

AN EFFICIENT FREQUENCY-INDEPENDENT NUMERICAL METHOD FOR COMPUTING THE FAR-FIELD PATTERN INDUCED BY POLYGONAL OBSTACLES

A. GIBBS* AND S. LANGDON†

Abstract. For problems of time-harmonic scattering by rational polygonal obstacles, embedding formulae express the far-field pattern induced by any incident plane wave in terms of the far-field patterns for a relatively small (frequency-independent) set of canonical incident angles. Although these remarkable formulae are exact in theory, here we demonstrate that: (i) they are highly sensitive to numerical errors in practice, and (ii) direct calculation of the coefficients in these formulae may be impossible for particular sets of canonical incident angles, even in exact arithmetic. Only by overcoming these practical issues can embedding formulae provide a highly efficient approach to computing the far-field pattern induced by a large number of incident angles.

Here we address challenges (i) and (ii), supporting our theory with numerical experiments. Challenge (i) is solved using techniques from computational complex analysis: we reformulate the embedding formula as a complex contour integral and prove that this is much less sensitive to numerical errors. In practice, this contour integral can be efficiently evaluated by residue calculus. Challenge (ii) is addressed using techniques from numerical linear algebra: we oversample, considering more canonical incident angles than are necessary, thus expanding the set of valid coefficient vectors. The coefficient vector can then be selected using either a least squares approach or column subset selection.

Key words. Embedding formula, Far-field pattern, Scattering, Cauchy integral, Oversampling

MSC codes. 35J05, 78A45, 30E20, 65F20

1. Introduction. In problems of two-dimensional time-harmonic scattering of acoustic, electromagnetic or elastic waves, obtaining a full characterisation of the scattering properties of an obstacle may require a representation of the far-field behaviour induced by a large set, possibly thousands, of incident plane waves [15]. In this paper we propose a new method for calculating this representation efficiently across all frequencies for a broad class of polygonal scatterers.

First, we state the scattering problem, which must be solved in order to compute the far-field pattern. We denote each incident plane wave by $u^i(\mathbf{x}; \alpha) := e^{-ik(x_1 \cos \alpha + x_2 \sin \alpha)}$, where $\mathbf{x} := (x_1, x_2) \in \mathbb{R}^2$, the wavenumber $k > 0$, and the incident angle $\alpha \in [0, 2\pi)$. We consider the scattered wave field $u^s(\cdot; \alpha)$ induced by $u^i(\cdot; \alpha)$ and a sound-soft *rational* polygon $\Omega \subset \mathbb{R}^2$, with boundary $\partial\Omega$. Rational polygons have exterior angles that are rational multiples of π (see also Definition 1.1). The scattered field $u^s(\cdot; \alpha)$ satisfies the Dirichlet Helmholtz problem

$$(1.1) \quad (\Delta + k^2)u^s = 0, \quad \text{in } \mathbb{R}^2 \setminus \Omega;$$

$$(1.2) \quad u^s = -u^i, \quad \text{on } \partial\Omega;$$

$$(1.3) \quad \frac{\partial u^s(\mathbf{x}; \alpha)}{\partial r} - ik u^s(\mathbf{x}; \alpha) = o(r^{-1/2}), \quad r := |\mathbf{x}| \rightarrow \infty.$$

The *far-field pattern* (also called the *far-field diffraction coefficient* (e.g., [10]), *far-field directivity* (e.g., [12]), or simply *far-field coefficient* (e.g., [3])) is of practical interest and is central to this paper. Intuitively, it describes the distribution of energy of the scattered field, far away from the scatterer. We denote the far-field pattern at observation angle θ (where $\mathbf{x} = r(\cos \theta, \sin \theta)$) by

*University College London (andrew.gibbs@ucl.ac.uk).

†Brunel University London (stephen.langdon@brunel.ac.uk).

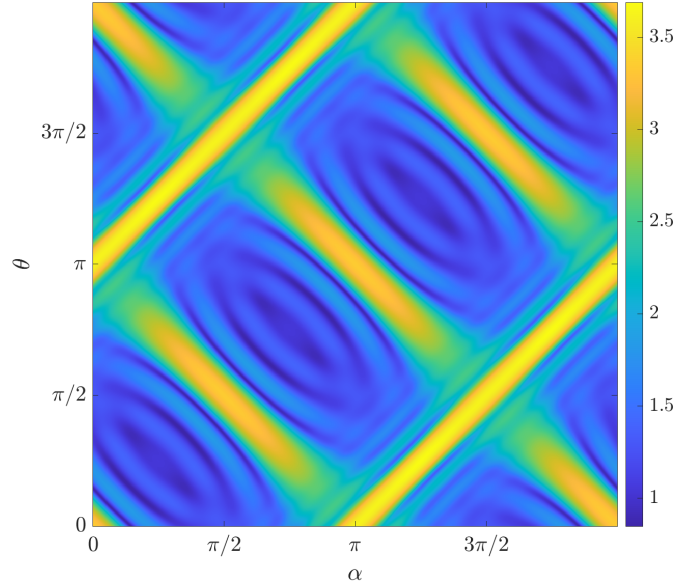


FIGURE 1.1. Full far-field characterisation $\log|D(\theta, \alpha)|$ for $(\theta, \alpha) \in \mathbb{T}$, where Ω is a square of diameter 2 and $k = 10$. This plot was computed using the approach described in this paper; see §5 for details.

$D(\theta, \alpha)$, defined by the asymptotic relationship (see, e.g., [10, Theorem 2.6])

$$(1.4) \quad u^s(\mathbf{x}; \alpha) = \frac{e^{i(kr + \pi/4)}}{\sqrt{2\pi kr}} \left(D(\theta, \alpha) + \mathcal{O}(r^{-1}) \right), \quad r \rightarrow \infty.$$

We will consider how $D(\theta, \alpha)$ depends on both the observation angle θ and the incident angle α , each considered as elements of the 2π -periodic set $\mathbb{S} := [0, 2\pi)$, as shown in Figure 1.1; we will often write $(\theta, \alpha) \in \mathbb{T}$, where $\mathbb{T} := \mathbb{S}^2$.

Numerous applications require an understanding of how $D(\theta, \alpha)$ varies over the full range $(\theta, \alpha) \in \mathbb{T}$. For example, in atmospheric physics, when modelling scattering of the sun's radiation by ice crystals in cirrus clouds, to simplify calculations it is common to consider the *orientation average*, where the far-field behaviour is averaged over $\alpha \in \mathbb{S}$. Details of the averaging technique and applications can be found in [29]. Another derived quantity of interest is the *monostatic cross section* $4\pi|D(\alpha, \alpha)|^2$ (sometimes referred to as *backscatter*), where the observer and the source are in the same direction. This appears frequently in underwater acoustics and sonar modelling, where it is common for both the signal transmitter and receiver to be positioned on the underside of a boat. A review of relevant applications is given in [16]. This paper considers $\Omega \subset \mathbb{R}^2$, which can provide an approximation for three-dimensional scattering problems on cylindrical obstacles [10, §3.4]. The potential extension to general three-dimensional obstacles is discussed in §6.

There are many numerical algorithms for solving (1.1)–(1.3) for fixed α , thereby producing an approximation $D_N(\cdot, \alpha)$ to $D(\cdot, \alpha)$. However, obtaining an approximation for a different incident angle $\alpha' \neq \alpha$ requires the prescription of new boundary data, and hence repetition of some or all of the numerical algorithm; this typically requires a much larger computational cost than varying θ . This paper is about efficient numerical approximation of $D(\theta, \alpha)$ over the whole range $(\theta, \alpha) \in \mathbb{T}$.

A popular approach for approximating $D(\theta, \alpha)$ for $(\theta, \alpha) \in \mathbb{T}$ is the *T-matrix* method (see, e.g., [28]). Once the *T-matrix* has been constructed, computing the far-field pattern for any given $\alpha \in \mathbb{S}$ requires multiplication by a single vector whose entries can be computed with an analytic formula. A drawback though of *T-matrix* methods is that numerically stable construction of the $O(k)$ -dimensional *T-matrix* requires, as an input, the numerical solution of $O(k)$ scattering problems [14, 15]. Hence, when k is large, *T-matrix* methods may be computationally prohibitive.

As with the stable *T-matrix* approach considered in [14], an input to our method is the numerical solution of a number of scattering problems. A key advantage of our method, when compared against *T-matrix* approaches, is that this number depends only on the geometry of Ω , and, crucially, is independent of the wavenumber k in the following sense: if the canonical scattering problems are solved by a numerical method with error \mathcal{E}_{in} , then our method will efficiently determine the far-field pattern induced by any incident angle with an error \mathcal{E}_{out} , where $\mathcal{E}_{\text{out}}/\mathcal{E}_{\text{in}}$ is bounded independently of k . In particular, this means that if the PDE (1.1)-(1.3) is solved by a numerical method for which any prescribed level of accuracy can be achieved with a number of degrees of freedom and computational cost that is independent of k , as is the case for HNABEM as described in §4.2, then our algorithm will share this key property of the underlying scheme, i.e. that the number of degrees of freedom and computational cost will be independent of k .

1.1. The embedding formula for rational polygons. Our method is based on an adaptation of the *embedding formula* derived in [3] (related ideas were first presented in [12]). First, we clarify our geometrical constraints.

DEFINITION 1.1 (Rational polygons). *An angle $\omega \in \mathbb{S}$ is rational if it can be expressed as π multiplied by a rational number. An S -sided polygon Ω is rational if its external angles $\{\omega_j\}_{j=1}^S$ are all rational angles. For a rational polygon Ω , we denote by p the smallest positive integer such that π/p divides ω_j exactly for $j = 1, \dots, S$, whilst $\{q_j\}_{j=1}^S$ denotes the set of integers such that $q_j\pi/p = \omega_j$, for $j = 1, \dots, S$.*

Some examples of rational polygons include: a square with $q_1 = \dots = q_4 = 3$ and $p = 2$; a right-angled isosceles triangle with $q_1 = q_2 = 7$, $q_3 = 6$ and $p = 4$; a screen $\Omega := [0, 1] \times \{0\}$ with $q_1 = q_2 = 2$ and $p = 1$. We note that Definition 1.1 does not require convexity, and this is the case for all of our theoretical results. For simplicity, and due to availability of high-frequency solvers, all of our examples are on convex polygons or screens.

As in [3], in the remainder of the paper, the formulae have been simplified by assuming that one edge of Ω is aligned with the horizontal axis. We now state the critical result of [3].

THEOREM 1.2. *Suppose that Ω is a rational polygon and that there exist distinct ‘canonical incident angles’ $\alpha_1, \dots, \alpha_M$, satisfying Assumption 1.3 (given below), where $M := \sum_{j=1}^S (q_j - 1)$ and q_j is as in Definition 1.1. Then there exist ‘embedding coefficients’ $b_m(\alpha)$ such that*

$$(1.5) \quad D(\theta, \alpha) = \frac{\sum_{m=1}^M b_m(\alpha) \Lambda(\theta, \alpha_m) D(\theta, \alpha_m)}{\Lambda(\theta, \alpha)}, \quad (\theta, \alpha) \in \mathbb{T},$$

where

$$(1.6) \quad \Lambda(\theta, \alpha) := \cos(p\theta) - (-1)^p \cos(p\alpha),$$

and p is as in Definition 1.1. If $\Lambda(\theta, \alpha) = 0$ then one or two applications of L’Hôpital’s rule may be used to express the right-hand side of (1.5) in terms of derivatives with respect to θ .

The beauty of Theorem 1.2 is that, given far-field patterns $D(\theta, \alpha_m)$ for distinct $\alpha_1, \dots, \alpha_M$, the embedding formula (1.5) provides an exact expression for $D(\theta, \alpha)$, valid for all $(\theta, \alpha) \in \mathbb{T}$. Referring to the example geometries above: $M = 8$ for the square, $M = 17$ for the right-angled isosceles triangle, and $M = 2$ for the screen.

Despite the remarkable implications of Theorem 1.2, to the best knowledge of the authors, embedding formulae have not found significant use in computational scattering applications. This may be surprising, as the formulae appear to provide a highly efficient and k -independent means for computing $D(\theta, \alpha)$ for all $(\theta, \alpha) \in \mathbb{T}$. However, although (1.5) is exact in principle, we will now see that these formulae are incredibly sensitive to numerical errors in the canonical far fields $D(\theta, \alpha_m)$.

We define the numerical analogue of (1.5)

$$(1.7) \quad \mathfrak{D}_N(\theta, \alpha) := \frac{\sum_{m=1}^M b_m(\alpha) \Lambda(\theta, \alpha_m) D_N(\theta, \alpha_m)}{\Lambda(\theta, \alpha)},$$

where $D_N(\cdot, \alpha)$ is some numerical approximation to the far-field pattern $D(\cdot, \alpha)$ for $\alpha \in \mathbb{S}$, such that $D_N(\cdot, \alpha) \rightarrow D(\cdot, \alpha)$ pointwise as $N \rightarrow \infty$. We reserve discussion about the approximation of the coefficients b_m until §3, for now we assume that they exist and are known exactly.

Algorithmically, (1.7) requires solution of M canonical problems corresponding to incident angles $\alpha_1, \dots, \alpha_M$, after which evaluation for any $(\theta, \alpha) \in \mathbb{T}$ is straightforward. However, for any given α , if we consider $\theta \approx \theta_0$, where $\Lambda(\theta_0, \alpha) = 0$, it is not hard to see that we immediately run into problems with the representation (1.7). For the exact theoretical formula (1.5), when $\theta = \theta_0$ the theorem states that the value of $D(\theta, \alpha)|_{\theta=\theta_0}$ is determined by the rate at which both the denominator and numerator go to zero as $\theta \rightarrow \theta_0$. In the approximate case (1.7), there are no guarantees that the numerator tends to zero as $\theta \rightarrow \theta_0$; a zero of the numerical approximation is likely close to θ_0 , but not at θ_0 . Instead, we expect that the numerator will tend to something small at θ_0 , approximately zero, but not zero. This numerical artifact is referred to as a *pole-zero pair*, and will lead to arbitrarily large errors in $\mathfrak{D}_N(\theta, \alpha)$, because this ‘small’ number is multiplied by unbounded values of $1/\Lambda(\theta, \alpha)$. In this sense (1.7) is numerically ill-conditioned; this effect can be seen in Figure 1.2. These qualitative statements are made precise later by Lemma 2.6. To address these pole-zero pairs, we will reformulate (1.5) and (1.7) as complex contour integrals. This technique is natural when dealing with removable singularities, see [2, §2.1] and the references [28], [51], and [76] therein.

We summarise this in the first fundamental challenge of this paper:

- (C1) The naive embedding approximation (1.7) is highly sensitive to numerical errors in the canonical far-field patterns.

Henceforth, to simplify the presentation we write

$$(1.8) \quad \hat{D}(\theta, \alpha) := \Lambda(\theta, \alpha) D(\theta, \alpha), \quad \hat{D}_N(\theta, \alpha) := \Lambda(\theta, \alpha) D_N(\theta, \alpha).$$

We now focus on calculating the coefficients $\mathbf{b} := [b_1, \dots, b_M]^T$, which, as we will see, poses a second practical challenge. Using (1.5) and the reciprocity principle (see, e.g., [10, Theorem 3.15]), which implies that $\hat{D}(\theta, \alpha) = (-1)^{p+1} \hat{D}(\alpha, \theta)$ for all $(\theta, \alpha) \in \mathbb{T}$, we see that the coefficients \mathbf{b} satisfy

$$(1.9) \quad A \mathbf{b} = \mathbf{d}, \quad \text{where } A := [\hat{D}(\alpha_m, \alpha_{m'})]_{m, m'=1}^M, \quad \mathbf{d} = (-1)^{p+1} [\hat{D}(\alpha, \alpha_m)]_{m=1}^M.$$

In Theorem 1.2 and [3] the following is assumed:

ASSUMPTION 1.3. *For any distinct canonical incident angles $\alpha_1, \dots, \alpha_M$, the system (1.9) has a unique solution.*

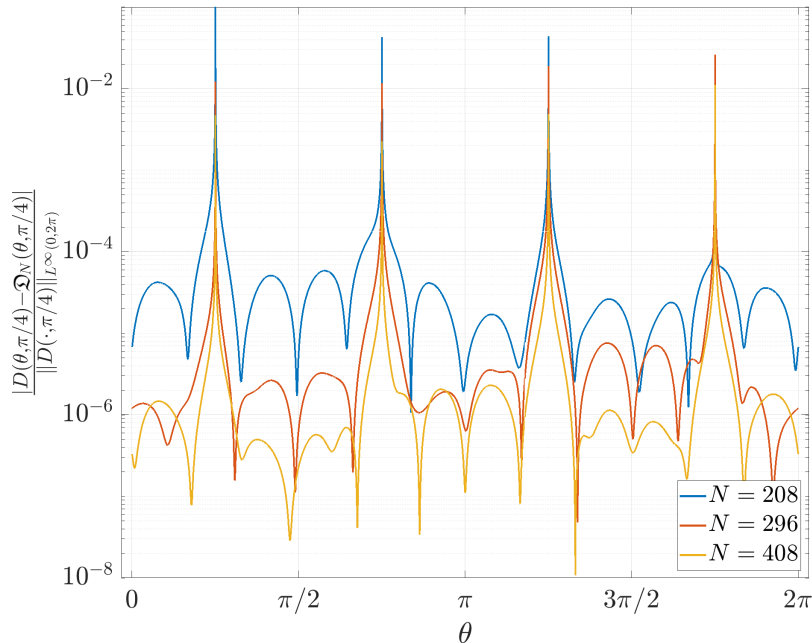


FIGURE 1.2. Example of unbounded errors which occur when applying the naive embedding approximation (1.7) directly. Here Ω is a square of diameter 2 and $k = 10$. The numerical solver used is described in §4.1.

Under this assumption, the coefficients \mathbf{b} can be derived from readily available quantities, namely the canonical far-field patterns $D(\theta, \alpha_m)$, for $m = 1, \dots, M$. We remark that we have conducted tens of thousands of numerical experiments on varying geometries and wavenumbers, and in each case we have found a set of canonical incident angles $\alpha_1, \dots, \alpha_M$ satisfying Assumption 1.3. However, in many of these experiments we have also found distinct canonical incident angles under which Assumption 1.3 is not satisfied, hence the coefficients \mathbf{b} cannot be found. For example, for the screen problem ($M = 2$), if we choose α_1 and α_2 such that $\cos \alpha_1 = -\cos \alpha_2$ then $\Lambda(\alpha_1, \alpha_2) = \Lambda(\alpha_2, \alpha_1) = 0$, thus A is the zero matrix and (1.9) cannot be solved. For a second example, consider the equilateral triangle ($M = 12$) with equispaced canonical incident angles $\alpha_m = a + 2(m-1)\pi/M$ for some $a \in \mathbb{S}$. Numerical approximation to A suggests that the condition number blows up as $a \rightarrow 0$; see Figure 1.3.

Even if Assumption 1.3 holds, A may have a large condition number. In this case, the coefficient vector \mathbf{b} may have large entries, amplifying numerical errors in (1.7). These issues are summarised in the second fundamental challenge below:

- (C2) In general, it is unclear how to choose the canonical incident angles so that Assumption 1.3 holds and hence how to determine the coefficients \mathbf{b} . Even if Assumption 1.3 holds, it is still unclear how to ensure that the coefficients \mathbf{b} have a small norm.

1.2. Aims and outline of this paper. In this paper, we address the fundamental challenges (C1) and (C2), reformulating the embedding formula (1.7) for rational polygons so that it can be of practical use with approximate far-field patterns.

In §2, we address (C1). We reformulate the embedding formula (1.7) as a complex contour

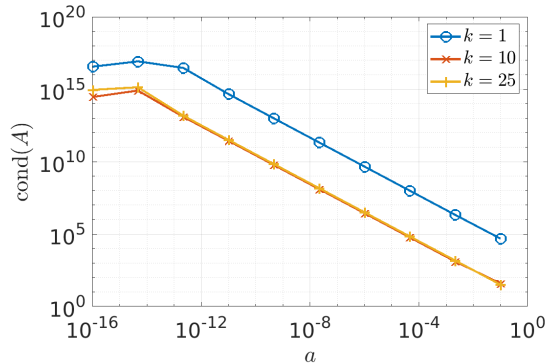


FIGURE 1.3. Blow-up of (approximations to) the condition number of A as $a \rightarrow 0$, for the equilateral triangle with $\alpha_m = a + (m-1)\pi/6$, $m = 1, \dots, 12$. The approximation was computed using the method described in §4.1, with $N = 375$.

integral to reduce the sensitivity to numerical errors. This is quantified by Theorem 2.3. This contour integral may be evaluated by residue calculus, except in a very small set of cases where rounding errors can affect the result. In this case, we interpolate the k -dependent part of the integrand by a quadratic polynomial so that the cost of evaluating the integral is independent of k . This interpolation is done in such a way that the value of the integral is unchanged. For simplicity, in §2 we assume that the embedding coefficients \mathbf{b} exist and are exact.

In §3, we address (C2), and consider the implications of numerical approximation of the embedding coefficients \mathbf{b} . We cannot solve this challenge theoretically, so instead, we oversample, taking more than M canonical far-field patterns, under the assumption that this gives us a broader space of coefficient vectors to choose from. We consider two strategies for selecting a coefficient vector from this enhanced space. The first strategy solves the redundant linear system via a regularising truncated singular value decomposition. This approach is partially supported by theory, which informs the choice of the truncation parameter. The second approach chooses a subset of M incident angles which are in some sense optimal, and discards the remaining redundant incident angles. This approach has less theoretical justification, but appears to be more efficient and accurate in practice.

In §5, we present numerical results. These demonstrate the effectiveness of our method via numerical examples, add empirical justification to the approach of §3 and demonstrate the frequency-independence of our method at high frequencies.

2. Reformulating the embedding formula. This section addresses (C1). The basic idea of our approach is as follows: It is well-known that the far-field pattern $D(\theta, \alpha)$ is an entire function with respect to observation angle θ (see, e.g., [10, §2.2]; note that by reciprocity [10, Theorem 3.17] it is also entire in α , but we do not require this here). It is then clear from (1.6) and (1.8) that any finite sum of $\hat{D}(\theta, \alpha)$ is entire in θ . Hence, we can complexify θ , and, using Cauchy's integral formula, express (1.5) as a complex contour integral

$$(2.1) \quad D(\theta, \alpha) = \frac{1}{2\pi i} \oint_{\gamma} \frac{\sum_{m=1}^M b_m(\alpha) \hat{D}(z, \alpha_m)}{\Lambda(z, \alpha)(z - \theta)} dz,$$

where γ is any closed contour containing θ , oriented anti-clockwise in \mathbb{C} . This is the only constraint on γ , recalling that the numerator of the integrand is entire and the singularities of $1/\Lambda(z, \alpha)$ are all removable (Theorem 1.2). The advantage of (2.1), compared to (1.5), is that we can choose γ in such a way that the magnitude of the denominator in the integrand remains bounded below, away from zero, for $z \in \gamma$. This is in contrast to (1.5), where the magnitude of the denominator has no lower bound, and small errors in the numerator lead to arbitrarily large errors overall.

We will show that the approximation

$$(2.2) \quad \mathcal{D}_N(\theta, \alpha; \gamma) = \frac{1}{2\pi i} \oint_{\gamma} \frac{\sum_{m=1}^M b_m(\alpha) \hat{D}_N(z, \alpha_m)}{\Lambda(z, \alpha)(z - \theta)} dz,$$

is well-conditioned in terms of numerical errors in $\hat{D}_N(z, \alpha_m)$, because we can always choose γ to be a safe distance from the poles where $\Lambda(z, \alpha) = 0$, and a safe distance from the pole at θ , so that the denominator does not get too large; hence the numerical errors are not significantly amplified. We now calculate the locations of these poles.

LEMMA 2.1. *Given $\alpha \in [0, 2\pi)$, the set of poles of $\Lambda(\cdot, \alpha)$ is given by*

$$\Theta_{\alpha} := \{\theta \in \mathbb{C} : \Lambda(\theta, \alpha) = 0\} = \begin{cases} \{\pm\alpha + (2n+1)\pi/p : n \in \mathbb{Z}\}, & p \text{ odd,} \\ \{\pm\alpha + 2n\pi/p : n \in \mathbb{Z}\}, & p \text{ even.} \end{cases}$$

Proof. Noting that

$$\Lambda(\theta, \alpha) = \begin{cases} 2 \cos\left(\frac{p(\theta+\alpha)}{2}\right) \cos\left(\frac{p(\theta-\alpha)}{2}\right), & p \text{ odd,} \\ -2 \sin\left(\frac{p(\theta+\alpha)}{2}\right) \sin\left(\frac{p(\theta-\alpha)}{2}\right), & p \text{ even,} \end{cases}$$

the location of the poles immediately follows. □

Note that the elements of Θ_{α} are all real.

DEFINITION 2.2 (Nearby poles). *Given θ , we define*

$$\theta_0 := \arg \min_{\tilde{\theta}_0 \in \Theta_{\alpha}} |\theta - \tilde{\theta}_0|_{2\pi},$$

the closest pole to θ , where $|\theta - \theta'|_{2\pi} := \min_{n \in \mathbb{Z}} \{|\theta - \theta' + 2n\pi|\}$. Similarly,

$$\theta'_0 := \begin{cases} \arg \min_{\tilde{\theta}'_0 \in \Theta_{\alpha} \setminus \{\theta_0\}} |\theta - \tilde{\theta}'_0|_{2\pi}, & \theta_0 \notin \{n\pi/p : n \in \mathbb{Z}\} \\ \theta_0, & \theta_0 \in \{n\pi/p : n \in \mathbb{Z}\} \end{cases}$$

is the closest pole to θ_0 in the case where $\theta'_0 \neq \theta_0$, corresponding to the poles being order one, and $\theta'_0 = \theta_0$ corresponding to a pole of order two.

Intuitively, it makes sense to choose γ such that it encloses θ , remaining as far as possible from the elements of Θ_{α} in order to avoid blow-up of the integrand. Doing so may require us to enclose θ_0 and possibly θ'_0 inside γ , if these poles are close to θ . This is the idea behind the following theorem.

THEOREM 2.3. *Suppose the approximations $D_N(z, \alpha_m)$, for $m = 1, \dots, M$, satisfy*

$$|D(z, \alpha_m) - D_N(z, \alpha_m)| \leq \epsilon_m, \quad \text{for } (z, \alpha_m) \in \Psi \times \mathbb{S}$$

where $\Psi := \{|\Im z| < \ln(3 + \pi^2/64)/p\}$ and ϵ_m , $m = 1, \dots, M$, are N -dependent constants. Then there exists a closed rectangular complex contour γ enclosing θ , θ_0 , and θ'_0 , such that

$$|D(\theta, \alpha) - \mathcal{D}_N(\theta, \alpha; \gamma)| \leq C \|\epsilon\|_2,$$

where $\|\epsilon\|_2 := \sqrt{\sum_{m=1}^M |\epsilon_m|^2}$ and

$$(2.3) \quad C := \frac{128 (5\pi + 4 \ln(3 + \pi^2/64)) (6 + \pi^2/64)}{\pi^4} \|\mathbf{b}\|_2.$$

This theorem is proved in §2.2, where details of the contour γ can also be found.

Theorem 2.3 should be interpreted in the following way: if the error in the canonical far-field approximations is bounded, then the error in the contour integral (2.2) is bounded. Therefore, the reformulated embedding formula (2.2) succeeds where our naive representation (1.7) failed, and we have addressed (C1). It is clear from (2.3) that the error will also depend on the size and accuracy of the coefficients \mathbf{b} - this is addressed in §3.

The sceptical reader may point out that: (i) we have analytically extended our approximations $D_N(\cdot, \alpha_m)$, a process which is known to be ill-conditioned [34], and (ii) quadrature evaluation of (2.2) may carry a frequency-dependent cost. Fortunately, if we further assume that our numerical approximation $D_N(\theta, \alpha)$ is analytic for θ in a complex neighbourhood of \mathbb{S} (justified below in Remark 2.4), we can choose γ to be a closed contour in this neighbourhood containing $\{\theta, \theta_0, \theta'_0\}$ and evaluate the integral (2.2) by residue calculus:

$$(2.4) \quad \mathcal{D}_N(\theta, \alpha, \gamma) = \frac{\sum_{m=1}^M b_m(\alpha) \hat{D}_N(\theta, \alpha_m)}{\Lambda(\theta, \alpha)} - \begin{cases} \sum_{\chi \in \{\theta_0, \theta'_0\}} \frac{\sum_{m=1}^M b_m(\alpha) \hat{D}_N(\chi, \alpha_m)}{p(\chi - \theta) \sin(p\chi)}, & \theta_0 \neq \theta'_0, \\ 2 \frac{\sum_{m=1}^M b_m(\alpha) \left[(\theta_0 - \theta) \frac{\partial \hat{D}_N(z, \alpha)}{\partial z} \Big|_{z=\theta_0} - \hat{D}_N(\theta_0, \alpha_m) \right]}{p^2(\theta_0 - \theta)^2 \cos(p\theta_0)}, & \theta_0 = \theta'_0, \end{cases} \quad \theta \notin \Theta_\alpha.$$

The second case on the right-hand side corresponds to the double pole. As in Theorem 1.2, at the points $\theta \in \Theta_\alpha$ we can compute \mathcal{D}_N using L'Hôpital's rule.

By representing the integral as (2.4), we address the two concerns above: (i) because all poles are in $\Theta_\alpha \cup \{\theta\}$, no analytic continuation is required, and (ii) we have evaluated the integral without quadrature.

It is instructive to notice that the first term on the right-hand side of (2.4) is precisely (1.7). Therefore, the second term on the right-hand side of (2.4) may be interpreted as a correction to (1.7). Conversely, in the exact case (where D_N is replaced by D), the formula (2.4) is exact, because the residues are zero. This is because the points at Θ_α are removable singularities in theory, manifesting as pole-zero pairs in practice.

Remark 2.4 (Analyticity of numerical solutions). The assumption required by (2.4), that $D_N(\theta, \alpha)$ is analytic for θ in a neighbourhood of \mathbb{S} , is entirely reasonable. In finite difference / element / volume and boundary element methods for solving (1.1)-(1.3), the far-field pattern (1.4) is approximated by integrating some (typically piecewise-analytic) data against an entire kernel (see e.g. [30] and [10, §3.5]). Therefore, the resulting approximation is entire, and the estimate of Theorem 2.3 also applies to the formula (2.4).

2.1. Evaluation of residues in finite precision arithmetic. Until now, we have not discussed the implications of rounding errors, implicitly assuming that all calculations are done in exact arithmetic. When two poles in (2.4) coalesce, the corresponding residues will grow, and there will be a large amount of numerical cancellation. In finite precision arithmetic, small rounding errors will be amplified; this is commonly known as *catastrophic cancellation*. We first remark that the region where this occurs is much smaller than the region where (1.7) breaks down (see the discussion around h and H in §4), so even if nothing is done to address this issue, (2.4) still offers a significant improvement over (1.7) in practice. Secondly, we remark that variable precision arithmetic (VPA) may be used to address catastrophic cancellation, whereas VPA would not fix the breakdown of (1.7). Here, we present a fix for this issue, which may be used without VPA.

Suppose we were to evaluate the integral representation (2.2) by numerical quadrature. This offers the advantage that we could choose γ such that the denominator remains bounded below, thus the samples at the quadrature nodes remain bounded, avoiding catastrophic cancellation between the quadrature samples (see [27] for a summary of relevant numerical integration techniques). However, we recall that a potential disadvantage of the quadrature approach (which motivated (2.4)) is that the numerator is an oscillatory function, and as $k \rightarrow \infty$ this may grow exponentially and/or oscillate increasingly rapidly along certain segments of γ . Here lies the dilemma: the contour integral approach can avoid catastrophic cancellation, and the residue approach avoids an $O(k)$ factor increase in computational cost due to quadrature. Can we avoid both?

Surprisingly, thanks to an idea we believe to be new, the answer is *yes*. The idea rests on the following key observation. For large k the *integrand* of (2.2) is highly oscillatory (on the real line), but the integral still only depends on the integrand's values at three points: θ , θ_0 and θ'_0 . Therefore, we can interpolate the k -dependent oscillatory numerator by a quadratic polynomial ρ_2 , constructed such that $\rho_2(\xi) = \sum_{m=1}^M b_m(\alpha) \hat{D}_N(\xi, \alpha_m)$ for $\xi \in \{\theta, \theta_0, \theta'_0\}$. In the case where $\theta = \theta_0$ or $\theta_0 = \theta'_0$ we add the additional constraint that $\rho'_2(\theta_0) = \sum_{m=1}^M b_m(\alpha) \frac{\partial \hat{D}_N}{\partial \theta}(\theta, \alpha_m)|_{\theta=\theta_0}$, since residues at double poles depend on the derivative of the numerator. Obtaining the (approximate) derivative of the far-field pattern is easy in practice, for reasons similar to those given in Remark 2.4.

Barycentric interpolation may be used for efficiency (see e.g. [25, 31]). However, it is not essential for accurate results, because γ can be chosen so that ρ_2 is only evaluated at a bounded distance from the interpolation points. It follows that

$$(2.5) \quad \mathcal{D}_N(\theta, \alpha; \gamma) = \frac{1}{2\pi i} \oint_{\gamma} \frac{\rho_2(z)}{\Lambda(z, \alpha)(z - \theta)} dz,$$

since the residues of the integral depend only on the value of the integrand (and possibly its derivative) at the poles $\xi = \theta, \theta_0, \theta'_0$. At these poles, the integrands and, where appropriate, the derivatives of the integrand in (2.2) and (2.5) are equal, and thus by the residue theorem (2.2) and (2.5) have the same value. However, (2.5) is independent of k , so accurate evaluation by quadrature requires an $O(1)$ cost, instead of $O(k)$, as $k \rightarrow \infty$.

2.2. Proof of Theorem 2.3. Before we can prove Theorem 2.3, we require some preliminary results. First, we note that for certain values of α , pairs of poles in Θ_α coalesce, forming a pole of order two.

DEFINITION 2.5 (Coalescence points).

We define the set of ‘coalescence points’ by:

$$(2.6) \quad \Theta_* := \left\{ \theta \in \mathbb{S} : \frac{\partial \Lambda}{\partial \theta}(\theta, \alpha) = 0 \right\} = \{ \theta \in \mathbb{S} : \theta = n\pi/p, n \in \mathbb{Z} \}.$$

With θ_0 and θ'_0 defined as in Definition 2.2, we define the ‘nearest coalescence point’ as

$$\theta_* = \arg \min_{\tilde{\theta}_* \in \Theta_*} |\theta_0 - \tilde{\theta}_*|_{2\pi}.$$

When θ_0 and θ'_0 are close to θ_* , the singularity will be stronger. Therefore these points play an important role when quantifying the breakdown of the naive embedding formula (1.7). The following result provides a lower bound on $\Lambda(\theta, \alpha)$, and will be useful when choosing the contour γ to avoid amplification of errors in the integral representation (2.2).

LEMMA 2.6. For Λ as in (1.6), θ_0 as in Definition 2.2 and θ_* as in Definition 2.5,

$$(2.7) \quad \frac{p^2}{8} |\theta - \theta_0| |\theta - \theta_*| \leq |\Lambda(\theta, \alpha)|, \quad (\theta, \alpha) \in \mathbb{T}.$$

Proof. Firstly, by Definition 2.2, we have $\Lambda(\theta, \alpha) = \Lambda(\theta, \alpha) - \Lambda(\theta_0, \alpha) = \cos(p\theta) - \cos(p\theta_0)$ and from standard trigonometric identities it follows that

$$(2.8) \quad |\Lambda(\theta, \alpha)| = 2 |\sin(p(\theta - \theta_0)/2)| \cdot |\sin(p(\theta_0 + \theta)/2)|.$$

We first focus on the lower bound of (2.7), for which we will require the inequality

$$(2.9) \quad |\sin(p(\theta_* + \theta)/2)| \leq |\sin(p(\theta_0 + \theta)/2)|.$$

To see why (2.9) holds, we note that the points $\theta_0 \in \Theta_\alpha$ are distributed symmetrically about the points $\theta_* \in \Theta_*$, and we have specified the condition that θ_0 must be the element of Θ_α closest to θ ; hence it follows that θ and θ_0 both lie on the same side of θ_* (in a local sense). Hence, it is clear that $p(\theta + \theta_0)/2$ is farther from $p\theta_*$ than $p(\theta + \theta_*)/2$, and this distance is no more than $\pi/(2p)$ by (2.6).

Combining (2.8) with (2.9) yields

$$\begin{aligned} |\Lambda(\theta, \alpha)| &\geq 2 |\sin(p(\theta - \theta_0)/2)| \cdot |\sin(p(\theta_* + \theta)/2)| \\ &= 2 |\sin(p(\theta - \theta_0)/2)| \cdot |\sin(p(2\theta_* + (\theta - \theta_*))/2)| \\ &= 2 |\sin(p(\theta - \theta_0)/2)| \cdot |\sin(p\theta_*) \cos(p(\theta - \theta_*)/2) + \sin(p(\theta - \theta_*)/2) \cos(p\theta_*)|. \end{aligned}$$

By the definition (2.6) of Θ_* , we have that $\sin(p\theta_*) = 0$ and $|\cos(p\theta_*)| = 1$, so

$$(2.10) \quad |\Lambda(\theta, \alpha)| \geq 2 |\sin(p(\theta - \theta_0)/2)| \cdot |\sin(p(\theta - \theta_*)/2)|.$$

From Definition 2.2 and (2.6), it follows that the farthest θ can be from the nearest θ_0 or θ_* is $\pi/(2p)$. Hence the argument of both sines of (2.10) is at most $\pi/4$, so we may use the identity $|\sin(x)| \geq |x/2|$ for $0 \leq |x| \leq \pi/4$ to obtain the lower bound on $|\Lambda(\theta, \alpha)|$ as claimed. \square

The above result provides an explanation for the blow-up observed in, for example, Figure 1.2. With this in mind, we aim to construct the contour γ in (2.2) so that the denominator in the integrand never gets too close to zero. To do this, we will also need a lower bound on Λ in the complex plane.

LEMMA 2.7. For Λ as in (1.6) and $c \in \mathbb{R}$,

$$|\Lambda(\theta, \alpha)| \leq |\Lambda(\theta + ic, \alpha)|, \quad (\theta, \alpha) \in \mathbb{T}.$$

Proof. Define $J(\theta, c) := |\Lambda(\theta + ic, \alpha)|$. We are interested in how Λ changes as we move from the real line in the positive imaginary direction. Therefore, the following quantity will be useful:

$$\frac{\partial J}{\partial c}(\theta, c) = \frac{p \left((-1)^{p+1} \cos(p\alpha) \cos(p\theta) + \cosh(pc) \sinh(pc) \right)}{\sqrt{(\cos(p\alpha) \cosh(pc) + (-1)^{p+1} \cos(p\alpha))^2 + (\sin(p\theta) \sinh(pc))^2}},$$

which can be obtained by using the representation

$$\Lambda(\theta + ic, \alpha) = \cos(p\theta) \cosh(pc) - i \sin(p\theta) \sinh(pc) - (-1)^p \cos(p\alpha).$$

The denominator is clearly positive, because θ, α and c are all real. We focus on the sign of the numerator. We have $\cos(p\alpha) \cos(p\theta) \geq -1$ and, for all $c \neq 0$, $\cosh(pc) > 1$, hence

$$\left((-1)^{p+1} \cos(p\alpha) \cos(p\theta) + \cosh(pc) \sinh(pc) \right) \begin{cases} > 0 & \text{when } c > 0 \\ = 0 & \text{when } c = 0 \\ < 0 & \text{when } c < 0. \end{cases}$$

Hence $|\Lambda(\theta + ic, \alpha)|$ is minimised when $c = 0$, proving the assertion. \square

The above result tells us that $|\Lambda(z, \alpha)|$ increases as we move vertically into the complex plane. This a helpful result for evaluating the contour integral (2.2), because numerical errors are amplified when $|\Lambda|$ is small. Considering that Λ is in the numerator and the denominator of (2.2), we also require the following bound.

LEMMA 2.8. *For Λ as in (1.6),*

$$\frac{e^{p|\Im z|} - 3}{2} \leq |\Lambda(z, \alpha)| \leq \frac{e^{p|\Im z|} + 3}{2}, \quad \text{for } (z, \alpha) \in \mathbb{C} \times \mathbb{S}.$$

Proof. By the triangle inequality and the definition (1.6),

$$|\Lambda(z, \alpha)| \leq |\cos(pz)| + 1 = \left| \frac{e^{ipz} + e^{-ipz}}{2} \right| + 1.$$

Similarly, by the negative triangle inequality,

$$|\Lambda(z, \alpha)| \geq |\cos(pz)| - 1 = \left| \frac{e^{ipz} + e^{-ipz}}{2} \right| - 1.$$

The assertion follows by considering the cases $\Im z$ positive and negative. \square

We are now ready to prove the main result of this section.

Proof of Theorem 2.3. We aim to continuously deform onto a rectangular complex contour γ (as in Figure 2.1), such that we can bound below the denominator of (2.1), and (2.2), uniformly on γ . We denote this bound by $\mathcal{M}(\gamma) := \min_{z \in \gamma} \{\Lambda(z, \alpha)(z - \theta)\}$. Using this bound, which will be derived shortly, together with Lemma 2.8, taking the absolute value of the integrand and applying the Cauchy-Schwarz inequality, we can obtain the following estimate:

$$(2.11) \quad \left| \frac{1}{2\pi i} \oint_{\gamma} \frac{\sum_{m=1}^M b_m(\alpha) \Lambda(z, \alpha_m) [D(z, \alpha_m) - D_N(z, \alpha_m)]}{\Lambda(z, \alpha)(z - \theta)} dz \right| \leq \frac{|\gamma|(3 + e^{p \max |\Im \gamma|})}{4\pi \mathcal{M}(\gamma)} \sqrt{\sum_{m=1}^M |b_m(\alpha)|^2 \epsilon_m^2}.$$

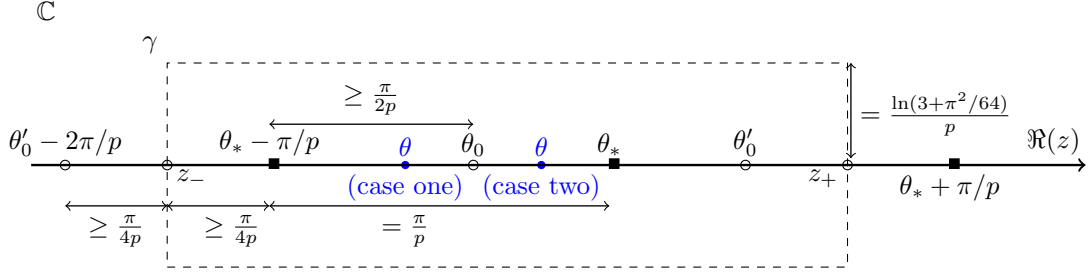


FIGURE 2.1. Rectangular contour, covering both cases one and two of the proof of Theorem 2.3. The arrows are labelled with values (or bounds) of the length of the regions to which the arrows are parallel.

It is sufficient to consider two cases, as depicted in Figure 2.1, the rest follow by symmetry.

Case one is where $\theta \leq \theta_0 \leq \theta_*$. By definition, θ_* is closer to θ_0 than $\theta_* - \pi/p$ is to θ_0 , hence $|(\theta_* - \pi/p) - \theta_0| \geq \pi/(2p)$. As the elements of Θ_α are symmetric about the elements of Θ_* , we can then deduce that $|(\theta_* - \pi/p) - (\theta'_0 - 2\pi/p)| \geq \pi/(2p)$. If we choose γ to bisect the real line at z_- , the point halfway between $(\theta_* - \pi/p)$ and $(\theta'_0 - 2\pi/p)$, then we have that $\text{dist}(\Theta_\alpha, z_-) \geq \pi/(4p)$ and $\text{dist}(\Theta_*, z_-) \geq \pi/(4p)$. Since $\theta \geq (\theta_* - \pi/p)$, we also have $\text{dist}(z_-, \theta) \geq \pi/(4p)$. A similar construction holds in the region to the right, choosing γ to intersect \mathbb{R} at the point z_+ , which is halfway between θ'_0 and $\theta_* + \pi/p$. This is indicated by Figure 2.1, and the bound $\text{dist}(z_+, \theta) \geq \pi/(4p)$ follows similar arguments.

Case two is where $\theta_0 \leq \theta \leq \theta_*$. If we use the same contour as in case one, we can make the stronger statement $\text{dist}(z_\pm, \theta) \geq 3\pi/(4p)$. By combining both cases, we have

$$(2.12) \quad \text{dist}(\Theta_\alpha, z_\pm) \geq \pi/(4p), \quad \text{dist}(\Theta_*, z_\pm) \geq \pi/(4p), \quad |\theta - z_\pm| \geq \pi/(4p).$$

Combining the above statement with Lemma 2.6, we obtain

$$(2.13) \quad \mathcal{M}(\{z_-, z_+\}) \geq \frac{p^2}{8} \left(\frac{\pi}{4p} \right)^3 = \frac{\pi^3}{512p}.$$

Denote by γ_v the union of the vertical components of the complex rectangle γ . By combining (2.13) with Lemma 2.7, we have $\mathcal{M}(\gamma_v) \geq \pi^3/(512p)$.

We know that the integrand will oscillate along the horizontal components, which we denote by γ_h . We want to choose the length of γ_v to be sufficiently large that $\mathcal{M}(\gamma_h)$ is bounded below, where γ_h are the horizontal components of the rectangle γ . From Lemma 2.8 we have

$$|\Lambda(\theta + ic, \alpha)| \geq \frac{e^{|c|p}}{2} - \frac{3}{2},$$

and it follows that if we choose $c := \pm \ln(3 + \pi^2/64)/p$, then

$$(2.14) \quad |\Lambda(\theta + ic, \alpha)| \geq \frac{\pi^2}{128}, \quad \text{for } \theta \in \mathbb{R}.$$

Thus we choose $\Im\gamma_h = \{-c, c\}$, where $c := \pm \ln(3 + \pi^2/64)/p$. Since $c \geq \pi/(4p)$, we have $\text{dist}(\gamma_h, \theta) \geq \pi/(4p)$ and hence, from (2.12), $\text{dist}(\gamma, \theta) \geq \pi/(4p)$. Combining this with (2.14)

gives

$$(2.15) \quad \mathcal{M}(\gamma) \geq \frac{\pi^3}{512p}.$$

The final ingredient required in the bound (2.11) is a bound on the contour length $|\gamma|$. Recalling that the distance between the elements of Θ_* is π/p , and our choice of z_{\pm} , as stated above, we have $z_- \geq \theta_* - \pi/p - \pi/(2p)$, and $z_+ \leq \theta_* + \pi/p$, so $|z_+ - z_-| \leq 5\pi/(4p)$. Thus we can bound the length of the rectangle as follows:

$$(2.16) \quad |\gamma| = 2|z_+ - z_-| + 4c \leq \frac{5\pi + 4 \ln(3 + \pi^2/64)}{4p}.$$

Inserting (2.15) and (2.16) into (2.11) completes the proof. \square

We do not expect the rectangular choice of γ to be optimal; other choices of γ may yield a smaller constant C in Theorem 2.3.

3. Computing the embedding coefficients. We now focus on (C2). Intelligently choosing the canonical incident angles $\alpha_1, \dots, \alpha_M$ such that Assumption 1.3 holds is a challenging problem. For standard bases of polynomials or trigonometric functions, *unisolvence* theorems (see for e.g. [13, §2.4]) tell us that an M -dimensional function can be reconstructed uniquely from a set of M distinct points. The problem (1.9) may be interpreted similarly, as a reconstruction problem in the non-standard basis $\{\hat{D}(\cdot, \alpha_m)\}_{m=1}^M$; here we have no theoretical guarantees about reconstruction from M distinct points. Assuming \mathbf{b} is unique and we can construct it, (2.3) adds a further constraint that $\|\mathbf{b}\|_2$ cannot get too large (recall the discussion after Theorem 2.3). These difficulties are already significant before we consider that, in practice, we must work with an approximation to the matrix and right-hand side of (1.9).

To address (C2), we consider two empirical approaches inspired by the philosophy of *oversampling*. This approach has proven to be effective in cases with non-orthogonal bases which satisfy the *frame* condition in [1], and was observed to be effective for problems using high-frequency bases which do not satisfy the frame condition in [20]. The idea here is to incorporate $\tilde{M} > M$ canonical far-field patterns into our algorithm, more than are strictly necessary according to [3]. The larger we choose \tilde{M} , the more likely it is that there exists a subset of $\{\alpha_m\}_{m=1}^{\tilde{M}}$ which satisfies Assumption 1.3. Likewise, we expect that if \tilde{M} is large enough, there are many subsets which satisfy Assumption 1.3, and we are free to choose a subset which minimises $\|\mathbf{b}\|_2$. Thus, by introducing redundancy, we have more samples and basis functions available, reducing the risk of aliasing and the effects of ill-conditioning. Hence, we expect that this gives us a better chance of solving (C2).

Suppose we use a *sampling strategy*, which chooses $\{\alpha_m\}_{m \in \mathbb{N}}$ dense in \mathbb{S} . Then for $\tilde{M} > M$, we consider the *oversampled system*

$$(3.1) \quad \tilde{A}\tilde{\mathbf{b}} = \tilde{\mathbf{d}}, \quad \text{where } \tilde{A} := [\hat{D}(\alpha_m, \alpha_{m'})]_{m, m'=1}^{\tilde{M}}, \quad \tilde{\mathbf{d}} = (-1)^{p+1} [\hat{D}(\alpha, \alpha_m)]_{m=1}^{\tilde{M}}.$$

It is easy to see that for any $\tilde{M} > M$, we have $\text{rank}(\tilde{A}) \leq M$, and we expect that there are multiple $\tilde{\mathbf{b}}$ which satisfy (3.1).

We have added redundancy to our problem, under the assumption that this increases the chance of finding a coefficient vector which addresses (C2). We now consider two different strategies to select the coefficient vector.

3.1. Strategy One. For the first strategy, we assume that \tilde{M} is chosen sufficiently large that there are multiple solutions $\tilde{\mathbf{b}}$ to (3.1), each satisfying

$$(3.2) \quad \hat{D}(\cdot, \alpha) = \sum_{m=1}^{\tilde{M}} \tilde{b}_m(\alpha) \hat{D}(\cdot, \alpha).$$

When multiple solutions exist to (3.1), the system is under-determined, and Algorithm 3.1 (see, e.g., [21]) with $\delta = 0$ provides a pseudo-inverse which will find the solution with minimal norm - this is consistent with our aim to address (C2).

Algorithm 3.1 Pseudo-inverse via truncated singular value decomposition

- 1: **Inputs:** $X \in \mathbb{C}^{\tilde{M}, \tilde{M}}$, $\delta > 0$
- 2: Compute the singular value decomposition,

$$X = U\Sigma V^*$$

- 3: Denote by σ_m the m th entry of Σ . Define Σ^\dagger as the diagonal matrix with entries

$$\sigma_m^\dagger \leftarrow \begin{cases} 1/\sigma_m & \text{if } \sigma_m > \delta, \\ 0 & \text{if } \sigma_m \leq \delta \end{cases}, \quad m = 1, \dots, \tilde{M}$$

- 4: **return** Return pseudo-inverse $X^\dagger \leftarrow V\Sigma^\dagger U^*$
-

However, we must consider the added implications of working with numerical approximations. Hence, we define another problem:

$$(3.3) \quad \tilde{A}_N \tilde{\mathbf{b}}_N \approx \tilde{\mathbf{d}}_N, \quad \text{where } \tilde{A}_N := [\hat{D}_N(\alpha_m, \alpha_{m'})]_{m, m'=1}^{\tilde{M}} \approx \tilde{A}, \quad \tilde{\mathbf{d}}_N = (-1)^{p+1} [\hat{D}_N(\alpha, \alpha_m)]_{m=1}^{\tilde{M}} \approx \tilde{\mathbf{d}}.$$

To determine $\tilde{\mathbf{b}}_N$, we will use Algorithm 3.1. The following lemma, simply an application of [11, Lemma 3.3], describes how the error is balanced between the residual and coefficient norms.

LEMMA 3.1. *Suppose that the pseudo-inverse \tilde{A}_N^\dagger is obtained using Algorithm 3.1 with matrix \tilde{A}_N and threshold δ . Then*

$$(3.4) \quad \tilde{\mathbf{b}}_N := \tilde{A}_N^\dagger \tilde{\mathbf{d}}_N,$$

satisfies

$$\|\tilde{\mathbf{d}}_N - \tilde{A}_N \tilde{\mathbf{b}}_N\|_2 \leq \inf_{\mathbf{v} \in \mathbb{C}^{\tilde{M}}} \left\{ \|\tilde{\mathbf{d}}_N - \tilde{A}_N \mathbf{v}\|_2 + \delta \|\mathbf{v}\|_2 \right\}.$$

Lemma 3.1 informs us how to choose δ so as to balance the residual error with the coefficient norm. A small choice of δ means that the residual error is relatively small, whereas a large choice of δ means that the coefficient norm is relatively small. A balance of both is desirable. When balancing these two, we first consider that for fixed N , even if $\|\tilde{\mathbf{d}}_N - \tilde{A}_N \tilde{\mathbf{b}}_N\|_2 \rightarrow 0$ for $\delta \rightarrow 0$, this does not imply $\|\tilde{\mathbf{b}}_N - \tilde{\mathbf{b}}\|_2 \rightarrow 0$, for any $\tilde{\mathbf{b}}$ solving (3.2). This is simply due to the problems (3.1) and (3.3) having different solutions. Obviously, as $N \rightarrow \infty$, we expect $\|\tilde{\mathbf{b}}_N - \tilde{\mathbf{b}}\|_2$ to be small,

but it can only be made *so small* by decreasing δ . There will exist a threshold as $\delta \rightarrow 0$, beyond which there is no practical advantage in decreasing δ any further. Moreover, choosing δ too small is a disadvantage, because the norm $\|\tilde{\mathbf{b}}_N\|_2$ is permitted to become very large as $\delta \rightarrow 0$, which is inconsistent with our aim to address (C2). This threshold will depend on the accuracy of $D \approx D_N$, but can be estimated using the following result.

LEMMA 3.2. *For \tilde{A}_N and $\tilde{\mathbf{d}}_N$ of the approximate system (3.3), $\tilde{\mathbf{b}}$ of the exact system (3.1) and $\|\epsilon\|_2$ as in Theorem 2.3,*

$$\inf_{\mathbf{v} \in \mathbb{C}^{\tilde{M}}} \left\{ \|\tilde{\mathbf{d}}_N - \tilde{A}_N \mathbf{v}_N\|_2 \right\} \leq \|\tilde{\mathbf{d}}_N - \tilde{A}_N \tilde{\mathbf{b}}\|_2 \leq 2(\tilde{M} \|\tilde{\mathbf{b}}\|_2 + 1) \|\epsilon\|_2.$$

Proof. The infimum is taken over all $\mathbf{v} \in \mathbb{C}^{\tilde{M}}$, we proceed by choosing $\mathbf{v} = \tilde{\mathbf{b}}$. We have

$$\begin{aligned} \tilde{\mathbf{d}}_N - \tilde{A}_N \tilde{\mathbf{b}} &= \tilde{\mathbf{d}}_N + (\tilde{\mathbf{d}} - \tilde{\mathbf{d}}) - \left[\tilde{A}_N + (\tilde{A} - \tilde{A}) \right] \tilde{\mathbf{b}} \\ &= \left[\tilde{\mathbf{d}}_N - \tilde{\mathbf{d}} - (\tilde{A}_N - \tilde{A}) \tilde{\mathbf{b}} \right] + \left[\tilde{\mathbf{d}} - \tilde{A} \tilde{\mathbf{b}} \right] \\ &= \tilde{\mathbf{d}}_N - \tilde{\mathbf{d}} - (\tilde{A}_N - \tilde{A}) \tilde{\mathbf{b}}, \end{aligned}$$

by (3.1). Elementary norm manipulation then gives

$$\|\tilde{\mathbf{d}}_N - \tilde{A}_N \tilde{\mathbf{b}}\|_2 \leq \|\tilde{\mathbf{d}}_N - \tilde{\mathbf{d}}\|_2 + \|\tilde{A}_N - \tilde{A}\|_F \|\tilde{\mathbf{b}}\|_2,$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The assertion follows by bounding each entry of $\tilde{\mathbf{d}}_N - \tilde{\mathbf{d}}$ and $\tilde{A}_N - \tilde{A}$, in terms of ϵ_m for $m = 1, \dots, \tilde{M}$. The factor of two follows because $|\Lambda| \leq 2$. \square

In some sense, $\tilde{\mathbf{b}}$ is the perfect approximation to the solution of (3.4), because there will be no error in our embedding coefficients. Lemma 3.2 quantifies that even with this *perfect solution*, there will still be a residual error; therefore, we should not waste too much effort minimising the residual, especially if this comes at the cost of $\|\tilde{\mathbf{b}}_N\|_2$ growing. Assuming that $\|\tilde{\mathbf{b}}_N\|_2 \approx \|\tilde{\mathbf{b}}\|_2$, Lemmas 3.1 and 3.2 suggest that to balance the residual error with the coefficient norm, a sensible choice is

$$(3.5) \quad \delta \geq \tilde{M} \|\epsilon\|_2.$$

This will minimise the coefficient norm $\|\mathbf{b}\|_2$, subject to the constraint that the residual error is no smaller than necessary. In §5.2, we experiment with different \tilde{M} and δ , providing solid numerical evidence that this approach addresses (C2).

3.2. Strategy Two. In Strategy One, we addressed (C2) by increasing the number of canonical far-field patterns in the embedding formula from M to \tilde{M} , minimising the least-squares error. Strategy Two considers \tilde{M} canonical far-field patterns only as a preprocessing step, before restricting to a subset of M indices $\mathcal{I} \subset \{1, \dots, \tilde{M}\}$ which are in some sense optimal. Then in the embedding formula, we use the canonical incident angles $\{\alpha_m\}_{m \in \mathcal{I}}$, thus choosing A (of (1.9)) as a square submatrix of \tilde{A} , keeping only the rows and columns in the index set \mathcal{I} . Strategy Two aims to choose this submatrix A such that (C2) is addressed.

To achieve this, we use Algorithm 3.2, initially proposed in [5]. This greedy algorithm is designed for problems which require a subset of matrix columns which maximises volume and equivalently (see [26]), minimises condition number. Although these are NP-hard optimisation problems, this algorithm achieves near-optimal results [6].

Algorithm 3.2 Column subset selection

-
- 1: **Inputs:** $\tilde{A} = [\mathbf{a}_1 | \dots | \mathbf{a}_{\tilde{M}}] \in \mathbb{C}^{\tilde{M} \times \tilde{M}}$, $M < \tilde{M}$
 - 2: Define an empty array $\mathcal{I} \leftarrow \{\}$
 - 3: **while** \mathcal{I} contains fewer than M elements **do**
 - 4: Assign

$$m^* \leftarrow \arg \max_{m=1, \dots, \tilde{M}} \{|\mathbf{a}_m|\}$$

- 5: Update $\mathcal{I} \leftarrow \mathcal{I} \cup m^*$
- 6: Update each vector

$$\mathbf{a}_m \leftarrow \mathbf{a}_m - \mathbf{a}_{m^*} \frac{\langle \mathbf{a}_m, \mathbf{a}_{m^*} \rangle}{\langle \mathbf{a}_{m^*}, \mathbf{a}_{m^*} \rangle}, \quad m = 1, \dots, \tilde{M}$$

- 7: **end while**
 - 8: **return** \mathcal{I}
-

Informally, at each iteration of the **while** loop, Algorithm 3.2 chooses the column vector which is (in some sense) the most orthogonal to the columns which have already been logged in the array \mathcal{I} , via the same computation used in Gram-Schmidt orthogonalisation.

To the best knowledge of the authors, current theoretical results for Algorithm 3.2 apply to column subset selection, whereas our application requires us to choose a subset of columns and rows. Despite the lack of available theory, our numerical experiments of §5 suggest that Strategy Two is highly effective in practice, provided \tilde{M} is chosen sufficiently large; our experiments suggest that $\tilde{M} = \lceil 3M/2 \rceil$ is typically more than sufficient. In terms of accuracy, we observe that it outperforms Strategy One at high frequencies. Moreover, considering that most CPU time is spent on far-field evaluations $D_N(\theta, \alpha)$, Strategy Two is more computationally efficient, by a factor of roughly M/\tilde{M} .

4. The algorithm. The algorithm we use for the numerical experiments in §5 is based upon the ideas presented in §1–§3, with some modifications in order to minimise unnecessary flops (floating point operations). It can be seen from Figure 1.2 that for some values of θ the naive approximation (1.7) is sufficient. With this in mind, we introduce a threshold $H > 0$, such that if $|\theta - \theta_0|_{2\pi} < H$, then we consider it necessary to correct the naive approximation in some way, otherwise we simply use (1.7). When a correction is considered necessary, all of the relevant integrals and sums are based on equations (2.4) and (2.5), with the following exception: the residue θ'_0 is excluded from the sum (2.4) when $|\theta_0 - \theta'_0|_{2\pi} \geq H$, as it is considered to have a negligible contribution.

Similarly, in the vast majority of cases, there will be no issues with rounding errors, and we introduce a second threshold $h < H$, such that if $|\theta_0 - \theta'_0|_{2\pi} < h$, we use the interpolation approach described in §2.1. For these contour integrals, we use a rectangular contour $\mathcal{R}(X, h)$, where X is the set of poles which the integral encloses, chosen so that $\mathcal{R}(X, h)$ is the smallest rectangle with X in its interior, satisfying $\text{dist}(X, \mathcal{R}(X, h)) = h$. These contour integrals are evaluated using a 20-point Gaussian quadrature rule along each edge.

Algorithm 4.1 summarises the key steps of our implementation, excluding details such as quadrature (discussed above), for brevity. Our algorithm has been implemented in Matlab and is available at [18]. This implementation is intended to be a proof of concept — developing a streamlined

Algorithm 4.1 Main routine

-
- 1: **Inputs:** $\theta, \alpha, \{D_N(\cdot, \alpha_m)\}_{m=1}^{\tilde{M}}, \{\frac{\partial}{\partial \theta} D_N(\cdot, \alpha_m)\}_{m=1}^{\tilde{M}}, \{\frac{\partial^2}{\partial \theta^2} D_N(\cdot, \alpha_m)\}_{m=1}^{\tilde{M}}, p, \text{Strategy} \in \{1, 2\}$
 - 2: **Adjustable parameters:** $\delta = 10^{-8}, H = 0.1, h = 10^{-3}$
 - 3: **First pre-computation step:** Construct $\{\hat{D}_N(\cdot, \alpha_m)\}_{m=1}^{\tilde{M}}$ (and the first and second derivatives) using (1.8).
 - 4: **Second pre-computation step - determining coefficient vector.** Note that this step does not need to be repeated for future values of $(\theta, \alpha) \in \mathbb{T}$:
 - 5: **if Strategy = 1 then**
 - 6: Construct \tilde{A}_N using (3.3), then construct and store \tilde{A}_N^\dagger using Algorithm 3.1 with inputs \tilde{A}_N and δ .
 - 7: **else if Strategy = 2 then**
 - 8: Construct a submatrix A_N using Algorithm 3.2 with inputs \tilde{A}_N and M . Store A_N^{-1} .
 - 9: **end if**
 - 10: Construct $\tilde{\mathbf{d}}_N$ using (3.3) and hence the coefficient vector $\tilde{\mathbf{b}}_N$
 - 11: **if $|\theta - \theta_0|_{2\pi} > h$ then**
 - 12: Assign the naive approximation (1.7): $I \leftarrow \mathfrak{D}(\theta, \alpha)$.
 - 13: **if $|\theta - \theta_0|_{2\pi} \in [h, H)$ then**
 - 14: A correction will be necessary - assign

$$I_{\text{correction}} \leftarrow \begin{cases} \frac{1}{2\pi i} \oint_{\mathcal{R}(\{\theta_0, \theta'_0\}, h)} \frac{\rho_2(z)}{\Lambda(z, \alpha)(z - \theta)} dz, & |\theta - \theta'_0|_{2\pi} < h \\ \sum_{\chi \in \{\theta_0, \theta'_0\}} \frac{\sum_{m=1}^{\tilde{M}} b_m(\alpha) \hat{D}_N(\chi, \alpha_m)}{p(\chi - \theta) \sin(p\chi)}, & |\theta - \theta'_0|_{2\pi} \in [h, H) \\ \frac{\sum_{m=1}^{\tilde{M}} b_m(\alpha) \hat{D}_N(\theta_0, \alpha_m)}{p(\theta_0 - \theta) \sin(p\theta_0)}, & |\theta - \theta'_0|_{2\pi} > H \end{cases}$$

- 15: Update $I \leftarrow I + I_{\text{correction}}$
- 16: **end if**
- 17: **else if $\theta = \theta_0 = \theta_*$ then**
- 18: Apply L'Hôpital's rule (twice):

$$I \leftarrow \frac{\sum_{m=1}^{\tilde{M}} \tilde{b}_m(\alpha) \frac{\partial^2 \hat{D}_N(\theta_0, \alpha_m)}{\partial \theta^2}}{-p^2 \cos(p\theta)}$$

- 19: **else**
- 20: Risk of rounding errors, use polynomial interpolation at the poles.
- 21: Construct ρ_2 , interpolating as described in §2.1.

$$I \leftarrow \begin{cases} \frac{1}{2\pi i} \oint_{\mathcal{R}(\{\theta, \theta_0, \theta'_0\}, h)} \frac{\rho_2(z)}{\Lambda(z, \alpha)(z - \theta)} dz, & |\theta_0 - \theta'_0|_{2\pi} < h \\ \frac{1}{2\pi i} \oint_{\mathcal{R}(\{\theta, \theta_0\}, h)} \frac{\rho_2(z)}{\Lambda(z, \alpha)(z - \theta)} dz + \frac{\sum_{m=1}^{\tilde{M}} b_m(\alpha) \hat{D}_N(\theta'_0, \alpha_m)}{p(\theta'_0 - \theta) \sin(p\theta'_0)}, & |\theta_0 - \theta'_0|_{2\pi} \in [h, H) \\ \frac{1}{2\pi i} \oint_{\mathcal{R}(\{\theta, \theta_0\}, h)} \frac{\rho_2(z)}{\Lambda(z, \alpha)(z - \theta)} dz, & |\theta_0 - \theta'_0|_{2\pi} > H \end{cases}$$

- 22: **end if**
 - 23: **Output:** I
-

software package is reserved for future work.

A key input to Algorithm 4.1 is $\{D_N(\cdot, \alpha_m)\}_{m=1}^{\tilde{M}}$, which must be obtained by some numerical method. We use two different numerical methods (outlined below) to broaden the range of the following experiments. In both methods, we reformulate the problem (1.1)-(1.3) as the standard first kind boundary integral equation, discretise using a boundary element method, approximating $\partial u/\partial n$ on the boundary of Ω , solving using oversampled collocation, as described in [20]. The code used for both solvers is available at [17].

4.1. Standard hp Boundary Element Method. The first solver is a standard hp boundary element method (BEM, see, e.g., [32]). We use a piecewise polynomial approximation space, defined on a graded mesh. The mesh is chosen so that the largest mesh element is no more than π/k (half a wavelength) long. Towards the corners of the polygon, we use a geometric grading with grading parameter 0.15, and $2P$ layers, where P is the maximal polynomial degree, reducing the polynomial degree linearly towards corners, as described in, e.g., [19].

It is well known that standard BEMs such as this need to increase degrees of freedom like $O(k)$ to maintain accuracy as k increases (see, e.g., [32]). Therefore, we use this method in examples with moderate to low k , where Ω is a polygon.

4.2. Hybrid Numerical-Asymptotic Boundary Element Method (HNA BEM). This method is a non-standard hp -BEM, where the basis consists of piecewise polynomials multiplied by k -dependent oscillatory functions; details are given in [20]. The mesh is graded towards the corners of the obstacle in the same way as for the standard BEM described above. The key difference here is that the mesh width is not k -dependent because the oscillations are resolved by the basis functions [23]. The numerical implementation of [20] is available at [17], and enjoys k -independent cost for screen problems. Using this solver, we are able to experiment for extremely large k for the case where Ω is a screen.

4.3. Default parameters. Through extensive experimentation, the following parameters were found to provide a good balance between efficiency and accuracy:

- For the thresholds introduced above, we choose $H = 0.15$ and $h = 0.01$.
- We oversample with $\tilde{M} = \lceil 3M/2 \rceil$.
- We use Strategy Two, unless explicitly stated otherwise.
- We choose \tilde{M} canonical incident angles equispaced on \mathbb{S} , with $\alpha_1 = 0$.
- When evaluating the contour integrals over the rectangular contours $\mathcal{R}(X, h)$, we use a 20-point Gaussian quadrature rule along each edge of the rectangle.
- If Strategy One is chosen instead of the default Strategy Two, we use $\delta = 10^{-8}$.

Unless mentioned otherwise in the following experiments, it can be assumed that these parameters have been used. It appears that smaller values of H are sufficient when k is large, potentially reducing unnecessary flops. We do not investigate this link here, and use a fixed value for all k .

5. Numerical examples and experiments. In this section, we present numerical experiments demonstrating the effectiveness of Algorithm 4.1.

5.1. Example applications of Algorithm 4.1. The first experiment we present compares the far-field pattern produced by the naive embedding approximation (1.7) and Algorithm 4.1.

Right-angled isosceles triangle, $k = 10$. We consider the far-field induced by a plane wave with angle $\alpha = 5\pi/4$, $k = 10$, and Ω the right-angled isosceles triangle with vertices $\mathbf{P}_1 = (0, 0)$, $\mathbf{P}_2 = (0, 1)$ and $\mathbf{P}_3 = (1, 0)$. By Definition 1.1 we have $q_1 = 6$, $q_2 = q_3 = 7$, $p = 4$ and $M = 17$.

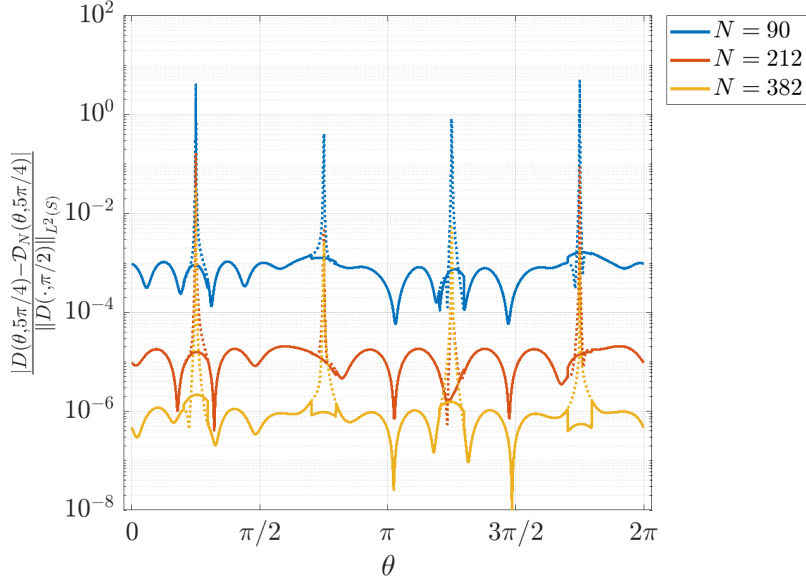


FIGURE 5.1. The errors in the naive approximation (1.7) (dotted) and the output of Algorithm 4.1 (solid), for the right-angled triangle with $k = 10$ and $\alpha = 5\pi/4$. In all cases, the spikes of the naive approximation do not appear using our method.

We use the default parameters of §4.3 and compute D_N using the standard hp -BEM solver of §4.1, with a reference solution of $N_{\text{ref}} = 600$. Figure 5.1 shows the errors for both approximations for a range of N . The naive approximation is shown as a dotted line, and the output of Algorithm 4.1 is shown as a thick line. The relative error is measured as

$$(5.1) \quad \frac{\|\mathcal{D}_N(\cdot, \alpha) - D_{N_{\text{ref}}}(\cdot, \alpha)\|_{L^\infty(\mathbb{S})}}{\|D_{N_{\text{ref}}}(\cdot, \alpha)\|_{L^\infty(\mathbb{S})}},$$

where each norm is approximated with 1000 equispaced points. This approximation will converge exponentially as the number of quadrature points increases, because the far-field is periodic and entire, see, e.g., [35]. The unbounded error of the naive approach is clearly visible and remedied by our new approach, as predicted by Theorem 2.3. Discontinuous jumps in the error of our method are visible. These could be smoothed out by choosing a larger value of H , although it is not necessary to achieve a uniformly low error.

Our method is most powerful when many incident angles are considered. Hence, for the next experiment, we consider one thousand equispaced $\alpha \in \mathbb{S}$. Figure 5.2 shows $\Re[\mathcal{D}_{212}]$ over \mathbb{T} and the pointwise relative error, estimated using

$$\frac{|\mathcal{D}_N(\theta, \alpha) - D_{N_{\text{ref}}}(\theta, \alpha)|}{|D_{N_{\text{ref}}}(\theta, \alpha)|}.$$

The relative error is fairly evenly distributed, peaking at around 10^{-4} , with no visible spikes.

Screen, $k = 100$. For a larger wavenumber $k = 100$, we now consider the far-field induced by the screen $\Omega = [0, 1] \times \{0\}$ and the incident angle $\alpha = \pi/4$. We use the HNA BEM solver described

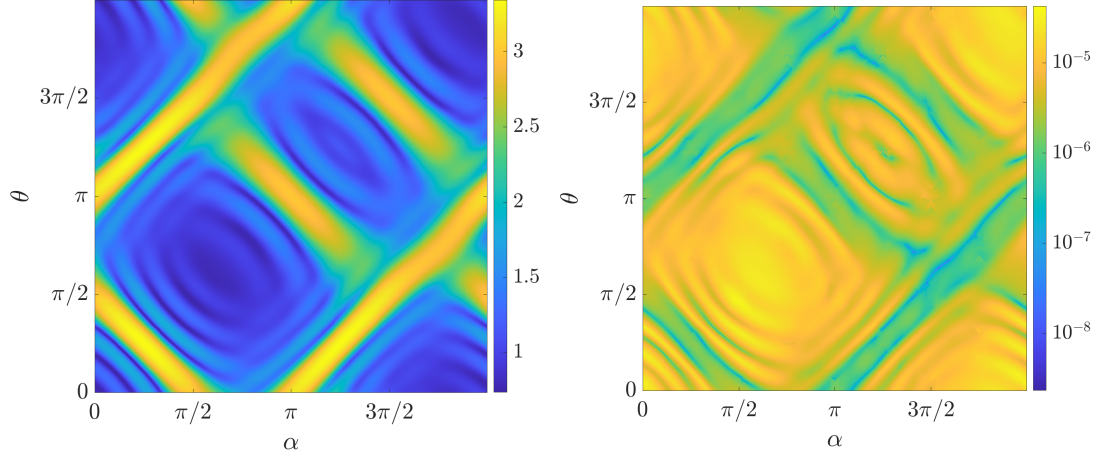


FIGURE 5.2. $\log |D(\theta, \alpha)|$ for right-angled isosceles triangle $k = 10$ (left), and corresponding pointwise relative error (right).

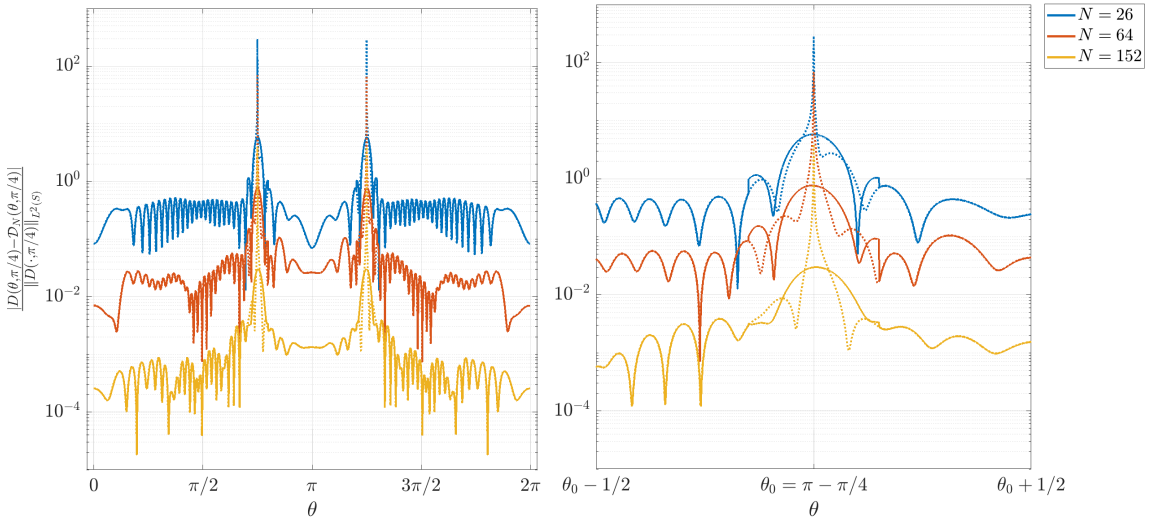


FIGURE 5.3. Left: the errors in the naive approximation (1.7) (dashed) and the output of Algorithm 4.1 (solid), for the screen with $k = 100$ and $\alpha = \pi/4$. Right: zoomed in around a point where (1.7) breaks down, $\theta_0 = \pi - \alpha$.

in §4.2, with a reference solution of $N_{\text{ref}} = 188$. The results are shown in Figure 5.3. Again, the naive approximation ((1.7), dotted lines) blows up at $\theta \in \Theta_\alpha$, which is fixed by our method. The region of blow-up appears to become narrower at higher frequencies, adding weight to our earlier claim in §4.3 that it may be possible to decrease H as k increases.

As in Figure 5.2, we consider the approximation of $D(\theta, \alpha)$ for 1000 values of α on the screen, with $N = 152$ and $N_{\text{ref}} = 188$. The results are shown in Figure 5.4. When compared against the

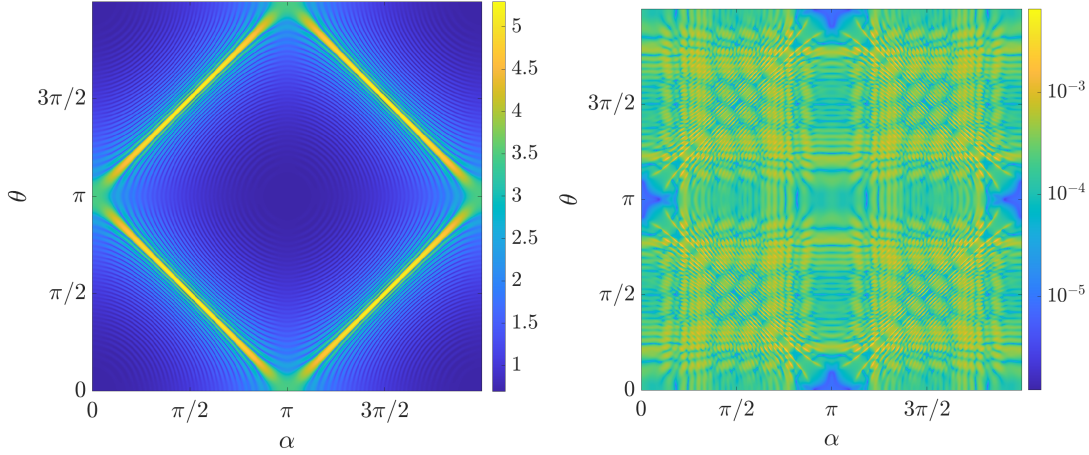


FIGURE 5.4. $\log |D(\theta, \alpha)|$ for screen $k = 100$ (left), and corresponding pointwise relative error (right).

low-frequency solution in Figure 5.2, we observe that the distribution of energy in the far-field has become more focused at $\theta = \pi \pm \alpha$, corresponding to the angles of reflection and propagation of the incident wave. This focusing is typical at higher frequencies.

5.2. Testing the strategies for (C2). Now we return to the two configurations identified in §1.1 for which the system (1.9) is singular or ill-conditioned. These configurations motivated solutions for (C2), and two strategies were proposed in §3. Here, we test these strategies for a range of oversampling parameters $\tilde{M} \geq M$ and truncation parameters δ (the latter only applies to Strategy One).

Screen, $k = 1000$. First, we consider the screen problem $\Omega = [0, 1] \times \{0\}$, with canonical incident angles $\alpha_1 = \frac{\pi}{2}$, $\alpha_2 = \frac{3\pi}{2}$. Recalling the discussion in §1.1, these incident angles correspond to A_N equal to the 2×2 zero matrix. This issue will occur at all wavenumbers, but we consider $k = 1000$. When oversampling, we ensure these two problematic incident angles are included, by considering the first \tilde{M} incident angles of

$$\alpha_1 = \frac{\pi}{2}, \quad \alpha_2 = \frac{3\pi}{2}, \quad \alpha_3 = \pi, \quad \alpha_4 = 0, \quad \alpha_5 = \frac{3\pi}{4}, \quad \alpha_6 = \frac{5\pi}{4}.$$

Figure 5.5 shows the effect of oversampling over the range $\tilde{M} = M = 2$ up to $\tilde{M} = 6$. We test Strategy One with truncation parameters $\delta \in \{10^{-12}, 10^{-8}, 10^{-4}\}$, and compare against Strategy Two. For the solver, we have used HNA BEM (§4.2), with $N = 118$, and for the reference solution $N_{\text{ref}} = 188$. The relative error is measured using (5.1) and the ‘input error’ is measured as

$$(5.2) \quad \mathcal{E}_{\text{in}} := \frac{\max_{m=1, \dots, \tilde{M}} \|D_N(\cdot, \alpha_m) - D_{N_{\text{ref}}}(\cdot, \alpha_m)\|_{L^\infty(\mathbb{S})}}{\max_{m=1, \dots, \tilde{M}} \|D_{N_{\text{ref}}}(\cdot, \alpha_m)\|_{L^\infty(\mathbb{S})}}.$$

The norms in (5.1) and (5.2) are approximated using 1000 equispaced samples.

As expected, for $\tilde{M} = M = 2$, i.e. without oversampling, the relative error is close to one. It should be noted that when $\tilde{M} = M$, Strategy Two does nothing, because the only valid submatrix is the full matrix. Therefore, the blow-up in the coefficient norm and large relative error is what

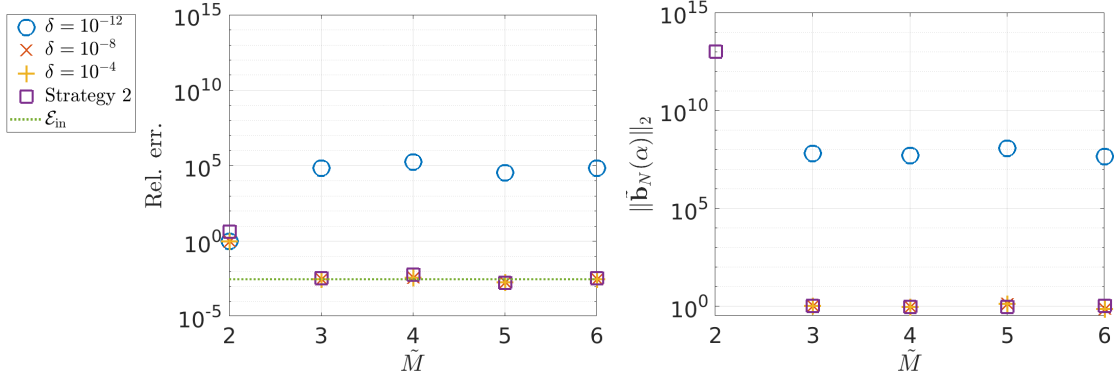


FIGURE 5.5. Error and coefficient norm, varying δ and \tilde{M} in Algorithm 4.1, for the screen with $k = 1000$.

we expect. On the other hand, Strategy One responds by minimising the coefficient norm, which is zero.

For $\tilde{M} \geq 3$ Strategy Two performs well, and over the same range with $\delta \in \{10^{-8}, 10^{-4}\}$, Strategy One performs well. The experiments show that $\delta = 10^{-12}$ is too low for accurate results, which is consistent with the choice (3.5).

Equilateral triangle, $k = 10$. Next we consider the problem where Ω is the equilateral triangle with side length $\sqrt{12}/2$ (such that the vertices lie on the unit circle), with wavenumber $k = 10$. Recalling Figure 1.3, we observed that the set of canonical incident angles $\alpha_m = 2(m-1)\pi/12$, for $m = 1, \dots, M = 12$, caused $\text{cond}(A_N)$ to blow up. We ensure these problematic incident angles are included, and oversample with up to an additional four angles, equispaced between consecutive angles in the above set, like so: $\alpha_m = 2\pi/24 + \alpha_{m-12}$, for $m = 13, \dots, 16$. Now we use a standard hp -BEM, with $N = 375$ and $N_{\text{ref}} = 591$, and as for the screen, the relative error is measured using (5.1) and the input error is measured using (5.2).

As for the screen problem, choosing $\tilde{M} = M + 1$ appears to be sufficient, and for this experiment all values of δ appear to work well, see Figure 5.6. The coefficient norm is slightly increasing for $\delta = 10^{-12}, 10^{-8}$ and $\tilde{M} = 15, 16$, but this does not affect the relative error.

In both of these experiments and in others which are not reported, we have observed that Strategy Two performs at least as well as Strategy One, provided we oversample sufficiently. Considering that Strategy Two has numerous advantages (stated at the end of §3.2), we recommend this as the default choice.

5.3. Conditioning estimates. The constant C in Theorem 2.3 explains the relationship between input error (in the canonical far-field approximation) and output error (of our embedding formula (2.2)). In this sense, C describes the conditioning of our embedding formula. Now we consider another quantity to measure conditioning, measured in terms of the ratio of \mathcal{E}_{out} and \mathcal{E}_{in} , where

$$\mathcal{E}_{\text{out}} := \frac{\|D_N - D_{N_{\text{ref}}}\|_{L^\infty(\mathbb{T})}}{\|D_{N_{\text{ref}}}\|_{L^\infty(\mathbb{T})}}.$$

We approximate the norms using a tensor product trapezoidal rule with 1000×1000 points (again, this converges exponentially by arguments in, e.g., [35]). Here we investigate the relationship between the input and output errors, over a range of wavenumbers, scatterers and solvers.

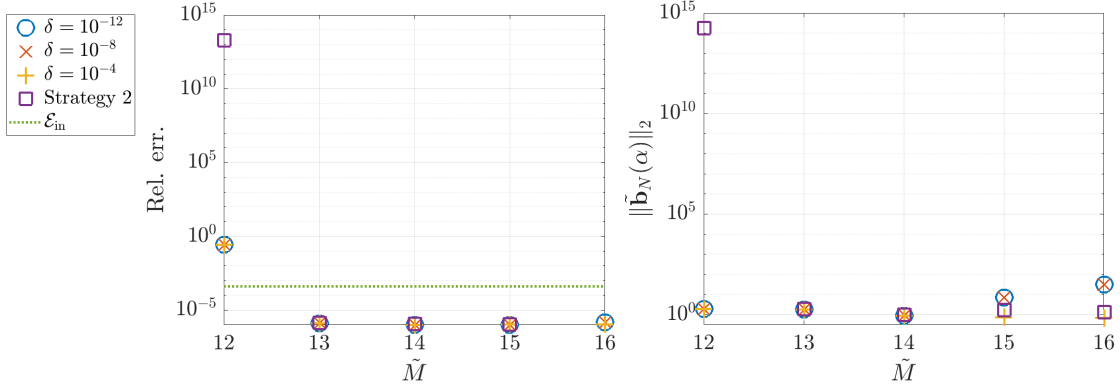


FIGURE 5.6. Error and coefficient norm, varying δ and \tilde{M} in Algorithm 4.1, for the equilateral triangle with $k = 10$.

k	Ω	N	\mathcal{E}_{in}	\mathcal{E}_{out}	$\mathcal{E}_{\text{out}}/\mathcal{E}_{\text{in}}$	$\text{cond}(A_N)$
5	triangle	78	4.3×10^{-4}	1.0×10^{-2}	2.4×10^1	4.3×10^0
	triangle	192	3.6×10^{-6}	7.5×10^{-5}	2.1×10^1	4.3×10^0
	triangle	354	9.7×10^{-8}	1.3×10^{-6}	1.3×10^1	4.3×10^0
	square	104	1.0×10^{-4}	3.5×10^{-2}	3.4×10^2	2.1×10^0
	square	256	1.2×10^{-6}	2.8×10^{-4}	2.4×10^2	2.1×10^0
	square	472	4.0×10^{-8}	1.1×10^{-5}	2.8×10^2	2.1×10^0
	pentagon	130	4.1×10^{-5}	2.3×10^{-3}	5.6×10^1	1.2×10^5
	pentagon	320	5.1×10^{-7}	2.6×10^{-5}	5.0×10^1	1.2×10^5
	pentagon	590	1.9×10^{-8}	2.1×10^{-7}	1.1×10^1	1.2×10^5
25	triangle	168	9.8×10^{-4}	2.7×10^{-2}	2.8×10^1	2.8×10^1
	triangle	342	4.3×10^{-6}	4.0×10^{-4}	9.3×10^1	1.4×10^1
	triangle	564	8.3×10^{-8}	1.1×10^{-5}	1.3×10^2	1.4×10^1
	square	200	6.8×10^{-4}	7.1×10^{-1}	1.1×10^3	1.3×10^1
	square	416	3.4×10^{-6}	3.9×10^{-3}	1.1×10^3	1.3×10^1
	square	696	5.6×10^{-8}	2.5×10^{-5}	4.6×10^2	1.3×10^1
	pentagon	235	4.8×10^{-4}	1.9×10^0	4.0×10^3	3.7×10^2
	pentagon	495	2.1×10^{-6}	7.6×10^{-3}	3.6×10^3	6.2×10^2
	pentagon	835	4.4×10^{-8}	5.8×10^{-5}	1.3×10^3	6.2×10^2

TABLE 5.1

L^∞ input and output errors for a range of regular polygons. Error values and condition numbers are reported to two significant figures.

Regular polygons, low wavenumbers. Low-frequency results for when Ω is a regular polygon with vertices positioned on the unit circle, where D_N is the standard BEM solver (§4.1), are given in Table 5.1. The same experiment was performed for Strategy One, with similar results. For the triangle, $N_{\text{ref}} = 942$, for the square, $N_{\text{ref}} = 1400$, and for the pentagon, $N_{\text{ref}} = 1660$. In all cases, we observe convergence as N increases, and our method can achieve a high accuracy for all incident angles with a relatively low N . There appears to be no obvious rule for predicting $\mathcal{E}_{\text{out}}/\mathcal{E}_{\text{in}}$.

k	N	\mathcal{E}_{in}	\mathcal{E}_{out}	$\mathcal{E}_{\text{out}}/\mathcal{E}_{\text{in}}$	$\text{cond}(A_N)$
500	44	4.2×10^{-4}	8.4×10^{-1}	2.0×10^3	2.8
	90	4.3×10^{-5}	2.5×10^{-1}	5.7×10^3	2.8
	152	1.0×10^{-5}	6.0×10^{-2}	6.0×10^3	2.8
	230	2.9×10^{-6}	1.6×10^{-2}	5.6×10^3	2.8
1000	44	2.4×10^{-4}	1.9×10^0	8.1×10^3	2.1
	90	6.2×10^{-5}	5.8×10^{-1}	9.4×10^3	2.2
	152	4.8×10^{-6}	3.2×10^{-2}	6.7×10^3	2.1
	230	2.9×10^{-6}	5.7×10^{-3}	2.0×10^3	2.1
5000	44	9.6×10^{-5}	6.5×10^{-1}	6.7×10^3	2.6
	90	1.3×10^{-5}	4.7×10^{-2}	3.6×10^3	2.3
	152	2.6×10^{-6}	1.1×10^{-2}	4.1×10^3	2.3
	230	1.3×10^{-6}	1.9×10^{-3}	1.5×10^3	2.3
10000	44	8.2×10^{-5}	5.4×10^{-1}	6.6×10^3	2.6
	90	1.2×10^{-5}	7.4×10^{-2}	6.1×10^3	2.9
	152	1.9×10^{-6}	7.5×10^{-3}	3.9×10^3	3.0
	230	1.0×10^{-6}	2.1×10^{-3}	2.0×10^3	3.0

TABLE 5.2

L^∞ input and output errors, for large k when Ω is a screen. Error values and condition numbers are reported to two significant figures.

Screen, high wavenumbers. High-frequency results on the screen, where D_N is the HNA BEM of §4.2, are given in Table 5.2. Here $N_{\text{ref}} = 188$. Again, we observe convergence in each case as N increases. Convergence was observed to be much slower when the same experiments were run for Strategy One. For $N = 230$, we observe $\approx 1\%$ error or less for all wavenumbers tested. This suggests that N does not need to be very large to accurately represent the far-field pattern for all incident angles at high frequencies. This is a very encouraging result, suggesting that when paired with the HNA BEM, the error and cost of our method remain fixed for large k .

6. Conclusions and future work. Embedding formulae describe the fascinating theoretical connection between the far-field patterns induced by different incident plane waves. We have shown that with careful modifications, some of these formulae can be of practical use, significantly reducing the cost in numerical scattering models for two-dimensional sound-soft polygons.

It is natural to ask if the techniques of this paper may be generalised to different scattering configurations. Focusing on two natural extensions, Table 6.1 places this current work within the context of necessary related results. The table is intended to highlight gaps elsewhere in the current scattering literature, which must be filled before the work of this paper can (possibly) be generalised.

It is clear from Table 6.1 that the main gap in the current literature is *Step two* - embedding formulae *in terms of far-field patterns*. In principle one could skip Step 2, applying the ideas of this paper to the embedding formulae of [12], which also hold for sound-hard problems and contain $1/\Lambda(\theta, \alpha)$ -type removable singularities; this is a possible area for future work. A similar approach may be possible for the three dimensional structures in [33]. However, a key practical advantage of Step 2 is that (to the best knowledge of the authors) there are many existing solvers for computing far-field patterns, and far fewer for computing edge Green's functions.

We remark that Step 4 is not essential for embedding formulae to be of practical use; any problem which requires the far-field pattern induced for a large number of incident waves may

Step	Sound-soft polygons	Sound-hard polygons	Sound-soft polyhedra
1. Edge Green Embedding formulae	Craster and Shanin [12].	Also in Craster and Shanin [12].	Some cases in [33], for example cubes.
2. Far-field Embedding formulae	Biggs [3].	In [3, 4], Biggs states that sound-hard formulae can be derived with minor modifications to sound-soft problem, but no results have yet been published.	
3. Numerically robust modification	This paper.		
4. Corresponding high-frequency solver	Screens: [20, 17]. Convex polygons: requires frequency-independent implementation of [8, 24] (work in progress).	Requires frequency-independent implementation of [9] (work in progress).	Initial ideas were discussed in [7, §7.6]. Initial experiments on square screens were presented in [22]. No frequency-independent solver is available.

TABLE 6.1

Overview of existing literature on embedding formulae and high-frequency solvers for exterior scattering problems.

enjoy a reduced computational cost using a numerically robust embedding formula, if one exists. We expect to have developed a frequency-independent solver for convex polygons in the near future, and we are excited to combine it with Algorithm 4.1, and investigate the performance at high frequencies.

7. Acknowledgements. The authors thank Nicholas Biggs, Abi Gopal, Stuart Hawkins, Dave Hewett, Daan Huybrechs, Andrea Moiola, Jennifer Scott and Marcus Webb for helpful conversations. We greatly appreciate the valuable comments and suggestions from Nick Trefethen and the anonymous referees. AG is grateful for support from EPSRC grants EP/S01375X/1 and EP/V053868/1.

REFERENCES

- [1] B. ADCOCK AND D. HUYBRECHS, *Frames and numerical approximation II: Generalized sampling*, J. Fourier Anal. Appl., 26 (2020).
- [2] A. P. AUSTIN, P. KRAVANJA, AND L. N. TREFETHEN, *Numerical algorithms based on analytic function values at roots of unity*, SIAM J. Numer. Anal., 52 (2014), pp. 1795–1821.
- [3] N. R. T. BIGGS, *A new family of embedding formulae for diffraction by wedges and polygons*, Wave Motion, 43 (2006), pp. 517–528.
- [4] ———, *Embedding formulae for scattering in a waveguide containing polygonal obstacles*, Quart. J. Mech. Appl. Math., 69 (2016), pp. 409–429.
- [5] P. BUSINGER AND G. H. GOLUB, *Handbook series linear algebra. Linear least squares solutions by Householder transformations*, Numer. Math., 7 (1965), pp. 269–276.

- [6] A. ÇIVRIL AND M. MAGDON-ISMAIL, *On selecting a maximum volume sub-matrix of a matrix and related problems*, Theoret. Comput. Sci., 410 (2009), pp. 4801–4811.
- [7] S. N. CHANDLER-WILDE, I. G. GRAHAM, S. LANGDON, AND E. A. SPENCE, *Numerical-asymptotic boundary integral methods in high-frequency acoustic scattering*, Acta Numer., 21 (2012), pp. 89–305.
- [8] S. N. CHANDLER-WILDE AND S. LANGDON, *A Galerkin boundary element method for high frequency scattering by convex polygons*, SIAM J. Numer. Anal., 45 (2007), pp. 610–640.
- [9] S. N. CHANDLER-WILDE, S. LANGDON, AND M. MOKGOLELE, *A high frequency boundary element method for scattering by convex polygons with impedance boundary conditions*, Commun. Comput. Phys., 11 (2012), pp. 573–593.
- [10] D. COLTON AND R. KRESS, *Inverse acoustic and electromagnetic scattering theory*, vol. 93 of Applied Mathematical Sciences, Springer, New York, third ed., 2013.
- [11] V. COPPÉ, D. HUYBRECHS, R. MATTHYSEN, AND M. WEBB, *The AZ algorithm for least squares systems with a known incomplete generalized inverse*, SIAM J. Matrix Anal. Appl., 41 (2020), pp. 1237–1259.
- [12] R. V. CRASTER AND A. V. SHANIN, *Embedding formulae for diffraction by rational wedge and angular geometries*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 461 (2005), pp. 2227–2242.
- [13] P. J. DAVIS, *Interpolation and approximation*, Dover Publications, Inc., New York, 1975.
- [14] M. GANESH AND S. C. HAWKINS, *A far-field based T -matrix method for two dimensional obstacle scattering*, ANZIAM J., 51 (2009), pp. C215–C230.
- [15] M. GANESH, S. C. HAWKINS, AND R. HIPTMAIR, *Convergence analysis with parameter estimates for a reduced basis acoustic scattering T -matrix method*, IMA J. Numer. Anal., 32 (2012), pp. 1348–1374.
- [16] G. C. GAUNAURD AND M. F. WERBY, *Acoustic Resonance Scattering by Submerged Elastic Shells*, Appl. Mech. Rev., 43 (1990), pp. 171–208.
- [17] A. GIBBS, *HNABEMLAB*, <https://github.com/AndrewGibbs/HNABEMLAB>, (2019).
- [18] ———, *Reef: Residue enhanced embedding formulae*, <https://github.com/AndrewGibbs/REEF>, (2021).
- [19] A. GIBBS, S. N. CHANDLER-WILDE, S. LANGDON, AND A. MOIOLA, *A high-frequency boundary element method for scattering by a class of multiple obstacles*, IMA J. Numer. Anal., 41 (2021), pp. 1197–1239.
- [20] A. GIBBS, D. P. HEWETT, D. HUYBRECHS, AND E. PAROLIN, *Fast hybrid numerical-asymptotic boundary element methods for high frequency screen and aperture problems based on least-squares collocation*, SN Partial Differential Equations and Applications, 1 (2020), p. 21.
- [21] G. H. GOLUB AND C. F. VAN LOAN, *Matrix computations*, Johns Hopkins University Press, Baltimore, MD, fourth ed., 2013.
- [22] J. HARGREAVES, Y. W. LAM, S. LANGDON, AND D. P. HEWETT, *A high-frequency bem for 3d acoustic scattering*, in The 22nd International Conference on Sound and Vibration, 2015.
- [23] D. P. HEWETT, S. LANGDON, AND S. N. CHANDLER-WILDE, *A frequency-independent boundary element method for scattering by two-dimensional screens and apertures*, IMA J. Numer. Anal., 35 (2015), pp. 1698–1728.
- [24] D. P. HEWETT, S. LANGDON, AND J. M. MELENK, *A high frequency hp boundary element method for scattering by convex polygons*, SIAM J. Numer. Anal., 51 (2013), pp. 629–653.
- [25] N. J. HIGHAM, *The numerical stability of barycentric Lagrange interpolation*, IMA J. Numer. Anal., 24 (2004), pp. 547–556.
- [26] Y. P. HONG AND C.-T. PAN, *Rank-revealing QR factorizations and the singular value decomposition*, Math. Comp., 58 (1992), pp. 213–232.
- [27] N. I. IOAKIMIDIS, K. E. PAPADAKIS, AND E. A. PERDIOS, *Numerical evaluation of analytic functions by Cauchy’s theorem*, BIT, 31 (1991), pp. 276–285.
- [28] P. A. MARTIN, *Multiple scattering: Interaction of time-harmonic waves with N obstacles*, vol. 107 of Encyclopedia of Mathematics and its Applications, Cambridge University Press, Cambridge, 2006.
- [29] M. I. MISHCHENKO, J. W. HOVENIER, AND L. D. TRAVIS, *Light scattering by nonspherical particles: Theory, measurements, and applications*, Measurement Science and Technology, 11 (2000), p. 1827.
- [30] P. MONK, *The near field to far field transformation*, COMPEL, 14 (1995), pp. 41–56.
- [31] B. SADIQ AND D. VISWANATH, *Barycentric Hermite interpolation*, SIAM J. Sci. Comput., 35 (2013), pp. A1254–A1270.
- [32] S. A. SAUTER AND C. SCHWAB, *Boundary element methods*, vol. 39 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 2011.
- [33] E. A. SKELTON, R. V. CRASTER, A. V. SHANIN, AND V. VALYAEV, *Embedding formulae for scattering by three-dimensional structures*, Wave Motion, 47 (2010), pp. 299–317.
- [34] L. N. TREFETHEN, *Quantifying the ill-conditioning of analytic continuation*, BIT Numerical Mathematics, 60 (2020), pp. 901–915.
- [35] L. N. TREFETHEN AND J. A. C. WEIDEMAN, *The exponentially convergent trapezoidal rule*, SIAM Rev., 56 (2014), pp. 385–458.