A Posteriori Error Estimation for Elliptic Partial Differential Equations
Lecture Notes of the Zürich Summerschool 2012.

S.A. Sauter*

August 23, 2012

Abstract

These lecture notes comprise the talks of the author on "A posteriori error estimates for modelling errors" and on "A posteriori error estimates for highly indefinite problems" given at the Zürich Summerschool 2012.

1 Lecture 1: Combined A Posteriori Modeling - Discretization Error Estimate for Elliptic Problems with Complicated Interfaces

Remark. This part of the lecture notes is a slightly extended and modified version of [53].

1.1 Introduction

This lecture is concerned with the solution of elliptic boundary value problems with complicated coefficients. As a model problem we choose the diffusion equation $\operatorname{div}(\mathbf{A}\operatorname{grad} u) = f$ in a two- or three-dimensional bounded domain with homogeneous Dirichlet boundary conditions. From the physical point of view, this equation can be regarded as a model of a stationary diffusion. Our focus is on applications, where

- the diffusion matrix $\mathbf{A}$ is piecewise smooth but, possibly, highly oscillatory and/or discontinuous along interfaces with, possibly, very rough and complicated structure,
- the target accuracy for the approximate solution is fairly moderate.

In this situation, the application of the "textbook" finite element method requires that the fine details of the interfaces are resolved by the finite element mesh. In particular for problems in 3D, the resulting linear system becomes very large, typically, much too large from the viewpoint of the moderate accuracy requirements.

*(stas@math.uzh.ch), Institut für Mathematik, Universität Zürich, Winterthurerstr 190, CH-8057 Zürich, Switzerland

We will introduce a *defeaturing strategy* for the diffusion matrix $\mathbf{A}$ which will be combined with the numerical discretization of the partial differential equation (PDE). In the following, we will briefly sketch the idea of the methodology.

The methods starts with a very simple (e.g., constant) approximation $\mathbf{A}_0$ of $\mathbf{A}$ and a very coarse finite element space S_0. The exact solution for the boundary value problem with $\mathbf{A}$ being replaced by $\mathbf{A}_0$ is denoted by u_0 and its Galerkin approximation with respect to S_0 is denoted by $u_{0,0}$. The error of this computable approximation is $\|\mathbf{A}\nabla(u - u_{0,0})\|$, where $\|\cdot\|$ denotes the L^2 norm. We will derive an a posteriori error majorant which is the sum of two terms

$$\|u - u_{0,0}\| \leq E_{\text{disc}} + E_{\text{mod}}.$$

Both terms, E_{mor} and E_{disc}, depend on $u_{0,0}$. The quantity E_{mor} measures the error caused by the simplification $\mathbf{A} \leftarrow \mathbf{A}_0$ and E_{disc} measures the error of the numerical error $\|u_0 - u_{0,0}\|$. The improvement strategy is now driven by this splitting: If the part E_{mor} dominates the error majorant, the diffusion matrix $\mathbf{A}_0$ is replaced by an improved approximation $\mathbf{A}_1$ and the finite element space is unchanged while, in the reversed case, $\mathbf{A}_0$ is unchanged and the space S_0 is enriched, e.g., by mesh refinement.

This adaptive modeling-discretization strategy can be iterated and results in a sequence $(u_{m_\ell,n_\ell})_\ell$ of computable solutions with balanced modelling and numerical errors.

Historically, the subject of a posteriori error estimation was mainly focused on the indication of discretization errors (e.g., see [3], [59], and references therein). In these cases, the error is measured by the quantity $\|u - u_h\|$, where u is the exact solution, u_h is the Galerkin approximation, and $\|\cdot\|$ is a certain norm associated with the problem (see, e.g., [2], [3], [7], [8], [9], [16], [22], [57], [59]).

Our method differs from these approaches and its derivation is based the publications (see [40] - [49]) in which estimates of the difference between the exact solution of boundary value problems and arbitrary functions from the corresponding energy space has been derived by purely functional methods, i.e., without requiring specific information on the approximating subspace and the numerical method used. As a result, the estimates contain no mesh dependent constants and are valid for any conforming approximation from the respective energy space. In [47, 49], these properties have been used for the analysis of numerical discretization errors.

1.2 Setting

We consider the elliptic problem

$$\begin{aligned} -\operatorname{div}(\mathbf{A}\nabla u) &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{1.1}$$

where Ω is a bounded domain in $\mathbb{R}^d$ $(d = 2, 3)$ with Lipschitz boundary $\partial\Omega$. The diffusion matrix $\mathbf{A}(x)$ belongs to the set $\mathbb{R}^{d\times d}$ of $d \times d$ matrices with real coefficients. We assume that

$$\mathbf{A} \text{ is symmetric}, \quad \mathbf{A}(x) \in L^\infty(\Omega, \mathbb{R}^{d\times d}), \quad f \in L^2(\Omega),$$

and

$$0 < c_1^2 := \operatorname*{ess\,inf}_{x\in\Omega} \inf_{v\in\mathbb{R}^d\setminus\{0\}} \frac{\mathbf{A}(x)v \cdot v}{v \cdot v} \leq \operatorname*{ess\,sup}_{x\in\Omega} \sup_{v\in\mathbb{R}^d\setminus\{0\}} \frac{\mathbf{A}(x)v \cdot v}{v \cdot v} =: \rho(\mathbf{A}) < \infty. \tag{1.2}$$

The norm in $L^2(\Omega)$ is denoted by $\|u\|_\Omega$ and "$\cdot$" stands for the Euclidean scalar product in $\mathbb{R}^d$. The notation $L^2(\Omega, \mathbb{R}^d)$ is used for the vector-valued functions with components in $L^2(\Omega)$ and

$$H_0^1(\Omega) := \left\{ u \in H^1(\Omega) \mid u|_{\partial\Omega} = 0 \text{ in the sense of traces} \right\}.$$

Also we introduce the space

$$H(\Omega, \mathrm{div}) := \{ q \in L^2(\Omega, \mathbb{R}^d) \mid \mathrm{div}\ q \in L^2(\Omega) \}$$

which is a Hilbert space endowed with the scalar product

$$(p, q)_{\mathrm{div}} := \int_\Omega (p \cdot q + \mathrm{div}\ p\ \mathrm{div}\ q)$$

and the norm $\|q\|_{\mathrm{div}} := (q, q)_{\mathrm{div}}^{1/2}$. For functions in $L^2(\Omega, \mathbb{R}^d)$, the energy and complementary energy norms are given by

$$\|q\|_{\mathbf{A}}^2 := \int_\Omega \mathbf{A}\, q \cdot q \quad \text{and} \quad \|q\|_{\mathbf{A}^{-1}}^2 := \int_\Omega \mathbf{A}^{-1} q \cdot q. \tag{1.3}$$

The generalized solution of (1.1) is the solution of the variational problem

$$\text{Find } u \in H_0^1(\Omega) \text{ such that} \qquad a(u, v) = \int_\Omega f\, v, \quad \forall v \in H_0^1(\Omega), \tag{1.4}$$

where $a(u, v) := \int_\Omega \mathbf{A} \nabla u \cdot \nabla v$ is the bilinear form generated by $\mathbf{A}$.

1.3 Combined Error Majorant

Consider the following simplified problem $\mathcal{P}_\varepsilon$: Find $u_\varepsilon \in H_0^1(\Omega)$ such that

$$a_\varepsilon(u_\varepsilon, v) := \int_\Omega \mathbf{A}_\varepsilon \nabla u_\varepsilon \cdot \nabla v = \int_\Omega f\, v \quad \text{for all } v \in H_0^1(\Omega), \tag{1.5}$$

where $\mathbf{A}_\varepsilon \in L^\infty(\Omega, \mathbb{R}^{d\times d})$ is a certain approximation of $\mathbf{A}$. We will always assume that for any ε, the matrix $\mathbf{A}_\varepsilon$ is positive definite and

$$c_{1\varepsilon}^2 |\zeta|^2 \le \mathbf{A}_\varepsilon(x)\, \zeta \cdot \zeta \le \rho(\mathbf{A}_\varepsilon)\, |\zeta|^2 \quad \text{for all } x \in \Omega \quad \text{and } \zeta \in \mathbb{R}^d. \tag{1.6}$$

Let $\mathcal{T}_h$ be a simplicial finite mesh in the sense of Ciarlet [19], where h denotes the maximal simplex diameter. Let S_h denote the continuous, piecewise affine finite element space

$$S_h := \left\{ u \in C^0(\overline{\Omega}) \mid \forall K \in \mathcal{T}_h \quad u|_K \in \mathbb{P}_1 \right\}.$$

The corresponding $H_0^1(\Omega)$ conforming space and the vector-valued version are given by

$$S_{h,0} := S_h \cap H_0^1(\Omega) \quad \text{and} \quad S_h^2 := S_h \times S_h.$$

The Galerkin finite element solution to the simplified problem $\mathcal{P}_\varepsilon$ is defined by

$$\text{Find } u_{\varepsilon,h} \in S_{\mathcal{T},0} \text{ such that } \quad a_\varepsilon(u_{\varepsilon,h},\, v_h) := \int\limits_\Omega \mathbf{A}_\varepsilon\, \nabla u_{\varepsilon,h} \cdot \nabla v_h \,=\, \int\limits_\Omega f\, v_h \quad \text{for all } v_h \in S_{h,0}. \tag{1.7}$$

In order to estimate the *discretization* error $\|\nabla(u_\varepsilon - u_{\varepsilon,h})\|_{\mathbf{A}_\varepsilon}$, we use a posteriori error estimates of the functional type (see [36] - [45], [47, 49] and the references therein). In our case, the estimate takes the form (we refer here to the ZSS12 lectures of Prof. Repin)

$$\|\nabla(u_\varepsilon - u_{\varepsilon,h})\|^2_{\mathbf{A}_\varepsilon} \le \mathcal{M}^2_\Omega(u_{\varepsilon,\,h}\,, y,\, \beta) := (1+\beta)\|\mathbf{A}_\varepsilon\, \nabla u_{\varepsilon,\,h} - y\|^2_{\mathbf{A}_\varepsilon^{-1}} + \\ + \left(1 + \frac{1}{\beta}\right) C^2_\Omega \| \operatorname{div} y + f\|^2_\Omega. \tag{1.8}$$

Here, y is an arbitrary vector-valued function from $H(\Omega, \operatorname{div})$, β is an arbitrary positive number, and $C^2_\Omega := c^{-2}_{1\varepsilon}\, C^2_{F\Omega}$, where $c_{1\varepsilon}$ is as in (1.6) and $C_{F\Omega}$ is the Friedrichs constant for the domain Ω, i.e.,

$$C_{F\Omega} := \sup_{w \in H^1_0(\Omega)\setminus\{0\}} \frac{\|w\|_\Omega}{\|\nabla w\|_\Omega}.$$

Theorem 1.1 *The total error is bounded from above by the sum*

$$\|\nabla(u \,-\, u_{\varepsilon,h})\|_{\mathbf{A}} \le E^{\varepsilon,\,h}_{\text{disc}} \,+\, E^{\varepsilon}_{\text{mod}}, \tag{1.9}$$

where $E^{\varepsilon,\,h}_{\text{disc}}$ and $E^{\varepsilon}_{\text{mod}}$ represent the discretization and modeling parts of the error, respectively, and are defined and estimated as follows:

$$E^{\varepsilon,\,h}_{\text{disc}} := \|\nabla(u_\varepsilon \,-\, u_{\varepsilon,\,h})\|_{\mathbf{A}} \,\le\, \kappa_1\, \mathcal{M}_\Omega(u_{\varepsilon,h},\, y,\, \beta), \tag{1.10}$$

$$E^{\varepsilon}_{\text{mod}} := \|\nabla(u \,-\, u_\varepsilon)\|_{\mathbf{A}} \le\, \kappa_\varepsilon \left(\frac{\kappa_2}{2}\, \mathcal{M}^2_\Omega(u_{\varepsilon,h},\, y,\, \beta) + \int\limits_\Omega f\, u_{\varepsilon,\,h}\right)^{1/2} \tag{1.11}$$

where $\kappa^2_1 := 1 + \rho(\Lambda_\varepsilon - I)$, $\kappa^2_\varepsilon := \frac{2\,\kappa_2}{2\,\kappa_2 - 1}\, \rho(\Lambda_\varepsilon + \Lambda^{-1}_\varepsilon - 2I)$, $\Lambda_\varepsilon := \mathbf{A}^{-1/2}_\varepsilon \mathbf{A} \mathbf{A}^{-1/2}_\varepsilon$, I is the identity matrix, ρ is defined by (1.2), and κ_2 in (1.17).

Proof. By the triangle inequality, we obtain

$$\|\nabla(u - u_{\varepsilon,\,h})\|_{\mathbf{A}} \le \|\nabla(u_\varepsilon - u_{\varepsilon,\,h})\|_{\mathbf{A}} + \|\nabla(u - u_\varepsilon)\|_{\mathbf{A}} \,=\, E^{\varepsilon,\,h}_{\text{disc}} + E^{\varepsilon}_{\text{mod}}. \tag{1.12}$$

We estimate the term $E^{\varepsilon,\,h}_{\text{disc}} = \|\nabla(u_\varepsilon \,-\, u_{\varepsilon,\,h})\|_{\mathbf{A}}$, as follows:

$$\begin{aligned}
\left(E^{\varepsilon,\,h}_{\text{disc}}\right)^2 &= \|\nabla(u_\varepsilon \,-\, u_{\varepsilon,\,h})\|^2_{\mathbf{A}_\varepsilon} + \int\limits_\Omega (\mathbf{A} - \mathbf{A}_\varepsilon)\, \nabla(u_\varepsilon - u_{\varepsilon,\,h}) \cdot \nabla(u_\varepsilon - u_{\varepsilon,\,h}) \\
&= \|\nabla(u_\varepsilon \,-\, u_{\varepsilon,\,h})\|^2_{\mathbf{A}_\varepsilon} + \int\limits_\Omega (\Lambda_\varepsilon - I)\mathbf{A}^{1/2}_\varepsilon \nabla(u_\varepsilon - u_{\varepsilon,\,h}) \cdot \mathbf{A}^{1/2}_\varepsilon\, \nabla(u_\varepsilon - u_{\varepsilon,\,h}) \\
&\le (1 + \rho\,(\Lambda_\varepsilon - I))\, \|\nabla(u_\varepsilon \,-\, u_{\varepsilon,\,h})\|^2_{\mathbf{A}_\varepsilon}.
\end{aligned}$$

Since the last norm is estimated by (1.8), we arrive at (1.10).

To estimate the term $E^{\varepsilon}_{\rm mod}$, we note that

$$0 = a\,(u\,-\,u_\varepsilon,\,v)\,+\,(a\,-\,a_\varepsilon)\,(u_\varepsilon,\,v), \qquad \forall\, v \in H^1_0\,(\Omega)\,,$$

and choose $v\,=\,u\,-\,u_\varepsilon$. Then,

$$\begin{aligned}(E^{\varepsilon}_{\rm mod})^2 = \|\nabla(u\,-\,u_\varepsilon)\|^2_{\mathbf{A}} = a\,(u\,-\,u_\varepsilon,\,u\,-\,u_\varepsilon)\,&=\,(a_\varepsilon\,-\,a)(u_\varepsilon,\,u\,-\,u_\varepsilon)\\ &= \int\limits_\Omega (\mathbf{A}_\varepsilon\,-\,\mathbf{A})\,\nabla u_\varepsilon\,\cdot\,\nabla(u\,-\,u_\varepsilon).\end{aligned}$$

By the Cauchy-Schwarz inequality, we find that

$$\|\nabla(u\,-\,u_\varepsilon)\|^2_{\mathbf{A}} \le \|(\mathbf{A}_\varepsilon\,-\,\mathbf{A})\,\nabla u_\varepsilon\|_{\mathbf{A}^{-1}}\,\|\nabla(u\,-\,u_\varepsilon)\|_{\mathbf{A}}.$$

Hence,

$$\begin{aligned}\|\nabla(u-u_\varepsilon)\|^2_{\mathbf{A}} \le \|(\mathbf{A}_\varepsilon\,-\,\mathbf{A})\,\nabla u_\varepsilon\|^2_{\mathbf{A}^{-1}} &= \left\|(I-\Lambda_\varepsilon)\,\mathbf{A}_\varepsilon^{1/2}\nabla u_\varepsilon\right\|^2_{\Lambda_\varepsilon^{-1}}\\ = \int\limits_\Omega (\Lambda_\varepsilon + \Lambda_\varepsilon^{-1} - 2I)\mathbf{A}_\varepsilon^{1/2}\nabla u_\varepsilon\cdot\mathbf{A}_\varepsilon^{1/2}\nabla u_\varepsilon\,&\le\,\rho\,(\Lambda_\varepsilon+\Lambda_\varepsilon^{-1}-2I)\,\|\nabla u_\varepsilon\|^2_{\mathbf{A}_\varepsilon}.\end{aligned}$$

Further, by the Young inequality with an arbitrary $\mu > 0$, we derive

$$\begin{aligned}\|\nabla u_\varepsilon\|^2_{\mathbf{A}_\varepsilon} = \int\limits_\Omega f\,(u_\varepsilon - u_{\varepsilon,h})\,+\,\int\limits_\Omega f\,u_{\varepsilon,h}\,&=\,a_\varepsilon\,(u_\varepsilon, u_\varepsilon - u_{\varepsilon,h})\,+\,\int\limits_\Omega f\,u_{\varepsilon,h}\\ \le \frac{1}{2\mu}\,\|\nabla\,u_\varepsilon\|^2_{\mathbf{A}_\varepsilon}\,+\,\frac{\mu}{2}\,\|\nabla(u_\varepsilon - u_{\varepsilon,h})\|^2_{\mathbf{A}_\varepsilon}\,&+\,\int\limits_\Omega f\,u_{\varepsilon,h},\end{aligned}$$

and obtain for $\mu > 1/2$

$$\|\nabla u_\varepsilon\|^2_{\mathbf{A}_\varepsilon} \le \frac{\mu^2}{2\,\mu - 1}\,\|\nabla\,(u_\varepsilon - u_{\varepsilon,h})\,\|^2_{\mathbf{A}_\varepsilon}\,+\,\frac{2\,\mu}{2\,\mu-1}\int\limits_\Omega f\,u_{\varepsilon,h}. \tag{1.13}$$

Therefore,

$$\|\nabla(u-u_\varepsilon)\|^2_{\mathbf{A}} \le \rho\,(\Lambda_\varepsilon+\Lambda_\varepsilon^{-1}-2I)\left(\frac{\mu^2}{2\,\mu-1}\,\|\nabla\,(u_\varepsilon - u_{\varepsilon,h})\,\|^2_{\mathbf{A}_\varepsilon}\,+\,\frac{2\,\mu}{2\,\mu-1}\int\limits_\Omega f\,u_{\varepsilon,h}\right). \tag{1.14}$$

Finally, we estimate the first term of (1.14) by the error majorant and obtain for the modeling error estimate from (1.9)

$$(E^{\varepsilon}_{\rm mod})^2 = \|\nabla(u\,-\,u_\varepsilon)\|_{\mathbf{A}} \le \frac{2\,\mu}{2\,\mu-1}\,\rho(\Lambda_\varepsilon + \Lambda_\varepsilon^{-1} - 2I)\,\left(\frac{\mu}{2}\,\mathcal{M}^2_\Omega(u_{\varepsilon,h},\,y,\,\beta) + \int\limits_\Omega f\,u_{\varepsilon,h}\right). \tag{1.15}$$

Hence, $E^{\varepsilon}_{\rm mod}$ can be minimized with respect to $\mu > 1/2$. Straightforward calculations show that $E^{\varepsilon}_{\rm mod}$ has the unique local minimum in

$$\mu = \mu_{min} = \frac{1}{2}\,+\,\left(\frac{1}{4}\,+\,\frac{\int\limits_\Omega f\,u_{\varepsilon,h}}{\mathcal{M}^2_\Omega(u_{\varepsilon,h},\,y,\,\beta)}\right)^{1/2}, \tag{1.16}$$

provided that $\mathcal{M}^2_\Omega(u_{\varepsilon,h},\, y,\, \beta)$ is positive (instead of 0), otherwise

$$(E^\varepsilon_{\mathrm{mod}})^2 \leq \rho(\Lambda_\varepsilon + \Lambda_\varepsilon^{-1} - 2I) \int\limits_\Omega f\, u_{\varepsilon,h},$$

which is also encompassed in (1.16) if we formally set $\mu_{min} = +\infty$. Now, one obtains (1.11) by using (1.16) in (1.15) and setting

$$\kappa_2 := \mu_{min}. \tag{1.17}$$

■

Remark 1.2 *From (1.9), it follows that*

$$\|\nabla(u - u_{\varepsilon,h})\|_{\mathbf{A}} \leq \left(\kappa_1 + \frac{\sqrt{\kappa_2}}{\sqrt{2}}\,\kappa_\varepsilon\right) \mathcal{M}_\Omega(u_{\varepsilon,h},\, y,\, \beta) + \kappa_\varepsilon \left(\int\limits_\Omega f\, u_{\varepsilon,h}\right)^{1/2}. \tag{1.18}$$

The quantity κ_ε measures the approximation quality of the matrix $\mathbf{A}_\varepsilon$ to the original matrix $\mathbf{A}$. In order to obtain a converging algorithm the sequence of simplified interfaces and the correspondingly chosen averaging strategy for the definition of $\mathbf{A}_\varepsilon$ should imply that $\kappa_\varepsilon \to 0$ as $\varepsilon \to 0$. Note that κ_ε is defined as a supremum over local quantities and, hence, its definition directly gives insight in which local parts of the domain the approximation $\mathbf{A}_\varepsilon$ should be improved so that κ_ε becomes smaller. If $\mathbf{A}$ and $\mathbf{A}_\varepsilon$ are diagonal matrices, then $\Lambda_\varepsilon = \{\lambda^\varepsilon_{ij}\}$ is also diagonal and $\lambda^\varepsilon_{ii} = \frac{a_{ii}}{a^\varepsilon_{ii}}$. In this case, we can easily find the quantities

$$\kappa_1^2 = 1 + \rho\,(\Lambda_\varepsilon - I) = 1 + \sup_{x\in\Omega} \max_{i=1,\ldots,d} \frac{|a_{ii}\,(x) - a^\varepsilon_{ii}\,(x)\,|}{a^\varepsilon_{ii}\,(x)}, \tag{1.19}$$

$$\kappa_\varepsilon^2 = \kappa_2\,\rho\left(\Lambda_\varepsilon + \Lambda_\varepsilon^{-1} - 2I\right) = \kappa_2 \sup_{x\in\Omega} \max_{i=1,\ldots,d} \frac{(a_{ii}\,(x) - a^\varepsilon_{ii}\,(x))^2}{a_{ii}\,(x)\,a^\varepsilon_{ii}\,(x)}. \tag{1.20}$$

1.4 Modeling-Discretization Adaptivity and A Posteriori Error Estimation

1.4.1 Localization of the Error Estimation

After the discrete solution $u_{\varepsilon,h}$ as well as a test function $y \in H\,(\Omega, \mathrm{div})$ (cf. Section 1.4.3) are determined, the optimal value of β in the error majorant $\mathcal{M}^2_\Omega(u_{\varepsilon,h}\,, y,\, \beta)$ is given by

$$\beta_{\mathrm{opt}} := \frac{\|\mathbf{A}_\varepsilon\,\nabla u_{\varepsilon,h} - y\|_{\mathbf{A}_\varepsilon^{-1}}}{C_\Omega\,\|\mathrm{div}\; y + f\|_\Omega}. \tag{1.21}$$

For a subset $\omega \subset \Omega$, we define the local quantities

$$\mathcal{M}^2_\omega\,(u_{\varepsilon,h}, y) := (1+\beta_{\mathrm{opt}}) \int_\omega \mathbf{A}_\varepsilon^{-1}\,(\mathbf{A}_\varepsilon\,\nabla u_{\varepsilon,h} - y)\cdot(\mathbf{A}_\varepsilon\,\nabla u_{\varepsilon,h} - y)$$
$$+ \left(1 + \frac{1}{\beta_{\mathrm{opt}}}\right) C^2_\Omega \int_\omega |\mathrm{div}\; y + f|^2\,.$$

In particular, this implies for a finite element mesh $\mathcal{T}_h$ of Ω the additive splitting

$$\mathcal{M}^2_\Omega(u_{\varepsilon,\,h}\,,\,y,\,\beta_{\rm opt}) = \sum_{K\in\mathcal{T}_h} M^2_K\,(u_{\varepsilon,h},y)\,.$$

In order to localize the modeling error we define for a subset $\omega\subset\Omega$ the localized version of ρ for any $B\in L^\infty\left(\Omega,\mathbb{R}^{d\times d}\right)$ by

$$\rho_\omega\,(B) := \operatorname*{ess\,sup}_{x\in\omega}\sup_{v\in\mathbb{R}^d\setminus\{0\}}\frac{B\,(x)\,v\cdot v}{v\cdot v}.$$

Then,

$$\kappa^2_{\varepsilon,\omega} := \frac{2\,\kappa_2}{2\,\kappa_2-1}\rho_\omega(\Lambda_\varepsilon+\Lambda_\varepsilon^{-1}-2I)$$

and, if $\mathcal{T}_h$ is a disjoint partition of Ω we have

$$E^\varepsilon_{\rm mod} = \left(\max_{K\in\mathcal{T}_h}\kappa_{\varepsilon,K}\right)\left(\frac{\kappa_2}{2}\,\mathcal{M}^2_\Omega(u_{\varepsilon,h},\,y,\,\beta)+\int_\Omega f\,u_{\varepsilon,\,h}\right)^{1/2}.$$

In view of Remark 1.2 we obtain the a posteriori estimate

$$\|\nabla\,(u\,-\,u_{\varepsilon,h})\|^2_{\mathbf{A}} \le 2\left(\kappa_1+\frac{\sqrt{\kappa_2}}{\sqrt{2}}\,\kappa_\varepsilon\right)^2\left(\sum_{K\in\mathcal{T}_h}M^2_K\,(u_{\varepsilon,h},y)\right)+2\left(\max_{K\in\mathcal{T}_h}\kappa_{\varepsilon,K}\right)^2\int_\Omega f\,u_{\varepsilon,\,h}$$

1.4.2 Sequence of Simplified Models

The algorithm generates two sequences of simplicial finite element meshes $\left(\mathcal{T}^{\rm disc}_\ell\right)_\ell$ and $\left(\mathcal{T}^{\rm mod}_\ell\right)_\ell$: the mesh $\mathcal{T}^{\rm disc}_\ell$ is employed for the definition of the (for simplicity conforming $\mathbb{P}_1$) finite element spaces $S_\ell\subset H^1_0\,(\Omega)$ and the mesh $\mathcal{T}^{\rm mod}_\ell$ for the definition of the approximation $\mathbf{A}_\ell=\mathbf{A}_{\varepsilon_\ell}$ of the diffusion tensor $\mathbf{A}$. We assume that the discretization meshes are nested, i.e., $\mathcal{T}^{\rm disc}_{\ell+1}$ is a refinement of $\mathcal{T}^{\rm disc}_\ell$ which implies that the finite element spaces S_ℓ are nested

$$S_0\subset S_1\subset\ldots\subset S_\ell\subset\ldots\subset H^1_0\,(\Omega)\,.$$

For simplicity, we assume that the coarsest meshes coincide, $\mathcal{T}^{\rm disc}_0=\mathcal{T}^{\rm mod}_0$, and that the simplices K of the refined mesh $\mathcal{T}^{\rm disc}_\ell$ are linked to a subset $\sigma\,(K)=\{Q_i\}$ of simplices in the corresponding mesh $\mathcal{T}^{\rm mod}_\ell$ as follows:

$$\forall\ell\ge0\quad\forall K\in\mathcal{T}^{\rm disc}_\ell\quad\exists\sigma\,(K)\subset\mathcal{T}^{\rm mod}_\ell\quad\text{s.t.}\quad\begin{cases}K\subset Q & \text{if } \sigma\,(K)=\{Q\}\ \text{ for some } Q\in\mathcal{T}^{\rm mod}_\ell,\\ \overline{K}=\bigcup\limits_{Q\in\sigma(K)}\overline{Q} & \text{otherwise.}\end{cases}$$

This means, either $K\in\mathcal{T}^{\rm disc}_\ell$ is fully contained in some simplex $Q\in\mathcal{T}^{\rm mod}_\ell$ or K is the union of some simplices in $\mathcal{T}^{\rm mod}_\ell$.

The computation on level ℓ then is structured as follows (we replace the indices ε, h, e.g., in $E^\varepsilon_{\rm mod}$ and $u_{\varepsilon,h}$, by the counting index ℓ to avoid double indices and write, e.g., u_ℓ short for $u_{\varepsilon_\ell,h_\ell}$).

1. **Modelling.** For any $Q \in \mathcal{T}_\ell^{\mathrm{mod}}$, an approximation $\mathbf{A}_Q$ of $\mathbf{A}|_Q$ has to be defined. One possible choice which, e.g., is very common for homogenization problems is given by the harmonic integral mean

$$\mathbf{A}_Q := \left(\frac{1}{|Q|} \int_Q \mathbf{A}^{-1} \right)^{-1},$$

while other choices might be preferable in other situations.

2. **Generate System.** The linear system for the Galerkin finite element discretization is generated in the usual fashion by computing the element system matrices simplexwise and then updating the global system matrix. If, for $K \in \mathcal{T}_\ell^{\mathrm{disc}}$, the set $\sigma(K)$ consists of only one element, say, $Q \in \mathcal{T}_\ell^{\mathrm{mod}}$, then for the quadrature over K the simplified diffusion tensor $\mathbf{A}_Q|_K$ is used. If $\sigma(K) \subset \mathcal{T}_\ell^{\mathrm{mod}}$ contains more than one element the integration over K is split into a composite quadrature rule over the simplices in $\sigma(K)$.[1]

3. **Solve.** For the new discretization mesh $\mathcal{T}_\ell^{\mathrm{disc}}$ and corresponding space S_ℓ, the Galerkin solution $u_\ell \in S_\ell$ is computed by solving the linear system corresponding to the diffusion coefficient $\mathbf{A}_\ell$ of the modelling mesh $\mathcal{T}_\ell^{\mathrm{mod}}$.

4. **Error Estimation.** The local quantities

$$\eta_K^{\mathrm{disc}} := \left(\kappa_1 + \frac{\sqrt{\kappa_2}}{\sqrt{2}}\, \kappa_\ell \right) M_K(u_\ell, y) \quad \text{and} \quad \eta_K^{\mathrm{mod}} := \kappa_{\varepsilon,K} \left(\int_\Omega f\, u_{\varepsilon,h} \right)^{1/2}$$

are computed for all $K \in \mathcal{T}_\ell^{\mathrm{disc}}$ as well as the total error majorant

$$E_\ell^{\mathrm{tot}} := \sqrt{\sum_{K \in \mathcal{T}_\ell^{\mathrm{disc}}} \left(\eta_K^{\mathrm{disc}}\right)^2} + \max_{K \in \mathcal{T}_\ell^{\mathrm{disc}}} \eta_K^{\mathrm{mod}}.$$

If E_ℓ^{tot} is smaller than the target accuracy then the solution process is terminated. Otherwise one proceeds with the step "mark".

5. **Mark.** For given threshold parameters $\gamma_{\mathrm{disc}}, \gamma_{\mathrm{mod}} \in (0,1)$ and for $\eta_{\max}^{\mathrm{disc}} := \max_K \eta_K^{\mathrm{disc}}$ and $\eta_{\max}^{\mathrm{mod}} := \max_{K \in \mathcal{T}_\ell^{\mathrm{disc}}} \eta_K^{\mathrm{mod}}$ we define for all $K \in \mathcal{T}_\ell^{\mathrm{disc}}$ the functions disc and mod by

$$\mathrm{disc}(K) := \begin{cases} \text{true} & \text{if } \eta_K^{\mathrm{disc}} \geq \gamma_{\mathrm{disc}} \eta_{\max}^{\mathrm{disc}}, \\ \text{false} & \text{otherwise,} \end{cases} \quad \text{and} \quad \mathrm{mod}(K) := \begin{cases} \text{true} & \text{if } \eta_K^{\mathrm{mod}} \geq \gamma_{\mathrm{mod}} \eta_{\max}^{\mathrm{mod}}, \\ \text{false} & \text{otherwise} \end{cases}$$

and extend $\mathrm{mod}(\cdot)$ to all $Q \in \mathcal{T}_\ell^{\mathrm{mod}}$ via

$$\mathrm{mod}(Q) := \begin{cases} \text{true} & \exists K \in \mathcal{T}_\ell^{\mathrm{disc}} \quad \text{with} \quad \mathrm{mod}(K) = \text{true} \quad \text{s.t.} \quad Q \in \sigma(K) \\ \text{false} & \text{otherwise.} \end{cases}$$

6. **Refine.** The mesh $\mathcal{T}_\ell^{\mathrm{disc}}$ is refined according to the marking $\mathrm{disc}(\cdot)$ and the mesh $\mathcal{T}_\ell^{\mathrm{mod}}$ is refined according to the marking $\mathrm{mod}(\cdot)$ of the simplices $Q \in \mathcal{T}_\ell^{\mathrm{mod}}$.

[1] We emphasize that our error majorant is by no means restricted to discretizations via the Galerkin finite element method. The only requirement is that some conforming approximation $u_\ell \in S_\ell \in H_0^1(\Omega)$ has been computed.

1.4.3 Computation of the Error Majorant

In this section we will explain how the free function $y \in H(\Omega, \mathrm{div})$ and the parameter β in $\mathcal{M}^2_\Omega(u_\ell, y, \beta)$ are determined. This question has been considered in the literature (see, e.g., [36, 40, 42, 46, 45, 50, 58]). Below we will briefly discuss the application to our case. Note that the computational cost for determining y and β has to be balanced with the gain of a sharper a posteriori error estimate.

For given $\mathbf{A}$, $\mathbf{A}_\ell$, f, and C_Ω, the squared majorant $\mathcal{M}^2_\Omega(u_\ell, y, \beta)$ is a quadratic functional. Our goal is to find some $y_\ell \in S^2_\ell$ and $\beta \in \mathbb{R}$ such that $\mathcal{M}^2_\Omega(u_\ell, y_\ell, \beta)$ is close to the minimum over $y \in H(\Omega, \mathrm{div})$ and $\beta \in \mathbb{R}$. Note that the pair $y = \mathbf{A}_\ell \nabla u$ and β as in (1.21) are the minimizers of $\mathcal{M}^2_\Omega(u, y, \beta)$ and motivates the following starting guess in our recursive algorithm. Let $b_{\ell,j}$, $1 \le j \le N_\ell$, denote the $\mathbb{P}_1$ nodal basis ("hat" functions) of the space S_ℓ. Then, the starting guess is given by the *Clément interpolation* of $\mathbf{A}_\ell \nabla u_\ell$, i.e.,

$$y_\ell^{(0)} := \sum_{j=1}^{N_\ell} \alpha_j b_{\ell,j} \quad \text{with} \quad \alpha_j := \frac{1}{|\omega_{\ell,j}|} \int_{\omega_{\ell,j}} \mathbf{A}_\ell \nabla u_\ell \quad \text{and} \quad \omega_{\ell,j} := \operatorname{supp} b_{\ell,j}.$$

The recursion is defined for $\nu = 1, 2, \ldots, \nu_{\max}$ by:

- Compute
$$\beta_\ell^{(\nu)} := \frac{\| \mathbf{A}_\ell \nabla u_\ell - y_\ell^{(\nu-1)} \|_{\mathbf{A}_\ell^{-1}}}{C_\Omega \left\| \mathrm{div}\, y_\ell^{(\nu-1)} + f \right\|_\Omega};$$

- Find $y_\ell^{(\nu)} \in S^2_\ell$ such that

$$\left(1 + \beta_\ell^{(\nu)}\right) \left(y_\ell^{(\nu)}, w\right)_{\mathbf{A}_\ell^{-1}} + \left(1 + \frac{1}{\beta_\ell^{(\nu)}}\right) \left(\mathrm{div}\, y_\ell^{(\nu)}, w\right)_\Omega$$
$$= \left(1 + \beta_\ell^{(\nu)}\right) \left(\mathbf{A}_\ell \nabla u_\ell, w\right)_{\mathbf{A}_\ell^{-1}} - \left(1 + \frac{1}{\beta_\ell^{(\nu)}}\right) C_\Omega^2 (f, w)_\Omega$$

  for all $w \in S^2_\ell$.

Remark 1.3 *Numerical experiments are reported in [53] and show that the choice $\nu_{\max} = 1$ is sufficient for all the considered cases. Note that the global minimization requires the generation and solution of a linear system of dimension $2N$. On the one hand, we expect that the arising computational cost is of the same order as the cost for computing u_ℓ.*

The numerical tests of the combined modeling-discretization a posteriori error majorant show the sharpness of the majorant for various test problems – for the details we refer to [53].

2 Lecture 2: A Posteriori Estimates of the Modeling Error for Elliptic Homogenization Problems

Remark. This part of the lecture notes is a slightly modified and extended version of the paper [54].

2.1 Introduction

In this lecture, we will consider boundary value problems with periodic structures which arise in various applications. Such structures are well known in industry (e.g., in composite materials). Homogenization theory is a well established tool to analyse media with periodic structures. Within the framework of the theory (see, e.g., [18], [28]), the behaviour of a heterogeneous media is described with the help of a certain homogenized problem, which is typically a boundary value problem with smooth coefficients, and the solution of a specially constructed problem with periodic boundary conditions. It has been proved that the functions reconstructed by this procedure converge to the exact solution as the cell size ε tends to zero. Moreover, known a priori error estimates qualified the convergence rate in terms of ε. The goal of this lecture is to derive an a posteriori estimate of the modeling error generated by homogenization, i.e., to estimate the difference between the exact solution of the original problem and its approximation obtained by the corresponding homogenized model. The error majorant employs the solution of the homogenized problem and, thus, is an a posteriori estimate.

The method is based on the theory of functional a posteriori estimates (see [40] - [46]), in which estimates of the difference between the exact solution of boundary value problems and arbitrary functions from the corresponding energy space has been derived by purely functional methods, i.e., without any restriction to a specific discretization. As a result, the estimates contain no mesh dependent constants and are applicable for any function from the corresponding energy space. In [48] - [51] these properties have been used for the analysis of various types of modeling errors. In the previous lecture (see [53]), it was suggested a combined adaptive numerical strategy, which is based on simplification (defeaturing) of problems having complicated and irregular coefficients. This strategy takes into account both, modeling and approximation errors. It was demonstrated that it is efficient for problems having rapidly changing (oscillating but non-periodical) diffusion coefficients.

Here, we consider a different case related to fine periodical structures, i.e., we are concerned with homogenized models of an elliptic boundary value problem with periodical coefficients.

Let $\Omega \subseteq \mathbb{R}^d$ be a bounded domain with Lipschitz boundary $\partial\Omega$, and $\Omega = \bigcup_{\mathbf{i}} \Pi^{\varepsilon}_{\mathbf{i}}$, where

$$\Pi^{\varepsilon}_{\mathbf{i}} = \mathbf{x}_{\mathbf{i}} + \varepsilon\, \widehat{\Pi} = \left\{ \mathbf{x} \in \mathbb{R}^d \;\middle|\; \frac{\mathbf{x} - \mathbf{x}_{\mathbf{i}}}{\varepsilon} \in \widehat{\Pi} \right\},$$

denotes the dilation and translation of the basic "cell" $\hat{\Pi}$, $\mathbf{x}_{\mathbf{i}}$ is the *reference* point of $\Pi^{\varepsilon}_{\mathbf{i}}$. By $\mathbf{x}$ we denote the global (Cartesian) coordinate system in $\mathbb{R}^d$ and by $\mathbf{i} = (i_1, i_2, ...i_d)$ the counting multi-indices for the cells. The notations $\bigcup_{\mathbf{i}}$ and $\sum_{\mathbf{i}}$ are shorthands for the union and summation over all cells. It is assumed that the total number of cells $\Pi^{\varepsilon}_{\mathbf{i}}$ in Ω is bounded from above by the quantity

$$c_0\, \varepsilon^{-d}, \quad \text{where} \quad c_0 = \mathcal{O}(1). \tag{2.1}$$

In the basic cell we use local Cartesian coordinates $\mathbf{y} \in \mathbb{R}^d$. For any $\Pi_{\mathbf{i}}^{\varepsilon}$, local and global coordinates are related by

$$\mathbf{y} = \frac{\mathbf{x} - \mathbf{x_i}}{\varepsilon} \in \widehat{\Pi} \qquad \forall \mathbf{x} \in \Pi_{\mathbf{i}}^{\varepsilon}\ \forall \mathbf{i}.$$

The diffusion matrix in the periodic setting is given via the cell matrix function $\widehat{\mathbf{A}} \in L^{\infty}(\widehat{\Pi},\, \mathbb{R}^{d\times d}_{\rm sym})$, where $\mathbb{R}^{d\times d}_{\rm sym}$ denotes the set of symmetric $d \times d-$ matrices. We assume that

$$c_1|\xi|^2 \,\le\, \widehat{\mathbf{A}}(\mathbf{y})\xi\cdot\xi \le \rho\left(\widehat{\mathbf{A}}\right)|\xi|^2 \qquad \forall\, \xi \in \mathbb{R}^d \quad \forall \mathbf{y} \in \widehat{\Pi}\,\text{a.e.}, \tag{2.2}$$

where $0 < c_1 \le c_2 < \infty$. The global matrix $\mathbf{A}_{\varepsilon}(\mathbf{x})$ defines the periodic structure on Ω

$$\mathbf{A}_{\varepsilon}(\mathbf{x}) := \widehat{\mathbf{A}}\left(\frac{\mathbf{x} - \mathbf{x_i}}{\varepsilon}\right) \qquad \forall \mathbf{x} \in \Pi_{\mathbf{i}}^{\varepsilon}\ \forall \mathbf{i}, \tag{2.3}$$

where ε is a small parameter (geometrical size of a cell). Note that the ellipticity estimate (2.2) for $\widehat{\mathbf{A}}$ is inherited to $\mathbf{A}_{\varepsilon}$.

For $f \in L^2(\Omega)$ we consider the second-order elliptic equation

$$-\operatorname{div}\left(\mathbf{A}_{\varepsilon}\nabla u_{\varepsilon}\right) = f \quad \text{in } \Omega$$

with homogeneous Dirichlet boundary conditions. The corresponding generalized solution is defined by the variational formulation

$$\int_{\Omega} \mathbf{A}_{\varepsilon}\nabla u_{\varepsilon}\cdot\nabla w = \int_{\Omega} f w \qquad \forall w \in H_0^1(\Omega). \tag{2.4}$$

For any $\varepsilon > 0$, the solution $u_{\varepsilon} \in H_0^1(\Omega)$ exists and is unique. It is known (see, e.g., [13], [18], [28]) that there exists a *homogenized matrix* $\mathbf{A}_0 \in \mathbb{R}^{d\times d}_{sym}$ (cf. (2.14)) which satisfies the estimate (2.2) such that

$$u_{\varepsilon} \to u_0 \quad \text{in } L^2(\Omega) \qquad \text{and} \qquad u_{\varepsilon} \rightharpoonup u_0 \quad \text{in } H_0^1(\Omega) \quad \text{for } \varepsilon \to 0,$$

where $u_0 \in H_0^1(\Omega)$ is the solution of the homogenized variational problem

$$\int_{\Omega} \mathbf{A}_0\nabla u_0\cdot\nabla w = \int_{\Omega} f w \qquad \forall w \in H_0^1(\Omega). \tag{2.5}$$

The homogenized problem (2.5) is well studied in the context of asymptotic analysis (see, e.g., [13], [28]). In particular, it was shown that it is possible to find the approximation

$$u_{\varepsilon}^1(\mathbf{x}) \,=\, u_{01}\left(\mathbf{x},\, \frac{\mathbf{x} - \mathbf{x_i}}{\varepsilon}\right) \qquad \forall \mathbf{x} \in \Pi_{\mathbf{i}}^{\varepsilon}\ \forall \mathbf{i}, \tag{2.6}$$

where

$$u_{01}(\mathbf{x},\, \mathbf{y}) \,=\, u_0(\mathbf{x}) \,+\, \varepsilon\, u_1(\mathbf{x},\, \mathbf{y}) \qquad \forall \mathbf{x} \in \Omega,\ \forall \mathbf{y} \in \widehat{\Pi} \tag{2.7}$$

and $u_1(\mathbf{x},\cdot)$ is a $\widehat{\Pi}-$periodic function such that (cf. [28])

$$\|u_{\varepsilon} - u_{\varepsilon}^1\|_{H^1(\Omega)} \,\le\, c\,\sqrt{\varepsilon}.$$

Derivation of error indicators for homogenized problems is a topic of vivid research. Here, we first of all mention residual type error indicators that develop the ideas suggested in [7, 8] for finite element approximations. Since our approach is based on a different technique, we will sketch here only briefly some relevant literature on residual based estimation and refer for a detailed review, e.g., to [25]. A posteriori error estimates for the heterogeneous multiscale discretization (HMM) of elliptic problems in a periodic setting can be found in [38] and [25]. In [1], an a posteriori estimate of residual type for general, possibly non-periodic, diffusion tensors with micro-scales is presented while a residual-type a posteriori error estimate for more general diffusion tensors has been developed in [25]. Also, we mention the papers [6, 10, 15, 16, 37, 56, 57], which are closely related to the topic.

Our goal is to deduce estimates of a different type, which provide guaranteed upper bounds of the modeling error and does not contain unknown constants. This *error majorants* reflects the decomposition (2.7). The majorant is based on the homogenized problem and its solution and, in addition, depends on free functions defined on the cell of periodicity. They should be chosen such that the majorant becomes as small as possible and can either be computed as the solution of a certain boundary value problem with periodic boundary conditions on the basic cell or by minimizing the error majorant. In general, the estimate has the form

$$\|\nabla \left(u_\varepsilon - u_\varepsilon^1\right)\|_{\mathbf{A}_\varepsilon} \;\leq\; \mathcal{M}_\oplus \left(u_\varepsilon^1;\; \boldsymbol{\eta},\, \boldsymbol{\lambda},\, s\right), \tag{2.8}$$

where

$$\|\boldsymbol{q}\|_{\mathbf{A}_\varepsilon} := \left(\int_\Omega \mathbf{A}_\varepsilon\, \boldsymbol{q}\cdot\boldsymbol{q}\right)^{1/2}. \tag{2.9}$$

The majorant $\mathcal{M}_\oplus$ depends on the solution of (2.5), the small parameter ε, and some other functions, defined on $\widehat{\Pi}$. Technically the derivation is based on a posteriori error estimates of functional type (see, e.g., [41]-[48]).

The structure of this lecture is as follows. In Section 2.2, we briefly overview the results in the homogenization theory of second order elliptic operators which are significant for subsequent analysis. In Section 2.2.1, we prove the main result, which yields computable upper and lower bounds of the modeling error. Numerical experiments have been performed and are reported in [54] which underpin the sharpness of the derived estimates.

2.2 Homogenization of second order elliptic operators

As a notation we associate to a sufficiently smooth function $\widehat{v} : \Omega \times \widehat{\Pi} \to \mathbb{R}$ the periodic version by

$$v\left(\mathbf{x}\right) := \widehat{v}\left(\mathbf{x},\, \frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right) \qquad \forall \mathbf{x} \in \Pi_i^\varepsilon \quad \forall \mathbf{i}.$$

On each cell $\Pi_{\mathbf{i}}^\varepsilon$, the operator $-\operatorname{div}\left(\mathbf{A}_\varepsilon \nabla\right)$ can be represented in a different form:

$$-\operatorname{div}\left(\mathbf{A}_\varepsilon \nabla v\right)\left(\mathbf{x}\right) = \left(\tilde{\mathcal{A}}_\varepsilon \widehat{v}\right)\left(\mathbf{x},\, \frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right) \qquad \forall \mathbf{x} \in \Pi_{\mathbf{i}}^\varepsilon \quad \forall \mathbf{i}.$$

To define the operator $\tilde{\mathcal{A}}_\varepsilon$ we need some notation. We write $\nabla_{\mathbf{x}}$, $\operatorname{div}_{\mathbf{x}}$ to indicate differentiation with respect to the "domain variable" $\mathbf{x} \in \Omega$ and $\nabla_{\mathbf{y}}$, $\operatorname{div}_{\mathbf{y}}$ for differentiation with respect to the "cell variable" $\mathbf{y} \in \widehat{\Pi}$. The notation $\nabla_{\mathbf{x}} \nabla_{\mathbf{x}}^{\intercal}$ is short for the matrix $\left(\frac{\partial^2}{\partial_{x_i}\partial_{x_j}}\right)_{i,j=1}^{d}$ and

$\nabla_{\mathbf{x}}\nabla_{\mathbf{y}}^{\intercal}$ and $\nabla_{\mathbf{y}}\nabla_{\mathbf{y}}^{\intercal}$ are defined analogously. Recall that the operator $\widehat{\mathbf{A}}$ is defined on $\widehat{\Pi}$ and hence solely depends on the cell variable. For $d \times d$ matrices $\mathbf{A}$ and $\mathbf{B}$, we write $\mathbf{A} : \mathbf{B}$ for the Euclidean product of matrices.

With this notation at hand the operator $\tilde{\mathcal{A}}_\varepsilon$ can be written in the form

$$\tilde{\mathcal{A}}_\varepsilon = -\left(\mathrm{div}_{\mathbf{x}} + \varepsilon^{-1}\,\mathrm{div}_{\mathbf{y}}\right)\left(\widehat{\mathbf{A}}\,\left(\nabla_{\mathbf{x}} + \varepsilon^{-1}\nabla_{\mathbf{y}}\right)\right) = \varepsilon^{-2}\mathcal{A}_1 + \varepsilon^{-1}\mathcal{A}_2 + \mathcal{A}_3,$$

and

$$\mathcal{A}_1 = -\,\mathrm{div}_{\mathbf{y}}\left(\widehat{\mathbf{A}}\nabla_{\mathbf{y}}\right), \quad \mathcal{A}_2 = -\,\mathrm{div}_{\mathbf{y}}\left(\widehat{\mathbf{A}}\nabla_{\mathbf{x}}\right) - \mathrm{div}_{\mathbf{x}}\left(\widehat{\mathbf{A}}\nabla_{\mathbf{y}}\right), \quad \mathcal{A}_3 = -\,\mathrm{div}_{\mathbf{x}}\left(\hat{\mathbf{A}}\nabla_{\mathbf{x}}\right).$$

Within the framework of the homogenization theory, the construction of an efficient approximation of the desired function u_ε is based on the form (2.6) - (2.7). It holds

$$-\,\mathrm{div}\left(\mathbf{A}_\varepsilon\,\nabla u_\varepsilon^1\right)(\mathbf{x}) =: \left(\tilde{\mathcal{A}}_\varepsilon u_{0,1}\right)\left(\mathbf{x}, \frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right) \qquad \forall \mathbf{x} \in \Pi_i^\varepsilon \quad \forall \mathbf{i},$$

where

$$\tilde{\mathcal{A}}_\varepsilon u_{01} = \varepsilon^{-1}\left(\mathcal{A}_1\,u_1 + \mathcal{A}_2\,u_0\right) + \left(\mathcal{A}_3\,u_0 + \mathcal{A}_2\,u_1\right) + \varepsilon\,\mathcal{A}_3\,u_1.$$

The natural requirement "$\tilde{\mathcal{A}}_\varepsilon u_\varepsilon^1$ must be uniformly bounded as ε tends to zero" leads to the condition $\mathcal{A}_1\,u_1 + \mathcal{A}_2\,u_0 = 0$. Hence,

$$-\,\mathrm{div}_{\mathbf{y}}\left(\widehat{\mathbf{A}}\nabla_{\mathbf{y}}u_1\right) = \mathrm{div}_{\mathbf{y}}\left(\widehat{\mathbf{A}}\nabla_{\mathbf{x}}u_0\right).$$

This equation is considered as a problem on the cell of periodicity depending on the domain variable $\mathbf{x} \in \Omega$ as a parameter.

In what follows, we assume that u_0 (solution of (2.5)) belongs to $H^2(\Omega)$. It is well known (e.g., [24]) that this assumption holds if $f \in L^2(\Omega)$ and, e.g., Ω is a bounded domain with a smooth boundary or (in the case $d = 2$) Ω is a convex bounded Lipschitz domain.

Let $\mathbf{N} = (N_k)_{k=1}^d$ be the unique solution of the auxiliary problem ($\widehat{\mathbf{a}}_k$ denotes the k-th row of the Matrix $\widehat{\mathbf{A}}$.)

$$\begin{aligned} \mathrm{div}_{\mathbf{y}}\,\widehat{\mathbf{A}}\nabla_{\mathbf{y}}N_k = \mathrm{div}_{\mathbf{y}}\,\widehat{\mathbf{a}}_k \quad &\text{in} \quad \widehat{\Pi}, \\ N_k \text{ satisfies periodic boundary conditions,} \\ \int_{\widehat{\Pi}} N_k = 0. \end{aligned} \tag{2.10}$$

Then $u_1(\mathbf{x}, \mathbf{y})$ can be written (cf., e.g., [11], [13], [28]) as

$$u_1(\mathbf{x}, \mathbf{y}) = -\left\langle \mathbf{N}(\mathbf{y}), \nabla_{\mathbf{x}}u_0(\mathbf{x})\right\rangle.$$

Therefore, u_ε^1 as defined in (2.6) has the form

$$u_\varepsilon^1(\mathbf{x}) = u_0(\mathbf{x}) - \varepsilon\left\langle \mathbf{N}\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right), \nabla_{\mathbf{x}}u_0(\mathbf{x})\right\rangle, \quad \forall \mathbf{y} \in \widehat{\Pi} \quad \forall \mathbf{x} \in \Pi_{\mathbf{i}}^\varepsilon\ \forall \mathbf{i}. \tag{2.11}$$

We always consider $\mathbf{N}$ and $\widehat{\mathbf{A}}$ as functions of $\mathbf{y} \in \widehat{\Pi}$ and u_0 as a function which depends solely on $\mathbf{x} \in \Omega$. Then, somewhat tedious calculations yield

$$\mathcal{A}_3\,u_0 + \mathcal{A}_2\,u_1 = \left(-\widehat{\mathbf{A}} + \widehat{\mathbf{A}}\nabla_{\mathbf{y}}\mathbf{N}^{\intercal}\right) : \nabla_{\mathbf{x}}\nabla_{\mathbf{x}}^{\intercal}u_0 + \mathrm{div}_{\mathbf{y}}\left(\widehat{\mathbf{A}}\nabla_{\mathbf{x}}\nabla_{\mathbf{x}}^{\intercal}u_0\mathbf{N}\right). \tag{2.12}$$

Let $\omega \subset \Omega$ be a measurable subset. For a function $\zeta \in L^1(\omega)$ the integral mean is given by

$$\langle \zeta \rangle_\omega := \frac{1}{|\omega|} \int_\omega \zeta. \tag{2.13}$$

If we write $\int_\omega \langle \zeta \rangle_\omega$ we consider this average as a constant function on ω (for vector-valued functions, we apply this definition componentwise). We denote the error caused by the average (2.13) by

$$\delta_\omega \zeta := \|\zeta - \langle \zeta \rangle_\omega\|_\omega,$$

where $\|\cdot\|_\omega$ denotes the standard L^2-norm on ω. For vector-valued functions $\boldsymbol{\zeta} = (\zeta_k)_{k=1}^d \in L^1\left(\omega, \mathbb{R}^d\right)$ and $\boldsymbol{\phi} = (\phi_k)_{k=1}^d \in L^1\left(\Omega, \mathbb{R}^d\right)$ we define the local and piecewise averages by

$$\delta_\omega \boldsymbol{\zeta} := \left(\|\zeta_k - \langle \zeta_k \rangle_\omega\|_\omega\right)_{k=1}^d, \qquad \delta_\Omega^{\mathrm{pw}} \boldsymbol{\phi} := \varepsilon^{d/2} \left(\sum_{\mathbf{i}} \|\phi_k - \langle \phi_k \rangle_{\Pi_\mathbf{i}^\varepsilon}\|_{\Pi_\mathbf{i}^\varepsilon} \right)_{k=1}^d$$

and

$$(\delta_\omega \boldsymbol{\zeta})^2 := \left(\|\zeta_k - \langle \zeta_k \rangle_\omega\|_\omega^2\right)_{k=1}^d, \qquad (\delta_\Omega^{\mathrm{pw}} \boldsymbol{\phi})^2 := \varepsilon^d \left(\left(\sum_{\mathbf{i}} \|\phi_k - \langle \phi_k \rangle_{\Pi_\mathbf{i}^\varepsilon}\|_{\Pi_\mathbf{i}^\varepsilon} \right)^2 \right)_{k=1}^d.$$

The mean value of the right-hand side of (2.12) with respect to $\mathbf{y}$ is given by

$$\langle \mathcal{A}_3\, u_0 + \mathcal{A}_2\, u_1 \rangle_{\widehat{\Pi}} = - \left\langle \widehat{\mathbf{A}} \left(\mathbf{I} - \nabla_\mathbf{y} \mathbf{N}^\intercal \right) \right\rangle_{\widehat{\Pi}} : \nabla_\mathbf{x} \nabla_\mathbf{x}^\intercal u_0 =: -\mathbf{A}_0 : \nabla_\mathbf{x} \nabla_\mathbf{x}^\intercal u_0 = - \operatorname{div}_\mathbf{x} \left(\mathbf{A}_0 \nabla_\mathbf{x} u_0 \right),$$

since, due to the periodicity of $\widehat{\mathbf{A}}$ and $\mathbf{N}$, the integral mean over the last term in (2.12) vanishes as a consequence of Gauss' theorem (with $\mathbf{n}$ denoting the outward normal to $\widehat{\Pi}$)

$$\int_{\widehat{\Pi}} \operatorname{div}_\mathbf{y} \left(\widehat{\mathbf{A}} \nabla_\mathbf{x} \nabla_\mathbf{x}^\intercal u_0 \mathbf{N} \right) = \int_{\widehat{\Pi}} \left\langle \mathbf{n}, \widehat{\mathbf{A}} \nabla_\mathbf{x} \nabla_\mathbf{x}^\intercal u_0 \mathbf{N} \right\rangle = 0.$$

The homogenization matrix is given by

$$\mathbf{A}_0 = \left\langle \widehat{\mathbf{A}} \left(\mathbf{I} - \nabla_\mathbf{y} \mathbf{N}^\intercal \right) \right\rangle_{\widehat{\Pi}} \tag{2.14}$$

with the solution $\mathbf{N} := (N_k)_{k=1}^d$ of the cell problem (2.10). In general, u_ε^1 defined by (2.11) does not satisfy the boundary conditions. We introduce the boundary corrected approximation w_ε^1 of u_ε by

$$w_\varepsilon^1(\mathbf{x}) := u_0(\mathbf{x}) \; - \; \varepsilon \psi^\varepsilon(\mathbf{x}) \left\langle \mathbf{N} \left(\frac{\mathbf{x} - \mathbf{x}_i}{\varepsilon} \right), \nabla u_0\left(\mathbf{x}\right) \right\rangle \quad \forall \mathbf{x} \in \Pi_\mathbf{i}^\varepsilon \; \forall \mathbf{i}, \tag{2.15}$$

where the cutoff function

$$\psi^\varepsilon(\mathbf{x}) := \min\{1, \frac{1}{\varepsilon} \operatorname{dist}(\mathbf{x}, \partial\Omega)\}.$$

satisfies the following conditions (for Lipschitz domains Ω):

$$\begin{array}{ll} \psi^\varepsilon \in W_0^{1,\infty}(\Omega), & \psi^\varepsilon \equiv 1 \text{ in } \Omega_\varepsilon^{in} := \{\, x \in \Omega \mid \operatorname{dist}(x, \partial\Omega) > \varepsilon\}, \\ 0 \le \psi^\varepsilon \le 1, & \varepsilon \left|\nabla \psi^\varepsilon\right| \le c \text{ in } \Omega \text{ for some } c \text{ independent of } \varepsilon. \end{array} \tag{2.16}$$

We summarize the three steps for computing the augmented approximation w_ε^1 of u_ε below. Note that our error majorant will depend on this precomputed function.

i) The solutions N_k of the cell problems

$$\begin{aligned} &\operatorname{div}\left(\widehat{\mathbf{A}}\,\nabla N_k\right) = \operatorname{div}\widehat{\mathbf{a}}_k \quad \text{in} \quad \widehat{\Pi}, \\ &N_k \text{ is periodic in } \widehat{\Pi}, \\ &\int_{\widehat{\Pi}} N_k = 0, \end{aligned} \tag{2.17}$$

have to be computed which allow to determine the homogenized matrix in the general case (cf. (2.14)):

$$\mathbf{A}_0 = \left\langle \widehat{\mathbf{A}}\,(\mathbf{I} - \nabla\mathbf{N}) \right\rangle_{\widehat{\Pi}}.$$

ii) The homogenized problem has to be solved: Find $u_0 \in H_0^1(\Omega)$ such that

$$\int_\Omega \mathbf{A}_0 \nabla u_0 \cdot \nabla w \;=\; \int_\Omega f\, w \qquad \forall w \in H_0^1(\Omega). \tag{2.18}$$

iii) With the help of u_0 and N_k, we obtain the approximation w_ε^1 of u_ε via

$$w_\varepsilon^1(\mathbf{x}) \;:=\; u_0(\mathbf{x}) \;-\; \varepsilon\psi^\varepsilon(\mathbf{x}) \left\langle \mathbf{N}\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right), \nabla_{\mathbf{x}} u_0(\mathbf{x}) \right\rangle \quad \forall \mathbf{x} \in \Pi_{\mathbf{i}}^\varepsilon \;\forall \mathbf{i} \tag{2.19}$$

with the cutoff function $\psi^\varepsilon := \min\{1, \frac{1}{\varepsilon}\operatorname{dist}(\mathbf{x}, \partial\Omega)\}$.

It is proved (see, e.g., in [20, 28]) that the following error estimate holds:

$$\|u_\varepsilon - w_\varepsilon^1\|_{H^1(\Omega)} \;\le\; \widetilde{c}\sqrt{\varepsilon}. \tag{2.20}$$

Relation (2.20) provides an a priori estimate of the modeling error evaluated in terms of the parameter ε. In the next section, we deduce a guaranteed a posteriori error majorant of

$$\|\nabla(u_\varepsilon - w_\varepsilon^1)\|_{A_\varepsilon}$$

which employ the computed functions N_k as well as the homogenized solution u_0.

2.2.1 Error estimate of the modeling error

In this section, we first prove a subsidiary result which states an upper bound of the L^2-product of a globally defined function and a periodic function defined on the cell. For a vector $\boldsymbol{\mu} = (\mu_i)_{i=1}^d \in (\mathbb{R}_{>0})^d$ and $s \in \mathbb{R}$ we denote by $\boldsymbol{\mu}^s$ the componentwise application of the power s, i.e., $\boldsymbol{\mu}^s = (\mu_i^s)_{i=1}^d$.

Lemma 2.1 *For all $\mathbf{g} \in L^2(\Omega)^d$, $\boldsymbol{\eta} \in L^2(\widehat{\Pi})^d$, and all $\boldsymbol{\lambda} = (\lambda_d)_{k=1}^d \in (\mathbb{R}_{>0})^d$ it holds*

$$\sum_{\mathbf{i}} \int_{\Pi_{\mathbf{i}}^\varepsilon} \mathbf{g}(\mathbf{x}) \cdot \boldsymbol{\eta}\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right) \mathbf{dx} \;\le |\Omega|\,\langle\mathbf{g}\rangle_\Omega \cdot \langle\boldsymbol{\eta}\rangle_{\widehat{\Pi}} \;+\; \frac{\boldsymbol{\lambda}}{2} \cdot (\delta_\Omega^{\mathrm{pw}}\mathbf{g})^2 \;+\; \frac{\boldsymbol{\lambda}^{-1}}{2} \cdot (\delta_{\widehat{\Pi}}\boldsymbol{\eta})^2. \tag{2.21}$$

Proof. For any $\mathbf{g} \in L^2(\Omega)^d$, we have

$$\mathcal{I} := \sum_{\mathbf{i}} \int_{\Pi_{\mathbf{i}}^\varepsilon} \mathbf{g}(\mathbf{x}) \cdot \boldsymbol{\eta}\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right) \mathbf{dx} = \sum_{k=1}^{d} \sum_{\mathbf{i}} \int_{\Pi_{\mathbf{i}}^\varepsilon} g_k(\mathbf{x})\, \eta_k\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right) \mathbf{dx}$$
$$= \sum_{k=1}^{d} \sum_{\mathbf{i}} \int_{\Pi_{\mathbf{i}}^\varepsilon} \left(g_k(\mathbf{x}) - \langle g_k\rangle_{\Pi_{\mathbf{i}}^\varepsilon}\right) \eta_k\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right) \mathbf{dx} \;+\; \sum_{k=1}^{d} \sum_{\mathbf{i}} \int_{\Pi_{\mathbf{i}}^\varepsilon} \langle g_k\rangle_{\Pi_{\mathbf{i}}^\varepsilon}\, \eta_k\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right) \mathbf{dx}.$$

Since

$$\sum_{\mathbf{i}} \int_{\Pi_{\mathbf{i}}^\varepsilon} \langle g_k\rangle_{\Pi_{\mathbf{i}}^\varepsilon}\, \eta_k\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right) \mathbf{dx} = \varepsilon^d \int_{\widehat{\Pi}} \eta_k \sum_{\mathbf{i}} \frac{1}{|\Pi_{\mathbf{i}}^\varepsilon|} \int_{\Pi_{\mathbf{i}}^\varepsilon} g_k \;=\; \varepsilon^d \int_{\widehat{\Pi}} \eta_k \sum_{\mathbf{i}} \frac{1}{\varepsilon^d\,|\widehat{\Pi}|} \int_{\Pi_{\mathbf{i}}^\varepsilon} g_k$$
$$= \int_{\widehat{\Pi}} \eta_k\, \frac{1}{|\widehat{\Pi}|} \int_{\Omega} g_k \;=\; |\Omega|\, \langle g_k\rangle_{\Omega}\, \langle \eta_k\rangle_{\widehat{\Pi}}$$

and

$$\sum_{\mathbf{i}} \int_{\Pi_{\mathbf{i}}^\varepsilon} \left(g_k(\mathbf{x}) - \langle g_k\rangle_{\Pi_{\mathbf{i}}^\varepsilon}\right) \eta_k\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right) \mathbf{dx} = \sum_{\mathbf{i}} \int_{\Pi_{\mathbf{i}}^\varepsilon} \left(g_k(\mathbf{x}) - \langle g_k\rangle_{\Pi_{\mathbf{i}}^\varepsilon}\right) \left(\eta_k\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right) - c_k\right) \mathbf{dx}$$
$$\le \left(\sum_{\mathbf{i}} \|g_k - \langle g_k\rangle_{\Pi_{\mathbf{i}}^\varepsilon}\|_{\Pi_{\mathbf{i}}^\varepsilon}\right) \left(\int_{\Pi_{\mathbf{i}}^\varepsilon} \left(\eta_k\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right) - c_k\right)^2 \mathbf{dx}\right)$$
$$= \left(\sum_{\mathbf{i}} \|g_k - \langle g_k\rangle_{\Pi_{\mathbf{i}}^\varepsilon}\|_{\Pi_{\mathbf{i}}^\varepsilon}\right) \;\varepsilon^{d/2}\, \|\eta_k - c_k\|_{\widehat{\Pi}},$$

we find that

$$\mathcal{I} \le \sum_{k} \left(|\Omega|\, \langle g_k\rangle_{\Omega}\, \langle \eta_k\rangle_{\widehat{\Pi}} \;+\; (\delta_{\Omega}^{\mathrm{pw}} g)_k \;\|\eta_k - c_k\|_{\widehat{\Pi}}\right)$$
$$\le \sum_{k} \left(|\Omega|\, \langle g_k\rangle_{\Omega}\, \langle \eta_k\rangle_{\widehat{\Pi}} \;+ \frac{\lambda_k}{2}\, (\delta_{\Omega}^{\mathrm{pw}} g)_k^2 \;+ \frac{1}{2\,\lambda_k}\, \|\eta_k - c_k\|_{\widehat{\Pi}}^2\right)$$
$$= |\Omega|\, \langle \mathbf{g}\rangle_{\Omega} \cdot \langle \boldsymbol{\eta}\rangle_{\widehat{\Pi}} \;+\; \frac{1}{2}\, \boldsymbol{\lambda} \cdot (\delta_{\Omega}^{\mathrm{pw}} \mathbf{g})^2 \;+\; \frac{1}{2} \int_{\widehat{\Pi}} \sum_{k} \frac{1}{\lambda_k}\, (\eta_k - c_k)^2,$$

where $\boldsymbol{\lambda} \in \mathbb{R}^d_{>0}$ and $(c_k)_{k=1}^d \in \mathbb{R}^d$ are arbitrary vectors. In particular, we set $c_k = \langle \eta_k\rangle_{\widehat{\Pi}}$, and this implies (2.21). ∎

Let $\nabla\nabla^{\intercal} u_0\,(\mathbf{x})$ and $\nabla \boldsymbol{N}^{\intercal}$ denote the Hessian matrix of u_0 and the Jacobian matrix of vector $\boldsymbol{N}$, respectively. In order to present the main estimate in a transparent form, we define the function

$$\mathbf{G} := \left(\nabla w_\varepsilon^1 - \mathbf{A}_\varepsilon^{-1}\mathbf{A}_0 \nabla u_0\right) = \left(\left(\mathbf{I} - \mathbf{A}_\varepsilon^{-1}\mathbf{A}_0\right) \nabla u_0 - \varepsilon \nabla \left(\psi^\varepsilon \mathbf{N}\left(\frac{\cdot - \mathbf{x_i}}{\varepsilon}\right) \cdot \nabla u_0\right)\right) \qquad \forall \mathbf{i}. \tag{2.22}$$

These functions allows to define part of the error majorant

$$\mathcal{F}\left(w_\varepsilon^1;\; \boldsymbol{\eta}, \boldsymbol{\lambda},\, s\right) := \|\mathbf{G}\|^2_{\mathbf{A}_\varepsilon} + 2\,\varepsilon^s\, |\Omega|\, \langle \mathbf{G}\rangle_{\Omega} \cdot\, \langle \boldsymbol{\eta}\rangle_{\widehat{\Pi}} +$$
$$+ \varepsilon^s \left(\boldsymbol{\lambda}^{-1} \cdot (\delta_{\widehat{\Pi}} \boldsymbol{\eta})^2 \;+\; \boldsymbol{\lambda} \cdot (\delta_{\Omega}^{\mathrm{pw}} \mathbf{G})^2\right) + c_0 \varepsilon^{2s}\; \|\boldsymbol{\eta}\|^2_{\widehat{\mathbf{A}}^{-1},}, \tag{2.23}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^d_{>0},\ s \in \mathbb{R}$,

$$\boldsymbol{\eta} \in H_0(\widehat{\Pi},\, \mathrm{div}) := \left\{ \boldsymbol{\vartheta} \in H(\widehat{\Pi},\, \mathrm{div}),\ \langle \mathrm{div}\ \boldsymbol{\vartheta} \rangle_{\widehat{\Pi}} = 0 \right\}$$

and

$$H(\widehat{\Pi},\, \mathrm{div}) := \left\{ \boldsymbol{\vartheta} \in \left(L^2\left(\widehat{\Pi}\right)\right)^d,\, \mathrm{div}\ \boldsymbol{\vartheta} \in L^2\left(\widehat{\Pi}\right) \right\}.$$

Now, we formulate of our main result.

Theorem 2.2 *Let $\mathbf{A}_\varepsilon$ be defined by (2.3) and let (2.1), (2.2) be satisfied. Let the reference cell $\widehat{\Pi}$ be convex. We assume that the right-hand side in (2.4) satisfies $f \in L^2(\Omega)$ and u_ε denotes the exact solution. The solution u_0 of the homogenized problem is required to be in $H^2(\Omega)$. The approximation defined by (2.19) with ψ^ε so that (2.16) holds is denoted by w^1_ε. Then, the error $u_\varepsilon - w^1_\varepsilon$ can be estimated by*

$$\|\nabla(u_\varepsilon - w^1_\varepsilon)\|_{\mathbf{A}_\varepsilon} \ \le\ \mathcal{M}_\oplus\left(w^1_\varepsilon,\, \boldsymbol{\eta},\, \boldsymbol{\lambda},\, s\right) := \mathcal{F}^{1/2}\left(w^1_\varepsilon;\ \boldsymbol{\eta}, \boldsymbol{\lambda},\, s\right)\ +\ \varepsilon^s\, \widetilde{C}\, \|\,\mathrm{div}\ \boldsymbol{\eta}\|_{\widehat{\Pi}}, \tag{2.24}$$

where $\mathcal{F}$ is defined by (2.23). The quantities $\boldsymbol{\eta} \in H_0(\widehat{\Pi},\, \mathrm{div})$, $\boldsymbol{\lambda} \in \mathbb{R}^d_{>0}$, and $s \in \mathbb{R}$ are free parameters and the constant $\widetilde{C}$ is defined by (2.29).

Proof. For any $v, w \in H^1_0(\Omega)$ and $\boldsymbol{\tau} \in H(\Omega, \mathrm{div})$, we have

$$\begin{aligned}\int_\Omega \mathbf{A}_\varepsilon \nabla(u_\varepsilon - v)\cdot\nabla w &= \int_\Omega (-\mathbf{A}_\varepsilon \nabla v\cdot\nabla w + f\, w)\\ &= \int_\Omega (\boldsymbol{\tau} - \mathbf{A}_\varepsilon\nabla v)\cdot\nabla w\ +\ \int_\Omega (\mathrm{div}\,\boldsymbol{\tau} + f)\, w.\end{aligned} \tag{2.25}$$

We set $w = u_\varepsilon - v$ and estimate the first term in (2.25) as follows:

$$\int_\Omega (\boldsymbol{\tau} - \mathbf{A}_\varepsilon\nabla v)\cdot\nabla(u_\varepsilon - v) \le \|\nabla(u_\varepsilon - v)\|_{\mathbf{A}_\varepsilon}\, \|\mathbf{A}_\varepsilon\nabla v - \boldsymbol{\tau}\|_{\mathbf{A}^{-1}_\varepsilon}. \tag{2.26}$$

We assume that $\boldsymbol{\tau}$, on any $\Pi^\varepsilon_{\mathbf{i}}$, is of the form

$$\boldsymbol{\tau}(\mathbf{x}) = \boldsymbol{\tau}_0(\mathbf{x}) - \varepsilon^s \boldsymbol{\eta}\left(\frac{\mathbf{x} - \mathbf{x_i}}{\varepsilon}\right) \tag{2.27}$$

with

$$\mathrm{div}\ \boldsymbol{\tau}_0 = -f \quad \text{and} \quad \boldsymbol{\eta} \in H_0\left(\widehat{\Pi}, \mathrm{div}\right).$$

Since

$$\mathrm{div}\ \boldsymbol{\tau}\,(\mathbf{x}) = \mathrm{div}\ \boldsymbol{\tau}_0\,(\mathbf{x}) - \varepsilon^{s-1}\,(\mathrm{div}\,\boldsymbol{\eta})\left(\frac{\mathbf{x} - \mathbf{x_i}}{\varepsilon}\right) = -f - \varepsilon^{s-1}\,(\mathrm{div}\,\boldsymbol{\eta})\left(\frac{\mathbf{x} - \mathbf{x_i}}{\varepsilon}\right) \quad \forall \mathbf{x} \in \Pi^\varepsilon_{\mathbf{i}}\ \forall \mathbf{i}$$

and

$$\left\langle (\mathrm{div}\,\boldsymbol{\eta})\left(\frac{\cdot - \mathbf{x_i}}{\varepsilon}\right)\right\rangle_{\Pi^\varepsilon_{\mathbf{i}}} = \varepsilon^d\, \langle \mathrm{div}\,\boldsymbol{\eta}\rangle_{\widehat{\Pi}} = 0$$

we obtain

$$\int_\Omega (\operatorname{div}\boldsymbol{\tau}+f)\,(u_\varepsilon - v) = -\sum_{\mathbf{i}} \int_{\Pi_{\mathbf{i}}^\varepsilon} \varepsilon^{s-1}\,(\operatorname{div}\boldsymbol{\eta})\left(\frac{\cdot-\mathbf{x_i}}{\varepsilon}\right)\,(u_\varepsilon - v)$$
$$\leq \varepsilon^s \sum_{\mathbf{i}} \varepsilon^{d/2-1} \|\operatorname{div}\boldsymbol{\eta}\|_{\widehat{\Pi}}\; C_{\Pi_{\mathbf{i}}^\varepsilon}\, \|\nabla(u_\varepsilon - v)\|_{\Pi_{\mathbf{i}}^\varepsilon},$$

where $C_{\Pi_{\mathbf{i}}^\varepsilon}$ is the constant in the Poincare's inequality for $\Pi_{\mathbf{i}}^\varepsilon$. For convex domains, $C_{\Pi_{\mathbf{i}}^\varepsilon} \leq \frac{\operatorname{diam}\Pi_{\mathbf{i}}^\varepsilon}{\pi}$ (for $d = 1,\, 2,\, 3$) (cf. [39]) and $\operatorname{diam}\Pi_{\mathbf{i}}^\varepsilon = \varrho\varepsilon$ for some $\varrho = O(1)$ depending on d and geometric properties of the basic cell (if, e.g., the cell is a cube, then $\varrho = \sqrt{d}$).

We use (2.1) and arrive at the estimate

$$\int_\Omega (\operatorname{div}\boldsymbol{\tau}+f)\,(u_\varepsilon - v) \;\leq \varepsilon^s\, \varepsilon^{\frac{d}{2}-1}\, \|\operatorname{div}\boldsymbol{\eta}\|_{\widehat{\Pi}}\, \sqrt{c_0}\, \varepsilon^{-\frac{d}{2}}\, \varepsilon\, \frac{\varrho}{\pi}\, \|\nabla(u_\varepsilon - v)\|$$
$$= \varepsilon^s \frac{\varrho}{\pi}\, \sqrt{c_0}\, \|\operatorname{div}\boldsymbol{\eta}\|_{\widehat{\Pi}}\, \|\nabla(u_\varepsilon - v)\|.$$

In view of (2.2), we obtain

$$\int_\Omega (\operatorname{div}\boldsymbol{\tau}+f)\,(u_\varepsilon - v) \;\leq \varepsilon^s\, \widetilde{C}\, \|\operatorname{div}\boldsymbol{\eta}\|_{\widehat{\Pi}}\, \|\nabla(u_\varepsilon - v)\|_{\mathbf{A}_\varepsilon}, \tag{2.28}$$

where

$$\widetilde{C} = \frac{\varrho}{\pi}\sqrt{\frac{c_0}{c_1}}. \tag{2.29}$$

Now (2.25), (2.26), and (2.28) imply the estimate

$$\|\nabla(u_\varepsilon - v)\|_{\mathbf{A}_\varepsilon} \;\leq\; \|\mathbf{A}_\varepsilon\nabla v - \boldsymbol{\tau}\|_{\mathbf{A}_\varepsilon^{-1}} \;+\; \varepsilon^s\, \widetilde{C}\, \|\operatorname{div}\boldsymbol{\eta}\|_{\widehat{\Pi}}. \tag{2.30}$$

Consider the first term in the right-hand side of the estimate (2.30) and set

$$v := w_\varepsilon^1 \quad \text{and} \quad \boldsymbol{\tau}_0 := \mathbf{A}_0\nabla u_0$$

It holds

$$\nabla w_\varepsilon^1 = \nabla u_0 - \varepsilon\nabla\left(\psi^\varepsilon \mathbf{N}\left(\frac{\cdot-\mathbf{x_i}}{\varepsilon}\right)\cdot\nabla u_0\right) \quad \forall \mathbf{i}$$

Hence,

$$\mathbf{A}_\varepsilon\nabla w_\varepsilon^1 - \boldsymbol{\tau} = \mathbf{A}_\varepsilon\left(\nabla u_0 - \varepsilon\nabla\left(\psi^\varepsilon\mathbf{N}\left(\frac{\cdot-\mathbf{x_i}}{\varepsilon}\right)\cdot\nabla u_0\right)\right) - \boldsymbol{\tau}_0 + \varepsilon^s\boldsymbol{\eta}\left(\frac{\cdot-\mathbf{x_i}}{\varepsilon}\right)$$
$$= \mathbf{A}_\varepsilon\mathbf{G} + \varepsilon^s\boldsymbol{\eta}\left(\frac{\cdot-\mathbf{x_i}}{\varepsilon}\right) \qquad \forall \mathbf{i}$$

with $\mathbf{G}$ as in (2.22). This leads to

$$\left\|\mathbf{A}_\varepsilon\left(\nabla w_\varepsilon^1 - \boldsymbol{\tau}\right)\right\|^2_{\mathbf{A}_\varepsilon^{-1}} = \sum_{\mathbf{i}} \int_{\Pi_{\mathbf{i}}^\varepsilon} \left\{\widehat{\mathbf{A}}^{-1}\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right)\left(\widehat{\mathbf{A}}\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right)\mathbf{G}(\mathbf{x}) + \varepsilon^s\boldsymbol{\eta}\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right)\right)\right.$$
$$\left.\cdot\left(\widehat{\mathbf{A}}\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right)\mathbf{G}(\mathbf{x}) + \varepsilon^s\boldsymbol{\eta}\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right)\right)\right\} d\mathbf{x}$$
$$= \|\mathbf{G}\|^2_{\mathbf{A}_\varepsilon} + \sum_{\mathbf{i}}\left(\varepsilon^{2s+d}\,\|\boldsymbol{\eta}\|^2_{\widehat{\mathbf{A}}^{-1}} + 2\varepsilon^s\int_{\Pi_{\mathbf{i}}^\varepsilon}\mathbf{G}(\mathbf{x})\cdot\boldsymbol{\eta}\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right)\right)$$

The result now follows from (2.1) and Lemma 2.1. ∎

Remark 2.3 *The right-hand side of the majorant (2.24) is the sum of two non-negative terms, which include a "free function" $\boldsymbol{\eta}$ defined on the cell of periodicity. Hence, the computation of the majorant is based on the flux of homogenized solution and a proper selection of the function $\boldsymbol{\eta}$ defined on the cell of periodicity. The scalar parameters λ_i and the power s can be selected in order to minimize the overall value of the majorant. We emphasize that the majorant does not require an approximation of the flux associated with the original periodic problem.*

The following remark concerns the effect of the term $\varepsilon^s \boldsymbol{\eta}$ in the ansatz for $\boldsymbol{\tau}$ in (2.27).

Remark 2.4 *If a periodic structure is coarse and consists of relatively few cells (e.g., 25-100) and/or the coefficients of the matrix $\widehat{\mathbf{A}}$ have jumps, oscillations, etc. then the term $\varepsilon^s \boldsymbol{\eta}$ may augment the homogenized flux substantially. If the periodic structure is fine, then the correction term is less significant and its influence can be diminished by increasing values of s. In the limit case, i.e., $s \to +\infty$, we obtain the following simplified version of the error majorant*

$$\|\nabla(u_\varepsilon - w_\varepsilon^1)\|_{\mathbf{A}_\varepsilon} \;\leq\; \overline{\mathcal{M}}_\oplus\,(u_0,\,\varepsilon) := \left| \sum_{\mathbf{i}} \int_{\Pi_{\mathbf{i}}^\varepsilon} \widehat{\mathbf{A}} \left(\frac{\mathbf{x} - \mathbf{x_i}}{\varepsilon} \right) \mathbf{G}(\mathbf{x}) \cdot \mathbf{G}(\mathbf{x})\, \mathbf{dx} \right|^{1/2}, \tag{2.31}$$

where $\mathbf{G}(\mathbf{x})$ is defined by (2.22). This majorant does not include any domain dependent constants or auxiliary functions and, hence, can be computed from N_k and u_0.

Remark 2.5 *In certain cases, we may know only numerical approximations to the solutions N_k and u_0 of the cell problem (cf. (2.17)) and of the homogenized equation (cf. (2.18)). The corresponding approximation errors can be estimated by error majorants of similar types (see [40] - [53] and references therein). Then, the overall error majorant will include both, approximation and modeling errors. A combined modeling-discretization strategy is suggested in [53] (where the modeling error is generated by defeaturing of a complicated structure) and should be used in this case. This topic deserves a separate investigation and lies beyond the framework of this paper which is focused on the principal structure of the guaranteed error bound for homogenized problems.*

3 Lecture 3: A Posteriori Error Estimation for Highly Indefinite Problems

Remark. This part of the lecture notes is an extended version of the paper [21] and includes details from [30], [32], [33].

3.1 Introduction

In this lecture we will introduce a new analysis for residual-based a posteriori error estimation. We consider the conforming Galerkin method with *hp*-finite elements applied to a class of highly indefinite boundary value problems, which arise, e.g., when electromagnetic or acoustic scattering problems are modelled in the frequency domain. As our model problem we consider a highly indefinite Helmholtz equation with oscillatory solutions.

Residual-based a posteriori error estimates for elliptic problems have been introduced in [7], [8] and their theory for elliptic problems is now fairly completely established (cf. [60], [3]). To sketch the principal idea and to explain our goal, let u denote the (unknown) solution of the weak formulation of an elliptic second order PDE with appropriate boundary conditions. Typically the solution belongs to some infinite-dimensional Sobolev space H. Let u_S denote a computed Galerkin solution based on a finite dimensional subspace $S \subset H$. A (reliable) a posteriori error estimator is a *computable* functional η which depends on u_S and the given data such that an estimate of the form

$$\|u - u_S\|_H \leq C\eta\,(u_S) \tag{3.1}$$

holds for a (minimal) constant C which either is known explicitly or sharp upper bounds are available. We emphasize that in the literature various refinements of this concept of a posteriori error estimation exist while, for the purpose of our introduction, this simple definition is sufficient.

In the classical theory the constant C depends linearly on the norm of the solution operator of the PDE in some appropriate function spaces, more precisely, it depends reciprocally on the *inf-sup constant* γ. In [31] it was proved for the Helmholtz problem with Robin boundary conditions that for certain classes of physical domains the reciprocal inf-sup constant $1/\gamma$ (and, hence, also the constant C in (3.1)) grows linearly with the wavenumber. See also [23] for further estimates of the inf-sup constant for the Helmholtz problem. However, this implies that for large wavenumbers the classical a posteriori estimation becomes useless because the error then typically is highly overestimated. Additional difficulties arise for the a posteriori error estimation for highly indefinite problems because the existence and uniqueness of the classical Galerkin solution is ensured only if the mesh width is sufficiently small.

In contrast to definite elliptic problems, there exist only relatively few publications in the literature on a posteriori estimation for highly indefinite problems (cf. [4], [5], [27]).

In [30] and [32] a new a priori convergence theory for Galerkin discretizations of highly indefinite boundary value problems has been developed which is based on new regularity estimates (the *splitting lemmas* as in [30] and [32]) where the solution is split into a "rough part" with wavenumber-independent regularity constant and a "smooth" part with high-order regularity in (weighted) Sobolev spaces but more critical dependence of the regularity constant on the wavenumber. This theory allows in the a priori convergence theory to "absorb" the L^2-error which depends critically on the wavenumber in the wavenumber-independent part of the equation.

We will develop a new a posteriori analysis based on similar ideas: The L^2-part of the a posteriori error will be estimated by the H^1-error and then can be compensated by an appropriate choice of the *hp*-finite element space.

This lecture is structured as follows. In Section 3.2, we will consider as our model problem the high frequency, time harmonic scattering of an acoustic wave at some bounded domain in an unbounded exterior domain and transform it to a finite domain by using a Dirichlet-to-Neumann boundary operator resp. some approximation to it. We define a conforming Galerkin *hp*-finite element discretization for its numerical approximation and formulate the a posteriori error estimator for *hp*-finite elements.

In Section 3.3, we summarize the a priori analysis as in [30] and [32] which will be needed a) to determine the minimal *hp*-finite element space for a stable Galerkin discretization and b) to estimate the *adjoint approximation property* which will appear as weights in our a posteriori error estimation.

In Section 3.4, we will present the a posteriori error analysis and prove the reliability and efficiency of our estimator. It will turn out that the optimal polynomial degree p will depend logarithmically on the wavenumber and, hence, the finite element interpolation theory has to be explicit with respect to the mesh width h and the polynomial degree p.

3.2 Model Helmholtz Problems and their Discretization

3.2.1 Model Problems

The Helmholtz equation describes wave phenomena in the frequency domain which, e.g., arises if electromagnetic or acoustic waves are scattered from or emitted by bounded physical objects. In this light, the computational domain for such wave problems, typically, is the unbounded complement of a bounded domain $\Omega^{\text{in}} \subset \mathbb{R}^d$, $d = 1, 2, 3$, i.e., $\Omega^{\text{out}} := \mathbb{R}^d \backslash \overline{\Omega^{\text{in}}}$. Throughout this paper, we assume that Ω^{in} has a Lipschitz boundary $\Gamma^{\text{in}} := \partial\Omega^{\text{in}}$.

The Helmholtz problem depends on the wavenumber k. In most parts of the paper (exceptions: Remarks 3.23, 3.13 and Corollaries 3.35, 3.36) we allow for variable wavenumber $k : \Omega^{\text{out}} \to \mathbb{R}$ but always assume that k is real-valued, nonnegative, and a positive constant outside a sufficiently large ball (cf. (3.11)).

For a given right-hand side $f \in L^2(\Omega^{\text{out}})$, the *Helmholtz problem* is to seek $U \in H^1_{\text{loc}}(\Omega^{\text{out}})$ such that

$$\left(-\Delta - k^2\right) U = f \qquad \text{in } \Omega^{\text{out}} \tag{3.2a}$$

is satisfied. Towards infinity, *Sommerfeld's radiation condition* is imposed

$$|\partial_r U - \mathrm{i}\, kU| = o\left(|x|^{\frac{1-d}{2}}\right) \qquad \text{for } |x| \to \infty, \tag{3.2b}$$

where ∂_r denotes differentiation in radial direction and $|\cdot|$ the Euclidian vector norm. For simplicity we restrict here to *homogeneous Dirichlet boundary condition* on Γ^{in}

$$U|_{\Gamma^{\text{in}}} = 0. \tag{3.2c}$$

Assumption 3.1 *The right-hand side f in (3.2a) is local in the sense that there exists some bounded, simply connected Lipschitz domain*[2] *$\Omega^\star$ such that a) $\overline{\Omega^{\text{in}}} \subset \Omega^\star$, b) $\operatorname{supp}(f) \subset \Omega^\star$, and c) k is constant in a neighbourhood of $\partial\Omega^\star$.*

[2]Since Ω^{in} is bounded, $\Omega^\star$ always can be chosen as a ball. Other choices of $\Omega^\star$ might be preferable in certain situations.

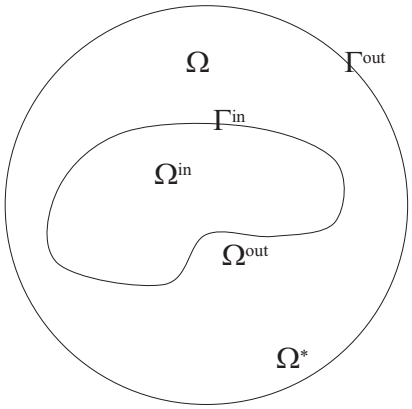

Figure 1: Scatterer Ω^{in} with boundary Γ^{in} and exterior domain Ω^{out}. The support of f is assumed to be contained in the bounded region $\Omega^\star$. The domain for the weak variational formulation is $\Omega = \Omega^\star \backslash \Omega^{\text{in}}$.

The *computational domain* (cf. Figure 1) will be

$$\Omega := \Omega^\star \backslash \overline{\Omega^{\text{in}}} \tag{3.3}$$

and, next, we will derive appropriate boundary conditions at the outer boundary $\Gamma^{\text{out}} := \partial\Omega^\star$. Problem (3.2) can be reformulated in an equivalent way as a *transmission problem* by seeking functions $u \in H^1(\Omega)$ and $u^{\text{out}} \in H^1_{\text{loc}}(\mathbb{R}^d \backslash \overline{\Omega^\star})$ such that

$$\begin{aligned} (-\Delta - k^2)\, u &= f && \text{in } \Omega, \\ (-\Delta - k^2)\, u^{\text{out}} &= 0 && \text{in } \mathbb{R}^d \backslash \overline{\Omega^\star}, \\ u &= 0 && \text{on } \Gamma^{\text{in}}, \\ u = u^{\text{out}} \quad &\text{and} \quad \partial_n u = \partial_n u^{\text{out}} && \text{on } \Gamma^{\text{out}}, \\ |\partial_r u^{\text{out}} - \mathrm{i}\, k u^{\text{out}}| &= o\left(|x|^{\frac{1-d}{2}}\right) && \text{for } |x| \to \infty. \end{aligned} \tag{3.4}$$

Here, n denotes the normal vector pointing into the *exterior domain* $\mathbb{R}^d \backslash \overline{\Omega^\star}$ and ∂_n denotes differentiation in normal direction.

It can be shown that, for given $g \in H^{1/2}(\Gamma^{\text{out}})$, the problem: Find $w \in H^1_{\text{loc}}\left(\mathbb{R}^d \backslash \overline{\Omega^\star}\right)$ such that

$$\begin{aligned} (-\Delta - k^2)\, w &= 0 && \text{in } \mathbb{R}^d \backslash \overline{\Omega^\star}, \\ w &= g && \text{on } \Gamma^{\text{out}}, \\ |\partial_r w - \mathrm{i}\, k w| &= o\left(|x|^{\frac{1-d}{2}}\right) && \text{for } |x| \to \infty \end{aligned} \tag{3.5}$$

has a unique weak solution. The mapping $g \mapsto w$ is called the *Steklov–Poincaré operator* and denoted by $S_P : H^{1/2}(\Gamma^{\text{out}}) \to H^1_{\text{loc}}\left(\mathbb{R}^d \backslash \overline{\Omega^\star}\right)$. The *Dirichlet-to-Neumann* (DtN) map is given by $T_k := \gamma_1 S_P : H^{1/2}(\Gamma^{\text{out}}) \to H^{-1/2}(\Gamma^{\text{out}})$, where $\gamma_1 := \partial_n$ is the normal derivative operator at Γ^{out}. Hence, problem (3.4) can be reformulated as: Find $u \in H^1(\Omega)$ such that

$$\begin{aligned} (-\Delta - k^2)\, u &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \Gamma^{\text{in}}, \\ \partial_n u &= T_k u && \text{on } \Gamma^{\text{out}}. \end{aligned} \tag{3.6}$$

The previous problems are posed in the *weak formulation* given by: Find

$$u \in \mathcal{H} := \left\{ u \in H^1(\Omega) : u|_{\Gamma^{\text{in}}} = 0 \right\} \tag{3.7}$$

such that

$$A_{\text{DtN}}(u,v) := \int_\Omega \left(\langle \nabla u, \nabla \bar{v} \rangle - k^2 u \bar{v} \right) - \int_{\Gamma^{\text{out}}} (T_k u)\, \bar{v} = \int_\Omega f \overline{v} \qquad \text{for all } v \in \mathcal{H}. \tag{3.8}$$

Since the numerical realization of the nonlocal DtN map T_k is costly, various approaches exist in the literature to approximate this operator by a local operator. The most simple one is the use of Robin boundary conditions leading to

$$\begin{aligned} (-\Delta - k^2)\, u &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \Gamma^{\text{in}}, \\ \partial_n u &= \mathrm{i}\, k u && \text{on } \Gamma^{\text{out}}. \end{aligned} \tag{3.9}$$

The weak formulation of this equation is given by: Find $u \in \mathcal{H}$ such that

$$A_{\text{Robin}}(u,v) := \int_\Omega \left(\langle \nabla u, \nabla \bar{v} \rangle - k^2 u \bar{v} \right) - \int_{\Gamma^{\text{out}}} \mathrm{i}\, k u \bar{v} = \int_\Omega f \overline{v} \qquad \text{for all } v \in \mathcal{H}. \tag{3.10}$$

In most parts of this paper we allow indeed that k is a function varying in Ω, while the following conditions are always assumed to be satisfied:

$$\begin{aligned} k \in L^\infty\left(\mathbb{R}^d, \mathbb{R}\right), \qquad & 0 \le \operatorname{essinf}_{x\in\Omega} k(x) \le \operatorname{esssup}_{x\in\Omega} k(x) =: k_{\max} < \infty, \\ & k = k_{\text{const}} \text{ outside a large ball}, \\ & k = k_{\text{const}} \text{ in an neighbourhood } \mathcal{U}^\star_{\text{const}} \text{ of } \Gamma^{\text{out}}. \end{aligned} \tag{3.11}$$

Let $\mathcal{U}_{\text{const}} := \mathcal{U}^\star_{\text{const}} \cap \overline{\Omega}$. The constants in the estimates in this paper will depend on $k_{\max}$, and $\mathcal{U}_{\text{const}}$ (through a trace inequality as in Lemma 3.3) but hold uniformly for all functions k satisfying (3.11).

3.2.2 Abstract Variational Formulation

Notation 3.2 *For a Lebesgue-measurable set $\omega \subset \mathbb{R}^d$ and $p \in [1,\infty]$, $m \in \mathbb{N}$, we denote by $L^p(\omega)$ the usual Lebesgue space with norm $\|\cdot\|_{L^p(\omega)}$ and by $H^m(\omega)$ the usual Sobolev spaces with norm $\|\cdot\|_{H^m(\omega)}$. The seminorm which contains only the derivatives of highest order is denoted by $|\cdot|_{H^m(\omega)}$. We equip the space $\mathcal{H}$ with the norm*

$$\|v\|_{\mathcal{H};\Omega} := \left(\|\nabla v\|^2_{L^2(\Omega)} + \|k_+ v\|^2_{L^2(\Omega)} \right)^{1/2} \quad \text{with} \quad k_+ := \max\{1, k\} \tag{3.12}$$

which is obviously equivalent to the $H^1(\Omega)$-norm.

Since Γ^{out} is a Lipschitz manifold and $\mathcal{U}_{\text{const}}$ is a Lipschitz domain, it is well known that the following trace estimates hold (see [14, (1.6.6) Theorem]).

Lemma 3.3 *There exists a constant C_{tr} depending only on $\mathcal{U}_{\text{const}}$ such that*

$$\forall u \in H^1(\Omega): \quad \|u\|_{H^{1/2}(\Gamma^{\text{out}})} \le C_{\text{tr}} \|u\|_{\mathcal{H};\mathcal{U}_{\text{const}}} \tag{3.13a}$$

and

$$\forall u \in H^1(\Omega): \quad \|u\|_{L^2(\Gamma^{\text{out}})} \le C_{\text{tr}} \|u\|^{1/2}_{L^2(\mathcal{U}_{\text{const}})} \|u\|^{1/2}_{H^1(\mathcal{U}_{\text{const}})}. \tag{3.13b}$$

Corollary 3.4 *For $u \in H^1(\Omega)$, we have*

$$\left\| \sqrt{k} u \right\|_{L^2(\Gamma^{\mathrm{out}})} \leq C_{\mathrm{tr}} \|u\|_{\mathcal{H};\mathcal{U}_{\mathrm{const}}} \leq C_{\mathrm{tr}} \|u\|_{\mathcal{H};\Omega}.$$

Proof. Since $k = k_{\mathrm{const}}$ on $\mathcal{U}_{\mathrm{const}}$, there holds

$$\begin{aligned} k_{\mathrm{const}} \|u\|^2_{L^2(\Gamma^{\mathrm{out}})} &\leq C_{\mathrm{tr}}^2 k_{\mathrm{const}} \|u\|_{L^2(\mathcal{U}_{\mathrm{const}})} \|u\|_{H^1(\mathcal{U}_{\mathrm{const}})} \\ &\leq \frac{C_{\mathrm{tr}}^2}{2} \left(k_{\mathrm{const}}^2 \|u\|^2_{L^2(\mathcal{U}_{\mathrm{const}})} + \|u\|^2_{H^1(\mathcal{U}_{\mathrm{const}})} \right) \\ &= \frac{C_{\mathrm{tr}}^2}{2} \left(\left(1 + k_{\mathrm{const}}^2\right) \|u\|^2_{L^2(\mathcal{U}_{\mathrm{const}})} + |u|^2_{H^1(\mathcal{U}_{\mathrm{const}})} \right) \\ &\leq C_{\mathrm{tr}}^2 \left(\|k_+ u\|^2_{L^2(\mathcal{U}_{\mathrm{const}})} + |u|^2_{H^1(\mathcal{U}_{\mathrm{const}})} \right). \end{aligned} \tag{3.14}$$

■

Both sesquilinear forms A_{DtN} (3.8) and A_{Robin} (3.10) belong to the following class of forms (see Proposition 3.7).

Assumption 3.5 (Variational formulation) *Let $\Omega \subset \mathbb{R}^d$, for $d \in \{2, 3\}$, be a bounded Lipschitz domain. Then $\mathcal{H}$, equipped with the norm $\|\cdot\|_{\mathcal{H};\Omega}$, is a closed subspace of $H^1(\Omega)$. We consider a sesquilinear form $A : \mathcal{H} \times \mathcal{H} \to \mathbb{C}$ that can be decomposed into $A = a - b$, where*

$$a(v, w) := \int_\Omega \left(\langle \nabla v, \nabla \bar{w} \rangle - k^2 v \bar{w} \right)$$

and the sesquilinear form b satisfies the following properties:

a. $b : \mathcal{H} \times \mathcal{H} \to \mathbb{C}$ is a continuous *sesquilinear form with*

$$|b(v, w)| \leq C_b \|v\|_{\mathcal{H};\Omega} \|w\|_{\mathcal{H};\Omega} \qquad \text{for all } v, w \in \mathcal{H}, \tag{3.15}$$

for some positive constant C_b.

b. There exist $\theta \geq 0$ and $\gamma_{\mathrm{ell}} > 0$ such that the following Gårding inequality *holds:*

$$\mathrm{Re}\left(a(v, v) - b(v, v)\right) + \theta \|k_+ v\|^2_{L^2(\Omega)} \geq \gamma_{\mathrm{ell}} \|v\|^2_{\mathcal{H};\Omega} \qquad \text{for all } v \in \mathcal{H}. \tag{3.16}$$

c. The adjoint problem*: Find $z \in \mathcal{H}$ such that*

$$a(v, z) - b(v, z) = (v, f)_{L^2(\Omega)} \qquad \text{for all } v \in \mathcal{H} \tag{3.17}$$

is uniquely solvable for every $f \in L^2(\Omega)$ with bounded solution operator $Q_k^\star : L^2(\Omega) \to \mathcal{H}$, $f \mapsto z$, more precisely, the (k-dependent) constant

$$C_k^{\mathrm{adj}} := \sup_{f \in L^2(\Omega) \setminus \{0\}} \frac{\|Q_k^\star \left(k_+^2 f\right)\|_{\mathcal{H};\Omega}}{\|k_+ f\|_{L^2(\Omega)}} \tag{3.18}$$

is finite.

Problem 3.6 *Let A be a sesquilinear form as in Assumption 3.5. For given $f \in L^2(\Omega)$, we seek $u \in \mathcal{H}$ such that*

$$a(u,v) - b(u,v) = \int_\Omega f\bar{v} \qquad \text{for all } v \in \mathcal{H}. \tag{3.19}$$

Proposition 3.7 *Both sesquilinear forms A_{Robin} (3.10) and A_{DtN} (3.8) (under the additional condition that Γ^{out} is a sufficiently large sphere) satisfy Assumption 3.5.*

Proof. The proof is a slight modification of the corresponding proofs for constant wavenumber k in [30] and [31]. Condition (a) for A_{Robin} follows from Corollary 3.4. For A_{DtN} we employ that k is constant in $\mathcal{U}_{\mathrm{const}}$ and Γ^{out} is a sphere of a radius $R > 0$. Hence, from the proof of [30, Lemma 3.3][3] it follows that

$$\begin{aligned}&\left|\int_{\Gamma^{\mathrm{out}}} (T_k u)\,\bar{v}\right| \\ &\qquad \le C\left(R^{-1}\|u\|_{H^{1/2}(\Gamma^{\mathrm{out}})}\|v\|_{H^{1/2}(\Gamma^{\mathrm{out}})} + k_{\mathrm{const}}\|u\|_{L^2(\Gamma^{\mathrm{out}})}\|v\|_{L^2(\Gamma^{\mathrm{out}})}\right).\end{aligned}$$

By using Corollary 3.4 we obtain

$$\left|\int_{\Gamma^{\mathrm{out}}} (T_k u)\,\bar{v}\right| \le C\left(1+\frac{1}{R}\right) C_{\mathrm{tr}}^2\,\|u\|_{\mathcal{H};\Omega}\,\|v\|_{\mathcal{H};\Omega}$$

and the continuity of A_{DtN} follows.

For condition (b) and Robin boundary conditions, we employ

$$\begin{aligned}\mathrm{Re}\,(A_{\mathrm{Robin}}(v,v)) + 2\|k_+ v\|^2_{L^2(\Omega)} &\ge \int_\Omega \left(|\nabla v|^2 + k_+^2\,|v|^2 + \left(k_+^2 - k^2\right)|v|^2\right) \\ &\ge \|v\|^2_{\mathcal{H};\Omega}\end{aligned}$$

and (3.16) holds with $\theta = 2$ and $\gamma_{\mathrm{ell}} = 1$.

For the sesquilinear form A_{DtN} we employ

$$\mathrm{Re}\left(\int_{\Gamma^{\mathrm{out}}} T_k v\,\bar{v}\right) \le 0 \qquad \forall v \in H^{1/2}\left(\Gamma^{\mathrm{out}}\right) \tag{3.20}$$

(proved [30, Lemma 3.3 (2)] by using spectral analysis) to obtain

$$\begin{aligned}&\mathrm{Re}\,(A_{\mathrm{DtN}}(v,v)) + 2\|k_+ v\|^2_{L^2(\Omega)} \\ &\qquad \ge \left(\int_\Omega \left(|\nabla v|^2 + k_+^2\,|v|^2 + \left(k_+^2 - k^2\right)|v|^2\right) - \mathrm{Re}\left(\int_{\Gamma^{\mathrm{out}}} T_k v\,\bar{v}\right)\right) \\ &\qquad \ge \|v\|^2_{\mathcal{H};\Omega}\end{aligned}$$

and (3.16) again holds with $\theta = 2$ and $\gamma_{\mathrm{ell}} = 1$.

For condition (c) we may apply Fredholm's theory and, hence, it suffices to prove that

$$a(u,v) - b(u,v) = 0 \qquad \text{for all } v \in \mathcal{H} \tag{3.21}$$

[3] In [30, Lemma 3.3] the Dirichlet-to-Neumann operator T_k has been analysed for a sphere by using the fact that spherical harmonics are the eigenfunctions of T_k with known eigenvalues. Then the proof follows by bounding these eigenvalues uniformly in k.

implies $u = 0$. For Robin boundary conditions we argue as in [31, (8.1.2)] and for DtN boundary conditions as in [30, Proof of Theorem 3.8] to see that (3.21) implies $u|_{\partial\Omega} = 0$ in the sense of traces. Hence, u solves

$$\int_\Omega \left(\langle \nabla u, \nabla \bar{v}\rangle - k^2 u\bar{v}\right) = 0 \qquad \text{for all } v \in \mathcal{H}. \tag{3.22}$$

Let $\Omega^{\star\star}$ be a bounded domain such that $\Omega \subset \Omega^{\star\star} \subset \mathbb{R}^d \backslash \overline{\Omega^{\text{in}}}$ and $\Gamma^{\text{out}} \subset \Omega^{\star\star}$. The extension of u by zero to $\Omega^{\star\star}$ is denoted by u_0. It satisfies $u \in \mathcal{H}\left(\Omega^{\star\star}\right) := \{u \in H^1\left(\Omega^{\star\star}\right) \mid u|_{\Gamma^{\text{in}}} = 0\}$ and

$$\int_\Omega \left(\langle \nabla u_0, \nabla \bar{v}\rangle - k^2 u_0\bar{v}\right) = 0 \qquad \text{for all } v \in \mathcal{H}\left(\Omega^{\star\star}\right).$$

Elliptic regularity theory implies that $u_0 \in H^2(Q)$ for any compact subset $Q \subset \Omega^{\star\star}$, in particular, in an open $\Omega^{\star\star}$ neighbourhood of Γ^{out}. The unique continuation principle (cf. [29, Ch. 4.3]) implies that $u_0 = 0$ in $\Omega^{\star\star}$ so that $u = 0$ in Ω. ∎

3.2.3 Discretization

Conforming Galerkin Discretization A *conforming Galerkin discretization* of Problem 3.6 is based on the definition of a finite dimensional subspace $S \subset \mathcal{H}$ and is given by: Find $u_S \in S$ such that

$$a(u_S, v) - b(u_S, v) = \int_\Omega f\bar{v} \qquad \text{for all } v \in S. \tag{3.23}$$

***hp*-Finite Elements** As an example for S as above, we will define *hp*-finite elements on a finite element mesh $\mathcal{T}$ consisting of simplices with maximal mesh width h and local polynomial degree p. Before formulating the conditions on the mesh in an abstract way, we give an example of a typical construction.

Example 3.8 (Patchwise construction of FE mesh.) *Let Ω denote a bounded domain.*

(a) We assume that a polyhedral (polygonal in 2D) domain $\widetilde{\Omega}$ along with a bi-Lipschitz mapping $\chi : \widetilde{\Omega} \to \Omega$ is given. Let $\widetilde{\mathcal{T}}^{\text{macro}} = \left\{\widetilde{K}_i^{\text{macro}} : 1 \le i \le q\right\}$ denote a conforming finite element mesh for $\widetilde{\Omega}$ consisting of open simplices. $\widetilde{\mathcal{T}}^{\text{macro}}$ is considered as a coarse partition of $\widetilde{\Omega}$, i.e., the diameters of the elements in $\widetilde{\mathcal{T}}^{\text{macro}}$ are of order 1. We assume that the restrictions $\chi_i := \chi|_{\widetilde{K}_i^{\text{macro}}}$ are analytic *for all $1 \le i \le q$.*

(b) The finite element mesh with step size h is generated by refining the mesh $\widetilde{\mathcal{T}}^{\text{macro}}$ in some standard (conforming) way and denoted by $\widetilde{\mathcal{T}} = \left\{\widetilde{K}_i : 1 \le i \le N\right\}$. The corresponding finite element mesh for Ω then is defined by $\mathcal{T} = \left\{K = \chi\left(\widetilde{K}\right) : \widetilde{K} \in \widetilde{\mathcal{T}}\right\}$.

Note that, for any $K = \chi\left(\widetilde{K}\right) \in \mathcal{T}$, there exists an affine bijection $A_K : \widehat{K} \to \widetilde{K}$ which maps the reference element $\widehat{K} := \left\{x \in \mathbb{R}^d_{>0} : \sum_{i=1}^d x_i < 1\right\}$ *to the simplex $\widetilde{K}$. A parametrization $F_K : \widehat{K} \to K$ can be chosen by $F_K := R_K \circ A_K$, where $R_K := \chi|_{\widetilde{K}}$ is independent of the mesh width $h := \max\{h_K : K \in \mathcal{T}\}$, where $h_K := \operatorname{diam}(K)$.*

Concerning the polynomial degree distribution, it will be convenient (cf.[33, (10)]) to assume that the polynomial degrees of neighbouring elements are comparable: There exists a constant $c_{\mathrm{p}} > 0$ such that

$$c_{\mathrm{p}}^{-1}\,(p_K + 1) \le p_{K'} + 1 \le c_{\mathrm{p}}\,(p_K + 1) \qquad \text{for all } K, K' \in \mathcal{T} \text{ with } \overline{K} \cap \overline{K'} = \emptyset. \tag{3.24}$$

To formulate the smoothness and scaling assumptions on R_K and A_K in an abstract way we have to introduce, for a function $v : \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}^d$, the notation

$$\frac{d^n}{n!} = \sum_{\alpha \in \mathbb{N}_0^d : |\alpha| = n} \frac{1}{\alpha!} \quad \text{and} \quad |\nabla^n v(x)|^2 = \sum_{\alpha \in \mathbb{N}_0^d : |\alpha| = n} \frac{n!}{\alpha!} |\partial^\alpha v(x)|^2. \tag{3.25}$$

Assumption 3.9 *Each element map F_K can be written as $F_K = R_K \circ A_K$, where A_K is an affine map and the maps R_K and A_K satisfy for constants C_{affine}, C_{metric}, $\gamma > 0$ independent of h_K:*

$$\begin{aligned} &\|A_K'\|_{L^\infty(\widehat{K})} \le C_{\mathrm{affine}} h_K, \qquad \|(A_K')^{-1}\|_{L^\infty(\widehat{K})} \le C_{\mathrm{affine}} h_K^{-1} \\ &\|(R_K')^{-1}\|_{L^\infty(\widetilde{K})} \le C_{\mathrm{metric}}, \qquad \|\nabla^n R_K\|_{L^\infty(\widetilde{K})} \le C_{\mathrm{metric}} \gamma^n n! \qquad \text{for all } n \in \mathbb{N}_0. \end{aligned}$$

Here, $\widetilde{K} = A_K(\widehat{K})$.

Remark 3.10 *Assumption 3.9 will be used in Section 3.3 for the a priori analysis and the derivation of the minimal hp-finite element space which leads to a stable discretization of the Helmholtz problem. It will turn out that the a posteriori estimate contains a weight which requires an a priori estimate. Since higher polynomial orders p are relevant for this, Assumption 3.9 also contains bounds on higher order derivatives of the element maps. The constants C_{affine}, C_{metric} describe the* shape-regularity *of the finite element mesh, i.e., they are a measure for possible distortions of the elements. The constants in the following estimates depend on the constants C_{affine}, C_{metric} and are moderately bounded if the shape regularity of the mesh is reasonably small.*

Definition 3.11 (*hp*-finite element space) *For meshes $\mathcal{T}$ with element maps F_K as in Assumption 3.9 the hp-finite element space of piecewise (mapped) polynomials is given by*

$$S^{p,1}(\mathcal{T}) := \{v \in \mathcal{H} : v|_K \circ F_K \in \mathbb{P}_p \text{ for all } K \in \mathcal{T}\}, \tag{3.26}$$

where $\mathbb{P}_p$ denotes the space of polynomials of degree p. For chosen $\mathcal{T}$ and p, we may let $S = S^{p,1}(\mathcal{T})$.

3.2.4 A Posteriori Error Estimator

The following Assumption collects the requirements for the a posteriori error estimation.

Assumption 3.12

a. *The continuous Helmholtz problem satisfies Assumption 3.5.*

b. *S is a hp-finite element space as explained in Section 3.2.3 and satisfies Assumption 3.9 and (3.24).*

c. $u_S \in S$ *is the computed solution satisfying the Galerkin equation.*

Remark 3.13 *Assumption 3.12 does not require the stability condition (3.31) to be satisfied which is only* sufficient *for existence and uniqueness of the discrete problem. We only assume that* u_S *exists, is computed, and solves the Galerkin equation for the specific problem. To be on the safe side in the case of constant wave number* k*, one can start the discretization process with the a priori choice (3.39) of* p *and* h *which implies (3.31) and, in turn, the existence and uniqueness of a Galerkin solution for any right-hand side in* $L^2(\Omega)$*.*

For the definition of the a posteriori error estimator we first have to introduce some notation. For a simplicial finite element mesh $\mathcal{T}$, the boundary of any element $K \in \mathcal{T}$ consists of $(d-1)$-dimensional simplices. We call (the relatively open interior of) these lower dimensional simplices the *edges* of K, although this terminology is related to the case $d = 2$. The set of all edges of all elements in $\mathcal{T}$ is denoted by $\mathcal{E}^\star$. The subset $\mathcal{E}^\partial \subset \mathcal{E}^\star$ consists of all edges which are contained in Γ^{out} while the subset $\mathcal{E}^\Omega \subset \mathcal{E}^\star$ consists of all edges that are contained in Ω. Finally, we set $\mathcal{E} := \mathcal{E}^\Omega \cup \mathcal{E}^\partial$, the set of all edges that are not in Γ^{in}. The set of simplex vertices that are not contained in $\overline{\Gamma^{\text{in}}}$ is denoted by $\mathcal{N}$ and, for the cardinality of a discrete set, we write $|\mathcal{N}|$, $|\mathcal{E}|$, etc. For a subset $\mathcal{M} \subset \Omega$ we define simplex neighborhoods about $\mathcal{M}$ by

$$\begin{aligned}
\omega^0_{\mathcal{M}} &:= \left\{\overline{\mathcal{M}}\right\}, \\
\omega^j_{\mathcal{M}} &:= \bigcup \left\{\overline{K} \mid K \in \mathcal{T} \text{ and } \overline{K} \cap \omega^{j-1}_{\mathcal{M}} \neq \emptyset\right\}, \quad j \geq 1, \\
h_{\mathcal{M}} &:= \max\left\{h_K \mid \overline{\mathcal{M}} \cap \overline{K} \neq \emptyset\right\}, \\
p_{\mathcal{M}} &:= \max\left\{p_K + 1 \mid \overline{\mathcal{M}} \cap \overline{K} \neq \emptyset\right\}, \\
\mathcal{E}_{\mathcal{M}} &:= \left\{E \in \mathcal{E}^\star \mid \overline{\mathcal{M}} \cap \overline{E} \neq \emptyset\right\}.
\end{aligned} \tag{3.27}$$

Definition 3.14 (Residual) *For* $v \in S$ *we define the* volume residual $\operatorname{res}(v) \in L^2(\Omega)$ *and the* edge residual $\operatorname{Res}(v) \in L^2(\cup_{E\in\mathcal{E}} E)$ *by*

$$\begin{aligned}
\operatorname{res}(v) &:= f + \Delta v + k^2 v \qquad \text{on } K \in \mathcal{T}, \\
\operatorname{Res}(v) &:= \begin{cases} [\partial_n v]_E & \text{on } E \in \mathcal{E}^\Omega, \\ -\partial_n v + \mathrm{i}\, k v & \text{on } E \in \mathcal{E}^\partial. \end{cases}
\end{aligned}$$

Here $[v]_E$ *is the* jump *of the given function* v *on the edge* E*, i.e., the difference of the limits in points* $x \in E$ *from both sides.*

In the definitions above we used exact data f, k. We will later, Section 3.4.3, replace these by approximations.

The residual $\operatorname{Res}(v)$ is defined for the Robin boundary condition (3.9) for simplicity. With an obvious modification of this definition, we could also insert a term $T_k v$ here, instead of $\mathrm{i}\, k v$, for the DtN boundary condition (3.6).

Definition 3.15 (Error estimator) *Given a set of* weights $\alpha = \{\alpha_K, \alpha_E : K \in \mathcal{T},\, E \in \mathcal{E}\}$*, we define for* $v \in S$ *the* error estimator

$$\eta(v,\alpha) := \left(\sum_{K\in\mathcal{T}} \alpha_K^2 \left\|\operatorname{res}(v)\right\|^2_{L^2(K)} + \sum_{E\in\mathcal{E}} \alpha_E^2 \left\|\operatorname{Res}(v)\right\|^2_{L^2(E)}\right)^{1/2}. \tag{3.28}$$

The choice of the weights α_K, α_E are related to an interpolation estimate which we explain next.

Assumption 3.16 (Interpolation operator) *Let $I_S : \mathcal{H} \to S$ denote a continuous linear operator that satisfies the* local approximation property: *There are constants $\alpha_K > 0$ for all $K \in \mathcal{T}$ and $\alpha_E > 0$ for all $E \in \mathcal{E}$ such that*

$$\|v - I_S v\|_{L^2(K)} \leq \alpha_K \|v\|_{\mathcal{H};\omega_K^m}, \tag{3.29a}$$

$$\|v - I_S v\|_{L^2(E)} \leq \alpha_E \|v\|_{\mathcal{H};\omega_E^m}, \tag{3.29b}$$

for some $m = O(1)$ independent of h_K, p_K.

The weights in (3.28) can be chosen as the minimal constants in (3.29) for any given operator I_S that satisfies the above mentioned properties. In [33, Thms 2.1, 2.2], a Clément-type *hp*-interpolation operator has been constructed which leads to specific choices of α_K, α_E.

Theorem 3.17 *Let $\Omega \subset \mathbb{R}^2$ and let $p = (p_K)_{K\in\mathcal{T}}$ denote a polynomial degree distribution satisfying (3.24). Let Assumption 3.12(a), (b) be satisfied. Then there exist $C > 0$, that depends only on the shape-regularity of the grid (cf. Remark 3.10), and a linear operator $I_S : H^1_{\mathrm{loc}}(\mathbb{R}^2) \to S$ such that for all simplices $K \in \mathcal{T}$ and all edges $E \in \mathcal{E}_K$ we have*

$$\|u - I_S u\|_{L^2(K)} + \frac{h_K}{p_K}\|\nabla I_S u\|_{L^2(K)} + \sqrt{\frac{h_K}{p_K}}\|u - I_S u\|_{L^2(E)} \leq C_0 \frac{h_K}{p_K}\|\nabla u\|_{L^2(\omega_K^4)}.$$

Proof. This result has been proven in [33] in a vertex oriented setting, but is easily reformulated as stated above using shape uniformity and quasi-uniformity in the polynomial degree (3.24). ∎

Corollary 3.18 *Let the Assumptions of Theorem 3.17 be satisfied. The constants α_K, α_E in Assumption 3.16 can be chosen according to*

$$\alpha_K := C_0 \frac{h_K}{p_K}, \qquad \alpha_E := C_0 \left(\frac{h_K}{p_K}\right)^{1/2}.$$

Theorem 3.33 will show that this $\eta(u_S, \alpha)$ can be used for a posteriori error estimation. That it estimates the error from above is called *reliability*, that it estimates the error from below is called *efficiency*.

3.3 A Priori Analysis

In this section, we collect those results on existence, uniqueness, stability, and regularity for the Helmholtz problem (3.6), which later will be used for the analysis of the a posteriori error estimator.

3.3.1 Well-posedness

Proposition 3.19 *Let $\Omega^{\text{in}} \subset \mathbb{R}^d$, $d = 2, 3$, in* (3.2a) *be a bounded Lipschitz domain which is star-shaped with respect to the origin. Let $\Gamma^{\text{out}} := \partial B_R$ for some $R > 0$. Then,* (3.8) *admits a unique solution $u \in \mathcal{H}$ for all $f \in \mathcal{H}'$ which depends continuously on the data.*

Proposition 3.20 *Let Ω be a bounded Lipschitz domain. For all $f \in (H^1(\Omega))'$, a unique solution u of problem* (3.10) *exists and depends continuously on the data.*

For the proofs of these propositions for constant k we refer, e.g., to [31, Prop. 8.1.3] and [17, Lemma 3.3], while for variable k one may argue as in Proposition 3.7.

3.3.2 Discrete Stability and Convergence

An essential role for the stability and convergence of the Galerkin discretization is played by the adjoint approximability which has been introduced in [32]; see also [55], [12].

Definition 3.21 (Adjoint approximability) *For a finite dimensional subspace $S \subset \mathcal{H}$, we define the* adjoint approximability *of Problem 3.6 by*

$$\eta_k^\star(S) := \sup_{f \in L^2(\Omega)\setminus\{0\}} \frac{\inf_{v\in S} \|Q_k^\star\left(k_+^2 f\right) - v\|_{\mathcal{H};\Omega}}{\|k_+ f\|_{L^2(\Omega)}}, \tag{3.30}$$

where $Q_k^\star$ is as in (3.18).

Theorem 3.22 (Stability and convergence) *Let Γ^{out} (cf. (3.4)) be the unit sphere. Let $\gamma_{\text{ell}}, \theta, C_b, C_k^{\text{adj}}$ be as in Assumption 3.5 and S as in Section 3.2.3. Then the condition*

$$\eta_k^\star(S) \le \frac{\gamma_{\text{ell}}}{2\theta(1 + C_b)} \tag{3.31}$$

implies the following statements:

(a) The discrete inf-sup condition is satisfied:

$$\inf_{v\in S\setminus\{0\}} \sup_{w\in S\setminus\{0\}} \frac{|a(v,w) - b(v,w)|}{\|v\|_{\mathcal{H};\Omega}\|w\|_{\mathcal{H};\Omega}} \ge \frac{\gamma_{\text{ell}}}{2 + \gamma_{\text{ell}}/(1 + C_b) + 2\theta C_k^{\text{adj}}} > \mathbf{0}. \tag{3.32}$$

(b) Let S satisfy (3.31). Then, the Galerkin method based on S is quasi-optimal, i.e., for every $u \in \mathcal{H}$ there exists a unique $u_S \in S$ with $a(u - u_S, v) - b(u - u_S, v) = 0$ for all $v \in S$, and there holds

$$\|u - u_S\|_{\mathcal{H};\Omega} \le \frac{2}{\gamma_{\text{ell}}}(1 + C_b) \inf_{v\in S} \|u - v\|_{\mathcal{H};\Omega}, \tag{3.33}$$

$$\|k_+(u - u_S)\|_{L^2(\Omega)} \le \frac{2}{\gamma_{\text{ell}}}(1 + C_b)^2 \eta_k^\star(S) \inf_{v\in S} \|u - v\|_{\mathcal{H};\Omega}. \tag{3.34}$$

Proof. The sesquilinear form $A(u,v) = a(u,v) - b(u,v)$ is continuous:

$$|A(u,v)| \overset{(3.15)}{\leq} (1+C_b)\,\|u\|_{\mathcal{H};\Omega}\,\|v\|_{\mathcal{H};\Omega} \qquad \forall u,v \in H^1(\Omega)\,. \tag{3.35}$$

Let $u \in S$ and set $z := \theta Q_k^\star\left(k_+^2 u\right)$. Then,

$$\begin{aligned} A(u,u+z) &= a(u,u) - b(u,u) + \theta\|k_+u\|^2_{L^2(\Omega)} + A(u,z) - \theta\|k_+u\|^2_{L^2(\Omega)} \\ &= A(u,u) + \theta\|k_+u\|^2_{L^2(\Omega)}. \end{aligned} \tag{3.36}$$

Let $z_S \in S$ denote the best approximation of z with respect to the $\|\cdot\|_{\mathcal{H}}$-norm. Then, by using (3.16) we get

$$\begin{aligned} \operatorname{Re} A(u,u+z_S) &\geq \operatorname{Re} A(u,u+z) - |A(u,z-z_S)| \overset{(3.36)}{=} \operatorname{Re}\left(A(u,u) + \theta\|k_+u\|^2_{L^2(\Omega)}\right) - |A(u,z-z_S)| \\ &\overset{(3.35)}{\geq} \gamma_{\mathrm{ell}}\,\|u\|^2_{\mathcal{H}} - (1+C_b)\,\|u\|_{\mathcal{H}}\,\|z-z_S\|_{\mathcal{H}} \\ &\geq \|u\|_{\mathcal{H}}\left(\gamma_{\mathrm{ell}}\,\|u\|_{\mathcal{H}} - \theta\,(1+C_b)\,\eta_k^\star(S)\,\|k_+u\|_{L^2(B_R)}\right) \geq \left(\gamma_{\mathrm{ell}} - \theta\,(1+C_b)\,\eta_k^\star(S)\right)\|u\|^2_{\mathcal{H}}\,. \end{aligned}$$

The stability of the continuous problem (cf. (3.18)) implies

$$\begin{aligned} \|u+z_S\|_{\mathcal{H}} &\leq \|u\|_{\mathcal{H}} + \|z-z_S\|_{\mathcal{H}} + \|z\|_{\mathcal{H}} \leq \|u\|_{\mathcal{H}} + \theta\eta_k^\star(S)\,\|k_+u\|_{L^2(B_R)} + \theta C_k^{\mathrm{adj}}\,\|k_+u\|_{L^2(B_R)} \\ &\leq \left(1 + \theta\eta_k^\star(S) + \theta C_k^{\mathrm{adj}}\right)\|u\|_{\mathcal{H}} \end{aligned}$$

so that

$$\operatorname{Re} A(u,u+z_S) \geq \frac{\gamma_{\mathrm{ell}} - \theta\,(1+C_b)\,\eta_k^\star(S)}{1 + \theta\eta_k^\star(S) + \theta C_k^{\mathrm{adj}}}\,\|u\|_{\mathcal{H}}\,\|u+z_S\|_{\mathcal{H}}\,.$$

Therefore, in view of the assumption (3.31), we have proved

$$\inf_{u\in S}\sup_{v\in S\setminus\{0\}} \frac{|A(u,v)|}{\|u\|_{\mathcal{H}}\,\|v\|_{\mathcal{H}}} \geq \frac{\gamma_{\mathrm{ell}} - \theta\,(1+C_b)\,\eta_k^\star(S)}{1 + \theta\eta_k^\star(S) + \theta C_k^{\mathrm{adj}}} \geq \frac{\gamma_{\mathrm{ell}}}{2 + \gamma_{\mathrm{ell}}/\,(1+C_b) + 2\theta C_k^{\mathrm{adj}}}.$$

Next, we will estimate the L^2-error by the H^1-error and employ the Aubin-Nitsche technique. The Galerkin error is denoted by $e = u - u_S$. We set $\psi := Q_k^\star\left(k_+^2 e\right)$ (cf. (3.18)) and denote by $\psi_S \in S$ the best approximation of ψ with respect to the $\mathcal{H}$-norm.

The L^2-error can be estimated by using the Galerkin orthogonality

$$\begin{aligned} \|k_+e\|^2_{L^2(B_R)} &= A(e,\psi) = A(e,\psi-\psi_S) \leq (1+C_b)\,\|e\|_{\mathcal{H}}\,\|\psi-\psi_S\|_{\mathcal{H}} \\ &\leq (1+C_b)\,\eta_k^\star(S)\,\|e\|_{\mathcal{H}}\,\|k_+e\|_{L^2(B_R)}\,, \end{aligned} \tag{3.37}$$

i.e.,

$$\|k_+e\|_{L^2(B_R)} \leq (1+C_b)\,\eta_k^\star(S)\,\|e\|_{\mathcal{H}}\,. \tag{3.38}$$

To infer from this a bound for $\|e\|_{\mathcal{H}}$, we notice that Galerkin orthogonality gives for arbitrary $v \in S$

$$\begin{aligned} \gamma_{\mathrm{ell}}\|e\|^2_{\mathcal{H}} &\leq \operatorname{Re}\left(a(e,e) - b(e,e) + \theta\|k_+e\|^2_{L^2(\Omega)}\right) = \operatorname{Re}\left(a(e,u-v) - b(e,u-v) + \theta\|k_+e\|^2_{L^2(\Omega)}\right) \\ &\leq (1+C_b)\|e\|_{\mathcal{H}}\|u-v\|_{\mathcal{H}} + \theta\|k_+e\|^2_{L^2(\Omega)} \\ &\leq (1+C_b)\|e\|_{\mathcal{H}}\|u-v\|_{\mathcal{H}} + \theta\|k_+e\|_{L^2(\Omega)}\|k_+e\|_{L^2(\Omega)} \\ &\leq (1+C_b)\|e\|_{\mathcal{H}}\|u-v\|_{\mathcal{H}} + \theta(1+C_b)\eta_k^\star(S)\|e\|_{\mathcal{H}}\|e\|_{\mathcal{H}} \\ &\leq (1+C_b)\|e\|_{\mathcal{H}}\|u-v\|_{\mathcal{H}} + \gamma_{\mathrm{ell}}/2\|e\|^2_{\mathcal{H}}. \end{aligned}$$

From this the error estimate (3.33) follows while the L^2 estimate (3.34) follows by combining (3.33) with (3.38). ∎

Remark 3.23 *In [30], [32], it is proved for the case of constant wave number k, that for S as in Section 3.2.3, i.e., hp-finite elements, the conditions*

$$p = O(\log(k)) \qquad \text{and} \qquad \frac{kh}{p} = O(1) \tag{3.39}$$

imply (3.31) and lead to the "minimal" finite element space for discretization of the Helmholtz equations. In this light, terms in the a-posteriori error estimates which grow polynomially in p are expected to grow, at most, logarithmically with respect to k and, hence, are moderately bounded, also for large wavenumbers.

3.4 Analysis of the A Posteriori Error Estimator

3.4.1 Estimate of the Adjoint Approximability

The constant in the a posteriori error estimate will contain the term $\eta_k^\star(S)$ as a factor. In order to get an explicit upper bound, an a priori estimate of the quantity is required which can be found for constant wavenumber in [30, Theorem 5.5] and [32, Prop. 5.3, Prop. 5.6]. Here we will outline the principal ideas of the analysis and refer for the more general setting to these two papers.

We restrict in this section 3.4.1 to the following model problem (cf. (3.8)):

- Ω is the unit ball in $\mathbb{R}^3$,
- $f \in L^2(\Omega)$ satisfies Assumption 3.1
- the wavenumber $k \geq 1$ is constant.

We consider the problem Find $u \in \mathcal{H}$ (with $\mathcal{H}$ as in (3.7)) such that

$$\int_\Omega \left(\langle \nabla u, \nabla \bar{v}\rangle - k^2 u\bar{v}\right) - \int_{\Gamma^{\text{out}}} (T_k u)\,\bar{v} = \int_\Omega f\bar{v} \qquad \text{for all } v \in \mathcal{H}. \tag{3.40}$$

The exact solution of (3.40) can be written as the (localized) acoustic volume potential. For this, let $\mu \in C^\infty(\mathbb{R}_{\geq 0})$ be a cutoff function such that

$$\begin{aligned} &\operatorname{supp}\mu \subset [0,4], && \mu|_{[0,2]} = 1, \quad |\mu|_{W^{1,\infty}(\mathbb{R}_{\geq 0})} \leq C, \\ &\forall x \in \mathbb{R}_{\geq 0} : 0 \leq \mu(x) \leq 1, && \mu|_{[4,\infty[} = 0, \quad |\mu|_{W^{2,\infty}(\mathbb{R}_{\geq 0})} \leq C, \end{aligned} \tag{3.41}$$

and let $g_k(r) := g_k(r) := \frac{\mathrm{e}^{\mathrm{i}kr}}{4\pi r}$. Define $G_k(z) := g_k(\|z\|)\,\mu(\|z\|)$ as the product of the fundamental solution to the operator $\mathcal{L}_k := -\Delta - k^2$ with the cutoff function. Then, the solution of (3.40) can be written by

$$u(x) := (N_k f)(x) := \int_\Omega G_k(x-y)\, f(y)\, dy \qquad \forall x \in \Omega.$$

The key ingredient of the analysis of the adjoint approximability is the following decomposition result:

Lemma 3.24 (decomposition lemma) *Let Ω be the unit ball in $\mathbb{R}^3$. Then there exists a constant $C > 0$ such that for $f \in L^2(\Omega)$ the function v given by*

$$v(x) = N_k f(x) = \int_\Omega G_k(x-y) f(y)\, dy, \qquad x \in \Omega,$$

satisfies

$$k^{-1}\|v\|_{H^2(\Omega)} + \|v\|_{H^1(\Omega)} + k\|v\|_{L^2(\Omega)} \leq C\|f\|_{L^2(\Omega)}. \tag{3.42}$$

Furthermore, for every $\lambda > 1$, there exists a λ- and k-dependent splitting $v = v_{H^2} + v_{\mathcal{A}}$ with

$$\|\nabla^p v_{H^2}\|_{L^2(\Omega)} \leq C\left(1 + \frac{1}{\lambda^2 - 1}\right)(\lambda k)^{p-2}\,\|f\|_{L^2(\Omega)} \qquad \forall p \in \{0,1,2\}, \tag{3.43a}$$

$$\|\nabla^p v_{\mathcal{A}}\|_{L^2(\Omega)} \leq C\lambda\left(\sqrt{3}\lambda k\right)^{p-1}\,\|f\|_{L^2(\Omega)} \qquad \forall p \in \mathbb{N}_0. \tag{3.43b}$$

Here, $\nabla^p v_{\mathcal{A}}$ stands for a sum over all derivatives of order p (see (3.25) for details).

Remark 3.25 *For $f \in L^2(\Omega)$ the function $v = N_k(f)$ cannot be expected to have more Sobolev regularity than H^2. The decomposition $v = v_{H^2} + v_{\mathcal{A}}$ of Lemma 3.24 splits v into an H^2-regular part v_{H^2} and an analytic part $v_{\mathcal{A}}$. The essential feature of this splitting is that the H^2-part v_{H^2} has a better H^2-regularity constant in terms of k than v itself, namely, (3.43a), (3.43b), and the triangle inequality $\|\nabla^2 v\|_{L^2(\Omega)} \leq \|\nabla^2 v_{H^2}\|_{L^2(\Omega)} + \|\nabla^2 v_{\mathcal{A}}\|_{L^2(\Omega)}$ imply*

$$\|\nabla^2 v_{H^2}\|_{L^2(\Omega)} \leq C\|f\|_{L^2(\Omega)} \quad \textit{versus} \quad \|\nabla^2 v\|_{L^2(\Omega)} \leq Ck\|f\|_{L^2(\Omega)}.$$

The fact that $\|v_{H^2}\|_{H^2} \leq C\|f\|_{L^2}$ for a $C > 0$ independent of k will be be essential for the stability and convergence analysis below.

Proof of Lemma 3.24. The estimates for v follow directly from those for v_{H^2} and $v_{\mathcal{A}}$ by fixing a parameter $\lambda > 1$. In order to construct the splitting $v = v_{H^2} + v_{\mathcal{A}}$, we start by recalling the definition of the Fourier transform for functions with compact support

$$\hat{u}\,(\xi) = (2\pi)^{-3/2} \int_{\mathbb{R}^3} \mathrm{e}^{-\,\mathrm{i}\langle \xi, x\rangle}\, u\,(x)\, dx \qquad \forall \xi \in \mathbb{R}^3$$

and the inversion formula

$$u\,(x) = (2\pi)^{-3/2} \int_{\mathbb{R}^3} \mathrm{e}^{\mathrm{i}\langle x, \xi\rangle}\, \hat{u}\,(\xi)\, d\xi \qquad \forall x \in \mathbb{R}^3.$$

We will define a decomposition of v_μ (which will determine the decomposition of v on B_Ω) by decomposing its Fourier transform, i.e.,

$$\widehat{v}_\mu = \widehat{v}_{H^2} + \widehat{v}_{\mathcal{A}}. \tag{3.44}$$

In order to define the two terms on the right-hand side of (3.44), we let $B_{\lambda k}(0)$ denote the ball of radius λk centred at the origin, where $\lambda > 1$ is the fixed constant (independent of k) selected in the statement of the lemma. The characteristic function of $B_{\lambda k}(0)$ is denoted by $\chi_{\lambda k}$. The Fourier transform of f is then decomposed as

$$\widehat{f} = \widehat{f}\chi_{\lambda k} + (1 - \chi_{\lambda k})\widehat{f} =: \widehat{f}_k + \widehat{f}_k^c.$$

By the inverse Fourier transformation, this decomposition of $\widehat{f}$ entails a decomposition of f into f_k and f_k^c given by

$$f_k(x) := (2\pi)^{-3/2} \int_{\mathbb{R}^3} \mathrm{e}^{\mathrm{i}\langle x,\xi\rangle}\, \chi_{\lambda k}(\xi)\, \hat{f}(\xi)\, d\xi \quad \text{and} \quad f_k^c(x) := f - f_k. \tag{3.45}$$

Accordingly, we define the decomposition of v_μ by

$$v_{\mu,H^2} := G_k \star f_k^c \quad \text{and} \quad v_{\mu,\mathcal{A}} := G_k \star f_k, \tag{3.46}$$

where "$\star$" denotes the convolution in $\mathbb{R}^3$. The functions v_{H^2} and $v_{\mathcal{A}}$ in (3.44) are then obtained by setting $v_{H^2} := v_{\mu,H^2}|_\Omega$ and $v_{\mathcal{A}} := v_{\mu,\mathcal{A}}|_\Omega$. We will obtain the desired estimates by showing the following, stronger estimates:

$$\|v_{\mu,H^2}\|_{H^2(\mathbb{R}^3)} \le C\|f\|_{L^2(\mathbb{R}^3)}, \tag{3.47a}$$

$$\|D^\alpha v_{\mu,\mathcal{A}}\|_{L^2(\mathbb{R}^3)} \le C\lambda\,(\lambda k)^{|\alpha|-1}\,\|f\|_{L^2(\mathbb{R}^3)}, \qquad \forall \alpha \in \mathbb{N}_0^3. \tag{3.47b}$$

The estimates (3.47) are obtained by Fourier techniques. To that end, we compute the Fourier transform of G_k:

$$\begin{aligned}
\widehat{G_k}(\xi) &= (2\pi)^{-3/2} \int_{\mathbb{R}^3} \mathrm{e}^{-\mathrm{i}\langle \xi,x\rangle}\, G_k(x)\, dx \\
&= (2\pi)^{-3/2} \int_0^\infty g_k(r)\,\mu(r)\, r^2 \left(\int_{\mathbb{S}_2} \mathrm{e}^{-\mathrm{i}\, r\langle \xi,\zeta\rangle}\, dS_\zeta \right) dr \\
&=: (2\pi)^{-3/2}\, \iota(\|\xi\|).
\end{aligned} \tag{3.49}$$

The inner integral in (3.49) can be evaluated analytically and we obtain

$$\iota(s) = 4\pi \int_0^\infty g_k(r)\,\mu(r)\, r^2 \frac{\sin(rs)}{(rs)} dr. \tag{3.50}$$

Applying the Fourier transform to the convolutions (3.46) leads to

$$\begin{aligned}
\widehat{v}_{\mu,H^2} &= (2\pi)^{3/2}\, \widehat{G_k}\widehat{f_k^c} = (2\pi)^{3/2}\, \widehat{G_k}\widehat{f}(1-\chi_{\lambda k}), \\
\widehat{v}_{\mu,\mathcal{A}} &= (2\pi)^{3/2}\, \widehat{G_k}\widehat{f_k} = (2\pi)^{3/2}\, \widehat{G_k}\widehat{f}\chi_{\lambda k}.
\end{aligned}$$

To estimate higher order derivatives of v_{μ,H^2} and $v_{\mu,\mathcal{A}}$ we define for a multi-index $\alpha \in \mathbb{N}_0^3$ the function $P_\alpha : \mathbb{R}^3 \to \mathbb{R}^3$ by $P_\alpha(\xi) := \xi^\alpha$ and obtain – by using standard properties of the Fourier transformation and the support properties of $\chi_{\lambda k}$ – for all $|\alpha| \le 2$

$$\begin{aligned}
\|\partial^\alpha v_{\mu,H^2}\|_{L^2(\mathbb{R}^3)} &= (2\pi)^{3/2} \left\| P_\alpha \widehat{G_k M}\,(1-\chi_{\lambda k})\,\widehat{f} \right\|_{L^2(\mathbb{R}^3)} \\
&\le (2\pi)^{3/2} \left(\max_{\xi\in\mathbb{R}^3:|\xi|\ge\lambda k} |P_\alpha I(\xi)| \right) \left\| (1-\chi_{\lambda k})\,\widehat{f} \right\|_{L^2(\mathbb{R}^3)} \\
&\le (2\pi)^{3/2} \left(\max_{s\ge\lambda k} \left| s^{|\alpha|}\iota(s) \right| \right) \|f\|_{L^2(\Omega)}.
\end{aligned} \tag{3.51}$$

The symbol $\iota(\cdot)$ is estimated in [30]. More precisely, [30, Lemma 3.7 (iv)] implies

$$\sup_{|s|\ge\lambda k} s^2\,|\iota(s)| \le C\left(1 + \frac{1}{\lambda^2 - 1}\right)$$

from which we conclude that

$$\max_{s\geq\lambda k}\left|s^{|\alpha|}\iota\left(s\right)\right|\leq C\left(\lambda k\right)^{|\alpha|-2}\left(1+\frac{1}{\lambda^2-1}\right)$$

holds for $|\alpha|\in\{0,1,2\}$. Thus,

$$\left\|\partial^{\alpha}v_{H^2}\right\|_{L^2(B_\Omega)}\leq C\left(\lambda k\right)^{|\alpha|-2}\left(1+\frac{1}{\lambda^2-1}\right)\left\|f\right\|_{L^2(\Omega)}$$

and (3.43a) follows.

Completely analogously, we derive for all $\alpha\in\mathbb{N}_0^3$

$$\left\|\partial^{\alpha}v_{\mu,\mathcal{A}}\right\|_{L^2(\mathbb{R}^3)}\leq\left(2\pi\right)^{3/2}\left(\max_{0\leq s\leq\lambda k}\left|s^{|\alpha|}\iota\left(s\right)\right|\right)\left\|f\right\|_{L^2(\Omega)}. \tag{3.52}$$

The proof of the lemma is completed by using the bound on the function ι:

$$\sup_{|s|\leq\lambda k}\left|s\right|^m\left|\iota\left(s\right)\right|\leq C\lambda\left(\lambda k\right)^{m-1}\quad\forall\lambda>0\quad\forall m\in\mathbb{N}_0$$

given in [30, Lemma 3.7 (v)] and using (3.25). ∎

The adjoint approximability is related to (3.40) via the adjoint problem: For given $f\in L^2\left(B_R\right)$, find $z\in H^1\left(\Omega\right)$ such that

$$A_{\mathrm{DtN}}^{\star}\left(z,v\right)=\left(v,f\right)_{L^2(B_R)}\qquad\forall v\in H^1\left(B_R\right) \tag{3.53}$$

Explicitly we have

$$A_{\mathrm{DtN}}^{\star}\left(z,v\right):=\int_{\Omega}\left\langle\nabla u,\nabla\bar{v}\right\rangle-k^2u\bar{v}-\int_{\partial B_R}u\left(\overline{T_kv}\right).$$

The solution of the adjoint problem can be expressed via the complex conjugate of the fundamental solution as

$$z\left(x\right)=\left(N_k^{\star}f\right)\left(x\right)=\int_{\Omega}\overline{G_k}\left(x-y\right)\bar{f}\left(y\right)dy\qquad\forall x\in\Omega.$$

For the estimate of the adjoint approximability in the context of *hp*-finite elements we have to investigate the approximability of the solutions $N_k^{\star}f$ for $f\in L^2\left(\Omega\right)$ by these finite elements.

For meshes $\mathcal{T}_h$ satisfying Assumption 3.9 with element maps F_K we denote the usual space of piecewise (mapped) polynomials by $S^{p,1}(\mathcal{T}_h):=\{u\in H^1(\Omega)\,|\,\forall K\in\mathcal{T}_h\colon u|_K\circ F_K\in\mathbb{P}_p\}$, where $\mathbb{P}_p$ denotes the space of polynomials of degree p. It is desirable to construct an approximant $Iu\in S^{p,1}(\mathcal{T}_h)$ of a given (sufficiently smooth) function u in an elementwise fashion. The C^0-continuity of an elementwise defined approximant Iu is most conveniently ensured if Iu is defined in such a way that for every topological entity E of the mesh (i.e., E is an element K, a face f, an edge e, or a vertex V) the restriction $(Iu)|_E$ is fully determined by $u|_{\overline{E}}$. There are many ways of realizing this construction principle. The construction employed in the present paper is based on the following concept.

Definition 3.26 (element-by-element construction) *Let $\widehat{K}$ be the reference simplex in $\mathbb{R}^d$, $d \in \{2,3\}$. A polynomial π is said to permit* an element-by-element *construction of polynomial degree p for $u \in H^s(\widehat{K})$, $s > d/2$, if:*

1. *$\pi(V) = u(V)$ for all $d+1$ vertices V of $\widehat{K}$,*
2. *for every edge e of $\widehat{K}$, the restriction $\pi|_e \in \mathbb{P}_p$ is the unique minimizer of*

$$\pi \mapsto p^{1/2}\|u-\pi\|_{L^2(e)} + \|u-\pi\|_{H^{1/2}_{00}(e)} \tag{3.54}$$

   under the constraint that π satisfies (1)[4];

3. *(for $d=3$) for every face f of $\widehat{K}$, the restriction $\pi|_f \in \mathbb{P}_p$ is the unique minimizer of*

$$\pi \mapsto p\|u-\pi\|_{L^2(f)} + \|u-\pi\|_{H^1(f)} \tag{3.56}$$

   under the constraint that π satisfies (1), (2) for all vertices and edges of the face f.

We are now in position to show that the solution $v = N_k^\star f$ can be approximated well by the FEM space $S^{p,1}(\mathcal{T}_h)$ provided that kh/p is sufficiently small and $p \geq c \ln k$.

Theorem 3.27 *Let $d \in \{1,2,3\}$ and $\Omega \subset \mathbb{R}^d$ be a bounded domain. Then there exist constants C, $\sigma > 0$ that depend solely on the constants appearing in Assumption 3.9 such that for every $f \in L^2(\Omega)$ the function $v := N_k^\star f$ satisfies*

$$\inf_{w \in S^{p,1}(\mathcal{T}_h)} k\|v-w\|_{\mathcal{H}} \leq C\|f\|_{L^2(\Omega)} \left(1 + \frac{kh}{p}\right) \left\{ \frac{kh}{p} + k\left(\frac{kh}{\sigma p}\right)^p \right\}.$$

Proof. We will only prove here the cases $d \in \{2,3\}$.

We note $v = N_k^\star f = \overline{N_k \overline{f}}$, fix $\lambda > 1$ in Lemma 3.24, and split with its aid $v = v_{H^2} + v_{\mathcal{A}}$ with $v_{H^2} \in H^2(\Omega)$ and $v_{\mathcal{A}}$ analytic; we have the following bounds

$$\|v_{H^2}\|_{H^2(\Omega)} \leq C\|f\|_{L^2(\Omega)}, \qquad \|\nabla^p v_{\mathcal{A}}\|_{L^2(\Omega)} \leq C(\lambda k)^{p-1}\|f\|_{L^2(\Omega)} \qquad \forall p \in \mathbb{N}_0.$$

We approximate v_{H^2} and $v_{\mathcal{A}}$ separately. [30, Theorem B.4] and a scaling argument provides an approximant $w_{H^2} \in S^{p,1}(\mathcal{T}_h)$ such that for every $K \in \mathcal{T}_h$ we have, for $q = 0, 1$,

$$\|v_{H^2} - w_{H^2}\|_{H^q(K)} \leq C\left(\frac{h}{p}\right)^{2-q} \|v_{H^2}\|_{H^2(K)} \qquad \forall K \in \mathcal{T}_h.$$

Hence, by summation over all elements, we arrive at

$$k\|v_{H^2} - w_{H^2}\|_{\mathcal{H}} \leq C\left(\frac{kh}{p} + \left(\frac{kh}{p}\right)^2\right)\|f\|_{L^2(\Omega)}.$$

[4]We recall the definition of the Sobolev space $H^{1/2}_{00}(\Omega)$. If Ω is an edge or a face of a triangle or a tetrahedron, then the Sobolev norm $\|\cdot\|_{H^{1/2}_{00}(\Omega)}$ is defined by

$$\|u\|^2_{H^{1/2}_{00}(\Omega)} := \|u\|^2_{H^{1/2}(\Omega)} + \left\| \frac{u}{\sqrt{\operatorname{dist}(\cdot, \partial\Omega)}} \right\|^2_{L^2(\Omega)}, \tag{3.55}$$

and the space $H^{1/2}_{00}(\Omega)$ is the complection of $C_0^\infty(\Omega)$ under this norm.

We now turn to the approximation of $v_{\mathcal{A}}$. Again, we construct the approximation $w_{\mathcal{A}} \in S^{p,1}(\mathcal{T}_h)$ in an element-by-element fashion. We start by defining for each element $K \in \mathcal{T}_h$ the constant C_K by

$$C_K^2 := \sum_{p \in \mathbb{N}_0} \frac{\|\nabla^p v_{\mathcal{A}}\|_{L^2(K)}^2}{(2\lambda k)^{2p}} \tag{3.57}$$

and we note

$$\|\nabla^p v_{\mathcal{A}}\|_{L^2(K)} \leq (2\lambda k)^p C_K \qquad \forall p \in \mathbb{N}_0, \tag{3.58}$$

$$\sum_{K \in \mathcal{T}_h} C_K^2 \leq \frac{4}{3} \left(\frac{C}{\lambda k} \right)^2 \|f\|_{L^2(\Omega)}^2. \tag{3.59}$$

Let the element map for K be $F_K = R_K \circ A_K$. From [30, Lemma C.1] we conclude that the function $\tilde{v} := v_{\mathcal{A}}|_K \circ R_K$ satisfies, for suitable constants $\tilde{C}$, C (which depend additionally on the constants describing the analyticity of the element maps R_K)

$$\|\nabla^p \tilde{v}\|_{L^2(\widetilde{K})} \leq C \tilde{C}^p \max\{p,k\}^p C_K \qquad \forall p \in \mathbb{N}_0.$$

Since A_K is affine, the function $\hat{v} := v_{\mathcal{A}}|_K \circ F_K = \tilde{v} \circ A_K$ therefore satisfies

$$\|\nabla^p \hat{v}\|_{L^2(\widehat{K})} \leq C h^{-d/2} \tilde{C}^p h^p \max\{p,k\}^p C_K \qquad \forall p \in \mathbb{N}_0.$$

Hence, the assumptions of [30, Lemma C.1] (with $R = 1$ there) are satisfied, and we get an approximation w on the element K by lifting an element-by-element construction on $\widehat{K}$ to K via F_K which satisfies for $q \in \{0,1\}$

$$\|v_{\mathcal{A}} - w\|_{H^q(K)} \leq C h^{d/2-q} h^{-d/2} C_K \left\{ \left(\frac{h}{h+\sigma} \right)^{p+1} + \left(\frac{kh}{\sigma p} \right)^{p+1} \right\}.$$

Summation over all elements $K \in \mathcal{T}_h$ gives

$$\|v_{\mathcal{A}} - w\|_{\mathcal{H}}^2 \leq \left[\left(\frac{h}{h+\sigma} \right)^{2p} + k^2 \left(\frac{h}{h+\sigma} \right)^{2p+2} + \frac{k^2}{p^2} \left(\frac{kh}{\sigma p} \right)^{2p} + k^2 \left(\frac{kh}{\sigma p} \right)^{2p+2} \right] \sum_{K \in \mathcal{T}_h} C_K^2. \tag{3.60}$$

The combination of (3.60) and (3.59) yields

$$k\|v_{\mathcal{A}} - w\|_{\mathcal{H}} \leq C \left[\left(\frac{h}{h+\sigma} \right)^p \left(1 + \frac{hk}{h+\sigma} \right) + k \left(\frac{kh}{\sigma p} \right)^p \left(\frac{1}{p} + \frac{kh}{\sigma p} \right) \right] \|f\|_{L^2(\Omega)}.$$

Furthermore, we estimate using $h \leq \mathrm{diam}\Omega$ and $\sigma > 0$ (independent of h)

$$\left(\frac{h}{h+\sigma} \right)^p \left(1 + \frac{kh}{\sigma + h} \right) \leq Ch(1+kh) \left(\frac{h}{\sigma+h} \right)^{p-1} \leq Ch(1+kh)p^{-2} \leq C \frac{h}{p} \left(\frac{1}{p} + \frac{kh}{p} \right).$$

We therefore arrive at

$$k\|v_{\mathcal{A}} - w\|_{\mathcal{H}} \leq C \left(\frac{1}{p} + \frac{kh}{p} \right) \left[\frac{kh}{p} + k \left(\frac{kh}{\sigma p} \right)^p \right] \|f\|_{L^2(\Omega)},$$

which completes the proof of the theorem. ∎

Combining Theorems 3.27, 3.22 produces the condition

$$\frac{kh}{p} + k\left(\frac{kh}{\sigma p}\right)^p \leq C$$

for quasi-optimality of the *hp*-FEM. We extract from Theorem 3.27 that quasi-optimality of the *h*-version FEM can be achieved under the side condition that $p \geq C \log k$:

Corollary 3.28 *Let Ω be the unit ball in $\mathbb{R}^3$. Let Assumption 3.9 be valid. Then there exist constants c_1, $c_2 > 0$ independent of k, h, and p such that (3.31) is implied by the following condition:*

$$\frac{kh}{p} \leq c_1 \qquad \textit{together with} \qquad p \geq c_2 \ln k. \tag{3.61}$$

Alternatively, the discrete stability follows from

$$p = O(1) \ \textit{fixed independent of } k \quad \textit{and} \quad kh + k(kh)^p \leq C \tag{3.62}$$

which is understood as a condition on the maximal step size h.

Proof. Theorem 3.27 implies

$$\eta(S) \leq C\left(1 + \frac{kh}{p}\right)\left(\frac{kh}{p} + k\left(\frac{kh}{\sigma p}\right)^p\right).$$

The right-hand side needs to be bounded by $1/C_c$. It is now easy to see that we can select c_1, c_2 such that this can be ensured. ∎

An easy consequence of the stability result Corollary 3.28 is:

Corollary 3.29 *Let the assumptions of Corollary 3.28 be satisfied and let (3.61) or (3.62) hold. Then, the Galerkin solution u_S exists and satisfies the error estimate*

$$\|u - u_S\|_{\mathcal{H}} \leq C_c\left(\frac{h}{p} + \left(\frac{kh}{\sigma p}\right)^p\right)\|f\|_{L^2(\Omega)}.$$

Remark 3.30 *To the best of the authors' knowledge, discrete stability in 2D and 3D has only been shown under much more restrictive conditions than (3.61), e.g., the condition $k^2 h \lesssim 1$. Even in one dimension, condition (3.61) improves the stability condition $kh \lesssim 1$ that was required in [26].*

3.4.2 Reliability

According to Assumption 3.12 the exact solution $u \in \mathcal{H}$ and the Galerkin solution $u_S \in S$ of (3.19) and (3.23), respectively, exist. In view of inequality (3.16), we estimate the error $e = u - u_S$, $\operatorname{Re}(a(e,e) - b(e,e))$, and $\|k_+ e\|_{L^2(\Omega)}$ separately in terms of $\eta(u_S, \alpha)$.

Lemma 3.31 *Let Assumption 3.12 be satisfied. Assume that there exists a linear and bounded linear operator $I_S : \mathcal{H} \to S$ as in Assumption 3.16. Then there is a constant $C_1 > 0$, that depends only on the shape-regularity of the grid (cf. Remark 3.10), such that*

$$|\operatorname{Re}(a(e,e) - b(e,e))| \leq C_1 \eta(u_S, \alpha)\|e\|_{\mathcal{H};\Omega}$$

with α as in (3.29).

Proof. Using the solution properties and integration by parts yields the error representation

$$\begin{aligned} a(e,e) - b(e,e) &= a(e, e - I_S e) - b(e, e - I_S e) \\ &= \sum_{K\in\mathcal{T}} \int_K \operatorname{res}(u_S)(e - I_S e) + \sum_{E\in\mathcal{E}} \int_E \operatorname{Res}(u_S)(e - I_S e). \end{aligned}$$

We use the assumed interpolation estimates (3.29) and get with the Cauchy–Schwarz inequality

$$\begin{aligned} &|\operatorname{Re}\left(a(e,e) - b(e,e)\right)| \\ &\quad\le \left(\sum_{K\in\mathcal{T}} \alpha_K^2 \left\|\operatorname{res}\left(u_S\right)\right\|_{L^2(K)}^2 + \sum_{E\in\mathcal{E}} \alpha_E^2 \left\|\operatorname{Res}\left(u_S\right)\right\|_{L^2(E)}^2\right)^{1/2} \left(\sum_{K\in\mathcal{T}} \|e\|_{\mathcal{H};\omega_K^4}^2\right)^{1/2} \\ &\quad\le C_1 \eta(u_S,\alpha) \|e\|_{\mathcal{H};\Omega}. \end{aligned}$$

■

Lemma 3.32 *Let Assumptions 3.12 and 3.16 be satisfied. Then, with C_1 from Lemma 3.31 and $\eta_k^\star(S)$ as in* (3.30)

$$\|k_+ e\|_{L^2(\Omega)} \le C_1 \eta_k^\star(S) \eta(u_S,\alpha). \tag{3.63}$$

Proof. We define z by (3.17) with $f := k_+^2 e$. Let $z_S \in S$ denote the best approximation of z with respect to the $\|\cdot\|_{\mathcal{H};\Omega}$-norm. We have, by using Galerkin's orthogonality and the arguments as in the proof of Lemma 3.31,

$$\begin{aligned} \|k_+ e\|_{L^2(\Omega)}^2 &= a(e,z) - b(e,z) = a(e, z - z_S) - b(e, z - z_S) \\ &= \sum_{K\in\mathcal{T}} \int_K \operatorname{res}(u_S)(z - z_S) + \sum_{E\in\mathcal{E}} \int_E \operatorname{Res}(u_S)(z - z_S). \end{aligned}$$

We further follow the arguments of the mentioned proof and, by using the definition of $\eta_k^\star(S)$, we get

$$\|k_+ e\|_{L^2(\Omega)}^2 \le C_1 \eta(u_S,\alpha) \|z - z_S\|_{\mathcal{H};\Omega} \le C_1 \eta(u_S,\alpha) \eta_k^\star(S) \|k_+ e\|_{L^2(\Omega)}$$

and this gives (3.63). ■

Theorem 3.33 (Reliability estimate) *Let Assumptions 3.12 and 3.16 be satisfied. Then, with C_1 from Lemma 3.31,*

$$\|e\|_{\mathcal{H};\Omega} \le \frac{1}{\gamma_{\mathrm{ell}}} C_1 \left(1 + (\gamma_{\mathrm{ell}}\theta)^{1/2} \eta_k^\star(S)\right) \eta(u_S,\alpha).$$

Proof. The combination of (3.16), (3.63) with the bounds obtained in Lemma 3.31 and 3.32 yields

$$\begin{aligned} \gamma_{\mathrm{ell}} \|e\|_{\mathcal{H};\Omega}^2 &\le \operatorname{Re}\left(a(e,e) - b(e,e)\right) + \theta \|k_+ e\|_{L^2(\Omega)}^2 \\ &\le C_1 \eta(u_S,\alpha) \|e\|_{\mathcal{H};\Omega} + \theta C_1^2 \eta_k^\star(S)^2 \eta(u_S,\alpha)^2 \end{aligned}$$

so that

$$\begin{aligned}\|e\|_{\mathcal{H};\Omega} &\leq \frac{1}{\gamma_{\text{ell}}} C_1 \eta(u_S, \alpha) + \left(\frac{\theta}{\gamma_{\text{ell}}}\right)^{1/2} C_1 \eta_k^\star(S) \eta(u_S, \alpha) \\ &\leq \frac{1}{\gamma_{\text{ell}}} C_1 \left(1 + (\gamma_{\text{ell}}\theta)^{1/2} \eta_k^\star(S)\right) \eta(u_S, \alpha).\end{aligned}$$

■

In the previous arguments res and Res were defined with exact data functions f, k. If we define $\widetilde{\eta}$ in terms of $\widetilde{\text{res}}$ and $\widetilde{\text{Res}}$, where f, k have been replaced by polynomial approximations $\widetilde{f}, \widetilde{k}$ the results holds with the following modification.

Corollary 3.34 *Let $\widetilde{f}, \widetilde{k}$ be approximations to f, k. Then*

$$\begin{aligned}\eta(u_S, \alpha) \leq \sqrt{3} \Bigg(&\widetilde{\eta}(u_S, \alpha) + \left(\sum_{K \in \mathcal{K}} \alpha_K^2 \|f - \widetilde{f}\|_{L^2(K)}^2 \right)^{1/2} \\ &+ \left(\sum_{K \in \mathcal{K}} \alpha_K^2 \|(k^2 - \widetilde{k}^2) u_S\|_{L^2(K)}^2 \right)^{1/2} \Bigg).\end{aligned}$$

Proof. We notice

$$\begin{aligned}\text{res}\,(u_S) &= f + k^2 u_S + \Delta u_S = \widetilde{f} + \widetilde{k}^2 u_S + \Delta u_S + f - \widetilde{f} + (k^2 - \widetilde{k}^2) u_S \\ &= \widetilde{\text{res}}\,(u_S) + f - \widetilde{f} + (k^2 - \widetilde{k}^2) u_S \\ \text{Res}\,(u_S) &= -\partial_n u_S + \mathrm{i}\, k u_S = \widetilde{\text{Res}}\,(u_S)\end{aligned}$$

since k is constant on Γ^{out}. We thus obtain

$$\begin{aligned}\eta(u_S, \alpha)^2 \leq 3\widetilde{\eta}(u_S, \alpha)^2 &+ 3 \sum_{K \in \mathcal{K}} \alpha_K^2 \|f - \widetilde{f}\|_{L^2(K)}^2 \\ &+ 3 \sum_{K \in \mathcal{K}} \alpha_K^2 \|(k^2 - \widetilde{k}^2) u_S\|_{L^2(K)}^2.\end{aligned}$$

■

An explicit estimate of the error by the error estimator requires an upper bound for the adjoint approximation property $\eta_k^\star(S)$. Such estimates for *hp*-finite elements spaces for constant wavenumbers k are derived in [30] and [32] for problem (3.8) and (3.10). We summarize the results as the following corollaries.

Corollary 3.35 (Robin boundary conditions) *Consider problem (3.10) with constant wavenumber k, where $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, is a bounded domain with analytic boundary. We use the approximation space S described in Section 3.2.3. Let $f \in L^2(\Omega)$ and $k \geq k_0 > 1$ and assume that $\Gamma^{\text{in}} = \emptyset$, i.e., we consider the pure Robin problem. Let Assumption 3.12 (a) and (b) as well as Assumption 3.16 be satisfied. Then there exist constants δ, $\tilde{c} > 0$ that are independent of h, p, and k such that the conditions*

$$\frac{kh}{p} \leq \delta \quad \textit{and} \quad p \geq 1 + \tilde{c} \log(k)$$

imply the k-independent a posteriori error estimate

$$\|e\|_{\mathcal{H};\Omega} \leq \frac{1}{\gamma_{\text{ell}}} C_1 \left(1 + (\gamma_{\text{ell}}\theta)^{1/2} \check{C}\right) \eta(u_S, \alpha),$$

where $\check{C}$ only depends on δ and $\tilde{c}$.

Corollary 3.36 (DtN boundary conditions) *Consider problem (3.8) for constant wavenumber k, where Ω has an analytic boundary. Let Assumption 3.12 (a) and (b) as well as Assumption 3.16 be satisfied and assume that the constant C_k^{adj} in (3.18) grows at most polynomially in k, i.e., there exists some $\beta \geq 0$ such that*[5] *$C_k^{\text{adj}} \leq Ck^\beta$. Let $f \in L^2(\Omega)$ and $k \geq k_0 > 1$. Then there exist constants δ, $\tilde{c} > 0$ that are independent of h, p, and k such that the conditions*

$$\frac{kh}{p} \leq \delta \quad \textit{and} \quad p \geq 1 + \tilde{c}\log(k)$$

imply the k-independent a posteriori error estimate

$$\|e\|_{\mathcal{H};\Omega} \leq \frac{1}{\gamma_{\text{ell}}} C_1 \left(1 + (\gamma_{\text{ell}}\theta)^{1/2} \check{C}\right) \eta(u_S, \alpha)$$

where $\check{C}$ only depends on δ and $\tilde{c}$.

3.4.3 Efficiency

The localized version of the error estimator is given by

$$\eta_K(v, \alpha) := \left(\alpha_K^2 \left\|\text{res}\,(v)\right\|_{L^2(K)}^2 + \frac{1}{2} \sum_{E \in \mathcal{E}(K)} \alpha_E^2 \left\|\text{Res}\,(v)\right\|_{L^2(E)}^2 \right)^{1/2},$$

where $\mathcal{E}(K) := \{E \in \mathcal{E} : E \subset \partial K\}$. Note that $\eta(v, \alpha) = \sqrt{\sum_{K \in \mathcal{T}} \eta_K^2(v, \alpha)}$.

In view of Corollary 3.34 let us define approximations $\widetilde{f}, \widetilde{k}$ to f, k, respectively, as local $L^2(K)$-projections onto a polynomial of degree p_K (or some $q_K \sim p_K$). In this case we use the notation $\widetilde{\text{res}}$ and $\widetilde{\eta}$ accordingly. Also we set

$$k_{K,+} := \max\{\|k\|_{L^\infty(K)}, 1\}$$

and, for any subset $\omega \subset \Omega$,

$$\delta_\omega^2 := \left\|f - \widetilde{f}\right\|_{L^2(\omega)}^2 + \left\|\left(k^2 - \widetilde{k}^2\right) u_S\right\|_{L^2(\omega)}^2.$$

Theorem 3.37 *Let Assumptions 3.12 and (3.11) be satisfied and let the mesh be shape regular (cf. Remark 3.10). We assume that Ω is either an interval ($d = 1$), or a polygonal domain ($d = 2$), or a Lipschitz polyhedron ($d = 3$), and that the element maps F_K are affine. We assume the resolution condition:*

$$\frac{k_{K,+} h_K}{p_K} \lesssim 1 \qquad \textit{for all } K \in \mathcal{T}. \tag{3.64}$$

[5] See [23] for sufficient conditions on the domain which implies this growth condition.

Then, there exists a constant C depending only on the constants in Assumption 3.9 and 3.5 — and in particular, is independent of k, p_K, h_K and u, u_S — so that

$$\widetilde{\eta}_K(u_S,\alpha) \leq C p_K^{3/2} \left(\alpha_K \frac{p_K}{h_K} + \alpha_E \left(\frac{p_K}{h_K}\right)^{1/2}\right) \left(\|u-u_S\|_{\mathcal{H};\omega_K} + \frac{\delta_{\omega_K}}{k_{K,+}}\right), \tag{3.65}$$

where α_K, α_E are weights in (3.28) such that (3.29a) and (3.29b) hold[6]*. For $d=2$, the choices as in Corollary 3.18 lead to*

$$\widetilde{\eta}_K(u_S,\alpha) \leq C p_K^{3/2} \left(\|u-u_S\|_{\mathcal{H};\omega_K} + \frac{\delta_{\omega_K}}{k_{K,+}}\right). \tag{3.66}$$

Proof. We apply the results [33, Lem. 3.4, 3.5]. There, the proofs are given for two space dimensions, i.e., $d=2$. They carry over to the case $d=1$ simply by using [33, Lem. 2.4] instead of [33, Thm. 2.5]. For the case $d=3$, are careful inspection of the proofs in [33, Thm. 2.5] (which is given in [34, Thm. D2]) and [33, Lem. 2.6] shows that these lemmata also hold for $d=3$. Hence, the proof of [33, Lem. 3.4, 3.5] can be used verbatim for the cases $d=1$ and $d=3$. We choose $\alpha=0$ in [33, Lem. 3.4, 3.5]. Following these lines of arguments we get for any $\varepsilon>0$, $K\in\mathcal{T}$, $E\in\mathcal{E}(K)$,

$$\begin{aligned}
&\frac{h_K^2}{p_K^2}\|\widetilde{\operatorname{res}}(u_S)\|_{L^2(K)}^2 \\
&\qquad \leq C(\varepsilon)\left(p_K^2\|\nabla(u-u_S)\|_{L^2(K)}^2 + p_K^{1+2\varepsilon}\frac{h_K^2}{p_K^2}\left(\left\|k^2(u-u_S)\right\|_{L^2(K)}^2 + \delta_K^2\right)\right)
\end{aligned}$$

and

$$\begin{aligned}
&\frac{h_K}{p_K}\left\|\widetilde{\operatorname{Res}}(u_S)\right\|_{L^2(E)}^2 \\
&\qquad \leq C(\varepsilon)p_K^{2\varepsilon}\left(p_K^2\|\nabla(u-u_S)\|_{L^2(\omega_K)}^2 + p_K^{1+2\varepsilon}\frac{h_K^2}{p_K^2}\left(\left\|k^2(u-u_S)\right\|_{L^2(\omega_K)}^2 + \delta_{\omega_K}^2\right)\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\alpha_K^2\|\widetilde{\operatorname{res}}(u_S)\|_{L^2(K)}^2 + \alpha_E^2\left\|\widetilde{\operatorname{Res}}(u_S)\right\|_{L^2(E)}^2 \\
&\qquad \leq \left(\alpha_K\frac{p_K}{h_K}\right)^2\frac{h_K^2}{p_K^2}\|\widetilde{\operatorname{res}}(u_S)\|_{L^2(K)}^2 + \left(\alpha_E^2\frac{p_K}{h_K}\right)\frac{h_K}{p_K}\left\|\widetilde{\operatorname{Res}}(u_S)\right\|_{L^2(E)}^2 \\
&\qquad \leq C(\varepsilon)p_K^2\left(\alpha_K^2\frac{p_K^2}{h_K^2} + \alpha_E^2\frac{p_K^{1+2\varepsilon}}{h_K}\right) \\
&\qquad\qquad \left(\|\nabla(u-u_S)\|_{L^2(\omega_K)}^2 + 4p_K^{2\varepsilon}\frac{k_{K,+}^2h_K^2}{p_K^3}\|k(u-u_S)\|_{L^2(\omega_K)}^2 + p_K^{2\varepsilon}\frac{h_K^2}{p_K^3}\delta_{\omega_K}^2\right).
\end{aligned} \tag{3.67}$$

For the special choice $\varepsilon=1/2$ and with condition (3.64) we finally get

$$\widetilde{\eta}_K^2(u_S,\alpha) \leq C p_K^3\left(\alpha_K^2\frac{p_K^2}{h_K^2} + \alpha_E^2\frac{p_K}{h_K}\right)\left(\|u-u_S\|_{\mathcal{H};\omega_K}^2 + k_{K,+}^{-2}\delta_{\omega_K}^2\right).$$

■

[6] Recall that in general α_K depends on h_K (cf. Corollary 3.18 for $d=2$).

Remark 3.38

(a) It is possible to choose any $\varepsilon > 0$ in (3.67) (with $C(\varepsilon) \sim 1/\varepsilon$). The factor $p_K^{3/2}$ in the estimates (3.65), (3.66) then can be replaced by $p^{1+\varepsilon}$, while condition (3.64) *has the weaker form $k_{K,+}h_K/p_K \leq p_K^{1/2-\varepsilon}$ (for $\varepsilon \leq 1/2$). However, in view of $p_K \sim \log(k)$ we think that this is of minor importance.*

(b) Theorem 3.37 could be completed by the data saturation condition, say in case of (3.66)*, $C\delta_{\omega_K} p_K^{3/2} k_{K,+} \leq 1/2$, which would then allow to bound $\widetilde{\eta}_K(u_S, \alpha)$ directly by the error.*

References

[1] A. Abdulle and A. Nonnenmacher. A posteriori error analysis of the heterogeneous multiscale method for homogenization problems. *C. R. Math. Acad. Sci. Paris*, 347(17-18):1081–1086, 2009.

[2] M. Ainsworth and A. Arnold. A reliable a posteriori error estimator for adaptive hierarchic modeling. In P. Ladevéze, editor, *Advances in Adaptive Computational Methods in Mechanics*, pages 101–114, NY, 1998. Elsevier.

[3] M. Ainsworth and J. T. Oden. *A Posteriori Error Estimation in Finite Element Analysis*. Wiley, 2000.

[4] I. Babuška, F. Ihlenburg, T. Strouboulis, and S. K. Gangaraj. A posteriori error estimation for finite element solutions of Helmholtz' equation I. The quality of local indicators and estimators. *Internat. J. Numer. Methods Engrg.*, 40(18):3443–3462, 1997.

[5] I. Babuška, F. Ihlenburg, T. Strouboulis, and S. K. Gangaraj. A posteriori error estimation for finite element solutions of Helmholtz' equation II. Estimation of the pollution error. *Internat. J. Numer. Methods Engrg.*, 40(21):3883–3900, 1997.

[6] I. Babuška, I. Lee, and C. Schwab. On the a posteriori estimation of the modeling error for the heat conduction in a plate and its use for adaptive hierarchical modeling. In *Proceedings of the Third ARO Workshop on Adaptive Methods for Partial Differential Equations (Troy, NY, 1992)*, volume 14, pages 5–21, 1994.

[7] I. Babuška and W. C. Rheinboldt. A-posteriori error estimates for the finite element method. *Internat. J. Numer. Meth. Engrg.*, 12:1597–1615, 1978.

[8] I. Babuška and W. C. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.*, 15:736–754, 1978.

[9] I. Babuška and R. Rodriguez. The problem of the selection of an a posteriori error indicator based on smoothing techniques. *Internat. J. Numer. Meth. Engrg.*, 36:539–567, 1993.

[10] I. Babuška and C. Schwab. A posteriori error estimation for hierarchic models of elliptic boundary value problems on thin domains. *SIAM, J. Numer. Anal.*, 33:221–246, 1996.

[11] N. Bakhvalov and G. Panasenko. *Homogenisation: averaging processes in periodic media*. Kluwer Academic Publishers Group, Dordrecht, 1989. Mathematical problems in the mechanics of composite materials, Translated from the Russian by D. Leĭtes.

[12] L. Banjai and S. Sauter. A Refined Galerkin Error and Stability Analysis for Highly Indefinite Variational Problems. *SIAM J. Numer. Anal.*, 45(1):37–53, 2007.

[13] A. Bensoussan, J.-L. Lions, and G. Papanicolaou. *Asymptotic Analysis for Periodic Structures*. North-Holland, Amsterdam, 1978.

[14] S. Brenner and L. Scott. *The Mathematical Theory of Finite Element Methods*. Springer-Verlag, New York, 1994.

[15] Z. Cai and S. Zhang. Recovery-based error estimator for interface problems: conforming linear elements. *SIAM J. Numer. Anal.*, 47(3):2132–2156, 2009.

[16] C. Carstensen and S. Sauter. A Posteriori Error Analysis for Elliptic PDEs on Domains with Complicated Structures. *Numer. Math.*, 96:691–712, 2004.

[17] S. Chandler-Wilde and P. Monk. Wave-Number-Eplicit Bounds in Time-Harmonic Scattering. *SIAM J. Math. Anal.*, 39:1428–1455, 2008.

[18] M. Chipot. *Elliptic Equations: An Introductory Course*. Birkhäuser Verlag, Basel, 2009.

[19] P. Ciarlet. *The finite element method for elliptic problems*. North-Holland, 1987.

[20] D. Cioranescu and P. Donato. *An introduction to homogenization*. The Clarendon Press Oxford University Press, New York, 1999.

[21] W. Doerfler and S. Sauter. A Posteriori Error Estimation for Highly Indefinite Helmholtz Problems. Technical Report 13-2011, Institut für Mathematik, Univ. Zürich, 2011.

[22] W. Dörfler and M. Rumpf. An adaptive strategy for elliptic problems including a posteriori controlled boundary approximation. *Math. Comp.*, 67(224):1361–1382, 1998.

[23] S. Esterhazy and J. Melenk. On stability of discretizations of the Helmholtz equation. In I. Graham, T. Hou, O. Lakkis, and R. Scheichl, editors, *Numerical Analysis of Multiscale Problems*, volume 83 of *Lect. Notes Comput. Sci. Eng.*, pages 285–324. Springer, Berlin, 2012.

[24] M. Friedman and R. Shaw. Diffraction of Pulses by Cylindrical Obstacles of Arbitrary Cross Section. *J. Appl. Mech.*, 29:40–46, 1962.

[25] P. Henning and M. Ohlberger. The heterogeneous multiscale finite element method for elliptic homogenization problems in perforated domains. *Numer. Math.*, 113(4):601–629, 2009.

[26] F. Ihlenburg. *Finite Element Analysis of Acousting Scattering*. Springer, New York, 1998.

[27] S. Irimie and P. Bouillard. A residual a posteriori error estimator for the finite element solution of the Helmholtz equation. *Comput. Methods Appl. Mech. Engrg.*, 190(31):4027–4042, 2001.

[28] V. Jikov, S. Kozlov, and O. Oleinik. *Homogenization of Differential Operators and Integral Functionals*. Springer, Berlin, 1994.

[29] R. Leis. *Initial Boundary Value Problems in Mathematical Physics*. Teubner, Wiley Sons, Stuttgart, Chichester, 1986.

[30] J. 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.

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

[32] J. M. Melenk and S. A. Sauter. Wave-Number Explicit Convergence Analysis for Galerkin Discretizations of the Helmholtz Equation. *SIAM J. Numer. Anal.*, 49(3):1210–1243, 2011.

[33] J. M. Melenk and B. I. Wohlmuth. On residual-based a posteriori error estimation in *hp*-FEM. *Adv. Comput. Math.*, 15(1-4):311–331 (2002), 2001.

[34] M. Melenk. HP-interpolation of non-smooth functions (extended version). Technical Report NI03050, Isaac Newton Institute for Mathematical Sciences, 2003.

[35] P. Neittaanmäki and S. Repin. A posteriori error estimates for boundary–value problems related to the biharmonic operator. *East-West J.Numer. Math.*, 9:157–178, 2001.

[36] P. Neittaanmäki and S. Repin. *Reliable methods for computer simulation*. Elsevier Science B.V., Amsterdam, 2004. Error control and a posteriori estimates.

[37] J. T. Oden and J. R. Cho. Adaptive *hpq*-finite element methods of hierarchical models for plate- and shell-like structures. *Comput. Methods Appl. Mech. Engrg.*, 136(3-4):317–345, 1996.

[38] M. Ohlberger. A posteriori error estimates for the heterogeneous multiscale finite element method for elliptic homogenization problems. *Multiscale Model. Simul.*, 4(1):88–114, 2005.

[39] L. E. Payne and H. F. Weinberger. An optimal Poincaré inequality for convex domains. *Arch. Rational Mech. Anal.*, 5:286–292, 1960.

[40] S. Repin. A posteriori error estimation for nonlinear variational problems by duality theory. *Zapiski Nauchn. Semin. POMI*, 243:201–214, 1997.

[41] S. Repin. A posteriori error estimation for variational problems with uniformly convex functionals. *Math. Comp.*, 69:481–500, 2000.

[42] S. Repin. *A posteriori estimates for partial differential equations*, volume 4. Walter de Gruyter GmbH & Co. KG, Berlin, 2008.

[43] S. Repin and S. Sauter. Functional a posteriori estimates for the reaction-diffusion problem. *C.R. Acad. Sci. Paris. Ser. I*, 343:349–354, 2006.

[44] S. Repin and S. Sauter. Computable estimates of the modeling error related to Kirchhoff-Love plate model. *Anal. Appl. (Singap.)*, 8(4):409–428, 2010.

[45] S. Repin, S. Sauter, and A. Smolianski. Duality Based A Posteriori Error Estimator for the Dirichlet Problem. *Proc. Appl. Math. Mech.*, 2:513–514, 2003.

[46] S. Repin, S. Sauter, and A. Smolianski. A Posteriori Error Estimation for the Dirichlet Problem with Account of the Error in Boundary Conditions. *Computing*, 70:205–233, 2003.

[47] S. Repin, S. Sauter, and A. Smolianski. A Posteriori Error Estimation for the Poisson Equation with Mixed Dirichlet/Neumann Boundary Conditions. *JCAM*, 164-165:601–612, 2004.

[48] S. Repin, S. Sauter, and A. Smolianski. A Posteriori Estimation of Dimension Reduction Errors for Elliptic Problems on Thin Domains. *SIAM J. Numer. Anal.*, 42:1435–1451, 2004.

[49] S. Repin, S. Sauter, and A. Smolianski. Two-sided a posteriori error estimates for mixed formulations of elliptic problems. *SIAM J. Numer. Anal.*, 45(3):928–945, 2007.

[50] S. Repin and J. Valdman. Functional a posteriori error estimates for problems with nonlinear boundary conditions. *J. Numer. Math.*, 16(1):51–81, 2008.

[51] S. I. Repin. Estimates for errors in two-dimensional models of elasticity theory. *J. Math. Sci. (New York)*, 106(3):3027–3041, 2001. Function theory and phase transitions.

[52] S. I. Repin. Two-sided estimates of deviation from exact solutions of uniformly elliptic equations. In *Proceedings of the St. Petersburg Mathematical Society, Vol. IX*, volume 209 of *Amer. Math. Soc. Transl. Ser. 2*, pages 143–171, Providence, RI, 2003. Amer. Math. Soc.

[53] S. I. Repin, T. Samrowski, and S. Sauter. Combined A Posteriori Modelling-Discretization Error Eestimate for Elliptic Problems with Variable Coefficients. *ESAIM: Math.Model.Numer.Anal.*, 46:1389–1405, 2012.

[54] S. I. Repin, T. Samrowski, and S. Sauter. Two-sided Estimates of the Modeling Error for Elliptic Homogenization Problems. Technical Report to appear, Institut für Mathematik, Univ. Zürich, 2012.

[55] S. Sauter. A Refined Finite Element Convergence Theory for Highly Indefinite Helmholtz Problems. *Computing*, 78(2):101–115, 2006.

[56] C. Schwab. A–posteriori modeling error estimation for hierarchic plate models. *Numer. Math.*, 74:221–259, 1996.

[57] C. Schwab and A.-M. Matache. Generalized FEM for homogenization problems. In *Multiscale and multiresolution methods*, volume 20 of *Lect. Notes Comput. Sci. Eng.*, pages 197–237. Springer, Berlin, 2002.

[58] J. Valdman. Minimization of functional majorant in a posteriori error analysis based on $H(\mathrm{div})$ multigrid-preconditioned CG method. *Adv. Numer. Anal.*, 2009.

[59] R. Verfürth. *A review of a posteriori error estimation and adaptive mesh refinement.* Wiley and Teubner, 1996.

[60] R. Verfürth. Robust a posteriori error estimators for singularly perturbed reaction-diffusion equations. *Numer. Math.*, 78:479–493, 1998.