An application of the re-iterated Galerkin approximation in 2-dimensions

L. G. Bennetts

Submitted to the Department of Mathematics,
University of Reading,
in partial fulfilment of the requirements
of the degree of
Master of Science.

I confirm that this is my own work
and the use of all material from other sources
has been properly and fully acknowledged.

L. G. Bennetts
September 2003.

Abstract

An acoustics problem, modelling the scattering of sound waves in a particular type of waveguide is formulated, and converted into a two-dimensional integral equation using a Green's function. The Galerkin, iterated Galerkin and re-iterated Galerkin methods of approximating the solution of integral equations are introduced and discussed. A link to variational principles is made that provides superconvergence results for the approximation of certain quantities. The acoustics problem is solved using the re-iterated Galerkin method and conclusions are drawn about its performance.

Acknowledgments

I would like to thank the following:

Prof. David Porter – for supervision, advice and guidance.

Dr. Peter Chamberlain – for always giving up time to answer my questions.

EPSRC – for financial backing throughout the year.

Contents

Chapter 1

An acoustics problem

Consider the following idealised acoustics problem.

In a region that will be known as a *waveguide*, of height b, and length and breadth of far greater dimension, a continual disturbance at one of the ends causes a disruption to the equilibrium state within the waveguide. The sound waves that are produced by the disturbance travel through the waveguide towards the opposite end. Within the waveguide, the relationship between pressure and density, and hence wave speed, is constant (as in air) in all but a fixed region. This region is referred to as the *obstacle*. The effect of the obstacle is to scatter the incident waves. At some non-specific point in time after the disturbance first occurs, the behaviour within the waveguide settles to a periodic steady state.

In addition, it is known that the geometry of the obstacle is that it resembles a uniform tube, lying in the breadth direction. That is, it has a constant height by length cross-section. The walls (i.e. boundaries of the waveguide) are highly absorbent, or so-called 'sound-soft'. This will be important later.

From this information, we aim to construct a mathematical model that will determine the structure of this periodic steady state. By concentrating our attention on the steady state problem, we are absolved from any need for an initial state, that could lead to an initial value problem. The steady state problem is of use in determining the energy transfer that results from the wave scattering. This issue will become better defined in due course.

The inherent symmetry in the problem means that, under some simplifying assumptions, one of the three spatial dimensions is redundant. Specifically, if it is assumed that the motion of the sound waves caused by the disturbance is perpendicular to the generators of the obstacle, that is, entirely without breadth direction, then the problem is unsensitive to breadth variation. This means it may be solved by working only over a length by height cross-sectional portion. This will be the case from now on.

For the purpose of simplicity, let the length dimension of the waveguide be of such significantly greater dimension than the height, that the effect of the length boundaries do not influence proceedings around the obstacle. In this case, we may relinquish the length dimensions altogether and assume that the waveguide has infinite length.

No generality is lost in assuming that the cross-section in question, which from now on will be referred to as the domain D, lies in the regular cartesian directions, so that

$$D = \{x, y : x \in \mathbb{R}, 0 < y < b\}.$$

For the present, the subdomain on which the wave speed varies will be undefined, suffice to say that it is connected and bounded. It will be denoted D', and obviously $D' \subset D$.

It is assumed that the disturbance is small in amplitude, in which case we may solve for the *linear* wave profile $\Phi = \Phi(x, y, t)$. It is well known (see [6] for example) that Φ satisfies the *wave equation*

$$\nabla^2 \Phi = \frac{1}{c^2} \Phi_{tt} \qquad \text{in } D, \forall t, \tag{1.1}$$

where $c = c(x, y)$ denotes the wave speed over D. As the wave equation is linear, Φ is the linear combination of solution waves. In particular, Φ may be constructed as the superposition of harmonic waves, as is manifest in a *Fourier series* representation

$$\Phi(x, y, t) = \sum_{n=-\infty}^{\infty} \phi_n(x, y) e^{-i\omega_n t}. \tag{1.2}$$

Each *Fourier mode* $e^{-i\omega_n t}$ represents the time dependence of a string of harmonic waves, travelling with the constant frequency ω_n. The representation (1.2) of Φ is in fact a mathematical device. The true solution to the acoustics problem involves only the *real part* of (1.2)

$$\Re\mathrm{e}(\Phi)(x, y, t) = \sum_{n=-\infty}^{\infty} \big(\Re\mathrm{e}(\phi_n)(x, y) \cos(\omega_n t) + \Im\mathrm{m}(\phi_n)(x, y) \sin(\omega_n t)\big). \tag{1.3}$$

As mathematicians, we need only solve for an arbitrary Fourier mode, allowing someone more closely related to the original problem to perform the reconstruction. So, let the solution of (1.1) be

$$\Phi(x, y, t) = \phi(x, y) e^{-i\omega t},$$

so that the problem's time dependence is fully defined. The wave equation now reduces to a *Helmholtz equation* for the unknown function $\phi(x, y)$,

$$\nabla^2 \phi + k^2 \phi = 0, \qquad \text{in } D,$$

where

$$k^2 = k^2(x, y) = \frac{\omega^2}{c^2(x, y)}$$

varies with the wave speed, i.e. k only varies over D'. This may be expressed by writing

$$\begin{array}{ll} k = k_0 & \text{in } D \backslash D', \\ k = k(x, y) & \text{in } D', \end{array} \tag{1.4}$$

where k_0 is a constant, and $k(x, y) \to k_0$ as $x, y \to \delta D'$, the boundary of D'.

The function ϕ is known as the *velocity potential.* This is as its gradient is equal to the vector field of sound wave velocities. This information will only be of use to us as a way of describing ϕ.

Before going any further, let us deal with the issue of non-dimensionalisation.

1.1 Non-dimensionalisation

By suitably redefining the independent variables in such a way that one dimension is set to a non-dimensional scale, we gain the practical advantage of removing a parameter from matters. Consistency then dictates that all dimensions are lost, and the problem is said to be *dimensionless*.

The choice of re-scaling is ours. So, let

$$x' = \frac{\pi}{b}x, \qquad y' = \frac{\pi}{b}y,$$

which rescales the breadth onto the dimensionless interval $(0, \pi)$.

Now, if

$$\phi(x, y) = \psi(x', y'), \qquad k(x, y) = \frac{\pi}{b}\kappa(x', y'),$$

then

$$\nabla^2\psi + \kappa^2\psi = 0 \qquad \text{in } x' \in \mathbb{R},\ 0 < y' < \pi.$$

Without wishing to confuse matters, yet eager to avoid the undesirable notation just introduced, let us revert to the previous notation by now re-defining

$$x = x', \qquad y = y'$$

$$\phi = \psi, \qquad k = \kappa$$

and the domain

$$D = \{x, y : x \in \mathbb{R}, 0 < y < \pi\}.$$

Our dimensionless problem stands as

$$\nabla^2\phi + k^2\phi = 0, \qquad \text{in } D$$

and k remains as in (1.4). Note that the original breadth b does not appear in this problem.

So far nothing has been said of the conditions that are needed to make this a well-posed problem.

1.2 Conditions

The 'sound-soft' property of the walls may be interpreted as

$$\Phi = 0 \qquad \text{on } y = 0, \pi \quad \forall x, t,$$

or, in the context of our problem

$$\phi = 0 \qquad \text{on } y = 0, \pi \quad \forall x. \tag{1.5}$$

Further conditions may be deduced as follows.

For large positive or negative x, that is x away from D', $\phi \sim \tilde{\phi}$ such that

$$\nabla^2 \tilde{\phi} + k_0^2 \tilde{\phi} = 0,$$

and $\tilde{\phi}$ satisfies condition (1.5). The separation solutions of this equation may easily be determined, but first let us make a further simplifying assumption. Although generality is lost, we will be content to restrict ourselves to the case in which

$$1 < k_0 < 2.$$

Now, let $\tilde{\phi}(x, y) = X(x)Y(y)$, then

$$X''Y + XY'' + k_0^2 XY = 0$$

hence

$$\frac{Y''}{Y} = -\frac{X'' + k_0^2 X}{X} = -\mu^2$$

where μ is the separation constant.

This yields the two ordinary differential equations

$$X'' + (k_0^2 - \mu^2)X = 0$$

and

$$Y'' + \mu^2 Y = 0,$$

where Y satisfies homogeneous boundary conditions.

There exist infinitely many solutions to each of these equations, which are

$$\begin{aligned} X_1(x) &= a_1 e^{i\beta_0 x} + b_1 e^{-i\beta_0 x}, \qquad & \beta_0 &= \sqrt{k_0^2 - 1}, \\ X_n(x) &= a_n e^{\gamma_n x} + b_n e^{-\gamma_n x}, \qquad & \gamma_n &= \sqrt{n^2 - k_0^2} \quad (n = 2, 3, \dots), \end{aligned}$$

where β_0 and γ_n are positive, real roots, and

$$Y_n(y) = c_n \sin(ny), \qquad n \in \mathbb{N},$$

all of which are determined up the arbitrary constants a_n, b_n or c_n.

The solutions X_n $(n \geq 2)$ are unbounded as $x \to \infty$, unless $a_n = 0$ and as $x \to -\infty$ unless $b_n = 0$. Hence, the only bounded solutions X_n $(n \geq 2)$ are trivial. This leaves

$$\tilde{\phi}(x, y) = (a_1 e^{i\beta_0 x} + b_1 e^{-i\beta_0 x}) c_1 \sin(y). \tag{1.6}$$

This result has significant implications for our problem.

1.3 Radiation Conditions

A solution

$$\phi \sim e^{i\beta_0 x} \sin(y) \quad \Rightarrow \quad \Phi \sim e^{i(\beta_0 x - \omega t)} \sin(y),$$

represents a wave travelling in the positive x direction. Similarly

$$\phi \sim e^{-i\beta_0 x} \sin(y) \quad \Rightarrow \quad \Phi \sim e^{-i(\beta_0 x + \omega t)} \sin(y),$$

is a wave travelling in the negative x direction. Therefore, from the solution (1.6) it may be deduced that, for large $|x|$, Φ behaves like the composition of two harmonic waves of the same frequency travelling in opposite directions parallel to the x-axis, with a sinusoidal y variation.

Suppose the disturbance occurs at $x = -\infty$, and we assume the wave in question has unit amplitude. After interaction with the domain D', part of the disturbance will be transmitted and part reflected. We are now in a position to derive and interpret the so called *radiation conditions* that

$$\begin{aligned} \phi(x,y) &\sim (\underbrace{e^{i\beta_0 x}}_{\text{incident wave}} + \underbrace{Re^{-i\beta_0 x}}_{\text{reflected wave}})\sin(y), \qquad x \to -\infty, \\ \phi(x,y) &\sim \underbrace{Te^{i\beta_0 x}}_{\text{transmitted wave}}\sin(y), \qquad x \to \infty, \end{aligned} \tag{1.7}$$

where R, the complex amplitude of the reflected wave, and T, the complex amplitude of the transmitted wave, are both unknown. A complex amplitude gives the *real amplitude* of a wave as its modulus, and the *phase* of a wave as its argument.

These unknown values give a measure of the energy transfer in the model, with $|R|^2$ proportional to the *reflected energy* and $|T|^2$ proportional to the *transmitted energy.* Note that, the model must satisfy

$$|R|^2 + |T|^2 = 1,$$

to conserve energy.

Figure 1.1 gives a diagrammatical representation of the model problem.

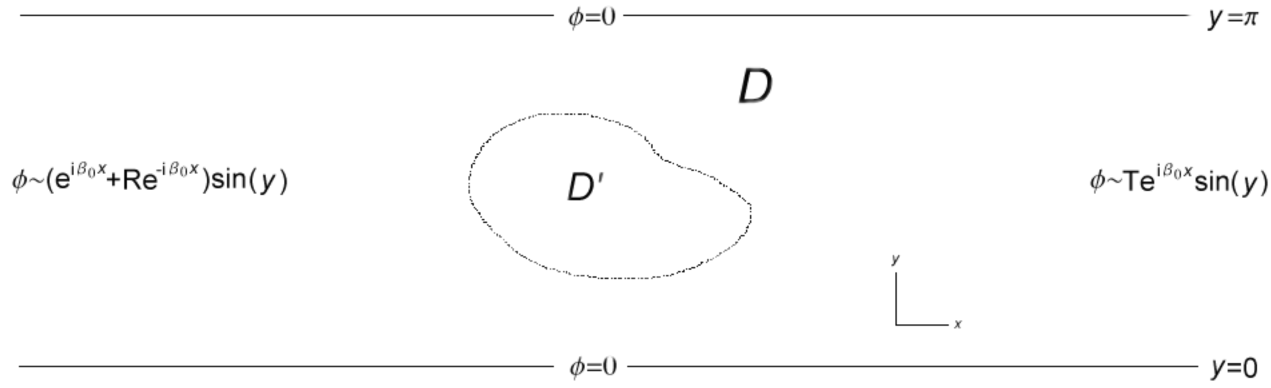
ϕ=0
$y=\pi$
D
ϕ~(e$^{i\beta_0 x}$+Re$^{-i\beta_0 x}$)sin(y)
D'
ϕ~Te$^{i\beta_0 x}$sin(y)
y
x
ϕ=0
y=0

Figure 1.1: The 2-dimensional model

Now that our model problem is fully posed, we must address the question of how to solve it.

No known explicit analytic expression for ϕ exists. Accordingly, approximate solutions are sought.

There is a variety of possible approximation techniques that may provide a solution. Some of the most familiar are *finite difference*, *finite element*, or *finite volume.* Disregarding these approaches, we look first to a technique used to produce solutions for similar but more basic equations.

1.4 Green's functions

A full treatment of *Green's functions* may be found in many texts, see for example [4]. Here, it will suffice to introduce a Green's function through our example.

Definition The *Dirac delta function* $\delta(z)$ is defined as

$$\int_a^b \psi(z)\delta(z-z_0)dz = \begin{cases} \psi(z_0) & \text{if } a < z_0 < b \\ 0 & \text{if } z_0 < a \text{ or } z_0 > b \end{cases}$$

for any sufficiently smooth function ψ.

Define a Green's function $G = G(x, y \mid x_0, y_0)$ as

$$\nabla^2 G + k_0^2 G = -\delta(x-x_0)\delta(y-y_0) \tag{1.8}$$

in $D \oplus D = \{x, y, x_0, y_0 : (x,y), (x_0,y_0) \in D\}$, with boundary conditions

$$G = 0 \qquad \text{on } y = 0, \pi, \quad \text{for } x \in \mathbb{R};\ (x_0, y_0) \in D \tag{1.9}$$

and radiation conditions

$$\begin{array}{ll} G = C_1(x_0,y_0)e^{-i\beta_0 x}\sin(y) & x \to -\infty, \\ G = C_2(x_0,y_0)e^{i\beta_0 x}\sin(y) & x \to \infty, \end{array} \tag{1.10}$$

where the $C_i(x_0, y_0)$ are functions that are constant with respect to x and y.

It is mentioned at this juncture that the expression $\delta(x-x_0)\delta(y-y_0)$ is undefined at the point $(x,y) = (x_0,y_0)$. In this two-dimensional case, the singularity it induces in the Green's function is *weak* (logarithmic) and in theory does not prove problematic. It will, however, require special treatment for numerical computation. This point is pursued in §6.3.

Let us determine G in its *Fourier series* representation by writing

$$G(x, y \mid x_0, y_0) = \sum_{n=1}^{\infty} A_n(x \mid x_0, y_0)\sin(ny),$$

where the cosine terms have been omitted to satisfy the homogeneous boundary conditions (1.9). By orthogonality of the Fourier series the variable coefficients A_n are given by

$$A_n(x \mid x_0, y_0) = \frac{2}{\pi}\int_0^{\pi} G(x, y \mid x_0, y_0)\sin(ny)\mathrm{d}y$$

Noticing that

$$A_n'' = \frac{2}{\pi}\int_0^{\pi} G_{xx}\sin(ny)\mathrm{d}y$$

where $'$ represents partial differentiation with respect to x, and

$$-n^2 A_n = \frac{2}{\pi}\int_0^{\pi} G_{yy}\sin(ny)\mathrm{d}y,$$

equation (1.8) may be manipulated to produce the set of equations

$$A_n'' - (n^2 - k_0^2)A_n = -\frac{2}{\pi}\sin(ny_0)\delta(x-x_0) \tag{1.11}$$

to be solved for the A_n. To make matters clearer, we write these equations as

$$A_1'' + \beta_0^2 A_1 = -\frac{2}{\pi}\sin(y_0)\delta(x - x_0)$$

and

$$A_n'' - \gamma_n^2 A_n = -\frac{2}{\pi}\sin(ny_0)\delta(x - x_0) \qquad (n \geq 2)$$

with β_0 and γ_n as previously defined. Using condition (1.10), we require

$$\begin{aligned} A_1 &= C_1(x_0, y_0)e^{-i\beta_0 x}, \qquad & x \to -\infty, \\ A_1 &= C_2(x_0, y_0)e^{i\beta_0 x}, \qquad & x \to \infty. \end{aligned} \tag{1.12}$$

All other A_n necessarily decay as $|x| \to \infty$.

Equations of the form

$$u''(x) + \kappa_1^2 u(x) = f(x)$$

and

$$v''(x) - \kappa_2^2 v(x) = g(x),$$

where the κ_i are constants with respect to x, have the solutions

$$u(x) = c_1 e^{i\kappa_1 x} + c_2 e^{-i\kappa_1 x} + \frac{1}{\kappa_1}\int_{a_1}^{x} \sin\big(\kappa_1(x - t)\big)f(t)\mathrm{d}t$$

and

$$v(x) = c_3 e^{\kappa_2 x} + c_4 e^{\kappa_2 x} + \frac{1}{\kappa_2}\int_{a_2}^{x} \sinh\big(\kappa_2(x - t)\big)g(t)\mathrm{d}t$$

where the c_i are constants, and the a_i lie in the domain of the appropriate solution.

Therefore, the solutions of (1.11) are

$$\begin{aligned} A_1(x \mid x_0, y_0) &= c_1 e^{i\beta_0 x} + c_2 e^{-i\beta_0 x} - \frac{2\sin(y_0)}{\beta_0 \pi}\int_{-\infty}^{x} \sin\big(\beta_0(x - t)\big)\delta(t - x_0)dt \\ &= \begin{cases} c_1 e^{i\beta_0 x} + c_2 e^{-i\beta_0 x} & (x_0 > x) \\ c_1 e^{i\beta_0 x} + c_2 e^{-i\beta_0 x} - \dfrac{2\sin(y_0)\sin\big(\beta_0(x - x_0)\big)}{\beta_0 \pi} & (x_0 < x) \end{cases} \end{aligned}$$

and, for $n \geq 2$

$$\begin{aligned} A_n(x \mid x_0, y_0) &= c_3 e^{\gamma_n x} + c_4 e^{\gamma_n x} - \frac{2\sin(ny_0)}{\gamma_n \pi}\int_{-\infty}^{x} \sinh\big(\gamma_n(x - t)\big)\delta(t - x_0)\mathrm{d}t \\ &= \begin{cases} c_3 e^{\gamma_n x} + c_4 e^{-\gamma_n x} & (x_0 > x) \\ c_3 e^{\gamma_n x} + c_4 e^{-\gamma_n x} - \dfrac{2\sin(ny_0)\sinh\big(\gamma_n(x - x_0)\big)}{\gamma_n \pi} & (x_0 < x). \end{cases} \end{aligned}$$

The only point of contention here is at $x = x_0$. This should not be surprising considering what has been said of the expected singularity in the Green's function. Taking the limits $x \to x_0^+$ and $x \to x_0^-$, it can be shown that the coefficients A_n are continuous functions of x at this point. However, similar calculations show that all A_n have a slope discontinuity at the point $x = x_0$.

Consider applying the appropriate conditions to the cases A_1 and A_n $(n \geq 2)$ separately.

In the limit $x \to -\infty$, the inequality $x_0 > x$ holds, hence

$$A_1 = c_1 e^{i\beta_0 x} + c_2 e^{-i\beta_0 x}, \qquad \text{as } x \to -\infty$$

and the first part of condition (1.12) implies that $c_1 = 0$.

Now, taking the limit $x \to \infty$, so that $x_0 < x$,

$$\begin{aligned} A_1(x \mid x_0, y_0) &= c_2 e^{-i\beta_0 x} - \frac{2\sin(y_0)\sin\big(\beta_0(x - x_0)\big)}{\beta_0 \pi} \\ &= c_2 e^{-i\beta_0 x} + i\frac{\sin(y_0)}{\beta_0\pi}(e^{i\beta_0(x-x_0)} - e^{-i\beta_0(x-x_0)}) \\ &= (c_2 - i\frac{\sin(y_0)}{\beta_0\pi}e^{i\beta_0 x_0})e^{-i\beta_0 x} + i\frac{\sin(y_0)}{\beta_0\pi}e^{i\beta_0(x-x_0)} \end{aligned}$$

and to satisfy the second part of condition (1.12) we must set

$$c_2 = i\frac{\sin(y_0)}{\beta_0\pi}e^{i\beta_0 x_0}.$$

This leaves

$$A_1(x \mid x_0, y_0) = \begin{cases} i\frac{\sin(y_0)}{\beta_0\pi}e^{-i\beta_0(x-x_0)} & (x_0 > x) \\ i\frac{\sin(y_0)}{\beta_0\pi}e^{-i\beta_0(x-x_0)} - \dfrac{2\sin(y_0)\sin\big(\beta_0(x-x_0)\big)}{\beta_0\pi} & (x_0 < x). \end{cases}$$

It is easily shown that G satisfies the symmetry property

$$G(x, y \mid x_0, y_0) = G(x_0, y_0 \mid x, y), \qquad \forall (x, y), (x_0, y_0) \in D : (x, y) \neq (x_0, y_0).$$

This enables us to write A_1 as a single expression (or, it at least saves much calculation in doing so)

$$A_1(x \mid x_0, y_0) = i\frac{\sin(y_0)}{\beta_0\pi}e^{i\beta_0|x-x_0|} \qquad (x, y), (x_0, y_0) \in D.$$

Similarly, as $x \to -\infty$

$$A_n(x \mid x_0, y_0) = c_3 e^{\gamma_n x} + c_4 e^{-\gamma_n x} \sim c_4 e^{-\gamma_n x},$$

so a bounded solution requires $c_4 = 0$. Letting $x \to \infty$

$$\begin{aligned} A_n(x \mid x_0, y_0) &= c_3 e^{\gamma_n x} - \frac{2\sin(ny_0)\sinh\big(\gamma_n(x-x_0)\big)}{\gamma_n\pi} \\ &= c_3 e^{\gamma_n x} - \frac{\sin(ny_0)}{\gamma_n\pi}\big(e^{\gamma_n(x-x_0)} - e^{-\gamma_n(x-x_0)}\big) \\ &= (c_3 - \frac{\sin(ny_0)}{\gamma_n\pi}e^{-\gamma_n x_0})e^{\gamma_n x} + \frac{\sin(ny_0)}{\gamma_n\pi}e^{-\gamma_n(x-x_0)} \\ &\sim c_3 - \frac{\sin(ny_0)}{\gamma_n\pi}e^{-\gamma_n x_0}e^{\gamma_n x} \end{aligned}$$

and we must set the constant

$$c_3 = \frac{\sin(ny_0)}{\gamma_n \pi} e^{-\gamma_n x_0}$$

for boundedness.

Thus

$$\begin{aligned} A_n(x \mid x_0, y_0) &= \begin{cases} \frac{\sin(ny_0)}{\gamma_n \pi} e^{\gamma_n (x-x_0)} & (x_0 > x) \\ \frac{\sin(ny_0)}{\gamma_n \pi} e^{\gamma_n (x-x_0)} - \frac{2\sin(ny_0)}{\gamma_n \pi} \sinh\big(\gamma_n(x-x_0)\big) & (x_0 < x) \end{cases} \\ &= \frac{\sin(ny_0)}{\gamma_n \pi} e^{-\gamma_n |x-x_0|} \qquad (x,y),(x_0,y_0) \in D. \end{aligned}$$

The required Green's function is now defined as

$$\begin{aligned} G(x,y \mid x_0,y_0) &= \frac{i}{\beta_0 \pi} \sin(y_0)\sin(y) e^{i\beta_0|x-x_0|} \\ &+ \sum_{n=2}^{\infty} \frac{1}{\gamma_n \pi} \sin(ny_0)\sin(ny) e^{-\gamma_n|x-x_0|}. \end{aligned} \tag{1.13}$$

This series expression is convergent everywhere except the point $(x,y) = (x_0,y_0)$, where G has the singularity spoken of. Now

$$\frac{1}{\gamma_n} \sim \frac{1}{n}$$

for large n implies that the singularity is logarithmic in nature (this will be made explicit in §6.3). Such functions are measurable and hence expressions containing the Green's function beneath an integral sign are well defined.

We proceed by applying Green's theorem in the plane

$$\int\!\!\int_D \big(u\nabla^2 v - v\nabla^2 u\big)\mathrm{d}x\mathrm{d}y = \int_{\delta D} \big(u\frac{\partial v}{\partial n} - v\frac{\partial u}{\partial n}\big)\mathrm{d}c \tag{1.14}$$

to our problem, where $\frac{\partial}{\partial n}$ denotes the outward unit normal derivative.
Substituting $u = \phi$ and $v = G$ in (1.14) gives

$$\int\!\!\int_D \big(\phi\nabla^2 G - G\nabla^2 \phi\big)\mathrm{d}x\mathrm{d}y = \int_{\delta D} \big(\phi\frac{\partial G}{\partial n} - G\frac{\partial \phi}{\partial n}\big)\mathrm{d}c.$$

Let us analyse each side of this equality in turn, beginning with the left hand side

$$\int\!\!\int_D \big(\phi\nabla^2 G - G\nabla^2 \phi\big)\mathrm{d}x\mathrm{d}y$$

$$= \int\!\!\int_D \big[\phi(x,y)\big(-k_0^2 G(x,y \mid x_0,y_0) - \delta(x-x_0)\delta(y-y_0)\big) - G(x,y \mid x_0,y_0)\big(-k^2(x,y)\phi(x,y)\big)\big]\mathrm{d}x\mathrm{d}y$$

$$= -\phi(x_0,y_0) + \int\!\!\int_{D'} (k^2(x,y) - k_0^2)\phi(x,y)G(x,y \mid x_0,y_0)\mathrm{d}x\mathrm{d}y.$$

In order to implement the conditions as $|x| \to \infty$, it is necessary to evaluate the right hand side using the finite domain $D_X = \{x, y : |x| < X, 0 < y < \pi\}$, with boundary $\delta D_X = \{x, y : x \leq |X| \text{ on } y = 0, \pi \text{ and } 0 \leq y \leq \pi \text{ on } x = \pm X\}$. For large X, (recall that $\phi = 0$ on $y = 0, \pi$)

$$\int_{\delta D_X} \Big(\phi \frac{\partial G}{\partial n} - G \frac{\partial \phi}{\partial n}\Big) dc$$

$$= - \int_0^\pi (\phi \frac{\partial G}{\partial x} - G \frac{\partial \phi}{\partial x})_{x=-X} \mathrm{d}y + \int_0^\pi (\phi \frac{\partial G}{\partial x} - G \frac{\partial \phi}{\partial x})_{x=X} \mathrm{d}y$$

$$\sim -\frac{1}{\pi} \int_0^\pi \big[(e^{i\beta_0 X} + Re^{-i\beta_0 X}) e^{-i\beta_0 (X+x_0)} - (e^{i\beta_0 X} - Re^{-i\beta_0 X}) e^{-i\beta_0 (X+x_0)}\big] \sin^2(y) \sin(y_0) \mathrm{d}y$$

$$-\frac{1}{\pi} \int_0^\pi \big[Te^{i\beta_0 X} e^{i\beta_0 (X-x_0)} - Te^{i\beta_0 X} e^{i\beta_0 (X-x_0)}\big] \sin^2(y) \sin(y_0) \mathrm{d}y$$

$$\to -e^{i\beta_0 x_0} \sin(y_0) \qquad \text{as } X \to \infty$$

Combining these expressions gives

$$\phi(x_0, y_0) = e^{i\beta_0 x_0} \sin(y_0) + \int\!\!\int_{D'} (k^2(x, y) - k_0^2) \phi(x, y) G(x, y \mid x_0, y_0) \mathrm{d}x \mathrm{d}y \tag{1.15}$$

for $(x_0, y_0) \in D$. If the independent variables are restricted to the domain D', this is known as an *integral equation* for the unknown function ϕ over D'. Once the integral equation is solved for ϕ in D', the full solution, ϕ in D, can be trivially obtained by substitution of ϕ in D' into the right hand side of (1.15).

The real interest is in solving the integral equation. It is with this issue that much of the proceedings will be concerned.

As alluded to earlier, the amplitudes R and T are of particular interest. In preparation of §4 we shall use the term *quantities of interest* to describe them. Using (1.15) we may derive similar expressions for these constants, as follows

$$\begin{aligned} x_0 \to -\infty \qquad \phi &\sim e^{i\beta_0 x_0} \sin(y_0) + \frac{i}{\beta_0 \pi} \int\!\!\int_{D'} (k^2(x, y) - k_0^2) e^{i\beta_0 (x - x_0)} \sin(y) \sin(y_0) \phi(x, y) \mathrm{d}x \mathrm{d}y \\ &= e^{i\beta_0 x_0} \sin(y_0) + \frac{i \sin(y_0)}{\beta_0 \pi} e^{-i\beta_0 x_0} \int\!\!\int_{D'} (k^2(x, y) - k_0^2) \sin(y) e^{i\beta_0 x} \phi(x, y) \mathrm{d}x \mathrm{d}y \end{aligned}$$

and

$$\begin{aligned} x_0 \to \infty \qquad \phi &\sim e^{i\beta_0 x_0} \sin(y_0) + \frac{i}{\beta_0 \pi} \int\!\!\int_{D'} (k^2(x, y) - k_0^2) e^{-i\beta_0 (x - x_0)} \sin(y) \sin(y_0) \phi(x, y) \mathrm{d}x \mathrm{d}y \\ &= \Big(1 + \frac{i}{\beta_0 \pi} \int\!\!\int_{D'} (k^2(x, y) - k_0^2) \sin(y) e^{-i\beta_0 x} \mathrm{d}x \mathrm{d}y\Big) e^{i\beta_0 x_0} \sin(y_0). \end{aligned}$$

Referring to (1.7), R and T must satisfy

$$R = \frac{i}{\beta_0 \pi} \int\!\!\int_{D'} (k^2(x, y) - k_0^2) e^{i\beta_0 x} \phi(x, y) \mathrm{d}x \mathrm{d}y \tag{1.16}$$

and

$$T = 1 + \frac{i}{\beta_0 \pi} \int\!\!\int_{D'} (k^2(x, y) - k_0^2) e^{-i\beta_0 x} \phi(x, y) \mathrm{d}x \mathrm{d}y, \tag{1.17}$$

both of which depend on the solution ϕ in D'.

Therefore, producing approximations to the reflected and transmitted energies may be regarded as our final goal, with the approximations to ϕ in D' that must be made en route to achieve this final goal, coming almost as a by-product.

Chapter 2

Integral equations

In §1 we found that an attempt to solve a differential equation, that models a realistic situation, using a Green's function, results in an implicit expression for the solution, in the form of an integral equation. This example will serve as both motivation and a suitable illustration for what follows.

Background reading on integral equations can be found in Porter and Stirling [1].

The term *integral equation* refers to an unknown function appearing beneath an integral sign. Integral equations can often appear as reformulations of initial and boundary value problems. Although the relative merits of using either formulation over the other will not be pursued, it is worth highlighting that integral equations have the advantage that they carry no extra conditions: all boundary and initial values are encompassed by the equation. It has also been noted that the 'smoothing' properties of integrals make them preferable to derivatives. The use of integral equations is now widespread, especially when solving problems in the vein of the one considered in the previous chapter.

Like differential equations, most integral equations of interest are intractable. Our attention must turn to approximation methods. It is with the specifics of this issue that the main body of this dissertation is concerned. After introducing the *re-iterated Galerkin* method of approximating integral equations, its application to the two-dimensional problem (1.15) will be explored, something that is not known to have previously been attempted. The re-iterated Galerkin method was devised by Porter and Stirling, and first published in [2]. It owes much to the work of Sloan, who provided the *iterated Galerkin* approximation in [9].

In particular, it will be shown how abstract techniques of functional analysis may be utilised for the practical purpose of approximating integral equations. In what follows a detailed knowledge of functional analysis is not expected; the nature of this dissertation is one of applied mathematics. More specifically, it has been written within the field of numerical analysis. However, a substantial portion of the material dealt with is theoretical.

A good starting point is to derive an abstract problem from a standard general integral equation.

2.1 Integral equations of the second kind

The one-dimensional, scalar integral equation of the *second kind* takes the form

$$\phi(x) = f(x) + \lambda \int_a^b k(x,t)\phi(t)dt \qquad (a \leq x \leq b). \tag{2.1}$$

It is to be solved for $\phi(x)$ $(a \leq x \leq b)$, given the *free term* $f(x)$, *kernel* $k(x,t)$ and constants a, b, λ.

Let

$$(K\phi)(x) = \int_a^b k(x,t)\phi(t)dt \qquad (a \leq x \leq b)$$

then K represents a *linear map* between function spaces, generated by the kernel $k(x,t)$. We may then consider the entire equation (2.1) as belonging to a function space, and write

$$\phi = f + \lambda K\phi. \tag{2.2}$$

This differs subtly from (2.1), as we are now dealing with fixed elements of a space defined by their associated function, and hence the equation's explicit dependence on the variable x may be dropped. As will be seen, it is necessary to re-introduce the independent variables when making an approximation. It is helpful to think of these elements as infinite dimensional vectors that contain the values of the associated function for all x in the interval in question. The reason for recasting the problem in this way is to make use of the properties of the space.

The operator K maps a space that contains the unknown ϕ into a space that also contains ϕ. This being the case, and assuming that f also belongs to this space, no generality is lost by considering all operations taking place in one function space.

The problem (2.1) lends itself most readily to the function space $L_2(a,b)$, the space of *Lebesgue*, square integrable functions, over the field $\mathbb{R}$ (or $\mathbb{C}$) . This is well known to be a *Hilbert space.* To maintain generality it will be assumed that we are working in an arbitrary Hilbert space and denote it by $\mathscr{H}$, and all theoretical results will hold for $\mathscr{H}$, although, when speaking of $\mathscr{H}$, we will almost exclusively be referring to $L_2(a,b)$ or a higher order extension. Hilbert spaces are the most structured class of function space; this will be particularly helpful as we are dealing with approximations. When approximating an element it is necessary to assign a value to its accuracy. By definition, a Hilbert space is equipped with an *inner product* $(\cdot,\cdot)$ and *norm* $\|\cdot\|$, that are used to assign a value to the abstract angle and distance between elements.

In the space $L_2(a,b)$

$$(u,v) = \int_a^b u\bar{v}$$

and

$$\|u\| = (u,u)^{\frac{1}{2}}.$$

The question of whether a solution to (2.2) (and (2.1)) exists and is unique, is a non-trivial one, but will not be addressed (find full discussion in [1]). It is dependent on the parameter λ and the *spectrum* of K. We are primarily concerned with approximation techniques, rather than the existence of a solution. For this reason it is assumed that approximations are made to a unique solution ϕ of (2.2). To emphasise that we are dealing with what would be a set problem let us relabel $K = \lambda K$, so that the scalar λ is absorbed by the operator.

Definition The *rank* of an operator is the dimension of its range.

If the operator K has finite rank, equation (2.2) is known as *degenerate*, and may be solved exactly. It would be illogical to attempt to approximate such equations, and so it is assumed that all K under consideration take infinite rank.

Definition An operator F is said to be *bounded* on $\mathscr{H}$ if

$$||F|| = \max\{\frac{||F\varphi||}{||\varphi||} : \varphi \in \mathscr{H}\}$$

is finite.

Definition An operator F is said to be *compact* on $\mathscr{H}$ if it is the limit of a sequence of finite rank operators $\{F_n\} \subset \mathscr{H}$, i.e. $\lim_{n\to\infty} F_n = F$.

It will also be assumed that the operator K is bounded and compact on $\mathscr{H}$. These assumptions essentially say that our kernel is 'respectable', which is most often the case in practice. Much can be said of operators that are bounded and compact. These properties will not regularly be referred to, but they are necessary for many of the results used.

When using the $L_2(a, b)$ space, a function $k(x, t)$ is said to be a L_2-kernel if

$$\int_a^b \int_a^b |k(x,t)|^2 \mathrm{d}x\mathrm{d}t < \infty.$$

Such a function induces a bounded, linear, compact integral operator K on $L_2(a, b)$, defined by

$$(K\varphi)(x) = \int_a^b k(x,t)\varphi(t)\mathrm{d}t.$$

Hence, (2.1) has been reduced to approximating the unique solution

$$\phi = (I - K)^{-1} f \tag{2.3}$$

of

$$\phi = f + K\phi \tag{2.4}$$

in a Hilbert space $\mathscr{H}$, where the operator K is linear, bounded, compact, and of infinite rank. For simplicity, on occasions (2.4) will be written as

$$A\phi = f \tag{2.5}$$

where $A = I - K$, and I is the identity map on $\mathscr{H}$. The exact solution (2.3) becomes $\phi = A^{-1}f$. Note that such expressions for ϕ are only notation. It has not been assumed that the inverse operator A^{-1} exists as a mapping from $\mathscr{H}$ into itself. All that is required for a unique solution to exist is that A is injective, and the free term f lies in its range $A(\mathscr{H})$.

Although the above Hilbert space problem was derived from (2.1), the connection is not exclusive. That is, we may easily recast any linear integral equation in the form (2.5), although the required assumptions can not be expected to hold in general. This point is being emphasised as methods that will be investigated in subsequent chapters seek to approximate the solution of (2.5) rather than (2.1). In particular, it will be shown that the two-dimensional example of the first chapter is applicable.

Before investigating approximation methods, it is nescessary to define what criteria they will be judged on.

2.2 A good approximation method?

These statements should be implicit, but for clarity it is sometimes helpful to to have them in print.

1. *A good approximation is one that minimises the norm of the pointwise error.*
2. *The method used is computationally inexpensive.*
3. *The method allows us to calculate an approximation to an arbitrary degree of accuracy with respect to 1 whilst not violating 2.*

Definition The *pointwise error* is

$$e = \phi - p$$

where p is the approximation to the exact solution ϕ.

Note that condition 1 is norm specific, that is, it is sensitive to the chosen space $\mathscr{H}$. Therefore, what may appear as a good approximation under one norm may not under another. What can be said though, is that the reduction of the norm of the pointwise error means that our approximation is getting closer to the exact solution, in some sense.

Condition 2 will often be alluded to, although no attempt will be made to measure the efficiency of a method.

Condition 3 essentially says that if we have a method of calculating successively improved approximations, then the n^{th} approximation is not sensitive to the value of n. The phrase ‘arbitrary degree of accuracy’ is somewhat limited by the error in calculations, introduced by computational round-off. The best achievable approximations are those whose error is of the same order as that of the computations. This last point is a prelude to §6.

Chapter 3

Galerkin Methods

Our main problem in finding an analytic solution to (2.4) is in our inability to deal with infinities. It is then natural to constuct our approximations in finite dimensional spaces where we may easily produce a general framework to solve equations. We seek to strike a balance between letting the finite dimension get large, as it will obviously play a role in increasing the accuracy of the approximation, and keeping the dimension small enough to simplify calculations.

3.1 The Galerkin method

This is a well-established and widely used form of *projection method.* We may attempt to make a Galerkin approximation in any finite dimensional subspace of $\mathscr{H}$. Let us fix this subspace to be E_N, where N denotes the dimension. As a subspace of a Hilbert space it possesses an orthonormal basis $\{\chi_1, \ldots, \chi_N\}$, so that

$$E_N = \text{span}\{\chi_1, \ldots, \chi_N\},$$

where

$$(\chi_i, \chi_j) = \begin{cases} 0 & i \neq j \\ 1 & i = j. \end{cases}$$

This basis may be used to define an *orthogonal projection* P_N from $\mathscr{H}$ into E_N

$$\begin{aligned} P_N : \mathscr{H} &\to E_N \\ \varphi &\mapsto \sum_{i=1}^{N} (\varphi, \chi_i)\chi_i. \end{aligned} \tag{3.1}$$

Now, note that

$$\varphi = \sum_{i=1}^{\infty} (\varphi, \xi_i)\xi_i$$

for any orthonormal basis $\{\xi_1, \xi_2, \ldots\}$ of $\mathscr{H}$. So P_N reduces each element in the Hilbert space into its components that lie in the chosen subspace, and hence

$$P_N p = p \qquad \forall p \in E_N. \tag{3.2}$$

The projection P_N has the property that it maps each element in $\mathscr{H}$ onto the element that best approximates it (with respect to the norm of the space), that is, for any φ in $\mathscr{H}$

$$||P_N\varphi - \varphi|| \leq ||p - \varphi|| \qquad \forall p \in E_N. \tag{3.3}$$

We solve equation (2.4) for its components within the subspace. This is achieved by using the projected forms of the terms on the right hand side of (2.4):

$$f \mapsto P_N f \qquad K \mapsto P_N K.$$

The solution p_N of this projected equation , which is obviously in the subspace, is the Galerkin approximation

$$\begin{aligned} p_N &= P_N f + P_N K p_N \\ \Rightarrow p_N &= (I - P_N K)^{-1} P_N f. \end{aligned} \tag{3.4}$$

It has been assumed that this solution exists, ie. the operator $(I - P_N K)^{-1}$ exists. We shall see that, this assumption is guarteed by another assumption made in subsequent approximations.

Definition The functions that correspond to the basis elements are known as *trial functions*, and the space of functions spanned by these functions the *trial space.*

Some properties of the Galerkin solution will now be highlighted.

Only in rare circumstances does the Galerkin approximation p_N coincide with the projection of the solution of (2.4) ϕ. To see this consider the difference between the two elements

$$\begin{aligned} P_N\phi - p_N &= P_N(I-K)^{-1}f - (I - P_N K)^{-1} P_N f \\ &= (P_N(I-K)^{-1} - (I - P_N K)^{-1} P_N) f \\ &= (I - P_N K)^{-1}((I - P_N)P_N - P_N(I-K))(I-K)^{-1} f \\ &= (I - P_N K)^{-1} P_N (K - K P_N)\phi, \end{aligned} \tag{3.5}$$

if $\phi = \sum_{i=1}^{\infty} \alpha_i \chi_i$, then (3.5) vanishes only if, for all $j > N$, either

$$a_j = 0$$

or

$$(K\chi_j, \chi_i) = 0 \qquad i = 1, \ldots, N.$$

Otherwise $p_N \neq P_N \phi$, and

$$\exists q \in E_N : \|q - \phi\| < \|p_N - \phi\|, \tag{3.6}$$

so it is not the optimal element in the subspace (with respect to the pointwise error). The choice $q = P_N\phi$ will satisfy (3.6).

Nor does the Galerkin approximation minimise the residual error. The approximation q_N that does this satisfies the projected equation

$$A q_N = Q_N f$$

where Q_N is the orthogonal projection onto the space AE_N.

The orthonormal basis $\{\chi_1, \ldots, \chi_N\}$ may be extended indefinitely to an orthonormal basis of $\mathscr{H}$. This is what is implied when $N \to \infty$ is written. Using (3.3), as $N \to \infty$

$$P_N \varphi \to \varphi, \qquad \forall \varphi \in \mathscr{H},$$

and in particular

$$\begin{aligned} P_N f \to f \qquad & \text{as } N \to \infty, \\ P_N K \to K \qquad & \text{as } N \to \infty. \end{aligned}$$

Therefore, despite (3.6), the Galerkin approximation satisfies (2.4) increasingly more closely as $N \to \infty$, and since

$$p_N = (I - P_N K)^{-1} P_N f \to (I - K)^{-1} f = \phi \qquad \text{as } N \to \infty,$$

the Galerkin approximation tends towards the exact solution.

To improve our approximation we need only extend our subspace. Unfortunately, as will be shown, in practice this approach on its own is flawed .

The dimension of the subspace is not the only factor that dictates the accuracy of the Galerkin approximation, the choice of basis relative to the problem in hand has a major role. In fact, the choice of basis is fundamental to all of the approximation methods we shall consider. In §7.1 we shall face the question of choosing a basis for the problem set in §1.

3.1.1 Evaluation of the Galerkin approximation

The advantage of an expression such as (3.4), which involves the inverse of a finite rank operator, over an expression such as (2.3), which involves the inverse of an operator of infinite rank, is that it can be written in an explicit form. To find this explicit form, (3.4) can be turned into a system of equations as follows.

Again, we consider the subspace

$$E_N = \text{span}\{\chi_1, \ldots, \chi_N\}$$

as fixed.

By definition the Galerkin approximation satisfies

$$p_N = P_N f + P_N K p_N \tag{3.7}$$

and as it belongs to the subspace, by (3.2)

$$P_N p_N = p_N \tag{3.8}$$

so (3.7) may be written

$$P_N p_N = P_N f + P_N K p_N \iff P_N(A p_N - f) = 0,$$

and hence the residual

$$A p_N - f \in \text{null}(P_N).$$

From the definition (3.1) of P_N we have

$$\begin{aligned} \varphi \quad & \in \quad \text{null}(P_N) \\ \iff \sum_{i=1}^{N} (\varphi, \chi_i) \chi_i \quad & = \quad 0 \\ \iff (\varphi, \chi_i) \quad & = \quad 0, \qquad i = 1, \ldots, N. \end{aligned}$$

This says that the null space of P_N is orthogonal to the basis $\{\chi_1, \ldots, \chi_N\}$ (and hence to the space E_N). We conclude that the Galerkin approximation must have a residual error $Ap_N - f$ that satisfies the N equalities

$$\begin{aligned} & (Ap_N - f, \chi_i) = 0 \qquad i = 1, \ldots, N \\ \iff & (Ap_N, \chi_i) = (f, \chi_i) \qquad i = 1, \ldots, N. \end{aligned} \tag{3.9}$$

As $p_N \in E_N$ it may be expressed in terms of the basis $\{\chi_1, \ldots, \chi_N\}$

$$p_N = \sum_{i=1}^{N} a_i \chi_i \tag{3.10}$$

where the scalar coefficients a_i are currently unknown (as p_N is currently unknown). Substituting (3.10) into (3.9) yields an $N \times N$ system of equations

$$\sum_{j=1}^{N} a_j (A\chi_j, \chi_i) = (f, \chi_i), \qquad i = 1, \ldots, N \tag{3.11}$$

that may in theory be solved (by the earlier assumption that $I - P_N K$ is invertible) for the unknown coefficients a_j. This fixes our Galerkin approximation p_N via (3.10).

The basis $\{\chi_1, \ldots, \chi_N\}$ was chosen in equations (3.9) to (3.11) for continuity of notation, although any basis of E_N would suffice.

Thus, every time we wish to find a Galerkin approximation we must solve an $N \times N$ system (3.11). Although achievable, this is computationally very expensive and sensitive to the value of N. Another drawback is that information is non-transferable, that is, work done to calculate p_N will not aid calculation of p_{N+1}. For these reasons we wish to keep N small, and so disregard the afore-mentioned approach, of expanding the subspace, for finding a sequence of approximations that tends to the exact solution as, in this approach, N is unrestricted.

Is it possible to improve on the Galerkin approximation p_N without increasing the dimension N? We address this in what remains of this chapter.

Definition If T is a bounded, linear operator on a Hilbert space, then the *adjoint* of T is the unique bounded linear operator T^* with the property

$$(Tu, v) = (u, T^* v)$$

for all $u, v \in \mathscr{H}$.

If $T = T^*$, then T is said to be *self-adjoint*.

3.2 Iterated Galerkin

The structure of (2.4) lends itself to an iteration process. Let us take the Galerkin approximation p_N as our 'initial guess', and iterate using (2.4) to produce the approximation

$$\hat{p}_N = f + K p_N. \tag{3.12}$$

This intuitively shares more of the exact solution's structure.

Before investigating the effects of this iteration, (3.12) is introduced independent of the Galerkin approximation.

3.2.1 Degenerate kernels

A kernel $k(x,t)$ is known as *degenerate* if it induces a linear operator K of finite rank. Such kernels take the form

$$k(x,t) = \sum_{i=1}^{M} b_i(x)k_i(t)$$

for functions b_i and k_i, and some finite integer M. As mentioned in §2 there exists an analytic framework for solving (2.4) if K has finite rank. Although it has been assumed from the start that operators under consideration take infinite rank, it has also been assumed that these operators are compact. As such, all K under consideration are the limit of a sequence of finite rank operators $\{K_n\}$. A natural approach to approximate ϕ would be to use the exact solution of

$$\phi_N = f + K_N\phi_N \tag{3.13}$$

where K_N is an operator of finite rank N that approximates K in a way that will shortly be defined.

The effect of any compact, bounded operator on an element may be expressed as the, possibly finite, series

$$K\varphi = \sum_{n=1}^{\infty} e_n(\varphi, v_n)\varphi_n,$$

where v_n are the eigenvectors of the operator K^*K, e_n^2 the corresponding eigenvalues, and $\varphi_n = e_nKv_n$[1]. For K of infinite rank, this series is infinite.

Now, we may form a degenerate, linear integral operator, K_N, by truncating this series in some fashion. Let us suppose

$$K_N\varphi = \sum_{i=1}^{N} e_n(\varphi, v_n)\varphi_n,$$

then

$$Kv_n = K_Nv_n, \qquad (n = 1, \dots N).$$

In this sense K_N approximates K. Further, $\{\varphi_1, \dots, \varphi_N\}$ is an N-dimensional basis for the operator K_N, and hence so is $\{Kv_1, \dots, Kv_N\}$. Therefore, for any v_j, there exist scalars α_i such that

$$\alpha_1K_Nv_1 + \dots \alpha_NK_Nv_N = K_Nv_j \qquad j \in \mathbb{N}. \tag{3.14}$$

Suppose $j > N$, by using the linearity of K_N

$$\begin{aligned} K_N(\alpha_1v_1 + \dots \alpha_Nv_N - v_j) = 0 \\ \Rightarrow \alpha_1v_1 + \dots \alpha_Nv_N - v_j \in \text{null}(K_N) \end{aligned}$$

but

$$v_i \notin \text{null}(K_N) \qquad i = 1, \dots, N.$$

Hence, the scalars α_i are necessarily zero, and (3.14) becomes

$$K_Nv_j = 0 \qquad (j > N). \tag{3.15}$$

[1] Taken from theorem 4.18 of [1].

If the approximation is written

$$\phi_N = v_0 + \sum_{i=1}^{\infty} a_i v_i,$$

where $Kv_0 = K_N v_0 = 0$, then by (3.15)

$$\begin{aligned} K_N \phi_N &= \sum_{i=1}^{\infty} a_i K_N v_i \\ &= \sum_{i=1}^{N} a_i K_N v_i. \end{aligned} \tag{3.16}$$

Substituting (3.16) into (3.13), we obtain the expression

$$\phi_N = f + \sum_{i=1}^{N} a_i K v_i \tag{3.17}$$

and all that must be done to find ϕ_N is to fix the coefficients $a_1, \ldots, a_N$. This can be acheived by taking the inner product of (3.17) with the first N v_j

$$\begin{aligned} (\phi_N, v_j) &= (f, \chi_j) + \sum_{i=1}^{N} a_i (K v_i, v_j) \quad j = 1, \ldots, N \\ \Rightarrow \sum_{i=1}^{\infty} a_i (v_i, v_j) &= (f, v_j) + \sum_{i=1}^{N} a_i (K v_i, v_j) \quad j = 1, \ldots, N \\ \Rightarrow \sum_{i=1}^{N} a_i \big((v_i, v_j) - (K v_i, v_j) \big) &= (f, v_j) \quad j = 1, \ldots, N. \end{aligned} \tag{3.18}$$

Comparison of (3.18) with (3.11) shows these a_i are exactly the coefficients used to define the Galerkin approximation p_N, where the same basis $\{v_1, \ldots, v_N\}$ is used. However, the approximations are not the same, as the coefficients are used in different expressions. In fact, ϕ_N is equal to the iterated Galerkin approximation

$$\begin{aligned} \phi_N &= f + \sum_{i=1}^{N} a_i K v_i \\ &= f + K p_N \\ &= \hat{p}_N \\ &\neq p_N. \end{aligned}$$

Note however, the strict conditions on the subspace. This does not exclude the possibility that there are other vectors that would satisfy the required relations.

If, instead the approximation ϕ_N is chosen to lie in the space $\{v_1, \ldots, v_N\}$, i.e.

$$\phi_N = \sum_{n=1}^{N} a_n v_n,$$

then the constraints on the v_i are not required. This approach is discussed by Burton [7] and credited to Sloan, Datyner and Burn.

In either case, (3.12) holds as an approximation ϕ_N to ϕ independently of the Galerkin approximation. For the purposes of this project it suits us to consider (3.12) as the iterate $\hat{p}_N$.

3.2.2 Analysis of $\hat{p}_N$

Can it be shown that the iterate $\hat{p}_N$ is an improvement over p_N as an approximation to ϕ?

To answer this we seek an expression for the new error $\hat{p}_N - \phi$ in terms of the old error $p_N - \phi$. Firstly, we need to derive an expression for $\hat{p}_N$ that will simplify this task. To do this, an auxiliary operator equality is required, namely

$$\begin{aligned} P_N(I - KP_N) &= P_N - P_NKP_N = (I - P_NK)P_N \\ \Rightarrow P_N(I - KP_N)^{-1} &= (I - P_NK)^{-1}P_N. \end{aligned} \tag{3.19}$$

Note that, this equality also tells us that the existence of the inverses of $I - KP_N$ and $I - P_NK$ are co-dependent.

Combining (3.4) and (3.12) gives

$$\hat{p}_N = f + K(I - P_NK)P_Nf,$$

then rearranging the second term on the right hand side using (3.19) we obtain

$$\begin{aligned} \hat{p}_N &= f + KP_N(I - KP_N)^{-1}f \\ &= \big((I - KP_N) + KP_N\big)(I - KP_N)^{-1}f \\ &= (I - KP_N)^{-1}f. \end{aligned} \tag{3.20}$$

From here on (3.20) will be used to define the iterate.

Now, analysing the error using (3.20) and (2.3) gives

$$\begin{aligned} \hat{p}_N - \phi &= (I - KP_N)^{-1}f - (I - K)^{-1}f \\ &= (I - KP_N)^{-1}\big((I - K) - (I - KP_N)\big)(I - K)^{-1}f \\ &= -(I - KP_N)^{-1}K(I - P_N)\phi. \end{aligned} \tag{3.21}$$

This is not quite what is required. It is an expression for the new error in terms of the exact solution only, but may be overcome by noting that from (3.8)

$$\begin{aligned} (I - P_N)p_N &= 0 \\ \Rightarrow (I - KP_N)^{-1}K(I - P_N)p_N &= 0. \end{aligned} \tag{3.22}$$

Thus, adding the left hand side of (3.22) to the right hand side of (3.21) we obtain

$$\begin{aligned} \hat{p}_N - \phi &= (I - KP_N)^{-1}K(I - P_N)p_N - (I - KP_N)^{-1}K(I - P_N)\phi \\ &= (I - KP_N)^{-1}K(I - P_N)(p_N - \phi), \end{aligned} \tag{3.23}$$

as required, expressing the new error in terms of the old. For brevity, (3.23) will be written

$$\hat{p}_N - \phi = S(p_N - \phi), \tag{3.24}$$

where

$$S = (I - KP_N)^{-1}K(I - P_N) \tag{3.25}$$

may be called the *Sloan operator.*

Note that, by(3.19), the existence of the Sloan operator implicitly requires the existence of the Galerkin approximation.

What can this expression tell us?

Taking norms on (3.24) and applying the Schwarz inequality

$$||\hat{p}_N - \phi|| \leq ||S|| ||p_N - \phi||. \tag{3.26}$$

Recall that a good approximation is one that minimises the norm of the error, then (3.26) tells us that $\hat{p}_N$ is an improvement to p_N, i.e.

$$||\hat{p}_N - \phi|| < ||p_N - \phi||$$

if the Sloan operator satisfies the inequality

$$||S|| < 1. \tag{3.27}$$

Sadly, this is not always the case.

Fortunately though, it is a condition that can be controlled through our choice of subspace E_N. This is due to the appearance of the orthogonal projection P_N in the definition (3.25) of S. Note that, for a fixed K, the Sloan operator is dependent only on P_N. In §3.1 we saw that P_N acts like the identity map on E_N. Thus, as the dimension N increases, P_N tends towards the identity map on $\mathscr{H}$. So, on any finite dimensional subspace

$$||I - P_N|| \to 0 \qquad \text{as } N \to \infty. \tag{3.28}$$

Taking norms and using the Schwarz inequality on (3.25)

$$||S|| \leq ||(I - KP_N)^{-1}|| ||K(I - P_N)||. \tag{3.29}$$

Although K has infinite rank, its compactness is enough to ensure[2] that from (3.28) on $\mathscr{H}$

$$||K(I - P_N)|| \to 0 \qquad \text{as } N \to \infty.$$

Therefore, the factor $||K(I - P_N)||$ can be made arbitrarily small via choice of a suitably large subspace.

In turn, $||S||$ can be made arbitrarily small, if the first factor in (3.29) does not become unbounded. This is equivalent to requiring the operator $I - KP_N$ to be invertible that by equality (3.19) is equivalent to the operator $I - P_N K$ being invertible. This was an assumption made in §3.1 for the existence of the Galerkin approximation. As $N \to \infty$, the operator $(I - P_N K)^{-1}$ tends to $(I - K)^{-1}$, which by assumption of uniqueness of a solution has a bounded norm. Hence, as we expand E_N, the norm of first factor, $(I - P_N K)^{-1}$, of S tends to a bounded limit.

So far in this section, it has been shown that under a condition we control we may iterate the Galerkin approximation p_N to produce an improved approximation $\hat{p}_N$. Computationally, this improvement is achieved at a low cost.

In practice, the dimension N does not have to be particularly large to ensure we have a suitable operator S.

[2] for proof see [1] Lemma 7.2.

3.2.3 Does the improvement repeat?

A natural question to ask is whether a further improvement will result if a second iteration of the form (3.12) is performed. Let the second iterate be

$$\begin{aligned} \hat{\hat{p}}_N &= f + K\hat{p}_N \\ &= f + K(I - KP_N)^{-1}f \end{aligned} \tag{3.30}$$

by (3.20).

Again we seek an expression for the new error $\hat{\hat{p}}_N - \phi$ in terms of the old error $\hat{p}_N - \phi$. Using (3.30) and (2.3)

$$\begin{aligned} \hat{\hat{p}}_N - \phi &= f + K(I - KP_N)^{-1}f - \phi \\ &= (I - K)\phi + K(I - KP_N)^{-1}f - \phi \\ &= -K\big(I - (I - KP_N)^{-1}(I - K)\big)\phi \\ &= -K(I - KP_N)\big((I - KP_N) - (I - K)\big)\phi \\ &= -K\underbrace{(I - KP_N)^{-1}K(I - P_N)\phi}_{=\phi - \hat{p} \text{ by } (3.21)}. \end{aligned}$$

Hence

$$\hat{\hat{p}}_N - \phi = K(\hat{p} - \phi).$$

Applying the argument used to deduce that $\hat{p}_N$ is an improvement over p_N if $\|S\| < 1$ from (3.24), from (3.31) we find that $\hat{\hat{p}}_N$ is an improvement over $\hat{p}_N$ if $\|K\| < 1$. However, this is where the similarity to the previous case ends. For, we may not manipulate the operator K; it is dictated by the problem in hand. Although in some circumstances we may be dealing with a problem (2.4) in which $\|K\| < 1$ and a second iteration will produce a further improvement, in general this is not the case. For this reason, iterated Galerkin does not provide a suitable means of calculating an approximation to an arbitrary degree of accuracy. In general we still rely on improvements to the Galerkin approximation to improve our iterated approximation, an approach that was earlier disregarded. The iterated Galerkin approximation does however provide an essential intermediate step for the re-iterated Galerkin approximation.

From now on it will be assumed that the dimension N is fixed such that $\|S\| < 1$. For reasons that will soon become clear, the approximations will be re-labeled as

$$\begin{aligned} p_0 &= p_N, \\ \hat{p}_0 &= \hat{p}_N. \end{aligned}$$

3.3 The re-iterated Galerkin method

The re-iterated Galerkin method picks up after the iterated Galerkin approximation $\hat{p}_0$, of the last section, has been made. It seeks to improve this approximation by adding on an approximation to the pointwise error e_0, such that

$$\phi = \hat{p}_0 + e_0. \tag{3.31}$$

How should this approximation be made? Firstly, a useable expression for e_0 must be found. The above equality (3.31) is not suitable, as it involves the unknown ϕ. However, (3.31) may be manipulated as follows. We note that

$$\begin{aligned}
\phi &= \hat{p}_0 + e_0 \\
\Rightarrow A\phi &= A\hat{p}_0 + Ae_0 \\
\Rightarrow Ae_0 &= A\phi - A\hat{p}_0 \\
&= f - A\hat{p}_0. \qquad (3.32)
\end{aligned}$$

The right hand side of (3.32) is the residual error in the approximation $\hat{p}_0$, that is, importantly, calculable. If this residual error is denoted by $\hat{r}_0$, (3.32) becomes

$$Ae_0 = \hat{r}_0, \qquad (3.33)$$

an integral equation to be solved for the pointwise error in the approximation $\hat{p}_0$.

Comparison with (2.4) shows (3.33) to be almost identical to our original equation, differing only in the free term. It is possible to follow the procedure that produced the approximation $\hat{p}_0$ to ϕ, merely substituting $\hat{r}_0$ for f, to find an approximation to e_0. That is, we first form the Galerkin approximation to e_0, using the subspace E_N. Call this p_1, from (3.4)

$$p_1 = (I - P_N K)^{-1}\hat{r}_0.$$

Then we iterate to find

$$\begin{aligned}
\hat{p}_1 &= \hat{r}_0 + Kp_1 \\
&= (I - KP_N)^{-1}\hat{r}_0.
\end{aligned}$$

Note that, as we have only changed the free term, and have done nothing to the operator K nor the subspace E_N, the Sloan operator S is the same one that appears in (3.24). By assumption $||S|| < 1$, and hence

$$||\hat{p}_1 - e_0|| < ||p_1 - e_0||,$$

i.e. the iterate $\hat{p}_1$ is an improvement over p_1 as an approximation to e_0.

Our new approximation is $\hat{p}_0 + \hat{p}_1$. Can it be shown that this is an improvement to $\hat{p}_0$? Again this is established by error analysis, since

$$\begin{aligned}
(\hat{p}_0 + \hat{p}_1) - \phi &= (I - KP_N)^{-1}\hat{r}_0 + (\hat{p}_0 - \phi) \\
&= (I - KP_N)^{-1}(K - I)(\hat{p}_0 - \phi) + (\hat{p}_0 - \phi) \\
&= (I - KP_N)^{-1}K(I - P_N)(\hat{p}_0 - \phi) \\
&= S(\hat{p}_0 - \phi). \qquad (3.34)
\end{aligned}$$

Therefore

$$\begin{aligned}
||(\hat{p}_0 + \hat{p}_1) - \phi|| &\leq ||S|| ||\hat{p}_0 - \phi|| \\
&\leq ||S||^2 ||p_0 - \phi||.
\end{aligned}$$

Again using the assumption that $||S|| < 1$, the new approximation $\hat{p}_0 + \hat{p}_1$ is an improvement over $\hat{p}_0$. More than this, the improvement has an upper bound of $||S||$. This is the same upper bound on the improvement of $\hat{p}_0$ over p_0, hence our new approximation is an improvement over the original Galerkin approximation with upper bound $||S||^2$.

The algebraic manipulations leading to (3.34) do not involve any properties of the original approximation $\hat{p}_0$, rather that we add on $\hat{p}_1$ to the original approximation, where $\hat{p}_1$ is the iterated Galerkin approximation to the pointwise error, made using (3.33). As such, it is clear that we are able to repeat the above procedure indefinitely, substituting in our latest approximation for $\hat{p}_0$, and the improvement will hold. This may be shown formally using induction.

Let the n^{th} approximation to ϕ be

$$\hat{\sigma}_n = \sum_{i=0}^{n} \hat{p}_i,$$

where p_0 and $\hat{p}_0$ are as previously defined, and

$$\begin{aligned} \hat{p}_j &= \hat{r}_{j-1} + Kp_j \\ &= (I - KP_N)^{-1}\hat{r}_{j-1} \end{aligned}$$

is the iterate of

$$p_j = (I - KP_N)P_N\hat{r}_{j-1}, \qquad (j = 1, \dots,)$$

the Galerkin approximation to the current error $e_{j-1} = \phi - \hat{\sigma}_{j-1}$ in E_N, obtained from the equation

$$Ae_{j-1} = \hat{r}_{j-1},$$

which says that the current pointwise error is the solution of an operator equation, identical to the original equation (2.4), excepting the free term that has been replaced by the current residual error.

Suppose that the current approximation is $\hat{\sigma}_n$, and up until this point each new approximation has been an improvement over the previous by a factor bounded above by $||S||$, i.e.

$$||\hat{\sigma}_j|| \leq ||S||||\hat{\sigma}_{j-1}||, \qquad (j = 1, \dots, n)$$

(the case $n = 1$ was proved in (3.34)).

What about $\hat{\sigma}_{n+1}$? Using the above definitions

$$\begin{aligned} \hat{\sigma}_{n+1} - \phi &= \hat{p}_{n+1} + \hat{\sigma}_n - \phi \\ &= (I - KP_N)^{-1}\hat{r}_n + \hat{\sigma}_n - \phi \\ &= (I - KP_N)^{-1}(K - I)(\hat{\sigma}_n - \phi) + \hat{\sigma}_n - \phi \\ &= (I - KP_N)^{-1}\big((K - I) + (I - KP_N)\big)(\hat{\sigma}_n - \phi) \\ &= (I - KP_N)^{-1}K(I - P)(\hat{\sigma}_n - \phi) \\ &= S(\hat{\sigma}_n - \phi), \end{aligned}$$

hence the Sloan operator maps the current error onto the new error.

Thus, by induction

$$\begin{aligned} ||\hat{\sigma}_{n+1} - \phi|| \leq & \quad ||S||||\hat{\sigma}_n - \phi|| && n \geq 0 \\ \leq & \quad ||S||^{n+1}||\hat{\sigma}_0 - \phi|| && n \geq 0 \\ \leq & \quad ||S||^{n+2}||p - \phi|| && n \geq 0, \end{aligned}$$

the method can be repeated as required.

It is important to emphasise the computational efficiency of the re-iterated Galerkin method. The expense of an approximation method is largely dictated by the number of and size of systems of equations that arise in their evaluation. In re-iterated Galerkin there are two types of approximation that must be made, the Galerkin approximation and the iterated Galerkin approximation. Only in the ordinary Galerkin approximation are we required to solve a system of equations. As the subspace E_N has been fixed, this will always be an $N \times N$ system of the form (3.11). In addition, as the operator A is the same in each step of the re-iterated Galerkin method, the co-efficients on the left hand side of (3.11) are the same. This means we need only calculate the inverse of the resulting matrix once and re-apply it to vectors resulting from the right hand side of (3.11) with the free term changed to the appropriate residual error. This constitutes a substantial computational saving.

This computational saving relies on only one subspace being used. Our ability to only use one subspace yet still incur improvements, is a consequence of the intermediate iterations. To see this, suppose we were to attempt the re-iteration process without using the iterations $\hat{p}_j$, then

$$r_0 = f - Ap_0$$

and

$$\begin{aligned} P_N r_0 &= P_N f - P_N A p_0 \\ &= \underbrace{P_N f + P_N K p_0}_{=p_0 \text{ by } (3.7)} - P_N p_0 \\ &= p_0 - p_0 \\ &= 0. \end{aligned}$$

Our Galerkin approximation p_1 in E_N to the current error satisfies

$$\begin{aligned} (I - P_N K)p_1 = P_N r_0 &= 0 \\ \Rightarrow p_1 &= 0 \end{aligned}$$

and the new approximation

$$p_0 + p_1 = p_0$$

is equal to the previous approximation. Thus, the re-iteration process breaks down at the first step without the use of the $\hat{p}_j$.

Another practical advantage of the re-iterated Galerkin method is that inaccuracies are not propagated. This is because improvements are calculated relative to the error in our current approximation.

3.3.1 A more accurate improvement factor

By using the Schwartz inequality to deduce condition (3.27), which if satisfied ensures iterations will improve our approximation, a certain amount of information is lost. For this reason, (3.27) is only a sufficient condition. We are only able to say that *if* the Sloan operator has a norm value less than one, *then* an iteration will improve matters. However, this does not preclude the possibility that iteration could improve the approximation although $\|S\| \geq 1$.

To remedy this situation, consider our approximation $\hat{\sigma}_n = \sum_{i=0}^{n} \hat{p}_i$. Rather than ask whether each successive approximation is an improvement, we seek to determine if the series representation of the approximation converges and if it does is its limit ϕ?

Analysis of the convergence is eased if $\hat{\sigma}_n$ is written as a geometric progression. First, note that

$$\begin{aligned} \hat{p}_i &= \hat{r}_{i-1} + Kp_i \\ \Rightarrow \quad P_N\hat{p}_i &= P_N\hat{r}_{i-1} + P_N K p_i \\ \Rightarrow \quad P_N\hat{p}_i &= p_i, \end{aligned}$$

so that the projection of the iterate is the Galerkin approximation, a result of independent interest. Now,

$$\begin{aligned} S\hat{p}_i &= (I - KP_N)^{-1}K(I - P_N)\hat{p}_i \\ &= (I - KP_N)^{-1}K(\hat{p}_i - p_i) \\ &= (I - KP_N)^{-1}(\hat{r}_{i-1} - \hat{p}_i + Kp_i) \\ &= (I - KP_N)^{-1}\hat{r}_i \\ &= \hat{p}_{i+1}. \end{aligned}$$

Therefore, by induction

$$\hat{p}_i = S^i \hat{p}_0 \tag{3.35}$$

and the approximation

$$\hat{\sigma}_n = \sum_{i=0}^{n} S^i \hat{p}_0. \tag{3.36}$$

Call $\mathscr{H}'$ the smallest Hilbert subspace of $\mathscr{H}$ (possibly $\mathscr{H}$ itself) containing all of the terms in the series, $S^i e_0$, on which S may be considered as an operator. By application of the root test, the condition

$$\lim_{n\to\infty} \|S^n\|^{\frac{1}{n}} < 1 \tag{3.37}$$

can be deduced for convergence of the series (3.36) as $n \to \infty$. It is well known (see [3] for example) that the limit to the left hand side of (3.37) is equal to (and sometimes used to define) $\rho_{\mathscr{H}'}(S) = \max\{|\nu| : \exists\varphi \in \mathscr{H}'$ such that $S\varphi = \nu\varphi\}$ the *spectral radius* of S over $\mathscr{H}'$. Hence, the condition that (3.36) converges is

$$\rho_{\mathscr{H}'}(S) < 1$$

This is a weaker condition than $\|S\| < 1$, as $\rho_{\mathscr{H}'}(S) \leq \|S\|$. In fact

$$\rho_{\mathscr{H}'}(S) \leq \rho_{\mathscr{H}}(S) \leq \|S\|,$$

and $\rho_{\mathscr{H}}(S) = \|S\|$ only if S is self-adjoint.

If (3.36) converges, necessarily $\hat{p}_n \to 0$ as $n \to \infty$. From the definition of $\hat{p}_n$, $\hat{p}_n \to 0$ is equivalent to $\hat{r}_{n-1} \to 0$. The residual error tending to zero is enough to guarantee that the limit of the series is ϕ, and the approximations $\hat{\sigma}_n$ converges to the exact solution.

In §3.3 we saw that $e_{n+1} = Se_n$. Using this equality, a similar expression to (3.36) is obtainable for the sum of the error terms

$$\sum_{i=0}^{n} e_i = \sum_{i=0}^{n} S^i e_0.$$

Assuming convergence, ϕ is the limit of the re-iterated Galerkin approximations, all of which belong to $\mathscr{H}'$ by definition, and the closed property of Hilbert spaces then ensures

$$\phi \in \mathscr{H}'.$$

Hence, for the same reasons, this series converges as $n \to \infty$ if and only if $\rho_{\mathscr{H}'}(S) < 1$. This is equivalent to saying that the series

$$\sum \|e_n\|$$

converges if and only if $\rho_{\mathscr{H}'}(S) < 1$. We may now deduce that the ratio of successive terms

$$\frac{\|e_{n+1}\|}{\|e_n\|} \to \rho_{\mathscr{H}'}(S) \qquad \text{as } n \to \infty.$$

At worst, we can expect

$$\frac{\|e_{n+1}\|}{\|e_n\|} \to \rho_{\mathscr{H}}(S) \qquad \text{as } n \to \infty.$$

Thus, the *theoretical improvement factor* is $\rho_{\mathscr{H}'}(S)$, the spectral radius of the Sloan operator over $\mathscr{H}'$. This result has been verified in practice (see [2]), although for reasons that will be discussed in §6.4 it is shown in terms of the ratio of the norm of successive residual errors. However, we must be careful to note that the theoretical improvement factor is an 'eventual trend'. Early iterations may not bear any relation to this value, with the ratio of successive error terms being bounded only by $\|S\|$. This means, if $\|S\| > 1$ and $\rho_{\mathscr{H}}(S) < 1$, the re-iterated Galerkin approximation could get worse before it starts improving.

It is important to remember that the last results have been made on a purely theoretical basis. In practice, where computational round-off occurs, and errors may be magnified by $\|S\|$, it would be unwise to expect the re-iterated Galerkin method to converge if the spectral radius of S is only narrowly less than one but $\|S\|$ is appreciably greater than one. Also, the above relations (3.35) and (3.36) are not suitable means of calculating approximations. They break down under the effects of round-off error, and the property of not propagating errors, described earlier, is lost.

For the benefit of the numerical analyst, the following section is an aside, in which the iteration step is introduced in what should be a more familiar context.

3.3.2 Iteration as a preconditioner

A preconditioner is a way of manipulating a difficult problem in order to ease an attempt to solve it. These are widely used in practical problems.

Note that

$$\begin{aligned} (I-K)\phi &= f \\ \Rightarrow (I-KP_N)^{-1}(I-K)\phi &= (I-KP_N)^{-1}f, \end{aligned}$$

and

$$\begin{aligned} I - S &= I - (I - KP_N)^{-1}K(I - P_N) \\ &= (I - KP_N)^{-1}(I - K). \end{aligned}$$

Combining these results gives

$$\begin{aligned} (I - S)\phi &= (I - KP_N)^{-1}f \\ &= \hat{p}. \end{aligned} \tag{3.38}$$

This says that the iteration step is equivalent to preconditioning the original equation by multiplying by the operator $(I - KP_N)^{-1}$. This has the effect of replacing the operator K with S. As S may always be chosen such that $\|S\| < 1$, the preconditioned equation (3.38) is open to repeated application of the Sloan iteration

$$\hat{p} = f + Kp. \tag{3.39}$$

This was earlier disregarded due to the requirement that $\|K\| < 1$. Indeed, it easily verified that

$$\hat{\sigma}_{n+1} = \hat{p}_0 + S\hat{\sigma}_n.$$

This confirms that the re-iterated Galerkin method is in fact the combination of Sloan iteration and preconditioning.

This point will remain as an aside, and will only be revisited briefly in §7.

Chapter 4

Quantities of interest

The links between deriving an approximation to the unknown function ϕ and approximating a quantity that involves this function, as one may expect, are strong. The latter however, is worthy of treatment in its own right. We shall focus on a method analogous to the re-iterated Galerkin method.

For the purposes of our example, the quantities of interest, which involve an unknown function, are the definitions (1.16) and (1.17) of the wave amplitudes R and T. In the next chapter it will be shown that they may be written as inner products where one of the arguments in each is the unknown ϕ.

4.1 Variational principles

Our goal is to approximate inner products of the form

$$(\phi, g) \tag{4.1}$$

where ϕ is the unknown solution of $A\phi = f$ in $\mathscr{H}$ and g is an arbitrary, known member of $\mathscr{H}$. To do this we may exploit methods of variational calculus. In particular, we will need to know that a functional J is defined to be *stationary* at p if *any* variation δp causes at most a second order variation in J. This may be written symbolically as

$$J(p + \delta p) = J(p) + O(\|\delta p\|^2) \tag{4.2}$$

The multi-dimensional definition follows in the obvious fashion.

Under suitable choice of functional, the stationary value coincides with the desired inner product. Hence, methods used to approximate this stationary value will also produce approximations to (4.1). The above definition (4.2) means that the error in an approximation to a stationary point becomes a second order error in the approximation to the stationary value. This extra degree of accuracy is gained with no extra effort, and is the reason for using this method rather than simply evaluating the inner product (4.1) with ϕ replaced by an approximation, for which we know nothing of its degree of accuracy. The process of obtaining an improvement in the order of accuracy of a calculation over the original approximation is known as *superconvergence.*

Any Hilbert space operator A under consideration may be used to define the operator L that maps any $(p, q, f, g) \in \mathscr{H}^4$ to $(f, q) + (p, g) - (Ap, q)$. If we consider f and g to be fixed, with f the free term of (2.4) and g as in (4.1), then

$$\begin{aligned} L : \mathscr{H} \oplus \mathscr{H} &\to \mathbb{C} \\ (p, q) &\mapsto (f, q) + (p, g) - (Ap, q). \end{aligned}$$

Although L is now only dependent on p and q, f and g are still arbitrary. To emphasise this, the operator will be written

$$L = L(p, q, f, g).$$

Consider the stationary point (ϕ, ψ) (the reason for the clash in notation will soon become clear) of L , and an arbitrary variation $(\phi + \delta\phi, \psi + \delta\psi)$, then

$$\begin{aligned} L(\phi + \delta\phi, \psi + \delta\psi, f, g) &= (f, \psi + \delta\psi) + (\phi + \delta\phi, g) - (A(\phi + \delta\phi), \psi + \delta\psi) \\ &= (f, \psi) + (f, \delta\psi) + (\phi, g) + (\delta\phi, g) \\ &\quad -(A\phi, \psi) - (A\phi, \delta\psi) - (A\delta\phi, \psi) - (A\delta\phi, \delta\psi) \\ &= L(p, q, f, g) \\ &\quad +(f, \delta\psi) + (\delta\phi, g) - (A\phi, \delta\psi) - (A\delta\phi, \psi) \qquad (4.3) \\ &\quad -(A\delta\phi, \delta\psi). \end{aligned}$$

To satisfy the requirements of a stationary point, the first order variations (line (4.3)) must vanish, therefore

$$\begin{aligned} 0 &= (f, \delta\psi) + (\delta\phi, g) - (A\phi, \delta\psi) - (A\delta\phi, \psi) \\ &= (f - A\phi, \delta\psi) + (\delta\phi, g - A^*q). \qquad (4.4) \end{aligned}$$

Equality (4.4) holds for all $(\delta\phi, \delta\psi) \in \mathscr{H} \oplus \mathscr{H}$ *if and only if*

$$A\phi = f \quad \text{and} \quad A^*\psi = g. \qquad (4.5)$$

By assumption, ϕ is unique, and it follows by a simple corollary of Theorem 4.11 of [1] that ψ is also unique. So, the functional L is stationary at the unique point (ϕ, ψ), such that ϕ and ψ are the unique solutions of (4.5).

The corresponding unique stationary value

$$\begin{aligned} L(\phi, \psi, f, g) &= (f, \psi) + (\phi, g) - (A\phi, \psi) \\ &= (f - A\phi, \psi) + (\phi, g) \\ &= (\phi, g) \end{aligned}$$

is the required inner product.

Finding the stationary value of L directly is no easier than solving (2.4) exactly. Again, our attention must turn to approximation methods. It should be clear that approximations to the stationary value of L will involve approximations to the ϕ.

4.1.1 The Rayleigh-Ritz method

An obvious candidate to approximate the stationary value of L (or any functional) consistent with our previous work, would be to reduce it to a problem of finite dimensions. Let $E_N \oplus F_M \subset$

$\mathscr{H} \oplus \mathscr{H}$, where E_N and F_M are finite dimensional subspaces of $\mathscr{H}$, of dimension N and M respectively, and

$$E_N = \operatorname{span}\{\chi_1, \ldots, \chi_N\}$$
$$F_M = \operatorname{span}\{\xi_1, \ldots, \xi_M\}.$$

We seek an approximation (p_N, q_M) to (ϕ, ψ), where

$$p_N = \sum_{i=1}^{N} a_i \chi_i \in E_N; \qquad q_M = \sum_{i=1}^{M} b_i \xi_i \in F_M.$$

This defines $p_N = p_N(a_1, \ldots, a_N)$ a function of the N variables a_i and $q_M = q_M(b_1, \ldots, b_M)$ a function of the M variables b_i. The functional L on the subspace $E_N \oplus F_M$ becomes the function $L = L(a_1, \ldots, a_N, b_1, \ldots, b_M)$ of the $N + M$ variables a_i, b_j $i = 1, \ldots, N; j = 1, \ldots, M$. Explicitly

$$L(a_1, \ldots, a_N, b_1, \ldots, b_M) = \sum_{j=1}^{M} \bar{b}_j (f, \xi_j) + \sum_{i=1}^{N} a_i (\chi_i, g) - \sum_{i=1}^{N} \sum_{j=1}^{M} a_i \bar{b}_j (A\chi_i, \xi_j).$$

Using basic calculus, the stationary values of L can be found by setting the derivatives with respect to the a_i and $\bar{b}_j$ to zero, i.e.

$$\begin{aligned} \frac{\partial L}{\partial a_i} &= (\chi_i, g) - \sum_{j=1}^{M} \bar{b}_j (A\chi_i, \xi_j) = 0, \quad i = 1, \ldots, N, \\ \frac{\partial L}{\partial \bar{b}_j} &= (f, \xi_j) - \sum_{i=1}^{N} a_i (A\chi_i, \xi_j) = 0, \quad j = 1, \ldots, M, \end{aligned}$$

or

$$\begin{aligned} (\chi_i, g - A^* q_N) = 0, \quad i = 1, \ldots, N, \\ (f - Ap_N, \xi_j) = 0, \quad j = 1, \ldots, M. \end{aligned}$$

Comparison with (3.9) shows that p_N and q_M are not the respective Galerkin approximations in E_N and F_M to the equations (4.4), unless $N = M$ and $\chi_i = \xi_i$, which would mean the subspaces E_N and F_M were dependent on one another. This can not generally be expected to be the case. As has been discussed already, a subspace is chosen to suit the specific equation, i.e. E_N and F_M are not dependent on each other.

Therefore, Rayleigh-Ritz does not provide a means for use of the re-iterated Galerkin method and will not be explored further.

4.1.2 The application of the re-iterated Galerkin method to the functional L

Rather than finding the stationary points of L within a finite dimensional subspace, let us examine the effect of evaluating L at the sequence of unrelated co-ordinates $(\hat{\sigma}_n, \hat{\tau}_n)$, where $\hat{\sigma}_n = \sum_{i=0}^{n} \hat{p}_i$ is the n^{th} re-iterated Galerkin approximation to $\phi = f + K\phi$ resulting from the Galerkin approximation $p_0 \in E_N$ and $\hat{\tau}_n = \sum_{i=0}^{n} \hat{q}_i$ is the n^{th} re-iterated Galerkin approximation to $\psi = g + K^*\psi$ resulting from the Galerkin approximation $q_0 \in F_M$. Let P_N and S be as in (3.1) and (3.25) respectively, and Q_M and T be the corrosponding orthogonal projection and Sloan operator for the subspace F_M, i.e.

$$T = (I - K^*Q_M)^{-1}K^*(I - Q_M).$$

Recall that, it has been assumed that $\|S\| < 1$. By the same argument, it can also be assumed that $\|T\| < 1$. Following the theory of §3.3 we have

$$\|\phi - \hat{\sigma}_n\| \leq \|S\|^{n+1}\|p_0 - \phi\| \tag{4.6}$$

and

$$\|\phi - \hat{\tau}_n\| \leq \|T\|^{n+1}\|q_0 - \phi\|. \tag{4.7}$$

The structure of the re-iterated Galerkin method is nicely reflected in the addition formulae

$$L(p + p', q + q', f, g) = L(p, q, f, g) + L(p', q', f - Ap, g - A^*q), \qquad \forall p, p', q, q' \in \mathscr{H}, \tag{4.8}$$

which says that improvements to an approximation of (ϕ, g) are made by approximating the stationary value $(p', q') = (\phi', \psi')$ of $L = L(p', q', f - Ap, g - A^*q)$, that must satisfy

$$A\phi' = f - Ap \qquad A^*\psi' = g - A^*q.$$

Moreover, equality (4.8) is of practical importance. A feature of the re-iterated Galerkin method, when approximating ϕ, was that the computational cost of calculating an improvent did not increase. Is there an analogous result here? Calculating directly, that is,

$$L(\hat{\sigma}_n, \hat{\tau}_n, f, g) = \sum_{i=0}^{n}(\hat{p}_i, g) + \sum_{i=0}^{n}(f, \hat{q}_i) - \sum_{i=0}^{n}\sum_{j=0}^{n}(A\hat{p}_i, \hat{q}_j)$$

constitutes $n^2 + 4n + 3$ inner products, of which $n^2 + 2n$ are known from previous calculations. So, $2n+3$ additional inner products must be calculated to make each improvement. Importantly $2n + 3 \to \infty$ as $n \to \infty$, and the cost of calculating improvements increases.

Compare this with calculating improvements using (4.8), so that

$$L(\hat{\sigma}_n, \hat{\tau}_n, f, g) = L(\hat{p}_0, \hat{q}_0, f, g) + \sum_{i=1}^{n} L(\hat{p}_i, \hat{q}_i, \hat{r}_{i-1}, \hat{s}_{i-1}),$$

where $\hat{r}_i = f - A\hat{\sigma}_i$ and $\hat{s}_i = g - A^*\hat{\tau}_i$ are the respective residual errors. Only the three inner products involved in $L(\hat{p}_n, \hat{q}_n, \hat{r}_{n-1}, \hat{s}_{n-1})$ are needed to calculate the improvement over the previous approximation $L(\hat{\sigma}_{n-1}, \hat{\tau}_{n-1}, f, g)$. As desired, we have a method that does not increase in cost as n increases.

This result is only of use when making calculations 'by hand'. In general, a machine will not recognise any difference in calculating an inner product involving the n^{th} re-iterated Galerkin approximation compared to an inner product involving the iterated Galerkin approximation. In fact, calculating L as described above increases the number of inner products that make up an approximation to the stationary value, and hence increases the chances of computational errors influencing results. For this reason, the computer program used in §7 simply calculates the functional L directly, *without* breaking up the re-iterated Galerkin approximations.

It would be reasonable to assume that as $n \to \infty$ our approximation $L(\hat{\sigma}_n, \hat{\tau}_n, f, g)$ will tend to the stationary value (ϕ, g). Can this be shown, and if so, can the improvements be described?

In §4.1, it was shown that

$$L(\phi + \delta\phi, \psi + \delta\psi, f, g) = (\phi, g) - (A\delta\phi, \delta\psi) \tag{4.9}$$

for any variations $\delta\phi, \delta\psi \in \mathscr{H}$. As, for any p and q in $\mathscr{H}$, there exist $\delta\phi$ and $\delta\psi$ in $\mathscr{H}$, such that

$$p = \phi + \delta\phi; \qquad q = \psi + \delta\psi,$$

(4.9) may be written

$$(\phi, g) - L(p, q, f, g) = (A(p - \phi), q - \psi) \qquad \forall p, q \in \mathscr{H}.$$

Substituting $\hat{\sigma}_n$ for p and $\hat{\tau}_n$ for q, and using the Schwartz inequality

$$|(\phi, g) - L(\hat{\sigma}_n, \hat{\tau}_n, f, g)| \leq \|A\| \|\phi - \hat{\sigma}_n\| \|\psi - \hat{\tau}_n\|,$$

using (4.6) and (4.7) gives

$$|(\phi, g) - L(\hat{\sigma}_n, \hat{\tau}_n, f, g)| \leq C\|S\|^{n+1}\|T\|^{n+1}$$

where $C = \|A\| \|p_0 - \phi\| \|q_0 - \phi\|$ is a constant (depending on the choice of subspaces). This shows that the re-iterated Galerkin method provides a means for calculating successive improvements to the staionary value (ϕ, g). The improvements are bounded above by $\|S\|\|T\| < \|S\| < 1$, although it must be noted that this tighter bound is obtained at the expense of a second system of equations to be solved, for the auxiliary, adjoint equation

$$A^*\psi = g.$$

Again, the discussion of §3.3.1 holds, and the improvement factor is expected to tend to $\rho_{\mathscr{H}'}(S)\rho_{\mathscr{H}''}(T)$, where $\mathscr{H}''$ is the equivalent of $\mathscr{H}'$ for the adjoint problem.

Chapter 5

The acoustics problem revisited

Having covered the theory of how to approximate the solution of certain operator equations in §3, we wish to use this theory to solve the problem set out in §1. In order to do this, it is necessary to classify the integral equation (1.15) in an appropriate setting. This process will lead neatly into making explicit how to use the theory of §4 to approximate the reflection and transmission amplitudes, R and T.

5.1 The Hilbert space $L_2(D')$

Recall that the integral equation in question is

$$\phi(x_0, y_0) = e^{i\beta_0 x_0}\sin(y_0) + \int\!\!\int_{D'} (k^2(x, y) - k_0^2)G(x, y \mid x_0, y_0)\phi(x, y)\mathrm{d}x\mathrm{d}y \qquad (5.1)$$

for $(x_0, y_0) \in D'$, where D' is a bounded subset of $D = \{x_0, y_0 : x_0 \in \mathbb{R}, 0 < y_0 < \pi\}$. Let us suppose that we wish to recast (5.1) in the Hilbert space of Lebesgue square integrable functions, over the domain D'. This space will be denoted $L_2(D')$. More generally, any space of two-dimensional Lebesgue integrable functions will be referred to as a $L_2 \oplus L_2$ space, i.e. the direct sum of one-dimensional Lebesgue spaces.

The space $L_2(D')$ is defined with the inner product

$$(u, v) = \int\!\!\int_{D'} u\bar{v}, \qquad \forall u, v \in L_2(D') \qquad (5.2)$$

and norm

$$\|u\| = (u, u)^{\frac{1}{2}}.$$

Note that, other Hilbert spaces could have been considered. However, the $L_2(D')$ space is the most natural choice in regards to the functions that are being dealt with. As we will soon see, the inner product (5.2) is particularly well suited to finding R and T by the method proposed in §4.

Let the $L_2(D')$ element ϕ represent the function $\phi(x_0, y_0)$ that satisfies the integral equation (5.1). Likewise, let the element f_0 represent the free term

$$e^{i\beta_0 x_0}\sin(y_0)$$

of (5.1). In addition, we define the multiplication operator M as

$$(M\varphi)(x,y) = \big(k^2(x,y) - k_0^2\big)\varphi(x,y)$$

and the integral operator $\mathcal{G}$ as

$$(\mathcal{G}\varphi)(x_0,y_0) = \int\!\!\int_{D'} G(x,y \mid x_0,y_0)\varphi(x,y)\mathrm{d}x\mathrm{d}y,$$

where $G(x,y \mid x_0,y_0)$ is the Green's function defined in (1.13). The integral equation (5.1) may now be solved by solving the equivalent operator equation

$$\phi = f_0 + \mathcal{G}M\phi \tag{5.3}$$

in the Hilbert space $L_2(D')$. This may be achieved via application of the re-iterated Galerkin method.

It is also required that the reflection and transmission amplitudes, R and T, be determined. In §1, it was shown that these unknowns may be written as

$$R = \frac{i}{\beta_0\pi}\int\!\!\int_{D'}(k^2(x,y) - k_0^2)e^{i\beta_0 x}\phi(x,y)\mathrm{d}x\mathrm{d}y$$

and

$$T = 1 + \frac{i}{\beta_0\pi}\int\!\!\int_{D'}(k^2(x,y) - k_0^2)e^{-i\beta_0 x}\phi(x,y)\mathrm{d}x\mathrm{d}y.$$

To access the theory of §4, these definitions must be expressed in terms of inner-products on $L_2(D')$. Our choice of function space makes this task particularly simple, with

$$R = \frac{i}{\beta_0\pi}(\phi, Mg_1) : g_1(x,y) = e^{-i\beta_0 x}\sin(y)$$

and

$$T = 1 + \frac{i}{\beta_0\pi}(\phi, Mg_2) : g_2(x,y) = e^{i\beta_0 x}\sin(y).$$

It is apparent that g_1 and g_2 are equal to the conjugate of f_0 and f_0 respectively. This point will be made use of later.

There is a purely theoretical issue that arises from the choice of $L_2(D')$ as our Hilbert space, that should be briefly addressed.

Elements of $L_2 \oplus L_2$ spaces define equivalence classes of two-dimensional square integrable functions, rather than individual functions. A set of suitable functions that differ only on a set of measure zero are all represented by the same $L_2 \oplus L_2$ element. It may be helpful to think of this as the correspondence between the set of $L_2 \oplus L_2$ functions and the set of $L_2 \oplus L_2$ elements being non-injective. It is, therefore, not unreasonable to foresee formal difficulties in re-mapping our Hilbert space approximation onto a function that approximates the exact solution of (5.1) in a pointwise sense. Recall though, from (1.15), that the full approximation, $\phi_{app}(x_0,y_0)$ on D, is defined as

$$\phi_{app}(x_0,y_0) = e^{i\beta_0 x_0}\sin(y_0) + \int\!\!\int_{D'}(k^2(x,y) - k_0^2)G(x,y \mid x_0,y_0)\hat{\sigma}(x,y)\mathrm{d}x\mathrm{d}y,$$

where $\hat{\sigma}(x,y)$ is any function that belongs to the equivalence class defined by the re-iterated Galerkin approximation $\hat{\sigma}$. By definition of the equivalence class, any $\hat{\sigma}(x,y)$ will provide the same approximation, ϕ_{app}. This may be described as one of the 'smoothing' properties of the integral.

In practice, when making a Hilbert space approximation, we must reintroduce the independent variables to define $L_2 \oplus L_2$ elements. In doing so, we will exclusively consider ourselves to be working with continuous, or, at worst, piecewise continuous functions. The question of re-mapping elements onto functions does not arise. For this reason, from now on we shall speak, in the main, of functions rather than elements and assume an implicit understanding of the analogous Hilbert space situation.

With the problem now fully defined, we may make some physical interpretations of previously abstract quantities. To begin with, let us consider the error.

5.1.1 Error

The re-iterated Galerkin method seeks to reduce the norm of the pointwise error

$$||e|| = ||\phi - \hat{\sigma}_n|| = \Big(\int\int_{D'} |\phi - \hat{\sigma}_n|^2 \mathrm{d}x\mathrm{d}y\Big)^{\frac{1}{2}}. \tag{5.4}$$

Again, the smoothing property of the integral removes any contention of definition of pointwise error.

The norm of the error is a measure of the error in the approximation, across the domain D'. A decrease in this value does not imply that the error is decreasing at all points in the domain, rather that the error is decreasing in an 'overall' sense. We can draw a link between the norm error and pointwise error as follows. It can be shown that

$$|(\mathcal{G}M\varphi)(x_0,y_0)| \leq C||\varphi|| \tag{5.5}$$

where $C^2 = \int\int_{D'} |(k^2(x,y) - k_0^2)G(x,y \mid x_0,y_0)|^2 \mathrm{d}x\mathrm{d}y$ is a finite constant. Now, (using the notation of §3)

$$KP_N(I - KP_N) = KP_N - KP_N(I - KP_N) = (I - KP_N)KP_N$$

$$\Rightarrow (I - KP_N)^{-1}KP_N = KP_N(I - KP_N)^{-1},$$

hence

$$\begin{aligned} K(I-P_N) + KP_NS &= K(I-P_N) + KP_N(I-KP_N)^{-1}K(I-P_N) \\ &= K(I-P_N) + (I-KP_N)^{-1}KP_NK(I-P_N) \\ &= (I-KP_N)^{-1}K(I-P_N) \\ &= S. \end{aligned}$$

This gives the equality

$$\hat{\sigma}_n - \phi = \big(K(I-P_N) + KP_NS\big)(\hat{\sigma}_{n-1} - \phi). \tag{5.6}$$

Combining (5.5) and (5.6) gives

$$\begin{aligned} |\hat{\sigma}_n(x_0,y_0) - \phi(x_0,y_0)| &\leq C||\big((I-P_N) + P_NS\big)(\hat{\sigma}_{n-1} - \phi)|| \\ &\leq C(1 + ||S||^2)^{\frac{1}{2}}||(\hat{\sigma}_{n-1} - \phi)||. \end{aligned}$$

Therefore, as the norm of the error tends to zero, so does the pointwise error at all points in the domain. If we had the time to estimate C, we could estimate the magnitude of the pointwise error, but this would inevitably be very crude.

It should also be noted that the value of (5.4) is relative to the size of the domain D'. This can be loosely thought of as, two approximations of equivalent average pointwise accuracy: the one defined on the larger domain will have a larger error norm.

5.2 Assumptions

Before attempting to use the re-iterated Galerkin method on the acoustics problem, which has been re-written as an operator equation, it is necessary to show that the assumptions outlined in §2 are satisfied by this equation. That is, it must be shown that equation (5.3):

- possesses an unique solution,

and the combined operator $\mathcal{G}M$:

- has infinite rank,
- is compact and bounded.

5.2.1 Existence and Uniqueness

To prove the existence and uniqueness of the solution of an integral equation or operator equation is a highly intricate matter, and worthy of consideration in its own right. A necessary condition for uniqueness of a solution is that the combined operator $\mathcal{G}M$ does not possess the unit eigenvalue. This property is sensitive to the many parameters left free in the model. An attempt to find conditions on the parameters, for which $\mathcal{G}M$ satisfies even this necessary condition, is deemed beyond the scope of this work. Instead, parameters will be varied on a trial and error basis. Any unexpected behaviour may be attributed to this eigenvalue issue.

The discovery of parameters that do not produce such a solution, is an indication of either some physical phenomenon or deficiency in the model. In other words, if we discover combinations of parameters for which we are confident that the re-iterated Galerkin method will not produce a solution, then there are two possibilities. Firstly, the mathematical model may be accurately representing the fact that there is no solution of the original acoustics problem. Secondly, in modelling the acoustics problem, discrepancies could have been introduced, causing the model to be unrepresentative for these combinations of parameters. Again, further investigation into this area is deemed far beyond the scope of our work, and will not be attempted.

5.2.2 Non-degeneracy

Extending what was said of degenerate kernels in §3.13 to two-dimensional integral equations, $\mathcal{G}M$ is an operator of infinite rank only if the kernel $(MG)(x, y \mid x_0, y_0) = \big(k^2(x,y) - k_0^2\big)G(x, y \mid x_0, y_0)$ cannot be written in the form

$$(MG)(x, y \mid x_0, y_0) = \sum_{i=1}^{N} b_i(x_0, y_0) k_i(x, y),$$

for some finite integer N. This property may be established by noting the non-separability of the Green's function G.

There is however another form of degeneracy that we must be aware of. This degeneracy would not effect the infinite rank of $\mathcal{G}M$, rather it would make the operator of finite rank in one of the independent variables. This would essentially mean that we would be solving a finite set of one-dimensional integral equations. To fully test the capabilities of the re-iterated Galerkin method on two-dimensional integral equations, this scenario should be avoided.

The non-separability of G spoken of is in the variable x_0, and we should therefore be suspicious of degeneracy only in y_0. As an example, consider the case where the multiplication operator M only introduces extra variation to the kernel in the x direction, and the domain D' is rectangular and covers the breadth of domain D. Assuming existence and uniqueness of a solution, the structure of the kernel implies a solution of the form

$$\phi(x_0, y_0) = \sum_{n=1}^{\infty} \phi_n(x_0) \sin(ny_0).$$

Substituting this solution into (5.3) and using orthogonality of the functions $\sin(ny)$ and $\sin(my)$ ($m \neq n$), gives the one dimension integral equations

$$\phi_1(x_0) = e^{i\beta_0 x_0} + \frac{i}{2\beta_0} \int_{l_1}^{l_2} M(e^{i\beta_0 |x-x_0|})(x)\phi_1(x)\mathrm{d}x \tag{5.7}$$

and

$$\phi_n(x_0) = \frac{1}{2\gamma_n} \int_{l_1}^{l_2} M(e^{-\gamma_n |x-x_0|})(x)\phi_n(x)\mathrm{d}x \quad (n \geq 2), \tag{5.8}$$

where l_1 and l_2 are the lower and upper limits on the length of the domain D'. Uniqueness implies that the solutions of the homogeneous integral equations (5.8) are the trivial solutions, hence

$$\phi(x_0, y_0) = \phi_1(x_0)\sin(y_0),$$

so that the y_0 dependence is fully defined and ϕ_1 is found by solving the one-dimensional integral equation (5.7).

5.2.3 Boundedness and compactness

The multi-dimensional extension of what was said of boundedness and compactness of L_2-kernels, in §2, holds . Hence, it need only be shown that the kernel $(MG)(x, y \mid x_0, y_0)$ is a $L_2 \oplus L_2$-kernel on D', i.e.

$$\int\int_{D'}\int\int_{D'} |(k^2(x,y) - k_0^2)G(x, y \mid x_0, y_0)|^2 \mathrm{d}x\mathrm{d}y\, \mathrm{d}x_0\mathrm{d}y_0 < \infty \tag{5.9}$$

and it may be deduced that $\mathcal{G}M$ is bounded and compact as an $L_2(D')$ operator.

It is reasonable to assume that the function $k^2(x,y) - k_0^2$ is itself a $L_2 \oplus L_2$-kernel, so the satisfaction of inequality (5.9) becomes a matter of showing that

$$\int\int_{D'}\int\int_{D'} |G(x, y \mid x_0, y_0)|^2 \mathrm{d}x\mathrm{d}y\, \mathrm{d}x_0\mathrm{d}y_0 < \infty, \tag{5.10}$$

where the Green's function, G, is fully defined as

$$G(x,y\mid x_0,y_0)=\frac{i}{\beta_0\pi}\sin(y_0)\sin(y)e^{i\beta_0|x-x_0|}+\sum_{n=2}^{\infty}\frac{1}{\gamma_n\pi}\sin(ny_0)\sin(ny)e^{-\gamma_n|x-x_0|}$$

and D' is a bounded sub-domain of D.

As in §1.4, let

$$D_X=\{x,y:|x|<X,0<y<\pi\},$$

but will now be considered only as a *finite* domain, and let $D'\subset D_X$. Obviously, if a function is a $L_2\oplus L_2$-kernel on D_X, then it is a $L_2\oplus L_2$-kernel on any subset of D_X. Thus, we may deduce (5.10) by showing that

$$\iint_{D_X}\iint_{D_X}|G(x,y\mid x_0,y_0)|^2\mathrm{d}x\mathrm{d}y\,\mathrm{d}x_0\mathrm{d}y_0<\infty. \tag{5.11}$$

Note that, the term

$$\frac{i}{\beta_0\pi}\sin(y_0)\sin(y)e^{i\beta_0|x-x_0|}$$

is bounded, whence, its contribution to the integral (5.11) is finite and it need not be considered for the current purpose. This leaves only the integral of the series

$$\iint_{D_X}\iint_{D_X}|\sum_{n=2}^{\infty}\frac{1}{\gamma_n\pi}\sin(ny_0)\sin(ny)e^{-\gamma_n|x-x_0|}|^2\mathrm{d}x\mathrm{d}y\,\mathrm{d}x_0\mathrm{d}y_0. \tag{5.12}$$

The orthogonality relation

$$\int_0^{\pi}\sin(my)\sin(ny)\mathrm{d}y=\begin{cases}0 & m\neq n\\ \frac{\pi}{2} & m=n\end{cases}$$

allows expression (5.12) to be simplified to

$$\int_{-X}^{X}\int_{-X}^{X}\sum_{n=2}^{\infty}\frac{\pi}{4\gamma_n^2}e^{-2\gamma_n|x-x_0|}\mathrm{d}x\mathrm{d}x_0. \tag{5.13}$$

Now, $e^{-2\gamma_n|x-x_0|}$ is bounded, and

$$\begin{aligned}\gamma_n^2 &= n^2-k_0^2\\ &\sim n^2\end{aligned}$$

for large n. So, for large N

$$\sum_{n=N}^{\infty}\frac{1}{\gamma_n^2}e^{-2\gamma_n|x-x_0|}\sim\sum_{N=2}^{\infty}\frac{1}{n^2}<\infty$$

at which point we have

$$\sum_{n=2}^{\infty}\frac{\pi}{4\gamma_n^2}e^{-2\gamma_n|x-x_0|}<\infty$$

and hence the integral (5.13) must also be finite.

This proves that the Green's function $G(x,y\mid x_0,y_0)$, and hence kernel $(MG)(x,y\mid x_0,y_0)$ are $L_2\oplus L_2$-kernels on D_X, and consequently $(MG)(x,y\mid x_0,y_0)$ induces the bounded and compact operator $\mathcal{G}M$ on D'.

Chapter 6

Numerical Considerations

This chapter deals with the factors that separate the task of actually carrying out the re-iterated Galerkin method from the theoretical issues of §3-4.

6.1 The discrete problem and numerical quadrature

In performing calculations on a computer, our continuous problem is necessarily reduced to a discrete problem. This means that instead of solving for elements that were earlier described as 'infinite vectors' that describe the shape of their associated function over a particular domain, we work with finite dimensional vectors that hold only certain point values of their associated functions.

This in itself does not effect our problem, as the continuous solution is, by definition, the solution of any discrete version of the equation. Also, as the solution is only being approximated, for a sufficiently refined discrete problem, some form of interpolation would be sufficient to provide a continuous solution if required.

On what basis should the discrete problem be set? This question is closely linked to the form of quadrature used in computations. In general, a computer program will employ a quadrature method to evaluate integrals. Quadrature methods are used to approximate definite integrals with a finite sum that takes the form

$$\int_{domain} f(\boldsymbol{t})\mathrm{d}\boldsymbol{t} \approx \sum_{i=1}^{N} w_i f(\boldsymbol{t}_i)$$

for some finite natural number N. Quadrature rules differ from one another in their abscissae (points $\boldsymbol{t}_i$) and weights w_i. In order to employ a quadrature rule, a program must necessarily carry around the values of functions at the appropriate abscissae points. It is therefore logical to base the discrete problem around the quadrature rule being used. As always, we seek to strike a balance between the extra accuracy that is gained by increasing the dimension of the quadrature rule and minimizing the computational expense.

In using a quadrature rule another form of error is introduced to our approximation, this will be discussed in §6.5.

By defining the wave profile as (1.2) complex variables were introduced into the problem. When making calculations, carrying around these complex variables can be cumbersome. The following section shows how this can be avoided, but first we make a simplifying assumption.

Let $k^2(x, y) - k_0^2$ be a real-valued function. This is physically valid as the function $k(x, y)$ and constant k_0 are produced by the ratio of the wave frequency and wave speed. Only in extremely rare circumstances would either of these measures involve an imaginary part.

6.2 A real valued kernel

In practice, it is of significant advantage to work with operators that map the real-valued elements of a space onto real-valued elements. Operators with this property allow the equation to be split into its real and imaginary parts, and solved independently. In terms of integral equations, such operators arise from real-valued kernels. Suppose

$$\phi = f + K\phi,$$

where $f = f_{re} + if_{im}$ and the kernel of K is a real-valued function, then

$$\phi = \phi_{re} + i\phi_{im},$$

such that $\phi_j \in \mathbb{R}$ and

$$\phi_j = f_j + K\phi_j, \qquad (j = re, im).$$

Thus, any program written to perform re-iterated Galerkin on a problem involving a real-valued kernel need only deal with real numbers.

Our kernel $(MG)(x, y \mid x_0, y_0) = (k^2(x, y) - k_0)G(x, y \mid x_0, y_0)$, unfortunately does not possess this property due to the appearance of the complex term $\frac{i}{\beta_0\pi}\sin(y_0)\sin(y)e^{i\beta_0|x-x_0|}$ in the Green's function G. However, this problem can be amended. Note that, by De-Moivre's theorem

$$\frac{i}{\beta_0\pi}\sin(y_0)\sin(y)e^{i\beta_0|x-x_0|} = \frac{i}{\beta_0\pi}\sin(y_0)\sin(y)\big[\cos\big(\beta_0|x - x_0|\big) + i\sin\big(\beta_0|x - x_0|\big)\big]. \qquad (6.1)$$

Using the even property of the cosine function (6.1) gives

$$\frac{i}{\beta_0\pi}\sin(y_0)\sin(y)\big[\cos\big(\beta_0(x - x_0)\big) + i\sin\big(\beta_0|x - x_0|\big)\big]$$

$$= \frac{1}{\beta_0\pi}\sin(y_0)\sin(y)\big[i\cos(\beta_0 x)\cos(\beta_0 x_0) + i\sin(\beta_0 x)\sin(\beta_0 x_0) - \sin\big(\beta_0|x - x_0|\big)\big].$$

We may then rewrite (5.3) as

$$\phi = f_0 + \frac{i}{\beta_0\pi}c_1 f_1 + \frac{i}{\beta_0\pi}c_2 f_2 + \tilde{\mathcal{G}}M\phi, \qquad (6.2)$$

where

$$f_1(x_0, y_0) = \cos(\beta_0 x_0)\sin(y_0) \in \mathbb{R},$$
$$f_2(x_0, y_0) = \sin(\beta_0 x_0)\sin(y_0) \in \mathbb{R},$$
$$c_1 = (\phi, Mf_1),$$
$$c_2 = (\phi, Mf_2),$$

and the integral operator

$$(\tilde{\mathcal{G}}\phi)(x_0, y_0) = \int\int_{D'} \tilde{G}(x, y \mid x_0, y_0)\phi(x, y)\mathrm{d}x\mathrm{d}y,$$

is such that

$$\tilde{G}(x, y \mid x_0, y_0) = \frac{1}{\beta_0\pi}\sin(y_0)\sin(y)\sin\big(\beta_0|x - x_0|\big) + \sum_{n=2}^{\infty}\frac{1}{\gamma_n\pi}\sin(ny_0)\sin(ny)e^{-\gamma_n|x-x_0|}. \quad (6.3)$$

It is easily seen from our assumptions on the function $k^2(x, y) - k_0^2$, that the operator $\tilde{\mathcal{G}}M$ is an integral operator arising from the real-valued kernel $(M\tilde{G})(x, y \mid x_0, y_0)$.

The arguments made in §5.2, which showed that the operator $\mathcal{G}M$ satisfies the relevant assumptions needed to apply the re-iterated Galerkin method, hold for the modified operator $\tilde{\mathcal{G}}M$.

The move made by redefining our problem as (6.2) is to remove the terms in the kernel that prevent it from being real-valued, and adding their contribution to the free term. In doing this, the original kernel, $(MG)(x, y \mid x_0, y_0)$, has been split into a degenerate kernel of rank 2

$$\frac{i}{\beta_0\pi}\big(k^2(x, y) - k_0^2\big)\sin(y_0)\sin(y)\cos(\beta_0|x - x_0|),$$

and a real kernel of infinite rank, $(M\tilde{G})(x, y \mid x_0, y_0)$. The gain in creating a operator with a real-valued kernel is made at the cost of a less straightforward sum of free terms, as the kernel of rank 2 has been moved into the free term. This however does not prove to be problematic.

First though, let

$$\phi = \phi_{re} + i\phi_{im},$$

and

$$\begin{aligned} c_j &= (\phi_{re}, Mf_j) + i(\phi_{im}, Mf_j) \\ &= c_{j_{re}} + ic_{j_{im}}, \qquad\qquad (j = 1, 2), \end{aligned}$$

then the equations

$$\left.\begin{aligned} \phi_{re} &= \Re\mathrm{e}(f_0) - \frac{1}{\beta_0\pi}c_{1_{im}}f_1 - \frac{1}{\beta_0\pi}c_{2_{im}}f_2 + \tilde{\mathcal{G}}M\phi_{re}, \\ \phi_{im} &= \Im\mathrm{m}(f_0) + \frac{1}{\beta_0\pi}c_{1_{re}}f_1 + \frac{1}{\beta_0\pi}c_{2_{re}}f_2 + \tilde{\mathcal{G}}M\phi_{im} \end{aligned}\right., \quad (6.4)$$

are both real-valued. Before proceeding, it will help to spot that

$$\Re\mathrm{e}(f_0) = \cos(\beta_0 x_0)\sin(y_0) = f_1; \qquad \Im\mathrm{m}(f_0) = \sin(\beta_0 x_0)\sin(y_0) = f_2.$$

Thus, equations (6.4) may be written

$$\phi_{re} = f_1 - \frac{1}{\beta_0\pi}c_{1_{im}}f_1 - \frac{1}{\beta_0\pi}c_{2_{im}}f_2 + \tilde{\mathcal{G}}M\phi_{re}$$

$$\phi_{im} = f_2 + \frac{1}{\beta_0\pi}c_{1_{re}}f_1 + \frac{1}{\beta_0\pi}c_{2_{re}}f_2 + \tilde{\mathcal{G}}M\phi_{im}.$$

To deal with the unknown constants that appear in these equations, define ϕ_i $(i = 1, 2)$ to be the solution of

$$\phi_i = f_i + \tilde{\mathcal{G}}M\phi_i,$$

so that

$$\phi_{re} = \big(1 - \frac{c_{1_{im}}}{\beta_0\pi}\big)\phi_1 - \frac{c_{2_{im}}}{\beta_0\pi}\phi_2$$

and

$$\phi_{im} = \frac{c_{1_{re}}}{\beta_0\pi}\phi_1 + \big(1 + \frac{c_{2_{re}}}{\beta_0\pi}\big)\phi_2.$$

Substituting these solutions into the definitions of the c_{i_j}

$$c_{1_{re}} = \big(1 - \frac{c_{1_{im}}}{\beta_0\pi}\big)(\phi_1, Mf_1) - \frac{c_{2_{im}}}{\beta_0\pi}(\phi_2, Mf_1),$$

$$c_{2_{re}} = \big(1 - \frac{c_{1_{im}}}{\beta_0\pi}\big)(\phi_1, Mf_2) - \frac{c_{2_{im}}}{\beta_0\pi}(\phi_2, Mf_2),$$

$$c_{1_{im}} = \frac{c_{1_{re}}}{\beta_0\pi}(\phi_1, Mf_1) + \big(1 + \frac{c_{2_{re}}}{\beta_0\pi}\big)(\phi_2, Mf_1),$$

$$c_{2_{im}} = \frac{c_{1_{re}}}{\beta_0\pi}(\phi_1, Mf_2) + \big(1 + \frac{c_{2_{re}}}{\beta_0\pi}\big)(\phi_2, Mf_2),$$

a 4 × 4 system to be solved for the unknown coefficients. It is unsurprising that, in separating the kernel into a degenerate part and a part of infinite rank, we are left to approximate the solutions involving the kernel of infinite rank, and then solve an equation of finite rank. The size of the system to be solved reflects that we are solving for the real and imaginary parts of a system arising from a kernel of rank 2. The inner products involved in this system may be approximated using the theory of §4.1.2. In turn, the solution of the system of equations, the coefficients, will be approximations.

Moreover, the structure of the operator $\tilde{\mathcal{G}}M$ allows these inner products to be approximated at a reduced cost.

6.2.1 The functional L revisited with the operator $\tilde{\mathcal{G}}M$

Both $\tilde{\mathcal{G}}$ and M are easily seen to be self-adjoint operators. However, only in rare circumstances will $\tilde{\mathcal{G}}M$ be self-adjoint.

Recall from §4 that approximations to an inner-product

$$(\phi, g),$$

where g is known and ϕ is unknown, can be made by approximating the stationary value of the functional L, such that

$$L(p, q, f, g) = (f, q) + (p, g) - (Ap, q).$$

This was achieved by approximating the solution of

$$A\phi = f$$

and the auxiliary equation

$$A^*\psi = g.$$

Our current goal is to approximate inner products of the form

$$(\phi_i, Mf_j),$$

where

$$\phi_i = f_i + \tilde{\mathcal{G}}M\phi_i.$$

With $A = I - \tilde{\mathcal{G}}M$, redefine the functional L as

$$L(p, q, f, g) = L(p, Mq, f, g)$$

and put $(f, g) = (f_i, g_j)$. As in §4, seek the stationary value $(p, q) = (\phi, \psi)$ of L:

$$L(\phi + \delta\phi, \psi + \delta\psi, f_i, Mf_j) = L(\phi + \delta\phi, M(\psi + \delta\psi), f_i, Mf_j)$$

$$= (f_i, M(\psi + \delta\psi)) + (\phi + \delta\phi, Mf_j) - ((I - \tilde{\mathcal{G}}M)(\phi + \delta\phi), M(\psi + \delta\psi)).$$

Isolating the first order terms, and using the self-adjointness of M

$$\begin{aligned} O(\delta\phi): &\qquad (\delta\phi, Mf_j - M\psi + M\tilde{\mathcal{G}}M\psi) = (M\delta\phi, f_j - \psi + \tilde{\mathcal{G}}M\psi), \\ O(\delta\psi): &\qquad (Mf_i - M\phi + M\tilde{\mathcal{G}}M\phi, \delta\psi) = (f_i - \phi + \tilde{\mathcal{G}}M\phi, M\delta\psi). \end{aligned}$$

Hence, the stationary value requires

$$\begin{aligned} \phi &= f_i + \tilde{\mathcal{G}}M\phi, \\ \psi &= f_j + \tilde{\mathcal{G}}M\psi, \end{aligned}$$

equations for which we have made approximations of the solutions. Thus, the approximation of the stationary value (ϕ_i, Mf_j) entails no auxiliary equation.

The solution ϕ may now be completely defined as the sum of solutions of assossiated real-valued operator equations.

6.2.2 The wave reflection and transmission amplitudes

Part of the problem set in §1 was to determine the wave reflection and transmission amplitudes R and T, such that

$$R = \frac{i}{\beta_0\pi}\int\int_{D'} \big(k^2(x,y) - k_0^2\big)e^{i\beta_0 x}\sin(y)\phi(x,y)\mathrm{d}x\mathrm{d}y$$

$$T = 1 + \frac{i}{\beta_0\pi}\int\int_{D'} \big(k^2(x,y) - k_0^2\big)e^{-i\beta_0 x}\sin(y)\phi(x,y)\mathrm{d}x\mathrm{d}y.$$

It has been shown that this may be achieved by interpreting these values as the innerproducts

$$R = \frac{i}{\beta_0\pi}(\phi, Mg_1): \qquad g_1(x,y) = e^{-i\beta_0 x}\sin(y)$$

$$T = 1 + \frac{i}{\beta_0\pi}(\phi, Mg_2): \quad g_2(x,y) = e^{i\beta_0 x}\sin(y)$$

Note that

$$\begin{aligned} g_1(x,y) &= \cos(\beta_0 x) - i\sin(\beta_0)\sin(y) \\ &= f_1(x,y) - if_2(x,y) \end{aligned}$$

and

$$\begin{aligned} g_2(x,y) &= \cos(\beta_0 x) + i\sin(\beta_0)\sin(y) \\ &= f_1(x,y) + if_2(x,y), \end{aligned}$$

so R and T may be written

$$\begin{aligned} R &= -\frac{1}{\beta_0\pi}(\phi, Mf_2) + \frac{i}{\beta_0\pi}(\phi, Mf_1) \\ &= -\frac{1}{\beta_0\pi}\big(c_2 - ic_1\big) \\ &= -\frac{1}{\beta_0\pi}\big((c_{1_{im}} + c_{2_{re}}) + i(c_{2_{im}} - c_{1_{re}})\big), \\ T &= 1 + \frac{1}{\beta_0\pi}(\phi, Mf_2) + \frac{i}{\beta_0\pi}(\phi, Mf_1) \\ &= 1 + \frac{1}{\beta_0\pi}\big(c_2 + ic_1\big) \\ &= 1 + \frac{1}{\beta_0\pi}\big((c_{2_{re}} - c_{1_{im}}) + i(c_{1_{re}} + c_{2_{im}})\big). \end{aligned}$$

Thus, we may deduce approximations to R and T by combining previous approximations.

We now return to the issue of the singularity in the Green's function, spoken of in §1.4. To deal with this problem, we must first lose some of the generality in our problem.

From now on, it is assumed that the domain D' is rectangular, such that

$$D' = \{x, y : a < x < b, c < y < d\}.$$

Before attempting to write a program some preliminary adjustments must be made to the integral operator $\tilde{\mathcal{G}}$ so that the singular kernel may be dealt with numerically.

The subsequent section owes much to Porter and Porter [5], and the work of Chamberlain [8].

6.3 Dealing with the singularity

Our operator is $\tilde{\mathcal{G}}M$ such that

$$(\tilde{\mathcal{G}}\varphi)(x_0, y_0) = \int\!\!\int_{D'} \tilde{G}(x, y \mid x_0, y_0)\varphi(x,y)\mathrm{d}x\mathrm{d}y, \tag{6.5}$$

where the kernel

$$\tilde{G}(x, y \mid x_0, y_0) = -a_0 + \sum_{2}^{\infty} s_n,$$

with

$$a_0 = a_0(x, y \mid x_0, y_0) = \frac{1}{\beta_0 \pi} \sin(y_0) \sin(y) \sin\big(\beta_0 |x - x_0|\big)$$

and

$$s_n = s_n(x, y \mid x_0, y_0) = \frac{1}{\gamma_n \pi} \sin(n y_0) \sin(n y) e^{-\gamma_n |x - x_0|}.$$

As already stated, a program will generally rely on a quadrature rule to approximate integrals. In what follows, it is assumed that the rectangular midpoint rule is used. To do this, let the required points be defined as

$$x_i = \frac{(2i-1)}{2X_p}(b-a) + a, \quad i = 1, \ldots, X_p; \qquad y_i = \frac{(2i-1)}{2Y_p}(d-c) + c, \quad i = 1, \ldots, Y_p$$

and for brevity

$$\boldsymbol{x}_{i+j(X_p-1)} = (x_i, y_j), \qquad i = 1, \ldots, X_p,\ j = 1, \ldots, Y_p.$$

From now on, this will be known as a $X_p \times Y_p$ *refinement.* The rectangular midpoint approximation of (6.5) is

$$(\tilde{\mathcal{G}} M \varphi)(x_0, y_0) \approx Q \sum_{i=1}^{X_p Y_p} \tilde{G}(\boldsymbol{x}_i \mid x_0, y_0)\big(k^2(\boldsymbol{x}_i) - k_0^2\big)\varphi(\boldsymbol{x}_i),$$

where $Q = (b-a)(d-c)/X_P Y_P$ is the area of one cell in the rectangular mesh defined by the points $\boldsymbol{x}_i$.

This simple expression is unfortunately invalid as a numerical device, due to the singularity in the kernel. More specifically (see §1.4), the singularity is logarithmic, occurring at the points $(x_0, y_0) = (x, y)$ in the function $\tilde{G}$. This means, an alternative expression for the above series at the points $(x_0, y_0) = \boldsymbol{x}_i$ must be found.

A second, related problem is in calculating approximations to the *infinite* series contained in $\tilde{G}$. When $(x_0, y_0) = (x, y)$ the terms

$$s_n \sim \frac{1}{n},$$

for large n. Hence, the logarithmic singularities is a result of a non-converging series at these points. Around the singularities the convergence of the series is extremely slow. This will obviously be a practical obstacle. Matters are much improved by writing

$$\sum_{n=2}^{\infty} s_n = \sum_{n=1}^{\infty} (s_n - t_n) + \sum_{n=1}^{\infty} t_n,$$

where

$$t_n = t_n(x, y \mid x_0, y_0) = \frac{1}{n\pi} \sin(n y_0) \sin(n y) e^{-n|x - x_0|}$$

and

$$s_1 = 0.$$

It is easily shown that

$$s_n - t_n \sim \frac{1}{n^3}$$

for large n. Hence, the series $\sum (s_n - t_n)$ contains no singularities (i.e. it will not diverge) and converges faster than $\sum s_n$ at all points in the domain.

The series $\sum t_n$ is summable, with

$$\sum_{n=1}^{\infty} t_n = \frac{1}{2\pi} \Re\mathrm{e}\big(\ln(1-\eta) - \ln(1-\xi)\big),$$

where

$$\eta = \eta(x, y \mid x_0, y_0) = e^{i(y+y_0)-|x-x_0|}, \quad \xi = \xi(x, y \mid x_0, y_0) = e^{i(y-y_0)-|x-x_0|}.$$

Using $\Re\mathrm{e}\big(\ln(z)\big) = \ln|z|$ gives

$$\sum_1^{\infty} t_n = \frac{1}{4\pi}\big(\ln(1-2e^{-|x-x_0|}\cos(y+y_0)+e^{-2|x-x_0|}) - \ln(1-2e^{-|x-x_0|}\cos(y-y_0)+e^{-2|x-x_0|})\big).$$

The logarithmic singularities are now clear. Note that, the former logarithm has singularities on the boundary

$$(x, y) = (x_0, y_0), \qquad y = 0, \pi,$$

the latter when

$$(x, y) = (x_0, y_0), \qquad \forall x, y.$$

As the points $\boldsymbol{x}_i$ never lie on the boundary of D', the integral involving $\ln|1-\eta|$ may be approximated using the rectangle midpoint rule. However, as $\ln|1-\xi|$ contains singularities within D', the integral arising from this function requires further attention before it may be approximated.

Before considering this problem, let us summarise.

The kernel of $\tilde{\mathcal{G}}$, $\tilde{G}$ may be written

$$\tilde{G}(x, y \mid x_0, y_0) = -a_0 + \sum_{n=1}^{\infty}(s_n - t_n) + \frac{1}{2\pi}\big(\ln|1-\eta| - \ln|1-\xi|\big).$$

So that, the operator $\tilde{\mathcal{G}}M$ may now be approximated as

$$\begin{aligned}(\tilde{\mathcal{G}}M\varphi)(x_0, y_0) \quad \approx \quad & Q\sum_{i=1}^{X_pY_p}\big(k^2(\boldsymbol{x}_i) - k_0^2\big)\big[-a_0(\boldsymbol{x}_i \mid x_0, y_0) + \mathcal{S}_N(\boldsymbol{x}_i \mid x_0, y_0)\dots \\ & \dots + \frac{1}{4\pi}\mathcal{L}_1(\boldsymbol{x}_i \mid x_0, y_0)\big]\varphi(\boldsymbol{x}_i) \\ & -\frac{1}{4\pi}\int\!\!\int_{D'}\mathcal{L}_2(x, y \mid x_0, y_0)(M\varphi)(x, y)\mathrm{d}x\mathrm{d}y \end{aligned} \tag{6.6}$$

where

$$\begin{aligned}\mathcal{L}_1(x, y \mid x_0, y_0) &= \ln(1-2e^{-|x-x_0|}\cos(y+y_0)+e^{-2|x-x_0|}), \\ \mathcal{L}_2(x, y \mid x_0, y_0) &= \ln(1-2e^{-|x-x_0|}\cos(y-y_0)+e^{-2|x-x_0|}),\end{aligned}$$

and $\mathcal{S}_N$ is some finite series that approximates $\sum_{n=1}^{\infty}(s_n - t_n)$.

We are still in need of an approximation to the remaining integral appearing in (6.6).

In preparation for what is to come let us write

$$\iint_{D'} \mathcal{L}_2(x,y \mid x_0,y_0)\varphi(x,y)\mathrm{d}x\mathrm{d}y$$

$$= \iint_{D'} \ln\Big(\frac{1-2e^{-|x-x_0|}\cos(y-y_0)+e^{-2|x-x_0|}}{(x-x_0)^2+(y-y_0)^2}\Big)\varphi(x,y)\mathrm{d}x\mathrm{d}y$$

$$+ \iint_{D'} \ln((x-x_0)^2+(y-y_0)^2)\big(\varphi(x,y)-\varphi(x_0,y_0)\big)\mathrm{d}x\mathrm{d}y$$

$$+\varphi(x_0,y_0)\iint_{D'} \ln\big((x-x_0)^2+(y-y_0)^2\big)\mathrm{d}x\mathrm{d}y.$$

By using the equality

$$\ln(1-2e^{-|x-x_0|}\cos(y-y_0)+e^{-2|x-x_0|})$$

$$= \ln\Big(\frac{1-2e^{-|x-x_0|}\cos(y-y_0)+e^{-2|x-x_0|}}{(x-x_0)^2+(y-y_0)^2}\Big) + \ln((x-x_0)^2+(y-y_0)^2)$$

we see that, from the limit

$$\lim_{\substack{x\to x_0\\ y\to y_0}} \ln\Big(\frac{1-2e^{-|x-x_0|}\cos(y-y_0)+e^{-2|x-x_0|}}{(x-x_0)^2+(y-y_0)^2}\Big) = 0, \tag{6.7}$$

the singularity has been transferred into the term

$$\ln\big((x-x_0)^2+(y-y_0)^2\big).$$

This is a less cluttered and hence more appealing term.

Matters have been further simplified by writing

$$\varphi(x,y) = \varphi(x,y)-\varphi(x_0,y_0)+\varphi(x_0,y_0)$$

and noting that for continuous $\varphi(x,y)$

$$\lim_{\substack{x\to x_0\\ y\to y_0}} \ln\big((x-x_0)^2+(y-y_0)^2\big)\big(\varphi(x,y)-\varphi(x_0,y_0)\big) = 0. \tag{6.8}$$

Thus, the unknown function $\varphi(x,y)$ has been removed from the integrand containing the singularities, this being

$$\iint_{D'} \mathcal{L}_3(x,y \mid x_0,y_0)\mathrm{d}x\mathrm{d}y. \tag{6.9}$$

where $\mathcal{L}_3(x,y \mid x_0,y_0) = \ln((x-x_0)^2+(y-y_0)^2)$.

Recall that the singularities occur at the points $(x_0,y_0)=(x,y)$, so that difficulties in using the rectangle midpoint rule occur when $(x_0,y_0)=\boldsymbol{x}_i$. At these points

$$\iint_{D'} \mathcal{L}_3(x,y \mid \boldsymbol{x}_j)\mathrm{d}x\mathrm{d}y \approx Q\sum_{\substack{i=1\\ i\neq j}}^{X_pY_p} \mathcal{L}_3(\boldsymbol{x}_i \mid \boldsymbol{x}_j) + \int_{y_{j_b}-Y}^{y_{j_b}+Y}\int_{x_{j_a}-X}^{x_{j_a}+X} \mathcal{L}_3(\boldsymbol{x}_j \mid \boldsymbol{x}_j)\mathrm{d}x\mathrm{d}y$$

where

$$j_a \equiv j \pmod{X_p}, \qquad j_b = \frac{j - j_a}{X_p}$$

and

$$X = \frac{b-a}{X_p}, \qquad Y = \frac{d-c}{Y_p}.$$

Now

$$\begin{aligned} A_{X,Y} &= \int_{y_{j_b}-Y}^{y_{j_b}+Y} \int_{x_{j_a}-X}^{x_{j_a}+X} \mathcal{L}_3(\boldsymbol{x}_j \mid \boldsymbol{x}_j) \mathrm{d}x\mathrm{d}y = \int_{-Y}^{Y} \int_{-X}^{X} \ln(u^2+v^2) \mathrm{d}u\mathrm{d}v \\ &= 4\big[XY\ln(X^2+Y^2) - 3XY + X^2\arctan(\frac{Y}{X}) + Y^2\arctan(\frac{X}{Y})\big] \end{aligned}$$

an analytic expression that holds for all $\boldsymbol{x}_j$.

Thus, the integral (6.9) may be approximated by the modified rectangle midpoint rule

$$\int\int_{D'} \mathcal{L}_3(x,y \mid x_0,y_0)\mathrm{d}x\mathrm{d}y \approx \left\{ \begin{array}{ll} \mathscr{L}(x_0,y_0) & (x_0,y_0) \neq \boldsymbol{x}_i, \\ \mathscr{L}_i(x_0,y_0) & (x_0,y_0) = \boldsymbol{x}_i, \end{array} \right.$$

where

$$\mathscr{L}(x_0,y_0) = Q \sum_{i=1}^{X_pY_p} \mathcal{L}_3(\boldsymbol{x}_i \mid x_0,y_0)$$

and

$$\mathscr{L}_j(x_0,y_0) = Q \sum_{\substack{i=1 \\ i\neq j}}^{X_pY_p} \mathcal{L}_3(\boldsymbol{x}_i \mid (x_0,y_0)) + A_{X,Y}.$$

The entire approximation is

$$\begin{aligned} (\tilde{\mathcal{G}}M\varphi)(x_0,y_0) \;\approx\; & Q\sum_{i=1}^{X_pY_p} \big(k^2(\boldsymbol{x}_i) - k_0^2\big)\Big[- a_0(\boldsymbol{x}_i \mid x_0,y_0) + \mathcal{S}_N(\boldsymbol{x}_i \mid x_0,y_0) + \ldots \\ & \ldots \frac{1}{4\pi}\mathcal{L}_1(\boldsymbol{x}_i \mid x_0,y_0) - \frac{1}{4\pi}\mathcal{L}_4(\boldsymbol{x}_i \mid x_0,y_0)\Big]\varphi(\boldsymbol{x}_i) \\ & -\frac{Q}{4\pi}\sum_{i=1}^{X_pY_p} \mathcal{L}_3(\boldsymbol{x}_i \mid (x_0,y_0)\big((M\varphi)(\boldsymbol{x}_i) - (M\varphi)(x_0,y_0)\big) \\ & -\frac{1}{4\pi}(M\varphi)(x_0,y_0) \left\{ \begin{array}{ll} \mathscr{L}(x_0,y_0) & (x_0,y_0) \neq \boldsymbol{x}_i \\ \mathscr{L}_i(x_0,y_0) & (x_0,y_0) = \boldsymbol{x}_i \end{array} \right. \end{aligned}$$

where

$$\mathcal{L}_4(x,y \mid x_0,y_0) = \ln\Big(\frac{1 - 2e^{-|x-x_0|}\cos(y-y_0) + e^{-2|x-x_0|}}{(x-x_0)^2 + (y-y_0)^2}\Big).$$

Bearing in mind what was said in §6.1, it is the approximation involving $\mathscr{L}_i$, rather than $\mathscr{L}$, that will be used.

6.3.1 The series approximation

For simplicity, let

$$\mathcal{S}_N = \sum_{n=1}^{N} \big(s_n - t_n\big) \approx \sum_{n=1}^{\infty} \big(s_n - t_n\big),$$

where N is some fixed positive integer. Increasing N improves the accuracy of $\mathcal{S}_N$, although it is expected that N should not have to be chosen particularly large, as the terms in the series are exponentially small accross the domain, except at the points $(x_0, y_0) = (x, y)$.

As an analytic solution is not at our disposal it is of importance to have as many ways of verifying the numerical solution as possible.

Check i. Recall that the functions that will be approximated by the re-iterated Galerkin method satisfy the real-valued operator equations

$$\phi_i = f_i + \tilde{\mathcal{G}} M \phi_i, \qquad (i = 1, 2),$$

which implies that

$$M f_i = (M - M\tilde{\mathcal{G}}M)\phi_i, \qquad (i = 1, 2). \tag{6.10}$$

From the known self-adjoint property of the operators $\tilde{\mathcal{G}}$ and M, it is easily deduced that the operator

$$M - M\tilde{\mathcal{G}}M$$

is also self-adjoint.

Now, consider the quantity

$$(\phi_i, M f_j), \qquad (i, j = 1, 2),$$

that must be approximated to find R and T. Using the above information

$$\begin{aligned} (\phi_i, M f_j) &= (\phi_i, (M - M\tilde{\mathcal{G}}M)\phi_j) \\ &= ((M - M\tilde{\mathcal{G}}M)\phi_i, \phi_j) \\ &= (M f_i, \phi_j) \\ &= (\phi_j, M f_i). \end{aligned}$$

This equality can be used to check the validity of the operator equations (6.10) used in computations.

There is a related result that can be used to reduce the amount of computation needed to approximate R and T. Our approximation of $(\phi_i, M f_j)$ is

$$L(\hat{\sigma}_{i,n}, M\hat{\sigma}_{j,n}, f_i, M f_j),$$

where $\hat{\sigma}_{i,n}$ is the n^{th} re-iterated Galerkin approximation to ϕ_i. Now

$$\begin{aligned} L(\hat{\sigma}_{i,n}, M\hat{\sigma}_{j,n}, f_i, M f_j) &= (f_i, M\hat{\sigma}_{j,n}) + (\hat{\sigma}_{i,n}, M f_j) \\ &\quad - (\hat{\sigma}_{i,n}, M\hat{\sigma}_{j,n}) + (\tilde{\mathcal{G}} M\hat{\sigma}_{i,n}, M\hat{\sigma}_{j,n}) \\ &= (\hat{\sigma}_{j,n}, M f_i) + (f_j, M\hat{\sigma}_{i,n}) \\ &\quad - (\hat{\sigma}_{j,n}, M\hat{\sigma}_{i,n}) + (\tilde{\mathcal{G}} M\hat{\sigma}_{j,n}, M\hat{\sigma}_{i,n}) \\ &= L(\hat{\sigma}_{j,n}, M\hat{\sigma}_{i,n}, f_j, M f_i), \end{aligned}$$

the approximation of $(\phi_j, M f_i)$. This shows that there is some symmetry in the 4 by 4 system described in §6.2, that may be taken advantage of.

Check ii Conservation of energy (see §1.3)

$$|R|^2 + |T|^2 = 1.$$

We shall come accross more checks as we progress.

In the preceeding theory, approximation methods were concerned with the pointwise error $e = \phi - \hat{\sigma}_n$. However, it is not possible to monitor what is happening to this quantity as it requires knowlege of the exact solution. Instead, we shall be content to look at the residual error $\hat{r} = f - \mathcal{G}M\hat{\sigma}$, a quantity that must be calculated when using the re-iterated Galerkin method.

6.4 The residual error

Let us consider some of the properties of the residual error, most of which are analogous to those already found for the pointwise error. To emphasise the generality of these results, the notation reverts to that of the earlier chapters.

It has already been shown that

$$Ae_n = \hat{r}_n,$$

from which we may deduce

$$\|\hat{r}_n\| \leq \|A\|\|e_n\|.$$

Thus, as $\|e_n\| \to 0$, $\|\hat{r}_n\| \to 0$.

Now, consider the effect of iteration on the residual error. Using the equality

$$\begin{aligned} \hat{p} &= f + Kp \\ &= r + p \end{aligned}$$

then

$$\begin{aligned} \hat{r} &= f - A\hat{p} \\ &= f - \hat{p} + K\hat{p} \\ &= f - (f + Kp) + K(r + p) \\ &= Kr. \end{aligned}$$

This says that the operator K maps the residual error in the Galerkin approximation onto the residual error in the iterate $\hat{p}$. Therefore, even if iteration improves the approximation with respect to the pointwise error, this does not necessarily entail an improvement in the residual error if $\|K\| > 1$.

Finally, what is the overall behaviour of the residual error? Using (3.19) and the similar equality

$$(I - KP_N)^{-1}K = K(I - P_NK)^{-1},$$

we may deduce that

$$K(I - P_N)(I - KP_N)^{-1} = (I - KP_N)^{-1}K(I - P_N).$$

Thus

$$\begin{aligned}\hat{r}_{n+1} &= f - A\hat{\sigma}_{n+1} \\ &= f - A\hat{\sigma}_n - A\hat{p}_{n+1} \\ &= \hat{r}_n - A(I-KP_N)^{-1}\hat{r}_n \\ &= K(I-P_N)(I-KP_N)^{-1}\hat{r}_n \\ &= (I-KP_N)^{-1}K(I-P_N)\hat{r}_n \\ &= S\hat{r}_n. \qquad (6.11)\end{aligned}$$

Therefore,

$$\begin{aligned}\hat{r}_{n+1} &= S\hat{r}_n \\ &= SAe_n\end{aligned}$$

and

$$\begin{aligned}\hat{r}_{n+1} &= Ae_{n+1} \\ &= ASe_n,\end{aligned}$$

which imply that

$$SAe_n = ASe_n.$$

As e_n is essentially arbitrary, this shows that the operators A and S commute. Equality (6.11) also produces the series expression

$$\begin{aligned}\sum_{i=0}^{n}\hat{r}_i &= \sum_{i=0}^{n} S^i\hat{r}_0 \\ &= \sum_{i=0}^{n} S^i A e_0 \\ &= A\sum_{i=0}^{n} S^i e_0.\end{aligned}$$

From §3.3.1, it is known that $\sum S^i e_0$ converges if $\rho_{\mathscr{H}'}(S) < 1$. So, this inequality must determine the convergence of $\sum \hat{r}_i = \sum S^i \hat{r}_0$. Mimicing the argument of §3.3.1, from which it was deduced that the ratio of the norms of sucessive point-wise errors tends to the spectral radius of the Sloan operator, this paragraph has shown that

$$\frac{||\hat{r}_{n+1}||}{||\hat{r}_n||} \to \rho_{\mathscr{H}'}(S) \qquad \text{as } n \to \infty. \qquad (6.12)$$

Appendices A-F, the MatLab code written to solve (5.3) using the re-iterated Galerkin method, are based on the work of §6.1-6.4. In particular, appendix D, the code that approximates the effect of the operator $\tilde{\mathcal{G}}M$, is based on §6.2-6.3.

To constuct bounds on the accuracy of any results would be a difficult and laborious task, and would inevitably prove to be crude and overly restrictive. Instead, we shall adopt a wholly pragmatic approach to the question of accuracy.

6.5 Accuracy

Let us summarise the errors that will effect our approximations:

- *The error inherant in an approximation.*
- *The rounding error in the computations.*
- *The error in the quadrature method.*
- *The error introduced by truncating the infinite series in the Green's function.*

Of these, we are able to control all but the rounding error.

In §3 it was shown that by suitable choice of basis, the re-iterated Galerkin method can be used to reduce the error in our approximation as required. However, in doing this, it was assumed that all calculations could be made exactly. We must now consider a more realistic position. Due to the series truncation error, the kernel with which we shall work is only an approximation of the true kernel. Thus, the operator used in computations is only an approximation to the true operator. Thus, we are no longer approximating the solution of the desired equation, rather the solution of a 'near-neighbour' equation. Furthermore, the quadrature and rounding errors will effect the accuracy of calculations made. In particular, the operator's influence on an approximation is manifest in a finite-dimensional matrix approximation to the operator. Computational errors in calculating the inner-products that make up this matrix cause innaccuracies in each entry. Therefore, there is a disparity between the operator and it's matrix approximation.

Recall the feature of the re-iterated Galerkin method, discussed briefly in §3.3, that it is sensitive to the error in the current approximation. The full importance of this property will now become clear. In the re-iterated Galerkin method, improvements are calculated using the residual error in the current approximation

$$\hat{r}_n = f - A\hat{\sigma}_n. \tag{6.13}$$

In using relation (6.13) to calculate the residual error, it will contain information about all errors in the current approximation $\hat{\sigma}_n$, including rounding and quadrature errors. Thus, the improvement calculated will, amongst other things, attempt to compensate for these errors. For this reason, the quadrature and rounding errors made at each stage of the re-iterated Galerkin method do not accumulate. That is, aside from the inaccuracies in the matrix approximation to the operator, the only computational errors that will effect our approximation, are those made at the last step.

If the errors under our control are reduced by performing more re-iterations, using more accurate quadrature methods, and taking more terms in the infinite series, then the problem on the computer becomes increasingly closer to the real problem, and in turn our approximation tends towards the desired solution. Now, suppose that a certain degree of accuracy is specified. If we make the error reductions spoken of, yet our approximation, to the specified degree of accuracy, does not change, then, to the specified degree of accuracy, the approximation is not sensitive to the errors in the approximation, and we have produced the required approximation. As the rounding error cannot be changed, it would not be possible to produce an approximation to a higher order of accuracy. In practice this would not prove to be an issue.

We shall suppose that we wish to obtain the quantities $|R|^2$ and $|T|^2$ accurate to three decimal places.

Chapter 7

Results

All calculations are made with MatLabv6, using the code given in appendices A-F.

It is now time to settle on a final problem, so that we may produce concrete results. This entails choosing the constant k_0, function $k(x,y)$ and domain D'. This choice is essentially our own, excepting the constraints previously outlined. These are

$$1 < k_0 < 2,$$

$$k(x,y) \to k_0 \qquad \text{as } (x,y) \to \delta D',$$

and the domain D' is rectangular. It should also be noted that, we should avoid letting $k = 0$ at any point, as this could only result from infinite wave speed, a physical impossibility.

So, let

$$k(x,y) = k_0 + \alpha \cos^2(\frac{mx}{2}) \sin^2(y)$$

for x, y in D', where

$$D' = \{x, y : |x| < \frac{\pi}{m}, 0 < y < \pi\}.$$

The constant parameters k_0, α and m have been left free, to allow comparison of results.

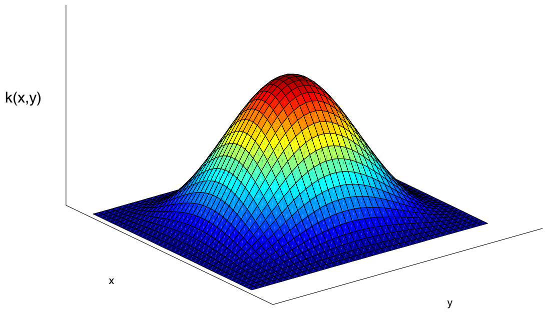

Figure 7.1: The function $k(x,y)$ over D'

It should be emphasised that the program written to perform the re-iterated Galerkin method on our acoustics problem, using the theory of the previous chapter, is adaptable to any choice that may have been made of the function $k(x, y)$ and rectangular domain D'.

The final influence that we must make over the problem is in constructing an appropriate subspace, in which the Galerkin approximation is made. The choice of this subspace will determine the ultimate success or failure of the application of the re-iterated Galerkin method.

7.1 Trial functions

The basis elements that form the subspace, are defined by their associated trial functions. As successive approximations will be built around these functions, it is logical that they share common traits with the exact solution. So, let us consider the solutions of

$$\phi_i(x_0, y_0) = F_i(x_0)\sin(y_0) + \int_0^\pi \int_{-\frac{\pi}{m}}^{\frac{\pi}{m}} \mathscr{M}(x, y)\tilde{G}(x, y \mid x_0, y_0)\phi_i(x, y)\mathrm{d}x\mathrm{d}y \qquad (i = 1, 2),$$

where

$$\begin{aligned} \mathscr{M}(x, y) &= \alpha k_0 \cos^2(\frac{mx}{2})\sin^2(y) + \alpha^2\cos^4(\frac{mx}{2})\sin^4(y), \\ F_1(x_0) = \cos(\beta_0 x_0); &\qquad F_2(x_0) = \sin(\beta_0 x_0), \end{aligned} \tag{7.1}$$

and $\tilde{G}(x, y \mid x_0, y_0)$ is as previously defined. Assuming that we are free to interchange the integrals and the infinite sum[1], then

$$\begin{aligned} \phi_i(x_0, y_0) &= F_i(x_0)\sin(y_0) + \sum_{n=1}^{\infty}\int_0^\pi \int_{-\frac{\pi}{m}}^{\frac{\pi}{m}} \mathscr{M}(x, y)\tilde{s}_n(x, y \mid x_0)\sin(ny_0)\phi_i(x, y)\mathrm{d}x\mathrm{d}y \\ &= \sum_{n=1}^{\infty} \tilde{\tilde{s}}_{i,n}(x_0)\sin(ny_0), \qquad (i = 1, 2), \end{aligned}$$

where

$$\begin{aligned} \tilde{s}_1(x, y \mid x_0) &= -\frac{1}{\beta_0\pi}\sin\left(\beta_0|x - x_0|\right)\sin(y) \\ \tilde{s}_n(x, y \mid x_0) &= \frac{1}{\gamma_n\pi}e^{-\gamma_n|x-x_0|}\sin(ny) \qquad (n = 2, 3, \dots) \end{aligned}$$

and

$$\begin{aligned} \tilde{\tilde{s}}_{i,1}(x_0) &= F_i(x_0) + \int_0^\pi \int_{-\frac{\pi}{m}}^{\frac{\pi}{m}} \mathscr{M}(x, y)\tilde{s}_1(x, y \mid x_0)\phi_i(x, y)\mathrm{d}x\mathrm{d}y \\ \tilde{\tilde{s}}_{i,n}(x_0) &= \int_0^\pi \int_{-\frac{\pi}{m}}^{\frac{\pi}{m}} \mathscr{M}(x, y)\tilde{s}_n(x, y \mid x_0)\phi_i(x, y)\mathrm{d}x\mathrm{d}y \qquad (n = 2, 3, \dots). \end{aligned}$$

Unlike the example of §5.2.1, no proof of degeneracy in this series expression is forthcoming. Therefore, both solutions $\phi_i(x_0, y_0)$ $(i = 1, 2)$ may be assumed to be expandable into infinite Fourier sine series in the variable y_0.

[1] This assumption is justified as the overall expression is known to be convergent.

Turning our attention to the x_0 variable, let our motivation come from noticing that the function F_1 is even (about the origin) and F_2 is odd (about the origin). Now, using the even property of $\mathscr{M}(x,y)$ in x, we find that

$$\begin{aligned}(\tilde{\mathcal{G}}M\varphi)(-x_0,y_0) &= \int_0^{\pi}\int_{-\frac{\pi}{m}}^{\frac{\pi}{m}} \mathscr{M}(x,y)\tilde{G}(x,y\mid -x_0,y_0)\varphi(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int_0^{\pi}\int_{-\frac{\pi}{m}}^{\frac{\pi}{m}} \mathscr{M}(x,y)\big[\big(-\frac{1}{\beta_0\pi}\sin\big(\beta_0|x+x_0|\big)\sin(y)\sin(y_0) \\ &\quad +\sum_{n=2}^{\infty}\frac{1}{\gamma_n\pi}e^{-\gamma_n|x+x_0|}\sin(ny)\sin(ny_0)\big)\big]\varphi(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int_0^{\pi}\int_{-\frac{\pi}{m}}^{\frac{\pi}{m}} \mathscr{M}(-\xi,y)\big[-\frac{1}{\beta_0\pi}\sin\big(\beta_0|\xi-x_0|\big)\sin(y)\sin(y_0) \\ &\quad +\sum_{n=2}^{\infty}\frac{1}{\gamma_n\pi}e^{-\gamma_n|\xi-x_0|}\sin(ny)\sin(ny_0)\big)\big]\varphi(-\xi,y)\mathrm{d}\xi\mathrm{d}y \\ &= \begin{cases}(\tilde{\mathcal{G}}M\varphi)(x_0,y_0) & \text{if } \varphi(x_0,y_0) \text{ is even in } x_0 \\ -(\tilde{\mathcal{G}}M\varphi)(x_0,y_0) & \text{if } \varphi(x_0,y_0) \text{ is odd in } x_0.\end{cases}\end{aligned}$$

Hence, the operator $\tilde{\mathcal{G}}M$ will preserve an odd or even property of a function in the x_0 variable.

Any function is the sum of its odd and even parts, so, if we let

$$\phi_i(x_0,y_0) = \phi_{i,ev_{x_0}}(x_0,y_0) + \phi_{i,odd_{x_0}}(x_0,y_0), \qquad (i=1,2)$$

where $\phi_{i,ev_{x_0}}$ and $\phi_{i,odd_{x_0}}$ are the even and odd parts of ϕ_i in the variable x_0, respectively. Therefore,

$$\begin{aligned}\phi_{i,ev_{x_0}} + \phi_{i,odd_{x_0}} &= f_i + \tilde{\mathcal{G}}M(\phi_{i,ev_{x_0}} + \phi_{i,odd_{x_0}}) \\ &= f_i + \tilde{\mathcal{G}}M\phi_{i,ev_{x_0}} + \tilde{\mathcal{G}}M\phi_{i,odd_{x_0}}. \qquad (7.2)\end{aligned}$$

As $\tilde{\mathcal{G}}M$ maps even functions in x_0 onto even functions in x_0, and odd functions in x_0 onto odd functions in x_0, we may split (7.2) into its odd and even components, giving

$$\begin{aligned}\phi_{i,ev_{x_0}} &= f_{i,ev_{x_0}} + \tilde{\mathcal{G}}M\phi_{i,ev_{x_0}}, \qquad (i=1,2), \\ \phi_{i,odd_{x_0}} &= f_{i,odd_{x_0}} + \tilde{\mathcal{G}}M\phi_{i,odd_{x_0}}, \qquad (i=1,2),\end{aligned}$$

where, as expected, $f_{i,ev_{x_0}}$ and $f_{i,odd_{x_0}}$ are the even and odd parts of f_i in the variable x_0, respectively. Formalising what has been said of the odd and even properties of the free terms, gives

$$f_{1,odd_{x_0}} = 0\,; \quad f_{2,ev_{x_0}} = 0,$$

in which case $\phi_{1,odd_{x_0}}$ and $\phi_{2,ev_{x_0}}$ must both satisfy the homogeneous equation

$$\varphi = \tilde{\mathcal{G}}M\varphi. \qquad (7.3)$$

However, uniqueness of solution, which is assumed, prevents any non-trivial solutions of (7.3), i.e.

$$\phi_{1,odd_{x_0}} = 0\,; \quad \phi_{2,ev_{x_0}} = 0.$$

This shows that $\phi_1 = \phi_{1,ev_{x_0}}$ is even in x_0, and $\phi_2 = \phi_{2,odd_{x_0}}$ is odd in x_0. This can and will be reflected in our choice of trial space. That is, the trial space corresponding to the solution ϕ_1 shall be even in x_0, and the trial space corresponding to ϕ_2 odd in x_0.

Recall that the Galerkin approximation is defined as

$$p = (I - P_N K)^{-1} P_N f.$$

Expanding the operator $(I - P_N K)^{-1}$ into series form gives

$$p = \sum_{j=0}^{\infty} (P_N K)^j P_N f.$$

With $K = \tilde{\mathscr{G}} M$ and $f = f_i$ $(i = 1, 2)$, the Galerkin approximation becomes

$$p_{i,0} = \sum_{j=0}^{\infty} (P_N \tilde{\mathscr{G}} M)^j P_N f_i.$$

As any orthogonal projection, onto a purely odd or even trial space, preserves the odd or even property of a function, and from what has been noted of the odd and even properties of the operator and free terms, it is clear that the Galerkin approximations to $\phi_1(x_0, y_0)$ and $\phi_2(x_0, y_0)$ will be even and odd in x_0, respectively.

In turn, the iterate

$$\hat{p}_{i,0} = f_i + \tilde{\mathscr{G}} M p_{i,0}, \qquad (i = 1, 2),$$

and residual

$$\hat{r}_{i,0} = f_i - \hat{p}_{i,0} + \tilde{\mathscr{G}} M \hat{p}_{i,0}, \qquad (i = 1, 2),$$

will preserve the even or odd property. An inductive argument can then be used to show that all approximations that are used to create the re-iterated Galerkin approximation will preserve the even or odd property. Thus, at every stage, the re-iterated Galerkin approximations to $\phi_1(x_0, y_0)$ and $\phi_2(x_0, y_0)$ will be even and odd in x_0, respectively.

Can anything similar be said of the variation of the approximations in y_0?

The function

$$\sin(ny) \qquad (n \in \mathbb{N})$$

is even about the point $\frac{\pi}{2}$, i.e.

$$\sin\big(n(\frac{\pi}{2} - y)\big) = \sin\big(n(\frac{\pi}{2} + y)\big),$$

if n is odd, and odd about $\frac{\pi}{2}$, i.e.

$$\sin\big(n(\frac{\pi}{2} - y)\big) = -\sin\big(n(\frac{\pi}{2} + y)\big),$$

if n is even. A simple consequence of this is that

$$(M\varphi)(x, y) = \mathscr{M}(x, y)\varphi(x, y)$$

is even or odd about the point $\frac{\pi}{2}$ in y, if φ has also has that property. Now, using the orthogonality of functions even about $\frac{\pi}{2}$ to functions odd about $\frac{\pi}{2}$, we have that if $\varphi(x_0, y_0)$ is even about $\frac{\pi}{2}$ in y_0 then

$$\begin{aligned}(\tilde{\mathcal{G}}M\varphi)(x_0,y_0) &= \int_0^{\pi}\int_{-\frac{\pi}{m}}^{\frac{\pi}{m}} \mathscr{M}(x,y)\big[\big(-\frac{1}{\beta_0\pi}\sin\big(\beta_0|x-x_0|\big)\sin(y)\sin(y_0)\\ &+ \sum_{n=1}^{\infty}\frac{1}{\gamma_{2n+1}\pi}e^{-\gamma_{2n+1}|x-x_0|}\sin\big((2n+1)y\big)\sin\big((2n+1)y_0\big)\big)\big]\varphi(x,y)\mathrm{d}x\mathrm{d}y,\end{aligned}$$

a function even in y_0, about the point $\frac{\pi}{2}$, and if $\varphi(x_0, y_0)$ is odd about $\frac{\pi}{2}$ in y_0

$$(\tilde{\mathcal{G}}M\varphi)(x_0,y_0) = \int_0^{\pi}\int_{-\frac{\pi}{m}}^{\frac{\pi}{m}} \mathscr{M}(x,y)\sum_{n=1}^{\infty}\frac{1}{\gamma_{2n}\pi}e^{-\gamma_{2n}|x-x_0|}\sin\big((2n+1)y\big)\sin\big((2n)y_0\big)\varphi(x,y)\mathrm{d}x\mathrm{d}y,$$

a function odd in y_0, about the point $\frac{\pi}{2}$. The former result is of importance to us, as both $f_i(x_0, y_0)$ $(i = 1, 2)$ are even in y_0, about $\frac{\pi}{2}$. Therefore, splitting ϕ_i $(i = 1, 2)$ into their even and odd parts about $\frac{\pi}{2}$ in y_0, we may deduce that the odd parts satisfy the homogeneous equation (7.3), and hence both solutions are even about $\frac{\pi}{2}$ in y_0, and the trial functions will be chosen to reflect this. This being the case, by exactly the same arguments used in the previous paragraph, the Galerkin approximation, iterate, residual, and hence all approximations must be even in y_0 about the point $\frac{\pi}{2}$.

From this discussion, we may deduce the following representations of the solutions

$$\phi_1(x_0,y_0) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} a_{i,j}\cos(\frac{j\pi x_0}{m})\sin\big((2i+1)y_0\big) \tag{7.4}$$

and

$$\phi_2(x_0,y_0) = \sum_{i=0}^{\infty}\sum_{j=1}^{\infty} b_{i,j}\sin(\frac{j\pi x_0}{m})\sin\big((2i+1)y_0\big), \tag{7.5}$$

where $a_{i,j}$ and $b_{i,j}$ are constants. It is, therefore, not unreasonable to create trial spaces from combinations of functions of the form

$$\cos(\frac{j\pi x_0}{m})\sin\big((2i+1)y_0\big), \quad (i,j = 0,1,2,\dots),$$

for the first problem, and

$$\sin(\frac{j\pi x_0}{m})\sin\big((2i+1)y_0\big), \quad (i = 0,1\dots;\ j = 1,\dots),$$

for the second problem. We shall proceed on this basis, increasing the size of the trial space in some logical fashion.

Both (7.4) and (7.5) are truncated versions of the double Fourier series. We have implicitly used the fact that the double Fourier series span $L_2 \oplus L_2$ spaces.

There is undoubtedly an optimal approach to increasing the size of our trial space. Moreover, there may well be a better set of trial functions to use. Analysis into this optimality could be expected to prove difficult, and may have to be approached as a practical trial and error investigation. This point is mentioned as possible extension to our work, and will not be pursued.

Check iii. The re-iterated Galerkin approximations must satisfy the even and odd properties spoken of.

This property could be used to quarter the computational domain. However, it is more useful to us as a device to help verify results.

Check iv. The inner-products that make up the matrix used in the re-iterated Galerkin method, are of the form

$$(A\chi_i, \chi_j),$$

where χ_i is a trial function. With $A = I - \tilde{\mathcal{G}}M$, and the specified trial functions, i.e.

$$\chi_i(x_0, y_0) = T_{i_p}(x_0)\sin(i_p y_0),$$

where $T_{i_p}(x_0)$ is equal to either $\cos(\frac{i_q \pi x_0}{m})$ or $\sin(\frac{i_q \pi x_0}{m})$, and $i_p, i_q \in \mathbb{N}$, this inner product becomes

$$(\chi_i, \chi_j) - (\tilde{\mathcal{G}}M\chi_i, \chi_j).$$

Concentrating our attentions to the term $(\tilde{\mathcal{G}}M\chi_i, \chi_j)$, we have

$$\begin{aligned}(\tilde{\mathcal{G}}M\chi_i, \chi_j) &= \int_0^{\pi}\int_{-\frac{\pi}{m}}^{\frac{\pi}{m}}\int_0^{\pi}\int_{-\frac{\pi}{m}}^{\frac{\pi}{m}} \mathscr{M}(x,y)\big[-\frac{1}{\beta_0\pi}\sin\big(\beta_0|x-x_0|\big)\sin(y)\sin(y_0) \\ &+ \sum_{n=2}^{\infty}\frac{1}{\gamma_n\pi}e^{-\gamma_n|x-x_0|}\sin(ny)\sin(ny_0)\big]T_{i_p}(x)\sin(i_p y)\mathrm{d}x\mathrm{d}y\, T_{i_p}(x_0)\sin(i_p y_0)\mathrm{d}x_0\mathrm{d}y_0 \\ &= \int_{-\frac{\pi}{m}}^{\frac{\pi}{m}}\int_0^{\pi}\int_{-\frac{\pi}{m}}^{\frac{\pi}{m}} \mathscr{M}(x,y)s_{j_p}(x,y \mid x_0)T_{i_p}(x)T_{j_p}(x_0)\sin(i_p y)\mathrm{d}x\mathrm{d}y\,\mathrm{d}x_0,\end{aligned}$$

thus removing the infinite series. This allows us to check the inner-products calculated with an approximation to the infinite series, with an integral in which we may use the exact integrand.

Check v. The approximations to ϕ_1 and ϕ_2 will be used in the functional L (see §6.2). In particular, we must calculate

$$L(\hat{\sigma}_{i,n}, M\hat{\sigma}_{j,n}, f_i, Mf_j) = (f_i, M\hat{\sigma}_{j,n}) + (\hat{\sigma}_{i,n}, Mf_j) - (\hat{\sigma}_{i,n}, M\hat{\sigma}_{j,n}) + (\tilde{\mathcal{G}}M\hat{\sigma}_{i,n}, M\hat{\sigma}_{j,n}),$$

where $\hat{\sigma}_{i,n}$ is the n^{th} re-iterated Galerkin approximation to ϕ_i. The odd and even properties of these approximations and the free terms, imply that terms of subscript i are orthogonal to terms of subscript j, for $i \neq j$. Therefore

$$L(\hat{\sigma}_{i,n}, M\hat{\sigma}_{j,n}, f_i, Mf_j) = 0, \qquad (i \neq j).$$

This check is of the functional, and should follow directly if *check iii* is satisfied.

There is another consequence of these odd and even properties.

7.2 The subspaces $E(D')$ and $O(D')$

It has just been shown that the operator $\tilde{\mathcal{G}}M$ maps the elements of $L_2(D')$ that correspond to functions even or odd in x_0 onto similar elements, and elements corresponding to functions even or odd about $\frac{\pi}{2}$ in y_0 onto similar elements. This implies that $\tilde{\mathcal{G}}M$ defines an integral operator on any of the related $L_2(D')$-subspaces, or combinations of such subspaces. In particular, the equation

$$\phi_1 = f_1 + \tilde{\mathcal{G}}M\phi_1,$$

may be considered in the $L_2 \oplus L_2$ space of elements corresponding to functions even in x_0 and even about $\frac{\pi}{2}$ in y_0, over the domain D'. Similarly, the equation

$$\phi_2 = f_2 + \tilde{\mathcal{G}}M\phi_2,$$

may be considered in the $L_2 \oplus L_2$ space of elements corresponding to functions odd in x_0 and even about $\frac{\pi}{2}$ in y_0, over the domain D'. These subspaces we denote by $E(D')$ and $O(D')$, respectively.

Therefore, we are essentially solving two equations in two separate Hilbert spaces. This links to the discussion of improvement factors of §3.3.1. A proof that $E(D')$ and $O(D')$ are the smallest Hilbert subspaces of $L_2(D')$, containing their respective equations, will not be attempted. There is, however, no reason to suspect otherwise, and for this reason we shall assume this property. Moreover, for any choice of the trial spaces defined, the Sloan operators are operators over $E(D')$ and $O(D')$. We shall also refer to $E(D')$ and $O(D')$ as the smallest subspaces containing all of the individual terms involved in our approximations. In any case, we have managed to refine the problem, and can deduce that the convergence of an application of the re-iterated Galerkin method depends, more so, on the spectral radius of the Sloan operator over the relevant $L_2(D')$-subspace, rather than $L_2(D')$ itself.

This is a good point at which to summarise the salient features of the problem to be solved. We will be approximating the solutions, ϕ_i $(i = 1, 2)$, of the two operator equations

$$\phi_i = f_i + \tilde{\mathcal{G}}M\phi_i, \qquad (i = 1, 2), \tag{7.6}$$

where

$$f_1(x_0, y_0) = \cos(\beta_0 x_0)\sin(y_0), \quad f_2(x_0, y_0) = \sin(\beta_0 x_0)\sin(y_0),$$

$$(\tilde{\mathcal{G}}\varphi)(x_0, y_0) = \int\int_{D'} \tilde{G}(x, y \mid x_0, y_0)\varphi(x, y)\mathrm{d}x\mathrm{d}y,$$

where $\tilde{G}$ is the modified Green's function (6.3), and

$$(M\varphi)(x, y) = \mathscr{M}(x, y)\varphi(x, y),$$

where $\mathscr{M}$ is defined by (7.1), over the space $L_2(D')$, where

$$D' = \{x, y : |x| < \frac{\pi}{m}, 0 < y < \pi\}.$$

These approximations are used to approximate the quantities

$$(\phi_i, Mf_j). \quad (i, j = 1, 2),$$

using the functional L (see §4). In turn, we use these approximations to approximate $|R|^2$ and $|T|^2$, quantities proportional to the reflected and transmitted energies, respectively. These values are required to three decimal places.

From now on, equation (7.6) will be known as problem 1 for $i = 1$ and problem 2 for $i = 2$. The respective trial spaces will be known as trial space 1 and trial space 2.

Let us clarify what we wish to investigate within the set problem:

1. Test the re-iterated Galerkin method.
2. Assess the controllable errors.
3. Solve the original problem.

As it stands, the problem contains more than enough parameters to allow us to deal effectively with these issues. A realistic position to put ourselves in, would be to allow some variation in the size (of the norm) of the operator $\tilde{\mathcal{G}}M$ for 1, and leave one parameter with which to compare the results in 3. Intuitively, the size of the operator will increase as the size of the domain increases, or the size of the kernel increases. These factors may be controlled by the parameters m, and k_0 and α, respectively. Let us choose to fix

$$m = 2; \quad k_0 = \sqrt{2},$$

and vary α.

Taking matters in the order presented, we begin by studying the application of the re-iterated Galerkin method.

7.3 Application of the re-iterated Galerkin method

The current section is exclusively a study of the re-iterated Galerkin method, rather than of any physical interpretation of results, or of the computational error. As such, throughout §7.3 a consistent refinement of 40 × 40, and series approximation $\mathcal{S}_3$, is used.

To give a full test of the re-iterated Galerkin method, it should be used on a problem in which the integral operator takes a norm value greater than 1. If this were not the case, the re-iterated Galerkin method would not be required, as repeated regular Sloan iteration (3.39) would converge.

7.3.1 Sloan iteration

A straightforward and practical test of whether the re-iterated Galerkin method is required, is to apply Sloan iteration to our problem, to see if it will converge.

Let us introduce the notation

$$\hat{r}_{i,n}$$

to be the residual error in the n^{th} re-iterated Galerkin approximation to ϕ_i $(i = 1, 2)$, and

$$\check{r}_{i,n},$$

to be the residual error in the n^{th} Sloan iterate to ϕ_i $(i = 1, 2)$.

The Sloan iteration should converge or diverge independently of the trial space (see §3.2.3). So, so let

$$\text{Trial Space 1: } 1 \in E(D') \tag{7.7}$$

and

$$\text{Trial Space 2: } x_0 \in O(D'). \tag{7.8}$$

Case $\alpha = 1$: Results of the application of Sloan iteration are recorded in table 7.1. Problem 1 displays evident divergence, whereas problem 2 is converging. In both series, the ratio of normed residuals settles within 10 iterations. Although the divergence in problem 1 is very slow, settling to 1.064, it is nevertheless divergence, and requires the re-iterated Galerkin method. As problem 2 does not diverge, and does in fact converge reasonably fast, it does not require the re-iterated Galerkin method. This does, however, provide us with another opportunity: to observe how the re-iterated Galerkin method affects the speed of convergence of an already convergent series.

Table 7.1: Sloan iteration

$\alpha = 1$; $\mathcal{S}_N = \mathcal{S}_3$; refinement: 40 × 40
Trial space 1 : (7.7); Trial space 2: (7.8)

Iterate	$\|\|\check{r}_{1,n}\|\|$	$\|\|\check{r}_{2,n}\|\|$	$\frac{\|\|\check{r}_{1,n}\|\|}{\|\|\check{r}_{1,n-1}\|\|}$	$\frac{\|\|\check{r}_{2,n}\|\|}{\|\|\check{r}_{2,n-1}\|\|}$
0	0.9119	0.8802		
1	1.177	0.2728	1.291	0.3099
2	0.9753	0.1274	0.828	0.4670
3	1.116	5.970×10^{-2}	1.144	0.4686
4	1.162	2.800×10^{-2}	1.041	0.4689
5	1.245	1.310×10^{-2}	1.079	0.4690
6	1.322	6.200×10^{-3}	1.062	0.4690
7	1.407	2.900×10^{-3}	1.065	0.4690
8	1.497	6.353×10^{-4}	1.064	0.4690
9	1.593	2.979×10^{-4}	1.064	0.4690
10	1.695	1.397×10^{-3}	1.064	0.4690

By making explicit a result that has been implicit thus far, we can extract further information from table 7.1. In §3.3.2, we saw that re-iterated Galerkin is a combination of Sloan iteration, with pre-conditioning replacing the operator K with the Sloan operator S. Therefore, all results on S in re-iterated Galerkin, are applicable to K in Sloan iteration. In particular, from (6.12), in repeated Sloan iteration

$$\frac{||\check{r}_{i,n}||}{||\check{r}_{i,n-1}||} \to \rho_{\mathscr{H}'}(K), \qquad \text{as } n \to \infty.$$

Therefore, table 7.1 also gives the approximations

$$\rho_{E(D')}(\tilde{\mathcal{G}}M) \approx 1.064,$$

and

$$\rho_{O(D')}(\tilde{\mathcal{G}}M) \approx 0.4690,$$

for $\alpha = 1$.

Although, there is no guarantee that, if the operator $\tilde{\mathcal{G}}M$ has small eigenvalues, a relatively small trial space will ensure S also has small eigenvalues, it is a good indication of this property. To see this, consider the two factors of

$$S = (I - \tilde{\mathcal{G}}MP_N)^{-1}(\tilde{\mathcal{G}}M - \tilde{\mathcal{G}}MP_N)$$

separately. The operator $(I - \tilde{\mathcal{G}}MP_N)$ has the eigenvalue 1 for all functions strictly outside the trial space, and $1 - \mu_n$, for all the eigenvalues μ_n of the operator $\tilde{\mathcal{G}}M$, where the corresponding eigenfunctions lie in the trial space. Hence, the operator $(I - \tilde{\mathcal{G}}MP_N)^{-1}$ has the eigenvalues 1 for all functions strictly outside the trial space, and $(1 - \mu_n)^{-1}$ for all eigenfunctions of $\tilde{\mathcal{G}}M$ within the trial space. Whereas, the operator $\tilde{\mathcal{G}}M - \tilde{\mathcal{G}}MP_N$ has the eigenvalues μ_n, for the eigenfunctions of $\tilde{\mathcal{G}}M$ orthogonal to the trial space. Therefore, the Sloan operator S has eigenvalues μ_n, the eigenvalues of $\tilde{\mathcal{G}}M$ where the corresponding eigenfunctions are orthogonal to the trial space. However, this is not necessarily the complete set of eigenvalues and eigenfunctions. It is now clear that, by picking a trial space containing parts of the eigenfunctions of $\tilde{\mathcal{G}}M$ corrosponding to the larger eigenvalues, minimises the spectral radius of S.

Table 7.2 shows the results[2] of the application of re-iterated Galerkin, with

$$\text{Trial Space 1: } \sin(y_0) \tag{7.9}$$

and

$$\text{Trial Space 2: } \sin(2x_0)\sin(y_0). \tag{7.10}$$

From the discussion of the previous paragraph, it is not surprising to find that these choices of trial spaces produce fast convergence. Again, the ratio of norms of successive residuals settle within 10 re-iterations, in both problems. Note that, we have improved the convergence rate in the second problem. Furthermore, problem 2 maintains a superior convergence rate. The approximations to the $|R|^2$ and $|T|^2$ also converge quickly. They are accurate to three decimal places (as required) by the second re-iterate, with the residual errors of orders 10^{-1} and 10^{-3} respectively. The program does, however, run into problems at later re-iterates. At the seventeenth re-iterate, the convergence rate of the second problem begins to deteriorate, and by the twenty-fifth re-iterate has begun to diverge at an near settled rate. As we shall see, this feature occurs consistently, in both problems. An explanation of this phenomena will be discussed in due course. For the present, it will not disturb our ability to observe the re-iterated Galerkin method working, and produce approximations to $|R|^2$ and $|T|^2$.

The effects of expanding the trial spaces to

$$\text{Trial Space 1: } \begin{cases} \sin(y_0), \\ \cos(2x_0)\sin(y_0), \end{cases} \tag{7.11}$$

and

$$\text{Trial Space 2: } \begin{cases} \sin(2x_0)\sin(y_0), \\ \sin(4x_0)\sin(y_0), \end{cases} \tag{7.12}$$

are recorded in table 7.3. Note that, in both problems, a larger subspace has produced faster convergence, and the values of $|R|^2$ and $|T|^2$ have converged by the first iterate (re-iterate 0).

[2] All tables are given to 4 significant figure accuracy.

Table 7.2: Re-iterated Galerkin

$\alpha = 1$; $\mathcal{S}_N = \mathcal{S}_3$; refinement: 40×40
Trial space 1 : (7.9); Trial space 2: (7.10)

Re-iterate	$\|\hat{r}_{1,n}\|$	$\|\hat{r}_{2,n}\|$	$\frac{\|\hat{r}_{1,n}\|}{\|\hat{r}_{1,n-1}\|}$	$\frac{\|\hat{r}_{2,n}\|}{\|\hat{r}_{2,n-1}\|}$	$\|R\|^2$	$\|T\|^2$
Galerkin	0.8793	0.7731				
0	0.9364	7.990×10^{-2}			4.202×10^{-5}	1.000
1	0.2885	7.700×10^{-3}	0.3081	$9.610{\times}10^{-2}$	2.500×10^{-3}	0.9975
2	0.1049	1.100×10^{-3}	0.3636	0.1455	2.900×10^{-3}	0.9971
3	3.900×10^{-2}	1.744×10^{-4}	0.3721	0.1562	3.000×10^{-3}	0.9700
4	1.460×10^{-2}	2.773×10^{-5}	0.3738	0.1590	3.000×10^{-3}	0.9700
5	5.500×10^{-3}	4.430×10^{-6}	0.3741	0.1598	3.000×10^{-3}	0.9700
6	2.200×10^{-3}	7.088×10^{-7}	0.3742	0.1600	3.000×10^{-3}	0.9700
7	7.641×10^{-4}	1.135×10^{-7}	0.3742	0.1601	3.000×10^{-3}	0.9700
			⋮			
16	1.092×10^{-7}	8.375×10^{-15}	0.3742	0.1701	3.000×10^{-3}	0.9970
17	4.019×10^{-8}	2.585×10^{-15}	0.3742	0.3087	3.00×10^{-3}	0.9970
18	1.537×10^{-8}	2.205×10^{-15}	0.3742	0.8530	3.000×10^{-3}	0.9970
19	5.752×10^{-9}	2.550×10^{-15}	0.3742	1.156	3.000×10^{-3}	0.9970
			⋮			
24	4.218×10^{-11}	4.741×10^{-15}	0.3742	1.115	3.000×10^{-3}	0.9700
25	1.578×10^{-11}	5.417×10^{-15}	0.3741	1.143	3.000×10^{-3}	0.9700

However, this is at the cost of a higher computational expense to achieve each new approximation. By the twenty-fifth re-iterate, both approximations are tending towards a settled divergence.

Table 7.3: Re-iterated Galerkin

$\alpha = 1$; $\mathcal{S}_N = \mathcal{S}_3$; refinement: 40×40
Trial space 1 : (7.11); Trial space 2: (7.12)

Re-iterate	$\|\|\hat{r}_{1,n}\|\|$	$\|\|\hat{r}_{2,n}\|\|$	$\frac{\|\|\hat{r}_{1,n}\|\|}{\|\|\hat{r}_{1,n-1}\|\|}$	$\frac{\|\|\hat{r}_{2,n}\|\|}{\|\|\hat{r}_{2,n-1}\|\|}$	$\|R\|^2$	$\|T\|^2$
Galerkin	0.1390	0.6340				
0	5.280×10^{-2}	2.670×10^{-2}			3.000×10^{-3}	0.9970
1	8.700×10^{-3}	2.500×10^{-3}	0.1641	0.0950	3.000×10^{-3}	0.9970
2	1.600×10^{-3}	3.00×10^{-4}	0.1838	0.1184	3.000×10^{-3}	0.9970
3	3.022×10^{-4}	3.61×10^{-5}	0.1900	0.1206	3.000×10^{-3}	0.9700
⋮						
14	4.943×10^{-12}	3.700×10^{-15}	0.1969	0.1598	3.000×10^{-3}	0.9970
15	9.724×10^{-13}	2.490×10^{-15}	0.1967	0.6729	3.00×10^{-3}	0.9970
16	1.919×10^{-13}	2.554×10^{-15}	0.1973	1.026	3.000×10^{-3}	0.9970
⋮						
18	7.502×10^{-15}	3.290×10^{-15}	0.1984	1.136	3.000×10^{-3}	0.9700
19	1.856×10^{-15}	3.654×10^{-15}	0.2475	1.111	3.000×10^{-3}	0.9700
20	9.415×10^{-16}	4.089×10^{-15}	0.5071	1.119	3.000×10^{-3}	0.9970
21	7.267×10^{-16}	4.539×10^{-15}	0.7719	1.110	3.000×10^{-3}	0.9970
22	6.880×10^{-16}	5.057×10^{-15}	0.9467	1.114	3.000×10^{-3}	0.9970
23	6.847×10^{-16}	5.635×10^{-15}	0.9952	1.114	3.000×10^{-3}	0.9970
24	6.999×10^{-16}	6.259×10^{-15}	1.022	1.111	3.00×10^{-3}	0.9970
25	7.155×10^{-16}	6.792×10^{-15}	1.022	1.085	3.00×10^{-3}	0.9970

Case $\alpha = 2$: Tables 7.4-7.8 are the corresponding results for $\alpha = 2$.

The results of Sloan iteration show that, as expected, an increase in the value of α has increased the spectral radius of the operator $\tilde{\mathcal{G}}M$, with

$$\rho_{E(D')}(\tilde{\mathcal{G}}M) \approx 2.46,$$

and

$$\rho_{O(D')}(\tilde{\mathcal{G}}M) \approx 1.099, \tag{7.13}$$

for $\alpha = 2$. This provides a sterner test for the re-iterated Galerkin method, and convergence should no longer be expected in problem 1 for a small trial space 1.

Indeed, the results of table 7.5, where

$$\text{Trial Space 1: } \sin(y_0)$$

and

$$\text{Trial Space 2: } \sin(2x_0)\sin(y_0).$$

Table 7.4: Sloan iteration

$\alpha = 2$; $\mathcal{S}_N = \mathcal{S}_3$; refinement: 40 × 40
Trial space 1 : (7.7); Trial space 2: (7.8)

Iterate	$\|\|\check{r}_{1,n}\|\|$	$\|\|\check{r}_{2,n}\|\|$	$\frac{\|\|\check{r}_{1,n}\|\|}{\|\|\check{r}_{1,n-1}\|\|}$	$\frac{\|\|\check{r}_{2,n}\|\|}{\|\|\check{r}_{2,n-1}\|\|}$
0	0.9528	1.244		
1	2.847	1.054	2.988	0.8476
2	5.318	1.158	1.868	1.099
3	14.23	1.273	2.675	1.099
4	34.18	1.399	2.403	1.099
5	84.84	1.5378	2.482	1.099
6	208.52	1.691	2.458	1.099
7	514.0	1.858	2.465	1.099
8	1.266×10^3	2.043	2.463	1.099
9	3.119×10^3	2.246	2.464	1.099
10	7.682×10^3	2.269	2.463	1.099

show divergence, although not in the predicted problem. In addition, table 7.5 provides the approximation $\rho_{O(D')}(S) \approx 3.505$, for the spectral radius of the Sloan operator in the second problem, which is far larger than the approximation to $\rho_{O(D')}(\tilde{\mathcal{G}}M)$, (7.13). Therefore, in this case, the divergence has worsened by using re-iterated Galerkin over regular Sloan iteration. This is the type of behavior that has never been ruled out, but that is difficult to attribute. As the spectral radius has grown from the operator $\tilde{\mathcal{G}}M$ to S, it would not appear that the divergence is related to the trial function $\sin(2x_0)\sin(y_0)$ being a poor approximation to the largest eigenfunction of $\tilde{\mathcal{G}}M$.

Table 7.5: Re-iterated Galerkin

$\alpha = 2$; $\mathcal{S}_N = \mathcal{S}_3$; refinement: 40 × 40
Trial space 1 : (7.9); Trial space 2: (7.10)

Iterate	$\|\|\hat{r}_{1,n}\|\|$	$\|\|\hat{r}_{2,n}\|\|$	$\frac{\|\|\hat{r}_{1,n}\|\|}{\|\|\hat{r}_{1,n-1}\|\|}$	$\frac{\|\|\hat{r}_{2,n}\|\|}{\|\|\hat{r}_{2,n-1}\|\|}$	$\|R\|^2$	$\|T\|^2$
Galerkin	0.9562	5.897				
0	2.521	4.351			0.4505	0.5495
1	1.535	15.23	0.6088	3.500	0.2122	0.7878
2	1.148	53.37	0.7483	3.505	0.1645	0.8355
3	0.8862	187.0	0.7718	3.505	0.1437	0.8563
			$\vdots$			
9	0.1961	3.465×10^5	0.7780	3.505	0.1197	0.8803
10	0.1526	1.215×10^6	0.7780	3.505	0.1192	0.8808

Expanding the trial spaces to

$$\text{Trial Space 1: } \begin{cases} \sin(y_0), \\ \cos(2x_0)\sin(y_0), \end{cases}$$

and

$$\text{Trial Space 2:}\ \begin{cases} \sin(2x_0)\sin(y_0), \\ \sin(4x_0)\sin(y_0), \end{cases}$$

improves matters (see table 7.6) and provides convergence in both problems. However, the extremely slow convergence in problem 2 is impractical.

Table 7.6: Re-iterated Galerkin

$\alpha = 2$; $\mathcal{S}_N = \mathcal{S}_3$; refinement: 40×40
Trial space 1 : (7.11); Trial space 2: (7.12)

Iterate	$\lVert\hat{r}_{1,n}\rVert$	$\lVert\hat{r}_{2,n}\rVert$	$\frac{\lVert\hat{r}_{1,n}\rVert}{\lVert\hat{r}_{1,n-1}\rVert}$	$\frac{\lVert\hat{r}_{2,n}\rVert}{\lVert\hat{r}_{2,n-1}\rVert}$	$\lvert R\rvert^2$	$\lvert T\rvert^2$
Galerkin	0.2795	2.557				
0	0.4454	1.179			7.020×10^{-2}	0.9298
1	0.2367	1.131	0.5314	0.9598	5.670×10^{-2}	0.9433
2	0.1311	1.118	0.5540	0.9882	5.190×10^{-2}	0.9481
3	7.300×10^{-2}	1.102	0.5564	0.9857	4.790×10^{-2}	0.9521
			$\vdots$			
9	2.100×10^{-3}	1.013	0.5548	0.9860	2.310×10^{-2}	0.9769
10	1.200×10^{-3}	0.9983	0.5547	0.9860	1.840×10^{-2}	0.9816

The results that come from further expanding the trial spaces to

$$\text{Trial Space 1:}\ \begin{cases} \sin(y_0), \\ \cos(2x_0)\sin(y_0), \\ \sin(3y_0) \end{cases} \tag{7.14}$$

and

$$\text{Trial Space 2:}\ \begin{cases} \sin(2x_0)\sin(y_0), \\ \sin(4x_0)\sin(y_0), \\ \sin(2x_0)\sin(3y_0), \end{cases} \tag{7.15}$$

are recorded in table 7.7. Observe the sharp increase in the rate of convergence in problem 2, that has been created by adding the trial function $\sin(2x_0)\sin(3y_0)$. Also note that, for the current and previous trial space 1, the norm of the residual error deteriorates between the Galerkin approximation and the iterated Galekin approximation. This is not unexpected, and does not indicate that the norm of the pointwise error is deteriorating (see §6.4).

Results of an even further expansion of the trial spaces to

$$\text{Trial Space 1:}\ \begin{cases} \sin(y_0), \\ \cos(2x_0)\sin(y_0), \\ \sin(3y_0), \\ \cos(4x_0)\sin(y_0), \\ \cos(2x_0)\sin(3y_0), \\ \sin(5y_0), \end{cases} \tag{7.16}$$

Table 7.7: Re-iterated Galerkin

$\alpha = 2$; $\mathcal{S}_N = \mathcal{S}_3$; refinement: 40×40
Trial space 1 : (7.14); Trial space 2: (7.15)

Iterate	$\|\|\hat{r}_{1,n}\|\|$	$\|\|\hat{r}_{2,n}\|\|$	$\frac{\|\|\hat{r}_{1,n}\|\|}{\|\|\hat{r}_{1,n-1}\|\|}$	$\frac{\|\|\hat{r}_{2,n}\|\|}{\|\|\hat{r}_{2,n-1}\|\|}$	$\|R\|^2$	$\|T\|^2$
Galerkin	0.2606	0.7044				
0	0.3026	0.1048			0.2127	0.7873
1	0.1403	2.400×10^{-3}	0.4636	2.260×10^{-2}	0.2011	0.7989
2	7.080×10^{-2}	4.588×10^{-4}	0.5048	0.1935	0.1985	0.8015
3	3.620×10^{-2}	4.365×10^{-5}	0.5117	9.510×10^{-2}	0.1978	0.8022
4	1.860×10^{-2}	5.720×10^{-6}	0.5135	0.1310	0.1976	0.8024
⋮						
9	6.695×10^{-4}	1.668×10^{-10}	0.5144	0.1253	0.1976	0.8024
10	3.444×10^{-4}	5.257×10^{-10}	0.5144	0.3152	0.1976	0.8024
11	1.772×10^{-4}	1.191×10^{-10}	0.5144	2.266	0.1976	0.8024
⋮						
19	8.691×10^{-7}	1.615×10^{-7}	0.5144	2.463	0.1976	0.8024
20	4.471×10^{-7}	3.978×10^{-7}	0.5144	2.463	0.1976	0.8024

and

$$\text{Trial Space 2: } \begin{cases} \sin(2x_0)\sin(y_0), \\ \sin(4x_0)\sin(y_0), \\ \sin(2x_0)\sin(3y_0), \\ \sin(6x_0)\sin(y_0), \\ \sin(4x_0)\sin(3y_0), \\ \sin(2x_0)\sin(5y_0), \end{cases} \tag{7.17}$$

can be found in table 7.8. The most noticeable effect is that the convergence rates of both problems are improved. The improvement is far greater in both size, and relative size, in problem 1. Also this improvement in problem 1 is far greater than the improvement gained by the last expansion of trial space (between table 7.6 and table 7.7). This could be a product of the extra number of terms added or the properties of the terms added. This relates to the possible investigation of the optimal expansion of the trial space that could be conducted.

Case $\alpha = 3$: As a further test of the re-iterated Galerkin method, and by way of confirmation of the behaviour we have witnessed thus far, we increase α again.

Table 7.9 produces the approximations

$$\rho_{E(D')}(\tilde{\mathcal{G}}M) \approx 4.194,$$

$$\rho_{O(D')}(\tilde{\mathcal{G}}M) \approx 1.892,$$

for $\alpha = 3$. Thus, we deduce that increasing the value of α has again increased the spectral radius of the operator $\tilde{\mathcal{G}}M$ in both problems.

Table 7.8: Re-iterated Galerkin

$\alpha = 2$; $\mathcal{S}_N = \mathcal{S}_3$; refinement: 40 × 40
Trial space 1 : (7.16); Trial space 2: (7.17)

Iterate	$\|\|\hat{r}_{1,n}\|\|$	$\|\|\hat{r}_{2,n}\|\|$	$\frac{\|\|\hat{r}_{1,n}\|\|}{\|\|\hat{r}_{1,n-1}\|\|}$	$\frac{\|\|\hat{r}_{2,n}\|\|}{\|\|\hat{r}_{2,n-1}\|\|}$	$\|R\|^2$	$\|T\|^2$
Galerkin	0.1450	0.5417				
0	4.450×10^{-2}	2.170×10^{-2}	0.4636	2.260×10^{-2}	0.1987	0.8013
1	9.200×10^{-3}	1.700×10^{-3}	0.2060	7.610×10^{-2}	0.1976	0.8024
2	2.000×10^{-3}	1.614×10^{-4}	0.2199	9.770×10^{-2}	0.1976	0.8024
3	4.521×10^{-4}	$1.680{\times}10^{-5}$	0.2242	0.1041	0.1976	0.8024
⋮						
9	6.044×10^{-8}	$2.735{\times}10^{-11}$	0.2263	0.1103	0.1976	0.8024
10	1.368×10^{-8}	$1.093{\times}10^{-11}$	0.2263	0.3995	0.1976	0.8024
11	3.095×10^{-9}	$2.594{\times}10^{-11}$	0.2263	2.373	0.1976	0.8024
⋮						
19	2.546×10^{-14}	$3.516{\times}10^{-8}$	0.2706	2.463	0.1976	0.8024
20	1.589×10^{-14}	$8.660{\times}10^{-8}$	0.6241	2.463	0.1976	0.8024

Table 7.9: Sloan iteration

$\alpha = 3$; $\mathcal{S}_N = \mathcal{S}_3$; refinement: 40 × 40
Trial space 1 : (7.7); Trial space 2: (7.8)

Iterate	$\|\|\check{r}_{1,n}\|\|$	$\|\|\check{r}_{2,n}\|\|$	$\frac{\|\|\check{r}_{1,n}\|\|}{\|\|\check{r}_{1,n-1}\|\|}$	$\frac{\|\|\check{r}_{2,n}\|\|}{\|\|\check{r}_{2,n-1}\|\|}$
0	0.9852	3.052		
1	5.057	3.294	5.133	1.605
2	15.88	6.247	3.139	1.896
3	72.71	11.82	4.580	1.892
4	296.9	22.36	4.084	1.892
5	1.256×10^{3}	42.31	4.229	1.892
6	5.253×10^{3}	80.05	4.184	1.892
7	2.205×10^{4}	151.5	4.197	1.892
8	9.245×10^{4}	286.6	4.193	1.892
9	3.878×10^{5}	542.2	4.195	1.892
10	1.627×10^{6}	1.026×10^{3}	4.194	1.892

Results with varying trial spaces are recorded in tables 7.10-7.14, with the largest trial spaces being

$$\text{Trial Space 1:} \begin{cases} \sin(y_0), \\ \cos(2x_0)\sin(y_0), \\ \sin(3y_0), \\ \cos(4x_0)\sin(y_0), \\ \cos(2x_0)\sin(3y_0), \\ \sin(5y_0), \\ \cos(6x_0)\sin(y_0), \\ \cos(4x_0x_0)\sin(3y_0), \\ \cos(2x_0)\sin(5y_0), \\ \sin(7y_0), \end{cases} \tag{7.18}$$

and

$$\text{Trial Space 2:} \begin{cases} \sin(2x_0)\sin(y_0), \\ \sin(4x_0)\sin(y_0), \\ \sin(2x_0)\sin(3y_0), \\ \sin(6x_0)\sin(y_0), \\ \sin(4x_0)\sin(3y_0), \\ \sin(2x_0)\sin(5y_0), \\ \sin(8x_0)\sin(y_0), \\ \sin(6x_0)\sin(5y_0), \\ \sin(4x_0)\sin(5y_0), \\ \sin(2x_0)\sin(7y_0). \end{cases} \tag{7.19}$$

Table 7.10: Re-iterated Galerkin

$\alpha = 3$; $\mathcal{S}_N = \mathcal{S}_3$; refinement: 40×40
Trial space 1 : (7.9); Trial space 2: (7.10)

Re-iterate	$\|\hat{r}_{1,n}\|$	$\|\hat{r}_{2,n}\|$	$\frac{\|\hat{r}_{1,n}\|}{\|\hat{r}_{1,n-1}\|}$	$\frac{\|\hat{r}_{2,n}\|}{\|\hat{r}_{2,n-1}\|}$	$\|R\|^2$	$\|T\|^2$
Galerkin	0.9969	1.199				
0	4.690	1.265			0.5816	0.4184
1	4.409	0.1003	0.9402	7.930×10^{-2}	0.6328	0.3672
2	5.262	2.360×10^{-2}	1.194	0.2352	0.4962	0.5038
3	6.515	2.100×10^{-3}	1.2380	8.840×10^{-2}	0.4334	0.5666
4	8.128	1.100×10^{-3}	1.2476	0.5499	0.3988	0.6012
⋮						
12	48.56	2.018×10^{-7}	1.250	0.7009	0.3469	0.6531
13	60.72	7.607×10^{-7}	1.250	3.770	0.3464	0.6536
14	75.93	3.188×10^{-6}	1.250	4.191	0.3461	0.6539
⋮						
20	209.2	1.740×10^{-2}	1.250	4.194	0.3455	0.6555

A new phenomenon that is discovered is that an increase in trial space 1 from (7.9) to (7.11), and (7.11) to (7.14) (between table 7.10, 7.11 and 7.12) causes the convergence rate of

Table 7.11: Re-iterated Galerkin

$\alpha = 3$; $\mathcal{S}_N = \mathcal{S}_3$; refinement: 40×40
Trial space 1 : (7.11); Trial space 2: (7.12)

Re-iterate	$\|\hat{r}_{1,n}\|$	$\|\hat{r}_{2,n}\|$	$\frac{\|\hat{r}_{1,n}\|}{\|\hat{r}_{1,n-1}\|}$	$\frac{\|\hat{r}_{2,n}\|}{\|\hat{r}_{2,n-1}\|}$	$\|R\|^2$	$\|T\|^2$
Galerkin	2.867	0.6383				
0	6.961	0.1789			0.3743	0.6257
1	24.96	4.150×10^{-2}	3.590	0.2322	0.3474	0.6526
2	89.16	1.100×10^{-2}	3.572	0.2646	0.3457	0.6543
$\vdots$						
12	2.948×10^{6}	1.196×10^{-7}	3.564	0.5112	0.3455	0.6545
13	1.051×10^{8}	4.005×10^{-7}	3.564	3.348	0.3455	0.6545
14	3.744×10^{8}	1.677×10^{-6}	3.564	4.188	0.3455	0.6545
15	1.334×10^{9}	7.035×10^{-6}	3.564	4.194	0.3455	0.6555

Table 7.12: Re -iterated Galerkin

$\alpha = 3$; $\mathcal{S}_N = \mathcal{S}_3$; refinement: 40×40
Trial space 1 : (7.14); Trial space 2: (7.15)

Re-iterate	$\|\hat{r}_{1,n}\|$	$\|\hat{r}_{2,n}\|$	$\frac{\|\hat{r}_{1,n}\|}{\|\hat{r}_{1,n-1}\|}$	$\frac{\|\hat{r}_{2,n}\|}{\|\hat{r}_{2,n-1}\|}$	$\|R\|^2$	$\|T\|^2$
Galerkin	3.336	0.6059				
0	6.809	4.570×10^{-2}			0.3523	0.6477
1	27.74	6.200×10^{-3}	4.074	0.1364	0.3460	0.6540
2	113.8	1.400×10^{-3}	4.101	0.2296	0.3455	0.6545
$\vdots$						
11	3.731×10^{7}	2.519×10^{-8}	4.101	0.4743	0.3455	0.6545
12	1.530×10^{8}	8.332×10^{-8}	4.101	3.308	0.3455	0.6545
$\vdots$						
15	1.055×10^{10}	6.138×10^{-6}	4.101	4.194	0.3455	0.6555

Table 7.13: Re-iterated Galerkin

$\alpha = 3$; $\mathcal{S}_N = \mathcal{S}_3$; refinement: 40×40
Trial space 1 : (7.16); Trial space 2: (7.17)

Re-iterate	$\|\hat{r}_{1,n}\|$	$\|\hat{r}_{2,n}\|$	$\frac{\|\hat{r}_{1,n}\|}{\|\hat{r}_{1,n-1}\|}$	$\frac{\|\hat{r}_{2,n}\|}{\|\hat{r}_{2,n-1}\|}$	$\|R\|^2$	$\|T\|^2$
Galerkin	0.3045	0.5268				
0	0.1985	2.800×10^{-2}			6.200×10^{-3}	0.9938
1	8.000×10^{-3}	4.000×10^{-3}	4.010×10^{-3}	0.1436	1.996×10^{-4}	0.9998
2	3.300×10^{-3}	7.442×10^{-4}	0.4193	0.1854	1.976×10^{-4}	0.9998
3	7.336×10^{-4}	1.472×10^{-4}	0.2196	0.1978	1.978×10^{-4}	0.9998
4	1.998×10^{-4}	2.985×10^{-5}	0.2723	0.2028	1.978×10^{-4}	0.9998
5	5.376×10^{-5}	6.121×10^{-6}	0.2691	0.2051	1.978×10^{-4}	0.9998
6	1.483×10^{-5}	1.262×10^{-6}	0.2758	0.2062	1.978×10^{-4}	0.9998
			⋮			
9	3.261×10^{-7}	1.121×10^{-8}	0.2818	0.2077	1.978×10^{-4}	0.9998
10	9.221×10^{-8}	4.487×10^{-6}	0.2827	0.4001	1.978×10^{-4}	0.9998
11	2.613×10^{-8}	1.613×10^{-6}	0.2834	3.595	1.978×10^{-4}	0.9998
12	7.419×10^{-9}	6.762×10^{-8}	0.2839	4.192	1.978×10^{-4}	0.9998
			⋮			
15	1.769×10^{-10}	4.989×10^{-6}	0.2947	4.194	1.978×10^{-4}	0.9998
16	1.031×10^{-10}	2.092×10^{-5}	0.5829	4.194	1.978×10^{-4}	0.9998
17	1.796×10^{-10}	8.7758×10^{-5}	1.741	4.194	1.978×10^{-4}	0.9998
			⋮			
20	1.370×10^{-9}	6.500×10^{-3}	1.972	4.194	1.978×10^{-4}	0.9998

Table 7.14: Re-iterated Galerkin

$\alpha = 3$; $\mathcal{S}_N = \mathcal{S}_3$; refinement: 40×40
Trial space 1 : (7.18); Trial space 2: (7.19)

Re-iterate	$\|\hat{r}_{1,n}\|$	$\|\hat{r}_{2,n}\|$	$\frac{\|\hat{r}_{1,n}\|}{\|\hat{r}_{1,n-1}\|}$	$\frac{\|\hat{r}_{2,n}\|}{\|\hat{r}_{2,n-1}\|}$	$\|R\|^2$	$\|T\|^2$
Galerkin	8.220×10^{-2}	0.4597				
0	1.860×10^{-2}	1.070×10^{-2}			6.200×10^{-3}	0.9938
1	3.000×10^{-3}	7.571×10^{-4}	0.1596	7.100×10^{-2}	1.977×10^{-4}	0.9998
2	5.491×10^{-4}	7.826×10^{-5}	0.1849	0.1034	1.978×10^{-4}	0.9998
			⋮			
8	3.257×10^{-8}	4.532×10^{-10}	0.1994	0.1699	1.978×10^{-4}	0.9998
9	6.497×10^{-9}	1.086×10^{-9}	0.1995	2.397	1.978×10^{-4}	0.9998
10	1.297×10^{-9}	4.551×10^{-9}	0.1995	4.189	1.978×10^{-4}	0.9998
			⋮			
13	1.521×10^{-11}	3.357×10^{-7}	0.2926	4.194	1.978×10^{-4}	0.9998
14	2.219×10^{-11}	1.408×10^{-6}	1.459	4.194	1.978×10^{-4}	0.9998
15	4.368×10^{-11}	5.906×10^{-6}	1.969	4.194	1.978×10^{-4}	0.9998
			⋮			
20	1.323×10^{-9}	7.700×10^{-7}	1.979	4.194	1.977×10^{-4}	0.9998

problem 1 to worsen, rather than improve. It is important to note that, the properties of the operators that have been spoken of, are, in the main, made according to a limit. In particular, it was never said that an expansion in the trial space would always cause the spectral radius of the Sloan operator to fall. Rather, in the limit that the dimension of the trial space tends to infinity, the spectral radius of the Sloan operator tends to zero. This is a good demonstration of this point. The increasing divergence does not continue indefinitely. As trial space 1 is further expanded, the convergence of problem 1 improves (see table 7.13 and 7.14), i.e. the approximation to the spectral radius of the Sloan operator gets smaller.

Tables 7.2-7.14 all contain approximations to two different problems, where the sizes of the trial spaces are equal, and the trial spaces themselves are analogous. A couple of comparisons are now made about the general differences between the two approximations. Firstly, for convergent approximations, the Galerkin approximation to problem 1 is generally superior to the Galerkin approximation of problem 2, or it at least appears to tend towards the exact solution faster as the trial spaces are expanded. Secondly, an effect which balances the first point, the convergence rate, of the re-iterated Galerkin method, in the second problem is faster than that of the first, in the main. This is surely connected with the fact that the approximations to the spectral radius of $\tilde{\mathcal{G}}M$ are considerably smaller in problem 2, than problem 1. These points will remain as observations. They fall under the remit of possible future investigations into optimality of the trial spaces.

Let us now return to an unresolved issue – what causes a sequence of re-iterated Galerkin approximations, that have a settled convergence rate, to become disrupted and subsequently diverge?

7.4 Eventual divergence

This problem is highlighted in tables 7.2, 7.3, 7.7, 7.8, 7.10, 7.11, 7.12, 7.13, 7.14, occurring in one or both of the problems. The problem appears to arise from a small residual error. We have already raised the point that an approximation may only be as good as the order of computational rounding error. However, this is not the answer as the rounding error is far smaller than the size of the residual error, at which the approximations go wrong. Also, re-iterating after rounding error has been reached, would cause a 're-shuffling' of errors, rather than the consistent divergence that we are finding.

Consider the following explanation.

The behaviour we are witnessing, seems to suggest that the re-iterated Galerkin method finds a new eigenvalue of the Sloan operator when the residual gets small – but why does this eigenvalue arise only for a small residual error, and why can it not be eliminated by expansion of the trial space? In §7.1, we discussed the properties of the Hilbert spaces in which each problem is set, and chose trial spaces accordingly. These Hilbert spaces we denoted $E(D')$ and $O(D')$, and are subspaces of the 'global' Hilbert space $L_2(D')$. The appearance of the subspaces was essentially a product of the various symmetries in the problems. By choosing trial functions from these subspaces, we are attempting to minimize the spectral radius of the Sloan operator over the subspace in question, and doing little, or nothing, to the spectral radius of the Sloan operator over the rest of $L_2(D')$. Whilst the various factors involved in the problem are relatively large, the effect of computational errors are negligible, relatively, and the problem maintains its true 'shape'. Hence, the re-iterated Galerkin method behaves as the theory suggests it would. However, as the residual error decreases, remembering that improved approximations are calculated from the residual error, computational errors become more pronounced, destroying the symmetries of the problem. Under these effects, the problem being dealt with on the computer moves from the subspace into the global Hilbert space, $L_2(D')$, and the convergence rate begins to seek out the spectral radius of the Sloan operator over $L_2(D')$, rather than the subspace. As our choice of trial space has not taken into account this possibility, the spectral radius of the Sloan operator over $L_2(D')$ can be expected to be large, and cause divergence.

The evidence for this hypothesis seems irrefutable. For example, note that, for a constant α, the eventual divergence rates appear to be settling to the same value (this is particularly clear in the tables presented for problem 2: $\alpha = 3$).

The effects of computational error on the residual are also visible. Figures 7.2-7.4 sample the residual error at three re-iterates for problem 2, with $\alpha = 1$, and trial space 2 as in (7.10). This corresponds to table 7.2. They catalogue the deterioration of the symmetry in the residual errors. Sample cross-sectional profiles are also included, all at the arbitrary value $y_0 = 1.9242$.

It is also evident that the higher the value of α, the larger the residual when the effects of computational error take hold. This is unsurprising, as, when calculating the residual error in an approximation, computational errors may be magnified by as much as $||\tilde{\mathcal{G}}M||$, which we may assume is generally larger for larger α.

If there were a requirement that approximations be made to a very high degree of accuracy, or we were using operators of very large size, we would need to investigate as to how the effects of computational error may be counteracted. A possible way of doing this would be to add in trial functions that would be of no use in the analytic problem, but would reduce the spectral radius of the Sloan operator over $L_2(D')$. As it is clear that we can already produce sufficiently

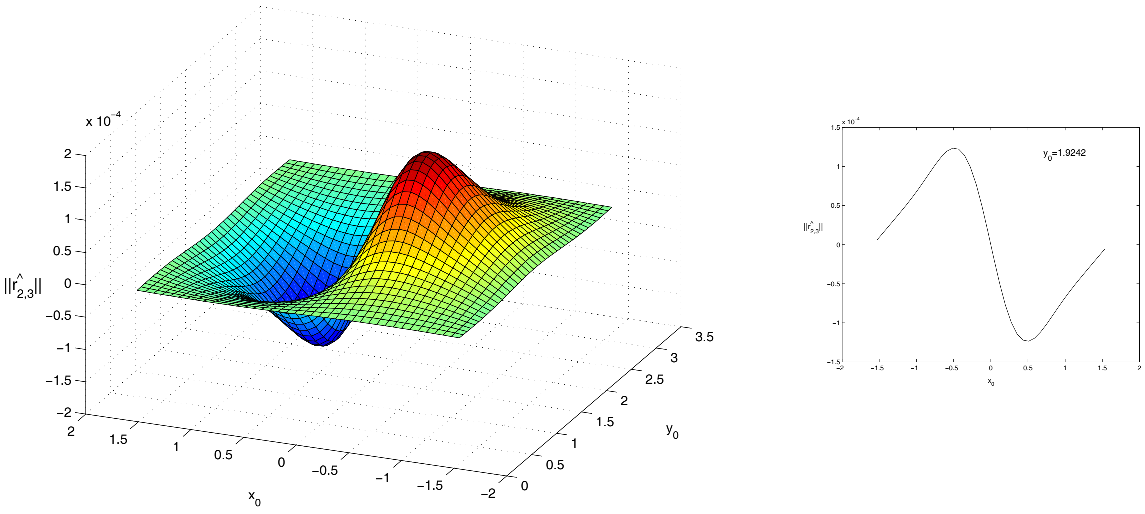

Figure 7.2: Residual error in problem 2 after 3 re-iterations and cross-section at $y_0 = 1.9242$
$\alpha = 1$; Trial space 2: (7.10)

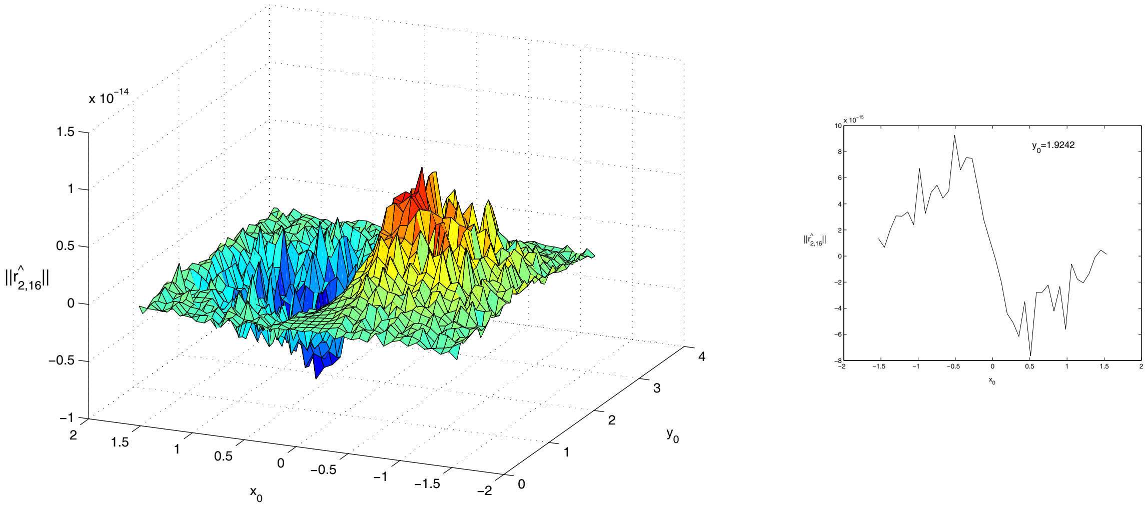

Figure 7.3: Residual error in problem 2 after 16 re-iterations and cross-section at $y_0 = 1.9242$
$\alpha = 1$; Trial space 2: (7.10)

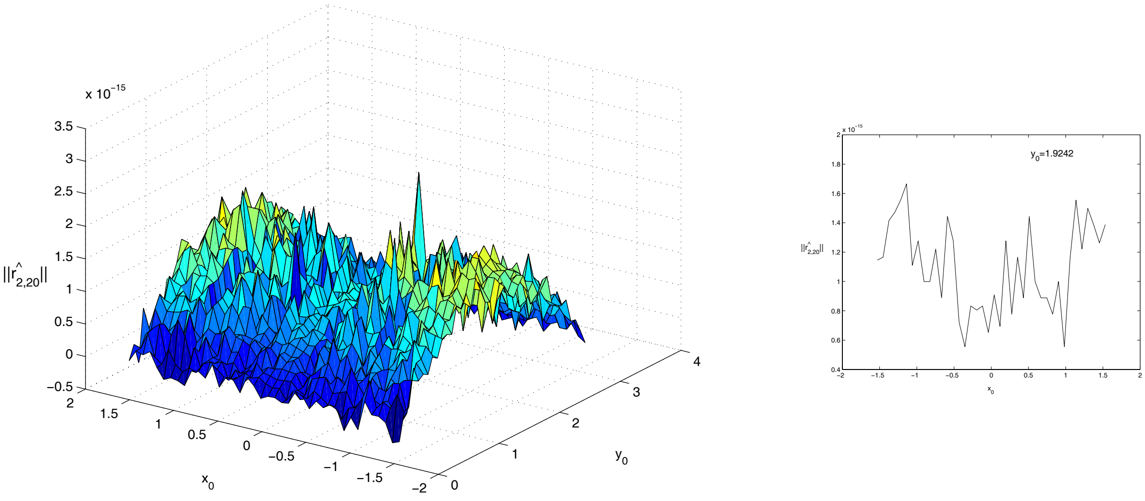

Figure 7.4: Residual error in problem 2 after 20 re-iterations and cross-section at $y_0 = 1.9242$ $\alpha = 1$; Trial space 2: (7.10)

good approximations to $|R|^2$ and $|T|^2$, this point will remain as only a possible avenue for further study.

In comparison, we seek only 3 decimal place accuracy in the quantities $|R|^2$ and $|T|^2$, and will use values of α no greater than 3. Tables 7.2-7.14 indicate that, to achieve convergence to 3 decimal places in $|R|^2$ and $|T|^2$, we require approximations with norm residual error of order approximately 10^{-2} to 10^{-3}, depending on the size of α, in both problems 1 and 2. To be certain, when making approximations, with α no greater than 3, we shall allow our approximations to ϕ_1 and ϕ_2 to reach a residual error of order 10^{-4} before terminating the re-iterations. Having carried out tests on the re-iterated Galerkin method, we turn our attention to the accuracy of the approximations that are produced, in relation to the original problem.

7.5 Accuracy of approximations

In the results produced thus far, we have been working with a fixed quadrature refinement and series approximation. By using a numerical quadrature and truncating the infinite series in the kernel, we introduce errors that affect our results (see §6.5). These errors diminish as we increase the accuracy of the quadrature, by increasing the refinement of the computational grid, and decrease the truncation error in the series approximation, by taking an increased number of terms. The effect of varying these errors may be judged from the approximate values of $|R|^2$ and $|T|^2$ that they produce. This will allow us to estimate how our program must be set to achieve approximations to 3 decimal place accuracy.

We treat these two causes of error separately.

7.5.1 Quadrature refinement

To make an accurate comparison, all other variables must be fixed. So, making use of prior results, we shall work with $\alpha = 1, 2$, and 3, choosing a pair of trial spaces, for each α, that produced convergence in $|R|^2$ and $|T|^2$ to four decimal places within 5 re-iterations. All other variables, excepting the refinement, are unchanged.

Re-iterated Galerkin approximations, where varying quadrature refinements are used, are contained in tables 7.15-7.17. Values of $|R|^2$ and $|T|^2$ are recorded after 5 re-iterations, by which time they had converged to at least 4 decimal places.

Table 7.15: Results of varying quadrature refinement after 5 re-iterations

$\alpha = 1$; $\mathcal{S}_N = \mathcal{S}_3$;
Trial space 1 : (7.11); Trial space 2: (7.10)

Refinement	$\|R\|^2$	$\|T\|^2$
10×10	3.200×10^{-3}	0.9968
20×20	3.000×10^{-3}	0.9700
30×30	3.000×10^{-3}	0.9700
40×40	3.000×10^{-3}	0.9700
50×50	3.000×10^{-3}	0.9700

Table 7.16: Results of varying quadrature refinement after 5 re-iterations

$\alpha = 2$; $\mathcal{S}_N = \mathcal{S}_3$;
Trial space 1 : (7.16); Trial space 2: (7.15)

Refinement	$\|R\|^2$	$\|T\|^2$
10×10	0.1958	0.8042
20×20	0.1973	0.8027
30×30	0.1975	0.8025
40×40	0.1976	0.8024
50×50	0.1976	0.8024

We have already seen how a larger operator, produced by a larger value of α, can cause problems with numerical results. This appears to be the case again. The quadrature error may be magnified by as much as $\|\tilde{\mathcal{G}}M\|$. It can be conjectured that this is the reason that approximations are further apart between different quadrature for larger α.

All three show apparent accuracy of $|R|^2$ and $|T|^2$ to 3 decimal places by 40×40 refinement.

7.5.2 Series truncation

Tables 7.18-7.20 give the re-iterated Galerkin approximations, after 5 iterations, with varying series truncation, and the refinement has been fixed at 40×40. Again, values of $|R|^2$ and $|T|^2$ had converged to at least 4 decimal places by the fifth re-iterate.

Table 7.17: Results of varying quadrature refinement after 5 re-iterations

$\alpha = 3$; $\mathcal{S}_N = \mathcal{S}_3$;
Trial space 1 : (7.16); Trial space 2: (7.15)

Refinement	$\|\|R\|\|^2$	$\|\|T\|\|^2$
10×10	6.000×10^{-3}	0.9940
20×20	8.748×10^{-4}	0.9991
30×30	3.323×10^{-4}	0.9997
40×40	1.978×10^{-3}	0.9998
50×50	1.978×10^{-4}	0.9998
60×60	1.216×10^{-4}	0.9999

Table 7.18: Results of varying series truncation after 5 re-iterations

$\alpha = 1$; 40×40
trial space 1 = 2; trial space 2 = 1

Terms	$\|\|R\|\|^2$	$\|\|T\|\|^2$
1	2.900×10^{-3}	0.9971
2	2.900×10^{-3}	0.9971
3	3.000×10^{-3}	0.9970
4	3.000×10^{-3}	0.9970
5	3.000×10^{-3}	0.9970
10	3.000×10^{-3}	0.9970

Table 7.19: Results of varying series truncation after 5 re-iterations

$\alpha = 2$; 40×40
trial space 1 = 2; trial space 2 = 1

Terms	$\|\|R\|\|^2$	$\|\|T\|\|^2$
1	0.2009	0.7991
2	0.2009	0.7991
3	0.1976	0.8024
4	0.1976	0.8024
5	0.1974	0.8026
10	0.1974	0.8026

Table 7.20: Results of varying series truncation after 5 re-iterations

$\alpha = 3$; 40×40
trial space 1 = 6; trial space 2 = 3

Terms	$\|\|R\|\|^2$	$\|\|T\|\|^2$
1	6.100×10^{-3}	0.9939
2	6.100×10^{-3}	0.9939
3	1.978×10^{-3}	0.9998
4	1.978×10^{-3}	0.9998
5	8.082×10^{-5}	0.9999
10	7.877×10^{-3}	0.9999

As terms in the infinite series are ordered in decreasing size, it seems a fair assumption that an increase of five to ten terms, having little relative effect, is a good indication that a further increase in terms used would have even less effect.

It is easily seen that, a low number of terms provides an approximation to the infinite series that produces 'good' approximations to the required quantities. This was predicted in §6.3.1, and somewhat vindicates the trouble taken to modify the original infinite series.

There are evidently terms in the series that, although they would be thought to have a significant size, do nothing to the approximation, for example, term 2. This is undoubtedly due to an orthogonality relationship, and could be pursued as further investigation.

Again, variations appear to have a greater net effect when the value of α is larger. This cannot be attributed to any magnifying effect the operator has on errors. It is likely to be simply that a larger value of α produces larger terms in the series.

The setting of an acoustics problem was used to fuel an investigation into the re-iterated Galerkin method. Having said this, we have arrived at a point at which we are able to produce results, and it will provide a satisfactory and neat end to our work to actually solve the original problem, albeit on a sample set.

7.6 Solutions of the acoustics problem

It is necessary to set limits on a sample set of parameters that we shall solve for. To maintain consistency with what has gone previously, only the parameter α will be varied. In order to keep the size of the operator used relatively small, we restrict α to the interval [-1,2]. This also allows us to investigate negative values of α, which has not been attempted previously as it was found that the operator $\tilde{\mathcal{G}}M$ that they induced allowed convergence from Sloan iteration. All other parameters remain fixed, as defined at the beginning of this chapter.

We may now make practical use of §7.5. As we are looking for 3 decimal place accuracy in $|R|^2$ and $|T|^2$, it seems reasonable to set the quadrature refinement to 40×40, and take $\mathcal{S}_N = \mathcal{S}_5$. To be absolutely certain of the accuracy of our approximations, we should perform similar tests to those in §7.5 on each separate value of α. However, for our purposes this would be highly over-zealous. Our choices of series approximation and quadrature refinement, would appear to be cautious enough to accept that we will achieve the required degree of accuracy.

Figure 7.5 plots the energies $|R|^2$ and $|T|^2$ against α, where values of α have been taken in steps of 0.05. Values of $|R|^2$ and $|T|^2$ were allowed to converge to 3 decimal place, and recorded at this accuracy. The trial spaces used in the re-iterated Galerkin method, were varied as required, although this point is immaterial to figure 7.5.

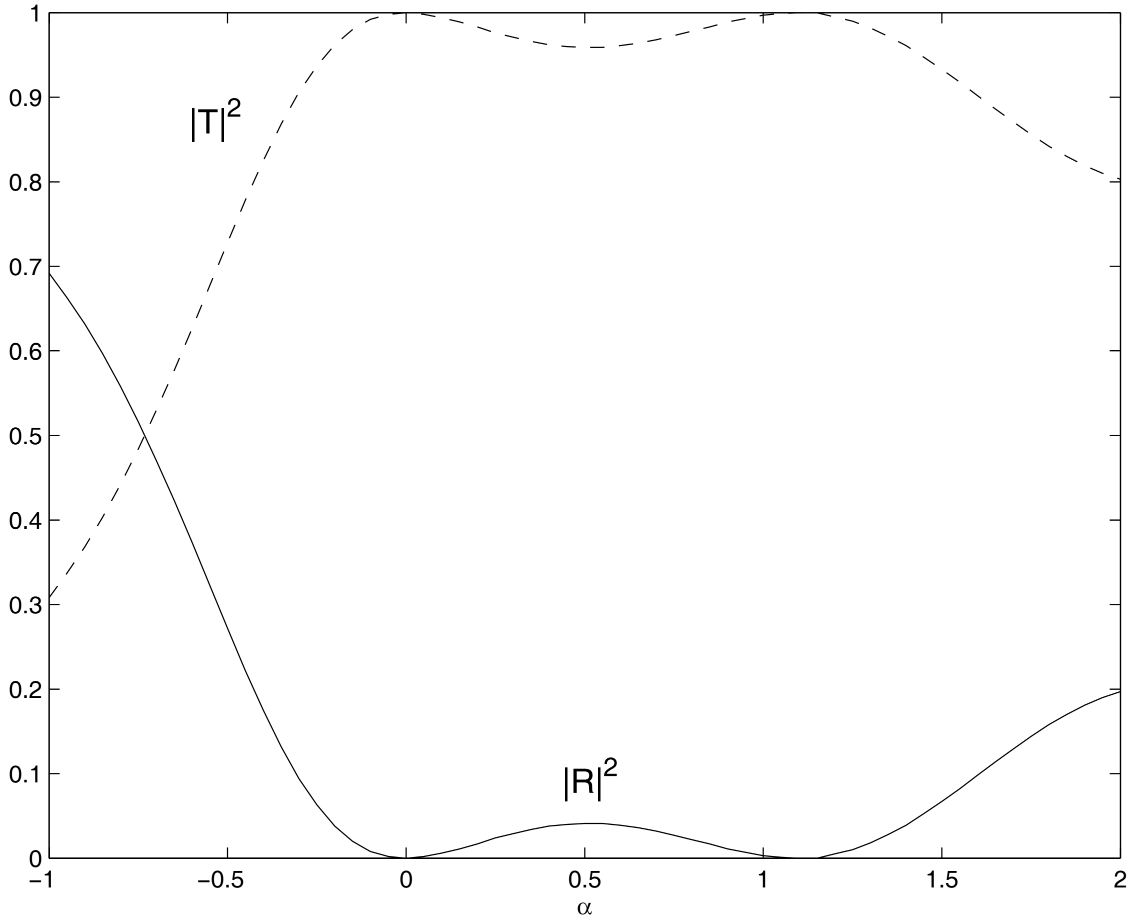

Figure 7.5: Energy proportion approximations against α

Note that, figure 7.5 shows that check ii, the conservation of energy property

$$|R|^2 + |T|^2 = 1,$$

is satisfied. Checks i and v were also satisfied by all approximations, to within a limit that can be attributed to the errors in the approximations and rounding error.

The graphs of $|R|^2$ and $|T|^2$ display 'wobbles'. This is akin to the findings of similar investigations, see for example [5].

With $\alpha = 0$ there is no obstacle, and hence nothing to reflect the incident wave. This is indicated by total transmission, i.e. $|T|^2 = 1$, in figure 7.5. Note that total transmission is produced at one other point, with α just greater than 1. Such occurrences are unsurprising, and attributed to 'interference effects' taking place across the obstacle.

A negative α indicates that the wave speed is greater in the domain D', than the rest of D. Our sample set indicates that a shift in α in the negative direction causes more energy to be reflected than an equivalent shift in the positive direction.

The overall tendency seems to be that, the greater the size of the obstacle, the greater the amount of reflected energy.

In figures 7.6-7.7, we return to investigating the behaviour of the re-iterated Galerkin method. Figure 7.6 plots the size of trial space one used against the number of re-iterations until the norm of the residual error in problem 1 reaches order 10^{-4}. Figure 7.7 is the corresponding graph for problem 2. Both figures 7.6 and 7.7, also plot the number of re-iterations needed for $|R|^2$ and $|T|^2$ to converge to 3 decimal places.

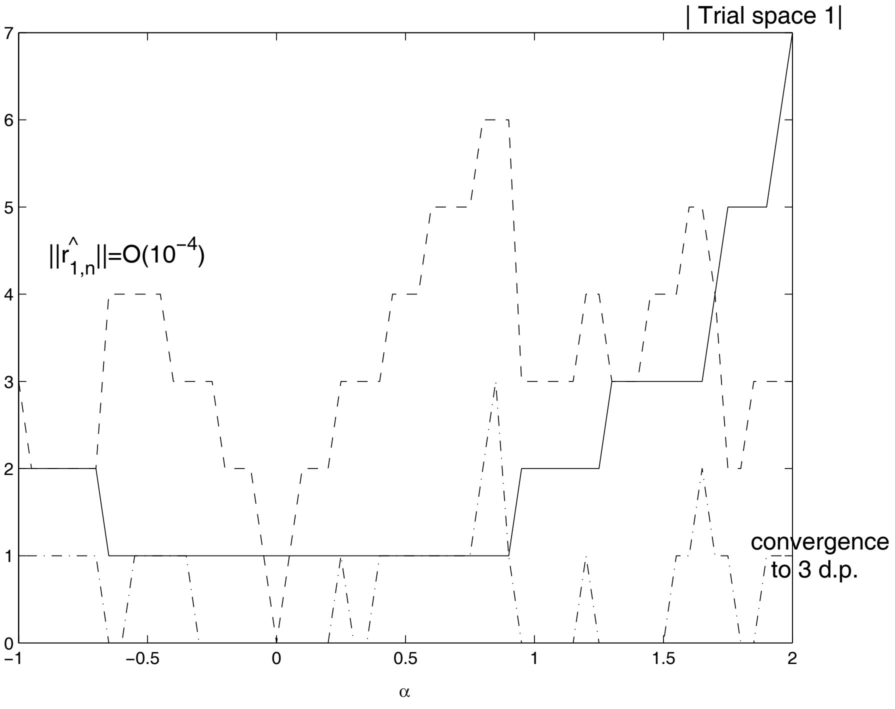

Figure 7.6: Convergence rates and trial spaces in problem 1

The trial spaces are built up in a 'triangular' fashion. This tactic has been adopted throughout this chapter. Essentially, both trial spaces are expanded by adding in more oscillatory terms. Figures 7.8-7.9 are diagrams that number the order in which the functions are added into their respective trial space, and are designed to convey the 'triangular' construction spoken of.

The expansion or contraction of subspaces, in figures 7.6-7.7, followed no exact pattern. Rather, changes were made to ensure convergence, and in a reasonably low number of re-iterations. Also, attempts were made to synchronize the convergence of the two problems.

Note that, in general, a lower sized trial space is required for fast convergence in problem 2. This is attributed to the faster convergence rate of the re-iterated Galerkin method in problem 2, which was also noted at the end of §7.3.

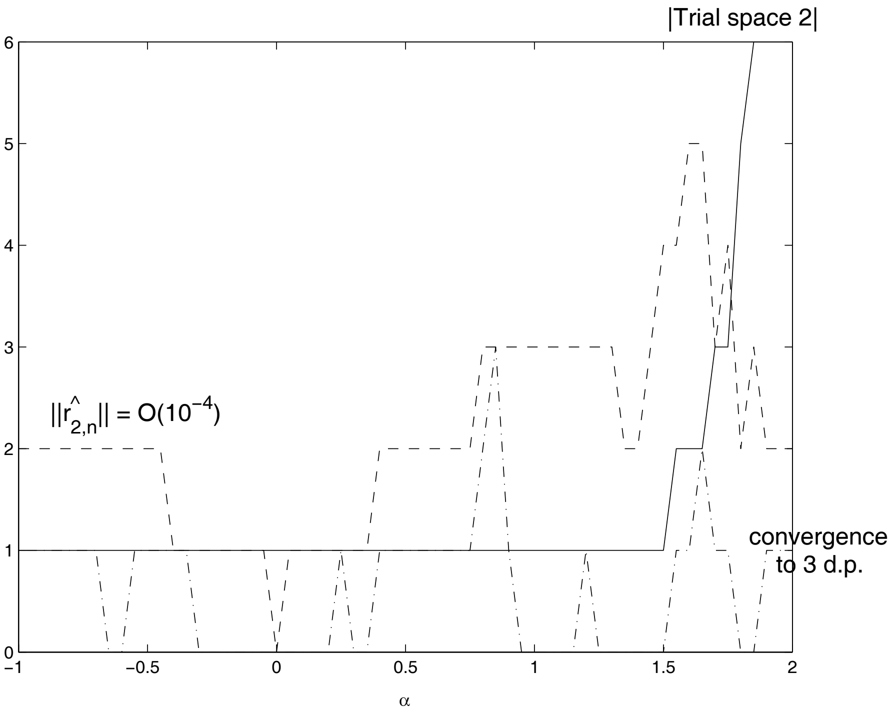

Figure 7.7: Convergence rates and trial spaces in problem 2

Figure 7.8: Trial space 1

$$\begin{tikzcd}
{}_{1}/\ \sin(y_0) \arrow[r, dotted] & {}_{2}/\ \cos(mx_0)\sin(y_0) \arrow[dl, dotted] & {}_{4}/\ \cos(2mx_0)\sin(y_0) \arrow[dl, dotted] \\
{}_{3}/\ \sin(3y_0) \arrow[urr, dotted] & {}_{5}/\ \cos(mx_0)\sin(3y_0) \arrow[dl, dotted] & \\
{}_{6}/\ \sin(5y_0) & &
\end{tikzcd}$$

Figure 7.9: Trial space 2

$$\begin{tikzcd}
{}_{1}/\ \sin(mx_0)\sin(y_0) \arrow[r, dotted] & {}_{2}/\ \sin(2mx_0)\sin(y_0) \arrow[dl, dotted] & {}_{4}/\ \sin(3mx_0)\sin(y_0) \arrow[dl, dotted] \\
{}_{3}/\ \sin(mx_0)\sin(3y_0) \arrow[urr, dotted] & {}_{5}/\ \sin(2mx_0)\sin(3y_0) \arrow[dl, dotted] & \\
{}_{6}/\ \sin(mx_0)\sin(5y_0) & &
\end{tikzcd}$$

It is of interest to see the approximations to ϕ_1 and ϕ_2. Figures 7.10-7.35 display these approximations, at steps of 0.25 in α. All approximations have a norm residual error of order 10^{-4} or less.

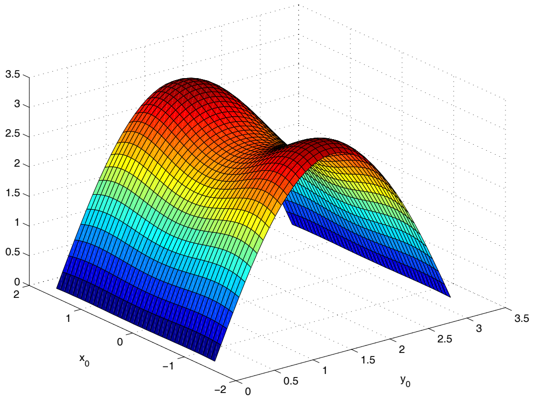

Figure 7.10: $\hat{\sigma}_{1,3}$; $\alpha = -1$

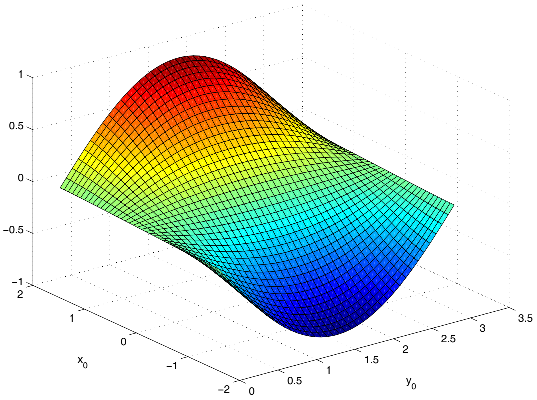

Figure 7.11: $\hat{\sigma}_{2,3}$; $\alpha = -1$

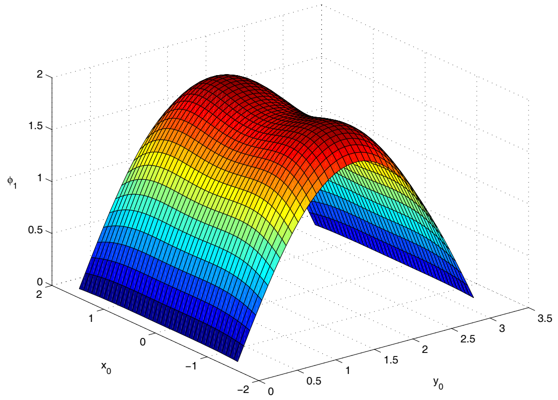

Figure 7.12: $\hat{\sigma}_{1,2}$; $\alpha = -0.75$

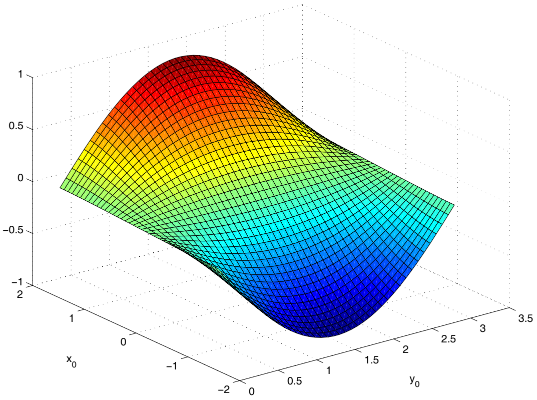

Figure 7.13: $\hat{\sigma}_{2,2}$; $\alpha = -0.75$

Note that all approximations satisfy the required symmetries (see §7.1).

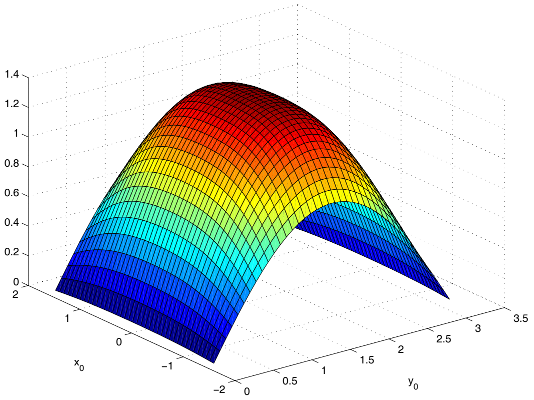

Figure 7.14: $\hat{\sigma}_{1,4}$; $\alpha = -0.5$

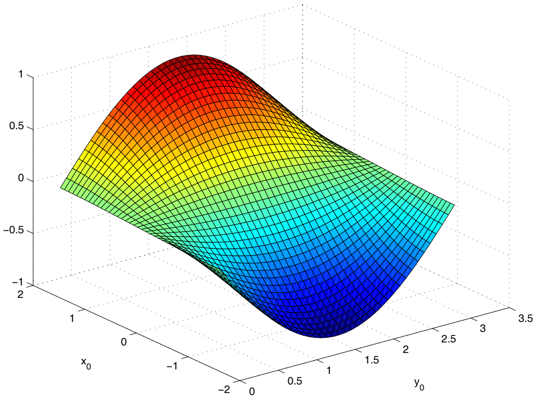

Figure 7.15: $\hat{\sigma}_{2,4}$; $\alpha = -0.5$

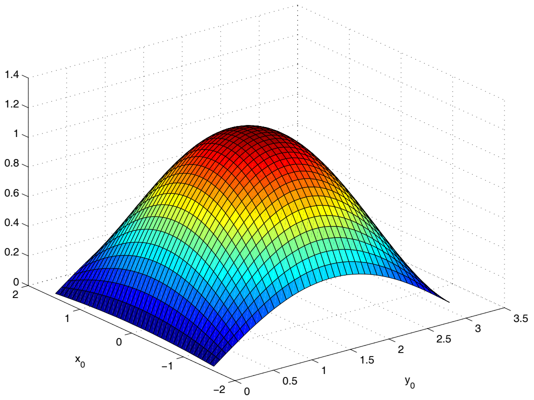

Figure 7.16: $\hat{\sigma}_{1,3}$; $\alpha = -0.25$

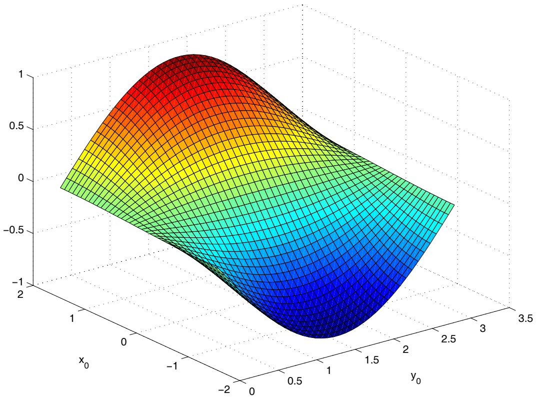

Figure 7.17: $\hat{\sigma}_{2,3}$; $\alpha = -0.25$

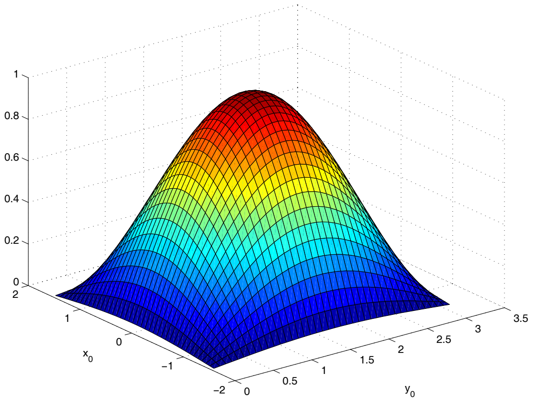

Figure 7.18: $\hat{\sigma}_{1,0}$; $\alpha = 0$

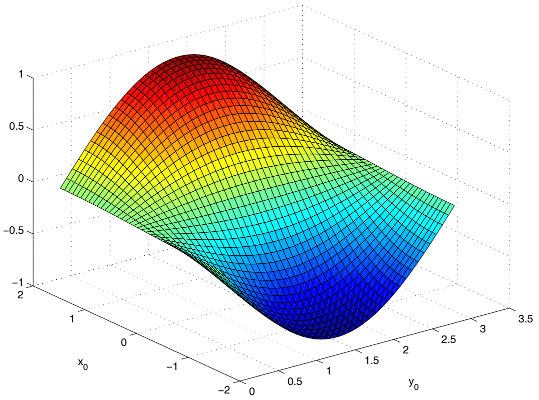

Figure 7.19: $\hat{\sigma}_{2,0}$; $\alpha = 0$

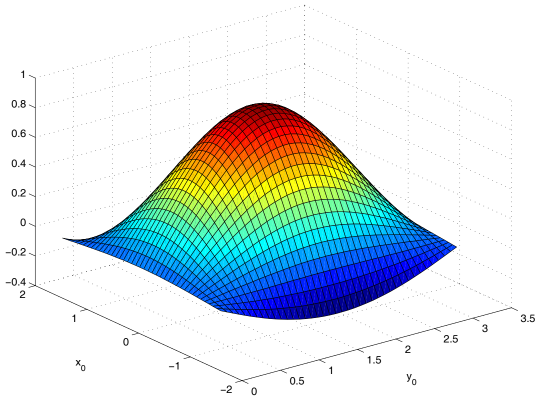

Figure 7.20: $\hat{\sigma}_{1,5}$; $\alpha = 0.25$

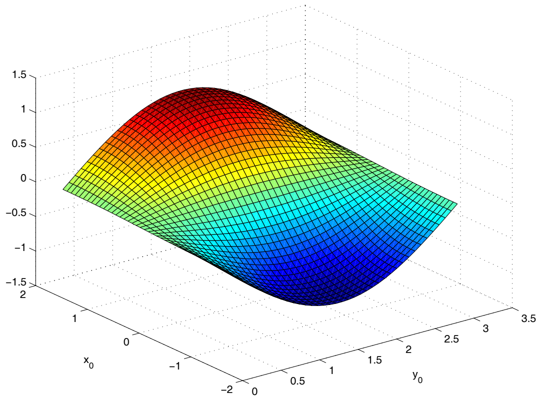

Figure 7.21: $\hat{\sigma}_{2,5}$; $\alpha = 0.25$

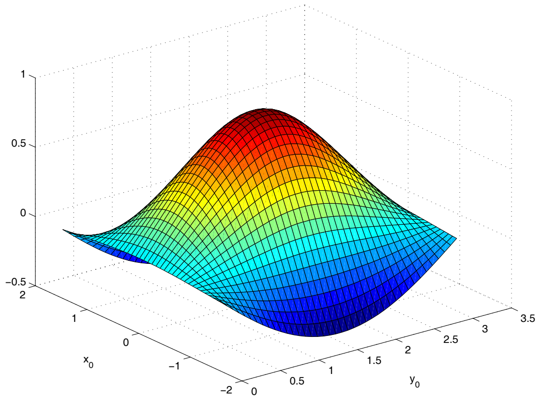

Figure 7.22: $\hat{\sigma}_{1,5}$; $\alpha = 0.5$

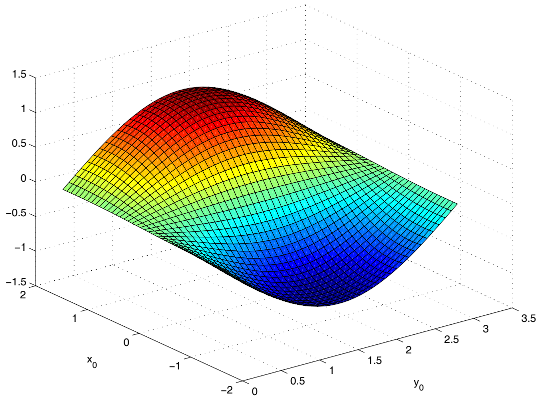

Figure 7.23: $\hat{\sigma}_{2,5}$; $\alpha = 0.5$

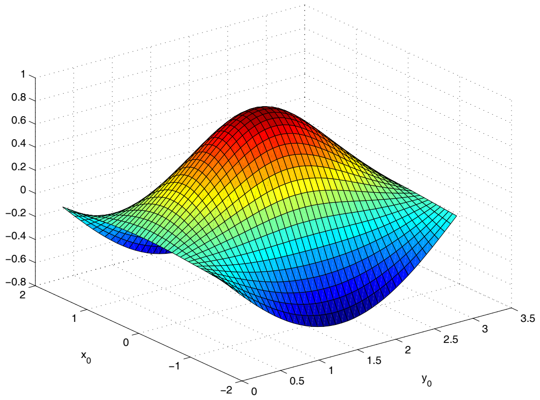

Figure 7.24: $\hat{\sigma}_{1,5}$; $\alpha = 0.75$

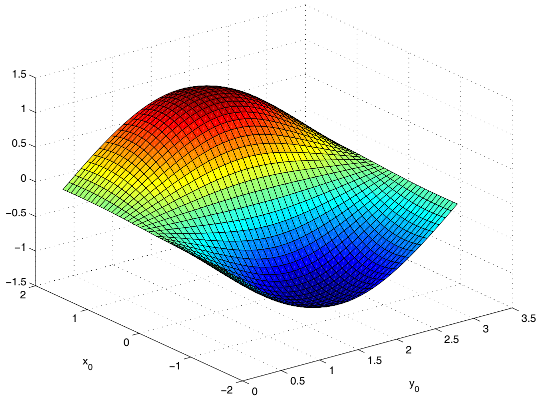

Figure 7.25: $\hat{\sigma}_{2,5}$; $\alpha = 0.75$

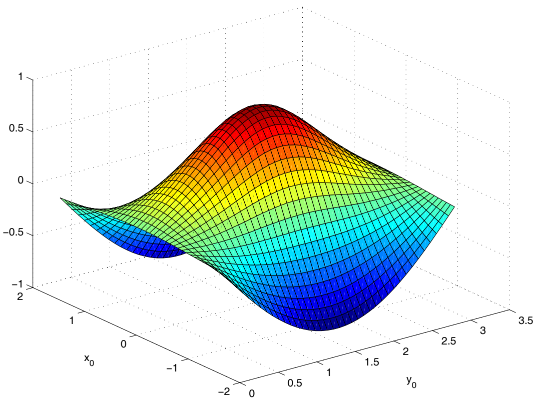

Figure 7.26: $\hat{\sigma}_{1,3}$; $\alpha = 1$

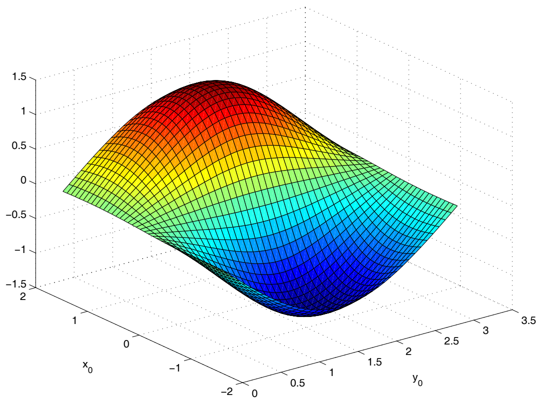

Figure 7.27: $\hat{\sigma}_{2,3}$; $\alpha = 1$

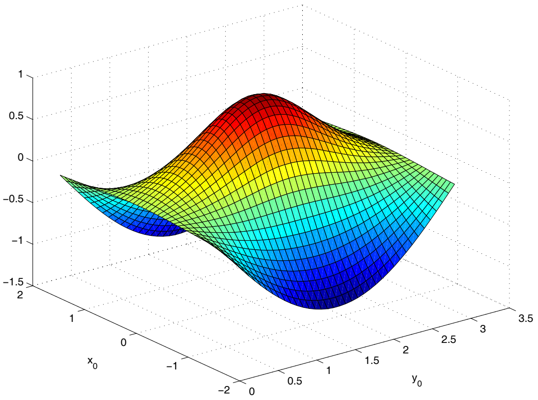

Figure 7.28: $\hat{\sigma}_{1,4}$; $\alpha = 1.25$

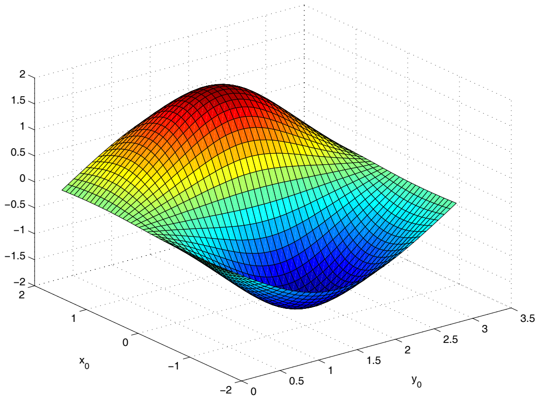

Figure 7.29: $\hat{\sigma}_{2,4}$; $\alpha = 1.25$

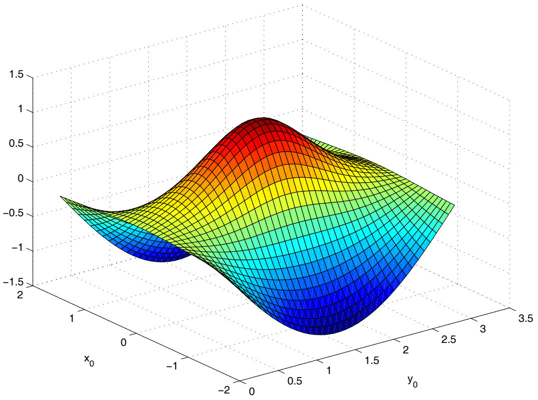

Figure 7.30: $\hat{\sigma}_{1,4}$; $\alpha = 1.5$

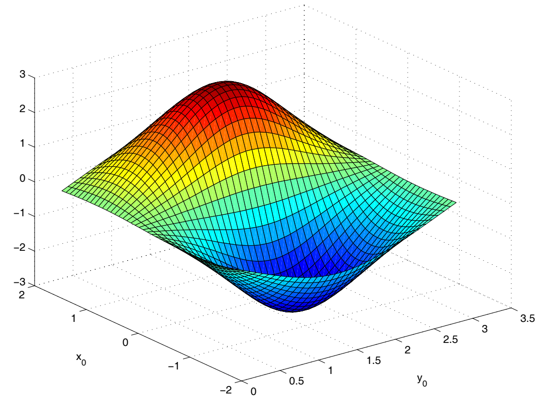

Figure 7.31: $\hat{\sigma}_{2,4}$; $\alpha = 1.5$

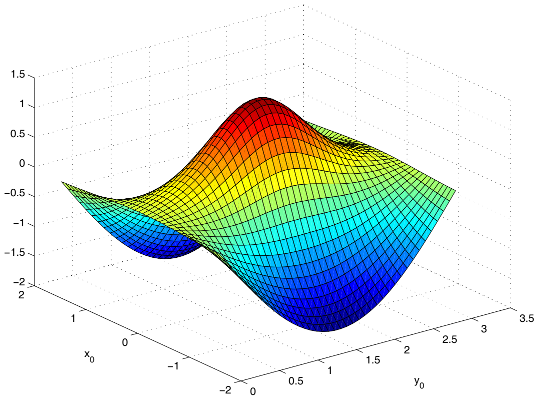

Figure 7.32: $\hat{\sigma}_{1,4}$; $\alpha = 1.75$

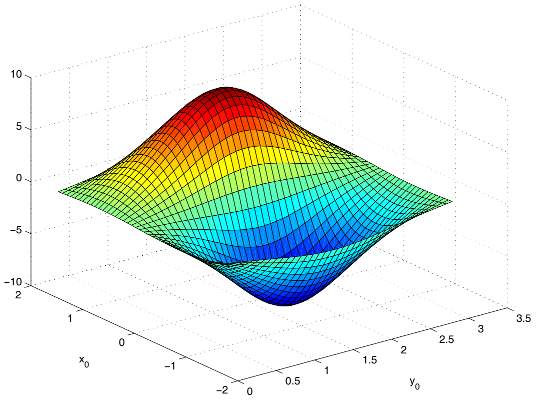

Figure 7.33: $\hat{\sigma}_{1,4}$; $\alpha = 1.75$

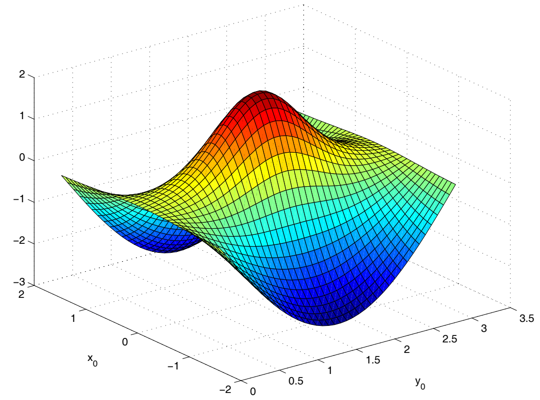

Figure 7.34: $\hat{\sigma}_{1,3}$; $\alpha = 2$

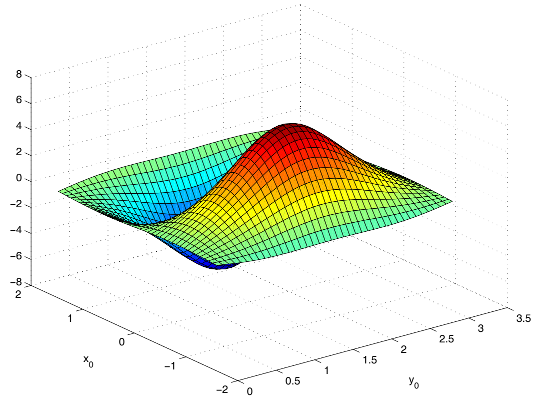

Figure 7.35: $\hat{\sigma}_{2,3}$; $\alpha = 2$

Each pair of approximations may be linearly combined, to give an approximation to the velocity potential, ϕ. The co-efficients needed to make these combinations, have already been calculated in making the approximations to $|R|^2$ and $|T|^2$ (see §6.2). This confirms the earlier assertion that an approximation to ϕ would come as a by-product of our goal to approximate $|R|^2$ and $|T|^2$.

The real and imaginary parts of ϕ are plotted side by side in figures 7.36-7.61.

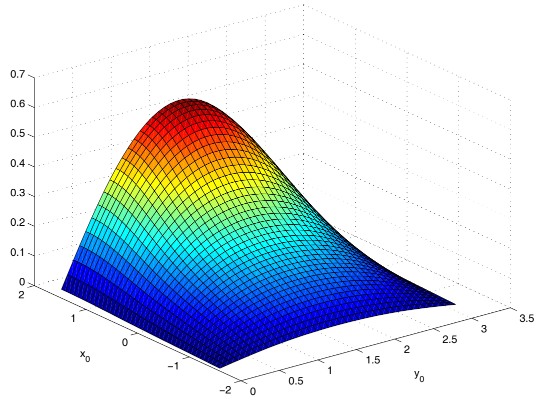

Figure 7.36: ϕ_{re}; $\alpha = -1$

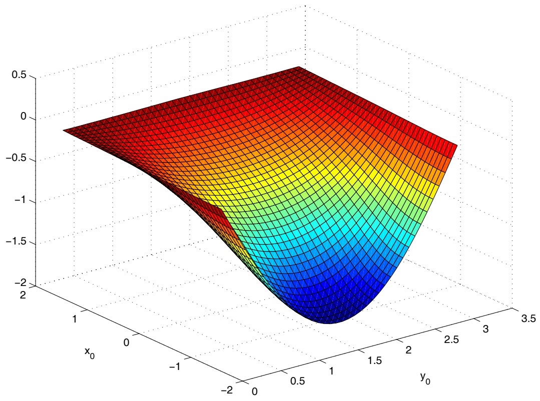

Figure 7.37: ϕ_{im}; $\alpha = -1$

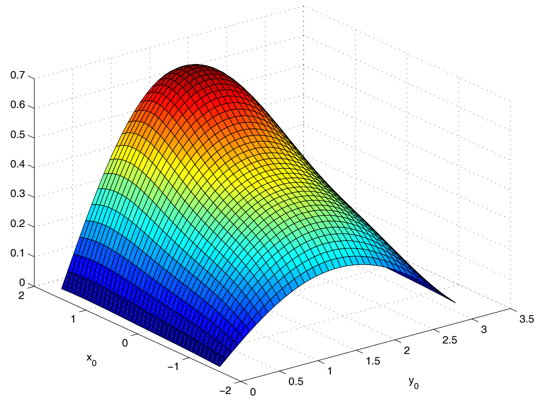

Figure 7.38: ϕ_{re}; $\alpha = -0.75$

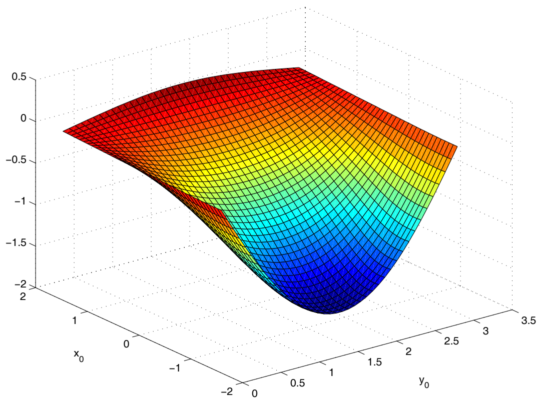

Figure 7.39: ϕ_{im}; $\alpha = -0.75$

Both the real and imaginary parts of ϕ contribute to Φ, the function defining the wave profile (see (1.3)). We shall not attempt any further reconstruction of the original acoustics problem. Our work gives a suitable base from which further investigation of the physical problem may be attempted.

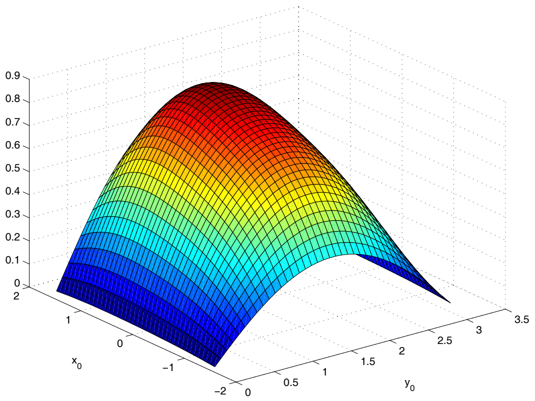

Figure 7.40: ϕ_{re}; $\alpha = -0.5$

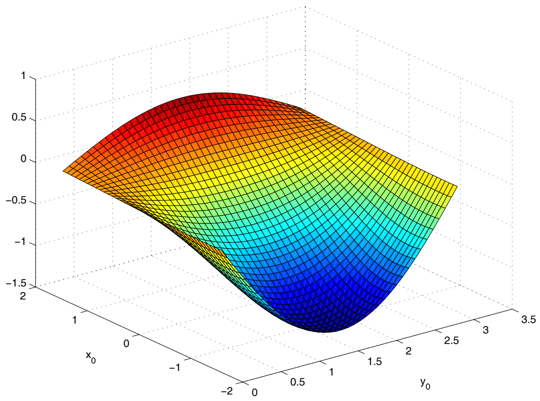

Figure 7.41: ϕ_{im}; $\alpha = -0.5$

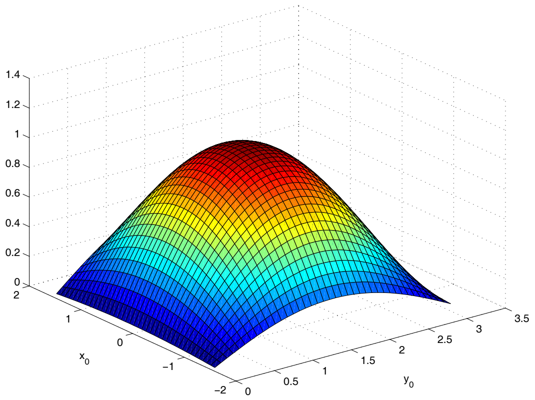

Figure 7.42: ϕ_{re}; $\alpha = -0.25$

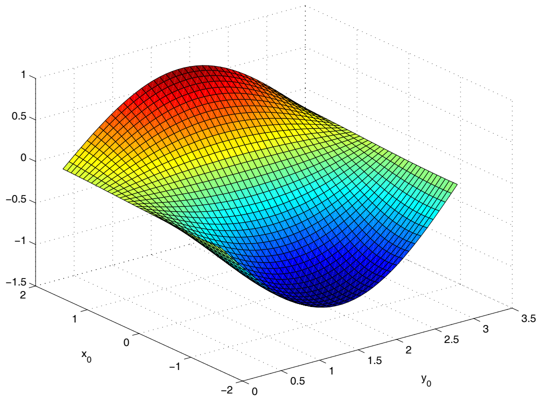

Figure 7.43: ϕ_{im}; $\alpha = -0.25$

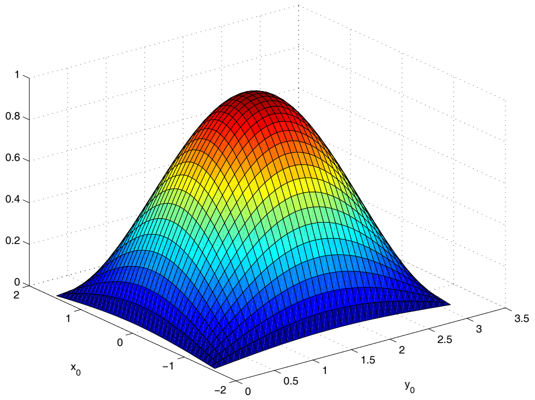

Figure 7.44: ϕ_{re}; $\alpha = 0$

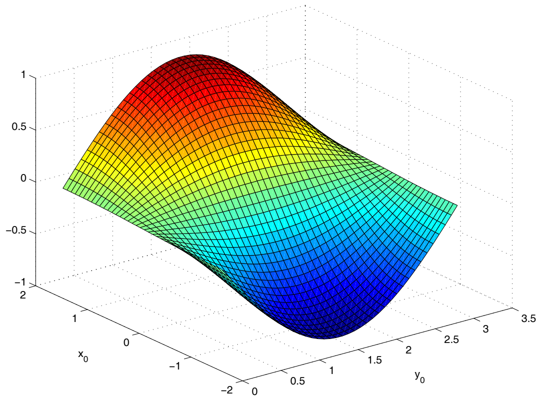

Figure 7.45: ϕ_{im}; $\alpha = 0$

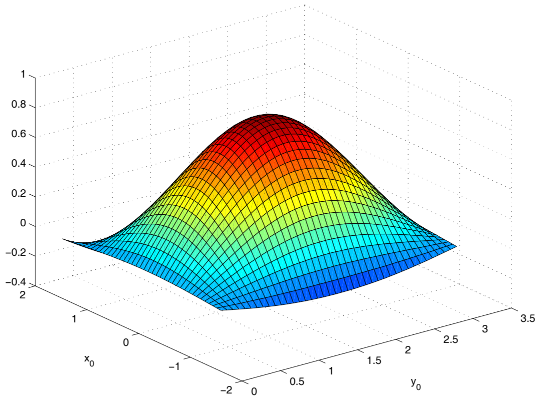

Figure 7.46: ϕ_{re}; $\alpha = 0.25$

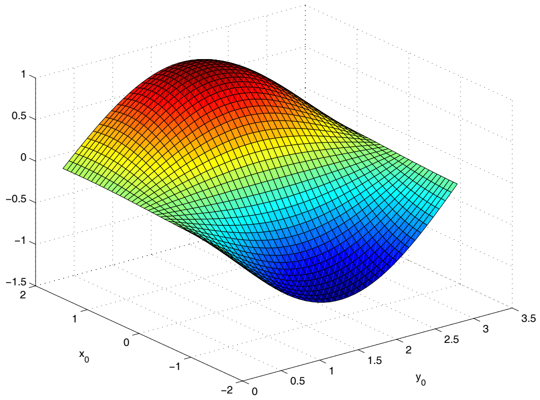

Figure 7.47: ϕ_{im}; $\alpha = 0.25$

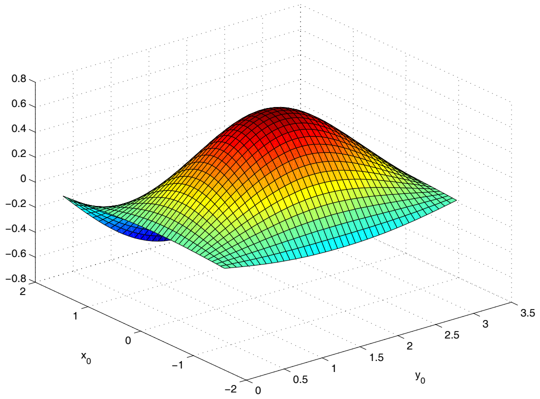

Figure 7.48: ϕ_{re}; $\alpha = 0.5$

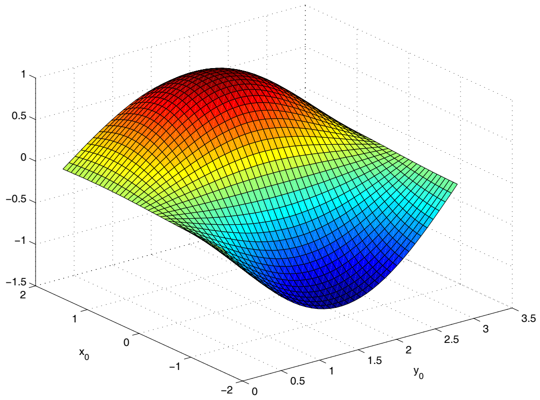

Figure 7.49: ϕ_{im}; $\alpha = 0.5$

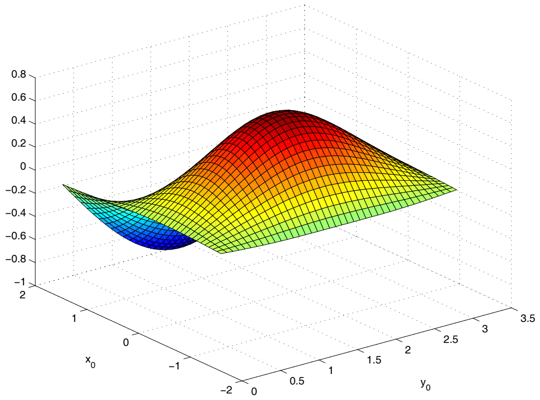

Figure 7.50: ϕ_{re}; $\alpha = 0.75$

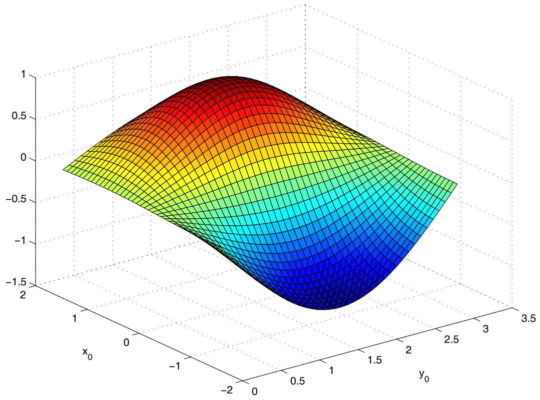

Figure 7.51: ϕ_{im}; $\alpha = 0.75$

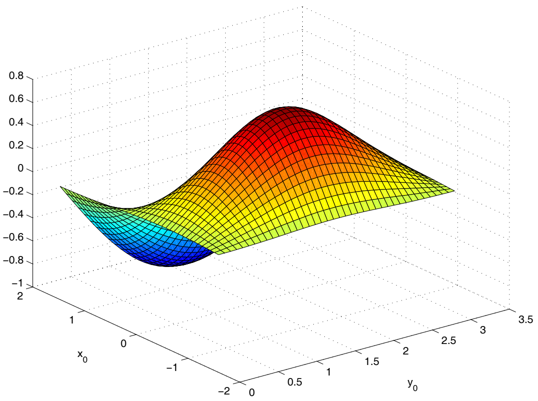

Figure 7.52: ϕ_{re}; $\alpha = 1$

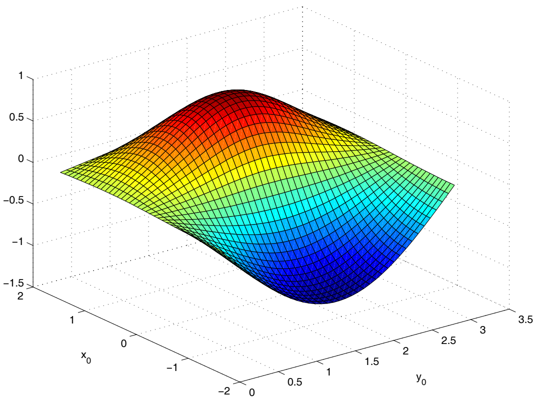

Figure 7.53: ϕ_{im}; $\alpha = 1$

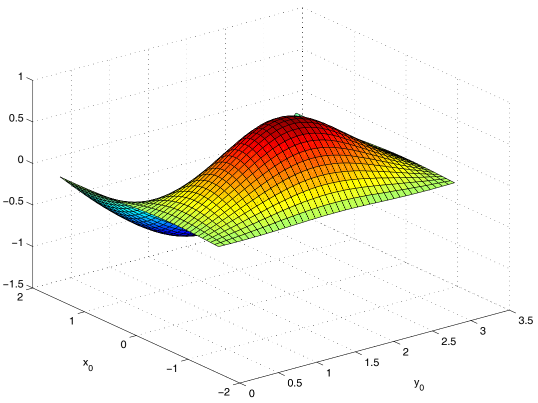

Figure 7.54: ϕ_{re}; $\alpha = 1.25$

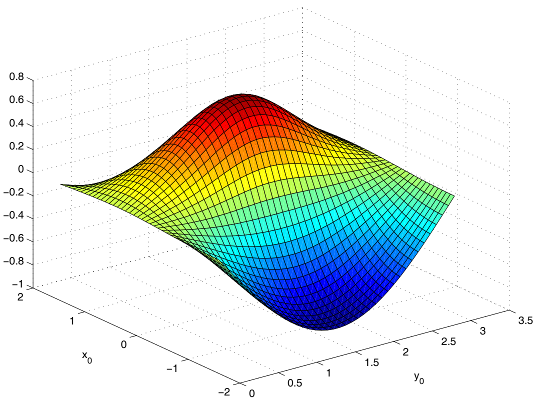

Figure 7.55: ϕ_{im}; $\alpha = 1.25$

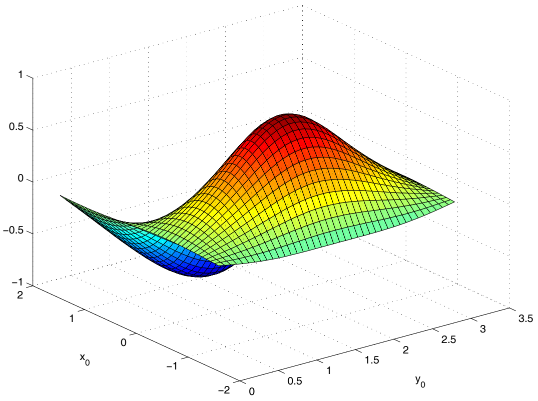

Figure 7.56: ϕ_{re}; $\alpha = 1.5$

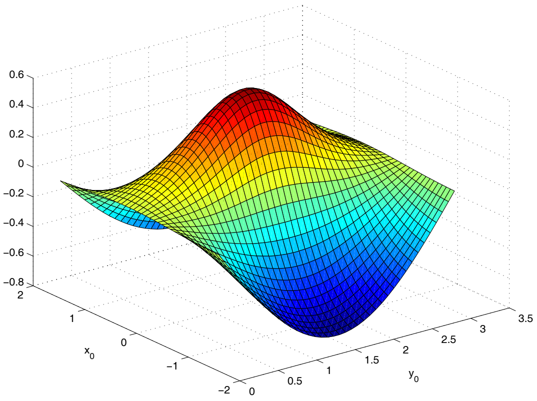

Figure 7.57: ϕ_{im}; $\alpha = 1.5$

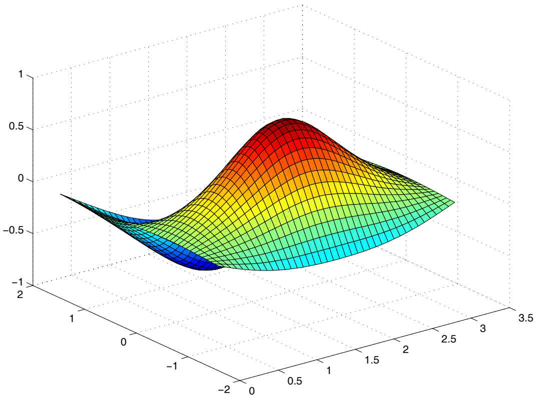

Figure 7.58: ϕ_{re}; $\alpha = 1.75$

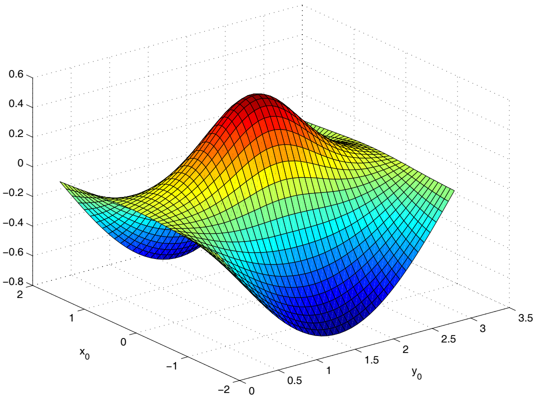

Figure 7.59: ϕ_{im}; $\alpha = 1.75$

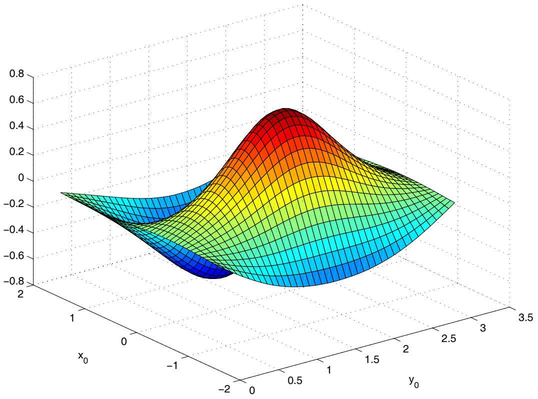

Figure 7.60: ϕ_{re}; $\alpha = 2$

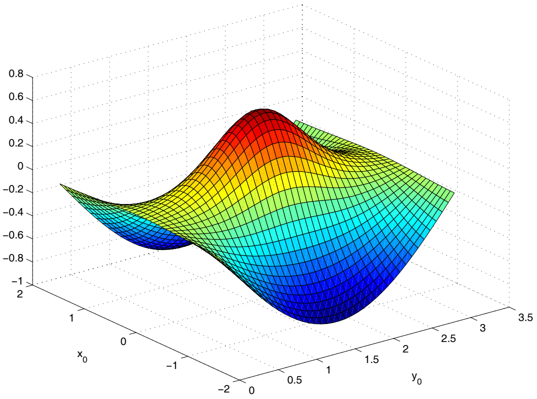

Figure 7.61: ϕ_{im}; $\alpha = 2$

In order to test whether the re-iterated Galerkin method, and the program that performs it, are robust, they are applied to a variation on the problem that has thus far been investigated in this chapter.

7.7 Parabolic x dependence

By making the subtle change

$$k(x,y) = k_0 + \frac{1}{\pi}\big((\frac{\pi}{m})^2 - x^2\big)\sin^2(y), \qquad (x,y) \in D' \tag{7.20}$$

we introduce a new property to the problem, in that the wave speed's x dependence, in the obstacle, is inversely parabolic, rather than inversely trigonometric. This change does little to the overall size and shape of k. In particular, all symmetries present in the previous problem remain, and all preparatory investigation into trial spaces are preserved. As such, it could quite reasonably be presumed that the re-iterated Galerkin method should work in a very similar fashion on this problem as it did on the previous.

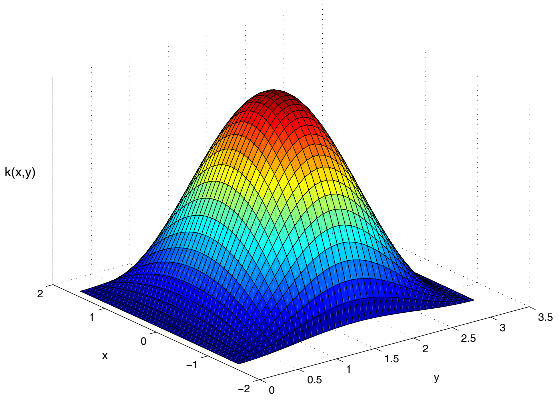

Figure 7.62: The new function $k(x,y)$ over D'

We are also presented with the opportunity to compare how a different wave speed variation effects the energy transference. Although it was just said that the new function $k(x,y)$ resembles its predecessor, there is a marked difference. At the points $x = \pm\frac{\pi}{m}$ (the lateral boundary of the obstacle D') the function $k(x,y)$, defined by (7.20) and k_0 in $D\backslash D'$, has a discontinuous derivative. It could be speculated that these discontinuities will prevent a gradual modulation of the wave, possible in the trigonometric case, and therefore increase the amount of reflected energy.

Similar experiments to those carried out on the previous problem were made here, with $m = 2$ and $k_0 = \sqrt{2}$ fixed. The behaviour of the re-iterated Galerkin method and variations of the errors were consistent with previous findings. To avoid appearing overly repetitious, only a sample of results is presented.

Taking $\alpha = 3$, tables 7.22-7.24 show how the re-iterated Galerkin method performs on the parabolic problem with varying subspaces.

As always, we begin by satisfying ourselves that the re-iterated Galerkin method is required, by applying Sloan iteration to the problem. Table 7.21 contains the results of this test. It is obvious that both problems are diverging, and we gain the approximations

$$\rho_{E(D')}(\tilde{\mathcal{G}}M) \approx 2.788,$$

and

$$\rho_{O(D')}(\tilde{\mathcal{G}}M) \approx 1.156,$$

for $\alpha = 2$. These are suitably large operators to provide a good test of the re-iterated Galerkin method.

Table 7.21: Sloan iteration

$\alpha = 2$; $\mathcal{S}_N = \mathcal{S}_3$; refinement: 40×40
Trial space 1 : (7.7); Trial space 2: (7.8)

Iterate	$\|\check{r}_{1,n}\|$	$\|\check{r}_{2,n}\|$	$\frac{\|\check{r}_{1,n}\|}{\|\check{r}_{1,n-1}\|}$	$\frac{\|\check{r}_{2,n}\|}{\|\check{r}_{2,n-1}\|}$
0	0.8654	1.503		
1	1.925	1.388	2.224	0.9235
2	3.930	1.612	2.024	1.161
3	11.87	1.863	3.022	1.156
4	32.20	2.153	2.712	1.156
5	90.44	2.488	2.809	1.156
6	251.3	2.875	2.779	1.156
7	700.7	3.322	2.788	1.156
8	1.951×10^3	3.839	2.785	1.156
9	5.436×10^3	4.436	2.786	1.156
10	1.514×10^4	5.126	2.786	1.156

Tables 7.22-7.24 contain the results of the application of re-iterated Galerkin to the parabolic problem, with $\alpha = 2$.

All observations on the re-iterated Galerkin method are present in tables 7.22-7.24. The re-iterated Galerkin method appears to perform better on the parabolic problem, in this case. Convergence is achieved in both problems, for all trial spaces used, despite the large size of the operators. Compare this situation with the case $\alpha = 2$, for the trigonometric x dependence.

To view the effect, the change to the parabolic problem, has on the transference of energies, the re-iterated Galerkin approximations to $|R|^2$ and $|T|^2$, are recorded over the interval $\alpha \in [-1, 2]$ in the same way as the previous problem. This will also allow us to test our hypothesis, that the discontinuous derivative will produce more reflection. Again, care was taken to try to ensure that results are accurate to 3 decimal places, and, as before, a refinement of 40×40 and series $\mathcal{S}_N = \mathcal{S}_5$, were found to be sufficient.

Figure 7.63 contains a graph of the approximations to $|R|^2$ and $|T|^2$, against α, calculated using the re-iterated Galerkin approximations, for the parabolic problem. Interestingly, there is almost total transmission over the interval $\alpha \in [0, 0.75]$.

Table 7.22: Re-iterated Galerkin

$\alpha = 2$; $\mathcal{S}_N = \mathcal{S}_3$; refinement: 40 × 40
Trial space 1 : (7.9); Trial space 2: (7.10)

Re-iterate	$\|\|\hat{r}_{1,n}\|\|$	$\|\|\hat{r}_{2,n}\|\|$	$\frac{\|\|\hat{r}_{1,n}\|\|}{\|\|\hat{r}_{1,n-1}\|\|}$	$\frac{\|\|\hat{r}_{2,n}\|\|}{\|\|\hat{r}_{2,n-1}\|\|}$	$\|R\|^2$	$\|T\|^2$
Galerkin	0.8563	2.688				
0	1.576	1.532		0.2831	0.7169	
1	1.131	1.017	0.7175	0.6638	0.1179	0.8821
2	0.9550	0.6733	0.8446	0.6623	0.3395	0.6605
⋮						
9	0.3544	0.0374	0.8688	0.6617	0.1435	0.8565.
10	0.3079	0.0247	0.8688	0.6617	0.1412	0.8588

Table 7.23: Re-iterated Galerkin

$\alpha = 2$; $\mathcal{S}_N = \mathcal{S}_3$; refinement: 40 × 40
Trial space 1 : (7.11); Trial space 2: (7.12)

Re-iterate	$\|\|\hat{r}_{1,n}\|\|$	$\|\|\hat{r}_{2,n}\|\|$	$\frac{\|\|\hat{r}_{1,n}\|\|}{\|\|\hat{r}_{1,n-1}\|\|}$	$\frac{\|\|\hat{r}_{2,n}\|\|}{\|\|\hat{r}_{2,n-1}\|\|}$	$\|R\|^2$	$\|T\|^2$
Galerkin	0.3743	1.462				
0	0.3886	0.6453			0.2195	0.7805
1	0.1914	0.1825	0.4925	0.2829	0.1422	0.8578
2	0.1020	0.0547	0.5328	0.2995	0.1362	0.8638
3	5.550×10^{-2}	0.0161	0.5396	0.2940	0.1351	0.8649
⋮						
8	2.600×10^{-3}	3.618×10^{-5}	0.5443	0.2954	0.1347	0.8653
9	1.400×10^{-3}	1.069×10^{-5}	0.5444	0.2954	0.1347	0.8653
⋮						
14	2.015×10^{-5}	1.776×10^{-8}	0.5445	1.867	0.1347	0.8653
15	1.097×10^{-5}	4.913×10^{-8}	0.5445	2.767	0.1347	0.8653
⋮						
25	8.474×10^{-8}	1.782×10^{-4}	0.5445	2.786	0.1347	0.8653

Table 7.24: Re-iterated Galerkin

$\alpha = 2$; $\mathcal{S}_N = \mathcal{S}_3$; refinement: 40 × 40
Trial space 1 : (7.14); Trial space 2: (7.15)

Re-iterate	$\|\hat{r}_{1,n}\|$	$\|\hat{r}_{2,n}\|$	$\frac{\|\hat{r}_{1,n}\|}{\|\hat{r}_{1,n-1}\|}$	$\frac{\|\hat{r}_{2,n}\|}{\|\hat{r}_{2,n-1}\|}$	$\|R\|^2$	$\|T\|^2$
Galerkin	0.3575	0.6565				
0	0.2443	0.0556			0.1494	0.8506
1	0.1091	2.900×10^{-3}	0.4464	5.220×10^{-2}	0.1384	0.8616
2	5.550×10^{-2}	2.566×10^{-4}	0.5092	8.850×10^{-2}	0.1357	0.8643
3	2.880×10^{-2}	2.737×10^{-5}	0.5186	0.1067	0.1349	0.8651
⋮						
6	4.100×10^{-3}	5.604×10^{-8}	0.5222	0.1343	0.1347	0.8653
⋮						
10	3.045×10^{-4}	4.949×10^{-11}	0.5224	0.3261	0.1347	0.8653
11	1.590×10^{-4}	1.247×10^{-10}	0.5224	2.519	0.1347	0.8653
⋮						
15	1.184×10^{-5}	7.507×10^{-9}	0.5224	2.786	0.1347	0.8653

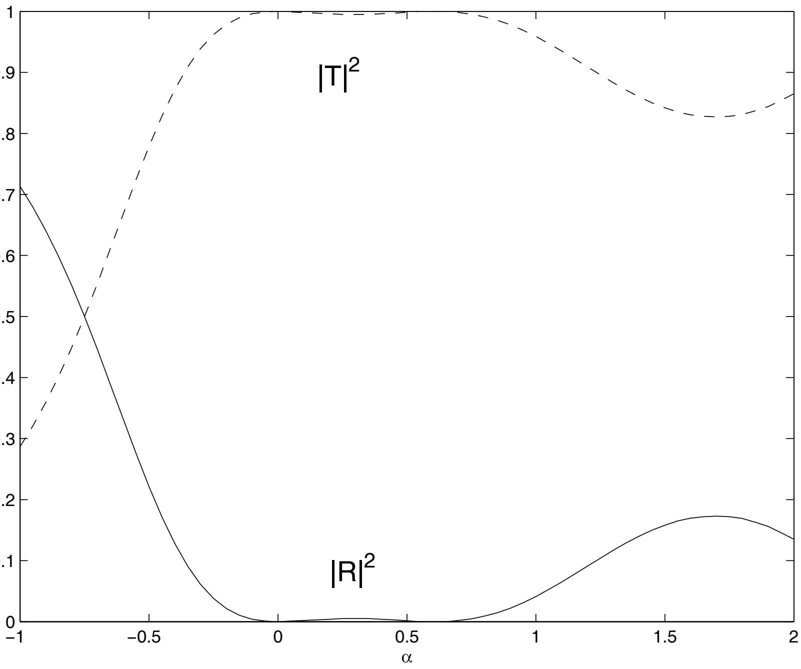

Figure 7.63: Energy proportion approximations against α

Figure 7.64 plots the approximations to $|R|^2$ from both the problem in which the x dependence of $k(x, y)$ was trigonometric, against the current problem, in which the dependence is parabolic. This is done specifically to test the hypothesis.

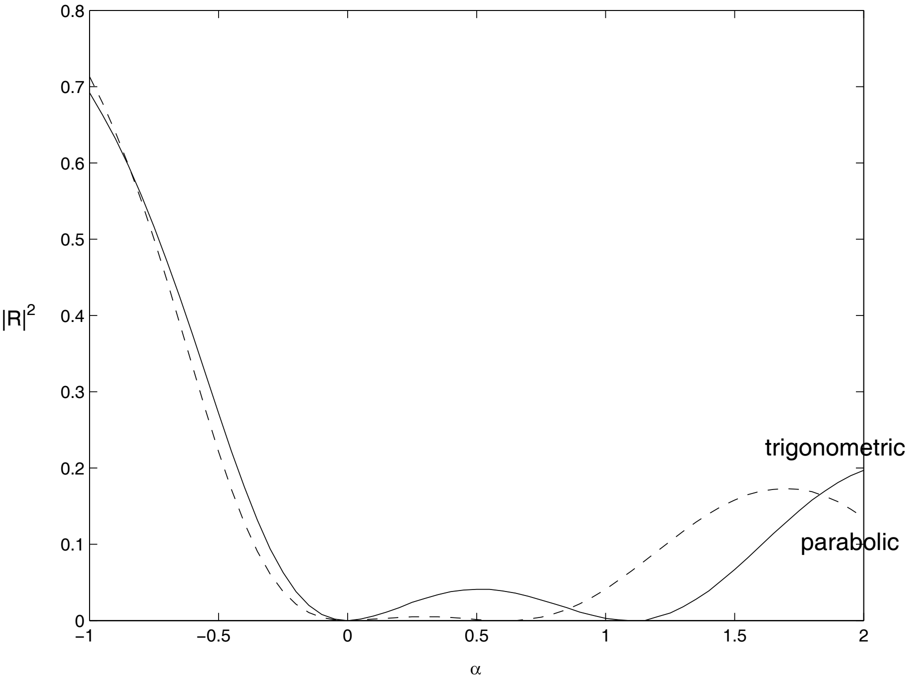

Figure 7.64: Comparison of $|R|^2$ between parabolic and trigonometric problems

Over the chosen interval, the two approximations, to $|R|^2$, do not differ vastly in size. The values of $|R|^2$ are particularly close for negative α. The average values of $|R|^2$ are: 0.1383 in the trigonometric case and 0.1429 in the parabolic case. This further demonstrates little difference between the two problems. Therefore, we have collected no evidence to support the hypothesis that a discontinuous derivative would cause more reflection.

Chapter 8

Conclusions

The body of our work has been in familiarising ourselves with the ways a particular aspect of approximation theory has been merged and influenced by results of functional analysis and spectral theory. It has been shown how this coalescence has brought about the Galerkin, iterated Galerkin (of Sloan) and re-iterated Galerkin (of Porter and Stirling) methods of approximating the solution of integral equations, and the related variational principles that can be used to approximate quantities that involve these solutions. In particular, the superiority of the re-iterated Galerkin approximation was expounded, and it was this approximation that we used in practice. The idealised acoustics problem, of §1, was introduced essentially as a platform on which to parade the re-iterated Galerkin approximation. As has already been noted, with a less stringent time restriction, this work could have induced an investigation with more physical conclusions. We attempted one, rather unsuccessful, investigation with physical implications, by asking whether a sharp change in wave speed would cause more reflection than allowing a gentle introduction to changing wave speed. The results of this investigation proved inconclusive but leaves an opportunity for a more expansive investigation. In seeking a function that satisfies the wave equation we have also gained the benefit of witnessing how a Green's function may be utilised in converting a differential equation into the integral equation used in our work. Many avenues were passed in making this transition, for example finite difference and finite element approximations. There is obviously a variety of possibilities for comparison of the re-iterated Galerkin method with one or more of these alternatives.

The latter chapters concentrated on the practical aspects of constructing an approximation. During this period, the acoustics problem, and hence integral equation, was refined to the point in which only a small set of parameters were available for variation. It was the aim of this work to show that the re-iterated Galerkin method is applicable to two-dimensional problems. This aim became somewhat lost at stages due to the presence of the acoustics problem. This added the limitation, that we only worked with a specific kernel which required much attention and rather consumed our attention. A full test of the re-iterated Galerkin method on two-dimensional would require consideration of far more varied problems. With a less stringent time restriction, we could have experimented more comprehensively with the multiplication operator M, introducing variations, such as asymmetries. It must, however, be noted that the kernel used has many strengths as a test for the re-iterated Galerkin method. It is singular, non-separable, involves an infinite series and induces a non-self-adjoint integral operator. These are all properties that could have highlighted inabilities in the method.

Although much of the material of the later chapters has generality, it was specifically with the integral equation posed in §1 that we dealt. This integral equation was equated to an operator equation in the Hilbert space of Lebesgue integrable functions. Most of the issues raised were of a computational nature, with attention centring on how an efficient program could be constructed to implement the re-iterated Galerkin method for the acoustics problem. In §1 we introduced the imaginary dimension to the problem, and this meant that when writing a program the integral equation had to be split into two, in order that the program worked only with real numbers. The integral operator also required special attention due to the singular infinite series that appears in the kernel. We also touched on the errors that are introduced by writing a program to perform computations, and the effects on the accuracy of approximations made using the re-iterated Galerkin method. It was seen that the re-iterated Galerkin method is well suited to the effects of these errors, as it is sensitive to the error in the current approximation, and hence does not accumulate errors. Given more time and effort, this could have been proved in a rigorous manner.

The program was implemented using trial spaces motivated by a discussion of the symmetries of the two integral equations. This discussion also helped us refine the setting of the integral equations. To counter the lack of an analytic solution, the results were verified against a set of checks, although it must be noted that these checks cannot be deemed as absolute assurance of the validity of results. Tests were made on the re-iterated Galerkin method and the errors introduced by numerical quadrature and truncation of the infinite series, before the original integral equation was solved for a sample set, in which only one parameter was varied. Most interest was reserved for the results that indicated the behaviour of the re-iterated Galerkin method. In the main, the program produced results consistent with the theory of the re-iterated Galerkin method, with the ratio of the norm of successive residuals settling, and operators with larger norm values requiring larger trial spaces to produce convergence, or a faster rate of convergence. It was, however, interesting to note that an expansion of a trial space did not always produce an improved convergence. The eventual divergence phenomenon was an incongruence that made for discussion that raised awareness of a fallibility of using the program to perform computations, but ultimately provided evidence that the program was working as it should.

Let us finish by briefly outlining areas of possible further research.

In the discussion of the existence and uniqueness of a solution (see §5.2.1), the question of eigenvalues and eigenvectors was introduced. Although no evidence of eigenvalue behaviour was found in our investigation, there are almost certainly parameters for which an eigenvalue will be present. By this it is meant that our choice of parameters prevent the existence and uniqueness of a solution. An eigenvalue can appear from two different sources. It may be an eigenvalue of the original problem, in that the chosen parameters cause the operator $\mathcal{G}M$ to have a unit eigenvalue. Such eigenvalues correspond to the eigenfunctions of the operator $\mathcal{G}M$, which are known as 'normal modes' of oscillation. In this case, there are either infinitely many solutions of the original equation or none. This indicates a probable flaw in the assumptions made in constructing the mathematical model for these parameters, for example the assumption of periodic steady state conditions. Recall that, the model can only ever be an approximation as it is built on the *linearised* wave equation. The re-iterated Galerkin approximations can be expected to 'blow-up' at an eigenvalue. For 'nearby' parameters, although there will exist an unique solution, the eigenvalue can cause problems in its approximation. An eigenvalue was found and discussed by Burton [7], for a one-dimensional example. It is still possible to produce Galerkin and iterated Galerkin approximations at such an eigenvalue, but these approximations are spurious. Conversely, it may not be possible to produce a Galerkin approximation when there exists an unique solution. This is if the chosen parameters coincide with an eigenvalue of the Galerkin approximation. More specifically, for a chosen trial space, and set of parameters, it may not be possible to invert the matrix required to produce the Galerkin approximation. The eigenvalues of this matrix are approximations of the eigenvalues of the exact equation. By changing the trial space these approximations will move, and the problem disappear, unless the problem was an accurate indication of an eigenvalue of the exact equation. The subject of eigenvalues is an ongoing area of research.

In §7, frequent reference was made to the possibility of an investigation of the optimality of trial spaces. The choice of trial space is fundamental to the success or failure of an application of the re-iterated Galerkin method. For the problems with which we dealt, it was possible to achieve a fast rate of convergence from a fairly small trial spaces, chosen in a logical rather than knowingly optimal fashion. However, by simple changes in certain parameters, the spectral radius of the operator $\tilde{\mathcal{G}}M$ could be made large enough that a more prudent choice of trial spaces would be required to maintain the low computational cost that the re-iterated Galerkin method emphasises. The difference in the effect of the re-iterated Galerkin method on problems that involve operators, with different size spectral radii, was well documented in the results of §7. An extension in this area could involve comparing the truncated double Fourier series used here, against other trial spaces, for instance, Legendre polynomials, or it could be in deciding on a more optimal way of extending the truncated Fourier series, than the 'triangular' approach we have adopted. The orthogonality relations of the trigonometric functions used in the kernel as well as the trial functions, would undoubtedly have an influence over this optimality question. As mentioned before, to attempt this extension from an analytic viewpoint could prove highly intricate, and would most likely involve the approximation of eigenvalues. It would, therefore, be more appropriate to tackle this issue as a practical investigation.

Appendix A

The MatLab code used to apply the re-iterated Galerkin method.

Functions are represented by finite dimensional vectors that hold point values taken at the nodes of a regular cartesian mesh. This structure was outlined in §6. The functions that appear are contained in appendices C-F.

```
clear;

global X2d
global rect
global alpha
global m
global k_0
global beta_0
global X_vals
global f_5
global basis1
global basis2
global mx_term
```

comment – Define the parameters here ...

```
alpha = 1.25;
k_0 = sqrt(2);
beta_0 = sqrt((k_0∧2)-1);
m = 2;
```

comment – This is number of terms used in the series, i.e. $S_N = S_{\texttt{mx\_term}}$.

```
mx_term = 5;
```

comment – `STORE` is used to store the required output and does not affect the approximation.

```
STORE = zeros(1,9);
STORE(1,1) = alpha;
STORE(1,2) = k_0;
STORE(1,3) = m;
STORE(1,4:8) = 0;
```

comment – Define the limits of the rectangular region D'...

```
A = -pi/m;
B = -A;
C = 0;
D = pi;
```

comment – xp=X_p and yp=Y_p.

```
xp = 40;
yp = 40;
```

comment – X_vals contains the values x_i. Likewise Y_vals contains the values y_i.

```
X_vals = zeros(xp,1);
for loop=1:xp
   X_vals(loop)=((2*loop -1)*(B-A)/(2*xp)) + A;
end

Y_vals = zeros(yp,1);
for loop=1:yp
   Y_vals(loop)=((2*loop -1)*(D-C)/(2*yp)) + C;
end
```

comment – X2d contains the points $\boldsymbol{x}_i$.

```
X2d = zeros(xp*yp,2);
for loop=1:yp
   X2d(xp*(loop-1) +1:xp*loop,1)=X_vals(:);
   X2d(xp*(loop-1) +1:xp*loop,2)=Y_vals(loop);
end
```

comment – rect=Q.

```
rect = ((B-A)/xp)*((D-C)/yp);

free_term(:,1) = cos(beta_0*X2d(:,1)).*sin(X2d(:,2));
free_term(:,2) = sin(beta_0*X2d(:,1)).*sin(X2d(:,2));
```

comment – mesh is the number of nodes in the rectangular mesh.

```
mesh = length(X2d(:,1));
```

comment – basis1 contains vectors defining the trial functions for problem 1.

```
basis1 = zeros(mesh,1);
```

```
basis1(:,1) = sin(X2d(:,2));
basis1(:,2) = cos(m*X2d(:,1)).*sin(X2d(:,2));
basis1(:,3) = sin(3*X2d(:,2));
```

comment – `basis2` contains vectors defining the trial functions for problem 2.

```
basis2 = zeros(mesh,1);
basis2(:,1) = sin(m*X2d(:,1)).*sin(X2d(:,2));
basis2(:,2) = sin(2*m*X2d(:,1)).*sin(X2d(:,2));
basis2(:,3) = sin(m*X2d(:,1)).*sin(3*X2d(:,2));
```

comment – Calculate the dimensions of the two trial spaces . . .

```
dim1 = length(basis1(1,:));
dim2 = length(basis2(1,:));
```

comment – f_5 = $A_{X,Y}$.

```
f_5 = log_5(X2d(:,1));
```

comment – The `G_basis` matrices contain vectors defining the trial functions having been operated on by $\tilde{\mathcal{G}}M$.

```
for loop=1:dim1
   G_basis1(:,loop) = tilde_G_modif(basis1(:,loop));
end

for loop=1:dim1
   A_basis1(:,loop) = basis1(:,loop) - G_basis1(:,loop);
end
```

comment – `matrix1` and `matrix2` are the matrices used to calculate the Galerkin approximations.

```
for loop = 1:dim1
for loop2 = 1:dim1
   matrix1(loop,loop2) = A_basis1(:,loop2)'*basis1(:,loop);
end
end

matrix1 = rect*matrix1;

for loop=1:dim2
   G_basis2(:,loop) = tilde_G_modif(basis2(:,loop));
end

for loop=1:dim2
```

```
   A_basis2(:,loop) = basis2(:,loop) - G_basis2(:,loop);
end

for loop = 1:dim2
for loop2 = 1:dim2
   matrix2(loop,loop2) = A_basis2(:,loop2)'*basis2(:,loop);
end
end

matrix2 = rect*matrix2;
```

comment – Define the vectors of inner products of the free term with the trial functions.

```
F = basis_ip1(free_term(:,1));
G = basis_ip2(free_term(:,2));
```

comment – Store the inverse matrices to save computations.

```
inv_mat1 = inv(matrix1);
inv_mat2 = inv(matrix2);
```

comment – **The Galerkin approximation**.
comment – Solve for the unknown co-efficients.

```
c1 = inv_mat1*F;
c2 = inv_mat2*G;

clear F;
clear G;
```

comment – Use coefficients to define Galerkin approximations...

```
p1 = basis1*c1;
p2 = basis2*c2;
```

comment – Calculate the residual errors in the Galerkin approximations...

```
residual_1 = free_term(:,1) - p1 + G_basis1*c1;
residual_2 = free_term(:,2) - p2 + G_basis2*c2;

Galerkin_res_error1 = l2_norm(residual_1)
Galerkin_res_error2 = l2_norm(residual_2)

STORE(2,:)  = 0;
STORE(3,1) = 0;
STORE(3,2) = Galerkin_res_error1;
STORE(3,3) = Galerkin_res_error2;
STORE(3,4:9) = 0;

clear c1; clear c2;
```

comment – From now on the approximations are `sigma1` and `sigma2`.

```
sigma1 = zeros(mesh,1);
sigma2 = zeros(mesh,1);
```

comment – **The iterated Galerkin approximation**.
comment – Initially the approximations are the iterated Galerkin approximations.
comment – The following definition is computationally economical.

```
sigma1 = residual_1 + p1;
sigma2 = residual_2 + p2;

clear residual_1; clear residual_2;
residual_1 = free_term(:,1) - sigma1 + tilde_G_modif(sigma1);
residual_2 = free_term(:,2) - sigma2 + tilde_G_modif(sigma2);

iterate_error1 = l2_norm(residual_1)
iterate_error2 = l2_norm(residual_2)

STORE(4,1) = 0;
STORE(4,2) = iterate_error1;
STORE(4,3) = iterate_error2;
```

comment – **The energy approximations**.
comment – The `A_sigma` are defined so as to save on calculations. They are vectors that represent the operator $I - \tilde{\mathcal{G}}M$ on our current approximations.

```
A_sigma_1 = free_term(:,1) - residual_1;
A_sigma_2 = free_term(:,2) - residual_2;

L(1,1) = 2*inner_product(sigma1 , Mult(free_term(:,1))) - ...
    inner_product(A_sigma_1, Mult(sigma1));
L(2,2) = 2*inner_product(sigma2 , Mult(free_term(:,2))) - ...
    inner_product(A_sigma_2, Mult(sigma2));
L(1,2) = inner_product(free_term(:,1), Mult(sigma2)) + ...
    inner_product(sigma1 , Mult(free_term(:,2))) - ...
    inner_product(A_sigma_1 , Mult(sigma2));
```

comment – This equality comes from a result proved in check i.

```
L(2,1) = L(1,2);
```

```
clear A_sigma_1; clear A_sigma_2;
```

comment – This follows the theory of §6.2.

```
energy_mat = eye(4);
energy_mat(1:2,3:4) = (1/(beta_0*pi))*L;
energy_mat(3:4,1:2) = -(1/(beta_0*pi))*L;
energy_vec(1:2,1) = L(1:2,1);
energy_vec(3:4,1) = L(1:2,2);
energy_c = energy_mat*energy_vec;

R = (1/(beta_0*pi))*(i*(energy_c(1)-energy_c(4)) - (energy_c(2) + energy_c(3)));
T = 1 + (1/(beta_0*pi))*(i*(energy_c(1) + energy_c(4)) - (energy_c(3)-energy_c(2)));
```

comment – Check ii…

```
energy_cons = abs(R)∧2 + abs(T)∧2;

energy_R = abs(R)∧2
energy_T = abs(T)∧2

STORE(4,6) = R;
STORE(4,7) = T;
STORE(4,8) = energy_R;
STORE(4,9) = energy_T;
STORE(4,4:5) = 0;
clear energy_mat; clear energy_vec;
```

comment – Checks i & v.

```
check_1 = inner_product(sigma1 , Mult(free_term(:,2)));
check_2 = inner_product(sigma2 , Mult(free_term(:,1)));
check_error = l2_norm(check_1 - check_2);

re_iterate_res_error1 = iterate_error1;
re_iterate_res_error2 = iterate_error2;
clear iterate_error1; clear iterate_error2;
```

comment – **Re-iteration**.
comment – Can do any number of re-iterations, here 1000 is the chosen number.
comment – The code within the loop mimics the above code.

```
 for loop=1:1000
```

   comment – The `previous` are used in the spectral radius approximations.

```
   previous1 = re_iterate_res_error1;
```

```
previous2 = re_iterate_res_error2;

clear re_iterate_res_error1; clear re_iterate_res_error2;

F = basis_ip1(residual_1);
G = basis_ip2(residual_2);

c1 = inv_mat1*F;
c2 = inv_mat2*G;

sigma1 = sigma1 + residual_1 + G_basis1*c1;
sigma2 = sigma2 + residual_2 + G_basis2*c2;

clear residual_1; clear residual_2

residual_1 = free_term(:,1) - sigma1 + tilde_G_modif(sigma1);
residual_2 = free_term(:,2) - sigma2 + tilde_G_modif(sigma2);

iterate = loop

re_iterate_res_error1 = l2_norm(residual_1)
re_iterate_res_error2 = l2_norm(residual_2)

STORE(4+loop,1) = loop;
STORE(4+loop,2) = re_iterate_res_error1;
STORE(4+loop,3) = re_iterate_res_error2;
```

comment – Spectral radius approximations…

```
spectral_rad1 = re_iterate_res_error1 / previous1
spectral_rad2 = re_iterate_res_error2 / previous2

STORE(4+loop,4) = spectral_rad1;
STORE(4+loop,5) = spectral_rad2;

clear previous1; clear previous2;

A_sigma_1 = free_term(:,1) - residual_1;
A_sigma_2 = free_term(:,2) - residual_2;

clear L; clear energy_mat; clear energy_vec; clear energy_c;
L(1,1) = 2*inner_product(sigma1 , Mult(free_term(:,1))) - ...
 inner_product(A_sigma_1, Mult(sigma1));
L(2,2) = 2*inner_product(sigma2 , Mult(free_term(:,2))) - ...
 inner_product(A_sigma_2, Mult(sigma2));
L(1,2) = inner_product(free_term(:,1), Mult(sigma2)) + ...
 inner_product(sigma1 , Mult(free_term(:,2))) - ...
```

```
    inner_product(A_sigma_1 , Mult(sigma2));
   L(2,1) = L(1,2);

   energy_mat = eye(4);
   energy_mat(1:2,3:4) = (1/(beta_0*pi))*L;
   energy_mat(3:4,1:2) = -(1/(beta_0*pi))*L;
   energy_vec(1:2,1) = L(1:2,1);
   energy_vec(3:4,1) = L(1:2,2);

   clear energy_c;
   energy_c = energy_mat / energy_vec;

   R = (1/(beta_0*pi))*(i*(energy_c(1)-energy_c(4)) - (energy_c(2) + energy_c(3)));
   T = 1 + (1/(beta_0*pi))*(i*(energy_c(1) + energy_c(4)) ...
    - (energy_c(3)-energy_c(2)));

   energy_cons = abs(R)∧2 + abs(T)∧2;
   energy_R = abs(R)∧2
   energy_T = abs(T)∧2

   STORE(4+loop,6) = R;
   STORE(4+loop,7) = T;
   STORE(4+loop,8) = energy_R;
   STORE(4+loop,9) = energy_T;

   clear energy_mat; clear energy_vec;

   check_1 = inner_product(sigma1 , Mult(free_term(:,2)));
   check_2 = inner_product(sigma2 , Mult(free_term(:,1)));
   check_error = l2_norm(check_1 - check_2);
end
```

Appendix B

The MatLab code used to apply repeated Sloan iteration.

comment – Up until iteration this has exactly the same structure as appendix A.

```
clear;

global X2d
global rect
global alpha
global m
global k_0
global beta_0
global X_vals
global f_5
global basis
global mx_term

alpha = 2;
k_0 = sqrt(2);
beta_0 = sqrt((k_0∧2)-1);
m=2;
mx_term = 3;

STORE = zeros(1,3);
STORE(1,1) = alpha;
STORE(1,2) = k_0;
STORE(1,3) = m;

A = -pi/m;
B = -A;

C = 0;
D = pi;

xp = 40;
```

```
yp = 40;

X_vals = zeros(xp,1);
for loop=1:xp
   X_vals(loop)=((2*loop -1)*(B-A)/(2*xp)) + A;
end

Y_vals = zeros(yp,1);
for loop=1:yp
   Y_vals(loop)=((2*loop -1)*(D-C)/(2*yp)) + C;
end

X2d = zeros(xp*yp,2);
for loop=1:yp
   X2d(xp*(loop-1) +1:xp*loop,1)=X_vals(:);
   X2d(xp*(loop-1) +1:xp*loop,2)=Y_vals(loop);
end

rect = ((B-A)/xp)*((D-C)/yp);
```

comment – This program only works on one problem at a time.

comment – The appropriate free term and trial function are chosen by the user by deleting the unwanted code.

```
free_term = cos(beta_0*X2d(:,1)).*sin(X2d(:,2));
free_term = sin(beta_0*X2d(:,1)).*sin(X2d(:,2));

mesh = length(X2d(:,1));

basis = zeros(mesh,1);
basis(:,1) = X2d(:,1);

n = length(basis(1,:));

f_5 = log_5(X2d(:,1));
for loop=1:n
   G_basis(:,loop) = tilde_G_modif(basis(:,loop));
end

for loop=1:n
   A_basis(:,loop) = basis(:,loop) - G_basis(:,loop);
end

for loop = 1:n
for loop2 = 1:n
   matrix(loop,loop2) = A_basis(:,loop2)'*basis(:,loop);
```

```
end
end

matrix = rect*matrix;

F = basis_ip(free_term);

inv_mat = inv(matrix);

c = inv_mat*F;
p = basis*c;

residual = free_term - p + G_basis*c;
Galerkin_res_error = l2_norm(residual)

STORE(3,2) = Galerkin_res_error;

iterate_error = Galerkin_res_error;
clear Galerkin_res_error;
```

comment – **Repeated Sloan iteration**

```
for loop=1:1000
   clear previous;
   previous = iterate_error;

   clear iterate_error;
   p = residual + p;

   clear residual;
   residual = free_term - p + tilde_G_modif(p);

   iterate_error = l2_norm(residual)

   spectral_rad_est = iterate_error / previous

   done = loop

   STORE(3+loop,1) = loop;
   STORE(3+loop,2) = iterate_error;
   STORE(3+loop,3) = spectral_rad_est;
end
```

Appendix C

The multiplication operator M. The unwanted function is merely deleted.

```
function var = Mult(f)

global X2d
global alpha
global m
global k_0
```

comment – The trigonometric x dependence…

```
var = 2*k_0*alpha*(cos(m*X2d(:,1)/2).∧2).*((sin(X2d(:,2))).∧2 ).*f +...
    (alpha∧2)*(cos(m*X2d(:,1)/2).∧4).*((sin(X2d(:,2))).∧4).*f;
```

comment – The parabolic x dependence…

```
l = pi/m;
parab = (l∧2)-(X2d(:,1).∧2);
parab = parab/pi;

var = 2*k_0*alpha*parab.*((sin(X2d(:,2))).∧2).*f +...
    (alpha∧2)*(parab.∧2).*((sin(X2d(:,2))).∧4).*f;
```

Appendix D

The approximation of the operator $\tilde{\mathcal{G}}M$.

The modifications that led to the structure of $\tilde{\mathcal{G}}M$ used is outlined in §6.3. Integrals are approximated using the rectangular midpoint rule.

D.1

comment – The function that approximates $\tilde{\mathcal{G}}M$.
comment – It merely combines other functions.

```
function var = tilde_G_modif(f)

global f_5

var = -sine_kernel(f) + ...
    modified_series(f)+...
    (log_2(f)/4)-...
    (log_3(f)/4)-...
    (log_4(f)/4)-...
    (Mult(f).*f_5)/4;

var = var/pi;
```

D.2

comment – The part of the kernel defined as $\pi M(a_0)$.

```
function var = sine_kernel(f)

global X2d
global rect
global beta_0

g = Mult(f);
dim = length(f);
```

```
var = zeros(dim,1);

for loop=1:dim
   var = var + sin(beta_0*abs(X2d(loop,1)-X2d(:,1)))*...
    sin(X2d(loop,2))*g(loop);
end

var = var.*sin(X2d(:,2));
var = var*rect;
var = var/beta_0;
```

D.3

comment – The approximate series $(M\mathcal{S}_N)(x, y \mid x_0, y_0)$.

```
 function var = modified_series(f)

 global X2d
 global rect
 global k_0
 global mx_term

 g = Mult(f);
 dim = length(f);
 var = zeros(dim,1);
```

comment – The first term is done separately. . .

```
for loop = 1:dim
   var = var - exp(-abs(X2d(loop,1)-X2d(:,1))).*sin(X2d(loop,2))*g(loop);
end

var = rect*sin(X2d(:,2)).*var;

for loop2=2:mx_term
   int_term =zeros(dim,1);
   gamma_n = sqrt((loop2land2)-(k_0land2));
for loop = 1:dim
   int_term = int_term + g(loop)*sin(loop2*X2d(loop,2))*sin(loop2*X2d(:,2)).*...
     (loop2*exp(-gamma_n*abs(X2d(loop,1)-X2d(:,1)))-...
    gamma_n*exp(-loop2*abs(X2d(loop,1)-X2d(:,1))));
end
   int_term = int_term*rect;
   int_term = int_term/(gamma_n*loop2);
   var = var + int_term;
end
```

D.4

comment – The part of the kernel defined as $(M\mathcal{L}_1)(x, y \mid x_0, y_0)$.

```
function var = log_2(f);

global X2d
global rect

dim = length(f);
g = Mult(f);
var = zeros(dim,1);

 for loop=1:dim
   var = var + g(loop)*log( 1 - 2*cos(X2d(loop,2) + X2d(:,2)).*...
    exp(-abs(X2d(loop,1)-X2d(:,1))) +...
    exp(-2*abs(X2d(loop,1)-X2d(:,1))));
end

var = var*rect;
```

D.5

comment – The part of the kernel defined as $(M\mathcal{L}_4)(x, y \mid x_0, y_0)$.

```
function var = log_3(f)

global X2d
global rect

dim = length(f);
g = Mult(f);
var = zeros(dim,1);

for loop = 1:dim
```

comment – The vectors `Z` and `Xv` are used to artificially implement the limit (6.7).

```
   Z=zeros(dim,1);
   Z(loop)= 1;
   Xv=X2d;
   Xv(loop,1)= X2d(loop,1) - 1;

   var = var + g(loop)*log(Z + (( 1 - 2*cos(X2d(loop,2)...
    - X2d(:,2)).*exp(-abs(X2d(loop,1)-X2d(:,1))) +...
    exp(-2*abs(X2d(loop,1)-X2d(:,1)))) ./(((X2d(loop,1)-Xv(:,1))∧2)...
    +((X2d(loop,2)-Xv(:,2)).∧2))));
end

var = var*rect;
```

D.6

comment – The part of the kernel defined as $\mathcal{L}_3(x, y \mid x_0, y_0)\big((M\psi)(x, y) - (M\psi)(x_0, y_0)\big)$.

```
function var = log_4(f)

global X2d
global rect

dim = length(f);
g = Mult(f);
var = zeros(dim,1);

for loop=1:dim
```

   comment – The vector `Xv` is used to artificially implement the limit (6.8).

```
   Xv=X2d;
   Xv(loop,1)= X2d(loop,1) - 1;

   var = var + ...
    (g(loop)-g(:)).*...
   log(((X2d(loop,1)-Xv(:,1)).∧2)+((X2d(loop,2)...
    -Xv(:,2)).∧2));
end

var = var*rect;
```

D.7

comment – The part of the kernel approximated by a modified rectangular mid-point rule.
comment – In §6.3 this function is denoted as $\mathscr{L}_i$.
comment – The value of this function depends only on the refinement.

```
function var = log_5(f)

global X2d
global rect
global X_vals

dim = length(f);
```

```
var = zeros(dim,1);
```

comment – Calculates the values `Xl`$=X$ and `Yl`$=Y$ (see §6.3)...

```
Xl=(X2d(2,1)-X2d(1,1))/2;
Yl=(X2d(1+length(X_vals),2)-X2d(1,2))/2;
```

comment – The cell containing the singularity is calculated analytically, `analytic`$=A_{X,Y}$.

```
analytic = 4*(Xl*Yl*log(Xl∧2 + Yl∧2) - 3*Xl*Yl + (Xl∧2)*atan(Yl/Xl) +...
 (Yl∧2)*atan(Xl/Yl));

for loop=1:dim
```

comment – The vector `Xv` is used to artificially remove the cell containing the singularity.

```
   Xv=X2d;
   Xv(loop,1)= X2d(loop,1) - 1;
   var = var + log(((X2d(loop,1)-Xv(:,1)).∧2)+((X2d(loop,2)-Xv(:,2)).∧2));
end

var = var*rect;
var = var + analytic;
```

Appendix E

Functions that output a vector of inner products of a function with either trial space 1 or trial space 2.

They are used to produce the 'right hand side' vectors that appear in the Galerkin approximations.

E.1

comment – Used in problem 1.

```
function var = basis_ip1(f)

global rect
global basis1

dim2 = length(basis1(1,:));
var = zeros(dim2,1);
for loop=1:dim2
   var(loop) = basis1(:,loop)'*f;
end
var = var*rect;
```

E.2

comment – Used in problem 2.

```
function var = basis_ip2(f)

global rect
global basis2

dim2 = length(basis2(1,:));
var = zeros(dim2,1);

for loop=1:dim2
   var(loop) = basis2(:,loop)'*f;
```

```
end

var = var*rect;
```

Appendix F

Functions that approximate a $L_2(D')$-inner product and $L_2(D')$-norm. As always, approximations are made using the rectangular midpoint rule.

F.1

```
function var = inner_product(f,g)

global rect
```

comment – Written knowing that all functions to be encountered are real-valued.

```
var = f'*g;
var = var*rect;
```

F.2

comment – The norm function is a trivial extension of the inner product function.

```
function var = l2_norm(f)

var = sqrt(inner_product(f,f));
```

Notation Index

Symbol	Description	Page
A	$I - K$	14
$A(\mathscr{H})$	range of operator A over $\mathscr{H}$	14
D, D'	domains	2
$E(D'), O(D')$	subspaces of $L_2(D')$	62
E_N	$N-$dimensional subspace	16
$e_n, \hat{r}_n$	pointwise, residual errors	$15, 25$
$G, \tilde{G}$	Green's function, modified Green's function	$6, 44$
$\mathcal{G}, \tilde{\mathcal{G}}$	specific integral operators	$37, 44$
$\mathscr{H}$	arbitrary Hilbert space	13
I	identity map	14
$f(x)$	free term	13
$L(p, q, f, g)$	specific functional	31
$k^2(x, y)$	$\frac{\omega^2}{c^2(x,y)}$	2
$k(x, t)$	kernel	13
K	integral operator	13
$L_2(D')$	space of Lebesgue square integrable functions over D'	36
M	operator of multiplication by $k^2(x, y) - k_0^2$	37
$\mathrm{null}(P_N)$	null space of P_N	18
$p_i, \hat{p}_i$	Galerkin, iterated Galerkin approximations	$17, 19$
P_N	orthogonal projection	16
R, T	complex amplitudes	5
S	Sloan operator	22
$\mathcal{S}_N$	approximate series	49
T^*	adjoint operator of T	19
x, y, x_0, y_0	cartesian co-ordinates	2
α, m, k_0	parameters	$56, 2$
χ_i	orthonormal basis element	16
ϕ	exact operator equation solution	13
$\phi(x, y)$	velocity potential	2
$\Phi(x, y, t)$	linear wave profile	2
ψ	solution of adjoint equation	32
$\rho_{\mathscr{H}}(S)$	specral radius of S over $\mathscr{H}$	28
$\hat{\sigma}_n$	re-iterated Galerkin approximation	26
φ	arbitrary element of $\mathscr{H}$	14
∇	$(\frac{\partial}{\partial x}, \frac{\partial}{\partial y})$	2
$(\cdot, \cdot)$	inner product	13
$\lvert\lvert \cdot \rvert\rvert$	norm	13

Bibliography

[1] Porter, D. & Stirling, D.S.G 1990 *Integral Equations: a practical treatment, from spectral theory to applications.* CUP

[2] Porter, D. & Stirling, D.S.G 1993 *The re-iterated Galerkin method, I.M.A. Journal of Numerical Analysis* **13,** pp. 125-139

[3] Reisz, F. & Sz-Nagy, B. 1955 *Functional Analysis* Frederick Ungar Pub. Co., New York

[4] Roach, G.F. 1970 *Green's functions: introductory theory with applications* Van Norstrand Reinhold

[5] Porter, R. & Porter, D. 2001 *Interaction of water waves with three-dimensional topography, Journal of Fluid Mechanics, vol.* **434** pp. 301-335

[6] Coulson, C.F. & Jeffery, A. 1977 *Waves: A mathematical approach to the common types of wave motion* Longman

[7] Burton, P.A. 1993 *Re-iterative methods for Integral Equations* MSc Dissertation, University of Reading

[8] Chamberlain, P.G. 1991 *Wave propagation on water of an uneven depth: An integral equation approach* PhD thesis, University of Reading

[9] Sloan, I.H. 1976 *Improvement by iteration for compact operator equations, Journal of Math. Comp., vol.* **30** pp. 758-764