative Estimate of superimposed buckling Coefficients

43/2011 *G. Hastermann, P. Lima, L. Morgado, E. Weinmüller*
Numerical Solution of the Density Profile Equation with p-Laplacians

42/2011 *M. Aurada, J.M. Melenk, D. Praetorius*
Mixed Conforming elements for the large-body limit in micromagnetics: a finite element approach

41/2011 *P. Amodio, T. Levitina, G. Settanni, E.B. Weinmüller*
On the Calculation of the Finite Hankel Transform Eigenfunctions

40/2011 *D.P. Hewett, S. Langdon, J.M. Melenk*
A high frequency hp boundary element method for scattering by convex polygons

39/2011 *A. Jüngel*
Semiconductor Device Problems

Institute for Analysis and Scientific Computing
Vienna University of Technology
Wiedner Hauptstraße 8–10
1040 Wien, Austria

**E-Mail:** admin@asc.tuwien.ac.at
**WWW:** http://www.asc.tuwien.ac.at
**FAX:** +43-1-58801-10196

ISBN 978-3-902627-05-6



ASC TU WIEN

# Generalized DG-Methods for Highly Indefinite Helmholtz Problems based on the Ultra-Weak Variational Formulation

J.M. Melenk*, A. Parsania†, and S. Sauter‡

**Abstract**

We develop a stability and convergence theory for the Ultra Weak Variational Formulation (UWVF) of a highly indefinite Helmholtz problem in $\mathbb{R}^d$, $d \in \{1,2,3\}$. The theory covers conforming as well as nonconforming generalized finite element methods. In contrast to conventional Galerkin methods where a minimal resolution condition is necessary to guarantee the unique solvability, it is proved that the UWVF admits a unique solution under much weaker conditions. As an application we present the error analysis for the *hp*-version of the finite element method explicitly in terms of the mesh width $h$, polynomial degree $p$ and wave number $k$. It is shown that the optimal convergence order estimate is obtained under the conditions that $kh/\sqrt{p}$ is sufficiently small and the polynomial degree $p$ is at least $O(\log k)$.

*AMS Subject Classifications:* 35J05, 65N12, 65N30
*Key words:* Helmholtz equation at high wavenumber, stability, convergence, discontinuous Galerkin methods, ultra-weak variational formulation, hp-finite elements

## 1 Introduction

In this paper we analyze the Discontinuous Galerkin method (DG) applied to the following model Helmholtz problem:

$$-\Delta u - k^2 u = f \quad \text{in} \quad \Omega, \tag{1.1}$$

$$\frac{\partial u}{\partial \mathbf{n}} + \mathrm{i}\, k u = g \quad \text{on} \quad \partial\Omega. \tag{1.2}$$

Here, $\Omega$ is a bounded Lipschitz domain in $\mathbb{R}^d$, $d \in \{2,3\}$, and $k$ is the real and positive wavenumber bounded away from zero, i.e., $k \geq k_0 > 0$. $\mathbf{n}$ is the outer normal vector to $\partial\Omega$ and $\mathrm{i} = \sqrt{-1}$ denotes the imaginary unit. We assume that the right-hand side $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$. By $H^s(\Omega)$ we denote the usual Sobolev space with norm $\|\cdot\|_{H^s(\Omega)}$, [1]. The seminorm which contains only the derivatives of order $s$ is denoted by $|\cdot|_{H^s(\Omega)}$.

*Institute für Analysis und Scientific Computing, Technische Universität Wien, Wiedner Hauptstrasse 8-10, A-1040 Wien, Austria; E-mail: melenk@tuwien.ac.at

†Institut für Mathematik, Universität Zürich, Zürich, Switzerland; E-mail: asieh.parsania@math.uzh.ch

‡Institut für Mathematik, Universität Zürich, Zürich, Switzerland; E-mail: stas@math.uzh.ch

The weak formulation for (1.1) is given by: Find $u \in V := H^1(\Omega)$ such that

$$a(u,v) = F(v) \qquad \forall v \in H^1(\Omega), \tag{1.3}$$

where

$$a(u,v) := \int_\Omega \left(\nabla u \nabla \bar{v} - k^2 u \bar{v}\right) + \mathrm{i}\, k \int_{\partial\Omega} u\bar{v}, \tag{1.4}$$

$$F(v) := \int_\Omega f\bar{v} + \int_{\partial\Omega} g\bar{v}. \tag{1.5}$$

Existence and uniqueness for the continuous problem have been proved in [22] for bounded Lipschitz domains.
Problems in high-frequency scattering of acoustic or electro-magnetic waves are highly indefinite – and the design of discretization methods that behave robustly with respect to the amount of indefiniteness is of great importance. For our model problem, the highly indefinite case arises for high wavenumbers $k$, and the solution $u$ is highly oscillatory. It is well known for such problems that low order finite elements suffer from the *pollution effect*, which mandates very fine meshes, [20]. For example, for a $\mathbb{P}_1$-finite element space, the number of degrees of freedom $N$ should be at least $N \gtrsim k^{2d}$, where $d$ is the spatial dimension. The conditions on the mesh size are less stringent for higher order FEM. A particular example is the analysis of [25, 26], where it has been shown in the context of high order methods that linking the polynomial degree $p$ logarithmically to the wavenumber can lead to a stable method with few degrees of freedom per wavelength. While, in particular, existence of discrete solutions is given in those circumstance, it is worth noting that a minimal resolution condition is required to ensure their existence.

This observation motivates the use of stabilized variational formulations which always guarantee the discrete stability of the method (existence and uniqueness of the discrete solution). Prominent examples of these types of methods are those incorporating least squares ideas, [2, 17, 18, 27], and Discontinuous Galerkin Methods, [12–14]. The *Ultra Weak Variational Formulation* (UWVF) of Cessenat and Després [6, 7, 9] belongs to the second class and permits using non-standard, discontinuous local discretization spaces such as plane waves (see [16, 19]).

The goal of this paper is to develop a theory for the UWVF that derives the convergence behavior of abstract conforming and non-conforming generalized finite element spaces from certain local approximation properties and local inverse estimates, which may be easy to check, possibly even at run-time.

This paper is structured as follows: In Section 2, we recall the UWVF for the Helmholtz problem (1.1). Section 3 is devoted to discrete stability and convergence. Particularly, the unified theory presented in Section 3 covers two popular choices of approximation spaces, namely, spaces consisting of piecewise plane waves and conforming as well as non-conforming *hp*-finite element spaces on affine simplicial meshes. Nevertheless, we also derive an abstract approximation criterion for general finite element spaces that implies existence and uniqueness of the discrete solution. Based on these results, we obtain quasi-optimal convergence in the DG-norm for general finite element spaces.

In Section 4 we apply the results of Section 3 to the *hp*-version of the FEM. We obtain a convergence theory that is explicit in the wavenumber $k$ as well as the discretization parameters; the mesh width $h$ and the polynomial degree $p$. These results may be viewed as an

extension of the results [25,26] for classical $H^1$-conforming discretizations to the DG-setting. In these papers, a scale resolution condition of the form

$$\frac{kh}{p} \leq c_1 \qquad \text{and} \qquad p \geq c_2 \log k \tag{1.6}$$

(for suitable $c_1$, $c_2$) is sufficient to guarantee quasi-optimality. For the *hp*-version of the DG-FEM on *regular* meshes, or, more generally, meshes that permit sufficiently rich $H^1$-conforming subspaces of the non-conforming DG-space, the same condition yields quasi-optimality. In the general case, the slightly stronger condition (4.17) is a sufficient condition for quasi-optimality. In particular, we show, for the first time for a DG-method, that quasi-optimality can be obtained for a fixed number of degrees of freedom per wavelength.

# 2 The Ultra Weak Variational Formulation

## 2.1 Meshes and Spaces

To formulate the UWVF we first introduce some notation. Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, denote a polygonal ($d=2$) or polyhedral ($d=3$) Lipschitz domain. The UWVF is based on a partition $\mathcal{T}$ of $\Omega$ into non-overlapping curvilinear polygonal/polyhedral subdomains ("finite elements") $K$ with possibly hanging nodes. The local and global mesh width is denoted by

$$h_K := \operatorname{diam} K \quad \text{and} \quad h := \max_{K \in \mathcal{T}} h_K. \tag{2.1}$$

In the case $d=3$, the boundary of $K$ can be split into *faces* and for $d=2$ into *edges*. For ease of notation we use the terminology "faces" in both cases. For $K \in \mathcal{T}$, we denote the set of faces by $\mathcal{E}(K)$. The subset of interior faces, i.e., the set of faces of $K$ which are not lying on $\partial\Omega$, is denoted by $\mathcal{E}^{\mathcal{I}}(K)$. For instance the number $\sharp\mathcal{E}(K) = d+1$ if $K$ is a simplex. As a convention we consider the finite elements $K \in \mathcal{T}$ always as open sets and the faces $e \in \mathcal{E}(K)$ as relatively open sets.

The interior skeleton $\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}}$ and the boundary skeleton $\mathfrak{S}_{\mathcal{T}}^{\mathcal{B}}$ are given by

$$\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}} := \bigcup_{K \in \mathcal{T}} \bigcup_{e \in \mathcal{E}^{\mathcal{I}}(K)} e, \qquad \mathfrak{S}_{\mathcal{T}}^{\mathcal{B}} := \bigcup_{K \in \mathcal{T}} \bigcup_{\substack{e \in \mathcal{E}(K) \\ e \subset \partial\Omega}} e.$$

Note that $\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}}$, $\mathfrak{S}_{\mathcal{T}}^{\mathcal{B}}$ are the union of the relative interior of the faces and, consequently, for any point $x \in \mathfrak{S}_{\mathcal{T}}^{\mathcal{I}}$, there exist exactly two elements in $\mathcal{T}$ (denoted by $K_x^+$, $K_x^-$) with $x \in \overline{K_x^+} \cap \overline{K_x^-}$.

Also define $\nabla_{\mathcal{T}}$ and $\Delta_{\mathcal{T}}$ as elementwise applications of the operators $\nabla$ and $\Delta$, respectively. The one-sided restrictions of some $\mathcal{T}$-piecewise smooth function $v$ for $x \in \mathfrak{S}_{\mathcal{T}}^{\mathcal{I}}$ are denoted by

$$v^+(x) := \lim_{\substack{y \in K_x^+ \\ y \to x}} v(y) \quad \text{and} \quad v^-(x) := \lim_{\substack{y \in K_x^- \\ y \to x}} v(y).$$

We use the same notation for vector-valued functions.

We define the averages and jumps for $\mathcal{T}$-piecewise smooth scalar-valued functions $v$ and vector-valued functions $\sigma_S$ on $\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}}$ by

$$\text{the averages:} \quad \{v\} := \frac{1}{2}\left(v^+ + v^-\right), \qquad \{\sigma_S\} := \frac{1}{2}\left(\sigma_S^+ + \sigma_S^-\right),$$

$$\text{the jumps:} \qquad [\![v]\!]_N := v^+\mathbf{n}^+ + v^-\mathbf{n}^-, \quad [\![\sigma_S]\!]_N := \sigma_S^+.\mathbf{n}^+ + \sigma_S^-.\mathbf{n}^-.$$

where $\mathbf{n}^+(x)$, $\mathbf{n}^-(x)$ denote the (outer) normal vectors of elements $K_x^+$, $K_x^-$.

Based on the partition $\mathcal{T}$ we can introduce *broken Sobolev spaces* in the standard way: For $s > 1$, we set

$$H^s_{\mathrm{pw}}(\Omega) := L^2(\Omega) \cap \prod_{K\in\mathcal{T}} H^s(K) \tag{2.2}$$

In particular, the case $k = 0$ corresponds to discontinuous functions.

## 2.2 Discrete Formulation

We approximate the solution of (1.3) from an *abstract* finite-dimensional space $S \subset H^2_{\mathrm{pw}}(\Omega)$, i.e., only the following two conditions are imposed:

$$S \subset L^2(\Omega) \quad \text{and} \quad S \subset \prod_{K\in\mathcal{T}} H^2(K) \tag{2.3}$$

are imposed.

We employ the formulation of the UWVF as derived in [3–6,16,19]. We assume that $\Omega$ is a bounded polygonal/polyhedral Lipschitz domain and that $\mathcal{T}$ is a shape-regular triangulation with possibly hanging nodes.

We denote by $(\cdot,\cdot)$ the $L^2$ inner product on $\Omega$, i.e., $(u,v) = \int_\Omega u\overline{v}dV$. Let $S$ be the discrete space as in (2.3). Let $\alpha \in L^\infty\left(\overline{\mathfrak{S}^{\mathcal{I}}_{\mathcal{T}}}\right)$, $\beta \in L^\infty\left(\overline{\mathfrak{S}^{\mathcal{I}}_{\mathcal{T}}}\right)$, and $\delta \in L^\infty\left(\overline{\mathfrak{S}^{\mathcal{B}}_{\mathcal{T}}}\right)$ be some positive and bounded functions on the mesh skeletons. (It will turn out that these functions can be chosen to be piecewise constant on a certain partition of the skeleton (cf. Remark 2.2).) Then, the UWVF can be written in the form:
Find $u_S \in S$ such that, for all $v \in S$,

$$a_{\mathcal{T}}(u_S,v) - k^2(u_S,v) = (f,v) - \int_{\mathfrak{S}^{\mathcal{B}}_{\mathcal{T}}} \delta\frac{1}{\mathrm{i}\,k} g\overline{\nabla_{\mathcal{T}} v\cdot\mathbf{n}}dS + \int_{\mathfrak{S}^{\mathcal{B}}_{\mathcal{T}}} (1-\delta)g\overline{v}dS =: F_{\mathcal{T}}(v) \tag{2.4}$$

where $a_{\mathcal{T}}(\cdot,\cdot)$ is the DG-bilinear form on $S\times S$ defined by

$$\begin{aligned} a_{\mathcal{T}}(u,v) := (\nabla_{\mathcal{T}}u,\nabla_{\mathcal{T}}v) &- \int_{\mathfrak{S}^{\mathcal{I}}_{\mathcal{T}}} [\![u]\!]_N\cdot\{\overline{\nabla_{\mathcal{T}}v}\}dS - \int_{\mathfrak{S}^{\mathcal{I}}_{\mathcal{T}}} \{\nabla_{\mathcal{T}}u\}\cdot[\![\overline{v}]\!]_N dS \\ &- \int_{\mathfrak{S}^{\mathcal{B}}_{\mathcal{T}}} \delta u\overline{\nabla_{\mathcal{T}}v\cdot\mathbf{n}}dS - \int_{\mathfrak{S}^{\mathcal{B}}_{\mathcal{T}}} \delta\nabla_{\mathcal{T}}u\cdot\mathbf{n}\overline{v}dS \\ &- \frac{1}{\mathrm{i}\,k}\int_{\mathfrak{S}^{\mathcal{I}}_{\mathcal{T}}} \beta[\![\nabla_{\mathcal{T}}u]\!]_N[\![\overline{\nabla_{\mathcal{T}}v}]\!]_N dS - \frac{1}{\mathrm{i}\,k}\int_{\mathfrak{S}^{\mathcal{B}}_{\mathcal{T}}} \delta\nabla_{\mathcal{T}}u\cdot\mathbf{n}\overline{\nabla_{\mathcal{T}}v\cdot\mathbf{n}}dS \\ &+ \mathrm{i}\,k\int_{\mathfrak{S}^{\mathcal{I}}_{\mathcal{T}}} \alpha[\![u]\!]_N[\![\overline{v}]\!]_N dS + \mathrm{i}\,k\int_{\mathfrak{S}^{\mathcal{B}}_{\mathcal{T}}} (1-\delta)u\overline{v}dS. \end{aligned} \tag{2.5}$$

Note that $a_{\mathcal{T}}(\cdot,\cdot)$ can be extended to a sesquilinear form on $H^{3/2+\epsilon}_{\mathrm{pw}}(\Omega)\times H^{3/2+\epsilon}_{\mathrm{pw}}(\Omega)$ for any $\epsilon > 0$. So far, the functions $\alpha$, $\beta$, $\delta$ are arbitrary, positive $L^\infty$ functions. Our analysis will rely on certain properties of $\alpha$ that depend on some trace inverse estimates for the space $S$. We therefore introduce:

**Definition 2.1** (inverse trace inequality)**.** *For each element $K$, the constant $C_{\mathrm{trace}}(S,K)$ is the smallest constant such that*

$$\|\nabla\left(v|_K\right)\|_{L^2(\partial K)} \leq C_{\mathrm{trace}}(S,K)\|\nabla v\|_{L^2(K)} \qquad \forall v \in S. \tag{2.6}$$

**Remark 2.2.** *The analysis of the continuity and coercivity will lead to the condition*

$$\alpha\left(x\right) \geq \frac{4}{3k} \max_{K\in\left\{K_x^+,K_x^-\right\}} C_{\mathrm{trace}}^2\left(S,K\right) \quad \forall x \in \mathfrak{S}_{\mathcal{T}}^{\mathcal{I}}. \tag{2.7}$$

*For the special case that $S$ is a conforming/nonconforming hp-finite element space, the estimate of the approximation property of $S$ with respect to the $\|\cdot\|_{DG}$ and $\|\cdot\|_{DG^+}$ norms, (cf. Section 4.2 ahead) leads to the choices*

$$\alpha\left(x\right) = \mathfrak{a} \max_{K\in\left\{K_x^+,K_x^-\right\}} \frac{p^2}{kh_K}, \quad \beta = \mathfrak{b}\frac{kh}{p}, \quad \delta = \quad \mathfrak{d}\frac{kh}{p}, \tag{2.8}$$

*where the parameter $\mathfrak{a}$ is selected fixed but sufficiently large below; the parameters $\mathfrak{b}$, $\mathfrak{d}$ are selected to be of size $O(1)$.*

**Remark 2.3.** *It is easy to see that $x \mapsto \alpha\left(x\right)$ can be chosen piecewise constant with respect to a sub-partition $\mathcal{E}$ of the set of all faces. More precisely, we define a subdivision of the set of inner edges by*

$$\mathcal{E}^{\mathcal{I}} := \left\{ \overset{\circ}{\partial} K \cap \overset{\circ}{\partial} K' \cap \Omega \mid \forall K \in \mathcal{T} \quad \forall K' \in \mathcal{T}\backslash\left\{K\right\} \right\},$$

*where $\overset{\circ}{\partial} K := \bigcup_{e\in\mathcal{E}(K)} e$. For any $e' \in \mathcal{E}^{\mathcal{I}}$, the maximum in (2.7) over $x \in e'$ can be chosen always as* one *fixed element $K$ so that the value of $\alpha$ is constant along $e'$. Hence, without loss of generality we may assume in the following that $\alpha$ is chosen as an $\mathcal{E}$-piecewise constant function. Note that the assumption "$\alpha$ is positive" then implies the existence of some $X \in \overset{\circ}{\partial} K \cap \Omega$ such that*

$$\alpha_{\partial K}^{\min} := \inf_{x\in\partial K} \alpha\left(x\right) = \alpha\left(X\right). \tag{2.9}$$

In the rest of this section we will show that the discretization given by the sesquilinear form $a_{\mathcal{T}}$ is consistent as well as adjoint consistent. The latter property will prove particularly useful for error estimates.

**Lemma 2.4** (consistency)**.** *Let the exact solution $u$ of (1.2) be in $H^{3/2+\epsilon}(\Omega)$ for some $\epsilon > 0$. Then $u$ satisfies the consistency condition*

$$a_{\mathcal{T}}(u,v) - k^2(u,v) = F_{\mathcal{T}}(v) \qquad \forall v \in S,$$

*where the right-hand side $F_{\mathcal{T}}$ is given in (2.4).*

*Proof.* It is enough to prove that $u$ satisfies the equation (2.4). From the $H^{3/2+\epsilon}$-regularity of $u$ it follows that $u$ and $\nabla u$ have well-defined traces on $\partial K$ for each $K \in \mathcal{T}$ and

$$[\![u]\!]_N = 0, \quad [\![\nabla u]\!]_N = 0, \quad \{\nabla u\} = \nabla u \quad \text{on} \quad \mathfrak{S}_{\mathcal{T}}^{\mathcal{I}}.$$

We multiply both sides of equation (1.1) by a test function $v \in S$, integrate elementwise and take the sum over all elements and finally apply integration by parts. We get

$$\sum_{K\in\mathcal{T}} \left( \int_{\partial K} (-\nabla u \cdot \mathbf{n})\bar{v} + \int_K \nabla u \cdot \nabla \bar{v} \right) - \int_\Omega k^2 u\bar{v} = \int_\Omega f\bar{v} \tag{2.10}$$

Using the definition of the jumps on the inner faces and inserting the boundary condition from equation (1.2), one gets

$$\begin{aligned} -\sum_{K\in\mathcal{T}} \int_{\partial K} (\nabla u \cdot \mathbf{n})\bar{v} dS = &- \int_{\mathfrak{S}_\mathcal{T}^\mathcal{B}} \delta \nabla u \cdot \mathbf{n}\overline{v} dS - \int_{\mathfrak{S}_\mathcal{T}^\mathcal{B}} (1-\delta) g\overline{v} dS \\ &+ \int_{\mathfrak{S}_\mathcal{T}^\mathcal{B}} \mathrm{i}\, k(1-\delta) u\overline{v} dS - \int_{\mathfrak{S}_\mathcal{T}^\mathcal{I}} \nabla u \cdot [\![\overline{v}]\!]_N dS. \end{aligned}$$

The boundary condition (1.2) gives us

$$\begin{aligned} = &- \int_{\mathfrak{S}_\mathcal{T}^\mathcal{B}} \delta \nabla u \cdot \mathbf{n}\overline{v} dS - \int_{\mathfrak{S}_\mathcal{T}^\mathcal{B}} (1-\delta) g\overline{v} dS + \int_{\mathfrak{S}_\mathcal{T}^\mathcal{B}} \mathrm{i}\, k(1-\delta) u\overline{v} dS - \int_{\mathfrak{S}_\mathcal{T}^\mathcal{I}} \nabla u \cdot [\![\overline{v}]\!]_N dS \\ &+ \frac{1}{\mathrm{i}\, k} \int_{\mathfrak{S}_\mathcal{T}^\mathcal{B}} \delta g \overline{\nabla_\mathcal{T} v \cdot \mathbf{n}} dS - \frac{1}{\mathrm{i}\, k} \int_{\mathfrak{S}_\mathcal{T}^\mathcal{B}} \delta \nabla u \cdot \mathbf{n} \overline{\nabla_\mathcal{T} v \cdot \mathbf{n}} dS - \int_{\mathfrak{S}_\mathcal{T}^\mathcal{B}} \delta u \overline{\nabla_\mathcal{T} v \cdot \mathbf{n}} dS. \end{aligned}$$

Inserting this result into equation (2.10) leads to

$$a_\mathcal{T}(u,v) - k^2(u,v) = (f,v) - \int_{\mathfrak{S}_\mathcal{T}^\mathcal{B}} \delta \frac{1}{\mathrm{i}\, k} g \overline{\nabla_\mathcal{T} v \cdot \mathbf{n}} dS + \int_{\mathfrak{S}_\mathcal{T}^\mathcal{B}} (1-\delta) g\overline{v} dS, \quad \forall v \in S$$

□

In Lemma 2.7 below we will consider the consistency with respect to the following adjoint problem.

**Definition 2.5** (adjoint solution operator $N_k^*$)**.** *The* adjoint Helmholtz problem *is given by:*

$$\textit{For } w \in L^2(\Omega) \textit{ find } \phi \in H^1(\Omega) \textit{ such that } a(v,\phi) = (v,w) \qquad \forall v \in H^1(\Omega). \tag{2.11}$$

*The* solution operator $N_k^* : L^2(\Omega) \to H^1(\Omega)$ *is characterized by the condition*

$$a(v, N_k^*(w)) = (v,w). \tag{2.12}$$

*Problem (2.11) has* $H^s(\Omega)$-regularity *for some* $s > 1$ *if for any given right-hand side* $w \in L^2(\Omega)$ *the solution* $\phi$ *of (2.11) is in* $H^s(\Omega)$ *and satisfies*

$$\|\phi\|_{H^s(\Omega)} \leq C_{\mathrm{reg}} \|w\|_{L^2(\Omega)}$$

*for some positive constant* $C_{\mathrm{reg}}$ *that is independent of* $w$ *and* $\phi$*.*

**Remark 2.6.** *The adjoint problem (2.11) is a well-posed problem, for which even* $k$*-explicit regularity is available. For example, if* $\Omega$ *convex (or smooth and star-shaped), then* $\phi \in H^2(\Omega)$ *and*

$$\begin{aligned} k\|\phi\|_{L^2(\Omega)} + \|\nabla\phi\|_{L^2(\Omega)} &\leq C_1(\Omega)\|w\|_{L^2(\Omega)}, \\ \|\nabla^2\phi\|_{L^2(\Omega)} &\leq C_2(\Omega)(1+k)\|w\|_{L^2(\Omega)}, \end{aligned}$$

*with $C_1(\Omega)$, $C_2(\Omega) > 0$ independent of $k \geq k_0 > 0$ ($k_0$ is arbitrary but fixed), [22, Prop. 8.1.4] for $d = 2$ and [8] for $d = 3$. For general Lipschitz domains, it was shown in [10, Thm. 2.4] that*

$$k\|\phi\|_{L^2(\Omega)} + \|\nabla\phi\|_{L^2(\Omega)} \leq C_3(\Omega)k^{5/2}\|w\|_{L^2(\Omega)}$$

*for a constant $C_3(\Omega)$ independent of $k \geq k_0$. If the Lipschitz domain $\Omega$ is polygonal/polyhedral, then classical elliptic regularity theory shows $\phi \in H^{3/2+\epsilon}(\Omega)$ for some $\epsilon > 0$, which depends on the geometry of $\Omega$.*

**Lemma 2.7** (adjoint consistency)**.** *Let the adjoint Helmholtz problem be $H^{3/2+\epsilon}(\Omega)$-regular for some $\varepsilon > 0$. Then for any $w \in L^2(\Omega)$, the solution $\phi := N_k^*$ of the adjoint problem (2.11) satisfies*

$$a_{\mathcal{T}}(v, \phi) - k^2(v, \phi) = (v, w) \qquad \forall v \in H^{3/2+\epsilon}_{\mathrm{pw}}(\Omega). \tag{2.13}$$

*Proof.* From the $H^{3/2+\epsilon}(\Omega)$-regularity of $\phi$ it follows that $\phi$ and $\nabla\phi$ have well-defined traces on $\partial K$ for each $K \in \mathcal{T}$ and

$$[\![\phi]\!]_N = 0, \quad [\![\nabla\phi]\!]_N = 0, \quad \{\nabla\phi\} = \nabla\phi \quad \text{on} \quad \mathfrak{S}_{\mathcal{T}}^{\mathcal{I}}.$$

The rest of the proof is just a repetition of the arguments in the proof of Lemma 2.4 by taking into account the zero Robin boundary conditions for the adjoint problem. □

We will use the mesh-dependent norms $\|\cdot\|_{DG}, \|\cdot\|_{DG^+}$ on $H^{3/2+\epsilon}_{\mathrm{pw}}(\Omega)$ for $\epsilon > 0$:

$$\begin{aligned}
\|v\|_{DG}^2 :=& \|\nabla_{\mathcal{T}} v\|_{L^2(\Omega)}^2 + k^{-1}\|\beta^{1/2}[\![\nabla_{\mathcal{T}} v]\!]_N\|_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}})}^2 + k\|\alpha^{1/2}[\![v]\!]_N\|_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}})}^2 \\
&+ k^{-1}\|\delta^{1/2}\nabla_{\mathcal{T}} v \cdot \mathbf{n}\|_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{B}})}^2 + k\|(1-\delta)^{1/2} v\|_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{B}})}^2 + k^2\|v\|_{L^2(\Omega)}^2, \\
\|v\|_{DG^+}^2 :=& \|v\|_{DG}^2 + k^{-1}\|\alpha^{-1/2}\{\nabla_{\mathcal{T}} v\}\|_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}})}^2.
\end{aligned}$$

# 3 Discrete Stability and Convergence Analysis

This section is devoted to the analysis of the discrete problem for the finite dimensional space $S$.

## 3.1 Continuity and Coercivity

**Proposition 3.1.** *Define $b_{\mathcal{T}}(u, v) := a_{\mathcal{T}}(u, v) + k^2(u, v)$. For any $0 < \delta < \frac{1}{3}$ and $\alpha$ satisfying (2.7), there exist constants $c_{\mathrm{coer}}$, $C_{\mathrm{c}} > 0$ independent of $h$, $k$, $\alpha$, $\beta$, $\delta$, and $C_{\mathrm{trace}}(S, K)$ such that*

**a)** *the sesquilinear form $b_{\mathcal{T}}(\cdot, \cdot)$ is coercive*

$$|b_{\mathcal{T}}(v, v)| \geq c_{\mathrm{coer}}\|v\|_{DG}^2 \qquad \forall v \in S.$$

**b)** *For any $\epsilon > 0$, the sesquilinear form $b_{\mathcal{T}}(\cdot, \cdot)$ satisfies the following continuity estimates*

$$|b_{\mathcal{T}}(v, w_S)| \leq C_{\mathrm{c}}\|v\|_{DG^+}\|w\|_{DG^+} \qquad \forall v, w \in H^{3/2+\epsilon}_{\mathrm{pw}}(\Omega), \tag{3.1}$$

$$|b_{\mathcal{T}}(v, w_S)| \leq C_{\mathrm{c}}\|v\|_{DG^+}\|w_S\|_{DG} \qquad \forall v \in H^{3/2+\epsilon}_{\mathrm{pw}}(\Omega), \quad \forall w_S \in S, \tag{3.2}$$

$$|b_{\mathcal{T}}(w_S, v)| \leq C_{\mathrm{c}}\|v\|_{DG^+}\|w_S\|_{DG} \qquad \forall v \in H^{3/2+\epsilon}_{\mathrm{pw}}(\Omega), \quad \forall w_S \in S. \tag{3.3}$$

*Proof.* a) The definition of $b_{\mathcal{T}}(.,.)$ leads to

$$\begin{aligned} b_{\mathcal{T}}(v,v) = \|\nabla_{\mathcal{T}} v\|^2_{L^2(\Omega)} - 2\,\mathrm{Re}\left(\int_{\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}}} [\![v]\!]_N \cdot \{\overline{\nabla_{\mathcal{T}} v}\} dS\right) - 2\,\mathrm{Re}\left(\int_{\mathfrak{S}_{\mathcal{T}}^{\mathcal{B}}} \delta v \overline{\nabla_{\mathcal{T}} v \cdot \mathbf{n}} dS\right) \\ + \mathrm{i}\, k^{-1} \|\beta^{1/2} [\![\nabla_{\mathcal{T}} v]\!]_N\|^2_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}})} + \mathrm{i}\, k^{-1} \|\delta^{1/2} \nabla_{\mathcal{T}} v \cdot \mathbf{n}\|^2_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{B}})} \\ + \mathrm{i}\, k \|\alpha^{1/2} [\![v]\!]_N\|^2_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}})} + \mathrm{i}\, k \|(1-\delta)^{1/2} v\|^2_{0,\mathfrak{S}_{\mathcal{T}}^{\mathcal{B}}} + k^2 \|v\|^2_{L^2(\Omega)}. \end{aligned}$$

By using Young's inequality for some positive function $s \in L^\infty\left(\overline{\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}}}\right)$ we get for the second term in the representation of $b_{\mathcal{T}}(\cdot,\cdot)$

$$\begin{aligned} \left|2\,\mathrm{Re}\int_{\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}}} [\![v]\!]_N \cdot \{\overline{\nabla_{\mathcal{T}} v}\} dS\right| &\le k \|\sqrt{\frac{s}{\alpha}} \alpha^{1/2} [\![v]\!]_N\|^2_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}})} \\ &\quad + \frac{1}{k} \|\frac{1}{\sqrt{s}} \nabla\left(v|_K\right)\|^2_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}})}. \end{aligned}$$

We choose $s := 4\alpha/5$. By using (2.7) we get

$$\begin{aligned} \left|2\,\mathrm{Re}\int_{\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}}} [\![v]\!]_N \cdot \{\overline{\nabla_{\mathcal{T}} v}\} dS\right| &\le \frac{4}{5} k \|\alpha^{1/2} [\![v]\!]_N\|^2_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}})} \\ &\quad + \sum_{K\in\mathcal{T}} \frac{5}{4k} \|\frac{1}{\alpha^{1/2}} \nabla\left(v|_K\right)\|^2_{L^2(\Omega\cap\partial K)}. \end{aligned}$$

For the second summand, we get with $\alpha_{\partial K}^{\min}$ as in (2.9)

$$\sum_{K\in\mathcal{T}} \frac{5}{4k} \|\frac{1}{\alpha^{1/2}} \nabla\left(v|_K\right)\|^2_{L^2(\Omega\cap\partial K)} \le \sum_{K\in\mathcal{T}} \frac{5}{4k} \frac{C^2_{\mathrm{trace}}(S,K)}{\alpha_{\partial K}^{\min}} \|\nabla v\|^2_{L^2(K)}.$$

Let $X \in \overset{\circ}{\partial K} \cap \Omega$ be defined as in Remark 2.3. Since $K \in \left\{K_X^+, K_X^-\right\}$, the condition on $\alpha$ (cf. (2.6)) implies

$$\alpha_{\partial K}^{\min} = \alpha(X) \ge \frac{4}{3k} \max_{K' \in \left\{K_X^+, K_X^-\right\}} C^2_{\mathrm{trace}}(S,K') \ge \frac{4}{3k} C^2_{\mathrm{trace}}(S,K). \tag{3.4}$$

Hence,

$$\sum_{K\in\mathcal{T}} \frac{5}{4k} \|\frac{1}{\alpha^{1/2}} \nabla\left(v|_K\right)\|^2_{L^2(\Omega\cap\partial K)} \le \frac{15}{16} \|\nabla_{\mathcal{T}} v\|^2_{L^2(\Omega)}.$$

All in all we have derived

$$\left|2\,\mathrm{Re}\int_{\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}}} [\![v]\!]_N \cdot \{\overline{\nabla_{\mathcal{T}} v}\} dS\right| \le \frac{4k}{5} \|\alpha^{1/2} [\![v]\!]_N\|^2_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}})} + \frac{15}{16} \|\nabla_{\mathcal{T}} v\|^2_{L^2(\Omega)}.$$

The third term in $b_{\mathcal{T}}(\cdot,\cdot)$ can be estimated in a similar fashion for any $t > 0$ by

$$\left|2\,\mathrm{Re}\int_{\mathfrak{S}_{\mathcal{T}}^{\mathcal{B}}} \delta v \overline{\nabla_{\mathcal{T}} v} \cdot \mathbf{n} dS\right| \le tk \frac{\delta}{1-\delta} \|(1-\delta)^{1/2} v\|^2_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{B}})} + \frac{1}{tk} \|\delta^{1/2} \nabla_{\mathcal{T}} v \cdot \mathbf{n}\|^2_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{B}})}.$$

By choosing $t = 3/2$ as well as $0 < \delta < \frac{1}{3}$ we obtain

$$\begin{aligned}
|b_{\mathcal{T}}(v,v)| &\geq \frac{1}{\sqrt{2}}\left(|\mathrm{Re}(b_{\mathcal{T}}(v,v))| + |\mathrm{Im}(b_{\mathcal{T}}(v,v))|\right) \\
&\geq \frac{1}{\sqrt{2}}\left(\frac{1}{16}\|\nabla_{\mathcal{T}} v\|^2_{L^2(\Omega)} + \frac{k}{5}\|\alpha^{1/2}[\![v]\!]_N\|^2_{L^2(\mathfrak{S}^{\mathcal{I}}_{\mathcal{T}})}\right. \\
&+ \frac{k}{4}\|(1-\delta)^{1/2} v\|^2_{L^2(\mathfrak{S}^{\mathcal{B}}_{\mathcal{T}})} + \frac{1}{3k}\|\delta^{1/2}\nabla_{\mathcal{T}} v\cdot\mathbf{n}\|^2_{0,\mathfrak{S}^{\mathcal{B}}_{\mathcal{T}}} \\
&\left.+k^{-1}\|\beta^{1/2}[\![\nabla_{\mathcal{T}} v]\!]_N\|^2_{L^2(\mathfrak{S}^{\mathcal{I}}_{\mathcal{T}})} + k^2\|v\|^2_{L^2(\Omega)}\right) \\
&\geq c_{\mathrm{coer}}\|v\|^2_{DG}.
\end{aligned}$$

b) Using Young's inequality we get

$$\begin{aligned}
|b_{\mathcal{T}}(v,w)| \leq\ & |(\nabla_{\mathcal{T}} v, \nabla_{\mathcal{T}} w)| + |\int_{\mathfrak{S}^{\mathcal{I}}_{\mathcal{T}}} [\![v]\!]_N \cdot \{\overline{\nabla_{\mathcal{T}} w}\} dS| \\
&+ |\int_{\mathfrak{S}^{\mathcal{I}}_{\mathcal{T}}} \{\nabla_{\mathcal{T}} v\}\cdot[\![\overline{w}]\!]_N dS| + |\int_{\mathfrak{S}^{\mathcal{B}}_{\mathcal{T}}} \delta v\overline{\nabla_{\mathcal{T}} w\cdot\mathbf{n}} dS| \\
&+ |\int_{\mathfrak{S}^{\mathcal{B}}_{\mathcal{T}}} \delta\nabla_{\mathcal{T}} v\cdot\mathbf{n}\overline{w} dS| + \frac{1}{k}|\int_{\mathfrak{S}^{\mathcal{I}}_{\mathcal{T}}} \left(\beta[\![\nabla_{\mathcal{T}} v]\!]_N[\![\overline{\nabla_{\mathcal{T}} w}]\!]_N\right) dS| \\
&+ \frac{1}{k}|\int_{\mathfrak{S}^{\mathcal{B}}_{\mathcal{T}}} \left(\delta\nabla_{\mathcal{T}} v\cdot\mathbf{n}\overline{\nabla_{\mathcal{T}} w\cdot\mathbf{n}}\right) dS| + |\int_{\mathfrak{S}^{\mathcal{I}}_{\mathcal{T}}} \left(k\alpha[\![v]\!]_N[\![\overline{w}]\!]_N\right) dS| \\
&+ k|\int_{\mathfrak{S}^{\mathcal{B}}_{\mathcal{T}}} (1-\delta) v\overline{w} dS| + k^2|(v,w)|.
\end{aligned} \tag{3.5}$$

For $0 < \delta < 1/2$ and for any $v$, $w \in H^{3/2+\epsilon}_{\mathrm{pw}}(\Omega)$ we finally obtain

$$|b_{\mathcal{T}}(v,w)| \leq C_{\mathrm{c}}\|v\|_{DG^+}\|w\|_{DG^+}.$$

Estimates in weaker norms are possible if one of these two functions is purely a finite element function, e.g., $w \in S$. A careful inspection of equation (3.5) shows that the only term which requires the $DG^+$-norm instead of $DG$-norm for $w$ in the continuity estimate is $\int_{\mathfrak{S}^{\mathcal{I}}_{\mathcal{T}}} [\![v]\!]_N \cdot \{\overline{\nabla_{\mathcal{T}} w}\} dS$. Using Young's inequality we get

$$\left|\int_{\mathfrak{S}^{\mathcal{I}}_{\mathcal{T}}} [\![v]\!]_N \cdot \{\overline{\nabla_{\mathcal{T}} w}\} dS\right| \leq \sum_{K\in\mathcal{T}} \left\{\|[\![v]\!]_N\|_{L^2(\Omega\cap\partial K)}\|\nabla\left(w|_K\right)\|_{L^2(\Omega\cap\partial K)}\right\}.$$

We apply the trace inequality in (2.6) and also (2.7) to obtain

$$\begin{aligned}
\left|\int_{\mathfrak{S}^{\mathcal{I}}_{\mathcal{T}}} [\![v]\!]_N \cdot \{\overline{\nabla_{\mathcal{T}} w}\} dS\right| &\leq \sum_{K\in\mathcal{T}} \left\{\frac{1}{\sqrt{\alpha^{\min}_{\partial K}}}\left\|\alpha^{\frac{1}{2}}[\![v]\!]_N\right\|_{L^2(\Omega\cap\partial K)} C_{\mathrm{trace}}(S,K)\|\nabla_{\mathcal{T}} w\|_{L^2(K)}\right\} \\
&\overset{(3.4)}{\leq} \sqrt{\frac{3k}{4}}\sum_{K\in\mathcal{T}}\left\{\left\|\alpha^{\frac{1}{2}}[\![v]\!]_N\right\|_{L^2(\Omega\cap\partial K)}\|\nabla_{\mathcal{T}} w\|_{L^2(K)}\right\} \\
&\leq \sqrt{\frac{3k}{2}}\left\|\alpha^{\frac{1}{2}}[\![v]\!]_N\right\|_{L^2(\mathfrak{S}^{\mathcal{I}}_{\mathcal{T}})}\|\nabla_{\mathcal{T}} w\|_{L^2(\Omega)}.
\end{aligned}$$

Hence, we finally obtain(3.2). The estimate (3.3) can be shown using the same techniques or derived from (3.2) by observing that for $v$, $w \in H^{3/2+\varepsilon}_{\mathrm{pw}}(\Omega)$ we have

$$b_{\mathcal{T},k}(v,w) = \overline{b_{\mathcal{T},-k}(w,v)}$$

where we have added the subscript $k$ (or $-k$) to emphasize how the parameter $k$ enters the definition. □

As a corollary of (3.3) we have the following continuity assertion, which is particularly suitable for adjoint problems:

**Corollary 3.2.** *For any $\epsilon > 0$, it holds*

$$|a_{\mathcal{T}}(v,u) - k^2(v,u)| \leq C_{\mathrm{c}}\|u\|_{DG+}\|v\|_{DG} \qquad \forall u \in H^{3/2+\epsilon}_{\mathrm{pw}}(\Omega) \quad \forall v \in S. \tag{3.6}$$

## 3.2 Quasi-Optimality

We start with a definition: We say that a pair $(u, u_S) \in H^{3/2+\varepsilon}_{\mathrm{pw}}(\Omega) \times S$ of functions satisfies the Galerkin orthogonality if

$$a_{\mathcal{T}}(u - u_S, v) = 0 \qquad \forall v \in S. \tag{3.7}$$

Our starting point for the analysis of the UVWF is a quasi-optimality result which is proved under the *assumption* that the above Galerkin orthogonality is valid. The existence and uniqueness of a solution $u_S$ of the discrete problem (2.4) is then shown in a second step based on the quasi-optimality result.

**Proposition 3.3.** *There exists a constant $\widetilde{C} > 0$ depending solely on the constants $C_c$, $c_{\mathrm{coer}}$ of Proposition 3.1 such that the following is true: Any pair $(u, u_S) \in H^{3/2+\varepsilon}_{\mathrm{pw}}(\Omega) \times S$ meeting the orthogonality condition (3.7) satisfies*

$$\|u - u_S\|_{DG} \leq \widetilde{C}\left(\inf_{v\in S}\|u - v\|_{DG^+} + \sup_{0\neq w_S\in S}\frac{k|(u-u_S,w_S)|}{\|w_S\|_{L^2(\Omega)}}\right).$$

*Proof.* We start with a triangle inequality

$$\|u - u_S\|_{DG} \leq \|u - v\|_{DG} + \|v - u_S\|_{DG} \qquad \forall v \in S \tag{3.8}$$

and employ the coercivity of $b_{\mathcal{T}}(\cdot,\cdot)$

$$\begin{aligned}
\|v - u_S\|^2_{DG} &\leq \frac{1}{c_{\mathrm{coer}}}|b_{\mathcal{T}}(v - u_S, v - u_S)| \\
&\leq \frac{1}{c_{\mathrm{coer}}}|b_{\mathcal{T}}(v - u, v - u_S)| + \frac{1}{c_{\mathrm{coer}}}|b_{\mathcal{T}}(u - u_S, v - u_S)| \\
&= \frac{1}{c_{\mathrm{coer}}}|b_{\mathcal{T}}(v - u, v - u_S)| + \frac{2k^2}{c_{\mathrm{coer}}}|(u - u_S, v - u_S)|,
\end{aligned} \tag{3.9}$$

where in the last inequality we employed the orthogonality condition (3.7). The continuity of $b_{\mathcal{T}}(\cdot,\cdot)$ expressed in (3.1) together with (3.9) implies

$$\|v - u_S\|^2_{DG} \leq \frac{C_{\mathrm{c}}}{c_{\mathrm{coer}}}\|v - u\|_{DG^+}\|v - u_S\|_{DG} + \frac{2k^2}{c_{\mathrm{coer}}}|(u - u_S, v - u_S)|.$$

We combine this result with (3.8) and obtain

$$\|u-u_S\|_{DG} \leq \|u-v\|_{DG} + \frac{C_{\mathrm{c}}}{c_{\mathrm{coer}}}\|v-u\|_{DG^+} + \frac{2k}{c_{\mathrm{coer}}} \sup_{0\neq w_S\in S} \frac{|(u-u_S,w_S)|}{\|w_S\|_{L^2(\Omega)}}.$$

□

Next, we will use the adjoint problem to gauge the contribution $\sup_{w_S\in S} \frac{k|(u-u_S,w_S)|}{\|w_S\|_{L^2(\Omega)}}$ in Proposition 3.3.

**Proposition 3.4.** *Assume that the adjoint Helmholtz problem is $H^{3/2+\varepsilon}(\Omega)$ regular for some $\varepsilon > 0$. Let the coefficients in the definition of $a_{\mathcal{T}}(\cdot,\cdot)$ satisfy $0 < \delta < \frac{1}{3}$ and (2.7). Then the following is true: For any pair $(u,u_S) \in H^{3/2+\varepsilon}_{\mathrm{pw}}(\Omega) \times S$ that satisfies (3.7) we have*

$$\sup_{0\neq w_S\in S} \frac{k|(u-u_S,w_S)_{L^2(\Omega)}|}{\|w_S\|_{L^2(\Omega)}} \leq (1+3C_{\mathrm{c}})\,\eta_k(S)\left(\inf_{v\in S}\|u-v\|_{DG^+} + \|u-u_S\|_{DG}\right),$$

*where the* adjoint approximation property *is defined by*

$$\eta_k(S) := \sup_{f\in L^2(\Omega)\setminus\{0\}} \inf_{\psi_S\in S} \frac{k\|N_k^* f - \psi_S\|_{DG^+}}{\|f\|_{L^2(\Omega)}}. \tag{3.10}$$

*Proof.* The solution of the adjoint problem (2.12) with right-hand side $w_S \in S \subset L^2(\Omega)$ is denoted by $\phi$. Our regularity assumption imply $\phi \in H^{3/2+\epsilon}(\Omega)$ for some $\epsilon > 0$ (cf. Remark 2.6). The adjoint consistency of the method stated in Lemma 2.7 then provides

$$(u-u_S,w_S) = a_{\mathcal{T}}(u-u_S,\phi) - k^2(u-u_S,\phi).$$

Using the definition of the sesquilinear form $a_{\mathcal{T}}$ and the Galerkin orthogonality, we get for any $v \in S$

$$\begin{aligned}
|(u-u_S,w_S)| &\leq |a_{\mathcal{T}}(u-v,\phi-\psi_S)| + |a_{\mathcal{T}}(v-u_S,\phi-\psi_S)| \\
&\quad + k^2|(u-u_S,\phi-\psi_S)| \\
&\leq (C_{\mathrm{c}}\|u-v\|_{DG^+} + C_{\mathrm{c}}\,\|v-u_S\|_{DG} \\
&\quad + \|u-u_S\|_{DG})\,\|\phi-\psi_S\|_{DG^+} \\
&\leq (2C_{\mathrm{c}}\|u-v\|_{DG^+} + (1+C_{\mathrm{c}})\|u-u_S\|_{DG})\,\|\phi-\psi_S\|_{DG^+}.
\end{aligned}$$

Since $v$, $\psi_S \in S$ are arbitrary, the statement follows. □

The combination of the previous results leads to the following wavenumber-explicit error estimate (still under the assumption of existence of a discrete solution).

**Theorem 3.5** (quasi-optimal convergence)**.** *Assume that the adjoint Helmholtz problem is $H^{3/2+\varepsilon}(\Omega)$ regular for some $\varepsilon > 0$. Let the coefficients in the definition of $a_{\mathcal{T}}(\cdot,\cdot)$ satisfy $0 < \delta < \frac{1}{3}$ and (2.7). If the condition*

$$\eta_k(S) < \frac{c_{\mathrm{coer}}}{4(1+C_c)}$$

*holds, then for any pair $(u,u_S) \in H^{3/2+\varepsilon}_{\mathrm{pw}}(\Omega) \times S$ that satisfies (3.7) we have*

$$\|u-u_S\|_{DG} \leq C \inf_{v\in S}\|u-v\|_{DG^+}, \tag{3.11}$$

*where $C$ depends solely on $C_c$ and $c_{\mathrm{coer}}$.*

*Proof.* By combining the results of Propositions 3.3 and 3.4, we get the following:

$$\|u-u_S\|_{DG} \leq \left(1+\frac{C_c}{c_{\text{coer}}}+\frac{4C_c}{c_{\text{coer}}}\eta_k(S)\right)\inf_{v\in S}\|u-v\|_{DG^+}+\frac{2(1+C_c)}{c_{\text{coer}}}\eta_k(S)\|u-u_S\|_{DG}.$$

The condition $\frac{2(1+C_c)}{c_{\text{coer}}}\eta_k(S) < 1/2$ allows us to absorb the error term on the right-hand side in the left-hand side. □

## 3.3 Discrete Stability

The preceding section provides an error analysis under the assumption of existence of the discrete solution $u_S \in S$ of (2.4). Extra conditions have to be imposed for existence as the following Example 3.6 shows. That is, the UWVF of the Helmholtz problem is not necessarily stable for an arbitrary discrete space $S$ that only satisfies the minimal condition (2.3).

**Example 3.6.** *Let* $\Omega := \operatorname{conv}\{(0,0)^\intercal,\ (1,0)^\intercal,\ (0,1)^\intercal\}$ *and let the mesh* $\mathcal{T}$ *consists of the single element* $\{\Omega\}$*. A (one-dimensional) space* $S$ *that satisfies condition (2.3) is defined by the span of the squared cubic bubble function,* $S = \operatorname{span}\{(27\lambda_1\lambda_2\lambda_3)^2\}$*, where* $\lambda_1 = \xi_1$*,* $\lambda_2 = \xi_2$*,* $\lambda_3 = 1-\xi_1-\xi_2$ *and* $0 \leq \xi_1 \leq 1$*,* $0 \leq \xi_2 \leq 1-\xi_1$*. In this case, equation (3.15) reduces to*

$$(\nabla w_S, \nabla v_S) - k^2(w_S, v_S) = 0 \qquad \forall v_S \in S. \tag{3.12}$$

*As* $S$ *is a one-dimensional space we get the following* $1\times 1$ *system* $(A-k^2B)w = 0$*, where* $A = \int_{\widehat{K}} \nabla b_1 \cdot \nabla b_1 = 5.1125$*,* $B = \int_{\widehat{K}} b_1^2 = 0.0843$ *and* $b_1 = (27\lambda_1\lambda_2\lambda_3)^2$*. Obviously, the value of* $k = \sqrt{\frac{A}{B}}$ *is a critical wavenumber where the system matrix becomes singular.*

In this section, we will study conditions under which the UWVF admits a unique solution in the discrete space $S$. One possible condition (3.13) is formulated in Theorem 3.7 and it is shown that this condition is always satisfied for plane waves methods as well as for conforming and non-conforming *hp*-finite element spaces on affine simplicial meshes (cf. Remark 3.8). Thus, Theorem 3.7 presents a unified stability theory for these types of methods and shows that a unique numerical solution always exists for these important choices of spaces. This is in contrast to conventional Galerkin methods applied to (1.3), where a minimal resolution condition on the finite element space, e.g., on the mesh width, has to be imposed in order to guarantee unique solvability of the discrete equations.

Alternatively, similarly to the classical Galerkin discretization, a condition on the adjoint approximation property on the abstract space can be employed to prove existence, uniqueness, and quasi-optimality of the discretization. This is proved in Theorem 3.9.

**Theorem 3.7.** *Let the discrete space* $S$ *satisfy (2.3). Let* $\beta \geq 0$*,* $0 < \delta < \frac{1}{3}$*, and choose* $\alpha$ *such that (2.7) is satisfied. Then, the UWVF problem (2.4) has a unique solution* $u_S \in S$ *if*

$$C_S < \frac{k}{2\,(1+C_c)} \quad \textit{with} \quad C_S := \sup_{w_S\in S\cap H^2_0(\Omega)\setminus\{0\}} \inf_{v_S\in S} \frac{\|\,\langle x, \nabla w_S\rangle - v_S\|_{DG^+}}{\|w_S\|_{L^2(\Omega)}}. \tag{3.13}$$

*Furthermore, let the exact solution of (1.3) satisfy* $u \in H^{3/2+\epsilon}(\Omega)$*, and let the adjoint Helmholtz problem be* $H^{3/2+\epsilon}(\Omega)$ *regular for some* $\epsilon > 0$*. Assume the adjoint approximation condition*

$$\eta_k(S) < \frac{c_{\text{coer}}}{4(1+C_c)}.$$

*Then, the quasi-optimal error estimate*

$$\|u - u_S\|_{DG} \leq C \inf_{v \in S} \|u - v\|_{DG^+}$$

*holds, where $C$ is independent of the choice of $k$, $h$, and the space $S$.*

*Proof.* If the discrete solution $u_S \in S$ of (2.4) exists, then the consistency statement Lemma 2.4 implies the orthogonality condition (3.7) so that the quasi-optimality assertion follows from Theorem 3.5. It therefore remains to assert existence of $u_S \in S$. By dimension arguments, existence of a solution $u_S \in S$ of (2.4) follows, if we can verify the following uniqueness assertion:

$$\forall w_S \in S \setminus \{0\} \quad \exists v_S \in S \quad \text{s.t.} \quad |a_{\mathcal{T}}(w_S, v_S) - k^2(w_S, v_S)| > 0. \tag{3.14}$$

We prove (3.14) indirectly, by showing the equivalent implication:
For any $w_S \in S$ it holds:

$$\big(\forall v_S \in S \quad a_{\mathcal{T}}(w_S, v_S) - k^2(w_S, v_S) = 0\big) \Rightarrow w_S = 0. \tag{3.15}$$

Our assumption in (3.15) implies for any $w_S \in S$

$$\operatorname{Im}\big(a_{\mathcal{T}}(w_S, v_S) - k^2(w_S, v_S)\big) = 0 \quad \text{and} \quad \operatorname{Re}\big(a_{\mathcal{T}}(w_S, v_S) - k^2(w_S, v_S)\big) = 0. \tag{3.16}$$

First we choose $v_S = w_S$ in (3.16). From the equation for the imaginary part we obtain

**(1)** $[\![\nabla_{\mathcal{T}} w_S]\!]_N = 0 \quad \text{on} \quad \mathfrak{S}_{\mathcal{T}}^{\mathcal{I}}$,

**(2)** $\nabla_{\mathcal{T}} w_S \cdot \mathbf{n} = 0 \quad \text{on} \quad \mathfrak{S}_{\mathcal{T}}^{\mathcal{B}}$,

**(3)** $[\![w_S]\!]_N = 0 \quad \text{on} \quad \mathfrak{S}_{\mathcal{T}}^{\mathcal{I}}$,

**(4)** $w_S = 0 \quad \text{on} \quad \mathfrak{S}_{\mathcal{T}}^{\mathcal{B}}$

and this implies $w_S \in H_0^2(\Omega) \cap S$ (in particular that $\nabla_{\mathcal{T}} w_S = \nabla w_S$ holds).
Hence, the real part of equation (3.16) gives us

$$\|\nabla w_S\|_{L^2(\Omega)}^2 - k^2 \|w_S\|_{L^2(\Omega)}^2 = 0. \tag{3.17}$$

Define $v_S^*(x) = \langle x, \nabla w_S \rangle$. From the real part of equation (3.16) it follows

$$\begin{aligned} 0 &= \operatorname{Re}\big(a_{\mathcal{T}}(w_S, v_S^*) - k^2(w_S, v_S^*)\big) + \operatorname{Re}\big(a_{\mathcal{T}}(w_S, v_S - v_S^*) - k^2(w_S, v_S - v_S^*)\big) \\ &\geq \operatorname{Re}\big(a_{\mathcal{T}}(w_S, v_S^*) - k^2(w_S, v_S^*)\big) - |a_{\mathcal{T}}(w_S, v_S^* - v_S)| - |k^2(w_S, v_S^* - v_S)|. \end{aligned}$$

By using $2\operatorname{Re}(w_S \nabla \overline{w_S}) = \nabla(|w_S|^2)$ for the first term, and continuity of $a_{\mathcal{T}}$, and applying Cauchy-Schwarz inequality we get (see also [22], [12])

$$\begin{aligned} 0 \geq& (2-d)\|\nabla w_S\|_{L^2(\Omega)}^2 + dk^2\|w_S\|_{L^2(\Omega)}^2 - 2C_c\|w_S\|_{DG}\|v_S^* - v_S\|_{DG^+} \\ &- 2k^2\|w_S\|_{L^2(\Omega)}\|v_S^* - v_S\|_{L^2(\Omega)} \\ \geq& (2-d)\|\nabla w_S\|_{L^2(\Omega)}^2 + dk^2\|w_S\|_{L^2(\Omega)}^2 - 2C_c C_S\|w_S\|_{DG}\|w_S\|_{L^2(\Omega)} \\ &- 2C_S k\|w_S\|_{L^2(\Omega)}^2. \end{aligned} \tag{3.18}$$

By using the definition of DG-norm and taking into account that $w_S \in H_0^2(\Omega) \cap S$ it follows $\|w_S\|_{DG} = \|w_S\|_{\mathcal{H}}$, where $\|w_S\|_{\mathcal{H}}^2 := \|\nabla w_S\|_{L^2(\Omega)}^2 + k^2 \|w_S\|_{L^2(\Omega)}^2$. For $d = 1$, we get

$$\begin{aligned} 0 &\geq \|w_S\|_{\mathcal{H}}^2 - 2C_c C_S \|w_S\|_{\mathcal{H}} \|w_S\|_{L^2(\Omega)} - 2C_S k \|w_S\|_{L^2(\Omega)}^2 \\ &\geq \left(1 - \frac{2C_c C_S}{k} - \frac{2C_S}{k}\right) \|w_S\|_{\mathcal{H}}^2. \end{aligned}$$

If $C_S < \frac{k}{2(1+C_c)}$ then it follows $w_S = 0 \quad \text{in} \quad \Omega$. For $d = 2, 3$ we add (3.17) to the equation (3.18) and then we do the same argument as in 1d. □

**Remark 3.8.** *For general finite-dimensional spaces $S$, condition (3.13) could be interpreted as a condition on the scale resolution. However, the condition (3.13) is always satisfied in the following two important cases:*

- *Typically in the UWVF, the discrete space $S$ consists of systems of (discontinuous) plane waves. In that setting, condition (3.13) is not imposed since it is trivially satisfied as then $S \cap H_0^2(\Omega) = \{0\}$ (this equality follows from the unique continuation principle for elliptic PDEs—see, e.g., the discussion in [10, Sec. 6.3] for details).*
- *DG-methods based on classical piecewise polynomials on affine triangulations (consisting of simplices) satisfy (3.13) automatically as $\langle x, \nabla_{\mathcal{T}} w_S \rangle \in S$.*

For new generalized finite element spaces, it might be complicated to verify condition (3.13). In the following theorem, we present a different criterion which also implies discrete stability.

**Theorem 3.9.** *Let the exact solution of (1.3) satisfy $u \in H^{3/2+\epsilon}(\Omega)$ and let the adjoint Helmholtz problem be $H^{3/2+\epsilon}(\Omega)$ regular for some $\epsilon > 0$. Assume that the coefficients in the definition of $a_{\mathcal{T}}(\cdot, \cdot)$ satisfy $0 < \delta < \frac{1}{3}$ and (2.7). If the condition*

$$\eta_k(S) < \frac{c_{\text{coer}}}{4(1 + C_c)}$$

*holds then the UWVF problem (2.4) has a unique solution $u_S \in S$ and satisfies the quasi-optimality property (3.11).*

*Proof.* The proof follows the lines in [21, Thm. 3.9]. We merely have to show existence of $u_S$. Since the (2.4) corresponds to a linear system of equations, it suffices to show uniqueness. Therefore, let $u_S \in S$ be in the kernel of the discrete operator, i.e., $a_{\mathcal{T}}(u_S, v) - k^2 (u_S, v) = 0$ for all $v \in S$. Then the pair $(0, u_S) \in H^{3/2+\varepsilon}(\Omega) \times S$ satisfies the orthogonality condition (3.7). Hence, Theorem 3.5 implies $\|0 - u_S\|_{DG} \leq C \inf_{v \in S} \|0 - v\|_{DG+} = 0$, which shows $u_S = 0$.

Again, the quasi-optimality follows as a combination of Theorem 3.5 and Lemma 2.4. □

# 4 Application to *hp*-Finite Elements

Theorem 3.5 provides a quasi-optimal error estimate for abstract approximation spaces $S$ that satisfy the conditions (2.3) and (3.13). The concrete choice of the space $S$ enters the analysis via a) the constant $C_{\text{trace}}(S, K)$, b) the estimate of the approximation error $\inf_{v \in S} \|u - v\|_{DG^+}$, c) the adjoint approximation property $\eta_k(S)$, and d) the constant $C_S$ in (3.13) – however,

as explained in Remark 3.8 the condition on $C_S$ is "automatically" satisfied for the *hp*-finite element spaces considered in this section. Further, we derive explicit estimates for these quantities in the context of *hp*-finite element space which are explicit with respect to the polynomial degree $p$ and the mesh size $h$.

## 4.1 Preliminaries

The simplicial finite element mesh $\mathcal{T}$ consists of elements $K$ which are the images of the reference element $\widehat{K}$, i.e., the reference triangle (in 2D) or the reference tetrahedron (in 3D), under the element map $F_K : \widehat{K} \to K$. The mesh width is denoted by $h := \max_{K\in\mathcal{T}} \operatorname{diam} K$ (cf. (2.1)).

We use the symbol $\nabla^n$ to denote derivatives of order $n$; more precisely, for a function $u : \Omega \to \mathbb{R}, \Omega \subset \mathbb{R}^d$, we set

$$|\nabla^n u(x)|^2 = \sum_{\alpha\in\mathbb{N}_0^d:\,|\alpha|=n} \frac{n!}{\alpha!} |D^\alpha u(x)|^2.$$

We will need some conditions on the element maps $F_K$ of the triangulations in order to capture the approximation properties of the *hp*-FEM spaces. The following assumption will make this more precise. We emphasize that, in contrast to the case of conforming subspaces, we do not require in the present context of DG-methods a "compatibility" condition for element maps of neighboring elements.

**Assumption 4.1.** *(simplicial finite element mesh). Each element map $F_K$ can be written as $F_K = R_K \circ B_K$, where $B_K$ is an affine map (containing the scaling by $h_K$) and $R_K$ is analytic. Let $\widetilde{K} := B_K(K)$. The maps $R_K$ and $B_K$ satisfy for* shape regularity *constants $C_{\text{affine}}, C_{\text{metric}}, \gamma > 0$ independent of $h$:*

$$\|B_K'\|_{L^\infty(\widehat{K})} \le C_{\text{affine}} h_K, \qquad \|(B_K')^{-1}\|_{L^\infty(\widehat{K})} \le C_{\text{affine}} h_K^{-1}$$
$$\|(R_K')^{-1}\|_{L^\infty(\widetilde{K})} \le C_{\text{metric}}, \qquad \|\nabla^n R_K\|_{L^\infty(\widetilde{K})} \le C_{\text{metric}} \gamma^n n! \quad \forall n \in \mathbb{N}_0.$$

**Remark 4.2.** *If the mapping $R_K$ in Assumption 4.1 are affine we say that $\mathcal{T}$ is an affine triangulation.*

*The constants $C$ in the estimates below may depend on the shape regularity constants in a continuous way and, possibly, increase with increasing values of $C_{\text{affine}}$, $C_{\text{metric}}$, and $\gamma$.*

For meshes $\mathcal{T}$ satisfying Assumption 4.1 we define the following nonconforming space of piecewise polynomials by

$$S^{p,0} := \{u \in L^2(\Omega) | \quad \forall K \in \mathcal{T} : \, u|_K \circ F_K \in \mathcal{P}_p\}, \tag{4.1}$$

where $\mathcal{P}_p$ denotes the space of polynomials of degree $p$. The *mesh size function* $h_\mathcal{T}$ is defined by $h_\mathcal{T}|_K := \operatorname{diam} K$ for all $K \in \mathcal{T}$. The estimate of $C_{\text{trace}}(S, K)$ in these cases is a local trace estimate for multivariate polynomials:

**Lemma 4.3.** *Let $\mathcal{T}$ satisfy Assumption 4.1. Then there exists $c_{\text{inv}} > 0$ independent of $K \in \mathcal{T}$ and $p$ such that for the hp-finite element space $S^{p,0}(\mathcal{T})$ we have (cf. (2.6))*

$$C_{\text{trace}}(S, K) \le \frac{c_{\text{inv}} p}{\sqrt{h_K}}$$

*Furthermore, for*

$$\mathfrak{a} > \frac{4}{3} c_{\text{inv}}^2 \tag{4.2}$$

*which is independent of $K$, $p$, and $k$, the choice of $\alpha$ given in (2.8) implies the condition (2.7).*

*Proof.* We merely prove the inverse estimate. On the reference element $\widehat{K}$, we have with the multiplicative trace inequality and a standard polynomial inverse estimate (see, e.g., [29, Thm. 4.76], where the case $d = 2$ is covered) for any $v \in \mathcal{P}_p$

$$\|v\|^2_{L^2(\partial\widehat{K})} \leq C\|v\|_{L^2(\widehat{K})}\|v\|_{H^1(\widehat{K})} \leq Cp^2\|v\|^2_{L^2(\widehat{K})}.$$

The assumptions on the element maps $F_K$ are such that the same $h$-dependence as in classical scaling argument are obtained, i.e., for $v \in S^{p,0}(\mathcal{T})$ we get for each $K \in \mathcal{T}$

$$\|v\|_{L^2(\partial K)} \leq Cph^{-1/2}\|v\|_{L^2(K)}. \tag{4.3}$$

For the actual estimate of interest, we let $v \in S^{p,0}(\mathcal{T})$, fix $K$, and set $\widehat{v} := v|_K \circ F_K$. We note $\nabla v = (\nabla\widehat{v}) \circ F_K \circ (F_K')^{-1}$ with, by the assumptions on the properties of $B_K$ and $R_K$,

$$\|(F_K')^{-1}\|_{L^\infty(\widehat{K})} \leq Ch_K^{-1}, \qquad \|(F_K')\|_{L^\infty(\widehat{K})} \leq Ch_K. \tag{4.4}$$

Applying the estimate (4.3) to the components of $\nabla\widehat{v} \circ F_K$ and observing (4.4), one can show the desired result. □

The trace inequality of Lemma 4.3 shows that the constant $\mathfrak{a}$ in (2.8) can be selected such that (2.7) is satisfied. This observation implies the following result:

**Theorem 4.4.** *Let $\alpha$, $\beta$, and $\delta$ be chosen according to (2.8) with $\mathfrak{a}$ sufficiently large. Let $S = S^{p,0}(\mathcal{T})$ be the hp-finite element space based on a mesh $\mathcal{T}$ that satisfies Assumption 4.1.*

- *If $C_S$ satisfies condition (3.13) then the UWVF has a unique solution in $S$.*
- *If $\mathcal{T}$ is an affine triangulation of $\Omega$ and satisfies Assumption 4.1, then the UWVF has a unique solution in $S$.*

## 4.2 Convergence Analysis

In this section we will show that the solution $u$ of the model boundary value problem (1.1), (1.2) can be approximated from the finite element space $S^{p,0}(\mathcal{T})$ provided that $kh/\sqrt{p}$ is small enough and $p \geq c \log k$ (with $c$ sufficiently large independent of $h$, $k$, $p$). Under more stringent conditions on the mesh, we will show that this condition can be relaxed to the condition that $kh/p$ be small enough and $p \geq c \log k$.

The proof of this approximation property is based on the following decomposition lemma, which is a generalization of [25, Theorem 4.10], where the special case $s = 0$ is covered:

**Theorem 4.5** (Decomposition Lemma)**.** *Let $\Omega \in \mathbb{R}^d$, $d \in \{2, 3\}$ be a bounded Lipschitz domain. Assume additionally that $\Omega$ has an analytic boundary. Assume furthermore that the solution operator $(f, g) \mapsto u := S_k(f, g)$ for the Helmholtz boundary value problem (1.1), (1.2) satisfies*

$$\|u\|_{\mathcal{H},\Omega} \leq C_{\text{stab}} k^\vartheta \left(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}\right) \tag{4.5}$$

*for some $C_{\mathrm{stab}}$ and $\vartheta \geq 0$ independent of $k$. Fix $s \in \mathbb{N}_0$. Then there exist constants $C$, $\lambda > 0$ independent of $k \geq k_0$ such that for every $f \in H^s(\Omega)$ and $g \in H^{s+1/2}(\partial\Omega)$ the solution $u = S_k(f,g)$ of the Helmholtz problem (1.3) can be written as $u = u_{H^{s+2}} + u_{\mathcal{A}}$, where, for all $n \in \mathbb{N}_0$*

$$\|u_{\mathcal{A}}\|_{\mathcal{H},\Omega} \leq Ck^{\vartheta}\left(\|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}(\partial\Omega)}\right) \tag{4.6}$$

$$\|\nabla^{n+2} u_{\mathcal{A}}\|_{L^2(\Omega)} \leq C\lambda^n k^{\vartheta-1} \max\{n,k\}^{n+2}\left(\|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}(\partial\Omega)}\right) \tag{4.7}$$

$$\|u_{H^{s+2}}\|_{H^{s+2}(\Omega)} + k^{s+2}\|u_{H^{s+2}}\|_{L^2(\Omega)} \leq C\left(\|f\|_{H^s(\Omega)} + \|g\|_{H^{s+1/2}(\partial\Omega)}\right). \tag{4.8}$$

*Proof.* The proof follows the lines of [25, Theorem 4.10]. The key modifications are collected in the Appendix. □

**Remark 4.6.** *For the present model problem (1.1), (1.2) the assumption (4.5) holds with $\vartheta = 5/2$ by [10, Thm. 2.4]. For star-shaped domains, $\vartheta = 0$ is possible as shown in [22, Prop. 8.1.4] for $d = 2$ and subsequently for $d = 3$ in [8].*

### 4.2.1 Convergence analysis for general non-conforming *hp*-finite elements

In this section we consider general non-conforming *hp*-finite elements. We stress that no conditions are imposed on the element maps $F_K$ that relate element maps of neighboring elements to each other. Hence, the conforming subspace $S \cap H^1(\Omega) \subset S$ may be small. As we will discuss in more detail in Section 5 below, better results can be expected if the conforming subspace $S \cap H^1(\omega) \subset S$ is sufficiently rich.

We start with a lemma that takes the role of the standard scaling argument:

**Lemma 4.7.** *Let $\mathcal{T}$ be a shape-regular mesh in the sense of Assumption 4.1. Fix $s \in \mathbb{N}_0$. Then for each $K \in \mathcal{T}$ and every sufficiently smooth $v$ the following relations between $v$ and $\widehat{v} := v|_K \circ F_K$ are true:*

$$\begin{aligned}
\|v\|_{L^2(K)} &\sim h^{d/2}\|\widehat{v}\|_{L^2(\widehat{K})},\\
\|\nabla v\|_{L^2(K)} &\sim h^{d/2-1}\|\nabla\widehat{v}\|_{L^2(\widehat{K})},\\
\|\nabla^{s+2}\widehat{v}\|_{L^2(\widehat{K})} &\leq Ch^{s+2-d/2}\|v\|_{H^{s+2}(K)},
\end{aligned}$$

*where $C$ and the implied constants depend solely on the constants appearing in Assumption 4.1.*

*Proof.* We will only consider the case of the $(s+2)$nd derivatives. We note the form $F_K = R_K \circ B_K$, where $B_K$ is affine. This implies the estimates

$$\|F_K'\|_{L^\infty(\widehat{K})} \leq Ch_K, \qquad \sum_{\alpha\in\mathbb{N}_0^2:|\alpha|=s+2} \|D^\alpha F_K\|_{L^\infty(\widehat{K})} \leq Ch_K^{s+2},$$

where the constants depend only on $s$ and the shape regularity constants appearing in Assumption 4.1. The chain rule then implies the estimates for $\|\nabla^{s+2}\widehat{v}\|_{L^2(\widehat{K})}$.

*Proof.* We will only consider the case of the $(s+2)$nd derivatives. We note the form $F_K = R_K \circ A_K$, where $A_K$ is affine. This implies the estimates

$$\|F_K'\|_{L^\infty(\widehat{K})} \leq Ch_K, \qquad \sum_{\alpha\in\mathbb{N}_0^2:|\alpha|=s+2} \|D^\alpha F_K\|_{L^\infty(\widehat{K})} \leq Ch_K^{s+2},$$

where the constants depend only on the constants appearing in Assumption 4.1. The chain rule then implies the estimates for $\|\nabla^{s+2}\widehat{v}\|_{L^2(\widehat{K})}$. □

For shape-regular triangulations with possibly hanging nodes we have the following result:

**Theorem 4.8.** *Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$ be a bounded Lipschitz domain with analytic boundary. Fix $s \in \mathbb{N}_0$. Let $\alpha$, $\beta$, $\delta$ be chosen according to (2.8). Fix $\overline{C} > 0$ and assume $p \geq s+1$ as well as $kh/p \leq \overline{C}$. Then the exist constants $C$, $\sigma > 0$ independent of $h$, $p$, and $p$ such that, for every $f \in H^s(\Omega)$ and $g \in H^{s+1/2}(\partial\Omega)$, there holds*

$$\inf_{v \in S} k\|u - v\|_{DG^+} \leq C_{f,g}\left(\left(\frac{h}{p}\right)^s \frac{kh}{\sqrt{p}} + k^{\vartheta}\left\{\left(\frac{h}{h+\sigma}\right)^p + k\left(\frac{kh}{\sigma p}\right)^p\right\}\right), \tag{4.9}$$

*where $C_{f,g} := \|f\|_{H^s(\Omega)} + \|g\|_{H^{s+1/2}(\partial\Omega)}$ and $\vartheta \geq 0$ is given by (4.5) (note also Remark 4.6).*

*Proof.* We employ the splitting $u = u_{H^{s+2}} + u_{\mathcal{A}}$ of Theorem 4.5 with $u_{H^{s+2}} \in H^{s+2}(\Omega)$ and the analytic part $u_{\mathcal{A}}$.

Following [26, Thm. 5.5], we approximate $u_{H^{s+2}}$ and $v_{\mathcal{A}}$ separately in the ensuing steps 1 and 2.

*1. step:* From, e.g., [26, Lemma B.3], we know that for every $s' > d/2$ and every $p \geq s'-1$ there exists a bounded linear operator $\pi_p : H^{s'}(\widehat{K}) \to \mathcal{P}_p$ such that

$$\|u - \pi_p u\|_{H^t(\widehat{K})} \leq C p^{-(s'-t)}|u|_{H^{s'}(\widehat{K})}, \qquad \text{for} \quad 0 \leq t \leq s', \tag{4.10}$$

$$\|u - \pi_p u\|_{H^t(\widehat{e})} \leq C p^{-(s'-1/2-t)}|u|_{H^{s'}(\widehat{K})} \quad \text{for} \quad 0 \leq t \leq s' - 1/2. \tag{4.11}$$

Here, the constant $C > 0$ depends only on $t$, $s'$. By $\widehat{K}$ we denote the reference element and by $\widehat{e}$ one of its edges (in 2D) or faces (in 3D). We apply this approximation result with $s' = s+2$. The elementwise application of the operator $\pi_p$ to $u_{H^{s+2}}$ (pulled back to the reference element $\widehat{K}$) defines an approximation $w_{H^{s+2}} \in S^{p,0}(\mathcal{T})$. By a scaling argument and summation over all elements, the bound (4.10) with $s' = s+2$ implies that $w_{H^{s+2}}$ satisfies

$$\begin{aligned} &k\left(k\|u_{H^{s+2}} - w_{H^{s+2}}\|_{L^2(\Omega)} + \|\nabla_{\mathcal{T}}(u_{H^{s+2}} - w_{H^{s+2}})\|_{L^2(\Omega)}\right) \leq \\ &C\left(k\left(\frac{h}{p}\right)^{s+1} + k^2\left(\frac{h}{p}\right)^{s+2}\right)\left(\|f\|_{H^s(\Omega)} + \|g\|_{H^{s+1/2}(\partial\Omega)}\right). \end{aligned}$$

In order to estimate the terms of the $DG^+$-norm associated with the skeleton, we employ the choice of the parameters $\alpha$, $\beta$, $\delta$ given in (2.8), viz.,

$$\alpha(x) = \frac{4}{3}\max_{K \in \{K_x^+, K_x^-\}} \frac{p^2}{kh_K} \qquad \forall x \in \mathfrak{S}_{\mathcal{T}}^{\mathcal{I}} \quad \text{and} \quad \beta = O(\frac{kh}{p}), \quad \delta = O(\frac{kh}{p}). \tag{4.12}$$

Recall the definition of $\alpha_{\partial K}^{\min}$ as in Remark 2.3 and estimate (3.4). On the inner skeleton $\mathfrak{S}_{\mathcal{T}}^I$ we get

$$k\|\alpha^{-1/2}\{\nabla_{\mathcal{T}}(u_{H^{s+2}} - w_{H^{s+2}})\}\|^2_{L^2(\mathfrak{S}_{\mathcal{T}}^I)} \leq \sum_{K \in \mathcal{T}} \frac{k}{\alpha_{\partial K}^{\min}}\|\{\nabla_{\mathcal{T}}(u_{H^{s+2}} - w_{H^{s+2}})\}\|^2_{L^2(\Omega \cap \partial K)}.$$

Let $X$ denote the minimizer as in (3.4). Then, with the definition (4.12) we get

$$\alpha_{\partial K}^{\min} = \alpha(X) = \frac{4}{3}\max_{K' \in \{K_X^+, K_X^-\}} \frac{p^2}{kh_{K'}} \geq \frac{4}{3}\frac{p^2}{kh_K}. \tag{4.13}$$

so that

$$k\|\alpha^{-1/2}\{\nabla_{\mathcal{T}}(u_{H^{s+2}}-w_{H^{s+2}})\}\|^2_{L^2(\mathfrak{S}^I_{\mathcal{T}})} \leq \sum_{K\in\mathcal{T}}\frac{3k^2h_K}{4p^2}\|\nabla((u_{H^{s+2}}-w_{H^{s+2}})|_K)\|^2_{L^2(\Omega\cap\partial K)}.$$

Thus, we get by scaling (4.10), (4.11) to the mesh $\mathcal{T}$

$$\begin{aligned}k\|\alpha^{-1/2}\{\nabla_{\mathcal{T}}(u_{H^{s+2}}-w_{H^{s+2}})\}\|^2_{L^2(\mathfrak{S}^I_{\mathcal{T}})} &\leq C\sum_{K\in\mathcal{T}}\frac{k^2h}{p^2}\left(\frac{h_K}{p}\right)^{2s+1}\sum_{e\in\mathcal{E}^{\mathcal{I}}(K)}\|u_{H^{s+2}}\|^2_{H^{s+2}(K)}\\ &\leq C\frac{k^2}{p}\left(\frac{h}{p}\right)^{2s+2}\|u_{H^{s+2}}\|^2_{H^{s+2}(\Omega)} \leq C\frac{k^2}{p}\left(\frac{h}{p}\right)^{2s+2}\left(\|f\|^2_{H^s(\Omega)}+\|g\|^2_{H^{s+1/2}(\partial\Omega)}\right).\end{aligned}$$

The following estimates can be obtained by similar arguments:

$$\begin{aligned}k^{1/2}\|\beta^{1/2}[\![\nabla_{\mathcal{T}}(u_{H^{s+2}}-w_{H^{s+2}})]\!]_N\|_{L^2(\mathfrak{S}^{\mathcal{I}}_{\mathcal{T}})} &\leq Ck\left(\frac{h}{p}\right)^{s+1}\left(\|f\|_{H^s(\Omega)}+\|g\|_{H^{s+1/2}(\partial\Omega)}\right),\\ k^{3/2}\|\alpha^{1/2}[\![u_{H^{s+2}}-w_{H^{s+2}}]\!]_N\|_{L^2(\mathfrak{S}^{\mathcal{I}}_{\mathcal{T}})} &\leq Ck\sqrt{p}\left(\frac{h}{p}\right)^{s+1}\left(\|f\|_{H^s(\Omega)}+\|g\|_{H^{s+1/2}(\partial\Omega)}\right),\\ k^{1/2}\|\delta^{1/2}\nabla_{\mathcal{T}}(u_{H^{s+2}}-w_{H^{s+2}})\cdot\mathbf{n}\|_{H^s(\mathfrak{S}^{\mathcal{B}}_{\mathcal{T}})} &\leq Ck\left(\frac{h}{p}\right)^{s+1}\left(\|f\|_{H^s(\Omega)}+\|g\|_{H^{s+1/2}(\partial\Omega)}\right),\\ k^{3/2}\|(1-\delta)^{1/2}(u_{H^2}-w_{H^2})\|_{L^2(\mathfrak{S}^{\mathcal{B}}_{\mathcal{T}})} &\leq Ck^{3/2}\left(\frac{h}{p}\right)^{s+3/2}\left(\|f\|_{H^s(\Omega)}+\|g\|_{H^{s+1/2}(\partial\Omega)}\right).\end{aligned}$$

In total, we get the following approximation property for the $H^{s+2}$-part:

$$k\|u_{H^{s+2}}-w_{H^{s+2}}\|_{DG^+} \leq C\left(\frac{h}{p}\right)^s\left(\frac{kh}{\sqrt{p}}+\left(\frac{kh}{p}\right)^{3/2}+\left(\frac{kh}{p}\right)^2\right)\left(\|f\|_{H^s(\Omega)}+\|g\|_{H^{s+1/2}(\partial\Omega)}\right).$$

Using the assumption $kh/p\leq\overline{C}$, this can be simplified to

$$k\|u_{H^{s+2}}-w_{H^{s+2}}\|_{DG^+} \leq C\left(\frac{h}{p}\right)^s\frac{kh}{\sqrt{p}}\left(\|f\|_{H^s(\Omega)}+\|g\|_{H^{s+1/2}(\partial\Omega)}\right).$$

*2. step:* For the approximation of the analytic part $u_{\mathcal{A}}$, we construct an element $w_{\mathcal{A}}\in S^{p,0}(\mathcal{T})$ as follows. For each $K\in\mathcal{T}$, let the constant $C_K$ by defined by

$$C_K^2 := \sum_{n\in\mathbb{N}_0}\frac{\|\nabla^n u_{\mathcal{A}}\|^2_{L^2(K)}}{(2\lambda\max\{n,k\})^{2n}}.$$

Then, we have

$$\begin{aligned}\|\nabla^n u_{\mathcal{A}}\|_{L^2(K)} &\leq (2\lambda\max\{n,k\})^n C_K \quad \forall n\in\mathbb{N}_0,\\ \sum_{K\in\mathcal{T}}C_K^2 &\leq C\left(\frac{1}{\lambda k}\right)^2 k^{2\vartheta}\left(\|f\|^2_{L^2(\Omega)}+\|g\|^2_{H^{1/2}(\partial\Omega)}\right). \qquad (4.14)\end{aligned}$$

For $q \in \{0,1,2\}$ we get the following estimate (see [26, Proof of Theorem 5.5]) for suitable $\sigma > 0$:

$$\|u_{\mathcal{A}} - w_{\mathcal{A}}\|_{H^q(K)} \leq C h_K^{-q} C_K \left\{ \left( \frac{h_K}{h_K + \sigma} \right)^{p+1} + \left( \frac{k h_K}{\sigma p} \right)^{p+1} \right\}. \tag{4.15}$$

It is convenient to define the abbreviations:

$$E(\sigma) := \left( \frac{h}{h+\sigma} \right)^p + k \left( \frac{kh}{\sigma p} \right)^p,$$
$$M := k^{\vartheta} \left( \|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}(\partial\Omega)} \right).$$

By summing over all elements, it follows as in [26] by suitably adjusting the constant $\sigma$

$$k\|u_{\mathcal{A}} - w_{\mathcal{A}}\|_{\mathcal{H}} \leq C \left( \frac{1}{p} + \frac{kh}{p} \right) E(\sigma) M. \tag{4.16}$$

In order to treat the terms associated with the skeleton $\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}} \cup \mathfrak{S}_{\mathcal{T}}^{\mathcal{B}}$ we use the multiplicative trace inequality

$$\|v\|^2_{L^2(\partial K)} \leq C \left( \|v\|_{L^2(K)} |v|_{H^1(K)} + h_K^{-1} \|v\|^2_{L^2(K)} \right)$$

to obtain

$$k\|\alpha^{-1/2} \{\nabla_{\mathcal{T}} (u_{\mathcal{A}} - w_{\mathcal{A}})\}\|^2_{L^2(\mathfrak{S}_{\mathcal{T}}^I)} \leq \sum_{K \in \mathcal{T}} \frac{k}{\alpha_{\partial k}^{\min}} \|\nabla_{\mathcal{T}} ((u_{\mathcal{A}} - w_{\mathcal{A}})|_K)\|^2_{L^2(\Omega \cap \partial K)}.$$

By using the estimate (4.13) we obtain

$$\begin{aligned}
&k \left\| \alpha^{-1/2} \{\nabla_{\mathcal{T}} (u_{\mathcal{A}} - w_{\mathcal{A}})\} \right\|^2_{L^2(\mathfrak{S}_{\mathcal{T}}^I)} \\
&\leq \sum_{K \in \mathcal{T}} \frac{3k^2 h_K}{4p^2} \left\| \nabla ((u_{\mathcal{A}} - w_{\mathcal{A}})|_K) \right\|^2_{L^2(\Omega \cap \partial K)} \\
&\leq \sum_{K \in \mathcal{T}} \frac{3}{4} \left( \frac{k^2 h_K}{p^2} \right) \left( \|\nabla (u_{\mathcal{A}} - w_{\mathcal{A}})\|_{L^2(K)} |\nabla (u_{\mathcal{A}} - w_{\mathcal{A}})|_{H^1(K)} + h_K^{-1} \|\nabla (u_{\mathcal{A}} - w_{\mathcal{A}})\|^2_{L^2(K)} \right).
\end{aligned}$$

By using the estimates in equation (4.15) we get

$$k\|\alpha^{-1/2} \{\nabla_{\mathcal{T}} (u_{\mathcal{A}} - w_{\mathcal{A}})\}\|^2_{L^2(\mathfrak{S}_{\mathcal{T}}^I)} \leq \sum_{K \in \mathcal{T}} \frac{3Ck^2}{4p^2} \left\{ h_K \left( \frac{h_K}{h_K + \sigma} \right)^{p-1} + \frac{k}{p} \left( \frac{k h_K}{\sigma p} \right)^p \right\}^2 C_K^2.$$

Finally equation (4.14) gives us after suitably adjusting the constant $\sigma$

$$k^{1/2} \|\alpha^{-1/2} \{\nabla_{\mathcal{T}} (u_{\mathcal{A}} - w_{\mathcal{A}})\}\|_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}})} \leq C \frac{1}{p^2} E(\sigma) M.$$

By the similar arguments we obtain the following estimates

$$\begin{aligned}
k^{1/2} \|\beta^{1/2} [\![\nabla_{\mathcal{T}} (u_{\mathcal{A}} - w_{\mathcal{A}})]\!]_N\|_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}})} &\leq C \frac{1}{p^{3/2}} E(\sigma) M, \\
k^{3/2} \|\alpha^{1/2} [\![u_{\mathcal{A}} - w_{\mathcal{A}}]\!]_N\|_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}})} &\leq C E(\sigma) M, \\
k^{1/2} \|\delta^{1/2} \nabla_{\mathcal{T}} (u_{\mathcal{A}} - w_{\mathcal{A}}) \cdot \mathbf{n}\|_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{B}})} &\leq C \frac{1}{p^{3/2}} E(\sigma) M, \\
k^{3/2} \|(1-\delta)^{1/2} (u_{\mathcal{A}} - w_{\mathcal{A}})\|_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{B}})} &\leq C \frac{(kh)^{1/2}}{p} E(\sigma) M.
\end{aligned}$$

The approximation property for the analytic part $u_{\mathcal{A}}$ with respect to the $DG^+$ norm is then

$$k\|u_{\mathcal{A}} - w_{\mathcal{A}}\|_{DG^+} \leq C\left(1 + \frac{1}{p} + \frac{kh}{p} + \frac{\sqrt{kh}}{p}\right) E(\sigma)M \leq CE(\sigma)M,$$

where, in the last estimate we used the assumption $kh/p \leq \overline{C}$. The combination of the estimates of steps 1 and 2 leads to the assertion. □

The approximation result Theorem 4.8 permits us to estimate the adjoint approximation property $\eta(S)$ of (3.10):

**Corollary 4.9.** *Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, be a bounded Lipschitz domain with analytic boundary. Let $\alpha$, $\beta$, $\delta$ be chosen according to (2.8). Fix $\overline{C} > 0$ and assume $kh/p \leq \overline{C}$. Then there exist constants $C$, $\sigma > 0$ such that $\eta_k(S)$ defined in (3.10) satisfies*

$$\eta_k(S) \leq C\left[\frac{kh}{\sqrt{p}} + k^{\vartheta}\left(\left(\frac{h}{h+\sigma}\right)^p + k\left(\frac{kh}{\sigma p}\right)^p\right)\right].$$

*Proof.* We apply Theorem 4.8 with $s = 0$ and $g = 0$. Given $f \in L^2(\Omega)$ let $v = N_k^* f = \overline{N_k \overline{f}}$. Hence, the regularity estimates of Theorem 4.5 (with $g = 0$) are applicable. The assumption $kh/p \leq \overline{C}$ allows us to estimate $(kh/p)^2 \leq Ckh/\sqrt{p}$. □

Finally, the convergence estimate for *hp*-FEM can be stated in the following theorem:

**Theorem 4.10** (Convergence Estimate)**.** *Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, be a bounded Lipschitz domain with analytic boundary. Fix $s \in \mathbb{N}_0$. Let $\alpha$, $\beta$, $\delta$ be chosen according to (2.8) with $\mathfrak{a}$ sufficiently large. Moreover, let $0 < \delta < \frac{1}{3}$. Then, there exist constants $c_1$, $c_2$, $C > 0$ independent of $k, h$ and $p$ such that under the assumptions*

$$\frac{kh}{\sqrt{p}} \leq c_1 \quad \text{together with} \quad p \geq c_2 \log(k) \quad \text{as well as} \quad p \geq s+1 \tag{4.17}$$

*there holds for $f \in H^s(\Omega)$ and $g \in H^{s+1/2}(\partial\Omega)$ the* a priori *estimate*

$$\|u - u_S\|_{DG} \leq C\left[\sqrt{p}\left(\frac{h}{p}\right)^{s+1} + k^{\vartheta-1}\left\{\left(\frac{h}{h+\sigma}\right)^p + k\left(\frac{kh}{\sigma p}\right)^p\right\}\right]\left(\|f\|_{H^s(\Omega)} + \|g\|_{H^{s+1/2}(\partial\Omega)}\right).$$

*In particular, under the additional assumption that $\mathfrak{b}$ and $\mathfrak{d}$ satisfy $\mathfrak{b}$, $\mathfrak{d} \geq c_0 > 0$, there holds*

$$\|\nabla_{\mathcal{T}}(u - u_S)\|_{L^2(\Omega)} + \sqrt{\frac{h}{p}}\|[\![\nabla_{\mathcal{T}}(u - u_S)]\!]_N\|_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}})} + \frac{p}{\sqrt{h}}\|[\![u - u_S]\!]_N\|_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{I}})} \leq C\|u - u_S\|_{DG}.$$

*Proof.* By taking the constant $\mathfrak{a}$ in (2.8) sufficiently large, we can ensure by Lemma 4.3 the condition (2.7). Hence the assertion is a combination of Theorems 3.9, 4.8, and Corollary 4.9. □

# 5 Conclusions

In this paper, we have formulated the ultra-weak variational formulation for abstract finite dimensional test and trial spaces (conforming and non-conforming ones). The concrete choice of the space $S$ enters the stability and convergence analysis via the following four quantities.

(a) Trace constant $C_{\mathrm{trace}}(S,K)$. Due to the formulation as a discontinuous Galerkin method, which contains integral jump terms across element faces, it is quite natural that local trace estimates for the space $S$ are required for the error analysis.

(b) Approximation property $\inf_{v\in S}\|u-v\|_{DG^+}$. In order to derive quantitative error estimates it is obvious that approximation results of $S$ for functions with higher Sobolev regularity are required. The trace estimate (cf. (a)) allows us to "transfer" the local approximation results for the elements $K\in\mathcal{T}$ to the skeleton norm.

(c) Adjoint approximation property $\eta_k(S)$. The decomposition lemma (Lemma 4.5) provides the regularity for the split solution so that, e.g., for *hp*-finite elements, interpolation operators can be constructed for the derivation of quantitative error estimates that are explicit in $k$, $h$, and $p$.

(d) The constant $C_S$ of (3.13). This condition ensures unique solvability of the discrete system (2.4) (see Theorem 3.7). For the important cases of *hp*-finite elements on affine, simplicial triangulations or plane wave approximation spaces, the condition (3.13) is automatically satisfied. If the adjoint approximation property can be controlled, then Theorem 3.9 provides an alternative way of ensuring unique solvability for (2.4).

As an application of our abstract theory we have derived sharp stability and convergence estimates for non-conforming *hp*-finite element spaces. The *a priori* estimate in Theorem 4.10 is optimal in $h$ (note that $f\in H^s(\Omega)$ with $g\in H^{s+1/2}(\partial\Omega)$ implies $u\in H^{s+2}(\Omega)$ by the assumed smoothness of $\partial\Omega$) but suboptimal in $p$ by half an order. This suboptimality is also present in the scale resolution condition (4.17). This is typical of $p$-explicit DG methods. While this suboptimality is sharp in the general case, [15], it can be removed (in both the scale resolution condition (4.17) as well as the *a priori* estimate of Corollary 4.9) by assuming that the approximation space either is conforming, i.e., $S\subset H^1(\Omega)$, or contains an $H^1$-conforming subspace that is sufficiently rich. The essential point in the argument is that in the conforming case the approximants $w_{H^{s+2}}$ and $w_{\mathcal{A}}$ in Theorem 4.8 can be chosen to be in $H^1(\Omega)$ (see, e.g., [26, Proof of Thm. 5.5]). As a consequence the following skeleton terms vanish

$$k^{3/2}\|\alpha^{1/2}[\![u_{H^{s+2}}-w_{H^{s+2}}]\!]_N\|_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{F}})}=0$$
$$k^{3/2}\|\alpha^{1/2}[\![u_{\mathcal{A}}-w_{\mathcal{A}}]\!]_N\|_{L^2(\mathfrak{S}_{\mathcal{T}}^{\mathcal{F}})}=0.$$

We do not work out these arguments here but refer to [28] for the details. This result can be generalized to non-conforming subspaces $S$ which are "close to a conforming subspace" in the sense that there exists a conforming subspace $S'\subset S\cap H^1(\Omega)$ which is sufficiently rich, i.e., allows for a conforming interpolant of the solution and has comparable mesh width. Such a situation arises, e.g., if a conforming *hp*-finite element mesh is further refined locally in a

non-conforming way (allowing for hanging nodes) without affecting the global mesh width $h$. Also here, we refer for the details to [28].

We have restricted the convergence analysis for *hp*-finite element spaces in Section 4 to Lipschitz domains with analytic boundaries in order not to further increase the technicalities in this paper. In [25], the case of polygonal domains for the standard variational formulation of the Helmholtz equation with conforming *hp*-finite element spaces has been considered and regularity estimates in weighted Sobolev spaces have been derived. We expect that the generalization of our theory for the UWVF for non-conforming finite element spaces to polygonal domains is possible along those lines.

# A Details for the proof of Theorem 4.5

We start with an extension of [25, Lemma 4.6] for the modified Helmholtz equation.

**Lemma A.1.** *Let $\Omega$ be a bounded Lipschitz domain with a smooth boundary. Let $S_k^\Delta$ be the solution operator for the boundary value problem*

$$-\Delta u + k^2 u = 0 \quad \text{in } \Omega, \qquad \partial_n u + \mathrm{i}\, k u = g \quad \text{on } \partial\Omega.$$

*Then, for every $s \in \mathbb{N}_0$ there exists $C > 0$ independent of $k \geq k_0$ such that*

$$\|S_k^\Delta(g)\|_{H^{s+2}(\Omega)} \leq C \left[ \|g\|_{H^{s+1/2}(\partial\Omega)} + k^{s+1/2} \|g\|_{L^2(\partial\Omega)} \right], \tag{A.1}$$

$$\|S_k^\Delta(g)\|_{H^1(\Omega)} + k\|S^\Delta(g)\|_{L^2(\Omega)} \leq C k^{-1/2} \|g\|_{L^2(\partial\Omega)}. \tag{A.2}$$

*Proof.* The case $s = 0$ in (A.1) as well as the estimate (A.2) is given in [25, Lemma 4.6]. For $s \geq 1$, we employ induction and the standard shift theorem for the Laplacian: Since $u$ solves

$$-\Delta u = -k^2 u \quad \text{in } \Omega, \qquad \partial_n u = g - \mathrm{i}\, k u \quad \text{on } \partial\Omega,$$

we have

$$\begin{aligned} \|u\|_{H^{s+2}(\Omega)} &\leq C \left[ k^2 \|u\|_{H^s(\Omega)} + \|g\|_{H^{s+1/2}(\partial\Omega)} + k\|u\|_{H^{s+1/2}(\partial\Omega)} \right] \\ &\leq C \left[ k^2 \|u\|_{H^s(\Omega)} + \|g\|_{H^{s+1/2}(\partial\Omega)} + k\|u\|_{H^{s+1}(\Omega)} \right], \end{aligned}$$

where we used a trace inequality. Using the induction hypothesis then leads to an estimate that involves norms of $g$ other than $\|g\|_{H^{s+3/2}(\partial\Omega)}$ and $\|g\|_{L^2(\partial\Omega)}$. These can be removed by an interpolation inequality (see, e.g., [11, Thm. 1.4.3.3]) and an appropriate use of the Young inequality. ∎

The analog of [25, Lemma 4.7] is the following (we use the operator $H_{\partial\Omega}^N$ defined in [25, (4.1c)]):

**Lemma A.2.** *Let $\Omega$ be a bounded Lipschitz domain with a smooth boundary. Fix $q \in (0,1)$ and $s \in \mathbb{N}_0$. Then, the operator $H_{\partial\Omega}^N$ can be selected such that the operator $S_k^\Delta \circ H_{\partial\Omega}^H$ satisfies for some $C > 0$ independent of $k$*

$$k^{s+2}\|S_k^\Delta(H_{\partial\Omega}^N g)\|_{L^2(\Omega)} + k^2 \|S_k^\Delta(H_{\partial\Omega}^N g)\|_{H^s(\Omega)} \leq q\|g\|_{H^{s+1/2}(\partial\Omega)} \tag{A.3}$$

$$\|S_k^\Delta(H_{\partial\Omega}^N g)\|_{H^{s+2}(\Omega)} \leq C\|g\|_{H^{s+1/2}(\partial\Omega)}. \tag{A.4}$$

*Proof.* Estimates (A.3) and (A.4) are shown in [25, Lemma 4.7] for the special case $s = 0$. For $s \geq 1$, these estimates are derived as in [25, Lemma 4.7] by combining Lemma A.1 with [25, Lemma 4.2]. We illustrate the procedure for the second term of the left-hand side of (A.3i) for the case $s \geq 2$: Lemma A.1 yields

$$\begin{aligned}\|S_k^{\Delta}(H_{\partial\Omega}^N)\|_{H^s(\Omega)} &\leq C\left[\|H_{\partial\Omega}^N g\|_{H^{s-3/2}(\partial\Omega)} + k^{s-3/2}\|H_{\partial\Omega}^N g\|_{L^2(\Omega)}\right] \\ &\leq C\left[(q/k)^2\|g\|_{H^{s+1/2}(\partial\Omega)} + k^{s-3/2}(q/k)^{s+1/2}\|g\|_{H^{s+1/2}(\partial\Omega)}\right],\end{aligned}$$

where we used [25, Lemma 4.2]. Rearranging terms yields the result. □

We also need properties of the Newton potential $N_k$, which generalize [25, Lemma 4.5]:

**Lemma A.3.** *Let $\Omega$ be a bounded Lipschitz domain. Fix $s \in \mathbb{N}_0$ and $q \in (0,1)$. Then the operator $H_\Omega$ of [25, (4.1b)] can be selected such that for $0 \leq s' \leq s+2$*

$$\|N_k(H_\Omega f)\|_{H^{s'}(\Omega)} \leq C(q/k)^{s+2-s'}\|f\|_{H^s(\Omega)} \tag{A.5}$$

*Proof.* Follows from the procedure in [25]; see also [24, Lemma 4.2]. The essential point is that [26, (3.35)] can be generalized (by using the notation therein) to

$$\|\partial^\alpha v_{\mu,H^2}\|_{L^2(\mathbb{R}^d)} = (2\pi)^{d/2}\left\|P_{\alpha-\beta}\widehat{G_k M}\,(1-\chi_{\lambda k})\,\widehat{\partial^\beta f}\right\|_{L^2(\mathbb{R}^d)}$$

for all $\alpha \in \mathbb{N}_0^d$ and $\beta \in \mathbb{N}_0^d$. By selecting $|\alpha| = s'$ and $|\beta| = s'-2$, we see that $|\alpha-\beta| = 2$ and this case is considered in [26, (3.35)]. By performing the same estimates as in [26, after (3.35)], we derive for $|\alpha-\beta| = 2$ the estimate

$$\|\partial^\alpha N_k(H_\Omega f)\|_{L^2(\Omega)} \leq C\left\|\partial^\beta H_\Omega f\right\|_{L^2(\Omega)}$$

so that

$$\|N_k(H_\Omega f)\|_{H^{s'}(\Omega)} \leq C\|H_\Omega f\|_{H^{s'-2}(\Omega)}$$

follows. The combination with [25, Lemma 4.2] leads to the assertion (A.5). □

The next lemma generalizes [25, Lemma 4.15]

**Lemma A.4.** *Let $\Omega$ be a bounded Lipschitz domain with a smooth boundary. Fix $s \in \mathbb{N}_0$. Assume that the solution operator $(f,g) \mapsto S_k(f,g)$ for (1.1), (1.2) satisfies (4.5). Then $S_k$ admits the following decomposition: $u = S_k(f,0) = u_{\mathcal{A}} + u_{H^{s+2}} + \widetilde{u}$, where*

$$\begin{aligned}\|u_{\mathcal{A}}\|_{H^1(\Omega)} + k\|u_{\mathcal{A}}\|_{L^2(\Omega)} &\leq Ck^{\vartheta}\|f\|_{L^2(\Omega)}, \\ \|\nabla^{n+2}u_{\mathcal{A}}\|_{L^2(\Omega)} &\leq Ck^{\vartheta-1}\gamma^n \max\{k,n\}^{n+2}\|f\|_{L^2(\Omega)} \qquad \forall n \in \mathbb{N}_0, \\ k^{s+2}\|u_{H^{s+2}}\|_{L^2(\Omega)} + \|u_{H^{s+2}}\|_{H^{s+2}(\Omega)} &\leq C\|f\|_{H^s(\Omega)}\end{aligned}$$

*for constants $C$, $\gamma > 0$ independent of $k$ and $n$, and the remainder $\widetilde{u} = S_k(\widetilde{f},0)$ satisfies the boundary value problem*

$$-\Delta\widetilde{u} - k^2\widetilde{u} = \widetilde{f} \quad \text{in } \Omega, \qquad \partial_n\widetilde{u} - \mathrm{i}\,k\widetilde{u} = 0 \quad \text{on } \partial\Omega$$

*for a right-hand side $\widetilde{f} \in H^s(\Omega)$ with*

$$\|\widetilde{f}\|_{H^s(\Omega)} \leq q\|f\|_{H^s(\Omega)}, \qquad \|\widetilde{f}\|_{L^2(\Omega)} \leq q\|f\|_{L^2(\Omega)}.$$

*Proof.* The proof follows that of [25, Lemma 4.15]. We only need to show the additional bound $\|u_{H^{s+2}}\|_{H^{s+2}(\Omega)} \leq C\|f\|_{H^s(\Omega)}$. To that end, we have to consider, in the notation of [25, Lemma 4.15] the terms

$$u^I_{H^2} = N_k(H_\Omega f), \tag{A.6}$$
$$u^{II}_{H^2} = S^\Delta_k(H^N_{\partial\Omega}(\mathrm{i}\,k u^I_{H^2} - \partial_n u^I_{H^2})). \tag{A.7}$$

For (A.6), we use Lemma A.3 to get

$$k^{s+2}\|N_k(H_\Omega f)\|_{L^2(\Omega)} + k\|N_k(H_\Omega f)\|_{H^{s+1}(\Omega)} + \|N_k(H_\Omega f)\|_{H^{s+2}(\Omega)} \leq C\|f\|_{H^s(\Omega)}$$
$$\|N_k(H_\Omega f)\|_{H^s(\Omega)} \leq C(q/k)^2\|f\|_{H^s(\Omega)}.$$

This implies in particular with a trace inequality that

$$\|\,\mathrm{i}\,k u^I_{H^2} - \partial_n u^I_{H^2}\|_{H^{s+1/2}(\partial\Omega)} \leq Ck\|u^I_{H^2}\|_{H^{s+1}(\Omega)} + C\|u^I_{H^2}\|_{H^{s+2}(\Omega)} \leq C\|f\|_{H^s(\Omega)},$$

so that also for (A.7), we can obtain, with the aid of Lemma A.2, the bounds

$$\|S^\Delta_k(H^N_{\partial\Omega}(\mathrm{i}\,k u^I_{H^2} - \partial_n u^I_{H^2}))\|_{H^{s+2}(\Omega)} \leq C\|f\|_{H^s(\Omega)},$$
$$k^{s+2}\|S^\Delta_k(H^N_{\partial\Omega}(\mathrm{i}\,k u^I_{H^2} - \partial_n u^I_{H^2}))\|_{L^2(\Omega)} + k^2\|S^\Delta_k(H^N_{\partial\Omega}(\mathrm{i}\,k u^I_{H^2} - \partial_n u^I_{H^2}))\|_{H^s(\Omega)} \leq q\|f\|_{H^s(\Omega)}.$$

From the above estimates follows the bound for $\|u_{H^{s+2}}\|_{H^{s+2}(\Omega)}$. The estimate for $\widetilde{f}$ follows also from the above observations by noting that we have to set $\widetilde{f} := 2k^2 u^{II}_{H^2}$ and then suitably adjust $q$ as in the proof [25, Lemma 4.15]. ☐

Finally, we formulate the analog of [25, Lemma 4.16]:

**Lemma A.5.** *Assume the hypotheses of Lemma A.4. Fix $q \in (0,1)$ and $s \in \mathbb{N}_0$. Then the solution $u = S_k(0,g)$ can be written as $u = u_{\mathcal{A}} + u_{H^{s+2}} + \widetilde{u}$, where*

$$\|u_{\mathcal{A}}\|_{H^1(\Omega)} + k\|u_{\mathcal{A}}\|_{L^2(\Omega)} \leq Ck^\vartheta\|g\|_{H^{1/2}(\partial\Omega)}, \tag{A.8}$$
$$\|\nabla^{n+2}u_{\mathcal{A}}\|_{L^2(\Omega)} \leq Ck^{\vartheta-1}\gamma^n\max\{n,k\}^{n+2}\|g\|_{H^{1/2}(\partial\Omega)} \qquad \forall n \in \mathbb{N}_0, \tag{A.9}$$
$$k^{s+2}\|u_{H^{s+2}}\|_{L^2(\Omega)} + \|u_{H^{s+2}}\|_{H^{s+2}(\Omega)} \leq C\|g\|_{H^{s+1/2}(\partial\Omega)}, \tag{A.10}$$

*where the constants $C$, $\gamma > 0$ are independent of $k$ and $n$. The remainder $\widetilde{u}$ satisfies the boundary value problem*

$$-\Delta\widetilde{u} - k^2\widetilde{u} = 0 \quad \text{in } \Omega, \qquad \partial_n\widetilde{u} - \mathrm{i}\,k\widetilde{u} = \widetilde{g} \quad \text{on } \partial\Omega$$

*for data $\widetilde{g} \in H^{s+1/2}(\partial\Omega)$ with*

$$\|\widetilde{g}\|_{H^{s+1/2}(\partial\Omega)} \leq q\|g\|_{H^{s+1/2}(\partial\Omega)}.$$

*Proof.* The proof follows [25, Lemma 4.16], and we will only discuss (A.8). To that end, we have to consider, in the notation of [25, Lemma 4.16], the terms

$$u^I_{H^2} = S^\Delta_k(H^N_{\partial\Omega}g), \tag{A.11}$$
$$u^{II}_{H^2} = N_k(H_\Omega(2k^2u^I_{H^2})). \tag{A.12}$$

For the term in (A.11), we use Lemma A.2 to get

$$k^{s+2}\|u^I_{H^2}\|_{L^2(\Omega)} + \|u^I_{H^2}\|_{H^{s+2}(\Omega)} \leq C\|g\|_{H^{s+1/2}(\partial\Omega)}$$
$$k^2\|u^I_{H^2}\|_{H^s(\Omega)} \leq q\|g\|_{H^{s+1/2}(\partial\Omega)}.$$

For the term in (A.12), we use Lemma A.3 to arrive at

$$k\|u^{II}_{H^2}\|_{H^{s+1}(\Omega)} + k^{s+2}\|u^{II}_{H^2}\|_{L^2(\Omega)} + \|u^{II}_{H^2}\|_{H^{s+2}(\Omega)} \leq Ck^2\|u^I_{H^2}\|_{H^s(\Omega)} \leq Cq\|g\|_{H^{s+1/2}(\partial\Omega)}.$$

As in the proof of [25, Lemma 4.16], we then set $\widetilde{g} := \mathrm{i}\,ku^{II}_{H^2} - \partial_n u^{II}_{H^2}$ and use the above estimates to get with the trace inequality

$$\|\widetilde{g}\|_{H^{s+1/2}(\partial\Omega)} \leq C\left[k\|u^{II}_{H^2}\|_{H^{s+1}(\Omega)} + \|u^{II}_{H^2}\|_{H^{s+2}(\Omega)}\right] \leq Cq\|g\|_{H^{s+1/2}(\partial\Omega)}.$$

Suitably adjusting the constant $q$ yields the result. □

# References


[1] R. A. Adams. *Sobolev Spaces*. Academic Press, 1975.

[2] M. Amara, H. Calandra, R. Djellouli, and M. Grigoroscuta-Strugaru. A stabilized DG-type method for solving efficiently Helmholtz problems. Technical Report 7461, INRIA, 2010.

[3] D. Arnold, F. Brezzi, B. Cockburn, and L. Marini. *Unified analysis of discontinuous Galerkin methods for elliptic problems*. SIAM J. Numer. Anal., 39:1749-1779, 2002.

[4] A. Buffa and P. Monk. *Error estimates for the Ultra Weak Variational Formulation of the Helmholtz Equation*. Math. Mod. Numer. Anal.42:925-940, 2008.

[5] P. Castillo, B. Cockburn, I. Perugia, and D. Schötzau. *An a priori error analysis of the local discontinuous Galerkin method for elliptic problems*. SIAM J. Numer. Anal., 38:1676-1706, 2000.

[6] O. Cessenat and B. Després. *Application of an ultra weak variational formulation of elliptic PDEs to the two-dimensional Helmholtz equation*. SIAM J. Numer. Anal., 35:255-299,1594-1607 1998.

[7] O. Cessenat and B. Després. *Using plane waves as base functions for solving time harmonic equations with the ultra weak variational formulation*. J. Computational Acoustics, 11:227-238, 2003.

[8] P. Cummings and X. Feng. Sharp regularity coefficient estimates for complex-valued acoustic and elastic Helmholtz equations. *Math. Models Methods Appl. Sci.*, 16(1):139–160, 2006.

[9] B. Després. *Sur une formulation variationelle de type ultra-faible*. C.R. Acad. Sci. Paris, Ser. I, 318:939-944, 1994.

[10] S. Esterhazy and J.M. Melenk. On stability of discretizations of the Helmholtz equation. In I.G. Graham, T.Y. Hou, O. Lakkis, and R. Scheichl, editors, *Numerical Analysis of Multiscale Problems*, volume 83 of *Lecture Notes in Computational Science and Engineering*, pages 285–324, Springer Verlag, 2012.

[11] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Pitman, 1985.

[12] X. Feng and H. Wu. *Discontinuous Galerkin methods for the Helmholtz equation with large wave number*. SIAM J. Numer. Anal., 47(4):2872-2896, 2009.

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

[14] X. Feng and Y. Xing. *Absolutely stable local Discontinuous Galerkin methods for the Helmholtz equation with large wave number*. Technical report, 2010.

[15] E. Georgoulis, E. Hall, and J.M. Melenk. On the suboptimality of the $p$-version interior penalty discontinuous galerkin method. *J. Sci. Comp.*, 42(1):54–67, 2010.

[16] C. J. Gittelson, R. Hiptmair and I. Perugia. *Plane Wave Discontinuous Galerkin Methods: Analysis of the h-version*. Math. Model. Numer. Anal. 43:297-331, 2009.

[17] I. Harari. *A Survey of finite element methods for time-harmonic acoustics*. Comput. Methods. Appl. Mech. Engrg., 195(13-16):1594-1607, 2006.

[18] I. Harari and T. J. R. Hughes. *Galerkin/least-squares finite element methods for the reduced wave equation with nonreflecting boundary conditions in unbounded domains*. Comput. Methods. Appl.Mech. Engrg., 98(3):411-454, 1992.

[19] T. Huttunen and P. Monk. *The use of plane waves to approximate wave propagation in anisotropic media*. J. Computational Mathematics, 25:350-367, 2007.

[20] F. Ihlenburg. *Finite Element Analysis of Acoustic Scattering*, volume 132 of *Applied Mathematical Sciences*. Springer Verlag, 1998.

[21] M. Löhndorf and J.M. Melenk. Wavenumber-explicit *hp*-BEM for high frequency scattering. *SIAM J. Numer. Anal.*, 49(6):2340–2363, 2011.

[22] J. M. Melenk. *On Generalized Finite Element Methods*. PhD thesis, University of Maryland at College Park, 1995.

[23] J. M. Melenk. *hp finite element methods for singular perturbations*, volume 1796 of *Lecture Notes in Mathematics*. Springer Verlag, 2002.

[24] J.M. Melenk. Mapping properties of combined field Helmholtz boundary integral operators. Technical Report 01/2010, Institute for Analysis and Scientific Computing, TU Wien, 2010 (revised version).

[25] J.M. Melenk and S. Sauter. Wavenumber explicit convergence analysis for finite element discretizations of the Helmholtz equation. *SIAM J. Numer. Anal.*, 49:1210–1243, 2011.

[26] J. M. Melenk and S. Sauter. *Convergence Analysis for Finite Element Discretizations of the Helmholtz equation with Dirichlet-to-Neumann boundary condition*. Math. Comp., 79:1871-1914, 2010.

[27] P. Monk and D.Q. Wang. A least squares methods for the Helmholtz equation. *Comput. Meth. Appl. Mech. Engrg.*, 175:121–136, 1999.

[28] A. Parsania. *Convergence Analysis for Finite Element Discretizations of Highly Indefinite Problems*. Doctoral thesis, Institut für Mathematik, Universität Zürich, 2012.

[29] C. Schwab. *p- and hp-Finite Element Methods*. Oxford University Press, 1998.