# Numerical analysis at the semiclassical analysis/numerical analysis interface: issues and case studies

Simon Chandler-Wilde, Euan Spence

## 1. Introduction

This is the first of three talks that kicked off this programme, introducing issues and problems at the interface between semiclassical analysis (SCA) and numerical analysis (NA) from the NA side, and exhibiting opportunities at the SCA/NA interface through case studies.

1.1. **The model problem.** We focus on a model problem of obstacle scattering in time-harmonic acoustics. Let $\Omega_- \subset \mathbb{R}^d$ $(d \geq 2)$ be a bounded Lipschitz open set (the obstacle) such that $\Omega := \mathbb{R}^d \setminus \overline{\Omega_-}$ is a connected Lipschitz domain. The *scattering problem* we consider is: given $k > 0$ (the wavenumber) and an incident plane wave $u^I(x) := \mathrm{e}^{\mathrm{i}kx\cdot d}$, travelling in the direction of the unit vector $d$, find $u \in C^2(\Omega) \cap H^1_{\mathrm{loc}}(\Omega)$ such that

$$\Delta u + k^2 u = 0 \text{ in } \Omega, \quad u = 0 \text{ on } \Gamma := \partial\Omega, \tag{1}$$

and such that the scattered field $u^S := u - u^I$ satisfies the standard Sommerfeld radiation condition (SRC)

$$\partial_r u^S(x) - \mathrm{i}ku^S(x) = o\big(r^{(1-d)/2}\big) \text{ as } r := |x| \to \infty, \text{ uniformly in } \widehat{x} := x/r.$$

Both SCA and NA seek to understand $u$, and solution operators, for the above problem. The key NA goal is: *compute $u$ for fixed but arbitrarily large $k$, to arbitrarily high accuracy, as efficiently as possible.*

1.2. **The Galerkin method.** The standard *Galerkin method (GM)* for solving the above problem starts from a variational formulation: find $v \in \mathcal{H}$ (some complex Hilbert space) such that

$$a(v,w) = F(w) \quad \forall w \in \mathcal{H}, \tag{2}$$

where $a(\cdot,\cdot)$ and $F(\cdot)$ are, respectively, some continuous sesquilinear form and continuous anti-linear functional on $\mathcal{H}$. We choose a sequence $(\mathcal{H}_N)_{N=1}^{\infty}$ of finite dimensional subspaces of $\mathcal{H}$, and, for each $N \in \mathbb{N}$, seek $v_N \in \mathcal{H}_N$ such that

$$a(v_N, w_N) = F(w_N) \quad \forall w_N \in \mathcal{H}_N. \tag{3}$$

To solve our model problem by the GM there are three choices to make:

i) The variational formulation, notably whether to use a *domain-based* formulation or a *boundary-based* formulation; see §2 below.

ii) The choice for $\mathcal{H}_N$. We discuss classical piecewise-polynomial (finite element) subspaces in §3; choices adapted to (1) are discussed in the articles by Ecevit, Chaumont-Frelet, and Moiola in this volume.

iii) How to solve the linear system associated to (3); see the discussion in the article by Gander.

## 2. Variational formulations

2.1. **Domain-based.** The standard domain-based variational formulation is set in a bounded Lipschitz domain $\Omega_R$ with $\overline{\Omega_-} \subset \Omega_R \subset \Omega$ (commonly $\Omega_R := \Omega \cap B_R$, where $B_R := \{x \in \mathbb{R}^d : |x| < R\}$, for some $R > 0$). The unknown is $v := u|_{\Omega_R} \in \mathcal{H} := H^1_0(\Omega_R)$, where $H^1_0(\Omega_R)$ is the closure in $H^1(\Omega_R)$ of $\mathcal{D}_R := \{\phi|_{\Omega_R} : \phi \in C_0^\infty(\Omega)\}$. To obtain (2) multiply the Helmholtz equation (1) by $w \in \mathcal{D}_R$ and integrate by parts. This gives (2) for $w \in \mathcal{D}_R$, so, by density, for all $w \in \mathcal{H}$, where

$$\begin{aligned} a(v,w) &:= \int_{\Omega_R} \nabla v \cdot \nabla \bar{w} - k^2 v \bar{w} - \int_{\Gamma_R} \mathrm{DtN}_k(\gamma v)\gamma \bar{w}\,\mathrm{d}s, \\ F(w) &:= \int_{\Gamma_R} \left(\partial_n u^{\mathrm{I}} - \mathrm{DtN}_k(\gamma u^{\mathrm{I}})\right)\gamma \bar{w}\,\mathrm{d}s \quad \forall v, w \in \mathcal{H}, \end{aligned} \tag{4}$$

$\gamma : H^1(\Omega_R) \to H^{1/2}(\partial\Omega_R)$ is the standard trace operator and $\Gamma_R := \partial\Omega_R \setminus \Gamma$ is the exterior boundary of $\Omega_R$. $\mathrm{DtN}_k$ denotes the exact Dirichlet to Neumann (DtN) map for the domain $\Omega_R^+ := \mathbb{R}^d \setminus \overline{\Omega_R \cup \Omega_-}$ exterior to $\Gamma_R$. Thus, for $g \in H^{1/2}(\Gamma_R)$, $\mathrm{DtN}_k g = \partial_n u$, where $u \in C^2(\Omega_R^+) \cap H^1_{\mathrm{loc}}(\Omega_R^+)$ is the unique solution to the Helmholtz equation (1) in $\Omega_R^+$ that satisfies the SRC and $u = g$ on $\Gamma_R$. If $\Gamma_R = \partial B_R$ the action of $\mathrm{DtN}_k$ can be calculated by separation of variables, but, even when $\Gamma_R = \partial B_R$, it can be attractive, for efficiency, to approximate $\mathrm{DtN}_k$ by a local absorbing boundary condition approximating the SRC, the simplest of which is the impedance boundary condition[1]

$$\partial_n u - \mathrm{i}ku = 0 \quad \text{on } \Gamma_R, \tag{5}$$

or to approximate $\mathrm{DtN}_k$ using PML (complex scaling in a layer around $\Omega_R$ with $u = 0$ on the outer boundary); see [7] and the references therein.

2.2. **Boundary-based.** Alternatively one can derive a variational formulation (2) via a boundary integral equation (BIE) formulation. The so-called *direct* route to a BIE is Green's representation theorem [2, Thm. 2.21], that, for $x \in \Omega$,

$$\begin{aligned} u^{\mathrm{S}}(x) &= -\int_\Gamma \left(\Phi(x,y)\partial_n u^{\mathrm{S}}(y) - \partial_{n(y)}\Phi(x,y)\gamma u^{\mathrm{S}}(x)\right)\,\mathrm{d}s(y) \\ &= -\int_\Gamma \left(\Phi(x,y)\partial_n u^{\mathrm{S}}(y) + \partial_{n(y)}\Phi(x,y) u^{\mathrm{I}}(x)\right)\,\mathrm{d}s(y), \end{aligned}$$

where we've used the boundary condition (1) to obtain the 2nd expression, and $\Phi(x,y)$ is the Helmholtz fundamental solution, $\Phi(x,y) = \exp(\mathrm{i}k|x-y|)/(4\pi|x-y|)$ for $d = 3$. Taking Dirichlet, Neumann, or impedance traces in the above equation gives a BIE (see, e.g., [2, §2.5, 2.6]), in operator form

$$A\,\partial_n u^{\mathrm{S}} = f, \tag{6}$$

where $A$ is a linear combination of boundary integral operators (BIOs) and the identity that is a bounded linear operator on some Hilbert space $\mathcal{H}$ ($\mathcal{H} = H^{-1/2}(\Gamma)$ and $L^2(\Gamma)$ are common choices). This leads to (2) with $v = \partial_n u^{\mathrm{S}}$ and $a(v,w) := (Av,w)_{\mathcal{H}}$, $F(w) := (f,w)_{\mathcal{H}}$, where $(\cdot,\cdot)_{\mathcal{H}}$ is the inner product on $\mathcal{H}$.

[1] Note that (1) with the SRC replaced by (5) is a classic NA model problem.

## 3. Piecewise polynomial spaces $\mathcal{H}_N$ for FEM/BEM

The standard NA choice for $\mathcal{H}_N$ is a space of piecewise polynomials. We construct on the bounded domain $G$ ($G = \Omega_R$ or $\Gamma$) a *mesh* $\mathcal{M}$, a finite collection of relatively open disjoint *elements* $\tau \subset G$, such that $G = \cup_{\tau \in \mathcal{M}} \overline{\tau}$. The standard setup is that each $\tau$ is the image of a fixed *reference element* $\mathcal{R}$ under a diffeomorphism $\chi_\tau : \mathcal{R} \to \tau$ (standard choices for $\mathcal{R}$ are a unit cube or a unit simplex, e.g., [9]). We choose $p \in \mathbb{N} \cup \{0\}$, denote by $\mathbb{P}_p$ the set of polynomials of (total or coordinate) degree $\leq p$ on $\mathcal{R}$ (e.g., [9]), and define $\mathcal{H}_N$ to be the set of $w_N : G \to \mathbb{C}$ such that, for each $\tau \in \mathcal{M}$, $w_N|_\tau = P \circ \chi_\tau^{-1}$, with $P \in \mathbb{P}_p$. Without further constraint the functions in this space $\mathcal{H}_N$ are, generically, discontinuous at the boundary of each $\tau$. If needed to ensure $\mathcal{H}_N \subset \mathcal{H}$ (e.g., if $\mathcal{H} = H_0^1(\Omega_R)$) we also require that each $w_N \in C(\overline{G})$. We term the GM (3) with this $\mathcal{H}_N$ the *finite element method (FEM)* when $G = \Omega_R$, the *boundary element method (BEM)* when $G = \Gamma$.

This construction is made for each $N \in \mathbb{N}$. With the hope of achieving that the GM solution $v_N \to v$ it is standard to require that i) $h := \max \operatorname{diam}(\tau) \to 0$ as $N \to \infty$ (this termed the $h$-FEM/BEM); or ii) $p \to \infty$ ($p$-FEM/BEM); or iii) $h \to 0$ and $p \to \infty$ simultaneously ($hp$-FEM/BEM). Crucial (and this is very much an endeavour at the SCA/NA interface) are sharp bounds for the *best approximation error* $\min_{w_N \in \mathcal{H}_N} \|v - w_N\|_\mathcal{H}$ as a function of $\Omega$, $k$, $h$ and $p$. By the Whittaker-Nyquist-Shannon criterion we expect that $\dim(\mathcal{H}_N) \sim k^m$, where $m$ is the dimension of $G$ ($m = d$ if $G = \Omega_R$, $= d-1$ if $G = \Gamma$) should be necessary and sufficient to ensure $\min_{w_N \in \mathcal{H}_N} \|v - w_N\|_\mathcal{H}$ remains small as $k \to \infty$. That $G$ is lower dimensional is a significant advantage for the boundary-based formulation, but the linear system associated to (3) is dense rather than sparse as in the domain-based formulation.

## 4. NA of the Galerkin method

The major goal in the NA of a particular Galerkin method is to prove *quasi-optimality*, that, for some constant $C_{\mathrm{qo}} > 0$ independent of $N$,

$$\|v - v_N\|_\mathcal{H} \leq C_{\mathrm{qo}} \min_{w_N \in \mathcal{H}_N} \|v - w_N\|_\mathcal{H}, \tag{7}$$

at least for all sufficiently large $N$, where $v$ and $v_N$ are the solutions of (2) and (3), respectively. The standard framework where this holds is where $a(\cdot,\cdot)$ is both *continuous* and *coercive*, i.e., for constants $C_{\mathrm{cont}}$, $C_{\mathrm{coer}} > 0$,

$$|a(u,w)| \leq C_{\mathrm{cont}} \|u\|_\mathcal{H} \|w\|_\mathcal{H} \quad \text{and} \quad |a(w,w)| \geq C_{\mathrm{coer}} \|w\|_\mathcal{H}^2 \quad \forall u, w \in \mathcal{H}. \tag{8}$$

By *Céa's lemma* (an extension of Lax-Milgram), it follows from (8) that (3) has exactly one solution $v_N \in \mathcal{H}_N$ for all $N \in \mathbb{N}$ and (7) holds with $C_{\mathrm{qo}} = C_{\mathrm{cont}}/C_{\mathrm{coer}}$. One reason why the FEM for Helmholtz is "hard" from an NA perspective is that $a(\cdot,\cdot)$, given by (4), is not coercive; if $w$ vanishes on $\Gamma_R$ then $a(w,w) = \|\nabla w\|_{L^2(\Omega_R)}^2 - k^2 \|w\|_{L^2(\Omega_R)}^2$ whereas $\|w\|_\mathcal{H}^2 = \|\nabla w\|_{L^2(\Omega_R)}^2 + \|w\|_{L^2(\Omega_R)}^2$.

## 5. Case studies at the SCA/NA interface

We finish with three examples of work at this interface.

5.1. **Hybrid NA-asymptotic methods.** Consider our model problem when $\Omega_-$ is $C^\infty$ and strictly convex. Melrose and Taylor [11] through SCA methods studied the $k \to \infty$ asymptotics of $\eta^{\text{slow}}(x) := k^{-1}\partial_n u(x)/\mathrm{e}^{\mathrm{i}kx\cdot d}$, for $x \in \Gamma$, especially near shadow boundaries. Combining these results with NA, Dominguez, Graham and Smyshlyaev [5] showed, in 2D, that a $k$-dependent mesh and $\dim(\mathcal{H}_N) \sim k^{1/9}$ keeps $\|\eta^{\text{slow}} - v_N\|_{L_2(\Gamma)}$ small as $k \to \infty$, where $v_N$ is a GM solution to a BIE formulation; this is improved to $k^\varepsilon$, $\forall \varepsilon > 0$, in [6], and see the article by Ecevit.

5.2. **"Pollution" in FEM/BEM.** If $a(\cdot,\cdot)$ is only *compactly perturbed coercive* (see, e.g., [3, §2.2]), then, provided (2) is uniquely solvable, (7) holds for $N \geq N_0$, for some sufficiently large $N_0$, but how do $C_{\text{qo}}$ and $N_0$ depend on $k$? To control $\min_{w_N \in \mathcal{H}_N} \|v - w_N\|_{\mathcal{H}}$, $\dim(\mathcal{H}_N) \sim k^d$ is sufficient for $h$-FEM, but $\dim(\mathcal{H}_N) \gg k^d$ is needed for (7) with $C_{\text{qo}}$ independent of $k$, the so-called "pollution effect" [1]. For $h$-BEM there is no pollution if $\Omega$ is $C^\infty$ and non-trapping [8]. Similarly, (7) holds for $hp$-FEM/BEM with $C_{\text{qo}}$ independent of $k$ provided $p \sim \log k$; see [10, 7] and the references therein, and the articles by Lafontaine and Melenk.

5.3. **$k$-dependence of BIOs.** A great SCA/NA question is how do the condition numbers $\text{cond}(A) := \|A\|\|A^{-1}\|$ of the BIOs $A$ arising in (6) depend on $k$ (and $\Omega$), and how does this translate to discretisations of $A$? A recent review is [4, §6.5].

## References

[1] I. M. Babuška, S. A. Sauter, *Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wavenumbers?* SIAM J. Numer. Anal. **34** (1997), 2392–2423.

[2] S. N. Chandler-Wilde, I. G. Graham, S. Langdon, E. A. Spence, *Numerical-asymptotic boundary integral methods in high-frequency acoustic scattering*, Acta Numerica **21** (2012), 89–305.

[3] S. N. Chandler-Wilde, D. P. Hewett, A. Moiola, J. Besson, *Boundary element methods for acoustic scattering by fractal screens*, Numer. Math. **147** (2021), 785–837.

[4] S. N. Chandler-Wilde, E. A. Spence, A. Gibbs, V. P. Smyshlyaev, *High-frequency bounds for the Helmholtz equation under parabolic trapping and applications in numerical analysis*, SIAM J. Math. Anal. **52** (2020), 845–893.

[5] V. Dominguez, I. G. Graham, V. P. Smyshlyaev, *A hybrid numerical-asymptotic boundary integral method for high-frequency acoustic scattering*, Numer. Math. **106** (2007), 471–510.

[6] F. Ecevit, H. C. Özen, *Frequency-adapted Galerkin boundary element methods for convex scattering problems*, Numer. Math. **135** (2017), 27–71.

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

[8] J. Galkowski, E. A. Spence, *Does the Helmholtz boundary element method suffer from the pollution effect?* SIAM Review, to appear, preprint at arXiv:2201.09721

[9] I. G. Graham, W. McLean, *Anisotropic mesh refinement: the conditioning of Galerkin boundary element matrices and simple preconditioners*, SIAM J. Numer. Anal. **44** (2006), 1487–1513.

[10] M. Bernkopf, T. Chaumont-Frelet, J. M. Melenk, *Wavenumber-explicit stability and convergence analysis of hp finite element discretizations of Helmholtz problems in piecewise smooth media* (2022), preprint at arXiv:2209.03601

[11] R. B. Melrose and M. E. Taylor, *Near peak scattering and the corrected Kirchhoff approximation for a convex obstacle*, Adv. Math. **55** (1985), 242–315.