SHARP PREASYMPTOTIC ERROR BOUNDS FOR THE HELMHOLTZ h-FEM

J. GALKOWSKI[*] AND E. A. SPENCE[†]

Abstract. In the analysis of the h-version of the finite-element method (FEM), with fixed polynomial degree p, applied to the Helmholtz equation with wavenumber $k \gg 1$, the *asymptotic regime* is when $(hk)^p C_{\rm sol}$ is sufficiently small and the sequence of Galerkin solutions are quasioptimal; here $C_{\rm sol}$ is the norm of the Helmholtz solution operator, normalised so that $C_{\rm sol} \sim k$ for nontrapping problems. The *preasymptotic regime* is when $(hk)^{2p} C_{\rm sol}$ is sufficiently small, and (for physical data) one expects the relative error of the Galerkin solution to be controllably small.

In this paper, we prove the natural error bounds in the preasymptotic regime for the variable-coefficient Helmholtz equation in the exterior of a Dirichlet, or Neumann, or penetrable obstacle (or combinations of these) and with the radiation condition approximated either by a radial perfectly-matched layer (PML) or impedance boundary condition. Previously, such bounds for $p > 1$ were only available for Dirichlet obstacles with the radiation condition approximated by an impedance boundary condition. Our result is obtained via a novel generalisation of the "elliptic-projection" argument (the argument used to obtain the result for $p = 1$) which can be applied to a wide variety of abstract Helmholtz-type problems.

AMS subject classifications. 35J05, 65N15, 65N30, 78A45

Key words. Helmholtz, FEM, high order, pollution effect, preasymptotic, perfectly-matched layer, elliptic projection.

1. Introduction.

1.1. Informal statement of the main result.

We consider the h-version of the finite-element method (h-FEM), where accuracy is increased by decreasing the meshwidth h while keeping the polynomial degree p constant, applied to the Helmholtz equation.

Theorem 1.1 (Informal statement of the main result). *Let u be the solution to the variable-coefficient Helmholtz equation, with wavenumber $k > 0$, in the exterior of a Dirichlet, or Neumann, or penetrable obstacle (or combinations of these) and with the radiation condition approximated either by a perfectly-matched layer (PML) or an impedance boundary condition. Let $C_{\rm sol}$ be the norm of the solution operator, normalised so that $C_{\rm sol} \sim k$ for nontrapping problems.*

Under the natural regularity assumptions on the domain and coefficients, if

$$(hk)^{2p} C_{\rm sol} \text{ is sufficiently small} \tag{1.1}$$

then the Galerkin solution u_h exists, is unique, and satisfies

$$\|u - u_h\|_{H^1_k(\Omega)} \leq C\Big(1 + hk + (hk)^p C_{\rm sol}\Big) \min_{v_h \in \mathcal{H}_h} \|u - v_h\|_{H^1_k(\Omega)}, \tag{1.2}$$

$$\|u - u_h\|_{L^2(\Omega)} \leq C\Big(hk + (hk)^p C_{\rm sol}\Big) \min_{v_h \in \mathcal{H}_h} \|u - v_h\|_{H^1_k(\Omega)}. \tag{1.3}$$

Furthermore, if the data is k-oscillatory (in a sense made precise below), then

$$\frac{\|u - u_h\|_{H^1_k(\Omega)}}{\|u\|_{H^1_k(\Omega)}} \leq C\Big(1 + hk + (hk)^p C_{\rm sol}\Big)(hk)^p; \tag{1.4}$$

[*] Department of Mathematics, University College London, 25 Gordon Street, London, WC1H 0AY, UK, J.Galkowski@ucl.ac.uk

[†] Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK, E.A.Spence@bath.ac.uk

i.e., the relative H^1_k error can be made controllably small by making $(hk)^{2p}C_{\rm sol}$ sufficiently small.

The norm $\|\cdot\|_{H^1_k(\Omega)}$ in the bounds above is defined by

$$\|v\|^2_{H^1_k(\Omega)} := k^{-2}\,\|\nabla v\|^2_{L^2(\Omega)} + \|v\|^2_{L^2(\Omega)}\,. \tag{1.5}$$

The fact that, for oscillatory data, the relative H^1_k error for the Helmholtz h-FEM is controllably small if $(hk)^{2p}C_{\rm sol}$ is sufficiently small was famously identified for 1-d nontrapping problems by the work of Ihlenburg and Babuška [25, 26]. The bounds (1.2) and (1.3) have previously been obtained (i) for the Dirichlet obstacle problem with impedance boundary conditions approximating the radiation condition [12, 40] and (ii) for PML with constant-coefficients, no obstacle, and $p = 1$ [32].

The present paper proves the bounds (1.2), (1.3), and (1.4) assuming only that the sesquilinear form is continuous, satisfies a Gårding inequality, and satisfies certain standard elliptic-regularity assumptions, therefore covering a variety of scatterers and methods for truncating the exterior domain (to approximate the radiation condition). Regarding the latter: in this paper we consider truncating with a PML or an impedance boundary condition, but truncating with the exact Dirichlet-to-Neumann map is also, in principle, covered by the abstract framework; see Remark 5.4 below.

1.2. Statement of the main abstract result. Let $\mathcal{H}\subset\mathcal{H}_0\subset\mathcal{H}^*$ be Hilbert spaces with $\mathcal{H}_0$ identified with its dual and $\mathcal{H}\subset\mathcal{H}_0$ compact. Let $a:\mathcal{H}\times\mathcal{H}\to\mathbb{C}$ be a continuous sesquilinear form, i.e.,

$$|a(u,v)|\le C_{\rm cont}\,\|u\|_{\mathcal{H}}\,\|v\|_{\mathcal{H}} \quad\text{and}\quad a(\lambda u,\mu v)=\lambda\bar{\mu}a(u,v)\quad\text{for all } u,v\in\mathcal{H}, \tag{1.6}$$

satisfying the Gårding inequality

$$\Re a(v,v)\ge C_{\rm G1}\,\|v\|^2_{\mathcal{H}} - C_{\rm G2}\,\|v\|^2_{\mathcal{H}_0}\qquad\text{for all } v\in\mathcal{H} \tag{1.7}$$

for some $C_{\rm G1}, C_{\rm G2}>0$. We assume further that $C_{\rm cont}, c, C$ and all the other constants in this section are independent of k.

Assumption 1.2 ("Elliptic regularity" assumptions on a). *Let $\mathcal{Z}_0=\mathcal{H}_0$, $\mathcal{Z}_1=\mathcal{H}$, and $\mathcal{Z}_j\subset\mathcal{Z}_{j-1}$ for $j=2,\dots,\ell+1$ such that $\mathcal{Z}_j$ is dense in $\mathcal{Z}_{j-1}$, and assume that for all $u\in\mathcal{H}$ with*

$$\sup_{v\in\mathcal{H},\,\|v\|_{(\mathcal{Z}_{j-2})^*}=1}|a(u,v)|<\infty,$$

$u\in\mathcal{Z}_j$ and

$$\|u\|_{\mathcal{Z}_j}\le C\Big(\,\|u\|_{\mathcal{H}_0}+\sup_{v\in\mathcal{H},\,\|v\|_{(\mathcal{Z}_{j-2})^*}=1}|a(u,v)|\Big),\quad j=2,\dots,\ell+1. \tag{1.8}$$

Assume further that for any $w\in\mathcal{H}$ such that

$$\sup_{w\in\mathcal{H},\,\|v\|_{(\mathcal{Z}_{j-2})^*}=1}|(\Re a)(u,v)|<\infty,$$

$w\in\mathcal{Z}_j$ with

$$\|w\|_{\mathcal{Z}_j}\le C\Big(\,\|u\|_{\mathcal{H}_0}+\sup_{v\in\mathcal{H},\,\|v\|_{(\mathcal{Z}_{j-2})^*}=1}|(\Re a)(u,v)|\Big),\quad j=2,\dots,\ell+1, \tag{1.9}$$

where the sesquilinear form $\Re a$ is defined by

$$(\Re a)(u,v) := \tfrac{1}{2}\big(a(u,v) + \overline{a(v,u)}\big). \tag{1.10}$$

Remark 1.3. *Note that $\Re a$ in (1.7) and (1.10) could be replaced by $\Re(\mathrm{e}^{\mathrm{i}\omega} a)$, so long as one uses the same value of ω in both conditions. Remark 4.4 below describes a situation where this is useful.*

Given $g \in \mathcal{H}^*$, suppose that $u \in \mathcal{H}$ satisfies

$$a(u,v) = \langle g, v\rangle \qquad \text{for all } v \in \mathcal{H}. \tag{1.11}$$

Given a sequence of finite dimensional subspace $\{\mathcal{H}_h\}_{h>0}$ with $\mathcal{H}_h \subset \mathcal{H}$, the sequence of Galerkin approximations of u, $\{u_h\}_{h>0}$, are defined by

$$a(u_h, v_h) = \langle g, v_h\rangle \quad \text{for all} \quad v_h \in \mathcal{H}_h. \tag{1.12}$$

Example 1.4. *For the Helmholtz equation outside a Dirichlet obstacle with PML truncation and Ω the truncated exterior domain, $\mathcal{H}_0 = L^2(\Omega)$, $\mathcal{H} = H^1_0(\Omega)$, and $\mathcal{Z}_j = H^j(\Omega) \cap H^1_0(\Omega)$. Assumption 1.2 is then elliptic regularity for the Helmholtz PML operator and its real part, which both hold if the coefficients of the Helmholtz equation are in $C^{\ell-1,1}$, the PML scaling function is $C^{\ell,1}$, and $\partial\Omega$ is $C^{\ell,1}$ (see Lemma 4.7 below).*

Theorem 1.5 (Abstract generalisation of the elliptic-projection argument). *Let $a : \mathcal{H} \times \mathcal{H} \to \mathbb{C}$ satisfy (1.6), (1.7), and Assumption 1.2. Suppose that $\mathcal{R}^* : \mathcal{H}^* \to \mathcal{H}$ defined by*

$$a(w, \mathcal{R}^* v) = \langle w, v\rangle \qquad \textit{for all } w \in \mathcal{H},\ v \in \mathcal{H}^*, \tag{1.13}$$

is well defined and let

$$\eta(\mathcal{H}_h) := \sup_{g \in \mathcal{H}_0, g\neq 0} \frac{\|(I - \Pi)\mathcal{R}^* g\|_{\mathcal{H}}}{\|g\|_{\mathcal{H}_0}}, \tag{1.14}$$

where $\Pi : \mathcal{H} \to \mathcal{H}_h$ is the orthogonal projection. Then the solution, u, to (1.11) exists and is unique and there exist $C_1, C_2, C_3 > 0$ such that if h satisfies

$$\eta(\mathcal{H}_h)\|I - \Pi\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}} \leq C_1, \tag{1.15}$$

then the solution u_h to (1.12) exists, is unique, and satisfies

$$\|u - u_h\|_{\mathcal{H}} \leq C_2\big(1 + \eta(\mathcal{H}_h)\big) \min_{w_h \in \mathcal{H}_h} \|u - v_h\|_{\mathcal{H}}\,, \tag{1.16}$$

$$\|u - u_h\|_{\mathcal{H}_0} \leq C_3\, \eta(\mathcal{H}_h) \min_{w_h \in \mathcal{H}_h} \|u - v_h\|_{\mathcal{H}}\,. \tag{1.17}$$

If, in addition,

$$\|g\|_{\mathcal{Z}_{\ell-1}} \leq C\,\|g\|_{\mathcal{H}^*} \tag{1.18}$$

for some $C > 0$, then there exists $C_4 > 0$ such that if h satisfies (1.15) then

$$\frac{\|u - u_h\|_{\mathcal{H}}}{\|u\|_{\mathcal{H}}} \leq C_4\big(1 + \eta(\mathcal{H}_h)\big)\,\|I - \Pi\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}}\,; \tag{1.19}$$

i.e., the relative error in $\mathcal{H}$ can be made controllably small by making $\eta(\mathcal{H}_h)\,\|I - \Pi\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}}$ sufficiently small.

Theorem 1.5 includes the result that the sequence of Galerkin solutions are quasioptimal with constant independent of k if $\eta(\mathcal{H}_h)$ is sufficiently small – with this the so-called *asymptotic regime* (see the discussion in §1.3).

The bounds (1.16), (1.17), and (1.19) and the meshthreshold (1.15) in Theorem 1.5 all involve the quantity $\eta(\mathcal{H}_h)$, which measures how well solutions of the adjoint problem are approximated in the space $\mathcal{H}_h$. Bounds on $\eta(\mathcal{H}_h)$ are given in [37, 38, 36, 13, 6, 29, 19, 20, 3]; see the discussion in §1.3. The following bound on $\eta(\mathcal{H}_h)$ is proved using the ideas in [6] (although the end result is phrased in a different way there); we include it here both for completeness, and because it holds under the assumptions of Theorem 1.5 (in fact, it only requires the regularity assumption (1.8) and not (1.9)).

Theorem 1.6 (Bound on $\eta(\mathcal{H}_h)$). *Under the assumptions of Theorem 1.5, there exists $C > 0$ such that*

(1.20)

$$\eta(\mathcal{H}_h) \leq C\bigg(\sum_{j=0}^{\lfloor \ell/2 \rfloor -1} \|(I-\Pi)\|_{\mathcal{Z}_{2(j+1)}\to\mathcal{H}} + \|(I-\Pi)\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}} \left(1 + \|\mathcal{R}^*\|_{\mathcal{H}_0\to\mathcal{H}}\right)\bigg).$$

Example 1.7. *In §4 and §5 below we show how Helmholtz problems with the radiation condition approximated by either a PML or an impedance boundary condition, respectively, fit into the abstract framework of Theorems 1.5 and 1.6. In both these cases, the norm of the adjoint solution operator, i.e., $\|\mathcal{R}^*\|_{\mathcal{H}_0\to\mathcal{H}}$, is the same as the norm of the solution operator of the original (non-adjoint) problem, which we denote by $C_{\rm sol}$. Furthermore, with $\{\mathcal{H}_h\}_{h>0}$ corresponding to the standard finite-element spaces of piecewise degree-p polynomials on shape-regular simplicial triangulations, indexed by the meshwidth h,*

$$\|(I-\Pi)\|_{\mathcal{Z}_{m+1}\to\mathcal{H}} \leq C(hk)^m \qquad \text{for } 0 \leq m \leq p.$$

The meshthreshold (1.15) then becomes that $(hk)^{2\ell}C_{\rm sol}$ is sufficiently small. Recall that ℓ is a parameter in the elliptic-regularity assumptions (Assumption 1.2). If the polynomial degree p is taken to be ℓ then (1.15) becomes (1.1). The bounds (1.16) and (1.17) then become (1.2) and (1.3), respectively.

1.3. Discussion of the context, novelty, and ideas behind Theorem 1.5.

The work of Ihlenburg and Babuška in 1-d. The celebrated work of [25, 26] studied the h-FEM applied to the constant-coefficient Helmholtz equation in 1-d (a nontrapping problem), and split the behaviour of the finite-element solutions as a function of h into the so-called asymptotic and preasymptotic regimes.

The *asymptotic regime* is when h is small enough, as a function of k, for the sequence of Galerkin solutions to be quasi-optimal uniformly in k, i.e.,

$$\|u-u_h\|_{H^1_k(\Omega)} \leq C \min_{v_h\in\mathcal{H}_h} \|u-v_h\|_{H^1_k(\Omega)}$$

with $C > 0$ independent of k. [26, Theorem 3.5] showed that a sufficient condition to be in the asymptotic regime is "hk^2/p sufficiently small", with later work (discussed below) then showing that a sufficient condition for nontrapping problems (when $C_{\rm sol} \sim k$) is "$(hk)^p k$ sufficiently small", with this condition then indicated to be necessary by numerical experiments. Therefore, the pollution effect for the h-FEM, i.e., the fact that one needs $h \ll k^{-1}$ to maintain accuracy, becomes less pronounced as p increases.

The *preasymptotic* regime is when the relative H^1_k error is controllably small, uniformly as $k \to \infty$, provided that the data is k-oscillatory, in the sense that it satisfies the bound (1.18) [1]. [26, Corollary 3.2] used the explicit form of the Helmholtz Green's function in 1-d to prove that if $(hk)^{2p}k$ sufficiently small then the finite-element solution is in the preasymptotic regime, with the numerical experiments in [26, Table 2] (for $p = 1, \ldots, 6$) indicating that this condition is also necessary. [26] also studied the phase difference between the exact and finite-element solutions (following [23, 43]), with [26, Theorem 3.2] showing that the difference between the true wavenumber and the numerical wavenumber is bounded by $C(hk)^{2p}k$. Thus the condition "$(hk)^{2p}k$ sufficiently small" also controls this phase difference; see also [1, Equation 3.5].

Error bounds in the asymptotic regime using the Schatz argument.. We now outline the argument that gives the condition "$(hk)^p C_{\rm sol}$ sufficiently small" for quasioptimality, with this argument also used in the proof of Theorem 1.5. We work in the setting of Examples 1.4 and 1.7; i.e., the PML approximation to the Helmholtz exterior Dirichlet problem, so that $\mathcal{H}_0 = L^2(\Omega)$ and $\mathcal{H} = H^1_0(\Omega)$. The Gårding inequality (1.7) is then

$$\Re a(w,w) \geq C_{\rm G1} \|w\|^2_{H^1_k(\Omega)} - C_{\rm G2} \|w\|^2_{L^2(\Omega)} \qquad \text{for all } w \in H^1_0(\Omega)$$

for $C_{\rm G1}, C_{\rm G2} > 0$ (see Corollary 4.6 below). Combining the Gårding inequality with the Galerkin orthogonality

$$a(u-u_h, v_h) = 0 \quad \text{ for all } v_h \in \mathcal{H}_h, \tag{1.21}$$

we find that, for all $v_h \in \mathcal{H}_h$,

$$\begin{aligned} \|u-u_h\|^2_{H^1_k(\Omega)} &\leq C_{\rm G1}^{-1}\big|a(u-u_h, u-v_h)\big| + C_{\rm G1}^{-1}C_{\rm G2}\|u-u_h\|^2_{L^2(\Omega)} \\ &\leq C_{\rm G1}^{-1}C_{\rm cont}\|u-u_h\|_{H^1_k(\Omega)}\|u-v_h\|_{H^1_k(\Omega)} + C_{\rm G1}^{-1}C_{\rm G2}\|u-u_h\|^2_{L^2(\Omega)}, \end{aligned} \tag{1.22}$$

where $C_{\rm cont}$ is the continuity constant of the sesquilinear form a. Therefore, (1.22) implies that a sufficient condition for quasioptimality is that the L^2 error is sufficiently small relative to the H^1_k error.

By the definition of $\mathcal{R}^*$ (1.13) (recalling that $\mathcal{H} = H^1_0(\Omega)$ here) and Galerkin orthogonality (1.21), for any $v_h \in \mathcal{H}_h$,

$$\begin{aligned} \|u-u_h\|^2_{L^2(\Omega)} = a\big(u-u_h, \mathcal{R}^*(u-u_h)\big) &= a\big(u-u_h, \mathcal{R}^*(u-u_h) - v_h\big) \\ &\leq C_{\rm cont}\|u-u_h\|_{H^1_k(\Omega)}\big\|\mathcal{R}^*(u-u_h)-v_h\big\|_{H^1_k(\Omega)}, \end{aligned} \tag{1.23}$$

and thus, by the definition of $\eta(\mathcal{H}_h)$ (1.14) (recalling that $\mathcal{H}_0 = L^2(\Omega)$),

$$\|u-u_h\|_{L^2(\Omega)} \leq C_{\rm cont}\eta(\mathcal{H}_h)\|u-u_h\|_{H^1_k(\Omega)}. \tag{1.24}$$

Combining this last inequality with (1.22), we see that a sufficient condition for quasioptimality is that $\eta(\mathcal{H}_h)$ is sufficiently small. Schatz [42] was the first to use the Aubin-Nitsche-type bound (1.24) with the Gårding inequality, and thus the argument above is often called the Schatz argument. The "adjoint approximability" concept, and associated definition of $\eta(\mathcal{H}_h)$, was introduced by Sauter in [41].

[1] The relative error can only be small for a certain subclass of data, since, given a finite-dimensional subspace $\mathcal{H}_h$, one can choose data such that the solution $v \in \mathcal{H}$ is orthogonal to $\mathcal{H}_h$. Then $\|u-u_h\|^2_{\mathcal{H}} = \|u\|^2_{\mathcal{H}} + \|u_h\|^2_{\mathcal{H}} \geq \|u\|^2_{\mathcal{H}}$.

The bound

$$\eta\big(\mathcal{H}_h\big) \leq C\big(hk + (hk)^p C_{\rm sol}\big) \tag{1.25}$$

under sufficient regularity of the coefficients and obstacle has now been proved for a wide variety of Helmholtz problems, with this bound sharp by the recent results of [17]. The bound (1.25) therefore gives the sufficient condition "$(hk)^p C_{\rm sol}$ sufficiently small" for quasioptimality, with this condition observed sharp for nontrapping problems in, e.g., [6, Figures 3, 5, and 8] for $p = 1, 2, 3, 4$.

For $p = 1$, the bound (1.25) can be proved using only H^2 regularity of the Helmholtz solution, with the condition "hk^2 sufficient small" for quasiopimality obtained for 1-d problems in [2, Theorem 3.1], [11, Lemma 2.6], [27, Theorem 3], and [33, Theorem 3.2], 2-d problems in [35, Proposition 8.2.7], and variable-coefficient problems in 2- and 3-d in [22, 21].

For $p > 1$ the bound (1.25) is proved by a judicious splitting of the solution in [37, 38, 13, 36] for constant-coefficient problems and [6, 29, 19, 20, 3] for variable-coefficient problems. All these papers apart from [6] make the constant C in (1.25) explicit in p under suitably analyticity/smoothness assumptions on the obstacle and coefficients, and thus give results about the *hp*-FEM (showing that quasioptimality holds if hk/p is sufficiently small and $p/\log k$ is sufficiently large). In addition, all these papers apart from [6] split the solution into "high-" and "low-" frequency components. In constrast, [6] instead expands the solution in a series whose terms increase with regularity, and with only the remainder satisfying a bound involving $C_{\rm sol}$; see Lemma 2.2 below.

Bounds in the preasymptotic regime. Numerical experiments indicate that, at least for nontrapping problems, the condition "$(hk)^{2p} C_{\rm sol}$ sufficiently small" for the relative H^1_k error to be controllably small is necessary and sufficient for 2- and 3-d Helmholtz problems; see, e.g., [12, Figure 3]. Nevertheless, despite the fact that sharp *asymptotic* error bounds have now been obtained for a variety of Helmholtz problems in 2- and 3-d and for arbitrary $p \in \mathbb{Z}^+$, until now the sharp *preasymptotic* error bounds were obtained only in the following cases.

1. $p = 1$, the constant-coefficient Helmholtz equation with an impedance boundary condition [44, Theorem 6.1] or PML (and no obstacle) [32, Theorem 4.4], the variable-coefficient Helmholtz equation with truncation via the exact Dirichlet-to-Neumann map [28, Theorem 4.1].
2. $p \in \mathbb{Z}^+$, the constant-coefficient Helmholtz equation with no obstacle and an impedance boundary condition approximating the radiation condition [12, Theorem 5.1],
3. $p \in \mathbb{Z}^+$, the variable-coefficient Helmholtz equation in the exterior of a Dirichlet obstacle with an impedance boundary condition approximating the radiation condition [40, Theorem 2.39].

The bounds in Point 1 for $p = 1$ come from the so-called elliptic projection argument, which proves error bounds under the condition "$(hk)^{p+1} C_{\rm sol}$ is sufficiently small"; i.e., the sharp condition when $p = 1$, but not when $p > 1$. The initial ideas behind this argument were introduced in the Helmholtz context in [15, 16] for interior-penalty discontinuous Galerkin methods, and then further developed for the standard FEM and continuous interior-penalty methods in [44, 45].

The bounds in Point 2 used an error-splitting argument (with this idea called "stability-error iterative improvement", and used earlier in [16, 44]) together with the idea of using discrete Sobolev norms in the duality argument. The bounds in Point 3 for variable-coefficients were obtained by repeating the constant-coefficient arguments in Point 2, but now keeping track of how the constants depend on the coefficients.

The elliptic-projection argument. Theorem 1.5 is proved by generalising the elliptic-projection argument, allowing it to prove error bounds under the sharp condition "$(hk)^{2p}C_{\rm sol}$ sufficiently small" for $p>1$. We therefore recap the main ideas of the elliptic-projection argument here, and then we explain below how we generalise this argument. Here, and in the rest of the paper, C is used for a constant, independent of h and k, but dependent on p, whose value may change line by line.

The bounds (1.2) and (1.3) come from the bounds

$$\|u-u_h\|_{H^1_k(\Omega)} \leq C\big(1+\eta(\mathcal{H}_h)\big) \min_{v_h\in\mathcal{H}_h} \|u-v_h\|_{H^1_k(\Omega)} \tag{1.26}$$

and

$$\|u-u_h\|_{L^2(\Omega)} \leq C\eta(\mathcal{H}_h) \min_{v_h\in\mathcal{H}_h} \|u-v_h\|_{H^1_k(\Omega)} \tag{1.27}$$

and the bound (1.25) on $\eta(\mathcal{H}_h)$. Observe that, by the consequence (1.22) of the Gårding inequality, the bound (1.26) follows from the bound (1.27).

To prove (1.27), the elliptic-projection argument writes (1.23) as

$$\begin{aligned}\|u-u_h\|^2_{L^2(\Omega)} &= a\big(u-u_h, \mathcal{R}^*(u-u_h)-v_h\big)\\ &= \widetilde{a}\big(u-u_h, \mathcal{R}^*(u-u_h)-v_h\big) - \big((1+c^{-2})(u-u_h), \mathcal{R}^*(u-u_h)-v_h\big)_{L^2(\Omega)},\end{aligned} \tag{1.28}$$

where

$$\widetilde{a}(u,v) := \int_\Omega k^{-2}A\nabla u\cdot\overline{\nabla v} + u\,\overline{v}.$$

Let $\widetilde{\Pi}: H^1_0(\Omega)\to\mathcal{H}_h$ be the solution of the variational problem

$$\widetilde{a}(w_h, \widetilde{\Pi}v) = \widetilde{a}(w_h, v) \quad \text{ for all } w_h\in\mathcal{H}_h.$$

Since $\widetilde{a}$ is coercive on $H^1_0(\Omega)$ and the continuity and coercivity constants of $\widetilde{a}$ in $\|\cdot\|_{H^1_k(\Omega)}$ are independent of k, $\widetilde{\Pi}$ is well-defined by the Lax–Milgram theorem and

$$\big\|(I-\widetilde{\Pi})v\big\|_{H^1_k(\Omega)} \leq C \min_{w_h\in\mathcal{H}_h}\|v-w_h\|_{H^1_k(\Omega)} \tag{1.29}$$

with $C>0$ independent of k by Céa's lemma. The definition of $\widetilde{\Pi}$ implies the Galerkin orthogonality

$$\widetilde{a}\big(w_h, (I-\widetilde{\Pi})v\big) = 0 \quad \text{ for all } w_h\in\mathcal{H}_h. \tag{1.30}$$

We now choose $v_h = \widetilde{\Pi}\mathcal{R}^*(u-u_h)$ in (1.28) so that, by (1.30), for all $w_h\in\mathcal{H}_h$,

$$\begin{aligned}\|u-u_h\|^2_{L^2(\Omega)} &= \widetilde{a}\big(v-w_h, (I-\widetilde{\Pi})\mathcal{R}^*(u-u_h)\big)\\ &\qquad - \big((1+c^{-2})(u-u_h), (I-\widetilde{\Pi})\mathcal{R}^*(u-u_h)\big)_{L^2(\Omega)}.\end{aligned} \tag{1.31}$$

For the first term on the right-hand side of (1.31) we use the continuity of $\widetilde{a}$, (1.29), and the definition of $\eta(\mathcal{H}_h)$ (1.14) to bound this term by

$$C\,\|v-w_h\|_{H^1_k(\Omega)}\,\eta(\mathcal{H}_h)\,\|u-u_h\|_{L^2(\Omega)}\,.$$

The second term on the right-hand side of (1.31) is bounded by

$$C \|u-u_h\|_{L^2(\Omega)} \big\|(I-\widetilde{\Pi})\mathcal{R}^*(u-u_h)\big\|_{L^2(\Omega)}.$$

Using the Schatz argument for $\widetilde{a}$, one can show that

$$\big\|(I-\widetilde{\Pi})\mathcal{R}^*(u-u_h)\big\|_{L^2(\Omega)} \leq Chk\big\|(I-\widetilde{\Pi})\mathcal{R}^*(u-u_h)\big\|_{H^1_k(\Omega)} \tag{1.32}$$

and then (1.29) and the definition of $\eta(\mathcal{H}_h)$ (1.14) imply that the second term on the right-hand side of (1.31) is bounded by

$$Chk\,\eta(\mathcal{H}_h)\,\|u-u_h\|^2_{L^2(\Omega)}\,, \tag{1.33}$$

which can be absorbed into the left-hand side if $hk\,\eta(\mathcal{H}_h)$ is sufficiently small, giving the result (1.27).

The ideas behind the proof of Theorem 1.5. We generalise the elliptic-projection argument based on the observation that if $\widetilde{a}(u,v) = a(u,v) + (Su,v)_{L^2(\Omega)}$ with S a self-adjoint smoothing operator, then the second term on the right-hand side of (1.31) is replaced by

$$\big(u-u_h, S^*(I-\widetilde{\Pi})\mathcal{R}^*(u-u_h)\big)_{L^2(\Omega)} \tag{1.34}$$

(see (2.14) below). Using the Schatz argument for $\widetilde{a}$ and the smoothing property of S, the modulus of this term is bounded by

$$\big\|S^*(I-\widetilde{\Pi})\mathcal{R}^*(u-u_h)\big\|_{L^2(\Omega)} \leq C(hk)^p\big\|(I-\widetilde{\Pi})\mathcal{R}^*(u-u_h)\big\|_{H^1_k(\Omega)} \tag{1.35}$$

(see (2.16) below). Provided that $\widetilde{\Pi}$ still satisfies (1.29), the term (1.34) is therefore bounded by

$$C(hk)^p\eta(\mathcal{H}_h)\,\|u-u_h\|^2_{L^2(\Omega)}\,. \tag{1.36}$$

Comparing (1.32) and (1.35), and also (1.33) and (1.36), we see that this new argument replaces the condition "$hk\eta(\mathcal{H}_h)$ sufficiently small" in the standard elliptic-projection argument by the condition "$(hk)^p\eta(\mathcal{H}_h)$ sufficiently small", which is the condition $(hk)^{2p}C_{\rm sol}$ sufficiently small" after using the bound (1.25) on $\eta(\mathcal{H}_h)$.

The challenge now is to ensure that the smoothing operator S is such that the projection $\widetilde{\Pi}$ is well-defined and satisfies (1.29). This is achieved in Lemma 2.1 below, where a suitable S such that $\widetilde{a}(u,v) = a(u,v) + (Su,v)_{L^2(\Omega)}$ is coercive is constructed. S is defined by an expansion in terms of the eigenfunctions of the (self-adjoint) operator associated with the real part of the sesquilinear form a (defined by (1.10)).

2. Proofs of the main results (Theorems 1.5 and 1.6).

2.1. Construction of a regularizing operator that produces coercivity when added to a.

LEMMA 2.1. *Suppose that $a : \mathcal{H}\times\mathcal{H}\to\mathbb{C}$ satisfies (1.6), (1.7), and Assumption 1.2. Then there exists $S:\mathcal{H}_0\to\mathcal{H}_0$ self adjoint and $c, C>0$ such that, with*

$$\widetilde{a}(u,v) := a(u,v) + \langle Su,v\rangle_{\mathcal{H}_0}, \tag{2.1}$$

$$\Re\widetilde{a}(v,v) \geq c\,\|v\|^2_{\mathcal{H}} \quad \textit{for all } v\in\mathcal{H}, \tag{2.2}$$

$$\|S\|_{\mathcal{H}_0\to\mathcal{Z}_j} \leq C, \qquad j=0,\ldots,\ell+1 \tag{2.3}$$

and $\widetilde{\mathcal{R}}:\mathcal{H}^*\to\mathcal{H}$ *defined by*

$$\widetilde{a}(\widetilde{\mathcal{R}}f,u) = \langle f,u\rangle \quad \textit{for all } u\in\mathcal{H},\ f\in\mathcal{H}^*, \tag{2.4}$$

is well defined with

$$\|\widetilde{\mathcal{R}}\|_{\mathcal{Z}_{j-2}\to\mathcal{Z}_j} \leq C, \qquad 2\leq j\leq \ell+1. \tag{2.5}$$

The proof of Lemma 2.1 uses the spectral theorem for bounded self-adjoint operators, $B:\mathcal{H}\to\mathcal{H}^*$, which we recap here. With $\mathcal{H}_0$ and $\mathcal{H}$ as in §1.2, let b be a sesquilinear form on $\mathcal{H}$ satisfying $b(u,v)=\overline{b(v,u)}$, with associated operator B; i.e., $b(u,v)=\langle Bu,v\rangle$ for all $u,v\in\mathcal{H}$. If b satisfies the Gårding inequality (1.7) (with a replaced by b) then there exist an orthonormal basis (in $\mathcal{H}_0$) of eigenfunctions of B, $\{\phi_j\}_{j=1}^\infty$, with associated eigenvalues satisfying $\lambda_1\leq\lambda_2\leq\ldots$ with $\lambda_j\to\infty$ as $j\to\infty$. Furthermore, for all $u\in\mathcal{H}$,

$$Bu=\sum_{j=1}^\infty \lambda_j\langle\phi_j,u\rangle\phi_j \tag{2.6}$$

(where the sum converges in $\mathcal{H}^*$); see, e.g., [34, Theorem 2.37]. Given a bounded function f, we define $f(B):\mathcal{H}_0\to\mathcal{H}_0$ by

$$f(B)u := \sum_{j=1}^\infty f(\lambda_j)\langle\phi_j,u\rangle\phi_j, \quad \text{so that} \quad \|f(B)\|_{\mathcal{H}_0\to\mathcal{H}_0}\leq \sup_{\lambda\in[\lambda_1,\infty)}|f(\lambda)|. \tag{2.7}$$

Proof of Lemma 2.1. Let $\mathcal{P}:\mathcal{H}\to\mathcal{H}^*$ be the operator associated with the sesquilinear form $\Re a$ defined by (1.10), i.e., $(\Re a)(u,v)=\langle\mathcal{P}u,v\rangle$ for all $u,v\in\mathcal{H}$; observe that $\mathcal{P}$ is self-adjoint. Since $(\Re a)$ also satisfies the Gårding equality satisfied by a (1.7), the spectral theorem recapped above applies. Let $\{\lambda_j\}_{j=1}^\infty$ be the eigenvalues of $\mathcal{P}$, let $\psi\in C^\infty_{\rm comp}(\mathbb{R};[0,\infty))$ be such that

$$x+\psi(x)\geq 1 \quad \text{for } x\geq -\lambda_1, \tag{2.8}$$

and let $S:=\psi(\mathcal{P})$, in the sense of (2.7).

We now use (1.9) to prove that $S:\mathcal{H}_0\to\mathcal{Z}_j$ satisfying (2.3). Since ψ has compact support, the function $t\mapsto t^m\psi(t)$ is bounded for any $m\geq 0$. Thus (2.7) implies that, for any $m\geq0$,

$$\|\mathcal{P}^m\psi(\mathcal{P})\|_{\mathcal{H}_0\to\mathcal{H}_0}\leq C_m. \tag{2.9}$$

By (1.9),

$$\|\psi(\mathcal{P})\|_{\mathcal{H}_0\to\mathcal{Z}_j}\leq C_\ell\Big(\|\psi(\mathcal{P})\|_{\mathcal{H}_0\to\mathcal{H}_0}+\|\mathcal{P}\psi(\mathcal{P})\|_{\mathcal{H}_0\to\mathcal{Z}_{j-2}}\Big),\quad j=2,\ldots,\ell+1,$$

so that, by induction and (2.9),

$$\|S\|_{\mathcal{H}_0\to\mathcal{Z}_{\ell+1}}=\|\psi(\mathcal{P})\|_{\mathcal{H}_0\to\mathcal{Z}_{\ell+1}}\leq C_\ell\sum_{j=0}^{\lceil(\ell+1)/2\rceil}\big\|\mathcal{P}^j\psi(\mathcal{P})\big\|_{\mathcal{H}_0\to\mathcal{H}_0}\leq C_\ell.$$

We now show that $\widetilde{a}$ satisfies (2.2). By the definitions of $\mathcal{P}$ and S, (2.6), (2.7), and the inequality (2.8), for all $v \in \mathcal{H}$,

$$\Re \widetilde{a}(v,v) = \Re a(v,v) + \langle \psi(\mathcal{P})v, v\rangle = \langle (\mathcal{P}+\psi(\mathcal{P}))v, v\rangle \geq \|v\|^2_{\mathcal{H}_0}.$$

Since $\psi \geq 0$, S is positive, and thus $\Re \widetilde{a}(v,v) \geq \Re a(v,v)$ for all $v \in \mathcal{H}$, for any $\epsilon > 0$ and for all $v \in \mathcal{H}$,

$$\Re \widetilde{a}(v,v) \geq \epsilon \Re a(v,v) + (1-\epsilon)\Re \widetilde{a}(v,v) \geq \epsilon C_{\rm G1} \|v\|^2_{\mathcal{H}} - C_{\rm G2}\epsilon \|v\|^2_{\mathcal{H}_0} + (1-\epsilon)\|v\|^2_{\mathcal{H}_0},$$

so that, choosing $\epsilon = \min(\frac{1}{2C_{\rm G2}}, \frac{1}{2})$, we have

$$\Re \widetilde{a}(v,v) \geq \frac{C_{\rm G1}}{2} \min\left(\frac{1}{C_{\rm G2}}, 1\right) \|v\|^2_{\mathcal{H}} + \frac{1}{2}\|v\|^2_{\mathcal{H}_0};$$

i.e., $\widetilde{a}$ is coercive. The existence of $\widetilde{\mathcal{R}} : \mathcal{H}^* \to \mathcal{H}$ satisfying (2.4) and $\|\widetilde{\mathcal{R}}\|_{\mathcal{H}^*\to\mathcal{H}} \leq C$ then follows from the Lax–Milgram theorem. Finally, to see that

$$\|\widetilde{\mathcal{R}}\|_{\mathcal{Z}_{j-2}\to\mathcal{Z}_j} \leq C, \qquad 2 \leq j \leq \ell+1,$$

observe that, since S is self-adjoint and satisfies (2.3), for $v \in (\mathcal{Z}_{j-2})^*$,

$$\begin{aligned}|a(\widetilde{\mathcal{R}}g, v)| = |\widetilde{a}(\widetilde{\mathcal{R}}g, v) - \langle S\widetilde{\mathcal{R}}g, v\rangle| &\leq |\widetilde{a}(\widetilde{\mathcal{R}}g, v)| + |\langle S\widetilde{\mathcal{R}}g, v\rangle| \\ &\leq |\langle v, g\rangle| + \|v\|_{(\mathcal{Z}_{j-2})^*}\|S\|_{\mathcal{H}\to\mathcal{Z}_{j-2}}\|(\widetilde{\mathcal{R}})^*\|_{\mathcal{H}^*\to\mathcal{H}}\|g\|_{\mathcal{H}^*} \\ &\leq \|v\|_{(\mathcal{Z}_{j-2})^*}(\|g\|_{\mathcal{Z}_{j-2}} + C\|g\|_{\mathcal{H}^*}),\end{aligned}$$

and the claim follows from (1.8). □

2.2. Proof Theorem 1.5 using Lemma 2.1. We claim it is sufficient to prove the bounds (1.16) and (1.17) under the assumption of existence. Indeed, by uniqueness of the variational problem (1.11), either of the bounds (1.16) or (1.17) under the assumption of existence implies uniqueness of u_h, and uniqueness implies existence for the finite-dimensional Galerkin linear system.

We next show that the bound (1.16) follows from (1.17). Now, by the Gårding inequality (1.7), Galerkin orthogonality (1.21), and (1.17), for any $v_h \in \mathcal{H}_h$,

$$\begin{aligned}\|u-u_h\|^2_{\mathcal{H}} &\leq C\Big[\big|a(u-u_h, u-v_h)\big| + \|u-u_h\|^2_{\mathcal{H}_0}\Big] \\ &\leq C\Big[\,\|u-u_h\|_{\mathcal{H}}\|u-v_h\|_{\mathcal{H}} + \Big(\eta(\mathcal{H}_h)\min_{w_h\in\mathcal{H}_h}\|u-w_h\|_{\mathcal{H}}\Big)^2\Big].\end{aligned} \tag{2.10}$$

The bound (1.16) on the error in $\mathcal{H}$ then follows by using the inequality $2ab \leq \epsilon a^2 + b^2/\epsilon$ for all $a, b, \epsilon > 0$ in the first term on the right-hand side of (2.10), and then using the inequality $a^2 + b^2 \leq (a+b)^2$ for $a, b > 0$.

We now prove (1.17) (using the ideas outlined in §1.3). By the definition of $\mathcal{R}^*$, Galerkin orthogonality (1.21), and the definition of $\widetilde{a}$ (2.1)

$$\begin{aligned}\|u-u_h\|^2_{\mathcal{H}_0} &= a\big(u-u_h, \mathcal{R}^*(u-u_h)\big) = a\big(u-u_h, \mathcal{R}^*(u-u_h) - v_h\big) \\ &= \widetilde{a}\big(u-u_h, \mathcal{R}^*(u-u_h)-v_h\big) - \big\langle S(u-u_h), \mathcal{R}^*(u-u_h)-v_h\big\rangle_{\mathcal{H}_0}.\end{aligned} \tag{2.11}$$

Let $\widetilde{\Pi} : \mathcal{H} \to \mathcal{H}_h$ be the solution of the variational problem

$$\widetilde{a}(w_h, \widetilde{\Pi}v) = \widetilde{a}(w_h, v) \quad \text{ for all } w_h \in \mathcal{H}_h.$$

Since $\widetilde{a}$ is continuous and coercive, with constants independent of k (see (2.2), (1.6), and (2.3)), by the Lax–Milgram lemma and Céa's lemma given $k_0 > 0$ there exists $C > 0$ such that for all $k \geq k_0$ and $v \in \mathcal{H}$, $\widetilde{\Pi}$ is well-defined with

$$\big\|(I - \widetilde{\Pi})v\big\|_{\mathcal{H}} \leq C \min_{w_h \in \mathcal{H}_h} \|v - w_h\|_{\mathcal{H}}\,. \tag{2.12}$$

The definition of $\widetilde{\Pi}$ implies the Galerkin orthogonality

$$\widetilde{a}\big(w_h, (I - \widetilde{\Pi})u\big) = 0 \quad \text{ for all } w_h \in \mathcal{H}_h. \tag{2.13}$$

We now choose $v_h = \widetilde{\Pi}\mathcal{R}^*(u - u_h)$ in (2.11) so that, by (2.13), for all $w_h \in \mathcal{H}_h$,

$$\begin{aligned} &\|u - u_h\|^2_{\mathcal{H}_0} \\ &= \widetilde{a}\big(u - w_h, (I - \widetilde{\Pi})\mathcal{R}^*(u - u_h)\big) - \big\langle u - u_h, S^*(I - \widetilde{\Pi})\mathcal{R}^*(u - u_h)\big\rangle_{\mathcal{H}_0} \\ &\leq C\, \|u - w_h\|_{\mathcal{H}} \big\|(I - \widetilde{\Pi})\mathcal{R}^*(u - u_h)\big\|_{\mathcal{H}} + \|u - u_h\|_{\mathcal{H}_0} \big\|S^*(I - \widetilde{\Pi})\mathcal{R}^*(u - u_h)\big\|_{\mathcal{H}_0}. \end{aligned} \tag{2.14}$$

By (2.12) and the definition of $\eta(\mathcal{H}_h)$ (1.14),

$$\big\|(I - \widetilde{\Pi})\mathcal{R}^*(u - u_h)\big\|_{\mathcal{H}} \leq C \min_{w_h \in \mathcal{H}_h} \|\mathcal{R}^*(u - u_h) - w_h\|_{\mathcal{H}} \leq C\eta(\mathcal{H}_h)\, \|u - u_h\|_{\mathcal{H}_0}\,. \tag{2.15}$$

We now claim that the bound (1.17) follows if we can prove that, for all $v \in \mathcal{H}$,

$$\big\|S^*(I - \widetilde{\Pi})v\big\|_{\mathcal{H}_0} \leq C\|I - \Pi\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}} \big\|(I - \widetilde{\Pi})v\big\|_{\mathcal{H}}. \tag{2.16}$$

Indeed, we use (2.15) in the first term on the right-hand side of (2.14), and then (2.16) with $v = \mathcal{R}^*(u - u_h)$ in the second term on the right-hand side of (2.14) to obtain

$$\begin{aligned} \|u - u_h\|^2_{\mathcal{H}_0} \leq C\eta(\mathcal{H}_h)\, &\|u - w_h\|_{\mathcal{H}}\, \|u - u_h\|_{\mathcal{H}_0} \\ &+ C\|I - \Pi\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}} \big\|(I - \widetilde{\Pi})\mathcal{R}^*(u - u_h)\big\|_{\mathcal{H}}\, \|u - u_h\|_{\mathcal{H}_0}\,. \end{aligned}$$

By (2.15), the last term on the right-hand side is $\leq C\|I - \Pi\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}}\, \eta(\mathcal{H}_h)\|u - u_h\|^2_{\mathcal{H}_0}$ and (1.17) follows.

We now prove (2.16) by using the duality argument described in §1.3 (as part of the Schatz argument). By the definition of $\widetilde{\mathcal{R}}$ (2.4) and Galerkin orthogonality (2.13), for all $w_h \in \mathcal{H}_h$,

$$\big\|S^*(I - \widetilde{\Pi})v\big\|^2_{\mathcal{H}_0} = \big\langle SS^*(I - \widetilde{\Pi})v, (I - \widetilde{\Pi})v\big\rangle_{\mathcal{H}_0} = \widetilde{a}\big(\widetilde{\mathcal{R}}SS^*(I - \widetilde{\Pi})v - w_h, (I - \widetilde{\Pi})v\big).$$

Then, by the bounds (2.5) and (2.3),

$$\begin{aligned} \big\|S^*(I - \widetilde{\Pi})v\big\|^2_{\mathcal{H}_0} &\leq C \min_{w_h \in \mathcal{H}_h} \big\|\widetilde{\mathcal{R}}SS^*(I - \widetilde{\Pi})v - w_h\big\|_{\mathcal{H}} \big\|(I - \widetilde{\Pi})v\big\|_{\mathcal{H}} \\ &\leq \|I - \Pi\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}} \big\|\widetilde{\mathcal{R}}SS^*(I - \widetilde{\Pi})v\big\|_{\mathcal{Z}_{\ell+1}} \big\|(I - \widetilde{\Pi})v\big\|_{\mathcal{H}}, \\ &\leq C\|I - \Pi\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}} \big\|SS^*(I - \widetilde{\Pi})v\big\|_{\mathcal{Z}_{\ell-1}} \big\|(I - \widetilde{\Pi})v\big\|_{\mathcal{H}}, \\ &\leq C\|I - \Pi\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}} \big\|S^*(I - \widetilde{\Pi})v\big\|_{\mathcal{H}_0} \big\|(I - \widetilde{\Pi})v\big\|_{\mathcal{H}} \end{aligned}$$

which implies the bound (2.16), and hence (1.17).

Finally, we prove (1.19). By (1.11), (1.18), and the abstract elliptic-regularity assumption (1.8), $u \in \mathcal{Z}_{\ell+1}$ with

$$\|u\|_{\mathcal{Z}_{\ell+1}} \leq C\big(\, \|u\|_{\mathcal{H}_0} + \|g\|_{\mathcal{Z}_{\ell-1}} \big) \leq C\big(\, \|u\|_{\mathcal{H}_0} + \|g\|_{\mathcal{H}^*} \big).$$

The variational problem (1.11) implies that

$$\|g\|_{\mathcal{H}^*} = \sup_{v\in\mathcal{H}^*, v\neq 0} \frac{|a(u,v)|}{\|v\|_{\mathcal{H}^*}} \leq C\,\|u\|_{\mathcal{H}}\,,$$

and thus $\|u\|_{\mathcal{Z}_{\ell+1}} \leq C\,\|u\|_{\mathcal{H}}$. The bound (1.16) then implies that

$$\|u-u_h\|_{\mathcal{H}} \leq C_2\big(1+\eta(\mathcal{H}_h)\big)\,\|I-\Pi\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}}\,\|u\|_{\mathcal{Z}_{\ell+1}}$$

and (1.19) follows.

2.3. Proof of Theorem 1.6. The following lemma is essentially [6, Theorem 2.6], rewritten in the abstract notation in §1.2.

Lemma 2.2. *Under the assumptions of Theorem 1.5, let $u=\mathcal{R}^*g$ with $\mathcal{R}^*$ defined by (1.13) and $g\in\mathcal{H}_0$. Let $u_m\in\mathcal{H}$, $m=0,\dots,\lfloor \ell/2\rfloor$, be defined by*

$$\widetilde{a}(v,u_0) = \langle v, g\rangle \quad \textit{for all } v\in\mathcal{H}, \tag{2.17}$$

and

$$\widetilde{a}(v,u_m) = \langle Sv, u_{m-1}\rangle \quad \textit{for all } v\in\mathcal{H},\ m=1,\dots,\lfloor \ell/2\rfloor. \tag{2.18}$$

Then

$$u_m\in\mathcal{Z}_{2(m+1)} \quad \textit{with} \quad \|u_m\|_{\mathcal{Z}_{2(m+1)}} \leq C\,\|g\|_{\mathcal{H}_0} \quad \textit{for } m=0,\dots,\lfloor \ell/2\rfloor-1, \tag{2.19}$$

and

$$u_{\lfloor \ell/2\rfloor}\in\mathcal{Z}_{\ell+1} \quad \textit{with} \quad \big\|u_{\lfloor \ell/2\rfloor}\big\|_{\mathcal{Z}_{\ell+1}} \leq C\,\|g\|_{\mathcal{H}_0}\,. \tag{2.20}$$

Furthermore, with

$$r_m := u - \sum_{j=0}^{m-1} u_j, \tag{2.21}$$

$$r_m\in\mathcal{Z}_{2(m+1)} \quad \textit{with} \quad \|r_m\|_{\mathcal{Z}_{2(m+1)}} \leq \big(1+\|\mathcal{R}^*\|_{\mathcal{H}_0\to\mathcal{H}}\big)\,\|g\|_{\mathcal{H}_0} \quad \textit{for } m=0,\dots,\lfloor \ell/2\rfloor-1, \tag{2.22}$$

and

$$r_{\lfloor \ell/2\rfloor}\in\mathcal{Z}_{\ell+1} \quad \textit{with} \quad \big\|r_{\lfloor \ell/2\rfloor}\big\|_{\mathcal{Z}_{\ell+1}} \leq \big(1+\|\mathcal{R}^*\|_{\mathcal{H}_0\to\mathcal{H}}\big)\,\|g\|_{\mathcal{H}_0}\,. \tag{2.23}$$

Proof. We first prove (2.19) by induction. By the definition of u_0 (2.17), continuity and coercivity of $\widetilde{a}$, and boundedness of S (2.3), $\|u_0\|_{\mathcal{H}} \leq C\,\|g\|_{\mathcal{H}_0}$. Then, by (1.8) with $j=2$,

$$\|u_0\|_{\mathcal{Z}_2} \leq C\big(\, \|u_0\|_{\mathcal{H}_0} + \|g\|_{\mathcal{H}_0} \big) \leq C\,\|g\|_{\mathcal{H}_0}\,,$$

which is (2.19) with $m = 0$.

Assume that (2.19) holds with $m = q$. By the definition of u_{q+1} (2.18), continuity and coercivity of $\widetilde{a}$, and boundedness of S (2.3),

$$\left\|u_{q+1}\right\|_{\mathcal{H}} \leq C \left\|u_q\right\|_{\mathcal{H}^*}. \tag{2.24}$$

By (1.8) with $j = 2(q+1)$ and the definition of u_{q+1} (2.18)

$$\left\|u_{q+1}\right\|_{\mathcal{Z}_{2(q+1)}} \leq C\Big(\left\|u_{q+1}\right\|_{\mathcal{H}_0} + \sup_{v\in\mathcal{H},\, \|v\|_{(\mathcal{Z}_{2q})^*}=1} \left|\langle Sv, u_q\rangle\right|\Big). \tag{2.25}$$

By duality

$$\|S\|_{(\mathcal{Z}_j)^*\to\mathcal{H}_0} \leq C \quad j = 0, \ldots, \ell+1,$$

and thus

$$\sup_{v\in\mathcal{H},\, \|v\|_{(\mathcal{Z}_{2q})^*}=1} \left|\langle Sv, u_q\rangle\right| \leq \|S\|_{(\mathcal{Z}_{2q})^*\to\mathcal{H}_0} \left\|u_q\right\|_{\mathcal{H}_0} \leq C \left\|u_q\right\|_{\mathcal{H}_0}. \tag{2.26}$$

Combining (2.25), (2.26), and (2.24), we find that

$$\left\|u_{q+1}\right\|_{\mathcal{Z}_{2(q+2)}} \leq C\big(\left\|u_{q+1}\right\|_{\mathcal{H}_0} + \left\|u_q\right\|_{\mathcal{H}_0} \big) \leq C \left\|u_q\right\|_{\mathcal{H}}.$$

Using (2.19) with $m = q$, we obtain (2.19) with $m = q + 1$, and the induction is complete.

If ℓ is odd, i.e., $\ell + 1$ is even, then this establishes both (2.19) and (2.20) since $2(\lfloor \ell/2\rfloor + 1) = \ell + 1$ (i.e., the highest-order case is even, and can be reached by increasing the regularity at each step by two). If ℓ is even, i.e., $\ell + 1$ is odd, then the argument above establishes (2.19). The bound for $u_{\lfloor \ell/2\rfloor}$ (i.e., (2.20)) then follows from elliptic regularity, using that $u_{\lfloor \ell/2\rfloor-1} = u_{\ell/2-1} \in \mathcal{Z}_\ell \subset \mathcal{Z}_{\ell-1}$ (i.e., at the last step, we only increase the regularity by one).

For the proof that $r_m \in \mathcal{Z}_{2(m+1)}$ and satisfies (2.22), observe that the definition of r_m (2.21) and the definition of u_m (2.18) implies that $r_0 = u$ and

$$\widetilde{a}(v, r_m) = \langle Sv, r_{m-1}\rangle \quad \text{ for all } v \in \mathcal{H},\ m = 1, \ldots, \lfloor \ell/2\rfloor.$$

The proof of (2.22) is then very similar to the proof of (2.19), with the first step being that, by (1.8), the fact that $u = \mathcal{R}^*g$, and the definition of $\mathcal{R}^*$ (1.13),

□

$$\|r_0\|_{\mathcal{Z}_2} = \|u\|_{\mathcal{Z}_2} \leq C\big(\|u\|_{\mathcal{H}_0} + \|g\|_{\mathcal{H}_0} \big) \leq C\big(1 + \|\mathcal{R}^*\|_{\mathcal{H}_0\to\mathcal{H}} \big) \|g\|_{\mathcal{H}_0}.$$

Proof of Theorem 1.6 using Lemma 2.2. As in Lemma 2.2, given $g \in \mathcal{H}_0$, let $u = \mathcal{R}^*g$. By (2.21),

$$\|(I-\Pi)\mathcal{R}^*g\|_{\mathcal{H}} \leq \sum_{j=0}^{\lfloor \ell/2\rfloor-1} \|(I-\Pi)\|_{\mathcal{Z}_{2(j+1)}\to\mathcal{H}} \|u_j\|_{\mathcal{Z}_{2(j+1)}} + \|(I-\Pi)\|_{\mathcal{Z}_{\ell+1}} \left\|r_{\lfloor \ell/2\rfloor}\right\|_{\mathcal{Z}_{\ell+1}}$$

so that, by the bounds (2.19), (2.20), and (2.23),

$$\begin{aligned}\|(I-\Pi)\mathcal{R}^*g\|_{\mathcal{H}} \leq C\Bigg(&\sum_{j=0}^{\lfloor \ell/2\rfloor-1} \|(I-\Pi)\|_{\mathcal{Z}_{2(j+1)}\to\mathcal{H}} \\ &+ \|(I-\Pi)\|_{\mathcal{Z}_{\ell+1}\to\mathcal{H}} \big(1 + \|\mathcal{R}^*\|_{\mathcal{H}_0\to\mathcal{H}} \big)\Bigg) \|g\|_{\mathcal{H}_0};\end{aligned}$$

the result (1.20) then follows from the definition of $\eta(\mathcal{H}_h)$ (1.14). □

3. Elliptic-regularity results. This section collects the elliptic-regularity results that are used to verify that Assumption 1.2 holds for Helmholtz problems with truncation of the exterior domain either by a PML (in §4) or an impedance boundary condition (in §5). Let

$$\mathcal{L}u = -k^{-2}\nabla\cdot(A\nabla u) - c^{-2}u,$$

with associated sesquilinear form

$$a(u,v) = \int_\Omega \Big(k^{-2}(A\nabla u)\cdot\overline{\nabla v} - c^{-2}u\,\overline{v}\Big),$$

where Ω be a bounded Lipschitz domain with outward-pointing unit normal vector n. The conormal derivative $\partial_{n,A}u$ is defined for $u\in H^2(\Omega)$ by $\partial_{n,A}u := n\cdot(A\nabla u)$; recall that $\partial_{n,A}u$ can be defined for $u\in H^1(\Omega)$ with $\mathcal{L}u\in L^2(\Omega)$ by Green's identity; see, e.g., [34, Lemma 4.3].

Assumption 3.1. *For all $x\in\Omega$, $A_{j\ell}(x)=A_{\ell j}(x)$ and*

$$\Re\sum_{j=1}^d\sum_{\ell=1}^d A_{j\ell}(x)\xi_k\overline{\xi_j}\geq c|\xi|^2 \quad \textit{for all } \xi\in\mathbb{C}^d.$$

Theorem 3.2 (Local elliptic regularity near a Dirichlet or Neumann boundary). *Let Ω be a Lipschitz domain and let G_1, G_2 be open subsets of $\mathbb{R}^d$ with $G_1\Subset G_2$ and $G_1\cap\partial\Omega\neq\emptyset$. Let*

$$\Omega_j := G_j\cap\Omega,\ j=1,2, \quad \textit{and} \quad \Gamma_2 := G_2\cap\partial\Omega. \tag{3.1}$$

Suppose that A satisfies Assumption 3.1, $A,c\in C^{m,1}(\overline{\Omega_2})$, $\Gamma_2\in C^{m+1,1}$, $u\in H^1(\Omega_2)$, and $\mathcal{L}u\in H^m(\Omega_2)$ for some $m\in\mathbb{N}$, and either $u=0$ or $\partial_{n,A}u=0$ on Γ_2. Then

$$\|u\|_{H_k^{m+2}(\Omega_1)}\leq C\Big(\|u\|_{H_k^1(\Omega_2)}+\|\mathcal{L}u\|_{H_k^m(\Omega_2)}\Big). \tag{3.2}$$

Proof. In unweighted norms, this follows from, e.g., [34, Theorems 4.7 and 4.16]; the proof in the weighted norms (4.11) is very similar. □

Theorem 3.3 (Local elliptic regularity for the transmission problem). *Let $\Omega_{\rm in}$ be a Lipschitz domain, and let $\Omega_{\rm out} := \mathbb{R}^d\setminus\overline{\Omega_{\rm in}}$. Let G_1, G_2 be open subsets of $\mathbb{R}^d$ with $G_1\Subset G_2$ and $G_1\cap\partial\Omega_{\rm in}\neq\emptyset$. Let*

$$\Omega_{\rm in/out,j} := G_j\cap\Omega_{\rm in/out}, \quad j=1,2, \quad \textit{and } \Gamma_2 := G_2\cap\partial\Omega_{\rm in}.$$

Suppose that A satisfies Assumption 3.1, $A|_{\Omega_{\rm in/out,2}}, c|_{\Omega_{\rm in/out,2}}\in C^{m,1}(\overline{\Omega_{\rm in/out,2}})$, $\Gamma_2\in C^{m+1,1}$, $u_{\rm in/out}\in H^1(\Omega_{\rm in/out})$, and $\mathcal{L}u\in H^m(\Omega_{\rm in/out,2})$ for some $m\in\mathbb{N}$. Suppose further that

$$u_{\rm in}=u_{\rm out} \quad \textit{and} \quad \partial_{n,A}u_{\rm in}=\beta\partial_{n,A}u_{\rm out} \quad \textit{on } \Gamma_2$$

for some $\beta>0$. Then

$$\begin{aligned}&\|u_{\rm in}\|_{H_k^{m+2}(\Omega_{\rm in,1})}+\|u_{\rm out}\|_{H_k^{m+2}(\Omega_{\rm out,1})}\\ &\quad\leq C\Big(\|u_{\rm in}\|_{H_k^1(\Omega_{\rm in,2})}+\|u_{\rm out}\|_{H_k^1(\Omega_{\rm out,2})}+\|\mathcal{L}u_{\rm in}\|_{H_k^m(\Omega_{\rm in,2})}+\|\mathcal{L}u_{\rm out}\|_{H_k^m(\Omega_{\rm out,2})}\Big).\end{aligned} \tag{3.3}$$

Proof. In unweighted norms, this is, e.g., [10, Theorem 5.2.1(i)] (and [34, Theorems 4.7 and 4.16] when $\beta = 1$); the proof in the weighted norms (4.11) is very similar. ∎

Theorem 3.4 (Local elliptic regularity for the impedance problem). *Let Ω be a Lipschitz domain and let G_1, G_2 be open subsets of $\mathbb{R}^d$ with $G_1 \Subset G_2$ and $G_1 \cap \partial\Omega \neq \emptyset$. Let Ω_j and Γ_2 be defined by (3.1). Suppose that, for some $m \in \mathbb{N}$, $\Gamma_2 \in C^{m+1,1}$, $u \in H^1(\Omega_2)$, and $\Delta u \in H^m(\Omega_2)$, and $(k^{-1}\partial_n - \mathrm{i})u = 0$ on Γ_2. Then*

$$\|u\|_{H_k^{m+2}(\Omega_1)} \leq C\Big(\|u\|_{H_k^1(\Omega_2)} + \big\|k^{-2}\Delta u\big\|_{H_k^m(\Omega_2)} \Big). \tag{3.4}$$

Proof. When $m = 0$, the result can be obtained from [7, Lemma 4.1] by multiplying by k^{-2} to switch to weighted norms, and using that the trace operator has norm bounded by $Ck^{1/2}$ from H^1_k to L^2 (which can be obtained from, e.g., [39, Theorem 5.6.4] since the weighted norms there are, up to a constant, the weighted norms (1.5)).

The proof that (3.4) follow for $m > 0$ is then standard and can be found e.g. in [14, §6.3.2, Theorem 5]. We repeat it here in the context of impedance boundary conditions for completeness.

We now prove that if the bound holds for $m = q$, then it holds for $m = q + 1$ (assuming the appropriate regularity of the coefficients and the domain). Without loss of generality, we can change coordinates and work with $U := B(0, s) \cap \{x_d > 0\}$ and $V := B(0, t) \cap \{x_d > 0\}$ for some $0 < t < s$. In these coordinates

$$\mathcal{L}u := (-k^{-2}a^{ij}\partial_{x_i}\partial_{x_j} - k^{-2}(b^i\partial_{x_i} - c))u = f, \qquad (-k^{-1}\partial_{x_d} - \mathrm{i})u = 0 \text{ on } \{x_d = 0\} \cap \overline{U}.$$

Suppose that for some $q \geq 0$, for any $0 < t < s$,

$$\|u\|_{H_k^{q+2}(V)} \leq C_t\big(\|u\|_{L^2(U)} + \|f\|_{H_k^q(U)} \big). \tag{3.5}$$

Now suppose that $f \in H_k^{q+1}(U)$ and $a, b, c \in C^{q+1,1}(\overline{U})$, and let $W := B(0, r) \cap \{x_d > 0\}$ with $t < r < s$. By (3.5),

$$\|u\|_{H_k^{q+2}(W)} \leq C\big(\|u\|_{L^2(U)} + \|f\|_{H_k^q(U)} \big), \tag{3.6}$$

and, by interior elliptic regularity, $u \in H^{q+3}_{\mathrm{loc}}(U)$.

The next step is to bound tangential derivatives of u: let $|\alpha| = q + 1$ with $\alpha_d = 0$ (so that ∂_x^α is a tangential derivative). Let

$$\widetilde{f} := \mathcal{L}\big(k^{-|\alpha|}\partial_x^\alpha u\big) \quad \text{ so that } \quad \widetilde{f} = [\mathcal{L}, k^{-|\alpha|}\partial_x^\alpha]u + k^{-|\alpha|}\partial_x^\alpha f$$

(where $[A, B] := AB - BA$) and, by (3.6) and the fact that the coefficients of $\mathcal{L}$ are $C^{q+1,1}(\overline{U})$,

$$\|\widetilde{f}\|_{L^2(W)} \leq C\big(\|u\|_{H^{q+2}(W)} + \|f\|_{H_k^{q+1}(W)} \big) \leq C\big(\|u\|_{L^2(U)} + \|f\|_{H_k^{q+1}(U)} \big). \tag{3.7}$$

Furthermore

$$(-k^{-1}\partial_{x_d} - \mathrm{i})k^{-|\alpha|}\partial_x^\alpha u|_{x_d=0} = k^{-|\alpha|}\partial_x^\alpha\big[(-k^{-1}\partial_{x_d} - \mathrm{i}u)|_{x_d=0}\big] = 0,$$

so that, by the analogue of (3.5) with $q = 0$ and U replaced by W, (3.6), and (3.7),

$$\big\|k^{-|\alpha|}\partial_x^\alpha u\big\|_{H_k^2(V)} \leq C\big(\big\|k^{-|\alpha|}\partial_x^\alpha u\big\|_{L^2(W)} + \big\|\widetilde{f}\big\|_{L^2(W)}\big) \leq C\big(\|u\|_{L^2(U)} + \|f\|_{H_k^{q+1}(U)} \big).$$

Therefore, by the definition of α,

$$\big\| k^{-|\beta|}\partial_x^\beta u\big\|_{L^2(V)} \leq C\big(\|u\|_{L^2(U)} + \|f\|_{H_k^{q+1}(U)} \big)$$
$$\text{for all } |\beta| = q+3 \text{ with } \beta_d \in \{0,1,2\}. \tag{3.8}$$

To prove that the bound (3.5) holds with q replaced by $q+1$, i.e.,

$$\|u\|_{H_k^{q+3}(V)} \leq C\big(\|u\|_{L^2(U)} + \|f\|_{H_k^{q+1}(U)}\big),$$

it is sufficient to prove that

$$\big\| k^{-|\beta|}\partial_x^\beta u\big\|_{L^2(V)} \leq C\big(\|u\|_{L^2(U)} + \|f\|_{H_k^{q+1}(U)} \big)$$
$$\text{for all } |\beta| = q+3 \text{ with } \beta_d \in \{0,\ldots,q+3\}.$$

We therefore now prove by induction that if

$$\big\| k^{-|\beta|}\partial_x^\beta u\big\|_{L^2(V)} \leq C\big(\|u\|_{L^2(U)} + \|f\|_{H_k^{q+1}(U)} \big) \tag{3.9}$$

for any $|\beta| = q+3$ with $\beta_d \in \{0,\ldots,j\}$ for some $j \in \{2,\ldots,q+2\}$, then (3.9) holds for $|\beta| = q+3$ with $\beta_d = j+1$. Combined with (3.8), this completes the proof.

We therefore assume that $|\beta| = q+3$ with $\beta_d = j+1$. Then, putting $\beta = \gamma+\delta$ with $\delta = (0,\ldots,0,2)$ and $|\gamma| = q+1$, and using that $u \in H^{q+3}_{\rm loc}(U)$, we have

$$k^{-|\gamma|}\partial^\gamma \mathcal{L} u = a^{dd} k^{-|\beta|}\partial^\beta u + Bu \quad \text{ in } V, \tag{3.10}$$

where

$$Bu = \sum_{|\alpha|\leq q+3,\, \alpha_d \leq j} a_\alpha k^{-|\alpha|}\partial_x^\alpha u.$$

By the induction hypothesis (3.9),

$$\|Bu\|_{L^2(V)} \leq C\big(\|u\|_{L^2(U)} + \|f\|_{H_k^{q+1}(U)} \big).$$

Dividing (3.10) by a^{dd}, taking the $L^2(V)$ norm, and using that $1/a^{dd}$ is bounded, we have

$$\|k^{-|\beta|}\partial^\beta u\|_{L^2(V)} \leq C\big(\|u\|_{L^2(U)} + \|f\|_{H_k^{q+1}(U)}\big);$$

i.e., we have proved that (3.9) holds for $|\beta| = q+3$ with $\beta_d = j+1$, and the proof is complete. □

4. Theorem 1.5 applied to the PML problem.

4.1. Definition of the PML problem.

Obstacles and coefficients for Dirichlet/Neumann/penetrable obstacle problem. Let $\Omega_{\rm p}, \Omega_- \subset B_{R_0} := \{x : |x| < R_0\} \subset \mathbb{R}^d$, $d = 2,3$, be bounded open sets with Lipschitz boundaries, $\Gamma_{\rm p}$ and Γ_-, respectively, such that $\Gamma_{\rm p} \cap \Gamma_- = \emptyset$, and $\mathbb{R}^d \overline{\setminus \Omega_-}$ is connected. Let $\Omega_{\rm out} := \mathbb{R}^d \overline{\setminus \Omega_- \cup \Omega_{\rm p}}$ and $\Omega_{\rm in} := (\mathbb{R}^d \overline{\setminus \Omega_-}) \cap \Omega_{\rm p}$.

Let $A_{\rm out} \in C^{0,1}(\Omega_{\rm out}, \mathbb{R}^{d\times d})$ and $A_{\rm in} \in C^{0,1}(\Omega_{\rm in}, \mathbb{R}^{d\times d})$ be symmetric positive definite, let $c_{\rm out} \in L^\infty(\Omega_{\rm out};\mathbb{R})$, $c_{\rm in} \in L^\infty(\Omega_{\rm in};\mathbb{R})$ be strictly positive, and let $A_{\rm out}$ and $c_{\rm out}$ be such that there exists $R_{\rm scat} > R_0 > 0$ such that

$$\overline{\Omega_-} \cup \operatorname{supp}(I - A_{\rm out}) \cup \operatorname{supp}(1 - c_{\rm out}) \Subset B_{R_{\rm scat}}.$$

The obstacle Ω_- is the impenetrable obstacle, on which we impose either a zero Dirichlet or a zero Neumann condition, and the obstacle $\Omega_{\rm in}$ is the penetrable obstacle, across whose boundary we impose transmission conditions.

For simplicity, we do not cover the case when Ω_- is disconnected, with Dirichlet boundary conditions on some connected components and Neumann boundary conditions on others, but the main results hold for this problem too (at the cost of introducing more notation).

Definition of the radial PML. Let $R_{\rm tr} > R_{{\rm PML},-} > R_{\rm scat}$ and let $\Omega_{\rm tr} \subset \mathbb{R}^d$ be a bounded Lipschitz open set with $B_{R_{\rm tr}} \subset \Omega_{\rm tr} \subset B_{CR_{\rm tr}}$ for some $C > 0$ (i.e., $\Omega_{\rm tr}$ has characteristic length scale $R_{\rm tr}$). Let $\Omega := \Omega_{\rm tr} \cap \Omega_+$ and $\Gamma_{\rm tr} := \partial\Omega_{\rm tr}$. For $0 \leq \theta < \pi/2$, let the PML scaling function $f_\theta \in C^3([0,\infty);\mathbb{R})$ be defined by $f_\theta(r) := f(r)\tan\theta$ for some f satisfying

(4.1)
$$\big\{f(r) = 0\big\} = \big\{f'(r) = 0\big\} = \big\{r \leq R_{{\rm PML},-}\big\}, \quad f'(r) \geq 0, \quad f(r) \equiv r \text{ on } r \geq R_{{\rm PML},+};$$

i.e., the scaling "turns on" at $r = R_{{\rm PML},-}$, and is linear when $r \geq R_{{\rm PML},+}$. We emphasize that $R_{\rm tr}$ can be $< R_{{\rm PML},+}$, i.e., we allow truncation before linear scaling is reached. Indeed, $R_{{\rm PML},+} > R_{{\rm PML},-}$ can be arbitrarily large and therefore, given any bounded interval $[0,R]$ and any function $\widetilde{f} \in C^3([0,R])$ satisfying

$$\big\{\widetilde{f}(r) = 0\big\} = \big\{\widetilde{f}'(r) = 0\big\} = \big\{r \leq R_{{\rm PML},-}\big\}, \qquad \widetilde{f}'(r) \geq 0,$$

our results hold for an f with $f|_{[0,R]} = \widetilde{f}$. Given $f_\theta(r)$, let

$$\alpha(r) := 1 + {\rm i} f_\theta'(r) \quad \text{and} \quad \beta(r) := 1 + {\rm i} f_\theta(r)/r. \tag{4.2}$$

and let

(4.3)
$$A := \begin{cases} A_{\rm in} & \text{in } \Omega_{\rm in}, \\ A_{\rm out} & \text{in } \Omega_{\rm out} \cap B_{R_{{\rm PML},-}}, \\ HDH^T & \text{in } (B_{R_{{\rm PML},-}})^c \end{cases} \text{ and } \frac{1}{c^2} := \begin{cases} c_{\rm in}^{-2} & \text{in } \Omega_{\rm in}, \\ c_{\rm out}^{-2} & \text{in } \Omega_{\rm out} \cap B_{R_{{\rm PML},-}}, \\ \alpha(r)\beta(r)^{d-1} & \text{in } (B_{R_{{\rm PML},-}})^c, \end{cases}$$

where, in polar coordinates,

(4.4)
$$D = \begin{pmatrix} \beta(r)\alpha(r)^{-1} & 0 \\ 0 & \alpha(r)\beta(r)^{-1} \end{pmatrix} \quad \text{and} \quad H = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \text{ for } d = 2,$$

and

(4.5)
$$D = \begin{pmatrix} \beta(r)^2\alpha(r)^{-1} & 0 & 0 \\ 0 & \alpha(r) & 0 \\ 0 & 0 & \alpha(r) \end{pmatrix} \text{ and } H = \begin{pmatrix} \sin\theta\cos\phi & \cos\theta\cos\phi & -\sin\phi \\ \sin\theta\sin\phi & \cos\theta\sin\phi & \cos\phi \\ \cos\theta & -\sin\theta & 0 \end{pmatrix}$$

for $d = 3$ (observe that then $A_{\rm out} = I$ and $c_{\rm out}^{-2} = 1$ when $r = R_{{\rm PML},-}$ and thus A and c^{-2} are continuous at $r = R_{{\rm PML},-}$).

We highlight that, in other papers on PMLs, the scaled variable, which in our case is $r + {\rm i} f_\theta(r)$, is often written as $r(1 + {\rm i}\widetilde{\sigma}(r))$ with $\widetilde{\sigma}(r) = \sigma_0$ for r sufficiently large; see, e.g., [24, §4], [4, §2]. Therefore, to convert from our notation, set $\widetilde{\sigma}(r) = f_\theta(r)/r$ and $\sigma_0 = \tan\theta$.

Let

$$\mathcal{H} := H_0^1(\Omega) \quad \text{or} \quad \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_{\rm tr}\}, \tag{4.6}$$

with the former corresponding to zero Dirichlet boundary conditions on Ω_- and the latter corresponding to zero Neumann boundary conditions on Ω_-.

Definition 4.1 (A variational formulation of the PML problem). *Given $G \in (\mathcal{H})^*$ and $\beta > 0$,*

$$\text{find } u \in \mathcal{H} \text{ such that } a(u,v) = G(v) \text{ for all } v \in \mathcal{H}, \tag{4.7}$$

where

$$a(u,v) := \left(\int_{\Omega\cap\Omega_{\rm out}} + \frac{1}{\beta} \int_{\Omega\cap\Omega_{\rm in}} \right) \left(k^{-2} (A\nabla u) \cdot \overline{\nabla v} - c^{-2} u \overline{v} \right). \tag{4.8}$$

When

$$G(v) := \left(\int_{B_{R_{\rm PML,-}}\cap\Omega_{\rm out}} + \frac{1}{\beta} \int_{\Omega\cap\Omega_{\rm in}} \right) c^{-2} g \overline{v} \tag{4.9}$$

for $g \in L^2(\Omega_+)$ with $\operatorname{supp} g \subset B_{R_{\rm PML,-}}$, the variational problem (4.7) is a weak form of the problem

$$\begin{aligned} k^{-2} c_{\rm out}^2 \nabla \cdot (A_{\rm out} \nabla u_{\rm out}) + u_{\rm out} &= -g \quad \text{in } \Omega_{\rm out}, \\ k^{-2} c_{\rm in}^2 \nabla \cdot (A_{\rm in} \nabla u_{\rm in}) + u_{\rm in} &= -g \quad \text{in } \Omega_{\rm in}, \\ u_{\rm in} = u_{\rm out} \quad \text{and} \quad \partial_{n,A_{\rm in}} u_{\rm in} &= \beta \partial_{n,A_{\rm out}} u_{\rm out} \quad \text{on } \partial\Omega_{\rm in}, \\ \text{either} \quad u_{\rm in} = 0 \quad \text{or} \quad \partial_{n,A_{\rm in}} u_{\rm in} &= 0 \quad \text{on } \partial\Omega_-, \end{aligned} \tag{4.10}$$

and with the Sommerfeld radiation condition approximated by a radial PML ((4.7) is obtained by multiplying the PDEs above by $c_{\rm in/out}^{-2} \alpha \beta^{d-1}$ and integrating by parts).

Using the fact that the solution of the true scattering problem exists and is unique with $A_{\rm out}, A_{\rm in}, c_{\rm out}, c_{\rm in}, \Omega_-$, and $\Omega_{\rm in}$ described above, the solution of (4.7) exists and is unique (i) for fixed k and sufficiently large $R_{\rm tr} - R_1$ by [30, Theorem 2.1], [31, Theorem A], [24, Theorem 5.8] and (ii) for fixed $R_{\rm tr} > R_1$ and sufficiently large k by [18, Theorem 1.5].

For the particular data G (4.9), it is well-known that, for fixed k, the error $\|u-v\|_{H_k^1(B_{R_{\rm PML,-}}\setminus\Omega)}$ decays exponentially in $R_{\rm tr} - R_{\rm PML,-}$ and $\tan\theta$; see [30, Theorem 2.1], [31, Theorem A], [24, Theorem 5.8]. It was recently proved in [18, Theorems 1.2 and 1.5] that the error $\|u-v\|_{H_k^1(B_{R_{\rm PML,-}}\setminus\Omega)}$ also decreases exponentially in k.

4.2. Showing that the PML problem fits in the abstract framework used in Theorem 1.5. Recall that $\mathcal{H}$ is defined by (4.6) and let $\mathcal{H}_0 = L^2(\Omega)$. We work with the norm $\|\cdot\|_{H_k^1(\Omega)}$ (1.5) on $\mathcal{H}$, and use below the higher-order norms

$$\|v\|_{H_k^m(\Omega)}^2 := \sum_{0\le|\alpha|\le m} k^{-2|\alpha|} \|\partial^\alpha v\|_{L^2(\Omega)}^2 . \tag{4.11}$$

The rationale for using these norms is that if a function v oscillates with frequency k, then $|(k^{-1}\partial)^\alpha v| \sim |v|$ for all α; this is true, e.g., if $v(x) = \exp(ikx\cdot a)$. We highlight that many papers on the FEM applied to the Helmholtz equation use the weighted H^1 norm $|||v|||^2 := \|\nabla v\|_{L^2(\Omega)}^2 + k^2 \|v\|_{L^2(\Omega)}^2$; we work with (1.5) instead, because weighting the jth derivative with k^{-j} is easier to keep track of than weighting the jth derivative with k^{-j+1}.

We first check that the sesquilinear form a (4.8) is continuous and satisfies a Gårding inequality, with constants uniform for $\epsilon \le \theta \le \pi/2 - \epsilon$.

Lemma 4.2 (Bounds on the coefficients A and c). *Given A and c as in (4.3), a scaling function $f(r)$ satisfying (4.1), and $\epsilon > 0$ there exist A_+ and c_- such that, for all $\epsilon \leq \theta \leq \pi/2 - \epsilon$, $x \in \Omega$, and $\xi, \zeta \in \mathbb{C}^d$,*

$$|(A(x)\xi, \zeta)_2| \leq A_+ \|\xi\|_2 \|\zeta\|_2 \quad \text{and} \quad \frac{1}{|c(x)|^2} \geq \frac{1}{c_-^2}.$$

Proof. This follows from the definitions of A and c in (4.3), the definitions of α and β in (4.2), and the fact that $f_\theta(r) := f(r)\tan\theta$. ∎

Continuity of a (1.6) with $C_{\text{cont}} := \max\{A_+, c_-^{-2}\}$ then follows from the Cauchy-Schwarz inequality and the definition of $\|\cdot\|_{H^1_k(\Omega)}$ (1.5).

Assumption 4.3. *When $d = 3$, $f_\theta(r)/r$ is nondecreasing.*

Assumption 4.3 is standard in the literature; e.g., in the alternative notation described above it is that $\widetilde{\sigma}$ is non-decreasing – see [4, §2].

Remark 4.4. *As noted above, the variational problem (4.7) is obtained by multiplying the PDEs in (4.10) by $c_{\text{in/out}}^{-2}\alpha\beta^{d-1}$ and integrating by parts (as in [9, §3]). If one integrates by parts the PDEs directly (as in, e.g., [24, Lemma 4.2 and Equation 4.8]), the resulting sesquilinear form satisfies Assumption 1.2 after multiplication by $\mathrm{e}^{\mathrm{i}\omega}$, for some suitable ω (see Remark 1.3), without the need for Assumption 4.3.*

Lemma 4.5. *Suppose that f_θ satisfies Assumption 4.3. With A defined by (4.3), given $\epsilon > 0$ there exists $A_- > 0$ such that, for all $\epsilon \leq \theta \leq \pi/2 - \epsilon$,*

$$\Re\big(A(x)\xi, \xi\big)_2 \geq A_- \|\xi\|_2^2 \quad \text{for all } \xi \in \mathbb{C}^d \text{ and } x \in \Omega_+.$$

Reference for the proof. See, e.g., [20, Lemma 2.3]. ∎

Corollary 4.6. *If f_θ satisfies Assumption 4.3 then*

$$\Re a(w, w) \geq A_- \|w\|^2_{H^1_k(\Omega)} - \big(A_- + c_{\min}^{-2}\big)\|w\|^2_{L^2(\Omega)} \quad \text{for all } w \in \mathcal{H}.$$

Let $\mathcal{R} : L^2(\Omega) \to \mathcal{H}$ be defined by $a(\mathcal{R}g, v) = (g, v)_{L^2(\Omega)}$ for all $v \in \mathcal{H}$; i.e., $\mathcal{R}$ is the solution operator of the PML problem. The definition of a and the facts that (with the matrices H and D defined by (4.4), (4.5)) H is real and the matrix D is diagonal (and hence symmetric) imply that $a(\overline{u}, v) = a(\overline{v}, u)$ for all $u, v \in \mathcal{H}$, and thus $\mathcal{R}g = \overline{\mathcal{R}^*\overline{g}}$. We therefore let

$$C_{\text{sol}} := \|\mathcal{R}\|_{L^2(\Omega)\to\mathcal{H}} = \|\mathcal{R}^*\|_{L^2(\Omega)\to\mathcal{H}}. \tag{4.12}$$

We highlight that (i) C_{sol} is bounded by the norm of the solution operator of the true scattering problem (i.e., with the Sommerfeld radiation condition) by [18, Theorem 1.6], (ii) $C_{\text{sol}} \sim k$ when the problem is nontrapping (with this the slowest-possible growth in k), and (iii) an advantage of working with the weighted norms (4.11) is that C_{sol} in fact describes the k-dependence of the Helmholtz solution operator between H_k^m and H_k^{m+2} for any m.

Lemma 4.7 (The PML problem satisfies Assumption 1.2). *Suppose that, for some $\ell \in \mathbb{Z}^+$, $A_{\text{out}}, A_{\text{in}}, c_{\text{out}}, c_{\text{in}} \in C^{\ell-1,1}$ and $f_\theta \in C^{\ell,1}$ on the closures of the domains on which they are defined, $\partial\Omega$ is $C^{\ell,1}$, and f_θ satisfies Assumption 4.3. Let*

$$\mathcal{Z}_j = \big\{v : v_{\text{out}} \in H^j(\Omega \cap \Omega_{\text{out}}),\, v_{\text{in}} \in H^j(\Omega_{\text{in}})\big\} \cap \mathcal{H} \tag{4.13}$$

with norm

$$\|v\|^2_{\mathcal{Z}_j} := \|v_{\rm out}\|^2_{H^j_k(\Omega_{\rm out}\cap\Omega)} + \|v_{\rm in}\|^2_{H^j_k(\Omega_{\rm in})}\,. \tag{4.14}$$

where the "out" and "in" subscripts denote restriction to $\Omega_{\rm out}\cap\Omega$ *and* $\Omega_{\rm in}$*, respectively.*

Then a *defined by* (4.8) *satisfies Assumption* 1.2 *and given* $\epsilon > 0$ *and* $k_0 > 0$ *there exists* $C > 0$ *such the bounds* (1.8) *and* (1.9) *hold for all* $k \geq k_0$ *and* $\epsilon \leq \theta \leq \pi/2 - \epsilon$*.*

Proof. First observe that Assumption 3.1 is satisfied by the definition (4.3) of A. Since

$$\sup_{v\in\mathcal{H},\,\|v\|_{(\mathcal{Z}_{j-2})^*}=1} |a(u,v)| = \|\mathcal{L}u\|_{\mathcal{Z}_{j-2}}\,,$$

the bound (1.9) holds by combining Theorem 3.2 (used near Γ_- and $\Gamma_{\rm tr}$) and Theorem 3.3 (used near $\Gamma_{\rm p}$) and using the fact that, by Green's identity, for $u \in H^1_0(\Omega)$ with $\mathcal{L}u \in L^2(\Omega)$ and $\partial_{n,A_{\rm in}} u_{\rm in} = \beta \partial_{n,A_{\rm out}} u_{\rm out}$ on $\partial\Omega_{\rm in}$,

$$\begin{aligned}&\|u_{\rm in}\|_{H^1_k(\Omega_{\rm in})} + \|u_{\rm out}\|_{H^1_k(\Omega_{\rm out})}\\ &\quad\leq C\Big(\|u_{\rm in}\|_{L^2(\Omega_{\rm in})} + \|u_{\rm out}\|_{L^2(\Omega_{\rm out})} + \|\mathcal{L}u_{\rm in}\|_{L^2(\Omega_{\rm in})} + \|\mathcal{L}u_{\rm out}\|_{L^2(\Omega_{\rm out})}\Big)\end{aligned}$$

(so that the H^1_k norms on the right-hand sides of (3.2) and (3.3) can be replaced by L^2 norms). Since the operator associated with the sesquilinear form $\Re a$ is

$$\left(\frac{\mathcal{L}+\mathcal{L}^*}{2}\right)u = -k^{-2}\nabla\cdot\left(\frac{A+\overline{A}}{2}\nabla u\right) - \left(\frac{c^{-2}+\overline{c}^{-2}}{2}\right)u$$

and the matrix A is symmetric, this operator also satisfies Assumption 3.1. The bound (1.8) then holds by a very similar argument. □

4.3. Theorem 1.5 applied to the PML problem.

Assumption 4.8. *Given* $p \in \mathbb{Z}^+$, $(\mathcal{H}_h)_{h>0}$ *are such that the following holds. There exists* $C > 0$ *such that, for all* $h > 0$, $0 \leq j \leq m+1 \leq p+1$*, and* $v \in \mathcal{H}\cap H^{\ell+1}(\Omega)$ *there exists* $\mathcal{I}_{h,p}v \in \mathcal{H}_h$ *such that*

$$\begin{aligned}&\big|v_{\rm out} - (\mathcal{I}_{h,p}v)_{\rm out}\big|_{H^j(\Omega_{\rm out}\cap\Omega)} + \big|v_{\rm in} - (\mathcal{I}_{h,p}v)_{\rm in}\big|_{H^j(\Omega_{\rm in})}\\ &\qquad\leq Ch^{m+1-j}\big(\|v_{\rm out}\|_{H^{m+1}(\Omega_{\rm out}\cap\Omega)} + \|v_{\rm in}\|_{H^{m+1}(\Omega_{\rm in})}\big).\end{aligned} \tag{4.15}$$

where the "out" and "in" subscripts denote restriction to $\Omega_{\rm out}\cap\Omega$ *and* $\Omega_{\rm in}$*, respectively.*

Assumption 4.8 holds when $(\mathcal{H}_h)_{h>0}$ consists of piecewise degree-p polynomials on shape-regular simplicial triangulations, indexed by the meshwidth; see, e.g., [8, Theorem 17.1], [5, Proposition 3.3.17].

Theorem 4.9 (Existence, uniqueness, and error bound in the preasymptotic regime for the PML problem). *Suppose that, for some* $\ell \in \mathbb{Z}^+$, $A_{\rm out}, A_{\rm in}, c_{\rm out}, c_{\rm in} \in C^{\ell-1,1}$ *and* $f_\theta \in C^{\ell,1}$ *on the closures of the domains where they are defined,* $\partial\Omega$ *is* $C^{\ell,1}$, f_θ *satisfies Assumption 4.3, and* $\beta > 0$*. Let* $C_{\rm sol}$ *be defined by* (4.12)*, and assume that* $\{\mathcal{H}_h\}_{h>0}$ *satisfy Assumption 4.8. Given* $\epsilon > 0$ *and* $p \in \mathbb{Z}^+$ *with* $p \geq \ell$*, there exists* $k_0 > 0$ *and* $C_j, j = 1,2,3,4$*, such that the following is true for all* $k \geq k_0$, $\epsilon \leq \theta \leq \pi/2 - \epsilon$*, and* $R_{\rm tr} > R_1 + \epsilon$*.*

The solution u *of the PML problem* (4.7) *exists and is unique, and if*

$$(hk)^{2\ell} C_{\rm sol} \leq C_1 \tag{4.16}$$

then the Galerkin solution u_h, exists, is unique, and satisfies

$$\left\|u-u_h\right\|_{H_k^1(\Omega)} \leq C_2\Big(1+hk+(hk)^\ell C_{\rm sol}\Big) \min_{w_h\in\mathcal{H}_h} \left\|u-v_h\right\|_{H_k^1(\Omega)}, \tag{4.17}$$

$$\left\|u-u_h\right\|_{L^2(\Omega)} \leq C_3\Big(hk+(hk)^\ell C_{\rm sol}\Big) \min_{w_h\in\mathcal{H}_h} \left\|u-v_h\right\|_{H_k^1(\Omega)}. \tag{4.18}$$

If, in addition, $g\in H^{p-1}(\Omega)\cap\mathcal{H}$ (with $\mathcal{H}$ defined by (4.6)) with

$$\left\|g\right\|_{H_k^{p-1}(\Omega)} \leq C\left\|g\right\|_{\mathcal{H}^*} \tag{4.19}$$

for some $C>0$, then there exists $C_4>0$ such that if h satisfies (4.16) then

$$\frac{\left\|u-u_h\right\|_{H_k^1(\Omega)}}{\left\|u\right\|_{H_k^1(\Omega)}} \leq C_4\Big(hk+(hk)^\ell C_{\rm sol}\Big)(hk)^\ell. \tag{4.20}$$

Theorem 4.9 is most interesting when $p=\ell$, i.e., the polynomial degree is the smallest possible covered by the theorem. In this case, (4.16) becomes the condition (1.1), and the bounds (4.17), (4.18), and (4.20) become (1.2), (1.3), and (1.4), respectively.

Proof of Theorem 4.9. By the results in §4.2, a defined by (4.8) satisfies the assumptions of Theorem 1.5. By (4.15), the definition of $\|\cdot\|_{\mathcal{Z}_j}$ (4.14), and the definition (4.11) of the weighted norms, $\|I-\Pi\|_{\mathcal{Z}_{m+1}\to\mathcal{H}}\leq C(hk)^m$. This bound along with Theorem 1.6 and (4.12) imply that

$$\eta(\mathcal{H}_h) \leq C\bigg(\sum_{j=0}^{\lfloor \ell/2\rfloor-1}(hk)^{2j+1}+(hk)^\ell C_{\rm sol}\bigg).$$

If $hk\leq C$, then $\eta(\mathcal{H}_h)\leq C(hk+(hk)^\ell C_{\rm sol})$; the result then follows from Theorem 1.5 and the fact that if the condition (4.16) holds, then $hk\leq C$ (since $C_{\rm sol}\geq Ck$). □

5. Theorem 1.5 applied to the impedance problem.

5.1. Definition of the impedance problem. Let $A_{\rm out}, A_{\rm in}, c_{\rm out}, c_{\rm in}, \Omega_-, \Omega_{\rm in}$, and $\Omega_{\rm tr}$ be as in §4.1. Let

$$A:=\begin{cases}A_{\rm in} & \text{in } \Omega_{\rm in},\\ A_{\rm out} & \text{in } \Omega_{\rm out}\cap\Omega,\end{cases} \qquad \text{and} \qquad \frac{1}{c^2}:=\begin{cases}c_{\rm in}^{-2} & \text{in } \Omega_{\rm in},\\ c_{\rm out}^{-2} & \text{in } \Omega_{\rm out}\cap\Omega.\end{cases}$$

Let

$$\mathcal{H}:=\{v\in H^1(\Omega)\,:\,v=0 \text{ on } \partial\Omega_-\} \quad \text{or} \quad H^1(\Omega), \tag{5.1}$$

with the former corresponding to zero Dirichlet boundary conditions on Ω_- and the latter corresponding to zero Neumann boundary conditions on Ω_-.

Definition 5.1 (Variational formulation of the impedance problem). *Given $G\in(\mathcal{H})^*$ and $\beta>0$,*

$$\text{find } u\in\mathcal{H} \text{ such that } a(u,v)=G(v) \text{ for all } v\in\mathcal{H}, \tag{5.2}$$

where

$$a(u,v):=\bigg(\int_{\Omega\cap\Omega_{\rm out}}+\frac{1}{\beta}\int_{\Omega\cap\Omega_{\rm in}}\bigg)\Big(k^{-2}(A\nabla u)\cdot\overline{\nabla v}-c^{-2}u\overline{v}\Big)-{\rm i}k^{-1}\int_{\Gamma_{\rm tr}}u\overline{v}. \tag{5.3}$$

The solution of this variational problem exists and is unique by, e.g., [22, Theorem 2.4].

5.2. Showing that the impedance problem fits in the abstract framework used in Theorem 1.5. The proofs that the sesquilinear form a is continuous and satisfies a Gårding inequality are very similar to those for the PML problem in §4.2 (in fact, they are simpler because there is no PML scaling parameter in which the bounds need to be uniform).

Lemma 5.2 (The impedance problem satisfies Assumption 1.2). *Suppose that, for some $\ell \in \mathbb{Z}^+$, $A_{\rm out}, A_{\rm in}, c_{\rm out}, c_{\rm in} \in C^{\ell-1,1}$ on the closures of the domains on which they are defined, and $\partial\Omega$ is $C^{\ell,1}$. With $\mathcal{Z}_j$ and its norm defined by (4.13) and (4.14), a defined by (5.3) satisfies Assumption 1.2 and given $k_0 > 0$ there exists $C > 0$ such the bounds (1.8) and (1.9) hold for all $k \geq k_0$.*

Proof. This is very similar to the proof of Lemma 4.7. The regularity assumption (1.8) follows by combining Theorem 3.2 used near $\partial\Omega_-$, Theorem 3.3 used near $\partial\Omega_{\rm in}$, and Theorem 3.4 used near $\Gamma_{\rm tr}$. The regularity assumption (1.9) follows by combining Theorem 3.2 used near $\partial\Omega_-$, Theorem 3.3 used near $\partial\Omega_{\rm in}$, and now Theorem 3.2 (with Neumann boundary condition) used near $\Gamma_{\rm tr}$. Indeed, near $\Gamma_{\rm tr}$, the operator associated with $(\Re a)$ is $-k^{-2}\Delta - 1$ with Neumann boundary conditions (coming from $A_{\rm out} = I$ and $c_{\rm out} = 1$ near $\Gamma_{\rm tr}$ and the fact that no boundary condition is imposed on $\Gamma_{\rm tr}$ in $\mathcal{H}$ (5.1)). □

5.3. Theorem 1.5 applied to the impedance problem.

Theorem 5.3 (Existence, uniqueness, and error bound in the preasymptotic regime for the impedance problem). *Suppose that, for some $\ell \in \mathbb{Z}^+$, $A_{\rm out}, A_{\rm in}, c_{\rm out}, c_{\rm in} \in C^{\ell-1,1}$ on the closures of the domains where they are defined, $\partial\Omega$ is $C^{\ell,1}$, and $\beta > 0$. Let $C_{\rm sol}$ be defined by (4.12), and assume that $\{\mathcal{H}_h\}_{h>0}$ satisfy Assumption 4.8. Given $p \in \mathbb{Z}^+$ with $p \geq \ell$, there exists $k_0 > 0$ and $C_j, j = 1, 2, 3, 4$, such that the following is true for all $k \geq k_0$.*

The solution u of the impedance problem (5.2) exists and is unique, and if (4.16) holds then the Galerkin solution u_h, exists, is unique, and satisfies the bounds (4.17) and (4.18). If, in addition, $g \in H^{p-1}(\Omega) \cap \mathcal{H}$ (with $\mathcal{H}$ defined by (5.1)) with (4.19) for some $C > 0$, then there exists $C_4 > 0$ such that if h satisfies (4.16) then the bound (4.20) holds.

Given Lemma 5.2, the proof of Theorem 5.3 is very similar to the proof of Theorem 4.9.

Remark 5.4 (Imposing the exact Dirichlet-to-Neumann map on $\Gamma_{\rm tr}$). *With the exact Dirichlet-to-Neumann map imposed on $\Gamma_{\rm tr}$, the Helmholtz sesquilinear form is continuous and satisfies a Gårding inequality (see, e.g., [37, Lemma 3.3 and Corollary 3.4]). To apply Theorem 1.5 to this problem, one therefore only needs to check the elliptic-regularity assumptions of Assumption 1.2. Using Theorems 3.2 and 3.3, this boils down to knowing the analogue of Theorem 3.4 with the impedance boundary condition replaced by $k^{-1}\partial_n u = \mathrm{DtN} u$ (for (1.8)) and also $k^{-1}\partial_n u = (\mathrm{DtN}+\mathrm{DtN}^*)u/2$ (for (1.9)). When $m = 0$ (i.e., the lowest-order regularity shift covered in Theorem 3.4), the first of these regularity results is given by [28, Theorem 6.1]. To prove this result for $m > 1$ one would need to make an argument similar to that in the proof of Theorem 3.4 except that, because DtN and DtN^* do not commute with tangential derivatives, one would need to obtain two additional estimates: 1) estimates on u with nontrivial boundary data, e.g., when $k^{-1}\partial_n u - (\mathrm{DtN})u = g \in H^s_k$ and 2) trace estimates for u that are needed to bound, e.g., $[T, \mathrm{DtN}]u$ where T is a vector field tangent to the boundary. The same strategy could also be used to handle higher-order* *impedance boundary conditions.*

Acknowledgements. EAS was supported by EPSRC grant EP/R005591/1 and JG was supported by EPSRC grants EP/V001760/1 and EP/V051636/1.

REFERENCES

[1] M. Ainsworth, *Discrete dispersion relation for hp-version finite element approximation at high wave number*, SIAM Journal on Numerical Analysis, 42 (2004), pp. 553–575.

[2] A. K. Aziz, R. B. Kellogg, and A. B. Stephens, *A two point boundary value problem with a rapidly oscillating solution*, Numer. Math., 53 (1988), pp. 107–121.

[3] M. Bernkopf, T. Chaumont-Frelet, and J. M. Melenk, *Stability and convergence of Galerkin discretizations of the Helmholtz equation in piecewise smooth media*, arXiv preprint arXiv:2209.03601, (2022).

[4] J. H. Bramble and J. Pasciak, *Analysis of a finite PML approximation for the three dimensional time-harmonic Maxwell and acoustic scattering problems*, Mathematics of Computation, 76 (2007), pp. 597–614.

[5] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, vol. 15 of Texts in Applied Mathematics, Springer, 3rd ed., 2008.

[6] T. Chaumont-Frelet and S. Nicaise, *Wavenumber explicit convergence analysis for finite element discretizations of general wave propagation problem*, IMA J. Numer. Anal., 40 (2020), pp. 1503–1543.

[7] T. Chaumont-Frelet, S. Nicaise, and J. Tomezyk, *Uniform a priori estimates for elliptic problems with impedance boundary conditions*, Communications on Pure & Applied Analysis, 19 (2020), p. 2445.

[8] P. G. Ciarlet, *Basic error estimates for elliptic problems*, in Handbook of numerical analysis, Vol. II, North-Holland, Amsterdam, 1991, pp. 17–351.

[9] F. Collino and P. Monk, *The perfectly matched layer in curvilinear coordinates*, SIAM Journal on Scientific Computing, 19 (1998), pp. 2061–2090.

[10] M. Costabel, M. Dauge, and S. Nicaise, *Corner Singularities and Analytic Regularity for Linear Elliptic Systems. Part I: Smooth domains.*, (2010). https://hal.archives-ouvertes.fr/file/index/docid/453934/filename/CoDaNi_Analytic_Part_I.pdf.

[11] J. Douglas Jr., J. E. Santos, D. Sheen, and L. S. Bennethum, *Frequency domain treatment of one-dimensional scalar waves*, Mathematical Models and Methods in Applied Sciences, 3 (1993), pp. 171–194.

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

[13] S. Esterhazy and J. M. Melenk, *On stability of discretizations of the Helmholtz equation*, in Numerical Analysis of Multiscale Problems, I. G. Graham, T. Y. Hou, O. Lakkis, and R. Scheichl, eds., Springer, 2012, pp. 285–324.

[14] L. C. Evans, *Partial differential equations*, American Mathematical Society Providence, RI, 1998.

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

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

[17] J. Galkowski, *Lower bounds for piecewise polynomial approximations of oscillatory functions*, arXiv preprint arXiv:2211.04757, (2022).

[18] J. Galkowski, D. Lafontaine, and E. A. Spence, *Perfectly-matched-layer truncation is exponentially accurate at high frequency*, arXiv preprint arXiv:2105.07737, (2021).

[19] J. Galkowski, D. Lafontaine, E. A. Spence, and J. Wunsch, *Decompositions of high-frequency Helmholtz solutions via functional calculus, and application to the finite element method*, arXiv preprint arXiv:2102.13081, (2021).

[20] J. Galkowski, D. Lafontaine, E. A. Spence, and J. Wunsch, *The hp-FEM applied to the Helmholtz equation with PML truncation does not suffer from the pollution effect*, arXiv preprint arXiv:2207.05542, (2022).

[21] J. Galkowski, E. A. Spence, and J. Wunsch, *Optimal constants in nontrapping resolvent estimates*, Pure and Applied Analysis, 2 (2020), pp. 157–202.

[22] I. G. Graham and S. A. Sauter, *Stability and finite element error analysis for the Helmholtz equation with variable coefficients*, Math. Comp., 89 (2020), pp. 105–138.

[23] I. Harari and T. J. R. Hughes, *Finite element methods for the Helmholtz equation in an exterior domain: model problems*, Computer methods in applied mechanics and engineering, 87 (1991), pp. 59–96.

[24] T. Hohage, F. Schmidt, and L. Zschiedrich, *Solving time-harmonic scattering problems based on the pole condition II: convergence of the PML method*, SIAM Journal on Mathematical Analysis, 35 (2003), pp. 547–560.

[25] F. Ihlenburg and I. Babuška, *Finite element solution of the Helmholtz equation with high wave number Part I: The h-version of the FEM*, Comput. Math. Appl., 30 (1995), pp. 9–37.

[26] F. Ihlenburg and I. Babuska, *Finite element solution of the Helmholtz equation with high wave number part II: the hp version of the FEM*, SIAM J. Numer. Anal., 34 (1997), pp. 315–358.

[27] F. Ihlenburg and I. Babuška, *Dispersion analysis and error estimation of Galerkin finite element methods for the Helmholtz equation*, Int. J. Numer. Meth. Eng., 38, Issue 22 (1995), pp. 3745–3774.

[28] D. Lafontaine, E. A. Spence, and J. Wunsch, *A sharp relative-error bound for the Helmholtz h-FEM at high frequency*, Numerische Mathematik, 150 (2022), pp. 137–178.

[29] D. Lafontaine, E. A. Spence, and J. Wunsch, *Wavenumber-explicit convergence of the hp-FEM for the full-space heterogeneous Helmholtz equation with smooth coefficients*, Comp. Math. Appl., 113 (2022), pp. 59–69.

[30] M. Lassas and E. Somersalo, *On the existence and convergence of the solution of PML equations*, Computing, 60 (1998), pp. 229–241.

[31] M. Lassas and E. Somersalo, *Analysis of the PML equations in general convex geometry*, Proceedings of the Royal Society of Edinburgh Section A: Mathematics, 131 (2001), pp. 1183–1207.

[32] Y. Li and H. Wu, *FEM and CIP-FEM for Helmholtz Equation with High Wave Number and Perfectly Matched Layer Truncation*, SIAM J. Numer. Anal., 57 (2019), pp. 96–126.

[33] C. H. Makridakis, F. Ihlenburg, and I. Babuška, *Analysis and finite element methods for a fluid-solid interaction problem in one dimension*, Mathematical Models and Methods in Applied Sciences, 6 (1996), pp. 1119–1141.

[34] W. McLean, *Strongly elliptic systems and boundary integral equations*, Cambridge University Press, 2000.

[35] J. M. Melenk, *On generalized finite element methods*, PhD thesis, The University of Maryland, 1995.

[36] J. M. Melenk, A. Parsania, and S. Sauter, *General DG-methods for highly indefinite Helmholtz problems*, Journal of Scientific Computing, 57 (2013), pp. 536–581.

[37] J. M. Melenk and S. Sauter, *Convergence analysis for finite element discretizations of the Helmholtz equation with Dirichlet-to-Neumann boundary conditions*, Math. Comp, 79 (2010), pp. 1871–1914.

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

[39] J. C. Nédélec, *Acoustic and electromagnetic equations: integral representations for harmonic problems*, Springer Verlag, 2001.

[40] O. R. Pembery, *The Helmholtz Equation in Heterogeneous and Random Media: Analysis and Numerics*, PhD thesis, University of Bath, 2020. https://researchportal.bath.ac.uk/en/studentTheses/the-helmholtz-equation-in-heterogeneous-and-random-media-analysis.

[41] S. A. Sauter, *A refined finite element convergence theory for highly indefinite Helmholtz problems*, Computing, 78 (2006), pp. 101–115.

[42] A. H. Schatz, *An observation concerning Ritz-Galerkin methods with indefinite bilinear forms*, Math. Comp., 28 (1974), pp. 959–962.

[43] L. L. Thompson and P. M. Pinsky, *Complex wavenumber Fourier analysis of the p-version finite element method*, Computational Mechanics, 13 (1994), pp. 255–275.

[44] H. Wu, *Pre-asymptotic error analysis of CIP-FEM and FEM for the Helmholtz equation with high wave number. Part I: linear version*, IMA J. Numer. Anal., 34 (2014), pp. 1266–1288.

[45] L. Zhu and H. Wu, *Preasymptotic error analysis of CIP-FEM and FEM for Helmholtz equation with high wave number. Part II: hp version*, SIAM J. Numer. Anal., 51 (2013), pp. 1828–1852.