On Integral Equation and Least Squares Methods for Scattering by Diffraction Gratings

Tilo Arens[1], Simon N. Chandler-Wilde[2,*] and John A. DeSanto[3]

[1] *Mathematisches Institut II, Univeristät Karlsruhe, 76128 Karlsruhe, Germany.*
[2] *Department of Mathematics, University of Reading, Whiteknights, PO Box 220, Berkshire RG6 6AX, United Kingdom.*
[3] *Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401, USA.*

Abstract. In this paper we consider the scattering of a plane acoustic or electromagnetic wave by a one-dimensional, periodic rough surface. We restrict the discussion to the case when the boundary is sound soft in the acoustic case, perfectly reflecting with TE polarization in the EM case, so that the total field vanishes on the boundary. We propose a uniquely solvable first kind integral equation formulation of the problem, which amounts to a requirement that the normal derivative of the Green's representation formula for the total field vanish on a horizontal line below the scattering surface. We then discuss the numerical solution by Galerkin's method of this (ill-posed) integral equation. We point out that, with two particular choices of the trial and test spaces, we recover the so-called SC (spectral-coordinate) and SS (spectral-spectral) numerical schemes of DeSanto et al., *Waves Random Media*, **8**, 315-414, 1998. We next propose a new Galerkin scheme, a modification of the SS method that we term the SS* method, which is an instance of the well-known *dual least squares Galerkin method*. We show that the SS* method is always well-defined and is optimally convergent as the size of the approximation space increases. Moreover, we make a connection with the classical least squares method, in which the coefficients in the Rayleigh expansion of the solution are determined by enforcing the boundary condition in a least squares sense, pointing out that the linear system to be solved in the SS* method is identical to that in the least squares method. Using this connection we show that (reflecting the ill-posed nature of the integral equation solved) the condition number of the linear system in the SS* and least squares methods approaches infinity as the approximation space increases in size. We also provide theoretical error bounds on the condition number and on the errors induced in the numerical solution computed as a result of ill-conditioning. Numerical results confirm the convergence of the SS* method and illustrate the ill-conditioning that arises.

Key words: Helmholtz equation; first kind integral equation; spectral method; condition number.

*Correspondence to: S. N. Chandler-Wilde, Department of Mathematics, University of Reading, Whiteknights, Berkshire RG6 6AX, United Kingdom. Email: `S.N.Chandler-Wilde@reading.ac.uk`

1 Introduction

We consider the scattering of a plane acoustic or electromagnetic wave by a perfectly reflecting, periodic surface. Adopting Cartesian coordinates $Oxyz$ we assume that the surface is invariant in the y direction and periodic in the x direction, specified by the equation $z = f(x)$, for some given continuous function f. The mathematical problem to be solved is two-dimensional. We assume throughout that the incident wave is time-harmonic ($e^{-i\omega t}$ time dependence), so that the total wave field u^t is a solution of the Helmholtz equation

$$\Delta u^t + k^2 u^t = 0 \quad \text{in } \Omega, \tag{1.1}$$

where $\Omega := \{\mathbf{r} = (x, z) \in \mathbb{R}^2 : z > f(x)\}$ is that part of the Oxz plane above the scattering surface. Throughout, we will assume that f is periodic with period $L > 0$ and that the incident field u^i is the plane wave

$$u^i(\mathbf{r}) = \exp(ik[x \sin\theta - z \cos\theta]), \tag{1.2}$$

where θ is the angle of incidence, measured from the z-axis, with $-\pi/2 < \theta < \pi/2$. It is the goal to determine the scattered field $u := u^t - u^i$ given the boundary condition

$$u^t = u^i + u = 0 \qquad \text{on } \partial\Omega, \tag{1.3}$$

where $\partial\Omega = \{(x, f(x)) : x \in \mathbb{R}\}$, and given that an appropriate radiation condition on u holds, expressing that u is outgoing from $\partial\Omega$. This problem models scattering of electromagnetic plane waves by a perfectly conducting diffraction grating in the TE polarization case. The same mathematics models acoustic scattering by a one-dimensional periodic sound soft surface.

Many different methods have been proposed for solving this problem. Alternative boundary integral equation methods to those proposed here are discussed in [1, 29, 35], standard differential equation (coupled-mode) based methods in [4, 31, 33], a coordinate-transformation-based differential equation method in [20, 21], and a method of variation of boundaries based on analytic continuation arguments in [5]. Many different specific surface examples are available [12] as well as the first treatment of the problem using spectral methods [13]. An extensive recent review of many of the different computational methods available is made in [14]. A classical method for solving this problem, on which we throw new light in Section 5, is the least squares method [28, 30], in which the scattered field is expressed as a linear combination of solutions of the Helmholtz equation (the Rayleigh expansion (2.1) below) and the coefficients in this expansion are determined by requiring that the boundary condition holds in a least squares sense. We note that, in the context of determining eigenfunctions of the Laplacian in 2D domains, the least squares method has recently been revived by Betcke and Trefethen [3]. The problem can also be tackled via a variational formulation in a part of the domain, truncated by the Rayleigh expansion which provides a non-local boundary condition, with the variational problem solved numerically by standard finite element methods (see e.g. [2, 17, 18]).

In Section 2 of the paper we formulate the scattering problem mathematically and derive the equivalent first kind integral equation formulation which is the basis of three of the numerical methods that we describe later in the paper.

In Section 3 we establish mapping properties of the integral operator that occurs in this formulation and properties of its adjoint operator, these properties key to analysing the stability and convergence of two of the numerical methods we will discuss. In this section we also study sets of basis functions which may be used to develop expansions for the solution of the integral equation. As a consequence of the mapping properties of the integral operators it follows, in particular, that one possible set of basis functions, the so-called *topological basis functions*, used in the numerical methods for diffraction gratings discussed in [15], are linearly independent and are complete in the space of square integrable functions.

The numerical solution, via Galerkin methods, of the first kind integral equation formulation we propose is the subject of Section 4. We point out that, applying a Galerkin scheme with a pulse basis (piecewise constant basis functions as the trial space) leads, after approximation of the integrals involved, to the version of the SC method implemented in [13, 15]. Applying a Galerkin method and expanding the solution in topological basis functions leads to the SS method of [15, 16]. The effectiveness of these methods for scattering by one-dimensional periodic surfaces has been investigated by careful numerical experiments in [15, 16], the experiments suggesting that both methods are very fast and accurate within certain parameter domains but can become ill-conditioned for surfaces with large slopes. We note further that no convergence proofs were given.

In Section 4 we also propose a new method, which we term the SS* method, based on a modification of the topological basis functions. As we point out, with this particular choice of basis functions the Galerkin method is an instance of the so-called *dual least squares method* [24]. The self-regularization properties of this method are well-known (see [24] and the references therein). Applying arguments from the theory of the dual least squares method [24] we are able to establish that the SS* method is convergent. Further we are able to establish precise estimates for the conditioning of the linear system to be solved, and how this conditioning depends on the surface and on the dimension of the approximation space, obtaining an upper bound on the loss of accuracy arising from evaluation of matrix entries by numerical quadrature.

In Section 5 we make connections between the SS* method we propose and the classic least squares method, in which the coefficients in the Rayleigh expansion (equation (2.1) below) for the scattered field are determined directly by the requirement that the boundary condition that the total field vanish is required to hold in a least squares sense. We describe a straightforward implementation of the least squares method which leads to solving the identical linear system to that solved in the SS* method. Given this connection, we are able to apply the results of Section 4 to deduce new information about the least squares method. In particular, we prove that the method is convergent in all cases (previous analyses, [28, 30], exclude certain combinations of the angle of incidence and the period). We also provide the first proof that the condition number of the matrix becomes unbounded

as N (the dimension of the approximation space) tends to infinity, and we provide an upper bound for the condition number as a function of N, the wavenumber, and the maximum surface height.

In the final Section 6 we present some numerical experiments, for scattering by sinusoidal surfaces, using parameter values (surface elevation, period, angle of incidence) selected from the examples for which results were computed previously in [15]. We compare the SC, SS, SS* and least squares methods, using a different type of method (the super-algebraically convergent Nyström method of [29], based on solving a boundary integral equation of the second kind), to provide accurate results for comparison. The limited numerical results illustrate why the methods we study in this paper are interesting for numerical computation, namely that, at least in some cases, very accurate results are obtained with the ratio of number of degrees of freedom to arc-length of boundary in the range 1-2. This compares very well with conventional boundary element methods where a ratio 5-10 is usually recommended in the engineering literature as the minimum requirement for acceptable accuracy. Our limited numerical results also suggest that the SS* and least squares methods have similar accuracy, and are more robust and reliable than the SC and the SS methods. This is in line with the theoretical results of Sections 4 and 5, where we are able to provide rigorous convergence proofs and error estimates for the SS* and least squares methods, while it is not clear theoretically that the SC and SS methods need be convergent as the number of degrees of freedom increases; indeed the numerical results suggest that these methods may not be convergent in all cases. We also use this section to investigate the conditioning of the linear system solved in the SS* and least squares methods. Our calculations confirm the ill-conditioning as $N \to \infty$. Indeed the condition number ultimately grows exponentially with N as our upper bound on the condition number predicts, though our upper bound overpredicts the rate of this exponential growth for the examples we look at.

We close this introduction with a brief list of notations used in the paper. Throughout $\lambda = 2\pi/k$ is the wavelength. The set of measurable functions that are square integrable on $(-L/2, L/2)$, usually denoted as $L^2(-L/2, L/2)$, will be abbreviated as X. The part of $\partial\Omega$ corresponding to a single period from $-L/2$ to $L/2$ will be denoted by Γ, i.e. $\Gamma := \{\mathbf{r} = (x, f(x)) : -L/2 \le x \le L/2\}$. Similarly, $\mathbb{R}_L^2$ denotes the part of the plane ($\mathbb{R}^2$) with $-L/2 < x < L/2$, i.e. $\mathbb{R}_L^2 := \{\mathbf{r} \in \mathbb{R}^2 : -L/2 < x < L/2\}$. It is also useful to have a notation for the finite horizontal line of height h in $\mathbb{R}_L^2$, namely $\Gamma_h := \{(x, h) : -L/2 \le x \le L/2\}$.

2   The scattering problem and a first kind integral equation

Given that the incident field u^i is the plane wave (1.2), we seek a scattered field u and total field $u^t = u + u^i$ which satisfy the Helmholtz equation (1.1) in Ω and the boundary condition (1.3). To capture fully the physics of the problem and ensure uniqueness of solution it is necessary to impose additional constraints, expressed in terms of the following definitions.

Definition 2.1. A function $u \in C(\Omega)$ is said to be *quasi-periodic (or Floquet periodic) with period L and phase-shift μ* if

$$u(x + L, z) = \exp(i\mu L)\, u(x, z),$$

for $\mathbf{r} = (x, z) \in \Omega$.

Let $f_- := \min f$ and $f_+ := \max f$, so that

$$f_- \leq f(x) \leq f_+.$$

Definition 2.2. A function $u \in C^2(\Omega)$ is said to satisfy the *Rayleigh expansion radiation condition (RERC)* if, for some complex constants u_n, which we will call the *Rayleigh coefficients*,

$$u(\mathbf{r}) = \sum_{n \in \mathbb{Z}} u_n \, \exp(ik[\alpha_n x + \beta_n z]), \qquad \text{for } z > f_+, \tag{2.1}$$

where $\alpha_n := \sin\theta + n\lambda/L$ (the Bragg condition) and

$$\beta_n := \begin{cases} \sqrt{1 - \alpha_n^2}, & |\alpha_n| \leq 1, \\ i\sqrt{\alpha_n^2 - 1}, & |\alpha_n| > 1. \end{cases}$$

The incident field is quasi-periodic with period L and phase-shift $\mu = k\sin\theta$, and it is appropriate, given the periodicity of f, to require that the scattered field u is also quasi-periodic with the same period and phase shift. It then follows, given that u satisfies the Helmholtz equation in every half-plane above $\partial\Omega$, that, for $z > f_+$, u is a linear combination of plane waves and inhomogeneous plane waves of the form $\exp(ik[\alpha_n x \pm \beta_n z])$. Discarding those waves which propagate downwards or increase exponentially with z leads to the requirement that u satisfy the RERC.

The complete formulation of the scattering problem is thus as follows. Note that we assume in this problem specification, to simplify later mathematical analysis, that the boundary curve $\partial\Omega$ has continuously varying tangent and curvature, equivalently that f lies in the set of functions $f \in C^2(\mathbb{R})$. For an analysis, including a proof of uniqueness and existence of solution, of the case when f is merely Lipschitz continuous see [18].

Problem 2.1. Given an L-periodic function $f \in C^2(\mathbb{R})$ and an incident field u^i, defined by (1.2), find $u \in C^2(\Omega) \cap C(\bar{\Omega})$, quasi-periodic with period L and phase-shift $\mu = k\sin\theta$, such that $u^t := u + u^i$ satisfies the Helmholtz equation (1.1) and the boundary condition (1.3), and u satisfies the RERC.

Remark 2.1. It is a well known result (e.g. [22, 34]) that Problem 2.1 has exactly one solution, and that the gradient of this solution is continuous up to the boundary $\partial\Omega$, so that $u \in C^1(\bar{\Omega})$, which allows the application below of Green's theorem. It is perhaps less well known that the assumption of quasi-periodicity is not required to ensure uniqueness. It is shown in [9] that a weaker radiation condition than the RERC, the *upward propagating*

radiation condition (UPRC) of [7], implies uniqueness of solution for scattering by general rough surfaces, if it is assumed that u is bounded in every horizontal strip above $\partial\Omega$. We also point out that it is shown in [8] that the weaker UPRC combined with an assumption of quasi-periodicity is equivalent to the RERC.

We proceed to derive a first kind integral equation formulation for Problem 2.1 via applications of Green's theorem. For this purpose, we introduce the quasi-periodic Green's function for the Helmholtz equation G_p, defined by

$$G_p(\mathbf{r}, \mathbf{r_0}) := \frac{i}{2kL} \sum_{n\in\mathbb{Z}} \frac{1}{\beta_n} \exp(ik[\alpha_n(x - x_0) + \beta_n|z - z_0|]), \tag{2.2}$$

for all $\mathbf{r} = (x, z)$, $\mathbf{r_0} = (x_0, z_0)$ with $\mathbf{r} - \mathbf{r_0}$ not a multiple of the vector $(L, 0)$. Of course, G_p is only well-defined in the case that $\beta_n \neq 0$ for all $n \in \mathbb{Z}$ and, for the moment, we assume that this is the case. We note that the quasi-periodic Green's function can be written in many equivalent forms, for example as the sum of Hankel functions

$$G_p(\mathbf{r}, \mathbf{r_0}) = \frac{i}{4} \sum_{n\in\mathbb{Z}} \exp(ik\alpha_0 nL) H_0^{(1)}(k|\mathbf{r} - \mathbf{r}_n|),$$

where $\mathbf{r}_n := (x_0 + nL, z_0)$. These representations and others more suited for numerical calculation are derived and discussed in [15, 27]. Note that it is clear from either of the above representations that $G_p(\mathbf{r}, \mathbf{r}_0)$, as a function of $\mathbf{r}$, is quasi-periodic with period L and phase shift $\mu = k\sin\theta$.

Now let $D \subset \overline{\mathbb{R}^2_L}$ denote any domain in which the divergence theorem holds. Further, let ν denote the outward drawn normal to D. Then, for any solution $u \in C^2(D) \cap C^1(\bar{D})$ of the Helmholtz equation, there holds

$$\int_{\partial D} \left\{ G_p(\mathbf{r}, \mathbf{r_0}) \frac{\partial u}{\partial \nu}(\mathbf{r_0}) - \frac{\partial G_p(\mathbf{r}, \mathbf{r_0})}{\partial \nu(\mathbf{r_0})} u(\mathbf{r_0}) \right\} ds(\mathbf{r_0}) = \begin{cases} u(\mathbf{r}), & \mathbf{r} \in D, \\ 0, & \mathbf{r} \in \overline{\mathbb{R}^2_L} \setminus \overline{D}. \end{cases} \tag{2.3}$$

Possible choices for the domain D are $D_{H_2}^+ := \{\mathbf{r} = (x, z) \in \mathbb{R}^2_L : f(x) < z < H_2\}$, with $H_2 > f_+$, and $D_{H_1}^- := \{\mathbf{r} = (x, z) \in \mathbb{R}^2_L : H_1 < z < f(x)\}$, with $H_1 < f_-$. If Green's formula (2.3) is applied in $D_{H_2}^+$ or $D_{H_1}^-$ to a field u that is quasi-periodic, then the integrals over the vertical lines $x = -L/2$ and $x = L/2$ cancel. In particular, suppose that $H_1 < f_-$ and $H_2 > f_+$. Then, with n denoting the downward drawn normal to Γ, applying (2.3) to u^i in $D_{H_1}^-$ we obtain that, for $\mathbf{r} = (x, z) \in \mathbb{R}^2_L$,

$$\int_\Gamma \left\{ G_p(\mathbf{r}, \mathbf{r_0}) \frac{\partial u^i}{\partial n}(\mathbf{r_0}) - \frac{\partial G_p(\mathbf{r}, \mathbf{r_0})}{\partial n(\mathbf{r_0})} u^i(\mathbf{r_0}) \right\} ds(\mathbf{r_0})$$

$$+ \int_{\Gamma_{H_1}} \left\{ G_p(\mathbf{r}, \mathbf{r_0}) \frac{\partial u^i}{\partial z_0}(\mathbf{r_0}) - \frac{\partial G_p(\mathbf{r}, \mathbf{r_0})}{\partial z_0} u^i(\mathbf{r_0}) \right\} dx_0 = \begin{cases} -u^i(\mathbf{r}), & \text{if } H_1 < z < f(x), \\ 0, & \text{if } z > f(x). \end{cases} \tag{2.4}$$

Similarly, applying (2.3) to u in $D_{H_2}^+$ yields

$$
\int_\Gamma \left\{ G_p(\mathbf{r}, \mathbf{r_0}) \frac{\partial u}{\partial n}(\mathbf{r_0}) - \frac{\partial G_p(\mathbf{r}, \mathbf{r_0})}{\partial n(\mathbf{r_0})} u(\mathbf{r_0}) \right\} ds(\mathbf{r_0})
$$
$$
- \int_{\Gamma_{H_2}} \left\{ G_p(\mathbf{r}, \mathbf{r_0}) \frac{\partial u}{\partial z_0}(\mathbf{r_0}) - \frac{\partial G_p(\mathbf{r}, \mathbf{r_0})}{\partial z_0} u(\mathbf{r_0}) \right\} dx_0 = \left\{ \begin{array}{ll} 0, & \text{if } z < f(x), \\ u(\mathbf{r}), & \text{if } f(x) < z < H_2. \end{array} \right. \tag{2.5}
$$

Since u satisfies the RERC and using the definition of the Green's function we see that, for $z < H_2$, the second integral in (2.5) has the value

$$
\int_{\Gamma_{H_2}} \left\{ G_p(\mathbf{r}, \mathbf{r_0}) \frac{\partial u}{\partial z_0}(\mathbf{r_0}) - \frac{\partial G_p(\mathbf{r}, \mathbf{r_0})}{\partial z_0} u(\mathbf{r_0}) \right\} dx_0
$$
$$
= \int_{\Gamma_{H_2}} \sum_{m \in \mathbb{Z}} u_m \exp(ik[\alpha_m x_0 + \beta_m H_2]) \left\{ G_p(\mathbf{r}, \mathbf{r_0}) ik\beta_m - \frac{\partial G_p(\mathbf{r}, \mathbf{r_0})}{\partial z_0} \right\} dx_0
$$
$$
= \frac{-1}{2L} \int_{-L/2}^{L/2} \sum_{m \in \mathbb{Z}} \sum_{n \in \mathbb{Z}} u_m \exp\left(ik \left[\alpha_n x + \beta_n(H_2 - z) + \beta_m H_2 + (m - n)\frac{\lambda x_0}{L} \right] \right) \left(\frac{\beta_m}{\beta_n} - 1 \right) dx_0
$$
$$
= 0.
$$

A similar calculation yields that the second integral in (2.4) also vanishes, provided $z > H_1$. Thus, adding (2.4) to (2.5) and applying the boundary condition (1.3) yields that

$$
\int_\Gamma G_p(\mathbf{r}, \mathbf{r_0}) \frac{\partial u^t}{\partial n}(\mathbf{r_0}) \, ds(\mathbf{r_0}) = \left\{ \begin{array}{ll} u(\mathbf{r}), & \text{if } z > f(x), \\ -u^i(\mathbf{r}), & \text{if } z < f(x), \end{array} \right. \tag{2.6}
$$

the lower part of this equation often referred to as the extinction theorem [32].

We see from (2.6) that, to compute the scattered field we need only find the normal derivative of the total field on Γ. Equation (2.6) provides integral equations for determining this normal derivative. In particular, we shall compute numerical solutions by solving a first kind integral equation obtained by differentiating (2.6). For $z < f_-$ and $\mathbf{r_0} \in \Gamma$ it is clear from the definition that $G_p(\mathbf{r}, \mathbf{r_0})$ is continuously differentiable with respect to z. Thus we can take the derivative on both sides of (2.6) with respect to z, exchanging the order of differentiation and integration, to obtain the integral equation

$$
-\frac{\partial u^i}{\partial z}(\mathbf{r}) = \int_\Gamma \frac{\partial G_p(\mathbf{r}, \mathbf{r_0})}{\partial z} \frac{\partial u^t}{\partial n}(\mathbf{r_0}) \, ds(\mathbf{r_0}), \qquad \mathbf{r} \in \Gamma_H, \tag{2.7}
$$

which holds for every $H < f_-$.

We will rewrite (2.7) as an integral equation on the interval $(-L/2, L/2)$. For $-L/2 \leq$

$x \leq L/2$, $-L/2 \leq x_0 \leq L/2$, set

$$
\begin{aligned}
\varphi(x_0) &:= \frac{1}{k} \frac{\partial u^t(\mathbf{r_0})}{\partial n}\Big|_{\mathbf{r_0}=(x_0,f(x_0))} \sqrt{1 + f'(x_0)^2}, \\
\psi(x) &:= -\frac{2}{k} \frac{\partial u^i(\mathbf{r})}{\partial z}\Big|_{\mathbf{r}=(x,H)} = 2i\beta_0 \exp(ik[\alpha_0 x - \beta_0 H]), \\
K(x,x_0) &:= 2L \frac{\partial G_p(\mathbf{r},\mathbf{r_0})}{\partial z}\Big|_{\mathbf{r}=(x,H),\,\mathbf{r_0}=(x_0,f(x_0))}.
\end{aligned}
\tag{2.8}
$$

Then (2.7) can be rewritten as the integral equation

$$
\psi(x) = \frac{1}{L} \int_{-L/2}^{L/2} K(x,x_0)\varphi(x_0)dx_0, \quad -L/2 \leq x \leq L/2, \tag{2.9}
$$

or, in operator form,

$$
D\varphi = \psi, \tag{2.10}
$$

where the integral operator D is defined by

$$
D\varphi(x) := \frac{1}{L} \int_{-L/2}^{L/2} K(x,x_0)\,\varphi(x_0)\,dx_0, \quad -L/2 \leq x \leq L/2. \tag{2.11}
$$

Explicitly, the kernel $K(x,x_0)$ is given by

$$
K(x,x_0) = \sum_{n \in \mathbb{Z}} \exp(ik[\alpha_n(x - x_0) + \beta_n(f(x_0) - H)]). \tag{2.12}
$$

The benefit of differentiating (2.6) is that (2.12) is well-defined even in the case that $\beta_n = 0$ for some $n \in \mathbb{Z}$. Our derivation of (2.9) assumed that $\beta_n \neq 0$ for all $n \in \mathbb{Z}$. However, using the result that the solution to Problem 2.1 depends continuously on the angle of incidence [22], it follows that (2.9) holds even when $\beta_n = 0$ for some n, by first perturbing θ slightly to make $\beta_n \neq 0$, so that (2.9) holds, and then taking the limit as this perturbation tends to zero.

In the results shown below we shall compute φ by solving (2.9). Note that (2.9) has exactly one solution in $X = L^2(-L/2, L/2)$. To see this, note first that the derivation above shows that (2.9) does have a solution, namely the normal derivative of the total field that satisfies Problem 2.1. Further, we show in the next section that the operator D is injective so that this solution is unique.

Once φ is obtained, and provided $\beta_n \neq 0$ for all $n \in \mathbb{Z}$, the scattered field is given by (2.6), which can be written as

$$
u(\mathbf{r}) = k \int_{-L/2}^{L/2} G_p(\mathbf{r}, (x_0, f(x_0)))\varphi(x_0)dx_0. \tag{2.13}
$$

An alternative representation for the scattered field can be obtained by reflecting equation (2.6) in the line $z = h$, for some $h < f_-$, to obtain that

$$k \int_{-L/2}^{L/2} G_p(\mathbf{r}, (x_0, 2h - f(x_0)))\varphi(x_0)dx_0 = -\exp(ik[\alpha_0 x - \beta_0(2h - z)]),$$

for $2h - z < f(x)$. In particular this equation holds for $z > f(x)$. Thus, subtracting the equation from (2.13), we find that

$$u(\mathbf{r}) = -\exp(ik[\alpha_0 x - \beta_0(2h - z)]) + k \int_{-L/2}^{L/2} G_{p,h}(\mathbf{r}, (x_0, f(x_0)))\varphi(x_0)dx_0, \qquad (2.14)$$

for $z > f(x)$, where

$$G_{p,h}(\mathbf{r}, \mathbf{r_0}) := G_p(\mathbf{r}, \mathbf{r_0}) - G_p(\mathbf{r}, \mathbf{r_0}'), \qquad (2.15)$$

and $\mathbf{r_0}' := (x_0, 2h - z_0)$ denotes the reflection of $\mathbf{r_0} = (x_0, z_0)$ in the line $z = h$. Note that $G_{p,h}$ is the quasi-periodic Dirichlet Green's function for the upper half-plane $z > h$, since $G_{p,h}(\mathbf{r}, \mathbf{r_0}) = 0$ on $z = h$.

The advantage of (2.14) compared to (2.13) is that, while G_p is undefined when $\beta_n = 0$ for some n, the definition of $G_{p,h}$ can be extended to this case by perturbing θ slightly so that $\beta_n \neq 0$ and then taking the limit in (2.15) as this perturbation tends to zero. From (2.2) and (2.15) we see that, explicitly, this leads to the equation

$$G_{p,h}(\mathbf{r}, \mathbf{r_0}) = \frac{i}{2kL} \sum_{n \in \mathbb{Z}} \exp(ik\alpha_n(x - x_0))c_n(z, z_0), \qquad (2.16)$$

where

$$c_n(z, z_0) := \begin{cases} \frac{1}{\beta_n}[\exp(ik\beta_n|z - z_0|) - \exp(ik\beta_n(z + z_0 - 2h))], & \text{if } \beta_n \neq 0, \\ ik(|z - z_0| - (z + z_0 - 2h)), & \text{if } \beta_n = 0. \end{cases}$$

Our derivation of (2.14) assumed, implicitly, that $\beta_n \neq 0$ for every n. But, in the same way we argued that (2.9) holds when $\beta_n = 0$ for some n, it follows that (2.14) holds in this case too.

From (2.2) and (2.13) we see that the coefficients in the Rayleigh expansion representation for $u(\mathbf{r})$, equation (2.1), are given by

$$u_n = \frac{i}{2L\beta_n} \int_{-L/2}^{L/2} \exp(-ik[\alpha_n x_0 + \beta_n f(x_0)])\varphi(x_0)dx_0, \qquad (2.17)$$

at least in the case that $\beta_n \neq 0$ for all $n \in \mathbb{Z}$. If $\beta_m = 0$ for some $m \in \mathbb{Z}$ we see, from the continuous dependence of both u and β_n on θ, that (2.17) still holds for $n \neq m$.

We can also find expressions for the coefficients u_n from (2.16) and (2.14). For $z > z_0$,

$$c_n(z, z_0) = -2ik(z_0 - h)\exp(ik\beta_n(z - h))\operatorname{sinc}(k\beta_n(z_0 - h)),$$

where

$$\operatorname{sinc} t := \begin{cases} \dfrac{\sin t}{t}, & t \neq 0, \\ 1, & t = 0. \end{cases}$$

Thus it follows from (2.16) and (2.14) that

$$
\begin{aligned}
u_n \;=\; & -\exp(-2ik\beta_0 h)\delta_{0,n} \\
& +\frac{k}{L}e^{-ik\beta_n h}\int_{-L/2}^{L/2} e^{-ik\alpha_n x_0}(f(x_0)-h)\operatorname{sinc}\left(k\beta_n(f(x_0)-h)\right)\varphi(x_0)dx_0, \quad (2.18)
\end{aligned}
$$

where $\delta_{m,n}$ is the Kronecker delta, with $\delta_{mn} = 1$ if $m = n$, $= 0$ otherwise. In particular, in the case that $\beta_n = 0$, (2.18) gives that

$$u_n = \frac{k}{L}\int_{-L/2}^{L/2}\exp(-ik\alpha_n x_0)(f(x_0)-h)\varphi(x_0)dx_0. \quad (2.19)$$

Note that, in deriving (2.18), we have assumed that $h < f_-$, but in fact, as the left and right hand sides of (2.18) are both analytic as functions of h in the whole complex plane, it follows by analytic continuation that (2.18) and its special case (2.19) in fact hold for all (real or complex!) values of h. We note that the expression (2.18) for the Rayleigh coefficients appears to be new.

3   Basis functions and properties of the integral operators

An understanding of the integral operator D relies to a great extent on properties of certain sets of basis functions for the space $X = L^2(-L/2, L/2)$, the set of square integrable functions, a Hilbert space with the inner product

$$\langle \phi, \psi \rangle := \frac{1}{L}\int_{-L/2}^{L/2}\phi(x)\overline{\psi(x)}\,dx$$

and norm $\|\phi\| := \langle \phi, \phi \rangle^{1/2}$. A standard set of basis functions for X is the Fourier basis. Taking into account that we wish to represent quasi-periodic functions, it is natural to shift the standard Fourier basis slightly, using the orthogonal basis functions ϕ_n, defined by

$$\phi_n(x) := \exp(ik\,\alpha_n\,x), \qquad -L/2 \leq x \leq L/2,\ n \in \mathbb{Z}. \quad (3.1)$$

The functions ϕ_n, $n \in \mathbb{Z}$, form a complete orthonormal system in X, so that

$$\langle \phi_n, \phi_m \rangle = \delta_{mn},$$

where δ_{mn} is the Kronecker delta. They appear quite naturally in connection with the integral operator D: from (2.11) and (2.12) we obtain that

$$D\varphi(x) = \sum_{n \in \mathbb{Z}} \sigma_n \langle \varphi, \psi_n \rangle \phi_n(x), \quad (3.2)$$

where

$$\sigma_n := \exp(ik\beta_n(f_- - H)), \quad \psi_n(x) := \exp\left(ik\left[\alpha_n x - \overline{\beta_n}(f(x) - f_-)\right]\right).$$

We point out that $|\psi_n(x)| \leq 1$, with equality at the point x where $f(x) = f_-$, and with equality for all x in the case that $|\alpha_n| \leq 1$. Note also that $|\sigma_n| \leq 1$, with equality for $|\alpha_n| \leq 1$, and that $\sigma_n \sim \exp(-k(f_- - H)|\alpha_n|)$ as $n \to \pm\infty$.

Let D^* denote the L^2-adjoint of the operator D, defined by the equation

$$\langle D\varphi, \psi \rangle = \langle \varphi, D^*\psi \rangle, \qquad \text{for all } \phi, \psi \in X. \tag{3.3}$$

Explicitly,

$$D^*\psi(x) = \frac{1}{L} \int_{-L/2}^{L/2} \overline{K(x_0, x)}\, \psi(x_0) dx_0, \quad -L/2 \leq x \leq L/2,$$

for $\psi \in X$, from which it follows that

$$D^*\psi = \sum_{n\in\mathbb{Z}} \overline{\sigma_n} \langle \psi, \phi_n \rangle\, \psi_n. \tag{3.4}$$

We note that, in the case when Γ is flat, i.e. $f_- = f(x) = f_+$, it holds that $\psi_n = \phi_n$. It then follows from (3.2) and (3.4) that ϕ_n is an eigenfunction of both D and D^*, with eigenvalues σ_n and $\overline{\sigma_n}$, respectively, and that $\{|\sigma_n| : n \in \mathbb{Z}\}$ are the singular values of D and D^*.

The numerical scheme we will propose for solving (2.9) will be a Galerkin scheme, based on expanding the solution of the integral equation in a finite sum of the functions ψ_n, and we will see in a moment that the set $\{\psi_n : n \in \mathbb{Z}\}$ is linearly independent and complete in X. A closely related basis for X is the set of so-called *topological basis functions* [15], $\{\tilde{\psi}_n : n \in \mathbb{Z}\}$, defined by

$$\tilde{\psi}_n(x) := \exp(ik\left[\alpha_n x - \beta_n(f(x) - f_-)\right]).$$

We will also discuss, following [15], using these functions to expand the solution, leading to the SS method of [15].

All three sets of basis functions are related to plane waves. Let $v_n(\mathbf{r}) = \exp(ik\left[\alpha_n x - \beta_n z\right])$, so that v_n is either a downwards propagating plane wave or an evanescent wave decaying exponentially as z decreases. Then ϕ_n is a multiple of v_n restricted to Γ_H, while $\tilde{\psi}_n$ is a multiple of v_n restricted to Γ. The functions ψ_n are also related to restrictions of plane waves on Γ. Let w_n be the plane wave travelling in the opposite direction to v_n, i.e. $w_n(\mathbf{r}) = \exp(ik\left[-\alpha_n x + \beta_n z\right])$. Then ψ_n is a multiple of the complex conjugate of w_n restricted to Γ.

We will now proceed by establishing some crucial properties of the sets of basis functions, and in fact giving some justification to the term basis function. Our first result was proved previously, for the case in which $\beta_n \neq 0$ for all $n \in \mathbb{Z}$, in [23], this paper making precise and completing the earlier argument in [30]. See [6, 36] for related comments on the completeness of plane wave bases in the case of non-periodic surfaces.

Lemma 3.1. *The set of functions $\{\psi_n : n \in \mathbb{Z}\}$ is complete in X.*

Proof. Let $\varphi \in X$ and assume that

$$\langle \varphi, \psi_n \rangle = \frac{1}{L} \int_{-L/2}^{L/2} \varphi(x) \overline{\psi_n(x)} \, dx = 0, \qquad n \in \mathbb{Z}.$$

Define $\phi \in L^2(\Gamma)$ by

$$\phi(\mathbf{r_0}) = \frac{k\varphi(x_0)}{\sqrt{1 + f'(x_0)^2}}, \qquad \mathbf{r_0} = (x_0, z_0) \in \Gamma,$$

choose $h < f_-$, and consider the function v, defined by

$$v(\mathbf{r}) := \int_\Gamma G_{p,h}(\mathbf{r}, \mathbf{r_0}) \, \phi(\mathbf{r_0}) \, ds(\mathbf{r_0}), \quad z > h,$$

where $G_{p,h}$ is the quasi-periodic Dirichlet Green's function for the half-plane $z > h$, given by (2.16). For $h < z < f_-$ it follows from (2.16) that, where $d_n(z)$ is defined by

$$d_n(z) := \begin{cases} \frac{1}{\beta_n}[\exp(-ik\beta_n z) - \exp(ik\beta_n(z - 2h))], & \text{if } \beta_n \neq 0, \\ 2ik(h - z), & \text{if } \beta_n = 0, \end{cases}$$

it holds that

$$v(\mathbf{r}) = \frac{i}{2L} \sum_{n \in \mathbb{Z}} \exp(ik[\alpha_n x + \beta_n f_-]) d_n(z) \int_{-L/2}^{L/2} \overline{\psi_n(x_0)} \, \varphi(x_0) dx_0 = 0.$$

Since solutions of the Helmholtz equation are analytic [10], it follows that $v(\mathbf{r}) = 0$ for $h < z < f(x)$. But v is a single layer potential with L^2 density and so is continuous in $\mathbb{R}^2$, so that $v = 0$ on $\partial\Omega$. Further, $v \in C^2(\Omega)$ and is quasiperiodic and satisfies the Helmholtz equation in Ω and the RERC. Thus, from the fact that Problem 2.1 has only one solution, it follows that $v = 0$ in Ω. However, from jump relations for single-layer potentials with L^2 densities [11], it follows that

$$\phi(\mathbf{r_0}) = \frac{\partial v^+}{\partial n}(\mathbf{r_0}) - \frac{\partial v^-}{\partial n}(\mathbf{r_0}),$$

for almost all $\mathbf{r_0} \in \Gamma$, where the superscripts $+$ and $-$ denote limiting values of the normal derivative as the boundary is approached from below and above, respectively. Thus $\varphi = 0$ in X. This completes the proof. $\qquad\square$

Of course the significance of the completeness of $\{\psi_n : n \in \mathbb{Z}\}$ is that it means that the linear span of $\{\psi_n : n \in \mathbb{Z}\}$ is dense in X, i.e. that every function in X can be

approximated arbitrarily closely by finite linear combinations of the functions ψ_n. In particular, the solution, φ, of the integral equation (2.9) can be approximated in this way.

Since the topological basis functions $\tilde{\psi}_n$ are so closely related to the functions ψ_n, precisely $\overline{\psi_n}$ is the restriction of a plane wave on Γ while $\tilde{\psi}_n$ is the restriction to Γ of the plane wave travelling in exactly the opposite direction, the completeness in X of the set of topological basis functions, $\{\tilde{\psi}_n : n \in \mathbb{Z}\}$, follows by symmetry from Lemma 3.1.

The previous lemma also has immediate consequences for the operator D defined in the previous section.

Lemma 3.2. *The operator D is injective.*

Proof. Assume $\varphi \in X$ and $D\varphi = 0$. Then all the Fourier coefficients of $D\varphi$ vanish. But the series on the right hand side of (3.2) is exactly the Fourier expansion of $D\varphi$. It follows that $\langle \varphi, \psi_n \rangle = 0$ for all $n \in \mathbb{Z}$. Thus, by Lemma 3.1, $\varphi = 0$ and the assertion is proved. $\square$

Lemma 3.3. *The operator D^* is injective.*

Proof. Suppose that $\psi \in X$ and $D^*\psi = 0$ and consider the double layer potential

$$v(\mathbf{r}) := 2 \int_{\Gamma_H} \frac{\partial \tilde{G}_p(\mathbf{r}, \mathbf{r_0})}{\partial z_0} \overline{\psi(x_0)} \, ds(\mathbf{r_0}),$$

for $\mathbf{r} = (x, z)$, $z > H$. Here $\tilde{G}_p$ denotes the quasi-periodic Green's function with θ replaced by $-\theta$, and so α_n replaced by $\tilde{\alpha}_n := -\sin\theta + n\lambda/L$ and β_n replaced with $\tilde{\beta}_n := \sqrt{1 - \tilde{\alpha}_n^2}$. Then, from (2.2), it follows that, for $z > H$,

$$
\begin{aligned}
v(\mathbf{r}) &= \frac{1}{L} \sum_{n \in \mathbb{Z}} \int_{-L/2}^{L/2} \exp(ik[\tilde{\alpha}_n(x - x_0) + \tilde{\beta}_n(z - H)]) \overline{\psi(x_0)} \, dx_0 \\
&= \sum_{n \in \mathbb{Z}} \exp(-ik[\alpha_n x - \beta_n(z - H)]) \langle \phi_n, \psi \rangle,
\end{aligned}
\tag{3.5}
$$

since $\tilde{\alpha}_{-n} = -\alpha_n$ and $\tilde{\beta}_{-n} = \beta_n$. In particular, for $\mathbf{r} = (x, f(x))$, $-L/2 < x < L/2$, we have

$$v(\mathbf{r}) = \sum_{n \in \mathbb{Z}} \exp(-ik[\alpha_n x - \beta_n(f(x) - H)]) \langle \phi_n, \psi \rangle$$

and, comparing with (3.4), we see that this expression is equal to $\overline{D^*\psi(x)}$. Since $D^*\psi = 0$ it follows that $v = 0$ on Γ. However, $v \in C^2(\bar{\Omega})$ and is quasiperiodic and a solution to the Helmholtz equation in Ω, and v satisfies the RERC. Thus, from the uniqueness result for Problem 2.1, it follows that $v = 0$ in Ω and, by analytic continuation, that $v(\mathbf{r}) = 0$ for $z > H$.

Thus and from (3.5), for $z > H$ and $m \in \mathbb{Z}$,

$$
\begin{aligned}
0 = \frac{1}{L} \int_{-L/2}^{L/2} v(\mathbf{r}) \phi_m(x) dx &= \sum_{n \in \mathbb{Z}} \langle \phi_m, \phi_n \rangle \exp(ik\beta_n(z - H)) \langle \phi_n, \psi \rangle \\
&= \exp(ik\beta_m(z - H)) \langle \phi_m, \psi \rangle,
\end{aligned}
$$

since $\{\phi_n : n \in \mathbb{Z}\}$ is orthonormal, so that $\langle \phi_m, \phi_n \rangle = 0$ for $n \neq m$. Thus $\langle \psi, \phi_m \rangle = \overline{\langle \phi_m, \psi \rangle} = 0$ for $m \in \mathbb{Z}$. Since $\{\phi_n : n \in \mathbb{Z}\}$ is complete, it follows that $\psi = 0$. $\qquad\square$

It is a standard result in functional analysis that if D is a bounded linear operator on a Hilbert space, then D is injective if and only if the range of D^*, the adjoint of D, is dense in X. Thus we have the following consequence of Lemmas 3.2 and 3.3.

Corollary 3.1. *The operators D and D^* have dense range.*

Next note the following relationship between ϕ_n and ψ_n which will be key to the effectiveness of the numerical scheme we propose. From (3.4), we have

$$D^* \phi_n = \sum_{m \in \mathbb{Z}} \overline{\sigma_m} \langle \phi_n, \phi_m \rangle \psi_m.$$

But the functions ϕ_n are orthonormal. Thus

$$D^* \phi_n = \overline{\sigma_n} \, \psi_n. \tag{3.6}$$

This relationship together with the injectivity of the operator D^* has the following consequence for the functions ψ_n.

Corollary 3.2. *The set of functions $\{\psi_n : n \in \mathbb{Z}\}$ is linearly independent.*

Proof. Suppose that $N \in \mathbb{N}$, that $j_1, j_2, ..., j_N \in \mathbb{Z}$ and that

$$a_1 \psi_{j_1} + a_2 \psi_{j_2} + ... + a_N \psi_{j_N} = 0$$

for some constants $a_1, ..., a_N$. Then, by (3.6) and the linearity of D^*,

$$D^* (\tilde{a}_1 \phi_{j_1} + \tilde{a}_2 \phi_{j_2} + ... + \tilde{a}_N \phi_{j_N}) = 0,$$

where $\tilde{a}_m := a_m / \overline{\sigma_{j_m}}$. Since D^* is injective from Lemma 3.3, it follows that

$$\tilde{a}_1 \phi_{j_1} + \tilde{a}_2 \phi_{j_2} + ... + \tilde{a}_N \phi_{j_N} = 0.$$

Since $\{\phi_n : n \in \mathbb{Z}\}$ is orthogonal and thus linearly independent, it follows that $\tilde{a}_m = 0$ for $m = 1, ..., N$, so that $a_m = 0$ for $m = 1, ..., N$. $\qquad\square$

4 Galerkin methods for the first kind integral equation

The first kind integral equation (2.10), in common with all first kind integral equations with continuous or weakly singular kernels, is ill-posed, that is the inverse operator D^{-1}, from the range of D onto X, is an unbounded operator. As a consequence, small changes in the function ψ in (2.10) and small changes to the operator D can lead to large changes in the solution φ. Great care has to be taken when solving (2.10) numerically, in particular as D

will be approximated in the discretisation process. It is essential to use a numerical scheme which incorporates regularisation, so that, at the discrete level, D^{-1} is approximated by a bounded operator: see [19,24,26] for a clear exposition of these issues. It is well known that certain Galerkin methods for solving first kind integral equations are self-regularizing, i.e. they have inbuilt regularization properties [24]. We will discuss certain Galerkin methods for solving (2.10) in this section.

Given two finite subspaces X_N, $Y_N \subset X$ of dimension N, the Galerkin method for (2.10) consists of finding a solution $\phi_N \in X_N$ of the variational equation

$$\langle D\varphi_N, \tilde{\psi}\rangle = \langle \psi, \tilde{\psi}\rangle, \qquad \text{for all } \tilde{\psi} \in Y_N. \tag{4.1}$$

Equation (4.1) is in fact equivalent to a finite system of simultaneous equations: given bases $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of X_N and $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ of Y_N, respectively, and setting $\varphi_N = \sum_{n=1}^{N} a_n^{(N)} \mathbf{x}_n$, (4.1) can be reformulated as the system

$$\sum_{n=1}^{N} \langle D\mathbf{x}_n, \mathbf{y}_m\rangle a_n^{(N)} = \langle \psi, \mathbf{y}_m\rangle, \qquad m = 1, \ldots, N. \tag{4.2}$$

Various versions of the Galerkin method can be obtained by specific choices of the subspaces X_N and Y_N and their bases. Noting the representation (3.2) for $D\varphi$, we see that a particularly convenient choice for Y_N is the space spanned by N distinct Fourier modes, $\phi_{j_1}, \ldots, \phi_{j_N}$. With this choice of Y_N it follows from (3.2) and (2.8), and since $\{\phi_n : n \in \mathbb{Z}\}$ is orthonormal, that

$$\langle D\mathbf{x}_n, \mathbf{y}_m\rangle = \sigma_{j_m} \langle \mathbf{x}_n, \psi_{j_m}\rangle \tag{4.3}$$

and

$$\langle \psi, \mathbf{y}_m\rangle = 2i\beta_0 \exp(-ik\beta_0 H)\delta_{0,j_m}, \tag{4.4}$$

where δ_{mn} is the Kronecker delta. In this case the linear system (4.2) is equivalent to

$$\sum_{n=1}^{N} A_{mn}^{(N)} a_n^{(N)} = b_m^{(N)}, \quad m = 1, \ldots, N, \tag{4.5}$$

with $A_{mn}^{(N)} := \langle \mathbf{x}_n, \psi_{j_m}\rangle$, $b_m^{(N)} := 2i\beta_0 \exp(-ik\beta_0 f_-)\delta_{0,j_m}$.

This choice of Y_N is the basis of several formulations investigated in [15]. The SC (spectral-coordinate) implementation in [15] can be derived by using for X_N a finite element space of piecewise constant functions. Precisely, for $n = 1, \ldots, N$ let

$$\mathbf{x}_n(x) := \begin{cases} 1, & \tilde{x}_{n-1} < x < \tilde{x}_n, \\ 0, & \text{otherwise,} \end{cases}$$

where $\tilde{x}_n := -L/2 + nL/N$, $n = 0, 1, \ldots, N$. Then

$$A_{mn}^{(N)} = \frac{1}{L} \int_{\tilde{x}_{n-1}}^{\tilde{x}_n} \exp(-ik[\alpha_{j_m} x - \beta_{j_m} f(x)])dx \, \exp(-ik\beta_{j_m} f_-). \tag{4.6}$$

Approximating the integrals in (4.6) by the midpoint rule, and defining $x_n := (\tilde{x}_{n-1} + \tilde{x}_n)/2 = -L/2 + (n-1/2)L/N$, we obtain from (4.2) the linear system

$$\sum_{n=1}^{N} \exp(-ik[\alpha_{j_m} x_n - \beta_{j_m} f(x_n)]) a_n^{(N)} = 2iN\beta_0 \delta_{0,j_m}, \quad m = 1, ..., N. \tag{4.7}$$

It is this linear system which is solved in the SC method implemented in [15].

A further method, the SS (spectral-spectral) method discussed in [15], is obtained by choosing X_N to be the space spanned by the N topological basis functions $\tilde{\psi}_{j_1}, \ldots, \tilde{\psi}_{j_N}$, so that $\mathbf{x}_n = \tilde{\psi}_{j_n}$. In this case we obtain from (4.5) the linear system

$$\sum_{n=1}^{N} \langle \tilde{\psi}_{j_n}, \psi_{j_n} \rangle a_n^{(N)} = 2i\beta_0 \exp(-ik\beta_0 f_-) \delta_{0,j_m}, \quad m = 1, \ldots, N. \tag{4.8}$$

This system is equivalent to equation (7.5) in [15], in fact is identical to this linear system (to within multiplication by a constant if the origin of the coordinate system is chosen so that $f_- = 0$. It is clearly less straightforward and requires more computation to set up the system matrix for equation (4.8) compared to (4.7), since calculation of the N^2 integrals $\langle \tilde{\psi}_{j_n}, \psi_{j_m} \rangle$ is required, where, explicitly,

$$\langle \tilde{\psi}_{j_n}, \psi_{j_m} \rangle = \frac{1}{L} \int_{-L/2}^{L/2} \exp(2\pi i(j_n - j_m)x/L) \exp(-ik(\beta_{j_n} + \beta_{j_m})(f(x) - f_-)) dx. \tag{4.9}$$

In the numerical results in Section 6 we will approximate these integrals using the trapezoidal rule with M panels, denoting the resulting approximations by $\langle \tilde{\psi}_{j_n}, \psi_{j_m} \rangle_M$. We note that, since the integrand in (4.9) is periodic with period L, this approximation is very rapidly convergent as $M \to \infty$ if f is smooth. Precisely, if $f \in C^l(\mathbb{R})$, for some integer $l \geq 2$, then, from the Euler-Maclaurin expansion [25], it follows that $\langle \tilde{\psi}_{j_n}, \psi_{j_m} \rangle_M = \langle \tilde{\psi}_{j_n}, \psi_{j_m} \rangle + \mathcal{O}(M^{-l})$ as $M \to \infty$.

We now propose a modification of the SS method, which we term the SS* method, based on choosing X_N to be the space spanned by the N functions $\psi_{j_1}, \ldots, \psi_{j_N}$. The significance of this choice is that it follows from (3.6) that $X_N = D^*(Y_N)$. As a consequence the SS* method is an instance of the so-called *Dual Least Squares Method* [24]. As we will prove below, based on the arguments presented in [24], this method combines a similar accuracy of approximation for the subspace X_N to that of the SS method with a much more stable algorithm.

The linear system to be solved in the SS* method is (4.5) with $\mathbf{x}_n = \psi_n$. This linear system can be written as

$$\mathbf{A}_N \mathbf{a}_N = \mathbf{b}_N, \tag{4.10}$$

where $\mathbf{a}_N = (a_1^{(N)}, ..., a_N^{(N)})^T$, $\mathbf{b}_N$ is the column vector with the single non-zero entry $2i\beta_0 \exp(-ik\beta_0 f_-)$ in the mth row, and $\mathbf{A}_N$ is the $N \times N$ matrix with entry

$$A_{mn}^{(N)} = \langle \psi_{j_n}, \psi_{j_m} \rangle = \frac{1}{L} \int_{-L/2}^{L/2} \exp(2\pi i(j_n - j_m)x/L) \exp(-ik(\overline{\beta}_{j_n} - \beta_{j_m})(f(x) - f_-)) dx \tag{4.11}$$

in row m of column n.

Clearly $\mathbf{A}_N$ is Hermitian. $\mathbf{A}_N$ is also positive definite, so that $\mathbf{A}_N$ is invertible. To see this it is convenient to introduce at this point the operator $M_N : \mathbb{C}^N \to X_N$, defined by

$$M_N \mathbf{a} := \sum_{m=1}^{N} a_m \psi_{j_m}, \qquad \text{for } \mathbf{a} = (a_1, ..., a_N)^T \in \mathbb{C}^N. \tag{4.12}$$

Let $M_N^* : X \to \mathbb{C}^N$ denote the adjoint of M_N, defined by

$$\langle M_N \mathbf{a}, \phi \rangle = (\mathbf{a}, M_N^* \phi), \qquad \text{for } \mathbf{a} \in \mathbb{C}^N, \ \phi \in X, \tag{4.13}$$

where $(\cdot, \cdot)$ is the standard scalar product on $\mathbb{C}^N$, defined by

$$(\mathbf{a}, \mathbf{b}) = \sum_{m=1}^{N} a_m \overline{b}_m.$$

Explicitly,

$$M_N^* \phi = (\langle \phi, \psi_{j_1} \rangle, ..., \langle \phi, \psi_{j_N} \rangle)^T, \tag{4.14}$$

from which we see that

$$\mathbf{A}_N \mathbf{a} = M_N^* M_N \mathbf{a}, \qquad \text{for } \mathbf{a} \in \mathbb{C}^N. \tag{4.15}$$

Thus

$$\overline{\mathbf{a}}^T \mathbf{A}_N \mathbf{a} = \overline{(\mathbf{a}, M_N^* M_N \mathbf{a})} = \langle M_N \mathbf{a}, M_N \mathbf{a} \rangle = ||M_N \mathbf{a}||^2 \geq 0, \tag{4.16}$$

with equality only if $\mathbf{a} = \mathbf{0}$, as $M_N \mathbf{a} = 0$ only if $\mathbf{a} = \mathbf{0}$ since the ψ_n are linearly independent by Corollary 3.2.

That $\mathbf{A}_N$ is invertible, so that the Galerkin solution is well-defined for every N and every selection of the mode numbers $j_1, ..., j_N$, is a first advantage of the SS* method. In operator terms, this means that there is a well-defined Galerkin method solution operator, $R_N : X \to X_N$, which maps $\psi \in X$ onto the solution, φ_N, of equation (4.1). This operator is bounded: we shall estimate its norm in Lemma 4.1 below. Further, we shall see shortly that, provided the sequence of spaces $X_1, X_2, ...$ is chosen in a natural way, the family of operators R_N is a regularisation strategy for the first kind equation (2.10), in the sense of [24], meaning that each R_N is bounded and $R_N D\psi \to \psi$ as $N \to \infty$ for every $\psi \in X$. Thus, as $N \to \infty$, the bounded operator R_N is an increasingly accurate approximation to the unbounded inverse operator D^{-1}. An attraction of this particular regularisation strategy is that an explicit error estimate holds for the SS* method, contained in the following lemma.

Lemma 4.1. *There holds* $\|R_N D\| \leq 1$ *and*

$$\|\varphi_N - \varphi\| \leq 2 \min_{\tilde{\varphi} \in X_N} \|\varphi - \tilde{\varphi}\|,$$

where $\varphi_N = R_N \psi \in X_N$ *is the numerical solution computed using the SS* method. Further,*

$$||R_N|| \leq \tau_N := \sup\{||\tilde{\psi}|| : \tilde{\psi} \in Y_N, ||D^*\tilde{\psi}|| = 1\}. \tag{4.17}$$

Proof. We follow the arguments in the proofs of [24, Theorems 3.7 and 3.11]. By (3.3) and since $\psi = D\varphi$, the variational equation (4.1) can be written in the form

$$\langle \varphi_N, D^*\tilde{\psi} \rangle = \langle \varphi, D^*\tilde{\psi} \rangle \qquad \text{for all } \tilde{\psi} \in Y_N. \tag{4.18}$$

As $X_N = D^*(Y_N)$, there exists $u_N \in Y_N$ such that $\varphi_N = D^*u_N$. Setting $\tilde{\psi} = u_N$ in (4.18) we obtain

$$||\varphi_N||^2 = \langle \varphi_N, \varphi_N \rangle = \langle \varphi, \varphi_N \rangle \leq ||\varphi|| \, ||\varphi_N||.$$

As $\varphi_N = R_N D\varphi$, it follows that $||R_N D|| \leq 1$.

Now if, in equation (4.1), $\psi = D\tilde{\varphi}$ for some $\tilde{\varphi} \in X_N$ then, by inspection, we see that the unique solution of (4.1) is $\varphi = \tilde{\varphi}$. Thus $R_N D\tilde{\varphi} = \tilde{\varphi}$ for $\tilde{\varphi} \in X_N$ and thus

$$\varphi_N - \varphi = (R_N D - I)\varphi = (R_N D - I)(\varphi - \tilde{\varphi}) \qquad \text{for all } \tilde{\varphi} \in X_N,$$

and hence

$$||\varphi_N - \varphi|| \leq 2 \, ||\varphi - \tilde{\varphi}|| \qquad \text{for all } \tilde{\varphi} \in X_N.$$

To see (4.17), note that if $\tilde{\psi} \in Y_N$ then $||D^*\hat{\psi}|| = 1$, where $\hat{\psi} := \tilde{\psi}/||D^*\tilde{\psi}||$. Since $\hat{\psi} \in Y_N$ it holds that $||\hat{\psi}|| \leq \tau_N$ so that $||\tilde{\psi}|| \leq \tau_N ||D^*\tilde{\psi}||$. Thus and by (4.18) and (3.3),

$$||\varphi_N||^2 = \langle \varphi, D^*u_N \rangle = \langle \psi, u_N \rangle \leq ||\psi|| \, ||u_N|| \leq \tau_N ||\psi|| \, ||\varphi_N||.$$

Thus $||\varphi_N|| \leq \tau_N ||\psi||$ for every $\psi \in X$ so that (4.17) holds.

$\square$

Obviously, we have the following corollary to Lemma 3.1 that, together with the previous lemma, yields convergence of the SS* method.

Corollary 4.1. *Provided the sequence of subspaces $X_1, X_2, ...$ is chosen so that, for every $n \in \mathbb{Z}$, $\varphi_n \in X_N$ for all sufficiently large N, then*

$$\min_{\tilde{\varphi} \in X_N} ||\varphi - \tilde{\varphi}|| \longrightarrow 0, \qquad \text{as } N \to \infty,$$

for every $\varphi \in X$.

In the numerical results in Section 6 we will choose $X_N = \{\psi_{I_N}, \psi_{I_N+1}, ..., \psi_{J_N}\}$, with $J_N := I_N + N - 1$ and I_N chosen so that $I_N \to -\infty$ and $J_N \to +\infty$ as $N \to \infty$. This satisfies the conditions of Corollary 4.1.

The error estimate in Lemma 4.1 shows that the solution generated by the SS* method has an accuracy very close to that achieved by the best approximation from the subspace X_N. Unfortunately, this accuracy is not always achieved in practice due to the ill-conditioned nature of the equation being solved, reflected in large values for the norms, $||R_N||$ and $||\mathbf{A}_N^{-1}||$, of the inverse operator and inverse matrix, respectively, involved in the SS* method, in the limit $N \to \infty$. This ill-conditioning leads to amplification of errors introduced in solving the linear system (4.10), the main source of error being numerical quadrature error inherent in the computation of the coefficients $A_{mn}^{(N)}$ of the matrix $\mathbf{A}_N$. In the numerical results in Section 6 we approximate these coefficients using the same quadrature rule as for the SS method, namely the trapezoidal rule with M panels, denoting the resulting approximation to $A_{mn}^{(N)} = \langle \psi_{j_n}, \psi_{j_m} \rangle$ by $\langle \psi_{j_n}, \psi_{j_m} \rangle_M$. As for the SS method, this approximation is very rapidly convergent if f is smooth. Precisely, if $f \in C^l(\mathbb{R})$, for some integer $l \geq 2$, then, from the Euler-Maclaurin expansion it follows that

$$\langle \psi_{j_n}, \psi_{j_m} \rangle_M = \langle \psi_{j_n}, \psi_{j_m} \rangle + \mathcal{O}(M^{-l})$$

as $M \to \infty$.

Due to this numerical quadrature error (and the additional small effects of rounding errors) we solve a perturbed version of equation (4.10), namely

$$\mathbf{A}_N^\delta \mathbf{a}_N^\delta = \mathbf{b}_N, \tag{4.19}$$

where $||\mathbf{A}_N^\delta - \mathbf{A}_N|| \leq \delta$, for some small $\delta > 0$. Standard matrix perturbation analysis [25] yields that, provided $\delta \, ||\mathbf{A}_N^{-1}|| < 1$, $\mathbf{A}_N^\delta$ is invertible, with

$$\left\| (\mathbf{A}_N^\delta)^{-1} \right\| \leq \frac{||\mathbf{A}_N^{-1}||}{1 - \delta \, ||\mathbf{A}_N^{-1}||}. \tag{4.20}$$

To analyse the effect of this inexact calculation, we introduce the operator $Q_N : X \to \mathbb{C}^N$, defined by $Q_N \phi := (\langle \phi, \phi_{j_1} \rangle, ..., \langle \phi, \phi_{j_N} \rangle)^T$. In terms of the operators M_N and Q_N, the matrix $\mathbf{A}_N$, and the diagonal matrix $\mathbf{D}_N := \mathrm{diag}(\sigma_{j_1}, ..., \sigma_{j_N})$, R_N can be expressed explicitly as

$$R_N = M_N \mathbf{A}_N^{-1} \mathbf{D}_N^{-1} Q_N.$$

We note that, for $\phi \in X$,

$$||Q_N \phi||^2 = \sum_{m=1}^N |\langle \phi, \phi_{j_m} \rangle|^2 \leq \sum_{m \in \mathbb{Z}} |\langle \phi, \phi_m \rangle|^2 = ||\phi||^2,$$

since $\{\phi_m : m \in \mathbb{Z}\}$ is complete and orthonormal. Thus $||Q_N \phi|| \leq ||\phi||$, for all $\phi \in X$, with equality if $\phi \in Y_N$, so that $||Q_N|| = 1$. We see also that

$$
\begin{aligned}
||R_N|| &= \sup_{\phi \in X, ||\phi||=1} ||M_N \mathbf{A}_N^{-1} \mathbf{D}_N^{-1} Q_N \phi|| \\
&= \sup_{\phi \in Y_N, ||\phi||=1} ||M_N \mathbf{A}_N^{-1} \mathbf{D}_N^{-1} Q_N \phi|| \\
&= \sup_{\mathbf{a} \in \mathbb{C}^N, ||\mathbf{a}||=1} ||M_N \mathbf{A}_N^{-1} \mathbf{D}_N^{-1} \mathbf{a}|| = ||M_N \mathbf{A}_N^{-1} \mathbf{D}_N^{-1}||.
\end{aligned}
$$

Further,

$$||\mathbf{D}_N|| = \max_{1 \le m \le N} |\sigma_{j_m}| = 1.$$

Thus, and by Lemma 4.1,

$$||M_N \mathbf{A}^{-1}|| \le ||M_N \mathbf{A}_N^{-1} \mathbf{D}_N^{-1}|| \, ||\mathbf{D}_N|| = ||R_N|| \le \tau_N,$$

where τ_N is as defined in Lemma 4.1.

In terms of $\mathbf{a}_N$, the solution of the linear system (4.10), the SS* method solution is

$$\varphi_N = M_N \mathbf{a}_N = \sum_{m=1}^{N} a_m^{(N)} \psi_{j_m}.$$

Let $\varphi_N^\delta = M_N \mathbf{a}_N^\delta$ be the approximation to φ_N calculated if (4.19) is solved in place of (4.10). Then, where the residual $\mathbf{e}_N := \mathbf{b}_N - \mathbf{A}_N \mathbf{a}_N^\delta$ measures by how much $\mathbf{a}_N^\delta$ fails to satisfy equation (4.10), it holds that

$$\varphi_N - \varphi_N^\delta = M_N \mathbf{A}_N^{-1} \mathbf{e}_N.$$

Taking norms in this equation we find that

$$||\varphi_N - \varphi_N^\delta|| \le ||M_N \mathbf{A}_N^{-1}|| \, ||\mathbf{e}_N|| \le \tau_N ||\mathbf{e}_N||. \tag{4.21}$$

The residual $\mathbf{e}_N$ can be bounded in terms of δ and the norm of $\mathbf{A}_N^{-1}$, since $\mathbf{e}_N = (\mathbf{A}_N^\delta - \mathbf{A}_N)(\mathbf{A}_N^\delta)^{-1} \mathbf{b}_N$ and $||\mathbf{b}_N|| = 2\beta_0$. Applying (4.20) it follows that

$$||\mathbf{e}_N|| \le ||\mathbf{A}_N^\delta - \mathbf{A}_N|| \, ||(\mathbf{A}_N^\delta)^{-1}|| \, ||\mathbf{b}_N|| \le \frac{2\delta\beta_0 \, ||\mathbf{A}_N^{-1}||}{1 - \delta \, ||\mathbf{A}_N^{-1}||}. \tag{4.22}$$

In the next two lemmas we obtain an upper bound for τ_N and explore further the relationship between R_N and $\mathbf{A}_N$. Lemma 4.3 shows that, unfortunately, $\mathbf{A}_N$ is badly conditioned as $N \to \infty$, unless the surface is flat in which case, as remarked earlier, $\psi_m = \phi_m$, $m \in \mathbb{Z}$, so that $\{\psi_m : m \in \mathbb{Z}\}$ is orthonormal, $\mathbf{A}_N$ is an identity matrix, and $\text{cond}\, \mathbf{A}_N = 1$, where $\text{cond}\, \mathbf{A}_N := ||\mathbf{A}_N|| \, ||\mathbf{A}_N^{-1}||$ denotes the condition number of the matrix $\mathbf{A}_N$.

Lemma 4.2. *For every $\epsilon > 0$ there exists $C_\epsilon' > 0$, depending only on f, k, and ϵ, such that*

$$||R_N|| \le \tau_N \le C_\epsilon' \exp(\beta_N^* k[f_+ - H + \epsilon]), \tag{4.23}$$

where $\beta_N^* := \max_{1 \le m \le N} \Im \beta_{j_m}$. In the special case that $f_+ = f_-$ (the surface is flat) then

$$||R_N|| \le \tau_N = \exp(\beta_N^* k[f_- - H]). \tag{4.24}$$

Proof. From Lemma 4.1 we have that $||R_N|| \leq \tau_N$. Recalling the definition of τ_N in Lemma 4.1, suppose that $\tilde{\psi} \in Y_N$ with $||D^*\tilde{\psi}|| = 1$. Using the notation in the proof of Lemma 3.3, consider the double layer potential

$$v(\mathbf{r}) := 2 \int_{\Gamma_H} \frac{\partial \tilde{G}_p(\mathbf{r}, \mathbf{r_0})}{\partial z_0} \overline{\tilde{\psi}(x_0)} \, ds(\mathbf{r_0}).$$

As was shown in the proof of Lemma 3.3, we have $v((x, f(x))) = \overline{D^*\tilde{\psi}(x)}$ and hence

$$\frac{1}{L} \int_{-L/2}^{L/2} |v((x, f(x)))|^2 dx = ||D^*\tilde{\psi}||^2 = 1. \tag{4.25}$$

As $\tilde{\psi} \in Y_N$, we have the representation $\tilde{\psi} = \sum_{m=1}^{N} \gamma_m \phi_{j_m}$, for some constants $\gamma_1, ..., \gamma_N$. Thus, and from (3.5),

$$v(\mathbf{r}) = \sum_{m=1}^{N} \overline{\gamma_m} \exp(-ik[\alpha_{j_m} x - \beta_{j_m}(z - H)]).$$

On the other hand, v satisfies Problem 2.1, except with different boundary data on Γ and with θ replaced by $-\theta$. From the well-posedness of this problem, we have the estimate that, for $\epsilon > 0$ and $h = f_+ + \epsilon$,

$$\int_{\Gamma_h} |v(\mathbf{r})|^2 dx \leq C_\epsilon \int_{\Gamma} |v(\mathbf{r})|^2 ds(\mathbf{r}), \tag{4.26}$$

where C_ϵ is a constant which depends only on f, k, and ϵ. But

$$\frac{1}{L} \int_{\Gamma_h} |v(\mathbf{r})|^2 dx = \sum_{m=1}^{N} |\gamma_m|^2 \exp(-2k(h - H)\Im\beta_{j_m}) \tag{4.27}$$

and, defining $C'_\epsilon := C_\epsilon \max \sqrt{1 + (f'(x))^2}$,

$$C_\epsilon \int_{\Gamma} |v(\mathbf{r})|^2 ds(\mathbf{r}) \leq C'_\epsilon \int_{-L/2}^{L/2} |v((x, f(x)))|^2 dx = LC'_\epsilon, \tag{4.28}$$

by (4.25). Now

$$||\tilde{\psi}||^2 = \sum_{m=1}^{N} |\gamma_m|^2 \leq \sum_{m=1}^{N} |\gamma_m|^2 \exp(2k(h - H)(\beta_N^* - \Im\beta_{j_m})).$$

Thus, and using (4.26)-(4.28),

$$||\tilde{\psi}|| \leq C'_\epsilon \exp(k(f_+ + \epsilon - H)\beta_N^*).$$

Since this holds for all $\tilde{\psi} \in Y_N$ with $||D^*\tilde{\psi}|| = 1$ we have shown the bound (4.23).

In the case $f_+ = f_-$, it holds that $\Gamma = \Gamma_h$ with $h = f_+ = f_-$. From equations (4.25) and (4.27) we deduce that

$$||\tilde{\psi}||^2 \le \sum_{m=1}^{N} |\gamma_m|^2 \exp(2k(f_- - H)(\beta_N^* - \Im\beta_{j_m})) = \exp(2k(f_- - H)\beta_N^*), \qquad (4.29)$$

and the inequality in (4.29) becomes an equality if $\gamma_m = 0$ for $m \ne m^*$, with $m^* \in \{1, ..., N\}$ chosen so that $\Im\beta_{j_{m^*}} \ge \Im\beta_{j_m}$ for $m = 1, ..., N$. $\qquad\square$

Lemma 4.3.
$$\kappa_N := \inf_{H < f_-} ||R_N|| = ||M_N \mathbf{A}_N^{-1}|| = ||\mathbf{A}_N^{-1}||^{1/2}. \qquad (4.30)$$

If the conditions of Corollary 4.1 are satisfied and $f_+ > f_-$, so that the surface is not flat, then $\kappa_N \to \infty$ and $\operatorname{cond}\mathbf{A}_N \to \infty$ as $N \to \infty$.

Proof. We have shown above that $||M_N \mathbf{A}_N^{-1}|| \le ||R_N||$ for all $H < f_-$. We have also shown that $||R_N|| = ||M_N \mathbf{A}_N^{-1} \mathbf{D}_N^{-1}|| \le ||M_N \mathbf{A}_N^{-1}|| \, ||\mathbf{D}_N^{-1}||$. Now

$$||\mathbf{D}_N^{-1}|| = \max_{1 \le m \le N} |\sigma_{j_m}|^{-1} = \max_{1 \le m \le N} \exp(k(f_- - H)\Im\beta_{j_m}).$$

Thus $\inf_{H < f_-} ||\mathbf{D}_N^{-1}|| = 1$ so that $||M_N \mathbf{A}_N^{-1}|| = \kappa_N$.

To see the rest of (4.30), note that, for $\mathbf{a} \in \mathbb{C}^N$, using (4.13) and (4.15),

$$\begin{aligned}
||M_N \mathbf{A}_N^{-1} \mathbf{a}||^2 &= \langle M_N \mathbf{A}_N^{-1}\mathbf{a}, M_N \mathbf{A}_N^{-1}\mathbf{a}\rangle \\
&= (\mathbf{A}_N^{-1}\mathbf{a}, M_N^* M_N \mathbf{A}_N^{-1}\mathbf{a}) \\
&= (\mathbf{A}_N^{-1}\mathbf{a}, \mathbf{a}) \\
&\le ||\mathbf{a}||^2 \, ||\mathbf{A}_N^{-1}||.
\end{aligned}$$

Thus $||M_N \mathbf{A}_N^{-1}\mathbf{a}|| \le ||\mathbf{a}|| \, ||\mathbf{A}_N^{-1}||^{1/2}$, so that $||M_N \mathbf{A}_N^{-1}|| \le ||\mathbf{A}_N^{-1}||^{1/2}$. Since $\mathbf{A}_N$ is Hermitian and positive definite, $||\mathbf{A}_N^{-1}||$ is the smallest eigenvalue of $\mathbf{A}_N$. Choosing $\mathbf{a}$ to be the associated eigenvector, so that $\mathbf{A}_N^{-1}\mathbf{a} = ||\mathbf{A}_N^{-1}||\mathbf{a}$, it follows that $||M_N \mathbf{A}_N^{-1}\mathbf{a}||^2 = (\mathbf{A}_N^{-1}\mathbf{a}, \mathbf{a}) = ||\mathbf{a}||^2 \, ||\mathbf{A}_N^{-1}||$. Thus

$$||M_N \mathbf{A}_N^{-1}|| = ||\mathbf{A}_N^{-1}||^{1/2}.$$

From Lemma 4.1 we have that, for every $\varphi \in X$,

$$||R_N|| \, ||D\varphi|| \ge ||R_N D\varphi|| \ge ||\varphi|| - 2\min_{\tilde{\varphi} \in X_N} ||\varphi - \tilde{\varphi}||.$$

If the conditions of Corollary 4.1 are satisfied it holds that $2\min_{\tilde{\varphi} \in X_N} ||\varphi - \tilde{\varphi}|| \le \frac{1}{2}||\varphi||$ for all sufficiently large N. Thus, for every non-zero $\varphi \in X$, there exists N_0 such that

$$||R_N|| \ge \frac{||\varphi||}{2||D\varphi||} \qquad (4.31)$$

for all $N \geq N_0$ and $H < f_-$.

If also $f_+ > f_-$, then we can define $\chi \in X$ with $||\chi|| \neq 0$ by $\chi(x) = 1$, if $f(x) > (f_+ + f_-)/2$, $\chi(x) = 0$, otherwise. For $m = 1, 2, ...$ let $\chi_m(x) := \chi(x)e^{imx}$. Then $||\chi_m|| = ||\chi||$ and

$$D\chi_m = \sum_{n \in \mathbb{Z}} \sigma_n \langle \chi_m, \psi_n \rangle \phi_n$$

so that

$$||D\chi_m||^2 = \sum_{n \in \mathbb{Z}} |\sigma_n|^2 |\langle \chi_m, \psi_n \rangle|^2 \leq \sum_{n \in \mathbb{Z}} |\langle \chi_m, \psi_n \rangle|^2, \tag{4.32}$$

since $|\sigma_n| \leq 1$. Note that

$$|\langle \chi_m, \psi_n \rangle| = \frac{1}{L} \int_{-L/2}^{L/2} \chi(x) \exp(-k(f(x) - f_-)\Im\beta_n)\, dx \leq \exp(-k(f_+ - f_-)\Im\beta_n/2),$$

so that the series (4.32) converges, uniformly in m. But also, by the Riemann-Lebesgue lemma, $\langle \chi_m, \psi_n \rangle \to 0$ as $m \to \infty$, for every $n \in \mathbb{Z}$. Thus

$$\sum_{n \in \mathbb{Z}} |\langle \chi_m, \psi_n \rangle|^2 \to 0$$

as $m \to \infty$. But, combining (4.31) and (4.32), we have that, for every $m \in \mathbb{N}$, it holds for all sufficiently large N that

$$\kappa_N^2 \geq ||\chi||^2 \left(4 \sum_{n \in \mathbb{Z}} |\langle \chi_m, \psi_n \rangle|^2 \right)^{-1}.$$

Thus $\kappa_N \to \infty$ as $N \to \infty$.

To finish the proof note that $\psi_0 \in X_N$ for all sufficiently large N. But if $\psi_0 \in X_N$ then $j_m = 0$ for some $m \in \{1, ..., N\}$, and then $A_{mm}^{(N)} = \langle \psi_0, \psi_0 \rangle = 1$ and $||\mathbf{A}_N \mathbf{a}|| \geq 1$, where $\mathbf{a}$ is the column vector with a 1 in row m as the only non-zero entry, so that $||\mathbf{a}|| = 1$. Thus $||\mathbf{A}_N|| \geq 1$ for all sufficiently large N, so that $\text{cond}\,\mathbf{A}_N = ||\mathbf{A}_N|| \kappa_N^2 \geq \kappa_N^2$ for all sufficiently large N. Thus $\text{cond}\,\mathbf{A}_N \to \infty$ as $N \to \infty$. $\qquad\square$

To finish this section we summarise, in a final theorem, the main results we have obtained in respect of the accuracy and convergence of the SS* method.

Theorem 4.1. *For every $\epsilon > 0$ there exists $C'_\epsilon > 0$, depending only on k, f, and ϵ, such that*

$$\kappa_N = ||\mathbf{A}_N^{-1}||^{1/2} \leq C'_\epsilon \exp(k(f_+ - f_- + \epsilon)\beta_N^*), \tag{4.33}$$

where $\beta_N^* := \max_{1 \leq m \leq N} \Im\beta_{j_m}$. If $||\mathbf{A}_N - \mathbf{A}_N^\delta|| \leq \delta$ with $\delta\kappa_N^2 < 1$, then $\mathbf{A}_N^\delta$ is invertible so that the linear system (4.19) has a unique solution, $\mathbf{a}_N^\delta$. Further, the approximate SS* method solution, $\varphi_N^\delta = M_N \mathbf{a}_N^\delta$, satisfies the error estimate

$$\left\| \varphi - \varphi_N^\delta \right\| \leq 2 \min_{\tilde{\varphi} \in X_N} ||\varphi - \tilde{\varphi}|| + \frac{2\delta\beta_0\kappa_N^3}{1 - \delta\kappa_N^2}. \tag{4.34}$$

If the conditions of Corollary 4.1 are also satisfied, then $\min_{\tilde{\varphi} \in X_N} \|\varphi - \tilde{\varphi}\| \to 0$ as $N \to \infty$ and, provided $f_+ > f_-$, $\kappa_N \to \infty$ as $N \to \infty$.

Proof. We have shown above that $\kappa_N^2 = \|\mathbf{A}_N^{-1}\|$ and that $\mathbf{A}_N^\delta$ is invertible provided $\delta\|\mathbf{A}_N^{-1}\| < 1$. Clearly $\|\varphi - \varphi_N^\delta\| \le \|\varphi - \varphi_N\| + \|\varphi_N - \varphi_N^\delta\|$, and $\|\varphi - \varphi_N\|$ is bounded in Lemma 4.1 while, from (4.21), $\|\varphi_N - \varphi_N^\delta\| \le \kappa_N\|\mathbf{e}_N\|$ and then, from (4.22) and Lemma 4.3, the bound (4.34) follows. The remainder of the results come from Corollary 4.1 and Lemma 4.2. □

We note that the bounds (4.33) and (4.34) show that a small amount of error in computing the matrix $\mathbf{A}_N$ will not have a significant effect on accuracy provided $\beta_N^* k(f_+ - f_-)$ is not large, in particular if $\beta_N^* = 0$ (X_N contains only propagating modes with $|\alpha_n| \le 1$).

5 The Rayleigh expansion and least squares methods

The SS* method we have proposed in Section 4 has close connections with methods for solving the diffraction grating problem based on the Rayleigh expansion (2.1). A question which has generated considerable debate over the years is whether the Rayleigh hypothesis holds. (The Rayleigh hypothesis is the supposition that the expansion (2.1) is valid not just in the half-plane above the diffraction grating but throughout Ω and on its boundary $\partial\Omega$.) If the Rayleigh hypothesis holds then the Rayleigh coefficients can be determined directly from the requirement that $u(\mathbf{r}) = -u^i(\mathbf{r})$ for $\mathbf{r}$ on $\partial\Omega$.

So as to relate this method more easily to the method of Section 4, it is convenient to consider the case when the angle of incidence is $-\theta$ rather than θ, so that $u^i(\mathbf{r}) = \exp(ik[-x\sin\theta - z\cos\theta])$. Then, if the Rayleigh hypothesis holds,

$$u(\mathbf{r}) = \sum_{n\in\mathbb{Z}} u_n w_n(\mathbf{r}), \qquad \text{for } \mathbf{r} \in \overline{\Omega}, \tag{5.1}$$

where, as defined earlier, $w_n(\mathbf{r}) = \exp(ik[-\alpha_n x + \beta_n z])$ and the Rayleigh coefficients u_n can be obtained from the requirement that $u(\mathbf{r}) = -u^i(\mathbf{r})$ for $\mathbf{r}$ on $\partial\Omega$.

When the Rayleigh expansion is used for computation, the sum (5.1) is truncated to a finite sum and a linear system is formed to find the finite number of coefficients u_n. This linear system is obtained by requiring that $u(\mathbf{r}) = -u^i(\mathbf{r})$ holds at a number of points equal to the number of unknown coefficients, in which case the method is termed the point collocation method, or by requiring that $u(\mathbf{r}) = -u^i(\mathbf{r})$ hold in a least squares sense. The following is an implementation of the least squares method that has close connections to the method of Section 4.

Note first that (5.1) is equivalent to

$$u(\mathbf{r}) = \sum_{n\in\mathbb{Z}} c_n \hat{w}_n(\mathbf{r}), \qquad \text{for } \mathbf{r} \in \overline{\Omega}, \tag{5.2}$$

where $\hat{w}_n(\mathbf{r}) := \exp(ik\left[-\alpha_n x + \beta_n(z - f_-)\right])$ and $c_n := u_n \exp(ik\,\beta_n f_-)$. The plane waves $\hat{w}_n$ are normalised so that the maximum value of $|\hat{w}_n(\mathbf{r})|$ on $\overline{\Omega}$ is 1 for each n; of course $\hat{w}_n(\mathbf{r}) = w_n(\mathbf{r})$ if the axes are chosen so that $f_- = 0$. Choosing, as for the Galerkin scheme of Section 4, N distinct integers $j_1, ..., j_N$, the least squares method we will consider is to approximate $u(\mathbf{r})$ by the finite linear combination of plane waves,

$$u_N(\mathbf{r}) = \sum_{n=1}^{N} c_n^{(N)} \hat{w}_{j_n}(\mathbf{r}), \tag{5.3}$$

and choose the coefficients $c_n^{(N)}$ so as to minimise

$$E_N := \int_{-L/2}^{L/2} \left| u_N((x, f(x))) + u^i((x, f(x))) \right|^2 dx.$$

Now note that $\hat{w}_n((x, f(x))) = \overline{\psi}_n(x)$. Let $\mathbf{c}_N = (c_1^{(N)}, ..., c_N^{(N)})^T$ and define

$$\chi(x) := u^i((x, f(x))) = \exp(ik[-x \sin\theta - f(x)\cos\theta]).$$

In terms of these notations and the operator $M_N : \mathbb{C}^N \to X_N$, defined by (4.12), we have that

$$E_N = ||M_N \overline{\mathbf{c}}_N - \overline{\chi}||^2.$$

Standard calculations (e.g. [24, Lemma 2.10]) yield that $\mathbf{c}_N$ minimises E_N if and only if $\mathbf{c}_N$ satisfies the normal equations

$$M_N^* M_N \overline{\mathbf{c}}_N = M_N^* \overline{\chi}. \tag{5.4}$$

Here $M_N^* : X_N \to \mathbb{C}^N$ is the adjoint of M_N, defined by (4.13) and given explicitly by (4.14). From (4.15) we have that (5.4) is the linear system

$$\mathbf{A}_N \overline{\mathbf{c}}_N = \mathbf{d}_N, \tag{5.5}$$

where

$$\mathbf{d}_N := M_N^* \overline{\chi} = (\langle \overline{\chi}, \psi_{j_1}\rangle, ..., \langle \overline{\chi}, \psi_{j_N}\rangle)^T,$$

and the elements of $\mathbf{A}_N$ are given explicitly by (4.11).

It was shown in Section 4 that $\mathbf{A}_N$ is Hermitian and positive definite, so that $\mathbf{A}_N$ is invertible. Thus the least squares method is well-defined: we solve (5.5) to obtain the vector $\mathbf{c}_N$ and then the scattered field is given approximately by (5.3).

For most grating profiles the Rayleigh hypothesis is not valid. This is a very crude statement and we refer the reader to [14, 28] for details. However, whether or not the Rayleigh hypothesis holds, it is known that the least squares method is convergent. Previous demonstrations of this fact (e.g. [28, 30]) are incomplete, in particular excluding the case when $\beta_n = 0$ for some n. We include here a proof, based on the results of Section 3, which is valid in all cases. In this theorem, as in Section 4, X_N denotes the linear space spanned by $\psi_{j_1}, ..., \psi_{j_N}$.

Theorem 5.1. *Suppose that the conditions of Corollary 4.1 are satisfied. Then, for every* $\mathbf{r} \in \Omega$, $u_N(\mathbf{r}) \to u(\mathbf{r})$ *as* $N \to \infty$, *and this convergence is uniform in* $\mathbf{r}$, *for* $\mathbf{r} \in S_\epsilon :=$ $\{(x,z) : z > f(x) + \epsilon\}$, *for every* $\epsilon > 0$. *Further, in the half-space* $z > f_+$, *above* $\partial\Omega$, $u(\mathbf{r})$ *is given by* (5.2), *with* $c_n = \lim_{N \to \infty} \tilde{c}_n^{(N)}$, *for every* $n \in \mathbb{Z}$, *where* $\tilde{c}_n^{(N)}$ *denotes the coefficient of* $\hat{w}_n(\mathbf{r})$ *in* (5.3).

Proof. Let $e_N(\mathbf{r}) := u(\mathbf{r}) - u_N(\mathbf{r})$. Then e_N satisfies Problem 2.1, except with different boundary data on Γ and with θ replaced by $-\theta$. From the well-posedness of the problem, we have that, for every $\epsilon > 0$,

$$|e_N(\mathbf{r})| \le C_\epsilon \int_\Gamma |e_N(\mathbf{r}_0)|^2 ds(\mathbf{r}_0), \quad \text{for } \mathbf{r} \in S_\epsilon,$$

where the constant C_ϵ depends only on ϵ, k, and f. But

$$\int_\Gamma |e_N(\mathbf{r}_0)|^2 ds(\mathbf{r}_0) \le \max_{x \in \mathbb{R}} \sqrt{1 + (f'(x))^2} \; \tilde{E}_N$$

where

$$\tilde{E}_N := \min_{\mathbf{c}_N \in \mathbb{C}^N} E_N = \min_{\tilde{\psi} \in X_N} ||\tilde{\psi} - \overline{\chi}||.$$

But, from Lemma 3.1, it follows that $\min_{\tilde{\psi} \in X_N} ||\tilde{\psi} - \overline{\chi}|| \to 0$ as $N \to \infty$. Thus $|u_N(\mathbf{r}) - u(\mathbf{r})| = |e_N(\mathbf{r})| \to 0$ as $N \to \infty$, uniformly on S_ϵ. From this and that

$$c_n - \tilde{c}_n^{(N)} = \frac{1}{L} \exp(-ik\beta_n(h - f_-)) \int_{\Gamma_h} e_N(\mathbf{r}) \exp(ik\alpha_n x) ds(\mathbf{r}),$$

for every $h > f_+$, it follows that $\tilde{c}_n^{(N)} \to c_n$ as $N \to \infty$. $\square$

Although the least squares method is, by the above result, theoretically convergent, computations indicate that it does not converge for all gratings due to problems of ill-conditioning [28]. The results of Section 4 provide, for the first time, a quantification of this ill-conditioning and allow us to estimate the effect of errors in solving (5.5) on the accuracy of the computed solution $\mathbf{c}_N$.

As in Section 4, we introduce errors when we estimate the coefficients of $\mathbf{A}_N$ and $\mathbf{d}_N$ by numerical integration. Due to this numerical quadrature error (and additional small rounding errors) we solve a perturbed version of equation (5.5), namely

$$\mathbf{A}_N^\delta \overline{\mathbf{c}}_N^\delta = \mathbf{d}_N^\delta. \tag{5.6}$$

We assume that $||\mathbf{A}_N - \mathbf{A}_N^\delta|| \le \delta_1$ and $||\mathbf{d}_N - \mathbf{d}_N^\delta|| \le \delta_2$, for some small $\delta_1, \delta_2 > 0$. As discussed in Section 4, $\mathbf{A}_N^\delta$ is invertible if $\delta_1 ||\mathbf{A}_N||^{-1} < 1$. If this condition holds we have further that

$$\overline{\mathbf{c}}_N - \mathbf{c}_N^\delta = (\mathbf{A}_N^\delta)^{-1} \mathbf{e}_N \tag{5.7}$$

with

$$
\begin{aligned}
\mathbf{e}_N &= \mathbf{A}_N^{\delta}\overline{\mathbf{c}}_N - \mathbf{d}_N^{\delta} \\
&= (\mathbf{A}_N^{\delta} - \mathbf{A}_N)\mathbf{A}_N^{-1}M_N^*\overline{\chi} + \mathbf{d}_N - \mathbf{d}_N^{\delta},
\end{aligned}
$$

since $\overline{\mathbf{c}}_N = \mathbf{A}_N^{-1}\mathbf{d}_N = \mathbf{A}_N^{-1}M_N^*\overline{\chi}$. Now $||\chi|| = 1$ and $\mathbf{A}_N^{-1}M_N^*$ is the adjoint of $M_N\mathbf{A}_N^{-1}$, so that $||\mathbf{A}_N^{-1}M_N^*|| = ||M_N\mathbf{A}_N^{-1}||$. But also, by Theorem 4.1,

$$
\kappa_N = ||\mathbf{A}_N^{-1}||^{1/2} = ||M_N\mathbf{A}_N^{-1}||.
$$

Thus

$$
||\mathbf{e}_N|| \le \delta_1\kappa_N + \delta_2. \tag{5.8}
$$

These bounds lead to our final theorem, concerned with the conditiong of the linear system (5.5) and the effects of errors in the entries of $\mathbf{A}_N$ and $\mathbf{d}_N$. As in Section 4, $\text{cond}\,\mathbf{A}_N$ denotes the condition number of $\mathbf{A}_N$, defined by $\text{cond}\,\mathbf{A}_N := ||\mathbf{A}_N||\,||\mathbf{A}_N^{-1}||$.

Theorem 5.2. *It holds that*

$$
\text{cond}\,\mathbf{A}_N \le N\kappa_N^2,
$$

where $\kappa_N = ||\mathbf{A}_N^{-1}||^{1/2}$. If $||\mathbf{A}_N - \mathbf{A}_N^{\delta}|| \le \delta_1$ and $||d_N - d_N^{\delta}|| \le \delta_2$, with $\delta_1\kappa_N^2 < 1$, then the linear system (5.6) has a unique solution, $\overline{\mathbf{c}}_N^{\delta}$, and

$$
||\mathbf{c}_N - \mathbf{c}_N^{\delta}|| \le \frac{\kappa_N^2}{1 - \delta_1\kappa_N^2}[\delta_1\kappa_N + \delta_2]. \tag{5.9}
$$

For every $\epsilon > 0$ there exists $C_{\epsilon}' > 0$, depending only on k, f, and ϵ, such that

$$
\kappa_N \le C_{\epsilon}' \exp(k(f_+ - f_- + \epsilon)\beta_N^*), \tag{5.10}
$$

where $\beta_N^* := \max_{1 \le m \le N} \Im\beta_{j_m}$. If the conditions of Corollary 4.1 are satisfied and $f_+ > f_-$, so that the surface is not flat, then $\kappa_N \to \infty$ and $\text{cond}\,\mathbf{A}_N \to \infty$ as $N \to \infty$.

Proof. From (4.11) we see that the entries of $\mathbf{A}_N$ satisfy $|A_{mn}^{(N)}| \le 1$. From this it follows that $||\mathbf{A}_N|| \le N$ so that $||\mathbf{A}_N||\,||\mathbf{A}_N^{-1}|| \le N\kappa_N^2$. The bound (5.9) follows from (5.7), (4.20), and (5.8). The remaining results are from Lemma 4.3 and Theorem 4.1. $\square$

We note that the bounds (4.33) and (5.9) show that a small amount of error in computing the matrix $\mathbf{A}_N$ and right hand side $\mathbf{d}_N$ will not have a significant effect on the accuracy of solving (5.5) provided $\beta_N^* k(f_+ - f_-)$ is not large, in particular if $\beta_N^* = 0$ (X_N contains only propagating modes with $|\alpha_n| \le 1$).

Table 1: Values of d/λ, L/λ and θ in the three examples. Also given are the number of propagating modes in the Rayleigh expansion.

Example	d/λ	L/λ	θ	# prop. modes	Corresponding example in [15]
1	4.7968	63.9587	20°	128	Example 1A
2	4.7877	63.8366	75°	128	Example 1B
3	0.2632	1.0526	20°	2	Example 2A

6   Numerical results

In this final section, we will examine the performance of the Galerkin and least squares methods we have discussed when they are applied to a particular model problem. We consider the case when the scattering surface is given by the function

$$f(x) = -\frac{d}{2}\cos\left(\frac{2\pi}{L}x\right),$$

for some $d > 0$. We choose values of d and L and the angle of incidence θ taken from [15] to allow direct comparison with the results obtained there. For each example we compute the solution to the scattering problem using four different methods: the SC, SS, SS* and least squares (LS) methods. Additionally, a super-algebraically convergent method presented in [29], based on a second kind integral equation formulation of the problem, is employed to provide accurate reference solutions.

A necessary condition for accuracy in scattering by a diffraction grating is based on energy conservation: the Rayleigh coefficients of the exact scattered field satisfy the relation

$$\beta_0 = \sum_{|\alpha_n| \leq 1} \beta_n |u_n|^2.$$

Hence, for each method, the particular method indicated by a superscript (XX), with XX = SC, SS, SS*, or LS, we compute the quantity

$$E_{\text{ener}} := \log_{10}\left|1 - \sum_{|\alpha_n| \leq 1} \frac{\beta_n}{\beta_0} |u_n^{(\text{XX})}|^2\right|. \tag{6.1}$$

Additionally, the individual Rayleigh coefficients are checked for accuracy by comparing against the reference solution. This is achieved by computing a second error functional,

$$E_{\text{coef}} := \log_{10}\left(2\sqrt{\sum_{|\alpha_n| \leq 1} \frac{\beta_n}{\beta_0} |u_n^{(\text{XX})} - u_n^{(\text{ref})}|^2}\right). \tag{6.2}$$

The weights in this definition are selected to make the values of E_{ener} and E_{coef} comparable. In particular, using the discrete Hölder inequality, we have that

$$
\begin{aligned}
1 - \sum_{|\alpha_n| \leq 1} \frac{\beta_n}{\beta_0} |u_n^{(\text{XX})}|^2 &= \sum_{|\alpha_n| \leq 1} \frac{\beta_n}{\beta_0} \left(|u_n|^2 - |u_n^{(\text{XX})}|^2 \right) \\
&\leq \sum_{|\alpha_n| \leq 1} \frac{\beta_n}{\beta_0} |u_n - u_n^{(\text{XX})}| \, |u_n + u_n^{(\text{XX})}| \\
&\leq \left(\sum_{|\alpha_n| \leq 1} \frac{\beta_n}{\beta_0} |u_n + u_n^{(\text{XX})}|^2 \right)^{1/2} \left(\sum_{|\alpha_n| \leq 1} \frac{\beta_n}{\beta_0} |u_n - u_n^{(\text{XX})}|^2 \right)^{1/2} \\
&\leq \left(4 + 2 \times 10^{E_{\text{ener}}} \right)^{1/2} \left(\sum_{|\alpha_n| \leq 1} \frac{\beta_n}{\beta_0} |u_n - u_n^{(\text{XX})}|^2 \right)^{1/2} .
\end{aligned}
$$

Hence, we expect $E_{\text{ener}} \leq E_{\text{coef}}$ for any reasonably accurate numerical method.

Table 1 gives the values of d, L (relative to the wavelength) and θ for the various examples. In each case, the calculations were carried out for a number of spaces (X_N, Y_N), starting with the spaces corresponding to the propagating modes and then increasing N by symmetrically adding evanescent modes. In the case of the SS, SS*, and LS methods, the coefficients in the linear system matrix have to be computed by numerical quadrature. As suggested in Section 4 we use an M-point trapezoidal rule which is rapidly convergent as M increases, and select values for M which ensure that the integrals are computed to machine accuracy.

The computed values for the functionals E_{ener} and E_{coef} for Examples 1 and 2 are displayed in Fig. 1. For the near normal incidence of Example 1, all methods perform equally well and compute the scattered field to high accuracy, even without any evanescent modes represented in the discrete spaces. Given that the total arc-length of Γ is approximately 130λ, we see that all methods are very efficient, achieving close to machine accuracy with $N = 128$, i.e. with less than one degree of freedom per wavelength of the boundary. For comparison, to achieve similar accuracy, the super-algebraically convergent Nyström method of [29] requires the solution of a linear system over 20 times larger. When the number of unknowns is increased, the SC method shows some signs of instability. It is worth noting that the functional E_{ener} is rather smaller in this example than E_{coef}, so that E_{ener} gives a somewhat misleading impression of the achieved accuracy.

In the case of the near grazing incidence of Example 2, the situation is somewhat different: all methods require a substantial number of evanescent modes to be included in the discrete spaces to compute the scattered field accurately. However, eventually all methods do provide accurate results which is a new observation compared to [15] where only a few evanescent modes were used. We note, moreover, that even with the largest value of N used ($N = 210$) the number of degrees of freedom per wavelength is very modest (≈ 1.6) given the high accuracy achieved.
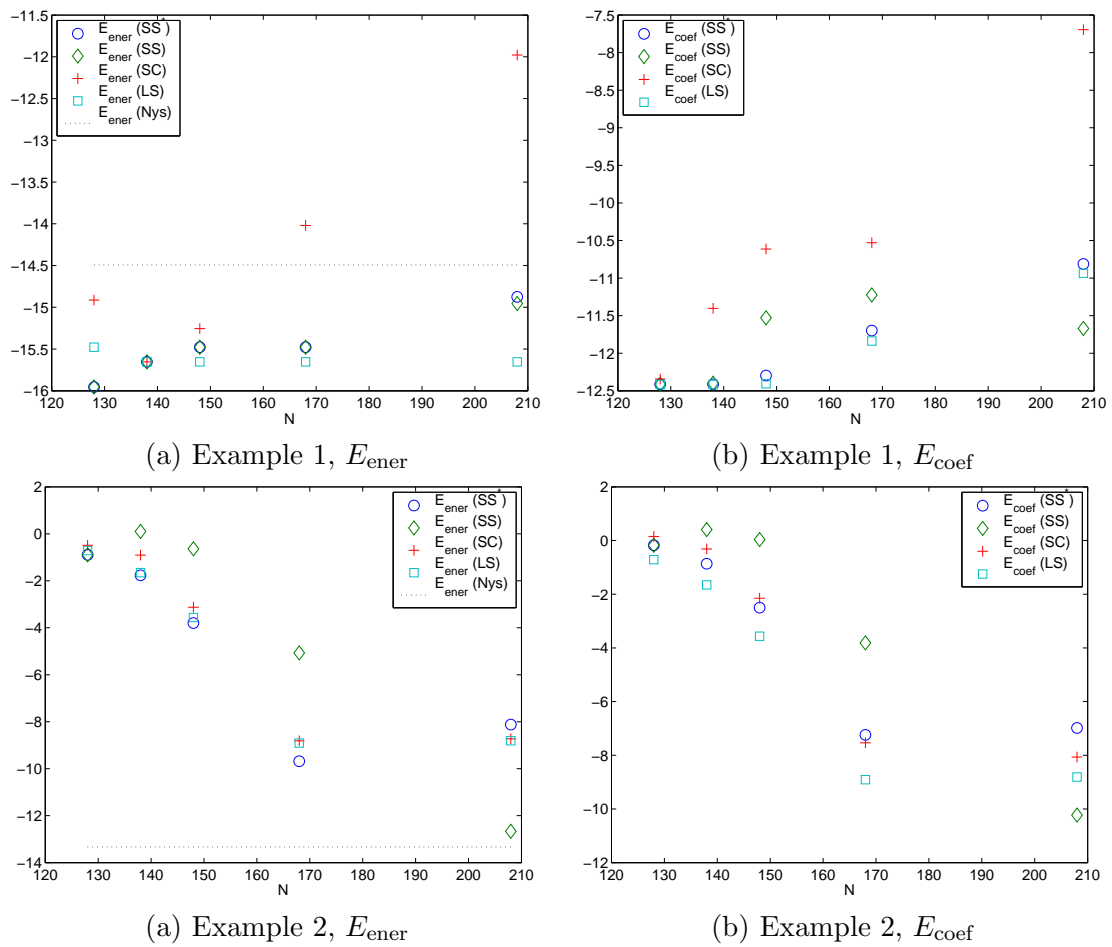
(a) Example 1, E_{ener}

(b) Example 1, E_{coef}

(a) Example 2, E_{ener}

(b) Example 2, E_{coef}

Figure 1: Values of the error functionals for Examples 1 and 2 for $N = 128, 138, 148, 168$ and 208, respectively.
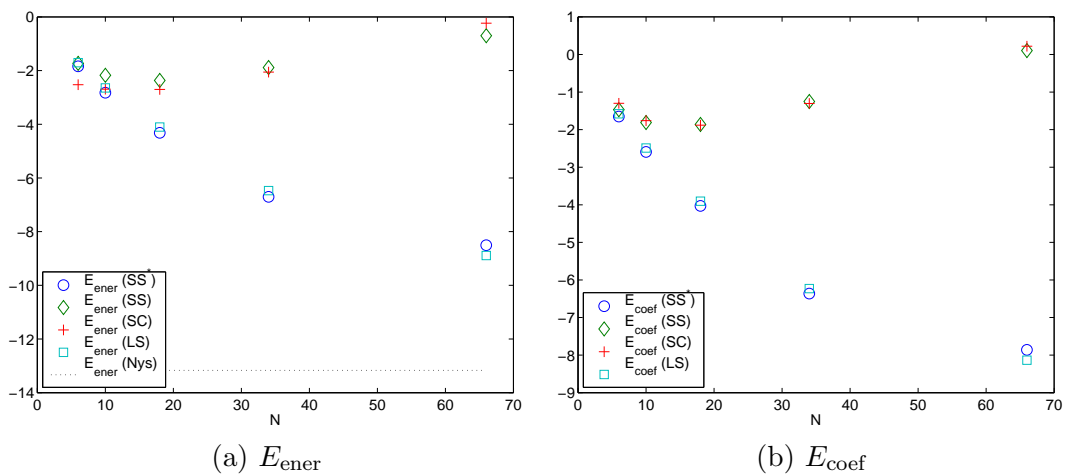

(a) E_{ener}

(b) E_{coef}

Figure 2: Values of the error functionals for Example 3 for $N = 6, 10, 18, 34$ and 66, respectively.

Table 2: Values of $\log_{10}\|\mathbf{A}_N\|$, $\log_{10}\|\mathbf{A}_N^{-1}\|$, and est, the approximate upper bound on $\log_{10}\|\mathbf{A}_N^{-1}\|$.

Example 1				Example 2			
N	$\log_{10}\|\mathbf{A}_N\|$	$\log_{10}\|\mathbf{A}_N^{-1}\|$	est	N	$\log_{10}\|\mathbf{A}_N\|$	$\log_{10}\|\mathbf{A}_N^{-1}\|$	est
128	0.44	0.42	0	128	0.43	0.40	0
138	0.45	4.91	10.46	138	0.44	4.85	10.35
148	0.45	8.82	15.13	148	0.44	8.76	15.04
168	0.45	14.84	22.21	168	0.44	14.83	22.13
208	0.45	16.89	33.50	208	0.44	17.11	33.43

Example 3			
N	$\log_{10}\|\mathbf{A}_N\|$	$\log_{10}\|\mathbf{A}_N^{-1}\|$	est
6	0.08	1.30	3.30
10	0.09	2.71	6.17
18	0.09	5.57	11.70
34	0.09	11.32	22.66
66	0.09	17.51	44.52

The situation changes significantly in Example 3, as illustrated in Fig. 2. For this case the SS and SC methods do not give accurate results for any N, while the SS* and LS methods converge as N increases, at least up to $N = 66$. However, as predicted theoretically, the condition numbers of the linear systems grow as N increases, reaching the value 6.7×10^{17} for $N = 66$. This large value of the condition number appears to be slowing the convergence rate, so that the results for $N = 66$ are not as accurate as would be expected from extrapolating the convergence rate from lower values of N. Nevertheless, the SS* and LS methods are pretty effective, achieving quite accurate results with $N = 34$, i.e. with ≈ 14 degrees of freedom per wavelength (Γ has arc-length $\approx 2.4\lambda$).

The conditioning of the system matrices is studied in a little more detail in Table 2, where we tabulate the norm of $\mathbf{A}_N$ and $\kappa_N^2 = \|\mathbf{A}_N^{-1}\|$. The estimate (4.33) predicts that, for every $\epsilon > 0$, $\kappa_N^2 \leq C\exp(2k(d+\epsilon)\beta_N^*)$, where the constant C depends only on ϵ, d, L and k, and β_N^* denotes the maximum value of $\Im\beta_n$ over the modes n included in the approximation space. To give some indication of the numerical value of this estimate we have tabulated $\log_{10}(\exp(2k\,d\,\beta_N^*)) \approx 0.869\,k\,d\,\beta_N^*$ in the column labelled *est*.

The results show that $\|\mathbf{A}_N^{-1}\|$ and the condition number cond $\mathbf{A}_N = \|\mathbf{A}_N\|\,\|\mathbf{A}_N^{-1}\|$ increase rapidly as N increases, once the approximation space starts to contain evanescent modes. Moreover, the principle of exponential growth of $\|\mathbf{A}_N^{-1}\|$, suggested by the bound (4.33), appears to be supported by our numerical results, though the bound overestimates the value of κ_N by orders of magnitude, in fact appears to overestimate the rate of exponential increase of $\|\mathbf{A}_N^{-1}\|$ by approximately a factor of two. We note that $\|\mathbf{A}_N\|$ remains bounded as N increases.

References

[1] H. Abe and T. Sato, Boundary integral equations from Hamilton's principle for surface acoustic waves under periodic metal gratings, IEEE T. Ultrason. Ferr. Freq. Control, 47 (2000), 1601–1603.

[2] G. Bao, D.C. Dobson and J.A. Cox, Mathematical studies in rigorous grating theory, J. Opt. Soc. Am. A, 12 (1995), 1029–1042.

[3] T. E. Betcke and L. N. Trefethen, Reviving the method of particular solutions, SIAM Rev., 47 (2005), 469–491.

[4] R. Brauer and O. Bryngdahl, Electromagnetic diffraction analysis of 2-dimensional gratings, Opt. Commun., 100 (1993), 1–5.

[5] O. P. Bruno and F. Reitich, Numerical solution of diffraction problems - a method of variation of boundaries. 2. Finitely conducting gratings, Padé approximants, and singularities, J. Opt. Soc. Am. A, 11 (1994), 2816–2828.

[6] M. Cadilhac, Some mathematical aspects of the grating theory, in: R. Petit (Ed.), Electromagnetic Theory of Gratings, Springer-Verlag, Berlin, 1980, pp. 53–62.

[7] S. N. Chandler-Wilde, The impedance boundary value problem for the Helmholtz equation in a half-plane, Math. Meth. Appl. Sci., 20 (1997), 813–840.

[8] S. N. Chandler-Wilde and B. Zhang, Electromagnetic scattering by an inhomogenous conducting or dielectric layer on a perfectly conducting plate, Proc. R. Soc. Lon. A, 454 (1998), 519–542.

[9] S. N. Chandler-Wilde and B. Zhang, A uniqueness result for scattering by infinite rough surfaces, SIAM J. Appl. Math., 58 (1998), 1774–1790.

[10] D. Colton and R. Kress, Integral Equation Methods in Scattering Theory, Wiley, New York, 1983.

[11] D. Colton and R. Kress, Inverse Acoustic and Electromagnetic Scattering Theory, Springer, Berlin, 1998, 2nd edition.

[12] J. A. DeSanto, Scattering from a perfectly reflecting arbitrary periodic surface: An exact theory, Radio Sci., 16 (1981), 1315–1326.

[13] J. A. DeSanto, Exact spectral formalism for rough-surface scattering, J. Opt. Soc. Am., A2 (1985), 2202–2207.

[14] J. A. DeSanto, Scattering from rough surfaces, in: R. Pike and P. Sabatier (Eds.), Scattering, Academic Press, 2002, pp. 15–36.

[15] J. A. DeSanto, G. Erdmann, W. Hereman and M. Misra, Theoretical and computational aspects of scattering from rough surfaces: One-dimensional perfectly reflecting surfaces, Waves Random Media, 8 (1998), 385–414.

[16] J. A. DeSanto, G. Erdmann, W. Hereman and M. Misra, Theoretical and computational aspects of scattering from periodic surfaces: One-dimensional transmission interfaces, Waves Random Media, 11 (2001), 425–453.

[17] J. Elschner and G. Schmidt, Diffraction in periodic structures and optimal design of binary gratings I: Direct problems and gradient formulas, Math. Meth. Appl. Sci., 21 (1998), 1297–1342.

[18] J. Elschner and M. Yamamoto, An inverse problem in periodic diffractive optics: reconstruction of Lipschitz grating profiles, Appl. Anal., 81 (2002), 1307–1328.

[19] H. W. Engl, M. Hanke and A. Neubauer, Regularization of Inverse Problems, Kluwer, Dordrecht, 1996.

[20] L. F. Li, Multilayer-coated diffraction gratings - differential equation method of Chandezon

et al revisited, J. Opt. Soc. Am. A, 11 (1994), 2816–2828.

[21] L. F. Li, J. Chandezon, G. Granet and J. P. Plumey, Rigorous and efficient grating-analysis method made easy for optical engineers, Appl. Optics, 38, 304–313 (1999).

[22] A. Kirsch, Diffraction by periodic structures, in: L. Pävärinta and E. Somersalo (Eds.), Inverse Problems in Mathematical Physics, Springer, Berlin, 1993, pp. 87–102.

[23] A. Kirsch, Uniqueness theorems in inverse scattering theory for periodic structures, Inverse Probl., 10 (1994), 145–152.

[24] A. Kirsch, An Introduction to the Mathematical Theory of Inverse Problems, Springer, Berlin, 1996.

[25] R. Kress, Numerical Analysis, Springer, Berlin, 1998.

[26] R. Kress, Linear Integral Equations, Springer, Berlin, 1999, 2nd edition.

[27] C. M. Linton, The Green's function for the two-dimensional Helmholtz equation in periodic domains, J. Eng. Math., 33 (1998), 377–402.

[28] D. Maystre, Rigorous vector theories of diffraction gratings, in: E. Wolf (Ed.), Progress in Optics XXI, Elsevier, Amsterdam, 1984, pp. 1–67.

[29] A. Meier, T. Arens, S. N. Chandler-Wilde and A. Kirsch, A Nyström method for a class of integral equations on the real line with applications to scattering by diffraction gratings and rough surfaces, J. Integral Equations Appl., 12 (2000), 281–321.

[30] R. F. Millar, The Rayleigh hypothesis and a related least-squares solution to scattering problems for periodic surfaces and other scatterers, Radio Sci., 8 (1973), 785–796.

[31] M. Neviere and E. Popov, Analysis of dielectric gratings of arbitrary profiles and thicknesses - comment, J. Opt. Soc. Am. A, 9 (1992), 2095–2096.

[32] M. Nieto-Vesperinas, Scattering and Diffraction in Physical Optics, Wiley, New York, 1991.

[33] A. F. Peterson, An outward-looking differential equation formulation for scattering from one-dimensional periodic diffraction gratings, Electromagnetics, 14 (1994), 227–238.

[34] R. Petit (Ed.), Electromagnetic Theory of Gratings, Springer, Berlin, 1980.

[35] A. Pomp, The integral method for coated gratings: Computational cost, J. Mod. Optics, 38 (1991), 109–120.

[36] A. G. Voronovich, Wave Scattering from Rough Surfaces, Springer, Berlin, 1998, 2nd edition.