THE UNIVERSITY OF READING

DEPARTMENT OF MATHEMATICS

Bounds on Quantities of Physical Interest

M. A. Wakefield

Thesis submitted for the degree of

Doctor of Philosophy

May 2003

Abstract

In many computational simulations there are identifiable quantities of physical interest which are weighted integrals of the solution of a boundary or initial value problem. Examples include lift and drag in aerodynamics and well productions in reservoir simulations. Methods that can bound these quantities sharply are therefore of considerable practical importance.

Although existing theory can sometimes be used to construct theoretical bounds on these quantities, for example those governed by a self-adjoint operator, it has not hitherto been exploited in a practical context. In this thesis, novel applications of these bounds have been devised and implemented for a number of model problems. In particular, the upscaling problem experienced in the oil industry, which is governed by the steady state diffusion equation, is addressed by finding bounds on the well outflow.

Extensions to the theory for problems involving non-self-adjoint operators, which enable computable bounds to be determined for quantities of physical interest, are developed using two approaches. Firstly, semi-discrete approximations are considered for the time-dependent diffusion equations and the advection-diffusion equation, although this approach does not lead to direct bounds on the analytic quantity of interest. However, such bounds are found by effectively squaring the operator, and this second method is successfully applied to the advection equation.

Finally, the construction of a grid refinement procedure to optimise the accuracy of the bounds and efficient numerical techniques for the determination of these physical quantities are considered.

Acknowledgements

I would first like to thank my supervisors Mike Baines and Dave Porter for their never ending support and enthusiasm. Your guidance has been invaluable. My thanks also extends to the other members of the Mathematics Department, past and present, and Chris Farmer, Schlumberger, for taking an active interest in this research and the many energetic discussions it prompted.

For their ever continuous love, encouragement and inspiration I have to thank Julie, Mum, Dad and Ian. All your help has been truly appreciated.

Fellow post grads including office mates Ken, Kev, Christina and Jo, thanks to you all for making the experience so enjoyable.

Finally, many thanks to all my friends and family who have tried to preserve a little balance in my life through the more hectic times of this project.

This work was supported by an EPSRC studentship.

Declaration

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

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

Concentrating on the evaluation of physical quantities expressed as weighted integrals of the solutions of partial differential equations, such as lift and drag, is a growing field in mathematical modelling and often generates numerical methods with an increased degree of accuracy for a given computational cost. In addition, identifying these quantities focuses the efforts of any approximation methods, effectively specifying the norm in which the error in the numerical computations will be judged. The quantities of interest also have a unifying effect since all the procedures involved in the solution method can be focused on a common objective. An example of such a procedure is grid refinement, based on increasing the accuracy of the quantity sought rather than unselectively resolving every feature of the solution.

A natural vehicle for this approach is optimisation. In an optimisation problem the solution is often judged on a low number of key criteria, for example the design of an airfoil section with improved lift and drag performance, as considered in [16]. Optimisation problems of this nature are often non-linear and an iterative solution procedure is adopted. During these iterations the flow around the airfoil is simulated many times, however, from these simulations the only quantities required are the lift and drag forces (and possibly the derivatives of these quantities with respect to variations in the geometry). Therefore, solving the flow around the airfoil, in some sense, can be considered a black-box procedure and in general the full flow fields generated may never actually be required by the user. Methods which enable these desired quantities to be obtained more accurately and efficiently hold obvious attraction.

Finally, it might be expected that by concentrating analytical and numerical efforts on resolving a low number of projections of the solution of a given problem, as opposed to the complete solution, greater returns may be available.

This thesis aims to investigate the common ground between numerical methods focused on resolving specific physical quantities and variational methods. In general these specific physical quantities, termed 'quantities of interest', are not derived from a variational principle, although the solutions to the governing equations may have been. However, a strong relationship exists between variational principles and the quantities of interest sought. Variational principles are concerned with relating the solution of an equation to the stationary point of an associated functional. Corresponding to the stationary point of the functional there exists the (scalar) stationary value of the functional, found by evaluating the functional at the stationary point. The stationary value of the functional is frequently found to hold physical significance and is often directly related to the quantity of interest discussed. Moreover, for a certain class of variational principles, bounds can be constructed on the stationary value of the functional, raising the possibility of bounding the related quantities of interest.

The focus of this research is to investigate under what circumstances such bounds can be constructed. The applications of the methods discussed are to model problems from the petroleum industry motivated by industrial contacts with Schlumberger, Abingdon. Due to the position of the research between classical variational principles and more specialised numerical methods the thesis rests on a broad foundation of literature categorized in the following sections.

Additional material specific to the problems considered will be referenced within the thesis.

1.2 Variational Methods

The relationship between variational principles and the solution of partial differential equations has proved a powerful theoretical and numerical tool. The application to models from fluid dynamics was initiated by Bateman [5], in which variational formulation for the governing equations of incompressible inviscid flow are presented. Bateman used the variational formulation of this problem to investigate the relationship between the boundary conditions imposed on the flow and the transition of flow from elliptic to hyperbolic form. Later, Luke [29] extended the variational formulation for an incompressible inviscid flow to include the appropriate boundary conditions for a free surface, enabling variational techniques to be applied to water wave problems.

The variational formulation of partial differential equations also enables consistent numerical methods to be constructed by replicating the continuous stationary equations at a discrete level. The best known of these techniques is the finite element family of methods to which extensive analysis has been applied, including [8, 26].

The construction of dual variational principles or complementary extremum principles for certain variational formulations, and the ability of such methods to obtain bounds on the stationary value is considered by Courant and Hilbert [13]. Dual variational principles are established by constraining the original 'free' principle and analysing the convexity present in the constrained functional.

The physical significance of the stationary point in many variational functionals was recognised by Synge [46]. Synge obtains bounds on the stationary value of physical functionals by considering the analytic solution to lie on a hypercircle defined by the approximate solutions, obtained from a pair of complementary variational principles, and an orthogonality relationship between the two approximation spaces. The hypercircle of Synge provides a useful geometric interpretation of the relationship between the pair of approximate solutions, the analytic solutions and the resulting bounds. A second geometric interpretation of the dual extremum principles is obtained by considering the convexity of the free principle. The ability to constrain the free functional in two different ways to obtain convex and concave functionals respectively, implies that the free principle is saddle-shaped. The saddle-shaped interpretation of the extremum principles is adopted and applied to a catalogue of examples by Sewell [44] and Arthurs [3]. For the purposes of this thesis the saddle-shaped interpretation will be adopted and the associated notation is introduced in section 1.2.1.

Dual extremum principles frequently arise when the governing operator is self-adjoint and positive. The positivity of the operator provides the required convexity in the governing functional and the self-adjointness enables the problem to be naturally split into two constrained variational principles. However, the desirability of dual variational principles to provide bounds on quantities of physical significance motivates extension of the methods to include non-self-adjoint problems. Gurtin [19] developed variational principles for (non-self-adjoint) linear initial-value problems in which the Laplace transform is applied to the governing equation in order to render the temporal derivatives purely algebraic. The variational principle is then constructed in the transform space. The method is considered with respect to the wave problem and the heat conduction problem although the concept of a quantity of interest is absent.

A quantity of interest is presented in Collins' [12] dual extremum formulation for the heat equation, in which an adjoint equation compensates for the energy dissipated by the primal and thus global conservation is retained in a particular sense. Although rigorous extremum principles are derived by Collins, the applicability of the method is restricted by the nature of the constraints required to generate the necessary saddle-shaped topology. The constraints on the trial functions takes the form of a Poisson equation and satisfying these exactly, as is required to guarantee valid bounds, is little easier than solving the original problem analytically. The applicability of these constraints associated with dual extremum principles is an important factor in determining the practicality of the method with respect to engineering applications. Considering methods in which the required constraints can be satisfied efficiently will be a key emphasis of the thesis.

The aim of the research is then to consider novel applications for existing dual variational principles and extend existing theory to include non-self-adjoint problems in a computationally efficient manner.

1.2.1 Saddle-Shaped Functionals

The existence of dual extremum principle relates directly to the saddle-shaped topology of the governing functional [3, 44]. The crucial property of the saddle-shaped topology is the existence of a pair of axes on which the functional is concave and convex respectively. These axes correspond to the pair of constrained functionals which also have opposing convexity properties. A representation of a saddle-shaped functional $\mathcal{G}(p,q)$ is shown in figure 1.1. Upper and lower bounds on the stationary value $\mathcal{G}(\widehat{p},\widehat{q})$ are found

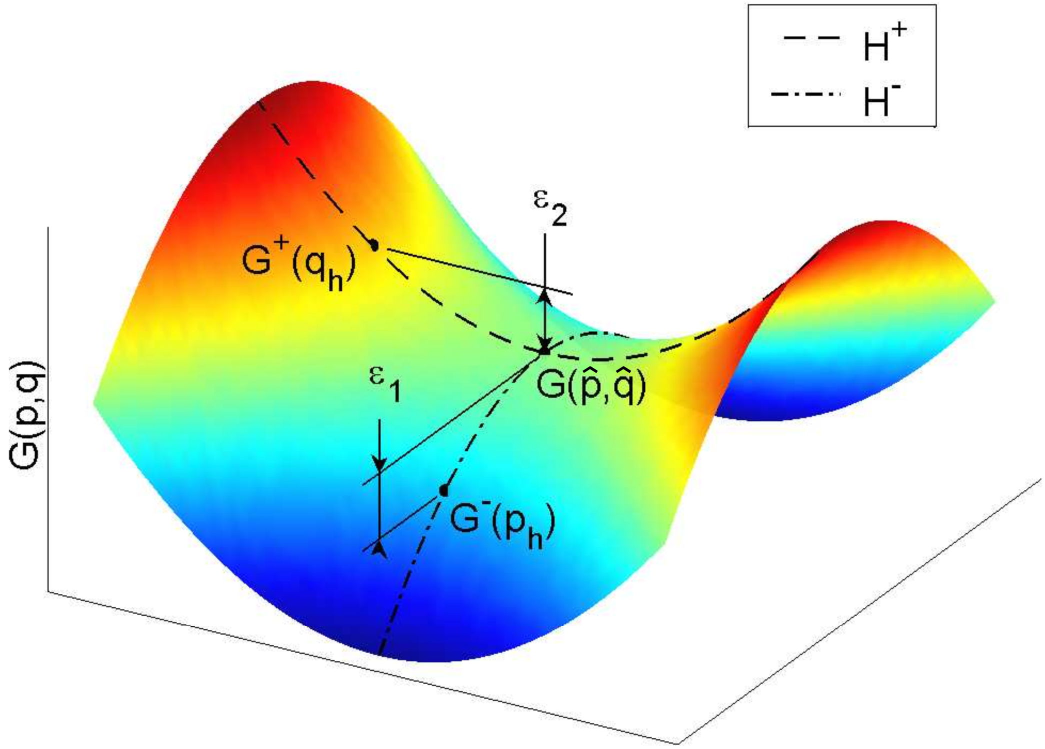

Figure 1.1: A saddle-shaped functional

by constraining the functional to lie on the axis for which the functional is either convex or concave respectively. The hatted variables denote the stationary point of the functional, or equivalently the analytic solution of the governing equation. The functional is constrained by satisfy particular relationships between the functions p and q, termed the constraints. For example, satisfying the constraint H^- which defines q in terms of p enables the dependence of the functional on p and q to be be replaced by dependence on p only. Consequently, the resulting constrained functional $\mathcal{G}^-(p)$ inherits the convexity associated with the unconstrained functional along the axis H^-. Moreover, the construction is such that the maximum $\mathcal{G}^-(\widehat{p})$ of the constrained functional $\mathcal{G}^-(p)$ coincides with the stationary point of the unconstrained functional $\mathcal{G}^-(\widehat{p},\widehat{q})$.

Approximations to the stationary point $p_h \approx \widehat{p}$ can therefore be made by maximising the value of the functional $\mathcal{G}^-(p_h)$. Whether or not the stationary point of the constrained functional is attained by maximising $\mathcal{G}^-(p_h)$ depends on whether the analytic solution $\widehat{p}$ is an element of the approximation space. If, as is more usual, $\widehat{p}$ is not an element of the approximations space, maximising $\mathcal{G}^-(p_h)$ results in the selection of the element p_h that minimises the error $\mathcal{G}^-(\widehat{p}) - \mathcal{G}^-(p_h) = \epsilon_1$. Crucially, due to the convex nature of the functional $\mathcal{G}^-(p)$, $\mathcal{G}^-(p) \leq \mathcal{G}^-(\widehat{p}) = \mathcal{G}(\widehat{p}, \widehat{q})$ for all p and hence $\mathcal{G}^-(p_h)$ is an optimal lower bound on the stationary value of the functional. The bound is optimal in the sense that the error ϵ_1 is minimised. Similarly, an optimal upper bound is found by constraining the functional $\mathcal{G}(p, q)$ by the constraint H^+ to obtain the functional $\mathcal{G}^+(q)$, and minimisation of $\mathcal{G}^+(q_h)$, yields an optimal upper bound and the approximation $q_h \approx \widehat{q}$ in an analogous manner.

Although, in practice satisfying the constraints H^- and H^+ may complicate implementing the dual extremum principles described, the resulting bounds are rewardingly sharp, due to their optimal nature. In particular, weak inequalities and approximations to the spectrum of the operator are avoided. A method of obtaining sharp bounds is a necessity if they are to be used in an engineering context.

The variational techniques cited provide a natural framework for bounding the stationary value of the governing functional, a quantity which is often of physical significance and is expressed as a weighted integral of the analytic solution. The properties of the stationary value have been known since before the work of Synge, but in contrast to classical variational methods, the development of numerical methods to efficiently resolve linear functionals of the solution is a relatively recent advance.

The definition of the quantity of interest is introduced in the next section. Determination of the bounds on this single quantity is the objective of the methods and techniques discussed in the following chapters.

1.3 Quantities of Interest

The quantity of interest Θ, an integral of the solution, can be considered as the inner product of a weighting function t with the solution $\widehat{\phi}$ of the governing equation

$$A\widehat{\phi} = s, \tag{1.1}$$

where A is a differential operator. Hence

$$\Theta = (\widehat{\phi}, t), \tag{1.2}$$

for some general inner product $(\cdot, \cdot)$. In practice the inner product will depend on the nature of the operator A and may include boundary terms. A specific inner product will be considered in chapter 2. The dual problem governed by the adjoint equation is found to be particularly useful in this context and is defined as

$$A^* \widehat{\sigma} = t. \tag{1.3}$$

The dual problem allows the inner product (1.2) to be expressed either in terms of the primal or dual solutions through

$$\Theta(\widehat{\phi}) = (\widehat{\phi}, t) = (\widehat{\phi}, A^* \widehat{\sigma}) = (A\widehat{\phi}, \widehat{\sigma}) = (s, \widehat{\sigma}). \tag{1.4}$$

An important application of the adjointness statement (1.4) is when the quantity of interest is required to be evaluated for many different forcing functions s_i of the primal equation, as outlined by Giles and Pierce [18]. In this context it is computationally cheaper to solve the dual problem and obtain $\widehat{\sigma}$, or an approximation to it, and then evaluate the quantity of interest using

$$\Theta_i(\widehat{\phi}) = (s_i, \widehat{\sigma}). \tag{1.5}$$

This situation occurs in many optimisation problems in which a solution is found iteratively and the forcing function for the primal problem (1.1) varies from iteration to iteration. Often however, the weight t in the quantity of interest (1.2) does not vary over the iterations and therefore neither does the solution $\widehat{\sigma}$. The evaluation of (1.5) is therefore considerably cheaper than $(\widehat{\phi}_i, t)$ for multiple functions s_i since the cost of obtaining multiple solutions of the primal equation (1.1) is substituted for the cost of obtaining a single solution of the dual equation (1.3).

The relationship (1.4) illustrates the role that the dual equation plays in the relationship between the primal equation and the quantity of interest. In general A is a partial differential operator and the analytic solutions $\widehat{\phi}$ and $\widehat{\sigma}$ are unavailable. Instead, approximations ϕ_h and σ_h are constructed using numerical methods and an approximation to the quantity of interest $\Theta(\phi_h)$ is found. Having obtained an approximation to the quantity of interest the error incurred in this quantity due to the approximate nature of ϕ_h and σ_h can be estimated using the 'error representation formula' of [45].

The 'error representation formula' is derived by considering the adjointness statement (1.4). Employing (1.4) the error in the quantity of interest when ϕ is approximated by ϕ_h can be equated to the residual of the primal problem (1.1) weighted by the dual solution, namely the error representation formula,

$$\Theta(\widehat{\phi}) - \Theta(\phi_h) = (\widehat{\phi} - \phi_h, t) = (\widehat{\phi} - \phi_h, A^*\widehat{\sigma}) = (A\widehat{\phi} - A\phi_h, \widehat{\sigma}) = (s - A\phi_h, \widehat{\sigma}). \quad (1.6)$$

Although this error representation formula is given in terms of $\widehat{\sigma}$, the analytic solution of the dual problem, and is therefore not directly accessible, approximations to the formula have formed the basis of much *a posteriori* error analysis concerning quantities of interest [7, 17, 38, 41, 45].

In contrast to the route suggested by the error representation formula, this research aims to bound the quantity of interest, and hence the error in the quantity of interest, using dual variational methods. These methods enable bounds to be constructed from a pair of numerical solutions and therefore the error bound is directly computable, as opposed to the error representation formula (1.6).

Having obtained an estimate for the error in the quantity of interest this information can be harnessed and used to improve the accuracy of the numerical simulation. Two main applications of this philosophy are in grid adaption and error correction.

1.3.1 Grid Adaption

The aim of a grid adaption algorithm is to efficiently construct an approximation space such that the error in the quantity of interest is less than a user defined tolerance

$$|\Theta(\widehat{\phi}) - \Theta(\phi_h)| \leq tol. \quad (1.7)$$

Grid adaption strategies arise naturally from *a posteriori* error estimation methods since the local contribution to the error representation formula can be evaluated from each element. Representing the error in the quantity of interest by the sum of the local error indicators η_i, so that

$$|\Theta(\widehat{\phi}) - \Theta(\phi_h)| = \sum_i \eta_i, \quad (1.8)$$

the grid can be refined in elements in which the local error indicator fails a local criteria. Various methods have been proposed to calculate the local error indicators including the Type I and II estimates of Suli and Houston [45].

The Type II *a posteriori* estimates are found by assuming that a finite element method is used to obtain the solutions ϕ_h and σ_h, where both solutions lie in the same approximation space. The error representation formula can then be written as

$$\begin{aligned} \Theta(\widehat{\phi}) - \Theta(\phi_h) &= (s - A\phi_h, \widehat{\sigma}), && (1.9) \\ &= (s - A\phi_h, \widehat{\sigma} - \sigma_h), && (1.10) \end{aligned}$$

through the well-known Galerkin orthogonality property. The Type II *a posteriori* estimates are then obtained by applying the Cauchy-Schwartz inequality to the error representation formula (1.10), and effectively bounding the norm $\|\widehat{\sigma}-\sigma_h\|$ by the product of an interpolation constant and a strong stability constant. However, in applying the Cauchy-Schwartz inequality and bounding the contribution from the adjoint equation the interplay between the primal and dual solutions is lost. As a result the Type II estimates prove poor at identifying regions in the solution to which the quantity of interest is sensitive. In addition, the ability to obtain the pair of constants, and the circumstances under which the bound is valid, in general is limited.

Instead, the Type I estimate is considered in which $\widehat{\sigma}$, or in practice a numerical approximation $\sigma_d \approx \widehat{\sigma}$, is present. The approximate solution of the dual problem σ_d must be found in a different space from that of the primal solution so that the error representation expression is non-zero. In addition, the error resulting from the substitution of $\widehat{\sigma}$ by σ_d must not adversely effect the accuracy of the error estimate. To ensure that σ_d is sufficiently well resolved the dual grid is refined to a tolerance less than that of primal grid. The Type I error estimate proves effective since local contributions to the inner product (1.10) are only substantial if both the primal residual and the dual solution are significant within the element. In this manner only regions in which the solution affects the quantity of interest are highlighted. Type I error estimates have been successfully implemented by Suli and Houston using the streamline-diffusion finite element method to model the stationary transport equation, and the discontinuous Galerkin finite element method to model the wave equation, Burger's equation and an aerofoil simulation.

The Type I error estimates generate highly efficient grids on which to resolve quantities of interest since solution details that are irrelevant to the quantity of interest are automatically ignored. The efficiency of the method attracts engineering applications and further extensions of the method including non-linear theory are discussed in [45] and constructed in [21]. In particular, the validity of the error representation formula and approximations to it in the proximity of shocks is considered.

The error representation formula is the basis of a family of *a posteriori* error estimation techniques including the Dual Weighted Residual Method of Becker and Rannacher [7, 41]. Becker and Rannacher advocate the use of a symmetrical form of the error representation formula

$$\begin{aligned}\Theta(\widehat{\phi}) - \Theta(\phi_h) &= (s - A\phi_h, \widehat{\sigma}), && (1.11)\\ &= \frac{1}{2}(s - A\phi_h, \widehat{\sigma} - \sigma_h) + \frac{1}{2}(\widehat{\phi} - \phi_h, t - A\sigma_h) && (1.12)\end{aligned}$$

in order to balance the contribution from the primal and dual residuals.

The mesh refinement procedure is naturally an iterative process in which the grid is progressively refined and the solution re-calculated until the desired accuracy is achieved. In contrast, error correction is a technique to recover additional accuracy in the quantity of interest, and can be applied only once.

1.3.2 Error Correction

The error correction method enables additional accuracy to be obtained in the quantity of interest by adding a term based on the weighted residual of the primal problem. The technique has been developed by Giles and Pierce [17, 18, 38], and is based on the expansion of the quantity of interest in the form

$$\begin{aligned}\Theta(\widehat{\phi}) &= (\widehat{\phi}, t), && (1.13)\\ &= (\phi_h, t) - (\phi_h - \widehat{\phi}, A^*\sigma_h) + (\phi_h - \widehat{\phi}, A^*\sigma_h - A^*\widehat{\sigma}), && (1.14)\\ &= (\phi_h, t) - (A\phi_h - A\widehat{\phi}, \sigma_h) + (\phi_h - \widehat{\phi}, A^*\sigma_h - A^*\widehat{\sigma}). && (1.15)\end{aligned}$$

The first term of the expansion is the value obtained for the quantity of interest when the approximate solution ϕ_h is used. The second term represents a computable correction to the first term. The correction is computable because $A\phi_h - A\widehat{\phi}$ is the residual of the primal problem and σ_h is the numerical solution of the dual problem. The third term is the remaining error and is not computable due to the presence of the analytic solution $\widehat{\phi}$. Numerical results generated using the error correction method have demonstrated that superconvergence can be obtained in the quantity of interest with respect to the order of the schemes used to calculate the solutions ϕ_h and σ_h. Superconvergence of the quantity of interest is an advantageous property reducing the computational cost of obtaining accurate results. Interestingly, finite element methods in which the primal and dual solutions are approximated in the same space already exhibit this convergence rate as $(A\phi_h - A\widehat{\phi}, \sigma_h) = 0$ by the Galerkin orthogonality property. However, when a non-optimal method is implemented, such as finite volumes, the error correction method enables this advantageous convergence property to be recovered.

The stationary value of the functionals considered naturally exhibit super-convergent characteristics due to the near stationary nature of the functional in the proximity of the stationary point. This property is explored in section 2.3.1.

1.4 Contents of Thesis

The content of thesis is based on constructing upper and lower bounds on quantities of interest. In general, the bounds are obtained by associating the quantity of interest with the stationary points of two saddle-shaped functionals. The saddle-shaped topology of these functionals enables bounds to be generated on the stationary values, which translate to bounds on the quantity of interest.

In chapter 2 the saddle-shaped structure of the functional associated with a self-adjoint problem is explored and an example governed by the diffusion equation is considered. The variational principles governing the diffusion equation can be found in [3] and [44]. From the diffusion functional the dual extremum principles that enable the upper and lower bounds to be found are demonstrated and the advantageous accuracy properties of the stationary point are considered. Numerical results are generated from the diffusion functional by considering the ‘upscaling’ problem encountered in the petroleum industry. The ‘upscaling’ problem involves the averaging of physical properties of the oil reservoirs in some sense to enable efficient numerical simulations to be implemented. The averaging is required because the original description is prohibitively large, but must be sensitive enough to preserve important features of the model.

In chapter 3 a review of existing upscaling methods is presented and a new upscaling philosophy generated. The new philosophy is based on preserving the bounds on the quantity of interest whilst implementing an upscaling method. The quantity of interest in these simulation is considered to be the flux out of the reservoir, equivalent to the well productions. This new method enables the performance of various existing upscaling methods to be contrasted. In addition, methods based on the new upscaling philosophy are also constructed, tested numerically, and compared with conventional upscaling methods. Extensions to permeability data with uncertainties and multiphase upscaling are also discussed.

The second half of the thesis is concerned with obtaining bounds on quantities of interest in which the problem is governed by a non-self-adjoint operator. The lack of self-adjointness in the governing equations removes the saddle-shaped topology of the associated functionals and upper and lower bounds of the nature sought are not immediately available. Various methods to regain self-adjointness are proposed. In chapter 4 the non-self-adjoint advection-diffusion equation is considered with applications as another prototype oil reservoir model. Motivated by this type of equation the time-dependent diffusion equation is initially considered. Discretising the time-dependent equation implicitly in time enables each time-step to be taken by solving a Helmholtz equation. The Helmholtz equation has an associated saddle-shaped functional, [3] and [44], and this presents the possibility of finding upper and lower bounds again. However, the discretisation of the equation in time is found to weaken these bounds to mere approximations to the analytic quantity of interest. A similar approach is found to be possible with the advection-diffusion equation, except that a Lagrangian discretisation is required. Again, however, the bounds obtained are only approximations to the analytic quantity of interest. For the advection-diffusion equation the approximations to the quantity of interest are found to deteriorate when advection dominated flows are considered, and therefore methods to obtain bounds on the analytic value of the quantity of interest are sought.

To obtain bounds on the analytic value of the quantity of interest the governing equations require modifying in such a manner as to re-introduce the required convexity into the associated functionals. Methods to achieve this modification are discussed in chapter 5. The approach adopted is essentially to “square” the operator in the governing equations and solve the new problem over the whole time-space domain. An example based on the advection equation is considered, motivated by the difficulties the advection terms caused the semi-discrete methods constructed in chapter 4. Solving the governing equations over the complete time-space domain can be expensive and therefore a grid refinement procedure based on improving the accuracy in the quantity of interest is presented and also applied to the advection equation.

Up to this point the focus of the methods considered have been on obtaining bounds on the quantity of interest. However, approximate solutions to the governing equations are also obtained during the course of these method. In chapter 6 an investigation is carried out into the possibility of obtaining the quantity of interest without having to solve the governing equations. In addition, a comparison between the relative costs of methods capable of calculating the quantity of interest is made.

The final chapter draws some general conclusions from the thesis and identifies areas for further research.

Chapter 2

The Self-Adjoint Operator Case

The aim of this chapter is to present a systematic method of obtaining upper and lower bounds on the stationary value of a functional which involves the solution of an equation containing a positive self-adjoint operator. The functional is found to be saddle-shaped with the underlying convexity generated through the positivity associated with the operator. Subsequently, it is found that the functional can be related to both the Lagrangian and Hamiltonian formalisms through which the convexity can be attributed to the concept of an 'energy' associated with the solution.

Initially we employ a general notation for clarity, assuming only that the functions involved are real valued.

The problem is to solve the equation

$$A\widehat{p} = r, \tag{2.1}$$

where A is a positive self-adjoint operator, and can therefore be written as $A = T^*T$ for some operator T. The governing equation (2.1) does not explicitly include boundary terms but they can be considered to be present in the definition of the operator.

Much of the theory which follows consists of obtaining weak solutions to (2.1) and therefore appropriate inner products are required. The inner products are associated with the two components, T and T^* of the operator A and satisfy the adjointness statement

$$\langle\langle q, Tp\rangle\rangle = \langle\langle Tp, q\rangle\rangle = \langle p, T^*q\rangle = \langle T^*q, p\rangle. \tag{2.2}$$

As with the operator, the inner products may in part be defined over the boundary of the domain.

In the following sections the saddle shaped structure of certain functionals will be illustrated. Then, having developed the theory in a general framework the method will be applied to a simplified oil reservoir model. In the oil reservoir model the relationship (2.2) will arise from the $-div$ / $grad$ adjointness statement embodied in the divergence theorem.

2.1 Bounds via a Saddle-Shaped Functional

In this section the saddle-shaped functional $\mathcal{G}(p,q)$ is introduced. The functional is found to be stationary at the solution of the governing equation (2.1) and the stationary value is found to be closely related in structure to the quantities of physical interest being pursued. Moreover, bounds can be established on the stationary value by applying constraints to the functional $\mathcal{G}(p,q)$ in the manner described in section 1.2.1.

2.1.1 The Governing Functional

Writing the governing equation (2.1) as two equations involving an intermediate function $\widehat{q}$ we obtain the pair

$$T^*T\widehat{p} = r \qquad \left\{ \begin{array}{rcl} T\widehat{p} &=& \widehat{q}, \\ T^*\widehat{q} &=& r. \end{array} \right. \tag{2.3}$$

The solution of the pair (2.3) is the stationary point of the functional

$$\mathcal{G}(p,q) = \frac{1}{2}\langle\langle q,q\rangle\rangle - \langle\langle Tp,q\rangle\rangle + \langle p,r\rangle \tag{2.4}$$

which is found by equating the first order variations of the functional with respect to p and q, to zero. Hence,

$$\begin{aligned} 0 &= \delta\mathcal{G}(p,q), && (2.5) \\ &= \langle\langle \delta q, q - Tp\rangle\rangle - \langle \delta p, T^*q - r\rangle, && (2.6) \end{aligned}$$

for all variations δq and δp, which implies that the pair (2.3) are satisfied at the stationary point and the solution $p = \widehat{p}$, $q = \widehat{q}$ is attained there.

2.1.2 The Stationary Value

The stationary value is obtained by substituting the natural conditions (2.3) into the functional to obtain

$$\begin{aligned}
\mathcal{G}(\widehat{p},\widehat{q}) &= \frac{1}{2}\langle\langle \widehat{q},\widehat{q}\rangle\rangle - \langle\langle T\widehat{p},\widehat{q}\rangle\rangle + \langle \widehat{p}, r\rangle, && (2.7)\\
&= -\frac{1}{2}\langle\langle T\widehat{p},\widehat{q}\rangle\rangle + \langle \widehat{p}, r\rangle, && (2.8)\\
&= -\frac{1}{2}\langle \widehat{p}, T^{*}\widehat{q}\rangle + \langle \widehat{p}, r\rangle, && (2.9)\\
&= \frac{1}{2}\langle \widehat{p}, r\rangle. && (2.10)
\end{aligned}$$

The stationary value is therefore in the form of a particular quantity of interest Θ, namely the inner product of the analytic solution $\widehat{p}$ of the governing equation with the weight function r, (although the quantity of interest has an additional multiplicative constant, $\frac{1}{2}$, in this case.) Upper and lower bounds on the stationary value and therefore bounds on the quantity of interest, are available in this case.

2.1.3 Upper and Lower Bounds on the Stationary Value

Upper and lower bounds on the stationary value are found by constraining the functional $\mathcal{G}(p,q)$ to obtain concave and convex functionals that are themselves stationary at the point $\mathcal{G}(\widehat{p},\widehat{q})$.

Firstly, the functional $\mathcal{G}(p,q)$ is constrained to lie on the hyperline H^{-} defined by the equation

$$q = Tp. \qquad (2.11)$$

Substituting (2.11) into the functional $\mathcal{G}(p,q)$ ensures that for all functions p, q is defined by the constraint. Making the substitution the functional $\mathcal{G}^{-}(p) = \mathcal{G}(p,Tp)$ is obtained,

$$\mathcal{G}^{-}(p) = \langle r, p\rangle - \frac{1}{2}\langle\langle Tp, Tp\rangle\rangle. \qquad (2.12)$$

A one-sided bound can be established by showing that the difference between the general value of the functional and its stationary value is non-negative. Considering the difference in the value of the functional evaluated at the stationary point $\widehat{p}$ and the value of the functional evaluated with *any* other suitable function p we obtain

$$\mathcal{G}^{-}(\widehat{p}) - \mathcal{G}^{-}(p) = \langle r,\widehat{p}\rangle - \frac{1}{2}\langle\langle T\widehat{p}, T\widehat{p}\rangle\rangle - \langle r, p\rangle + \frac{1}{2}\langle\langle Tp, Tp\rangle\rangle, \qquad (2.13)$$

$$
\begin{aligned}
&= \langle T^*\widehat{q}, \widehat{p} - p\rangle - \frac{1}{2}\langle\!\langle T\widehat{p}, T\widehat{p}\rangle\!\rangle + \frac{1}{2}\langle\!\langle Tp, Tp\rangle\!\rangle, && (2.14)\\
&= \langle\!\langle\, \widehat{q}, T(\widehat{p} - p)\,\rangle\!\rangle - \frac{1}{2}\langle\!\langle T\widehat{p}, T\widehat{p}\rangle\!\rangle, + \frac{1}{2}\langle\!\langle Tp, Tp\rangle\!\rangle, && (2.15)\\
&= \frac{1}{2}\langle\!\langle\, T(\widehat{p} - p), T(\widehat{p} - p)\,\rangle\!\rangle, && (2.16)\\
&\geq 0, && (2.17)
\end{aligned}
$$

and hence $\mathcal{G}^-(p)$ is an underestimate of $\mathcal{G}^-(\widehat{p})$ and a lower bound for the stationary value of the functional. The principle

$$\delta\mathcal{G}^-(p) = 0 \qquad p \in H^- \tag{2.18}$$

is referred to as a maximum principle since the maximum value of the functional occurs at the solution of the governing equation (2.3).

Similarly, by defining the hyperline H^+ by the equation

$$T^*q = r \tag{2.19}$$

and constraining the functional to lie on this line, an upper bound is found. In contrast with the lower bound the constraint H^+ cannot be directly substituted into the functional due to the lack of an explicit relationship expressing the function p in terms of q. Instead the approximation space is constrained and q is considered to belong to the set of functions H^+ satisfying (2.19). Under these assumptions we obtain the functional $\mathcal{G}^+(q)$,

$$\mathcal{G}^+(q) = \frac{1}{2}\langle\!\langle q, q\rangle\!\rangle. \tag{2.20}$$

The functional $\mathcal{G}^+(q)$ provides an upper bound on the stationary value, since for any suitable q and ϕ

$$
\begin{aligned}
\mathcal{G}^+(q) - \mathcal{G}^+(\widehat{q}) &= \frac{1}{2}\langle\!\langle q, q\rangle\!\rangle - \frac{1}{2}\langle\!\langle \widehat{q}, \widehat{q}\rangle\!\rangle, && (2.21)\\
&= \frac{1}{2}\langle\!\langle q, q\rangle\!\rangle - \frac{1}{2}\langle\!\langle \widehat{q}, \widehat{q}\rangle\!\rangle + \langle\, \phi, T^*(\widehat{q} - q)\,\rangle, && (2.22)\\
&= \frac{1}{2}\langle\!\langle q, q\rangle\!\rangle - \frac{1}{2}\langle\!\langle \widehat{q}, \widehat{q}\rangle\!\rangle + \langle\!\langle T\phi, \widehat{q} - q\rangle\!\rangle, && (2.23)
\end{aligned}
$$

and in particular when $\phi = \widehat{p}$

$$
\begin{aligned}
\mathcal{G}^+(q) - \mathcal{G}^+(\widehat{q}) &= \frac{1}{2}\langle\!\langle q, q\rangle\!\rangle - \frac{1}{2}\langle\!\langle \widehat{q}, \widehat{q}\rangle\!\rangle + \langle\!\langle T\widehat{p}, \widehat{q} - q\rangle\!\rangle, && (2.24)\\
&= \frac{1}{2}\langle\!\langle q, q\rangle\!\rangle - \frac{1}{2}\langle\!\langle \widehat{q}, \widehat{q}\rangle\!\rangle + \langle\!\langle \widehat{q}, \widehat{q} - q\rangle\!\rangle, && (2.25)\\
&= \frac{1}{2}\langle\!\langle q - \widehat{q}, q - \widehat{q}\rangle\!\rangle, && (2.26)\\
&\geq 0. && (2.27)
\end{aligned}
$$

The principle

$$\delta\mathcal{G}^+(q) = 0 \qquad q \in H^+ \tag{2.28}$$

is referred to as a minimum principle since the minimum of the functional occurs at the solution of the governing equation (2.3).

The contrasting nature of the two constraints (2.11) and (2.19) results in the numerics concerned with the upper and lower bounds having differing degrees of complexity. Satisfying the first constraint by constructing a pair of functions $(p, q) \in H^-$ is trivially achieved by applying the operator T to the function p. Satisfying the second constraint is non-trivial and in general obtaining an upper bound on the stationary value is more demanding. However, it is worth noting that in splitting the self-adjoint operator into its components T and T^*, it is not assumed that the constraint H^+,

$$T^*q = r \tag{2.29}$$

uniquely determines $q = \widehat{q}$, but rather implies a set of functions containing $\widehat{q}$ over which $\mathcal{G}^+(q)$ can be minimised.

The saddle-shaped topology of the functional $\mathcal{G}(p, q)$ is demonstrated by the convex and concave functionals, $\mathcal{G}^-(p)$ and $\mathcal{G}^+(q)$, existing along the "axes" H^- and H^+ respectively. A sketch of the axes is shown in figure 2.1, with the solution of the governing equation naturally occurs at the intersection of the two hyperlines H^- and H^+.

2.1.4 Finite Dimensional Approximations

The derivation of the upper and lower bounds holds for all $q \in H^+$ and all $p \in H^-$ respectively. To obtain numerical approximations the finite dimensional subspaces $H_h^+ \subseteq H^+$ and $H_h^- \subseteq H^-$ are introduced in which the approximations $p_h \approx \widehat{p}$ and $q_h \approx \widehat{q}$ can be found. Defining $\mu^\pm$ in terms of the optimum bounds for the given subspaces,

$$\mu^- = \max_{p_h \in H_h^-} \mathcal{G}^-(p_h), \tag{2.30}$$

$$\mu^+ = \min_{q_h \in H_h^+} \mathcal{G}^+(q_h), \tag{2.31}$$

the bounds on the quantity of interest

$$2\mu^- \leq \langle \widehat{p}, r \rangle \leq 2\mu^+ \tag{2.32}$$

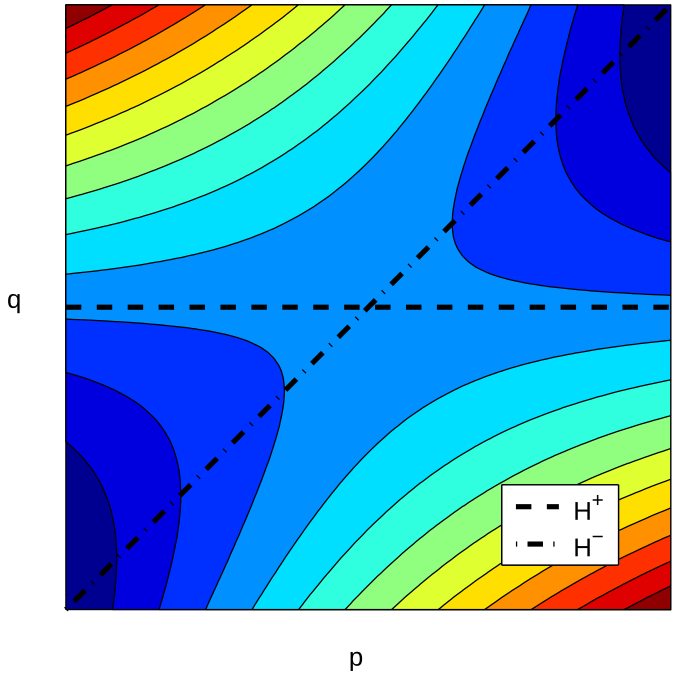

Figure 2.1: The saddle shaped functional $\mathcal{G}(p, q)$

are established.

The functionals $\mathcal{G}^-(p)$ and $\mathcal{G}^+(q)$ are quadratic and therefore have a unique maximiser and minimiser respectively. Correspondingly, obtaining the bounds μ^- and μ^+ in the finite-dimensional subspaces reduces to solving symmetric positive definite systems of linear normal equations, for which standard matrix methods can be used. The self-adjointness of the operator implies symmetry and positive definite properties in the matrix and this allows efficient methods such as conjugate gradients (CG) to be applied.

To summarise, the variational principle $\delta\mathcal{G}(p, q) = 0$ is found to deliver the required stationary conditions and in addition the stationary value has found to be directly related to a particular quantity of interest. The saddle-shaped topology of the functional $\mathcal{G}(p, q)$ is fundamental to obtaining the upper and lower bounds on the stationary value and is due to a positive 'energy' term in the Lagrangian formalism. The Lagrangian in turn enables the Hamiltonian formalism to be constructed which includes the introduction of the intermediate variable q. As a means of illuminating the foundations of the functional $\mathcal{G}(p, q)$ from this point of view, a derivation from classical mechanics is considered.

2.2 A Derivation from Classical Mechanics

There is a large body of literature on Lagrangian and Hamiltonian Mechanics (see e.g. the introductory text [30].) Examples of generating saddle-shaped functionals from Hamilton's equations can be found in [44].

Initially we consider the Lagrangian formalism. The Lagrangian is defined as a kinetic energy term minus a potential energy term. The kinetic energy term is defined in terms of the energy norm associated with the inner product $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ in the standard manner,

$$\|\phi\|^2_{\langle\!\langle\rangle\!\rangle} = \langle\!\langle \phi, \phi \rangle\!\rangle. \tag{2.33}$$

2.2.1 The Lagrangian Formalism

In the Lagrangian formalism we consider the concave Lagrangian functional $\mathcal{L}(p, Tp)$,

$$\begin{aligned} \mathcal{L}(p, Tp) &= \frac{1}{2}\|Tp\|^2_{\langle\!\langle\rangle\!\rangle} - \langle p, r \rangle, \\ &= \frac{1}{2}\langle\!\langle Tp, Tp \rangle\!\rangle - \langle p, r \rangle, \\ &\quad \text{Kinetic term} - \text{Potential term}. \end{aligned} \tag{2.34}$$

The first order variations of the functional are

$$\begin{aligned} \delta\mathcal{L} &= \langle\!\langle Tp, \delta(Tp) \rangle\!\rangle - \langle r, \delta p \rangle, \\ &= \langle T^*Tp - r, \delta p \rangle. \end{aligned} \tag{2.35}$$

which vanish at the stationary point $p = \widehat{p}$ for all variations in p. The criterion $\delta\mathcal{L}(p, Tp) = 0$ is simply Hamilton's principle and the governing equation (2.3) is the corresponding Euler-Lagrange equation.

The Lagrangian functional automatically provides the required convexity to obtain a one-sided bound on the stationary value $\mathcal{L}(\widehat{p}, T\widehat{p})$. The functional $\mathcal{L}(p, Tp)$ has the same form as the functional $\mathcal{G}^-(p)$ and the bound is demonstrated in the same manner. Considering the difference between the functional evaluated at the stationary point $(\widehat{p}, T\widehat{p})$ and the functional evaluated at any point (p, Tp), then

$$\mathcal{L}(\widehat{p}, T\widehat{p}) - \mathcal{L}(p, Tp) = \frac{1}{2}\langle\!\langle T\widehat{p}, T\widehat{p} \rangle\!\rangle - \langle \widehat{p}, r \rangle - \frac{1}{2}\langle\!\langle Tp, Tp \rangle\!\rangle + \langle p, r \rangle, \tag{2.36}$$

$$
\begin{aligned}
&= \frac{1}{2}\langle\langle T\widehat{p}, T\widehat{p}\rangle\rangle - \frac{1}{2}\langle\langle Tp, Tp\rangle\rangle + \langle p-\widehat{p}, r\rangle, && (2.37)\\
&= \frac{1}{2}\langle\langle T\widehat{p}, T\widehat{p}\rangle\rangle - \frac{1}{2}\langle\langle Tp, Tp\rangle\rangle + \langle\langle Tp - T\widehat{p}, T\widehat{p}\rangle\rangle, && (2.38)\\
&= -\frac{1}{2}\langle\langle T\widehat{p} - Tp, T\widehat{p} - Tp\rangle\rangle, && (2.39)\\
&= -\frac{1}{2}\|T\widehat{p} - Tp\|^2_{\langle\langle\rangle\rangle}, && (2.40)\\
&\leq 0, && (2.41)
\end{aligned}
$$

and hence the value of the Lagrangian functional is never less than the analytic stationary value of the functional and forms an upper bound on it.

The complementary bound is obtained by increasing the scope of the functional by introducing a second independent (conjugate) function q, through a Legendre transformation enabling the change of variables $(p, Tp) \mapsto (p, q)$. The change of variables introduces the Hamiltonian formalism and an associated functional that is stationary at the solution of Hamilton's equations.

2.2.2 The Hamiltonian Formalism

The Hamiltonian functional $\mathcal{H}(p, q)$ is defined by the Legendre transformation

$$\mathcal{L}(p, Tp) = \langle\langle Tp, q\rangle\rangle - \mathcal{H}(p, q), \tag{2.42}$$

between the pairs (p,Tp) and (p,q), where q is a conjugate variable. Equating the first order variations we obtain

$$\langle\frac{\partial\mathcal{L}}{\partial p}, \delta p\rangle + \langle\langle\frac{\partial\mathcal{L}}{\partial(Tp)}, \delta(Tp)\rangle\rangle = \langle\langle Tp, \delta q\rangle\rangle + \langle\langle\delta(Tp), q\rangle\rangle - \langle\frac{\partial\mathcal{H}}{\partial p}, \delta p\rangle - \langle\langle\frac{\partial\mathcal{H}}{\partial q}, \delta q\rangle\rangle. \tag{2.43}$$

Defining the conjugate function to be

$$q = \frac{\partial\mathcal{L}}{\partial(Tp)} = Tp, \tag{2.44}$$

the relationship

$$0 = \langle\langle Tp - \frac{\partial\mathcal{H}}{\partial q}, \delta q\rangle\rangle - \langle\frac{\partial\mathcal{H}}{\partial p} + \frac{\partial\mathcal{L}}{\partial p}, \delta p\rangle \tag{2.45}$$

is obtained from (2.43) and leads to Hamilton's equations

$$
\begin{aligned}
\frac{\partial\mathcal{H}}{\partial q} &= Tp, && (2.46)\\
\frac{\partial\mathcal{H}}{\partial p} &= -\frac{\partial\mathcal{L}}{\partial p}. && (2.47)
\end{aligned}
$$

From Hamilton's equations the Hamiltonian $\mathcal{H}(p,q)$ can be constructed,

$$\mathcal{H}(p,q) = \frac{1}{2}\langle\langle q,q\rangle\rangle + \langle p,r\rangle, \tag{2.48}$$

and is found to be consistent with the definition

$$\mathcal{H} = \langle\langle Tp,q\rangle\rangle - \mathcal{L}(p,Tp), \tag{2.49}$$

of (2.42) with

$$q = Tp. \tag{2.50}$$

At the stationary point the solution $(\widehat{p},\widehat{q})$ of the governing equation (2.3) satisfies Hamilton's equations

$$\begin{aligned} T\widehat{p} &= \frac{\partial\mathcal{H}(\widehat{p},\widehat{q})}{\partial\widehat{q}} = \widehat{q}, \qquad (H^-) && (2.51)\\ T^*\widehat{q} &= \frac{\partial\mathcal{H}(\widehat{p},\widehat{q})}{\partial\widehat{p}} = r, \qquad (H^+) && (2.52)\end{aligned}$$

which naturally splits the original governing equation into the two component system considered in section (2.1.1). The functional that is stationary at the solution of Hamilton's equations is

$$\begin{aligned} \overline{\mathcal{G}}(p,q) &= \langle\langle Tp,q\rangle\rangle - \mathcal{H}(p,q), && (2.53)\\ &= \langle\langle Tp,q\rangle\rangle - \frac{1}{2}\langle\langle q,q\rangle\rangle - \langle p,r\rangle, && (2.54)\end{aligned}$$

where p and q are considered independent variables. Hamilton's equations can be shown to be the natural conditions of the functional by equating the first variation of the functional to zero, giving

$$\begin{aligned} 0 &= \delta\overline{\mathcal{G}}(p,q), && (2.55)\\ &= \langle\langle Tp-q,\delta q\rangle\rangle + \langle\delta p, T^*q-r\rangle, \qquad \forall\delta p,\delta q && (2.56)\end{aligned}$$

implying $p=\widehat{p}$, $q=\widehat{q}$ at the stationary point.

The functional $\overline{\mathcal{G}}(p,q)$ is found to be saddle-shaped. The saddle-shape is generated by the Lagrangian functional introducing the required concavity in (2.42).

2.2.3 A Saddle-Shaped Functional

A saddle-shaped functional has been generated by a quadratic functional of two functions possessing two distinct axes, intersecting at the stationary point, and on which the functional is convex and concave respectively. The axis on which the functional is concave is generated by the Lagrangian functional $\mathcal{L}(p, Tp)$ along the hyperline given by Hamilton's equation $Tp = q$. Along this hyperline the functionals $\mathcal{L}(p, Tp)$ and $\overline{\mathcal{G}}(p, q)$ assume the same values. This is best demonstrated by separating the Lagrangian component in the functional to obtain

$$\begin{aligned} \overline{\mathcal{G}}(p,q) &= \langle\!\langle Tp, q\rangle\!\rangle - \frac{1}{2}\langle\!\langle q, q\rangle\!\rangle - \langle p, r\rangle, &(2.57)\\ &= \frac{1}{2}\langle\!\langle Tp, Tp\rangle\!\rangle - \langle p, r\rangle - \frac{1}{2}\langle\!\langle Tp - q, Tp - q\rangle\!\rangle, &(2.58)\\ &= \mathcal{L}(p, Tp) - \frac{1}{2}\langle\!\langle Tp - q, Tp - q\rangle\!\rangle, &(2.59)\end{aligned}$$

the first term of which is known to be concave and the second zero on the hyperline $Tp = q$.

The axis on which the functional is found to be convex is along the hyperline given by Hamilton's equation $T^*q = r$. This is demonstrated by writing the functional as

$$\begin{aligned} \overline{\mathcal{G}}(p,q) &= \langle\!\langle Tp, q\rangle\!\rangle - \frac{1}{2}\langle\!\langle q, q\rangle\!\rangle - \langle p, r\rangle, &(2.60)\\ &= \frac{1}{2}\langle\!\langle q, q\rangle\!\rangle + \langle\!\langle p, T^*q - r\rangle\!\rangle, &(2.61)\end{aligned}$$

the first term of which is convex by inspection and the second zero on the hyperline $T^*q = r$. The axes given by the hyperlines $Tp - q$ and $T^*q = r$ intersect at the stationary point $(\widehat{p}, \widehat{q})$ as given by Hamilton's equations (2.51) and (2.52) and we conclude that the functional $\overline{\mathcal{G}}(p, q)$ is endowed with a saddle-shaped topology. However, note that the axes considered are not necessarily the axes of maximum curvature.

The functionals $\overline{\mathcal{G}}(p, q)$ and $\mathcal{G}(p, q)$ of (2.4) and (2.54) differ only in sign and we choose to work with $\mathcal{G}(p, q)$ so that the stationary value has the same sign as the quantity of interest. The functional $\overline{\mathcal{G}}(p, q)$ was created as a result of adhering to convention with regard to the concavity of the Lagrangian. Contrary to convention the 'configuration coordinates' are labelled p and the 'conjugate momenta' are labelled q. This labelling is done for the benefit of the example in section (2.4.1) in which the function p will represent a pressure field and q a fluid flux, for which the labels are naturally suited.

Choosing to work with the functional $\mathcal{G}(p,q)$ rather than $\overline{\mathcal{G}}(p,q)$ implies that the functional switches convexity but retains the saddle-shaped topology.

2.3 Properties of the Stationary Value

The stationary value has already been associated with the quantity of physical interest that we are seeking to bound. In addition to being able to bound the stationary value, approximating the quantity of interest using functionals that are stationary at this point also yields advantageous accuracy properties.

2.3.1 Second Order Accurate Stationary Values

The additional accuracy present in the approximations to the stationary value is due to the stationary nature of the functional at the point $(\widehat{p},\widehat{q})$. The approximation to the stationary value is second order accurate with respect to first order errors in the approximation of the stationary point. The extra order of accuracy achieved is demonstrated by considering a Taylor series expansion around the analytic stationary value, utilising the property that the first derivative of the functional vanishes at the stationary point. Therefore writing $p_h = \widehat{p} + \epsilon$, where ϵ is error incurred in the approximation $p_h \approx \widehat{p}$,

$$\begin{aligned} \mathcal{G}^-(p_h) = \mathcal{G}^-(\widehat{p}+\epsilon) &= \mathcal{G}^-(\widehat{p}) + \frac{\partial \mathcal{G}^-(\widehat{p})}{\partial p}\epsilon + \frac{1}{2}\frac{\partial^2 \mathcal{G}^-(\widehat{p})}{\partial p^2}\epsilon^2 + h.o.t., && (2.62)\\ &= \mathcal{G}^-(\widehat{p}) + O(\epsilon^2) + h.o.t., && (2.63)\end{aligned}$$

and hence the bound exhibits superconvergent properties. The additional order of accuracy obtained in approximating the stationary value is schematically illustrated in figure 2.2. The upper bound based on the functional $\mathcal{G}^+(q)$ displays identical behaviour through a similar argument. In practice the accuracy of the quantity of interest will depend on the numerical method employed. For a finite element solution the order of accuracy will depend on the order of the polynomial. The error in the quantity of interest was investigated by constructing approximate solutions of the Poisson equation

$$\begin{aligned} -\frac{d^2\widehat{p}(x)}{dx^2} &= 1 \qquad 0 \le x \le 1, && (2.64)\\ \widehat{p}(0) = \widehat{p}(1) &= 0. && (2.65)\end{aligned}$$

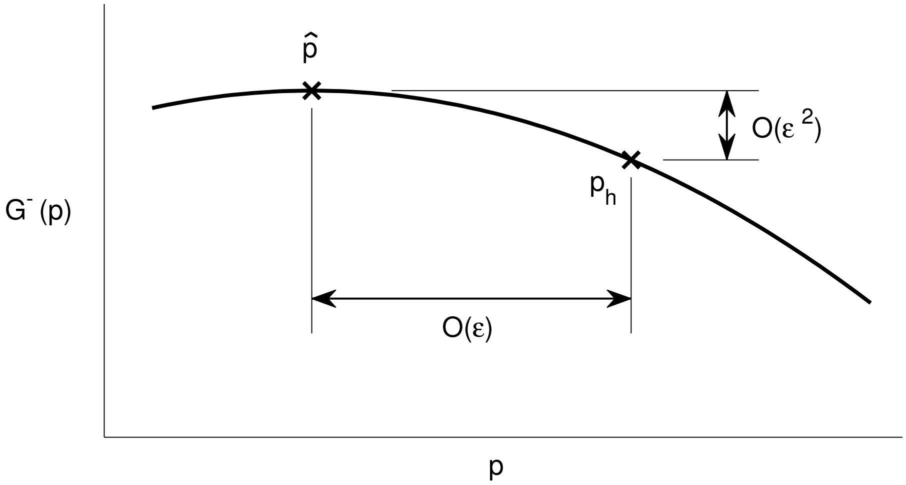

Figure 2.2: Superconvergence property of the lower bound

with the quantity of interest defined as the

$$\Theta(\widehat{p}) = \int_0^1 \widehat{p}\, dx. \tag{2.66}$$

The approximate solution $p_h \approx \widehat{p}$ and lower bounds on the quantity of interest were found using the maximum principle. The solutions p_h were constructed with linear finite elements and a Fourier sine expansion, the results obtained are plotted in figures 2.3 and 2.4 respectively.

From the result obtained using the linear finite element method, figure 2.3, the convergence of the solution in the L_2 norm is found to be $O(h^2)$, whilst the convergence of the solution in the energy norm is found to be an order of h less, where h is the element size. These results agree with the convergence theory for finite elements [8, 26]. However, the order of accuracy of the approximation to the quantity of interest is equal to the order of accuracy of the method in the energy norm, squared. This suggests that in general a greater return on computational effort can be achieved by employing higher order elements, and in models where this is practical quadratic elements will therefore be favoured. Quadratic elements could not be employed in this example however, as they are capable of representing the analytic solution. The convergence results using the Fourier method, figure 2.4, alludes to the possiblity of obtaining increased accuracy in the stationary value with the convergence rate of this quantity being greater than that of the solution in the L_2 norm.

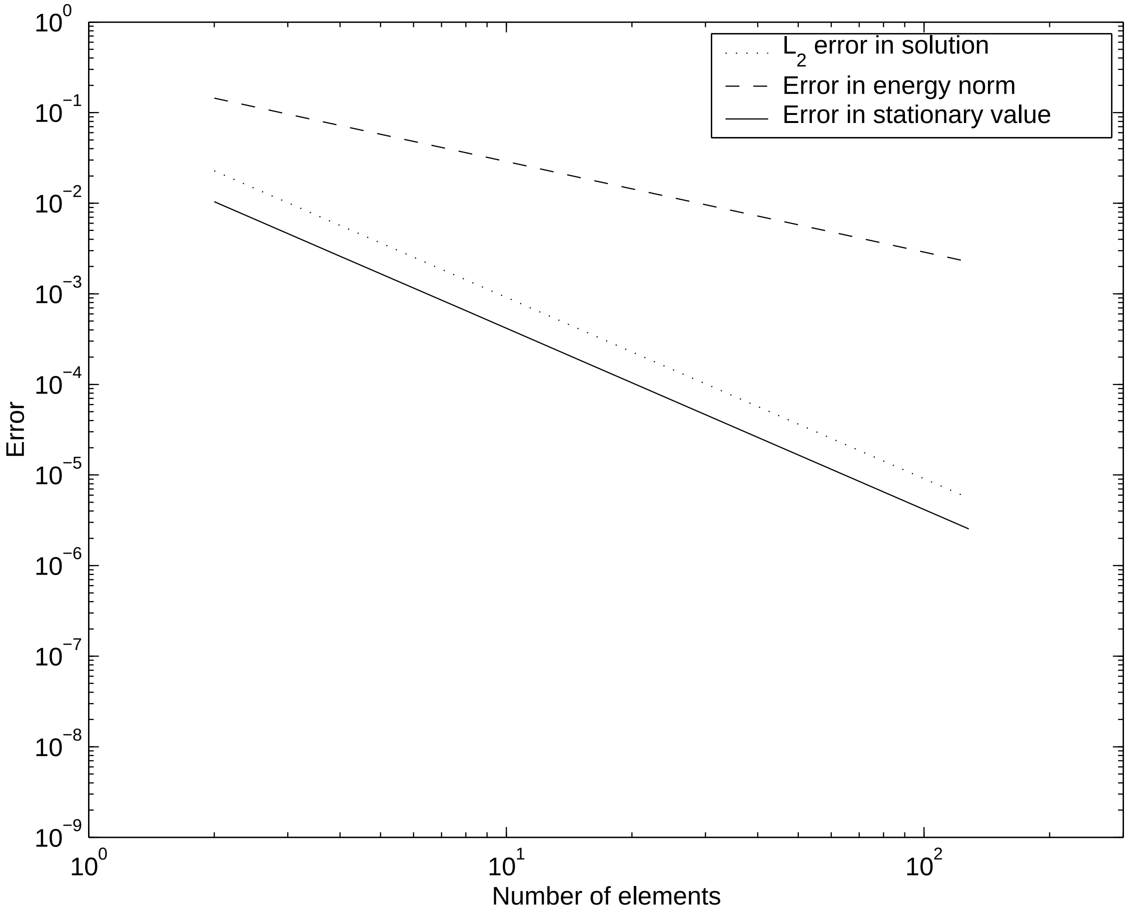

Figure 2.3: Convergence properties of the finite element method

2.3.2 The Twinning Method

In the preceding theory the quantity of interest has been considered to be of the form

$$\Theta(\widehat{p}) = \langle \widehat{p}, r \rangle, \tag{2.67}$$

$$= \langle\!\langle T\widehat{p}, T\widehat{p} \rangle\!\rangle, \tag{2.68}$$

where $\widehat{p}$ is the solution of the governing equation

$$T^*T\widehat{p} = r, \tag{2.69}$$

and r is both the forcing of the governing equation and the weight in the quantity of interest. This type of problem is referred to as self-dual. As a result of the self-duality, the quantity of interest is a positive quantity by (2.68) and can be directly identified with the stationary value of the functional $\mathcal{G}(\widehat{p}, \widehat{q})$ and the bounds

$$2\mu^- \leq \Theta(\widehat{p}) \leq 2\mu^+, \tag{2.70}$$

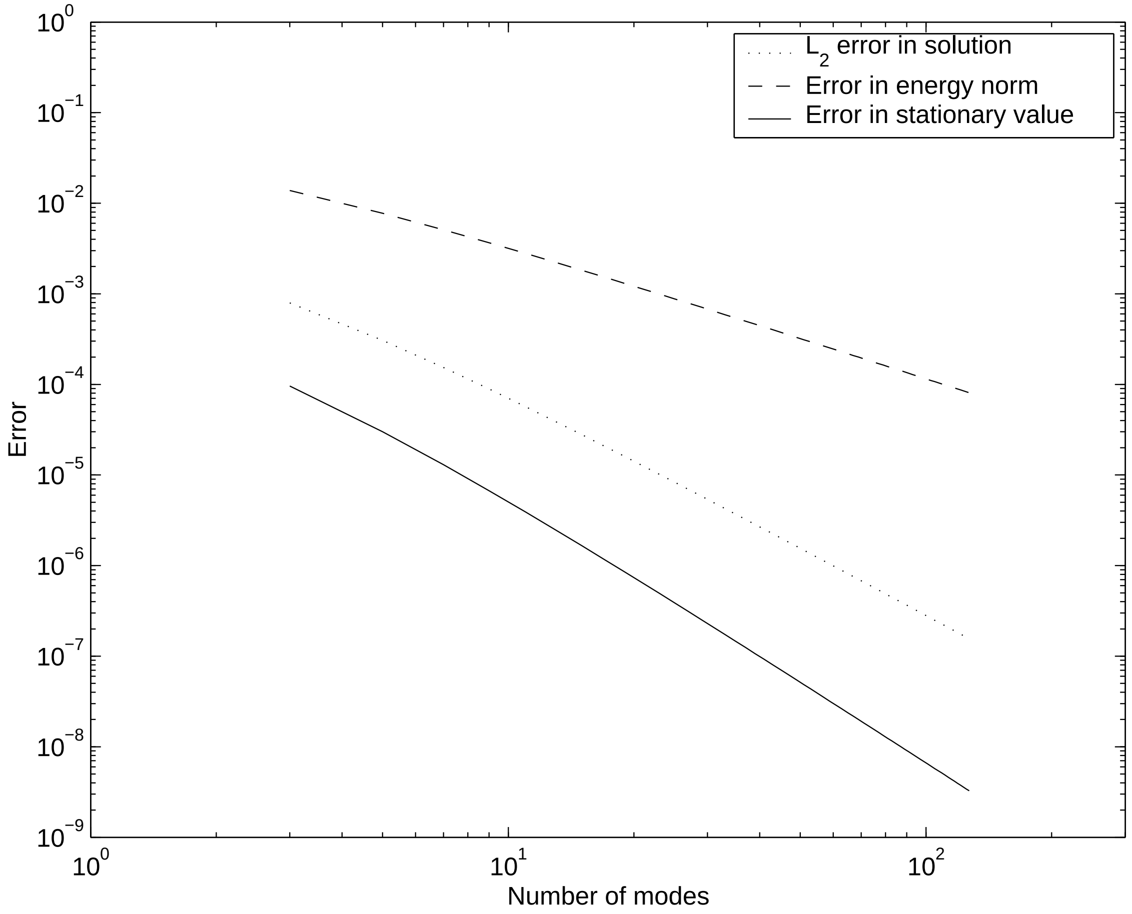

Figure 2.4: Convergence properties of the Fourier method

which have been found to be superconvergent.

In general however, the problem may not be self-dual and instead the quantity of interest may be the projection of the analytic solution with a completely different function t, (where t is not a constant multiple of r). In this situation the dual problem is required, forced by the function t. Superconvergent bounds on the quantity of interest can then be obtained using the 'twinning' method as described below.

2.3.3 The Dual Problem

The twinning method is implemented when the problem is not self-dual. The lack of self-duality requires a separate dual problem to be formulated. The pair of primal and dual problems is then

$$T^*T\widehat{u} \quad = \quad s \qquad \text{primal problem,} \tag{2.71}$$

$$T^*T\widehat{v} \;=\; t \qquad \text{dual problem,} \tag{2.72}$$

and the quantity of interest is

$$\begin{aligned}\Theta(\widehat{u}) &= \langle \widehat{u}, t\rangle, &(2.73)\\ &= \langle\!\langle T\widehat{u}, T\widehat{v}\rangle\!\rangle. &(2.74)\end{aligned}$$

The quantity of interest (2.74) is now no longer a positive quantity but can be decoupled into the difference of two positive quantities by 'twinning', in effect considering the difference of two squares,

$$\begin{aligned}\Theta(\widehat{u}) &= \langle\!\langle T\widehat{u}, T\widehat{v}\rangle\!\rangle, &(2.75)\\ &= \frac{1}{4}\langle\!\langle\, T(\widehat{u}+\widehat{v}), T(\widehat{u}+\widehat{v})\,\rangle\!\rangle - \frac{1}{4}\langle\!\langle\, T(\widehat{u}-\widehat{v}), T(\widehat{u}-\widehat{v})\,\rangle\!\rangle. &(2.76)\end{aligned}$$

Bounds on the quantity of interest can then calculated by introducing the transformations

$$\begin{aligned}\widehat{p}_1 &= \widehat{u}+\widehat{v} \qquad r_1 = s+t, &(2.77)\\ \widehat{p}_2 &= \widehat{u}-\widehat{v} \qquad r_2 = s-t, &(2.78)\end{aligned}$$

and applying the minimum and maximum principles to the pair of *self-dual* problems

$$\begin{aligned}T^*T\widehat{p}_1 &= r_1, &(2.79)\\ T^*T\widehat{p}_2 &= r_2. &(2.80)\end{aligned}$$

From the pair of problems (2.79) and (2.80) we obtain the bounds

$$2\mu_i^- \leq \langle\!\langle T\widehat{p}_i, T\widehat{p}_i\rangle\!\rangle \leq 2\mu_i^+, \tag{2.81}$$

say, and an expression for the quantity of interest in terms of the self-dual solutions p_1 and p_2,

$$\begin{aligned}\Theta(\widehat{u}) &= \langle\!\langle T\widehat{u}, T\widehat{v}\rangle\!\rangle, &(2.82)\\ &= \frac{1}{4}\langle\!\langle T\widehat{p}_1, T\widehat{p}_1\rangle\!\rangle - \frac{1}{4}\langle\!\langle T\widehat{p}_2, T\widehat{p}_2\rangle\!\rangle. &(2.83)\end{aligned}$$

The quantity of interest is therefore bounded by

$$\frac{1}{2}(\mu_1^- - \mu_2^+) \leq \Theta(\widehat{u}) \leq \frac{1}{2}(\mu_1^+ - \mu_2^-). \tag{2.84}$$

The bounds μ_i^- and μ_i^+ are both superconvergent in the sense described in section 2.3.1 and hence the order of accuracy extends to the upper and lower bounds on $\Theta(\widehat{u})$.

Having constructed the theory in a general operator notation the attention is turned to a particular application of the method in the context of oil reservoir simulation. The model considered is governed by a self-adjoint operator and upper and lower bounds on quantities of physical interest are available and computable. The combination of the quantity of interest required and the boundary conditions considered renders the problem effectively self-dual and the need to implement the twinning method is avoided.

2.4 Oil Reservoir Simulation

Oil reservoirs are complex systems involving multiple fluid phases and a broad range of time and length scales. The combination of these attributes makes it prohibitively complex to model the reservoirs in full. In order to obtain predictions to improve oil recovery strategies many simplified models have been constructed. A typical model is the black oil model in which the saturations of the oil, water and gas phase are modelled. The three phases are assumed to be miscible at a macroscopic level, with the associated velocities obeying Darcy's law. Darcy's law states that the fluid flux is proportional to the pressure gradient of the phase with the constant of proportionality defined to be the permeability of the medium. The black oil model is non-linear and one source of the non-linearities is the dependence of the permeability on the phase saturations. The permeability characteristics are obtained from physical testing and typical curves can be found in Mayer-Gürr [33]. The formulation of the black oil model can be found in the books of Muskat [35] and Amyx [1]. The reservoir model can be simplified by reducing the number of phases.

The inherent complexity in the reservoir models forces numerical solutions. A finite difference method is employed largely in the industry and its application to reservoir simulations is introduced in Aziz [4]. More recently weak solutions have been investigated, see e.g. the work of Ewing and Chen [10, 11], and in addition a growing number of industrial simulations are being based on the streamline method. In the streamline method the fluid paths are first approximated and then the phase components are advected along the paths using 1-D approximations. Consequently, the majority of the computations in the streamline method involve calculating the solution of many 1-D local problems, which is relatively quick, and an overall reduction in the simulation cost is achieved. The streamline method will not be considered in this research but an introduction can be found in the Society of Petroleum Engineers reports including [6],[24] and [39].

2.4.1 A Simple Oil Reservoir Model

The simplest model of an oil reservoir is found by reducing the number of fluid phases to one and consider single phase incompressible flow obeying Darcy's Law. The fluid flux $\mathbf{q}(\mathbf{x})$ through the medium is then proportional to the pressure gradient $\nabla p(\mathbf{x})$ with the permeability $\lambda(\mathbf{x})$ being the constant of proportionality. As the flow is considered to consist of a single phase the saturation is constant in the reservoir and the non-linearities associated with the phase saturations are removed. The incompressibility condition implies that the divergence of the flux $\mathbf{q}(\mathbf{x})$ is zero. In addition to the governing equations within the reservoir, boundary conditions are also required. On the boundary segment Γ^- the pressure will be specified and on the remaining segment Γ^+ the outward normal flux will be specified. The governing equation is then the diffusion equation with Dirichlet and Neumann boundary conditions, which can be written as

$$\begin{aligned}
\widehat{\mathbf{q}}(\mathbf{x}) &= -\lambda(\mathbf{x})\nabla\widehat{p}(\mathbf{x}) \quad \text{in } \Omega, && (2.85)\\
\nabla\cdot\widehat{\mathbf{q}}(\mathbf{x}) &= 0 \quad \text{in } \Omega, && (2.86)\\
\widehat{p}(\mathbf{x}) &= f(\mathbf{x}) \quad \text{on } \Gamma^-, && (2.87)\\
\widehat{\mathbf{q}}(\mathbf{x})\cdot\mathbf{n}(\mathbf{x}) &= g(\mathbf{x}) \quad \text{on } \Gamma^+, && (2.88)
\end{aligned}$$

where $\mathbf{n}(\mathbf{x})$ denotes the outward unit normal vector on Γ^+, and $\Gamma^-\cup\Gamma^+$ is the complete boundary of the given domain Ω. The permeability tensor λ is known to be symmetric, see [15], and positive definite in order that the flux $\widehat{\mathbf{q}}$ always has a positive component in the direction $-\nabla\widehat{p}$. These property ensure that λ^{-1}, $\lambda^{-\frac{1}{2}}$ and $\lambda^{\frac{1}{2}}$ also exist, are bounded, and real valued. Initially we consider $\lambda(\mathbf{x})$ to be a smooth C^1 function and as a result the pressure and flux components are also smooth and continuous, belonging to C^2 and C^1 respectively, in the domain.

In addition to a being a prototype oil reservoir model the diffusion equation is used to model groundwater flow, which is generally considered to be a true single phase problem. The solution can also be linked to an advection-diffusion equation in which the concentration of a substance is advected by the flow field $\mathbf{q}$, as well as diffusing across the domain. This pair of equations can then be used to model pollutants and tracers in fluid flows. The diffusion equation is of course also an appropriate model for heat conduction, and therefore the equations in general have other application aside from oil reservoir simulation.

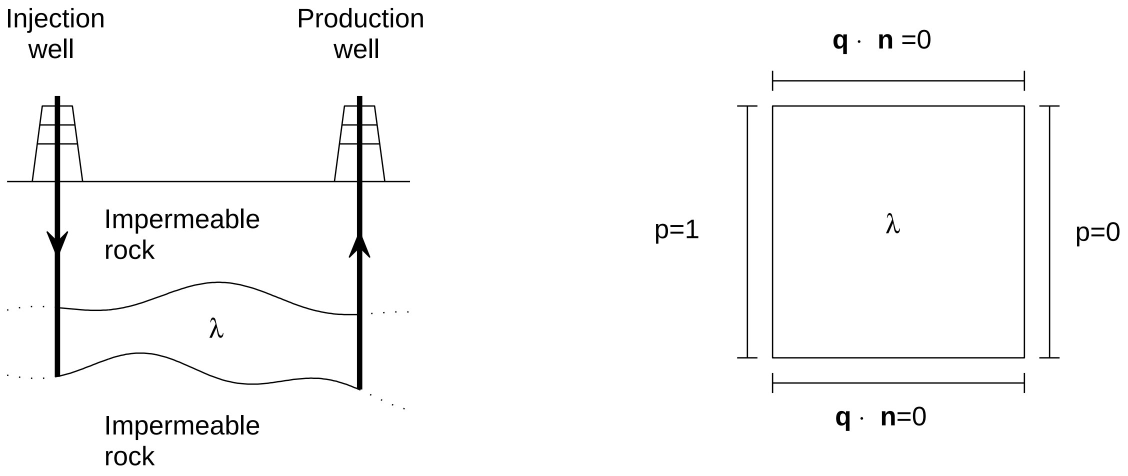

Figure 2.5: Model A

Two 2D model problems will be considered to illustrate the ability of the method to obtain upper and lower bounds on a quantity of physical interest. The physical problems and model abstractions are shown in figures 2.5 and 2.6. Model A represents flow along a porous seam sandwiched between impermeable layers. The flow is driven by the horizontal pressure gradient and the flow passing through the inlet and outlet is interpreted as flowing through the injection and production wells respectively. Model B forms a quarter of the '5-spot' problem in which a production well is located centrally and four injection wells are located in a surrounding square to form the pattern for five found on a die. The '5-spot' problem models a reservoir response, in the horizontal plane, to 'water flooding', in which water is pumped into the reservoir to displace the remaining oil. The no-flow boundary conditions in Model B represents a symmetry assumption on the solution and also tacitly assumes that the permeability field is symmetric. In

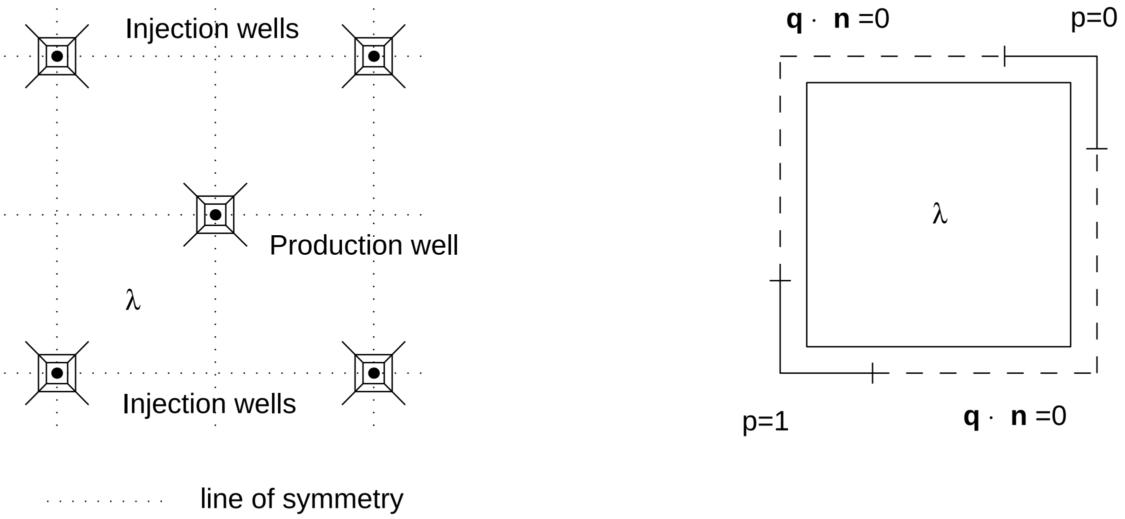

Figure 2.6: Model B

both models the flow passing through the inlet and outlet is interpreted as flow through the wells. The wells are modelled as segments of the boundary rather than as point sources and sinks as in industrial simulations a dedicated well model would model the flow in proximity to the well and be coupled to the rest of the reservoir simulation over a portion of the boundary. The quantities of interest in these simulations are the outflows from the production wells, since these represent the quantity of oil available for trading.

The reservoir model described fits into the framework of the previous chapters, since the diffusion operator is known to be self-adjoint. The self-adjointness of the diffusion operator is demonstrated by the divergence theorem

$$\begin{aligned} \langle\!\langle \mathbf{q}, Tp \rangle\!\rangle &= \iint_\Omega \mathbf{q}\cdot(\nabla p)\; d\Omega - \int_{\Gamma^-} \mathbf{q}\cdot(\mathbf{n}p)\; d\Gamma && (2.89)\\ &= -\iint_\Omega (\nabla\cdot\mathbf{q})p\; d\Omega + \int_{\Gamma^+} (\mathbf{q}\cdot\mathbf{n})p\; d\Gamma && (2.90)\\ &= \langle T^*\mathbf{q}, p\rangle && (2.91) \end{aligned}$$

(cf. (2.2)), where the inner products include contributions from the boundary. The operators T and T^* can then be defined as

$$Tp = \begin{cases} \nabla p & \text{in } \Omega \\ -\mathbf{n}p & \text{on } \Gamma^- \end{cases} \qquad T^*\mathbf{q} = \begin{cases} -\nabla\cdot\mathbf{q} & \text{in } \Omega \\ \mathbf{q}\cdot\mathbf{n} & \text{on } \Gamma^+ \end{cases} \qquad (2.92)$$

where we note that the Dirichlet boundary segment Γ^- is associated with the operator T and the Neumann boundary segment Γ^+ is associated with the adjoint operator T^*.

2.4.2 The Diffusion Functional

The diffusion functional $\mathcal{G}(p, \mathbf{q}; \lambda)$ is constructed in the same manner as in the general framework, that is

$$\mathcal{G}(p, \mathbf{q}; \lambda) = \mathcal{H}(p, \mathbf{q}; \lambda) - \langle\!\langle Tp, \mathbf{q} \rangle\!\rangle, \tag{2.93}$$

where the functions appearing after the semicolon are assumed known. The required partial derivatives of the Hamiltonian $\mathcal{H}(p, \mathbf{q}; \lambda)$ are then

$$\frac{\partial \mathcal{H}(\widehat{p}, \widehat{\mathbf{q}}; \lambda)}{\partial \widehat{\mathbf{q}}} = \begin{cases} \lambda^{-1}\widehat{\mathbf{q}} & \text{in } \Omega \\ \mathbf{n}f & \text{on } \Gamma^- \end{cases} \qquad H^- \tag{2.94}$$

$$\frac{\partial \mathcal{H}(\widehat{p}, \widehat{\mathbf{q}}; \lambda)}{\partial \widehat{p}} = \begin{cases} 0 & \text{in } \Omega \\ -g & \text{on } \Gamma^+ \end{cases} \qquad H^+ \tag{2.95}$$

generating the Hamiltonian

$$\mathcal{H}(p, \mathbf{q}; \lambda) = \frac{1}{2}\iint_\Omega \lambda^{-1}\mathbf{q}\cdot\mathbf{q}\, d\Omega + \int_{\Gamma^-} \mathbf{q}\cdot\mathbf{n}f\, d\Gamma - \int_{\Gamma^+} pg\, d\Gamma, \tag{2.96}$$

(cf. (2.48)) and the functional

$$\mathcal{G}(p, \mathbf{q}; \lambda) = \iint_\Omega \left\{\frac{1}{2}\lambda^{-1}\mathbf{q}\cdot\mathbf{q} + \mathbf{q}\cdot\nabla p\right\} d\Omega - \int_{\Gamma^-} (p-f)\mathbf{q}\cdot\mathbf{n}\, d\Gamma - \int_{\Gamma^+} pg\, d\Gamma. \tag{2.97}$$

It can be shown that the functional is stationary at the solution of the diffusion equation and boundary conditions, since the first variation of $\mathcal{G}(p, \mathbf{q}; \lambda)$ is

$$\begin{aligned} \delta\mathcal{G}(p, \mathbf{q}; \lambda) &= \iint_\Omega \left\{\delta p(-\nabla\cdot\mathbf{q}) + \delta\mathbf{q}\cdot\left(\lambda^{-1}\mathbf{q} + \nabla p\right)\right\} d\Omega & (2.98) \\ &\quad - \int_{\Gamma^-} (p-f)\delta\mathbf{q}\cdot\mathbf{n}\, d\Gamma + \int_{\Gamma^+} \delta p(\mathbf{q}\cdot\mathbf{n} - g)\, d\Gamma, & (2.99) \end{aligned}$$

and therefore the functional is stationary for any variations in p and $\mathbf{q}$ if and only if the diffusion equation and boundary conditions are satisfied. Thus the problem of determining the solution of (2.85) to (2.88) is equivalent to finding the functions $\widehat{p}$ and $\widehat{\mathbf{q}}$ which make $\mathcal{G}(p, \mathbf{q}; \lambda)$ stationary.

The stationary value of the functional is found by substituting the stationary conditions into the functional, to give

$$\begin{aligned} \mathcal{G}(\widehat{p}, \widehat{\mathbf{q}}; \lambda) &= \iint_\Omega \left\{\frac{\lambda^{-1}}{2}\widehat{\mathbf{q}}\cdot\widehat{\mathbf{q}} + \widehat{\mathbf{q}}\cdot\nabla\widehat{p}\right\} d\Omega - \int_{\Gamma^-} (\widehat{p}-f)\widehat{\mathbf{q}}\cdot\mathbf{n}\, d\Gamma - \int_{\Gamma^+} \widehat{p}g\, d\Gamma, \\ &= \frac{1}{2}\iint_\Omega \widehat{\mathbf{q}}\cdot\nabla\widehat{p}\, d\Omega - \int_{\Gamma^+} \widehat{p}g\, d\Gamma, \\ &= \frac{1}{2}\int_{\Gamma^-} f\widehat{\mathbf{q}}\cdot\mathbf{n}\, d\Gamma - \frac{1}{2}\int_{\Gamma^+} \widehat{p}g\, d\Gamma. \end{aligned} \tag{2.100}$$

The stationary value is therefore a weighted integral of the flux over the boundary which is an important physical quantity representing well production in a reservoir model. From the boundary conditions specified in the model problems, namely

$$f = \begin{cases} 1 & \text{at the inlet,} \\ 0 & \text{at the outlet,} \end{cases} \tag{2.101}$$

$$g = 0, \tag{2.102}$$

we obtain

$$\mathcal{G}(\widehat{p}, \widehat{\mathbf{q}}; \lambda) = \frac{1}{2}\int_{\Gamma^-} f\widehat{\mathbf{q}} \cdot \mathbf{n}\, d\Gamma - \frac{1}{2}\int_{\Gamma^+} \widehat{p}g\, d\Gamma, \tag{2.103}$$

$$= \frac{1}{2}\int_{In} \widehat{\mathbf{q}} \cdot \mathbf{n}\, d\Gamma, \tag{2.104}$$

$$= -\frac{1}{2}\int_{Out} \widehat{\mathbf{q}} \cdot \mathbf{n}\, d\Gamma. \tag{2.105}$$

The problem is therefore effectively self-dual with the factor of minus a half separating the stationary value of the functional and the quantity of interest. For the benefits of simplicity the dual problems is not introduced and bounds will be calculated on the stationary value of the functional. Naturally, bounds on the quantity of interest, the production well outflow, can be found by multiplying the results by minus two. If the reservoir model had included multiple production wells the dual problem could have been used to specify the flux out of a particular well or combination of wells, in which case the method of twinning would have been need.

2.4.3 The Maximum and Minimum Principles

The upper and lower bounds on the stationary value of the functional are obtained via maximum and minimum principles analogous to those derived in section 2.1.3. The free principle is constrained to satisfy one of the pair of Hamilton's equation and the resulting functional is found to have the desired convexity. Constraining $\mathcal{G}(p, \mathbf{q}; \lambda)$ to satisfy Hamilton's equation H^- we require the subset of natural conditions

$$\mathbf{q} = -\lambda \nabla p \quad \text{in } \Omega, \tag{2.106}$$

$$p = f \quad \text{on } \Gamma^-, \tag{2.107}$$

to hold. Substituting the constraints (2.106) and (2.107) into $\mathcal{G}(p,\mathbf{q};\lambda)$ we obtain the maximum principle

$$\mathcal{G}^{-}(p;\lambda) = -\frac{1}{2}\iint_{\Omega} \lambda\nabla p\cdot\nabla p\, d\Omega - \int_{\Gamma^{+}} pg\, d\Gamma \tag{2.108}$$

which has (2.86) and (2.88) as natural conditions. The maximum principle retains the same stationary value as the free principle, since the natural conditions of $\mathcal{G}^{-}(p;\lambda)$ along with the constraints imposed, (2.106) and (2.107), are equivalent to the natural conditions of the free principle.

A lower bound on $\mathcal{G}(\widehat{p},\widehat{\mathbf{q}};\lambda)$ is found by comparing the stationary value of $\mathcal{G}^{-}(\widehat{p};\lambda)$ and $\mathcal{G}^{-}(p;\lambda)$, where $\widehat{p}$ is the analytic solution satisfying the full set of stationary conditions, and p is *any* function satisfying the constraint (2.107), since

$$\begin{aligned}
\mathcal{G}^{-}(\widehat{p};\lambda) - \mathcal{G}^{-}(p;\lambda) &= \frac{1}{2}\iint_{\Omega}\{\lambda\nabla p\cdot\nabla p - \lambda\nabla\widehat{p}\cdot\nabla\widehat{p}\}\, d\Omega + \int_{\Gamma^{+}} g(p-\widehat{p})\, d\Gamma,\\
&= \frac{1}{2}\iint_{\Omega}\{\lambda\nabla p\cdot\nabla p - \lambda\nabla\widehat{p}\cdot\nabla\widehat{p}\}\, d\Omega - \int_{\Gamma} \lambda\nabla\widehat{p}(p-\widehat{p})\cdot\mathbf{n}\, d\Gamma,\\
&= \frac{1}{2}\iint_{\Omega}\{\lambda\nabla p\cdot\nabla p - \lambda\nabla\widehat{p}\cdot\nabla\widehat{p} - 2\lambda\nabla\widehat{p}\cdot\nabla(p-\widehat{p})\}\, d\Omega,\\
&= \frac{1}{2}\iint_{\Omega}(\lambda^{\frac{1}{2}}\nabla p - \lambda^{\frac{1}{2}}\nabla\widehat{p})^{2} d\Omega,\\
&\geq 0. \qquad (2.109)
\end{aligned}$$

(cf. (2.17)). Thus for any function p satisfying the constraints (2.106) and (2.107), the value of the functional $\mathcal{G}^{-}(p;\lambda)$ cannot exceed the exact stationary value $\mathcal{G}(\widehat{p},\widehat{\mathbf{q}};\lambda)$.

Similarly the minimum principle is defined by constraining $\mathcal{G}(p,\mathbf{q};\lambda)$ strongly by the subset of natural conditions, H^{+},

$$\begin{aligned}
\nabla\cdot\mathbf{q} &= 0 \quad \text{in } \Omega, \qquad (2.110)\\
\mathbf{q}\cdot\mathbf{n} &= g \quad \text{on } \Gamma^{+}. \qquad (2.111)
\end{aligned}$$

Substituting the constraints (2.110) and (2.111) into $\mathcal{G}(p,\mathbf{q};\lambda)$ we obtain the functional

$$\mathcal{G}^{+}(\mathbf{q};\lambda) = \frac{1}{2}\iint_{\Omega}\lambda^{-1}\mathbf{q}\cdot\mathbf{q}\, d\Omega + \int_{\Gamma^{-}} f\mathbf{q}\cdot\mathbf{n}\, d\Gamma, \tag{2.112}$$

which has (2.85) and (2.87) as natural conditions. An upper bound on $\mathcal{G}(\widehat{p},\widehat{\mathbf{q}};\lambda)$ is found by comparing the stationary value of $\mathcal{G}^{+}(\widehat{\mathbf{q}};\lambda)$ and $\mathcal{G}^{+}(\mathbf{q};\lambda)$, where $\widehat{\mathbf{q}}$ is the analytic solution satisfying the full set of stationary conditions, and $\mathbf{q}$ is *any* function satisfying (2.110) and (2.111), since

$$\begin{aligned}
\mathcal{G}^+(\widehat{\mathbf{q}};\lambda) - \mathcal{G}^+(\mathbf{q};\lambda) &= \frac{1}{2}\iint_\Omega \left\{\lambda^{-1}\widehat{\mathbf{q}}\cdot\widehat{\mathbf{q}} - \lambda^{-1}\mathbf{q}\cdot\mathbf{q}\right\} d\Omega + \int_{\Gamma^-} f(\widehat{\mathbf{q}} - \mathbf{q})\cdot\mathbf{n}\, d\Gamma, \\
&= \frac{1}{2}\iint_\Omega \left\{\lambda^{-1}\widehat{\mathbf{q}}\cdot\widehat{\mathbf{q}} - \lambda^{-1}\mathbf{q}\cdot\mathbf{q}\right\} d\Omega + \int_{\Gamma} \widehat{p}(\widehat{\mathbf{q}} - \mathbf{q})\cdot\mathbf{n}\, d\Gamma, \\
&= \frac{1}{2}\iint_\Omega \left\{\lambda^{-1}\widehat{\mathbf{q}}\cdot\widehat{\mathbf{q}} - \lambda^{-1}\mathbf{q}\cdot\mathbf{q} + 2(\widehat{\mathbf{q}} - \mathbf{q})\cdot\nabla\widehat{p}\right\} d\Omega, \\
&= -\frac{1}{2}\iint_\Omega (\lambda^{-\frac{1}{2}}\widehat{\mathbf{q}} - \lambda^{-\frac{1}{2}}\mathbf{q})^2 d\Omega, \\
&\leq 0, \qquad (2.113)
\end{aligned}$$

(cf. (2.27)) which can be considered the dual of the inequality (2.109).

The finite dimensional subspaces in which approximations to the stationary value of the functional are sought can then be defined. The space H_h^- in which the stationary value of the functional $\mathcal{G}^-(p_h;\lambda)$ is found is defined by the span of the basis functions ϕ_i which must be capable of satisfying the constraints of H^- namely:

- able to represent the Dirichlet boundary condition f
- once differentiable

Expanding the function in terms of the basis functions

$$p_h = \sum_{j=1}^{N} p_j\phi_j \qquad (2.114)$$

where at least one of the coefficients p_j is known from the necessary Dirichlet boundary condition, the normal equations are then

$$\iint_\Omega \lambda\nabla\phi_i\cdot\nabla\sum_{j=1}^{N} p_j\phi_j\, d\Omega + \int_{\Gamma^+} g\sum_{j=1}^{N} p_j\phi_j\, d\Gamma = 0 \qquad i = 1,\cdots,M \qquad (2.115)$$

to be solved for the remaining $M < N$ unknown coefficients. The set of equations (2.115) form an $M\times M$ positive definite matrix equation of the form

$$\mathsf{K}\mathbf{p} + \mathbf{g} = \mathbf{b}, \qquad (2.116)$$

where K is the $M\times M$ positive definite symmetric matrix

$$\mathsf{K}_{ij} = \iint_\Omega \lambda(\mathbf{x})\nabla\phi_i\cdot\nabla\phi_j\, d\Omega, \qquad (2.117)$$

$\mathbf{p}$ is the vector of unknown coefficients, $\mathbf{g}$ is the term arising from the Neumann boundary condition

$$\mathbf{g}_i = \int_{\Gamma^+} g\phi_i \, d\Gamma, \tag{2.118}$$

and $\mathbf{b}$ is the forcing resulting from overwriting the $N - M$ known Dirichlet nodes.

The space H_h^+ in which the stationary value of the functional $\mathcal{G}^+(\mathbf{q}_h; \lambda)$ is sought must be capable of satisfying the remaining constraints:

- able to represent the Neumann boundary condition $\mathbf{q} \cdot \mathbf{n} = 0$
- divergence free

For the 2D problems considered this is achieved using the stream function Ψ_h and defining the flux as

$$\mathbf{q} = \frac{\partial \Psi_h}{\partial y}\mathbf{i} - \frac{\partial \Psi_h}{\partial x}\mathbf{j}. \tag{2.119}$$

The boundary condition on Ψ_h is then interpreted as setting the tangential derivative of Ψ_h equal to zero, which again must be representable by the basis ϕ_i. The stream function is uniquely defined by assigning $\Psi_h = 0$ at one node on the Neumann boundary segment. The discrete matrix system arising is then of the form (2.116). To apply the method in higher dimensions, divergence-free finite elements could be employed of the type constructed by Raviart [42] or Gustafson [20].

2.4.4 Discontinuous Permeability Data

The description of the rock permeability by a piecewise constant function is a feature of many current reservoir simulations [15]. The discontinuities in the permeability function enable the simulation to be regarded as a multiple domain problem with appropriate interfacial conditions. The required interfacial conditions for conservation are continuity of the flux normal to the permeability discontinuity and continuity of the pressure over the discontinuity. As a result of the permeability discontinuity, the solutions $\widehat{p}$ and $\widehat{\mathbf{q}}$ are no longer smooth functions but experience discontinuities along the surface on which the permeability is discontinuous and where the governing equation in differential form is no longer valid.

Consider a λ field with a single discontinuity along the line Γ_I. The discontinuity splits the domain into the components Ω_1 and Ω_2 and similarly decomposes the permeability data into two components λ_1 and λ_2. An illustration of the domain is shown in figure 2.7, where $\Gamma_1^- \cup \Gamma_1^+ \cup \Gamma_I = \Gamma_1$ is the boundary of the sub domain Ω_1 and similarly with the second sub domain. To obtain the required functional over the union of the sub-domains Ω_m a sum of functionals of the form $\mathcal{G}_M(p_m, \mathbf{q}_m; \lambda_m)$ is considered with additional terms defined at the interfaces to impose the required continuity conditions. The required multiple domain functional is then $\mathcal{G}_M(p, \mathbf{q}; \lambda)$,

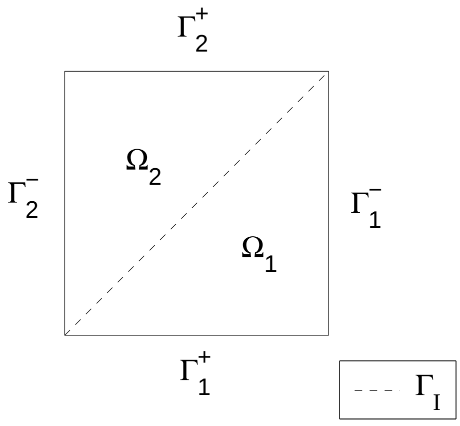

Figure 2.7: A domain with a permeability discontinuity

$$
\begin{aligned}
\mathcal{G}_M(p, \mathbf{q}; \lambda) \;=\;& \iint_{\Omega_1} \left\{ \frac{1}{2}\lambda_1^{-1}\mathbf{q}_1 \cdot \mathbf{q}_1 + \mathbf{q}_1 \cdot \nabla p_1 \right\} d\Omega + \iint_{\Omega_2} \left\{ \frac{1}{2}\lambda_2^{-1}\mathbf{q}_2 \cdot \mathbf{q}_2 + \mathbf{q}_2 \cdot \nabla p_2 \right\} d\Omega \\
& - \int_{\Gamma_1^-} (p_1 - f)\mathbf{q}_1 \cdot \mathbf{n}_1 \, d\Gamma - \int_{\Gamma_1^+} p_1 g \, d\Gamma - \int_{\Gamma_2^-} (p_2 - f)\mathbf{q}_2 \cdot \mathbf{n}_2 \, d\Gamma \\
& - \int_{\Gamma_2^+} p_2 g \, d\Gamma - \int_{\Gamma_I} \mathbf{q}_I \cdot (p_1 \mathbf{n}_1 + p_2 \mathbf{n}_2) \, d\Gamma \qquad (2.120)
\end{aligned}
$$

which has first variation

$$
\begin{aligned}
\delta\mathcal{G}_M(p, \mathbf{q}; \lambda) \;=\;& \iint_{\Omega_1} \left\{ \delta p_1(-\nabla \cdot \mathbf{q}_1) + \delta \mathbf{q}_1 \cdot \left(\lambda_1^{-1}\mathbf{q}_1 + \nabla p_1 \right) \right\} d\Omega \\
& + \iint_{\Omega_2} \left\{ \delta p_2(-\nabla \cdot \mathbf{q}_2) + \delta \mathbf{q}_2 \cdot \left(\lambda_2^{-1}\mathbf{q}_2 + \nabla p_2 \right) \right\} d\Omega \\
& - \int_{\Gamma_1^-} (p_1 - f)\delta\mathbf{q}_1 \cdot \mathbf{n}_1 \, d\Gamma + \int_{\Gamma_1^+} \delta p_1(\mathbf{q}_1 \cdot \mathbf{n}_1 - g) \, d\Gamma \\
& - \int_{\Gamma_2^-} (p_2 - f)\delta\mathbf{q}_2 \cdot \mathbf{n}_2 \, d\Gamma + \int_{\Gamma_2^+} \delta p_2(\mathbf{q}_2 \cdot \mathbf{n}_2 - g) \, d\Gamma \\
& + \int_{\Gamma_I} \{ \delta p_1 \mathbf{n}_1 \cdot (\mathbf{q}_1 - \mathbf{q}_I) + \delta p_2 \mathbf{n}_2 \cdot (\mathbf{q}_2 - \mathbf{q}_I) \} \, d\Gamma \\
& - \int_{\Gamma_I} \delta\mathbf{q}_I \cdot (p_1 \mathbf{n}_1 + p_2 \mathbf{n}_2) \, d\Gamma. \qquad (2.121)
\end{aligned}
$$

The natural conditions of the functional in the sub domains $m = 1, 2$ are therefore

$$\widehat{\mathbf{q}}_m = -\lambda_m \nabla \widehat{p}_m \quad \text{in } \Omega_m, \tag{2.122}$$

$$\nabla \cdot \widehat{\mathbf{q}}_m = 0 \quad \text{in } \Omega_m, \tag{2.123}$$

$$\widehat{p}_m = f \quad \text{on } \Gamma_m^-, \tag{2.124}$$

$$\widehat{\mathbf{q}}_m \cdot \mathbf{n}_m = g \quad \text{on } \Gamma_m^+, \tag{2.125}$$

and at the interface where $\mathbf{n}_1 = -\mathbf{n}_2 = \mathbf{n}_I$

$$\mathbf{q}_1 \cdot \mathbf{n}_I = \mathbf{q}_I \cdot \mathbf{n}_I \quad \text{on } \Gamma_I, \tag{2.126}$$

$$\mathbf{q}_2 \cdot \mathbf{n}_I = \mathbf{q}_I \cdot \mathbf{n}_I \quad \text{on } \Gamma_I, \tag{2.127}$$

$$p_1 = p_2 \quad \text{on } \Gamma_I, \tag{2.128}$$

as required. The global solutions $\widehat{p}$ and $\widehat{\mathbf{q}}$ are naturally composed of the solutions from the individual subdomains such that

$$\widehat{p}(\mathbf{x}) = \widehat{p}_m(\mathbf{x}) \quad \text{iff} \quad \mathbf{x} \in \Omega_m, \tag{2.129}$$

$$\widehat{\mathbf{q}}(\mathbf{x}) = \widehat{\mathbf{q}}_m(\mathbf{x}) \quad \text{iff} \quad \mathbf{x} \in \Omega_m. \tag{2.130}$$

Crucially the maximum and minimum principles obtained in section (2.4.2) also survive with the minimum of modifications. A maximum principle is established by constraining the free principle to satisfy (2.106) and (2.107), and in addition constraining the pressure to be continuous across the interface. The new set of constraints is then

$$\mathbf{q}_m = -\lambda_m \nabla p_m \quad \text{in } \Omega_m, \tag{2.131}$$

$$p_m = f \quad \text{on } \Gamma_m^-, \tag{2.132}$$

$$p_1 = p_2 \quad \text{on } \Gamma_I, \tag{2.133}$$

which when substituted into the free principle produces the functional

$$\mathcal{G}_M^-(p;\lambda) = -\frac{1}{2}\iint_{\Omega_1} \lambda_1 \nabla p_1 \cdot \nabla p_1 \, d\Omega - \int_{\Gamma_1^+} p_1 g \, d\Gamma \tag{2.134}$$

$$-\frac{1}{2}\iint_{\Omega_2} \lambda_2 \nabla p_2 \cdot \nabla p_2 \, d\Omega - \int_{\Gamma_2^+} p_2 g \, d\Gamma, \tag{2.135}$$

$$= \mathcal{G}_1^-(p_1;\lambda_1) + \mathcal{G}_2^-(p_2;\lambda_2). \tag{2.136}$$

The functional $\mathcal{G}_M^-(p;\lambda)$ is found to satisfy the maximum principle and this is established by adding the term

$$\int_{\Gamma_I} \left\{ \lambda_1 \nabla \widehat{p}(p_1 - \widehat{p}) \cdot \mathbf{n}_1 + \lambda_2 \nabla \widehat{p}(p_2 - \widehat{p}) \cdot \mathbf{n}_2 \right\} \, d\Gamma = 0, \tag{2.137}$$

which is zero as a result of the continuity of pressure over the interface and the property of the outward unit normals along Γ_I, namely $\mathbf{n}_1 = -\mathbf{n}_2$. Adding the term (2.137) permits the integral

$$\int_{\Gamma_m} \lambda_m \nabla \widehat{p}(p_m - \widehat{p}) \cdot \mathbf{n}_m \, d\Gamma \tag{2.138}$$

to be constructed for each m, from which positivity follows as in the derivation of (2.17). Hence, considering the difference between the multiple domain functional evaluated at the analytic solution $\widehat{p}$ and with *any* p satisfying the constraints (2.131) - (2.133) the required result is obtained,

$$\begin{aligned}
\mathcal{G}_M^-(\widehat{p};\lambda) - \mathcal{G}_M^-(p;\lambda) &= \frac{1}{2}\iint_{\Omega_1} \{\lambda_1 \nabla p_1 \cdot \nabla p_1 - \lambda_1 \nabla\widehat{p} \cdot \nabla\widehat{p}\} \, d\Omega \\
&\quad + \frac{1}{2}\iint_{\Omega_2} \{\lambda_2 \nabla p_2 \cdot \nabla p_2 - \lambda_2 \nabla\widehat{p} \cdot \nabla\widehat{p}\} \, d\Omega \\
&\quad + \int_{\Gamma_1^+} g(p_1 - \widehat{p}) \, d\Gamma + \int_{\Gamma_2^+} g(p_2 - \widehat{p}) \, d\Gamma, \\
&= \frac{1}{2}\iint_{\Omega_1} \{\lambda_1 \nabla p_1 \cdot \nabla p_1 - \lambda_1 \nabla\widehat{p} \cdot \nabla\widehat{p}\} \, d\Omega \\
&\quad + \frac{1}{2}\iint_{\Omega_2} \{\lambda_2 \nabla p_2 \cdot \nabla p_2 - \lambda_2 \nabla\widehat{p} \cdot \nabla\widehat{p}\} \, d\Omega \\
&\quad - \int_{\Gamma_1^+} \lambda_1 \nabla\widehat{p}(p_1 - \widehat{p}) \, d\Gamma - \int_{\Gamma_2^+} \lambda_1 \nabla\widehat{p}(p_2 - \widehat{p}) \, d\Gamma \\
&\quad - \int_{\Gamma_I} \{\lambda_1 \nabla\widehat{p}(p_1 - \widehat{p}) \cdot \mathbf{n}_1 + \lambda_2 \nabla\widehat{p}(p_2 - \widehat{p}) \cdot \mathbf{n}_2\} \; d\Gamma, \\
&= \frac{1}{2}\iint_{\Omega_1} \{\lambda_1 \nabla p_1 \cdot \nabla p_1 - \lambda_1 \nabla\widehat{p} \cdot \nabla\widehat{p}\} \, d\Omega \\
&\quad + \frac{1}{2}\iint_{\Omega_2} \{\lambda_2 \nabla p_2 \cdot \nabla p_2 - \lambda_2 \nabla\widehat{p} \cdot \nabla\widehat{p}\} \, d\Omega \\
&\quad - \int_{\Gamma_1} \lambda_1 \nabla\widehat{p}(p_1 - \widehat{p}) \cdot \mathbf{n}_1 \, d\Gamma - \int_{\Gamma_2} \lambda_2 \nabla\widehat{p}(p_2 - \widehat{p}) \cdot \mathbf{n}_2 \, d\Gamma, \\
&= \frac{1}{2}\iint_{\Omega_1} ({\lambda_1}^{\frac{1}{2}} \nabla p_1 - {\lambda_1}^{\frac{1}{2}} \nabla\widehat{p})^2 d\Omega \\
&\quad + \frac{1}{2}\iint_{\Omega_2} ({\lambda_2}^{\frac{1}{2}} \nabla p_2 - {\lambda_2}^{\frac{1}{2}} \nabla\widehat{p})^2 d\Omega, \\
&\geq 0.
\end{aligned} \tag{2.139}$$

Similarly enforcing flux continuity at the interface in addition to the constraints (2.110) and (2.111) recovers a minimum principle. The complete set of constraints is then

$$\nabla \cdot \mathbf{q}_m = 0 \quad \text{in } \Omega_m, \tag{2.140}$$

$$\mathbf{q}_m \cdot \mathbf{n}_m = g \quad \text{on } \Gamma_m^+, \tag{2.141}$$

$$\mathbf{q}_2 \cdot \mathbf{n}_I \quad = \quad \mathbf{q}_2 \cdot \mathbf{n}_I, \tag{2.142}$$

which generates the functional $\mathcal{G}_M^+(\mathbf{q};\lambda)$ when substituted into the free principle,

$$\begin{aligned}
\mathcal{G}_M^+(\mathbf{q};\lambda) &= \frac{1}{2}\iint_{\Omega_1} \lambda_1^{-1}\mathbf{q}_1 \cdot \mathbf{q}_1 \, d\Omega + \int_{\Gamma_1^-} f\mathbf{q}_1 \cdot \mathbf{n}_1 \, d\Gamma && (2.143)\\
&\quad + \frac{1}{2}\iint_{\Omega_2} \lambda_2^{-1}\mathbf{q}_2 \cdot \mathbf{q}_2 \, d\Omega + \int_{\Gamma^-} f\mathbf{q}_2 \cdot \mathbf{n}_2 \, d\Gamma, && (2.144)\\
&= \mathcal{G}_1^+(\mathbf{q}_1;\lambda_1) + \mathcal{G}_2^+(\mathbf{q}_2;\lambda_2). && (2.145)
\end{aligned}$$

Again the functional $\mathcal{G}_M^+(\mathbf{q};\lambda)$ is found to satisfy a minimum principle which is illustrated by adding the term

$$\int_{\Gamma_I} \{\widehat{p}(\widehat{\mathbf{q}} - \mathbf{q}_1) \cdot \mathbf{n}_1 + \widehat{p}(\widehat{\mathbf{q}} - \mathbf{q}_2) \cdot \mathbf{n}_2\} \, d\Gamma = 0 \tag{2.146}$$

along the internal boundary. This term is zero as a result of continuity of the normal flux over the interface and enables the integrals

$$\int_{\Gamma_m} \widehat{p}(\widehat{\mathbf{q}} - \mathbf{q}_m) \cdot \mathbf{n}_m \, d\Gamma \tag{2.147}$$

to be constructed for each sub-domain m, and from which the bound can be asserted. Considering the difference between the functional evaluated at the stationary point $\widehat{\mathbf{q}}$ and evaluated at *any* other point $\mathbf{q}$ satisfying the constraints (2.140) - (2.142) the minimum principle is demonstrated,

$$\begin{aligned}
\mathcal{G}_M^+(\widehat{\mathbf{q}};\lambda) - \mathcal{G}_M^+(\mathbf{q};\lambda) &= \frac{1}{2}\iint_{\Omega_1} \left\{\lambda_1^{-1}\widehat{\mathbf{q}} \cdot \widehat{\mathbf{q}} - \lambda_1^{-1}\mathbf{q}_1 \cdot \mathbf{q}_1\right\} d\Omega + \int_{\Gamma_1^-} f(\widehat{\mathbf{q}} - \mathbf{q}_1) \cdot \mathbf{n}_1 \, d\Gamma\\
&\quad + \frac{1}{2}\iint_{\Omega_2} \left\{\lambda^{-1}\widehat{\mathbf{q}} \cdot \widehat{\mathbf{q}} - \lambda^{-1}\mathbf{q}_2 \cdot \mathbf{q}_2\right\} d\Omega + \int_{\Gamma_2^-} f(\widehat{\mathbf{q}} - \mathbf{q}_2) \cdot \mathbf{n}_2 \, d\Gamma,\\
&= \frac{1}{2}\iint_{\Omega_1} \left\{\lambda_1^{-1}\widehat{\mathbf{q}} \cdot \widehat{\mathbf{q}} - \lambda_1^{-1}\mathbf{q}_1 \cdot \mathbf{q}_1\right\} d\Omega + \int_{\Gamma_1^-} \widehat{p}(\widehat{\mathbf{q}} - \mathbf{q}_1) \cdot \mathbf{n}_1 \, d\Gamma\\
&\quad + \frac{1}{2}\iint_{\Omega_2} \left\{\lambda^{-1}\widehat{\mathbf{q}} \cdot \widehat{\mathbf{q}} - \lambda^{-1}\mathbf{q}_2 \cdot \mathbf{q}_2\right\} d\Omega + \int_{\Gamma_2^-} \widehat{p}(\widehat{\mathbf{q}} - \mathbf{q}_2) \cdot \mathbf{n}_2 \, d\Gamma\\
&\quad + \int_{\Gamma_I} \{\widehat{p}(\widehat{\mathbf{q}} - \mathbf{q}_1) \cdot \mathbf{n}_1 + \widehat{p}(\widehat{\mathbf{q}} - \mathbf{q}_2) \cdot \mathbf{n}_2\} \, d\Gamma,\\
&= \frac{1}{2}\iint_{\Omega_1} \left\{\lambda_1^{-1}\widehat{\mathbf{q}} \cdot \widehat{\mathbf{q}} - \lambda_1^{-1}\mathbf{q}_1 \cdot \mathbf{q}_1\right\} d\Omega + \int_{\Gamma_1} \widehat{p}(\widehat{\mathbf{q}} - \mathbf{q}_1) \cdot \mathbf{n}_1 \, d\Gamma\\
&\quad + \frac{1}{2}\iint_{\Omega_2} \left\{\lambda^{-1}\widehat{\mathbf{q}} \cdot \widehat{\mathbf{q}} - \lambda^{-1}\mathbf{q}_2 \cdot \mathbf{q}_2\right\} d\Omega + \int_{\Gamma_2} \widehat{p}(\widehat{\mathbf{q}} - \mathbf{q}_2) \cdot \mathbf{n}_2 \, d\Gamma,\\
&= -\frac{1}{2}\iint_{\Omega_1} (\lambda_1^{-\frac{1}{2}}\widehat{\mathbf{q}} - \lambda_1^{-\frac{1}{2}}\mathbf{q}_1)^2 d\Omega - \frac{1}{2}\iint_{\Omega_2} (\lambda_2^{-\frac{1}{2}}\widehat{\mathbf{q}} - \lambda_2^{-\frac{1}{2}}\mathbf{q}_2)^2 d\Omega,\\
&\leq 0. && (2.148)
\end{aligned}$$

In general there may be many permeability discontinuities. However, the functionals $\mathcal{G}_M^-(p;\lambda)$ and $\mathcal{G}_M^+(\mathbf{q};\lambda)$ retain the required convexity and this can be demonstrated by adding terms of the form (2.137) and (2.146) for each discontinuity.

The similarities between the extremum principles constructed over smooth and discontinuous permeability data enables the functional $\mathcal{G}^-(p;\lambda)$ to be interpreted as the sum of functionals

$$\mathcal{G}^-(p;\lambda) = \sum_{m=1}^{M} \mathcal{G}_m^-(p_m;\lambda_m) \tag{2.149}$$

over the M sub domains Ω_m with the associated constraints (2.131) - (2.133). The sum naturally reduces to a single term in the case where the permeability data is continuous. Similarly for the minimum principle.

Chapter 3

Single-Phase Upscaling

3.1 Introduction to Upscaling

In the preceding chapter a simple reservoir model was constructed and variational methods described that enabled bounds to evaluated for the flux out of the domain. However it is common practice in reservoir simulations to substitute the original geological model for a lower resolution replacement in order to simplify the model. The process of determining this lower resolution replacement is termed upscaling. The motivation to upscale is as follows. The permeability of the reservoir rock can vary over the length scale of millimetres whilst the reservoir can extend over kilometres. As a result the permeability data set is huge and constructing numerical methods using it directly is found to be prohibitively uneconomical. Actually, the physical permeabilities are only known over a very small fraction of the domain, by way of core samples, and the remaining permeability data is constructed by seismic surveying and geo-statistical techniques. Here the interest is in analysing the upscaling stage from the fine to a coarse representation of the data.

In general it is preferable that the upscaled permeability field is regarded as a function of the original permeability data rather than the flow solution, as this allows the upscaling stage to be uniquely calculated in advance and a computational saving made. Methods in which the upscaled permeability additionally depends on the flow solution introduce non-linearities into the governing equations (2.85)-(2.88) of the reservoir model. These non-linearities require an iterative solution procedure in general and the possibility arises that the upscaled system of equations becomes more expensive to solve than the original linear system. Therefore, initially consideration is restricted to upscaling methods that can be separated from the flow solution. In addition we will also assume the original permeability data to be a scalar function, although the upscaled permeability field may be a tensor. The anisotropy in the upscaled data is introduced as a result of patterns in the original permeability data creating preferential directions of flow.

The aim of upscaling then is to construct an upscaled permeability field, $\Lambda(\lambda(\mathbf{x}))$, over which the associated pressure and flux functions $\widehat{P}(\mathbf{x})$ and $\widehat{\mathbf{Q}}(\mathbf{x})$ are solved. The upscaled governing equations are therefore

$$\begin{aligned}
\widehat{\mathbf{Q}}(\mathbf{x}) &= -\Lambda(\lambda(\mathbf{x}))\nabla\widehat{P}(\mathbf{x}) \quad \text{in } \Omega, && (3.1)\\
\nabla\cdot\widehat{\mathbf{Q}}(\mathbf{x}) &= 0 \quad \text{in } \Omega, && (3.2)\\
\widehat{P}(\mathbf{x}) &= f(\mathbf{x}) \quad \text{on } \Gamma^-, && (3.3)\\
\widehat{\mathbf{Q}}(\mathbf{x})\cdot\mathbf{n}(\mathbf{x}) &= g(\mathbf{x}) \quad \text{on } \Gamma^+, && (3.4)
\end{aligned}$$

to which the variational formulation of chapter 2 can be applied. Practically, the original permeability data λ can be regarded as a piecewise constant description which is coarsened through the process of upscaling by local calculations over each grid cell Ω_i, as shown in figure 3.1.

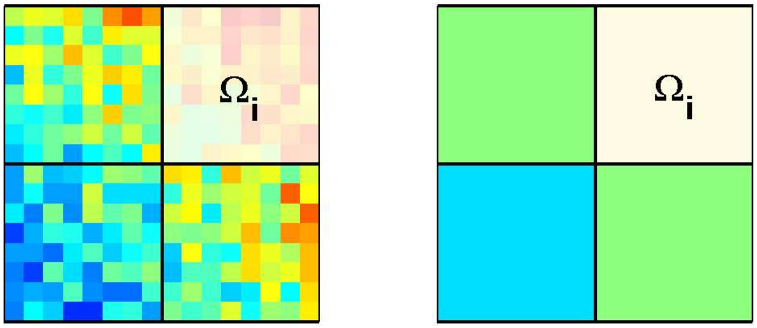

Figure 3.1: Upscaling over the area Ω_i

3.1.1 Review of Conventional Upscaling Methods

Many upscaling methods of differing complexity have been constructed and suggested [15]. This review contains references to material concerned with upscaling the permeability data with respect to the single phase governing equations (2.85)-(2.88). Another field of study related to upscaling is that of homogenisation, in which the properties of heterogeneous materials are sought in the limit as the length scale associated with the heterogeneity tends to zero. Homogenisation is not considered in this thesis since the length scales associated with permeability data will be considered non-zero. The application of homogenisation techniques to reservoir simulation can be found in Hornung [22] and Panfilov [37].

The advent of computing and numerical modelling enabled reservoir simulation to develop in importance. Early work by Warren and Price [49] investigated the possibility of deducing the probable configuration of the permeability field for a known probability distribution of the permeability values. Warren and Price considered steady state flow between opposing faces as well as radial flow into a well. Time dependent solutions were also calculated for a radial model around a well and the pressure 'build-up' curves, representing the well pressures over time, were extracted. The effective permeability values for the homogenous reservoirs were then calculated and Warren and Price concluded that out of the arithmetic, harmonic and geometric means of the permeability data, the geometric mean characterized the heterogeneous reservoir most closely for log-normal and exponentially distributed permeability values.

With advances in computing power and numerical techniques, upscaling methods based on solving local flow simulations (rather than simple averaging) were developed. These methods are motivated by numerically conducting the type of physical experiment that would be performed in the laboratory. This second category of methods attempt to model the response of a flow over sections of the original permeability data but their drawback is that they require assumptions to be made regarding the boundary conditions for these local flow simulations. These local flow simulations are normally implemented on a cuboidal subdomain with a pressure potential defined between opposing faces and either no-flow boundary conditions, or a linear pressure gradient, specified along the sides. These methods are therefore naturally able to produce a diagonal permeability tensor by solving local flow simulations in each coordinate direction. Upscaling methods of this type have been developed by King *et al* [27], who calculate a coarse permeability field of effective permeabilities in this way. The effective permeability is defined as the permeability of subdomain if it were homogeneous, such that the net boundary flux is equal to that of the solution over the fine scale permeability field in the subdomain, with identical boundary conditions.

Methods to address the arbitrary nature in which the boundary conditions are chosen for these local simulations have been constructed, including the work of Wallstrom *et al* [48]. The general practice is to embed the local simulations in a slightly larger problem so that the boundary conditions imposed do not act directly on the local domain. In the work of Wallstrom *et al* so called Effective Flux Boundary Condition, (EFBCs), are generated by drawing similarities with the classical problem of flow through an infinite homogeneous domain containing an homogeneous inclusion. In this context the coarse grid cell is approximated as an elliptical inclusion within a background medium of appropriate effective permeability.

A pseudo-local flow method bridging the two categories is also common. The arithmetic-harmonic method is implemented by taking the harmonic average along permeability strips in the direction of the flow and then arithmetically averaging the strips across the flow. This method also produces a diagonal tensor.

Numerical methods have also been developed that effectively achieve an upscaled flow solution without necessarily explicitly defining an upscaled permeability field first. The multiscale finite element method of Hou and Wu [23] is an example of such a numerical method. The multiscale finite element method employs basis functions constructed from local solutions of the governing equation over the original permeability data. This method enables flow features generated by the fine permeability field to enter into a coarse representation of the solution. However, the local flow solutions are not used to calculate an effective coarse permeability value, but are used solely as the expansion for the numerical solution. Solving the governing equations effectively over the original permeability data, but with complex basis functions, enables convergence analysis from the finite element literature to be applied to the method.

Moulton *et al* [34] base the method of upscaling on the multigrid process. The multigrid method automatically captures the many length scales associated with reservoir simulations, and in particular an effective permeability tensor can be extracted from the coarse grid operator. The coarse grid discrete operator is obtained by operator-induced coarsening of the fine-scale discretisation method. The effective permeability tensor can then be computed directly from the node weightings of the stencil associated with the coarse grid operator. In this manner the upscaled permeability field can be extracted by initialising the multigrid method but without using it to actually solve the governing equations (2.85)-(2.88).

A similar approach is taken by Arbogast *et al* [2] in which the fine scale solution is decomposed into a coarse solution and a remainder. Approximations to the remainder are then calculated on a fine grid locally and are found to modify the coarse grid operator, and therefore the coarse solution. In contrast to the method of Moulton *et al* the upscaled permeability field is not extracted but instead the modified coarse grid operator is used directly to generate the coarse solution.

The upscaling methods reviewed so far have had very little quantifiable error analysis associated with them. One method of introducing a measure of the quality of the upscaling procedure is to minimize a 'cost' or 'error' function. This philosophy is developed by Nielsen and Tveito [36] who define upscaling as an optimisation problem in which the error between the fine and coarse velocity fields is minimized with respect to a defined norm. The choice of norm permits the optimal coarse velocity field to be obtained without having to solve the fine scale problem.

The optimality of the upscaling may be reassuring but the error, or cost, of the upscaling method may not be measured in a practical norm. Obtaining meaningful error estimates in an upscaled solution is compounded by the drift between the solutions $\widehat{p}$, the exact pressure solution obtained using the original permeability data λ, and $\widehat{P}$, the exact pressure solution obtained using the upscaled permeability data Λ. As a result an error measure on the flux consists of a standard discretisation error $\|\Lambda\nabla P_h - \Lambda\nabla\widehat{P}\|$, where P_h is a numerical approximation to $\widehat{P}$, and a consistency error $\|\Lambda\nabla\widehat{P} - \lambda\nabla\widehat{p}\|$, due to the difference between the exact upscaled and original solutions. The flux error therefore satisfies the triangle inequality

$$\|\Lambda\nabla P_h - \lambda\nabla\widehat{p}\| \;\leq\; \underset{\text{Discretisation}}{\|\Lambda\nabla P_h - \Lambda\nabla\widehat{P}\|} + \underset{\text{Consistency}}{\|\Lambda\nabla\widehat{P} - \lambda\nabla\widehat{p}\|}. \tag{3.5}$$

The discretisation error can be estimated to some degree by the numerical method employed. Approximations to the consistency error are less obvious.

3.1.2 An Upscaling Philosophy

In contrast to existing methods we adopt a more abstract approach and define the *aim* of upscaling to be the replacement of the original permeability field λ with a coarser description Λ such that the quantity of interest is preserved, that is,

$$\mathcal{G}(\widehat{p}, \widehat{\mathbf{q}}; \lambda) = \mathcal{G}(\widehat{P}, \widehat{\mathbf{Q}}; \Lambda). \tag{3.6}$$

Recognising that we are unlikely to achieve the equality in (3.6) we introduce two coarsely defined permeability fields Λ^- and Λ^+ and treat upscaling as a comparison problem. The practical aim of this method is then to generate new upscaling methods which enable upper and lower bounds to be calculated for the original problem, defined in terms of λ the fine scale permeability description, using approximations defined over the coarse upscaled permeability fields Λ^- and Λ^+ . Hence bounds of the form

$$\mathcal{G}^-(P_h; \Lambda^-) \leq \mathcal{G}(\widehat{p}, \widehat{\mathbf{q}}; \lambda) \leq \mathcal{G}^+(\mathbf{Q}_h; \Lambda^+). \tag{3.7}$$

will be found where $\mathcal{G}(\widehat{p}, \widehat{\mathbf{q}}; \lambda)$ denotes the exact value of the quantity of interest defined over the fine scale permeability data.

The new method is called consistent upscaling since bounds are retained on the analytic solution over the *original* permeability data. Comparing solutions calculated using conventional upscaling methods with solutions obtained over the original permeability data will also permit the consistency and discretisation errors associated with conventional methods to be analysed. Schematically the bounds shown in figure 3.2 are obtained. The analytic value of the quantity of interest is bounded by the dual extremum principles constructed using the original permeability data. The analytic value of the quantity of interest is also bounded by the results obtained using the consistently upscaled permeability data Λ^- and Λ^+ in the maximum and minimum principles respectively. In addition, the approximations to the quantity of interest obtained using the conventionally upscaled permeability data Λ_{eff} in the maximum and minimum principles are also plotted.

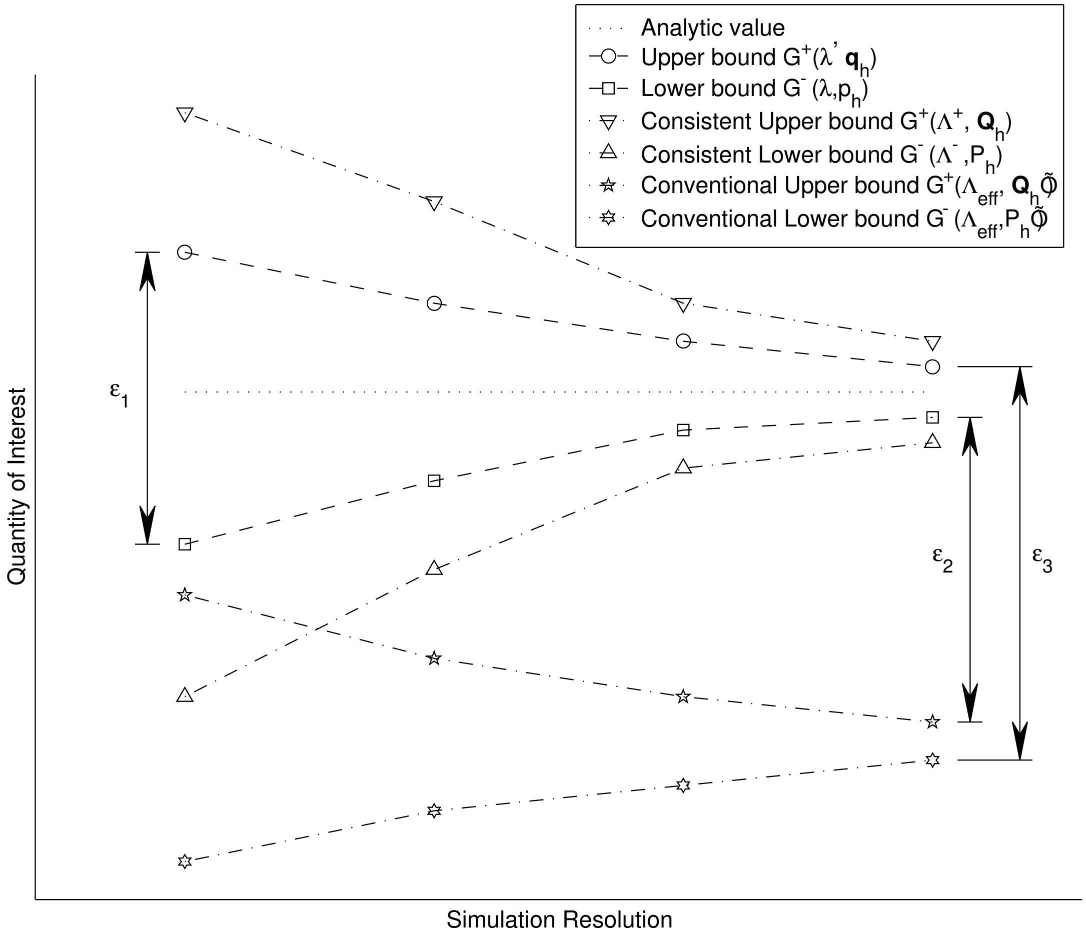

Figure 3.2: Schematic convergence of upper and lower bounds

From the various approximations to the quantity of interest the discretisation and consistency errors can be bounded. The discretisation error associated with a pair solutions of the minimum and maximum principles is bounded by the difference in the value of the quantity of interest. For example the discretisation error in the approximation to the quantity of interest when solved at the coarsest resolution considered, and using the original permeability data, is bounded by ϵ_1. The consistency error of the conventional upscaling method shown is bounded between ϵ_2 and ϵ_3.

Having discussed the motivation for a consistent upscaling method some examples are now generated.

3.2 Consistent Upscaling Methods

From the description of consistent upscaling it is possible to consider the method as the pair of constrained optimisation problems,

$$\min_{\Lambda^-}\left[\mathcal{G}(\widehat{p},\widehat{\mathbf{q}};\lambda)-\mathcal{G}^-(P_h;\Lambda^-)\right] \quad s.t. \quad \mathcal{G}(\widehat{p},\widehat{\mathbf{q}};\lambda)-\mathcal{G}^-(P_h;\Lambda^-)\geq 0 \quad \forall P_h\in H_h^-, \tag{3.8}$$

$$\min_{\Lambda^+}\left[\mathcal{G}^+(\mathbf{Q}_h;\Lambda^+)-\mathcal{G}(\widehat{p},\widehat{\mathbf{q}};\lambda)\right] \quad s.t. \quad \mathcal{G}^+(\mathbf{Q}_h;\Lambda^+)-\mathcal{G}(\widehat{p},\widehat{\mathbf{q}};\lambda)\geq 0 \quad \forall \mathbf{Q}_h\in H_h^+. \tag{3.9}$$

where P_h and $\mathbf{Q}_h$ are approximations to the solutions $\widehat{P}$ and $\widehat{\mathbf{Q}}$ respectively. The minimum and maximum principles discussed in chapter 2 present possible forms for Λ^- and Λ^+. Considering the maximum principle (2.109), but instead writing the difference between the functionals constructed using the original and upscaled permeability data,

$$\begin{aligned}
\mathcal{G}^-(\widehat{p};\lambda)-\mathcal{G}^-(P_h;\Lambda^-) &= \frac{1}{2}\iint_\Omega \left\{\Lambda^-\nabla P_h\cdot\nabla P_h-\lambda\nabla\widehat{p}\cdot\nabla\widehat{p}\right\}d\Omega+\int_{\Gamma^+} g(P_h-\widehat{p})\,d\Gamma,\\
&= \frac{1}{2}\iint_\Omega \left\{\Lambda^-\nabla P_h\cdot\nabla P_h-\lambda\nabla\widehat{p}\cdot\nabla\widehat{p}-2\lambda\nabla\widehat{p}\cdot\nabla(P_h-\widehat{p})\right\}d\Omega,\\
&= \frac{1}{2}\iint_\Omega \left\{(\lambda^{\frac{1}{2}}\nabla P_h-\lambda^{\frac{1}{2}}\nabla\widehat{p})^2+(\Lambda^- -\lambda)\nabla P_h\cdot\nabla P_h\right\}d\Omega.
\end{aligned} \tag{3.10}$$

is obtained. Minimisation and positivity of the right hand side of (3.10) can be ensured by any of the following techniques.

3.2.1 Inf-Sup Upscaling

The inf-sup upscaling method, producing a scalar function Λ^-, defines Λ^- to be everywhere greater than or equal to λ on each coarse region Ω_i. The method can be implemented with various basis functions for Λ^- such that, whilst retaining $\Lambda^-\geq\lambda$ everywhere, the difference is also minimised.

3.2.2 Piecewise-Constant Upscaling

A special upscaling procedure is possible when ∇P_h is modelled as a piecewise constant function. In this case the upscaled permeability is simply the arithmetic average and the second term in (3.10) does not contribute. This can be considered as a full decoupling of the upscaling procedure and the solution of the governing equations (3.1)-(3.4). In this situation prior knowledge of the coarse solution P_h would not help to construct the upscaled permeability field Λ^-. In the case of a two-dimensional simulation the method is implemented by discretising the pressure P_h with piecewise linear triangles. The discretisation can then be constructed on a regular square grid, with two triangles per square, so that the model still shares the same nodes as methods constructed with quadrilateral elements. The piecewise-constant method also produces scalar upscaled permeability data.

3.2.3 Spectral Upscaling

The spectral method is capable of producing upscaled permeability data as a symmetric tensor. In the spectral upscaling method we construct a consistent upscaled permeability field for all P_h belonging to the space spanned by a set of chosen basis functions $\{\phi_i\}$. To implement the method we expand P_h in terms of the set $\{\phi_i\}$, which must satisfy the constraints H^- (2.131)-(2.133),

$$P_h = \sum_{i=1}^{N} P_i \phi_i, \tag{3.11}$$

and expand Λ^- in terms of a set of piecewise constant basis functions $\{\theta_i\}$ corresponding to each coarse element Ω_i

$$\Lambda^- = \sum_{i=1}^{M} \Lambda_i^- \theta_i, \tag{3.12}$$

where

$$\theta_i(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in \ \Omega_i, \\ 0 & \mathbf{x} \notin \ \Omega_i. \end{cases} \tag{3.13}$$

The second term in the equation (3.10) can now be written as a sum of local contributions over the element Ω_i

$$\iint_{\Omega} (\Lambda^- - \lambda) \nabla P_h \cdot \nabla P_h \, d\Omega = \sum_{i=1}^{M} \mathbf{P}_i^T (\gamma_i \mathsf{R}_i - \mathsf{S}_i) \mathbf{P}_i \tag{3.14}$$

where $\mathbf{P}_i$ is the vector of node values P_i in the neighbourhood of θ_i, and R_i and S_i are the element stiffness matrices for the elements $\phi_{i,j}$ in the neighbourhood of θ_i. In two-dimensions

$$(\mathsf{R}_i)_{ij} \quad = \quad \iint_{\Theta_i} \left\{ \frac{\partial \phi_i}{\partial x} \cdot \frac{\partial \phi_j}{\partial x} + \alpha_i \left(\frac{\partial \phi_i}{\partial x} \cdot \frac{\partial \phi_j}{\partial y} + \frac{\partial \phi_i}{\partial y} \cdot \frac{\partial \phi_j}{\partial x} \right) + \beta_i \frac{\partial \phi_i}{\partial y} \cdot \frac{\partial \phi_j}{\partial y} \right\} d\Omega,$$

$$\tag{3.15}$$

$$(\mathsf{S}_i)_{ij} = \iint_{\Theta_i} \lambda \nabla \phi_i \cdot \nabla \phi_j \, d\Omega, \tag{3.16}$$

where α_i, β_i and γ_i are to be determined. Positivity of the second term in (3.10) is achieved if the matrices $\gamma_i \mathsf{R}_i - \mathsf{S}_i$ are all positive semi-definite in the sense that

$$\mathbf{P}_i^T (\gamma_i \mathsf{R}_i - \mathsf{S}_i) \mathbf{P}_i \geq 0 \quad \forall \mathbf{P}_i, i = 1, \cdots, M. \tag{3.17}$$

This is accomplished by solving the generalised eigenvalue problem

$$\mathsf{S}_r \mathbf{x} = \nu \mathsf{R}_r \mathbf{x}, \tag{3.18}$$

in each coarse element Ω_i, where S_r and R_r are S_i and R_i reduced by one row and column to remove the zero eigenvalue present. The eigenvalue shift required for positivity of (3.17) is achieved by choosing $\gamma_i = \max(\nu)$ since

$$(\gamma \mathsf{R}_r - \mathsf{S}_r) \mathbf{x} = (\gamma - \nu) \mathsf{R}_r \mathbf{x}. \tag{3.19}$$

The coefficients α_i and β_i of R are determined by clustering the eigenvalues η of $\mathsf{R}_r^{-1} \mathsf{S}_r$ in order that the spread of the eigenvalues $(\gamma - \mu)$ is minimised. In constructing the spectral upscaling method attempts are made to approximate the operator S_r by $\nu \mathsf{R}_r$ and therefore $\mathsf{R}_r^{-1} \mathsf{S}_r \approx \nu \mathsf{I}$. This provides the motivation to cluster the eigenvalues of $\mathsf{R}_r^{-1} \mathsf{S}_r$ in order to obtain an operator similar to a multiple of the identity. The clustering is achieved through a simple numerical minimisation of $\max(\nu) - \min(\nu)$ using a bracketing technique. The upscaled permeability coefficient matrix Λ_i^- over the coarse element Ω_i is then

$$\Lambda_i^- = \begin{bmatrix} \gamma_i & \gamma_i \alpha_i \\ \gamma_i \alpha_i & \gamma_i \beta_i \end{bmatrix}. \tag{3.20}$$

If, due to computational cost, Λ^- is required only as a diagonal matrix then the eigenvalue clustering stage can be avoided by setting $\alpha = 0$ and $\beta = 1$.

3.2.4 Tuned Basis Functions

In addition to ensuring that the second term of (3.10) is positive, the first term can be minimised by selecting a set of basis functions that efficiently represent the pressure solution. Due to the flux continuity condition across the permeability discontinuities, the exact pressure gradient $\nabla\widehat{p}$ also experiences discontinuities at these interfaces. Therefore in order to minimise the integral

$$\frac{1}{2}\iint_{\Omega}(\lambda^{\frac{1}{2}}\nabla P_h - \lambda^{\frac{1}{2}}\nabla\widehat{p})^2 \tag{3.21}$$

in (3.10), P_h should also exhibit similar pressure gradient discontinuities. Constructing a set of basis functions that feature the correct interfacial conditions along the permeability discontinuities enables a more accurate approximation to the solution to be formed. A set of tuned basis functions which satisfy the flux continuity condition more accurately can be constructed from local solutions over the λ field. This is implemented by expanding the tuned basis functions, $\tilde{\phi}_i$, in terms of a set of fine basis functions $\{\sigma_j\}$ through

$$\tilde{\phi}_i = \sum_j a_{ij}\sigma_j, \tag{3.22}$$

and then solving

$$\iint_{\Omega}\sigma_k\nabla\cdot(\lambda\nabla\sum_j a_{ij}\sigma_j - \nabla\phi_i)\,d\Omega = 0 \qquad \forall k, \tag{3.23}$$

or

$$\iint_{\Omega}\sigma_k\nabla\cdot(\lambda\nabla\sum_j a_{ij}\sigma_j - \lambda_{eff}\nabla\phi_i)\,d\Omega = 0 \qquad \forall k, \tag{3.24}$$

for the unknown coefficients of a_{ij}, for each i.

The use of any conventional upscaling method to produce a symmetric tensor Λ_{eff} allows the anisotropy of the media to be further incorporated into the tuned basis functions. Any upscaling method can be used to generate Λ_{eff} as (3.23) or (3.24) are only used to determine the basis functions

$$\tilde{\phi}_i = \sum_j a_{ij}\sigma_j, \tag{3.25}$$

which will then be used in the expansion of the coarse solution P_h. Therefore computing Λ_{eff} using a cheap method such as the arithmetic-harmonic upscaling technique is recommended. We propose to solve the stationary equations using continuous finite elements with compact support and hence enforce

$$\tilde{\phi}_i(\mathbf{x}) = \phi_i(\mathbf{x}) \quad \text{if} \quad \phi_i(\mathbf{x}) = 0. \tag{3.26}$$

Along the boundaries of the domain the fine permeability field is reflected so that the same routine can be used to solve for both boundary and internal nodes. A zero Dirichlet boundary condition is inherited from the regular basis functions and is imposed around the perimeter of the basis function. The compact support of $\tilde{\phi}_i$ enables (3.23) and (3.24) to be expressed as

$$\iint_\Omega \nabla\sigma_k \cdot (\lambda\nabla \sum_j a_{ij}\sigma_j - \nabla\phi_i)\, d\Omega = 0 \qquad \forall k, \tag{3.27}$$

or

$$\iint_\Omega \nabla\sigma_k \cdot (\lambda\nabla \sum_j a_{ij}\sigma_j - \lambda_{eff}\nabla\phi_i)\, d\Omega = 0 \qquad \forall k, \tag{3.28}$$

respectively, and implies that the flux generated by the tuned basis function should be weakly equivalent to that generated by the coarse basis functions, but satisfy the fine scale interfacial conditions. Having solved for the coefficients a_{ij} the basis functions require normalising so that they sum to unity and are therefore able to represent the boundary conditions. As a result of this adjustment they no longer satisfy (3.23) or (3.24) exactly but still retain much of the required detail. A typical tuned basis function obtained from this procedure is shown in figure 3.3.

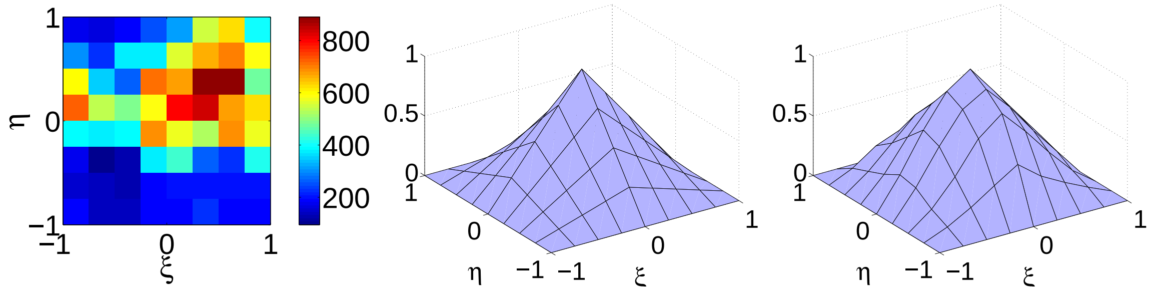

Figure 3.3: Underlying λ field, regular basis function ϕ_i and tuned basis function $\tilde{\phi}_i$

The coarse solution P_h is then loosely a linear sum of many local solutions over the fine grid and this displays similarities with the upscaling methods of King. However, computing local solutions with compact support removes the need to make predictions on the best boundary condition for a particular simulation.

The method is also closely related to the multiscale finite element method of Hou and Wu [23] which also involves basis functions constructed from local solutions. The method described here differs from that of Hou and Wu in the treatment of the boundary conditions with the tuned basis functions here effectively inheriting the boundary conditions from the equivalent coarse element.

Tuned basis functions are not directly applicable to the piecewise constant upscaling method. Tuned basis functions could be constructed for the piecewise constant upscaling method by expanding the basis functions in terms of fine linear triangles, but Λ^- would then no longer be determined by the simple arithmetic average of λ. Instead a weighted arithmetic average stemming from the non-constant gradient of the basis function within Ω_i would be required. In general it may even be required to determine a separate Λ^- value corresponding to each pair of interacting nodes. This approach has not been considered in this research.

3.2.5 Consistent Methods for the Minimum Principle

Consistent minimum principle upscaling methods may be constructed from the dual of (3.10), written as

$$\begin{aligned}
\mathcal{G}^+(\mathbf{Q}_h;\Lambda^+) - \mathcal{G}^+(\widehat{\mathbf{q}};\lambda) &= \frac{1}{2}\iint_\Omega \left\{(\Lambda^+)^{-1}\mathbf{Q}_h\cdot\mathbf{Q}_h - \lambda^{-1}\widehat{\mathbf{q}}\cdot\widehat{\mathbf{q}}\right\} d\Omega + \int_{\Gamma^-} f(\mathbf{Q}_h - \widehat{\mathbf{q}})\cdot\mathbf{n}\, d\Gamma, \\
&= \frac{1}{2}\iint_\Omega \left\{(\Lambda^+)^{-1}\mathbf{Q}_h\cdot\mathbf{Q}_h - \lambda^{-1}\widehat{\mathbf{q}}\cdot\widehat{\mathbf{q}} + 2(\mathbf{Q}_h - \widehat{\mathbf{q}})\cdot\nabla\widehat{p}\right\} d\Omega, \\
&= \frac{1}{2}\iint_D \left\{(\lambda^{-\frac{1}{2}}\widehat{\mathbf{q}} - \lambda^{-\frac{1}{2}}\mathbf{Q}_h)^2 + \left((\Lambda^+)^{-1} - \lambda^{-1}\right)\mathbf{Q}_h\cdot\mathbf{Q}_h\right\} d\Omega
\end{aligned} \tag{3.29}$$

from which direct parallels with the methods above can be constructed in order to ensure positivity and minimise the right hand side. If Λ^+ is chosen to be a scalar along with λ then the corresponding inf component of the inf-sup upscaling method is to construct Λ^+ to be everywhere less than λ.

The piecewise constant upscaling case is again an important method. Expanding the flux Q_h in terms of a piecewise constant set of basis functions results in the upscaled permeability field Λ^+ being defined simply as the harmonic mean.

The spectral method proceeds in the same manner as for the lower bound. The expansion of the coarse flux solution is via a stream function Ψ, discretised using the basis functions ϕ_i to give

$$\mathbf{Q}_h = \sum_{i=1}^{N} Q_i \left(\frac{\partial\phi_i}{\partial y}\mathbf{i} - \frac{\partial\phi_i}{\partial x}\mathbf{j}\right). \tag{3.30}$$

The inverse of the upscaled permeability field is expanded over the coarse elements as

$$(\Lambda^{+})^{-1} = \sum_{i=1}^{M} (\Lambda_i^{+})^{-1} \theta_i \tag{3.31}$$

and this enables the counterparts to the matrices R_i and S_i to be constructed, and the corresponding coefficients α, β and γ to be determined through the eigenvalue clustering and shifting procedure. Finally the inverse is taken for each coarse element Ω_i, on which $(\Lambda_i^{+})^{-1}$ is a constant, to obtain Λ^{+}.

Tuned basis functions for the coarse flux $\mathbf{Q}_h$ aim to minimise the integral

$$\iint_D (\lambda^{-\frac{1}{2}} \widehat{\mathbf{q}} - \lambda^{-\frac{1}{2}} \mathbf{Q}_h)^2 d\Omega. \tag{3.32}$$

To introduce fine scale structure into the solution $\mathbf{Q}_h$ we choose to weakly equate the flux generated by the tuned basis function times the inverse permeability with the flux generated by the regular basis function. This is a direct analogy of equating the fluxes in (3.27) and (3.28). Similarly, the tuned basis functions are expanded in terms of the finer basis functions by,

$$\tilde{\phi}_i = \sum_j c_{ij} \sigma_j. \tag{3.33}$$

Then for the unknown coefficients a_{ij} and each coarse node i, either

$$0 = \iint_\Omega \left(\frac{\partial \sigma_k}{\partial y} \mathbf{i} - \frac{\partial \sigma_{\mathbf{k}}}{\partial \mathbf{x}} \mathbf{j} \right) \cdot \left(\lambda^{-1} \sum_j a_{ij} \left(\frac{\partial \sigma_k}{\partial y} \mathbf{i} - \frac{\partial \sigma_{\mathbf{k}}}{\partial \mathbf{x}} \mathbf{j} \right) - \left(\frac{\partial \phi_i}{\partial y} \mathbf{i} - \frac{\partial \phi_{\mathbf{i}}}{\partial \mathbf{x}} \mathbf{j} \right) \right) d\Omega \ \forall k \tag{3.34}$$

$$= \iint_\Omega \nabla \sigma_k \cdot (\lambda^{-1} \nabla \sum_j a_{ij} \sigma_j - \nabla \phi_i) \, d\Omega \qquad \forall k \tag{3.35}$$

or

$$0 = \iint_\Omega \nabla \sigma_k \cdot (\lambda^{-1} \nabla \sum_j a_{ij} \sigma_j - \lambda_{eff}^{-1} \nabla \phi_i) \, d\Omega \qquad \forall k \tag{3.36}$$

is solved.

3.3 Results Obtained Using the Consistent Upscaling Methods

Results have been obtained by solving the set of equations (3.1)-(3.4) for a range of conventional upscaling methods and the consistent methods described in section 3.2.

The solutions were constructed over a randomly generated permeability field. The permeability field was constructed to include regions of high and low permeability values and thereby includes larger scale features. The regions of high and low permeability were generated using a random walk and the values in these regions where again generated randomly from linear distributions around 1000 ± 10 and 200 ± 50 respectively. Finally the data was smoothed using the five point Laplacian finite difference operator.

The boundary conditions considered are taken from the two model problems described in chapter 2 and are shown in figure 3.4 along with the permeability field λ. The physical interpretation of the stationary value is that it is equal to half the outward normal flux integrated over the inlet $x = 0$. The stationary value is therefore directly related to the flux out of the production well in these models.

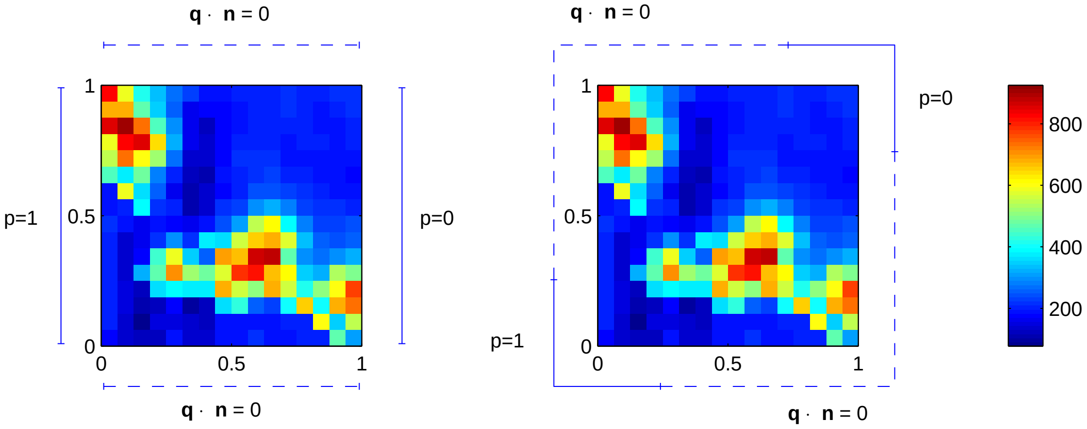

Figure 3.4: Boundary conditions and permeability data for Model A and Model B

The numerical methods were implemented on a mesh of $N \times N$ square elements. In all but the piecewise constant upscaling method the pressure and stream function were discretised using piecewise bilinear quadrilateral elements. The choice of basis functions ensures that the boundary conditions can be satisfied exactly, and although greater accuracy in the quantity of interest could have been expected if elements of a higher polynomial degree were selected, linear elements were chosen for simplicity especially with respect to the spectral upscaling and tuned basis functions procedures. The piecewise constant upscaling method requires the pressure and stream function to be discretised using piecewise linear triangles defined over the same grid. The pressure and stream function solutions obtained by solving Model A and Model B with a resolution of 16 × 16 elements are shown in figure 3.5.

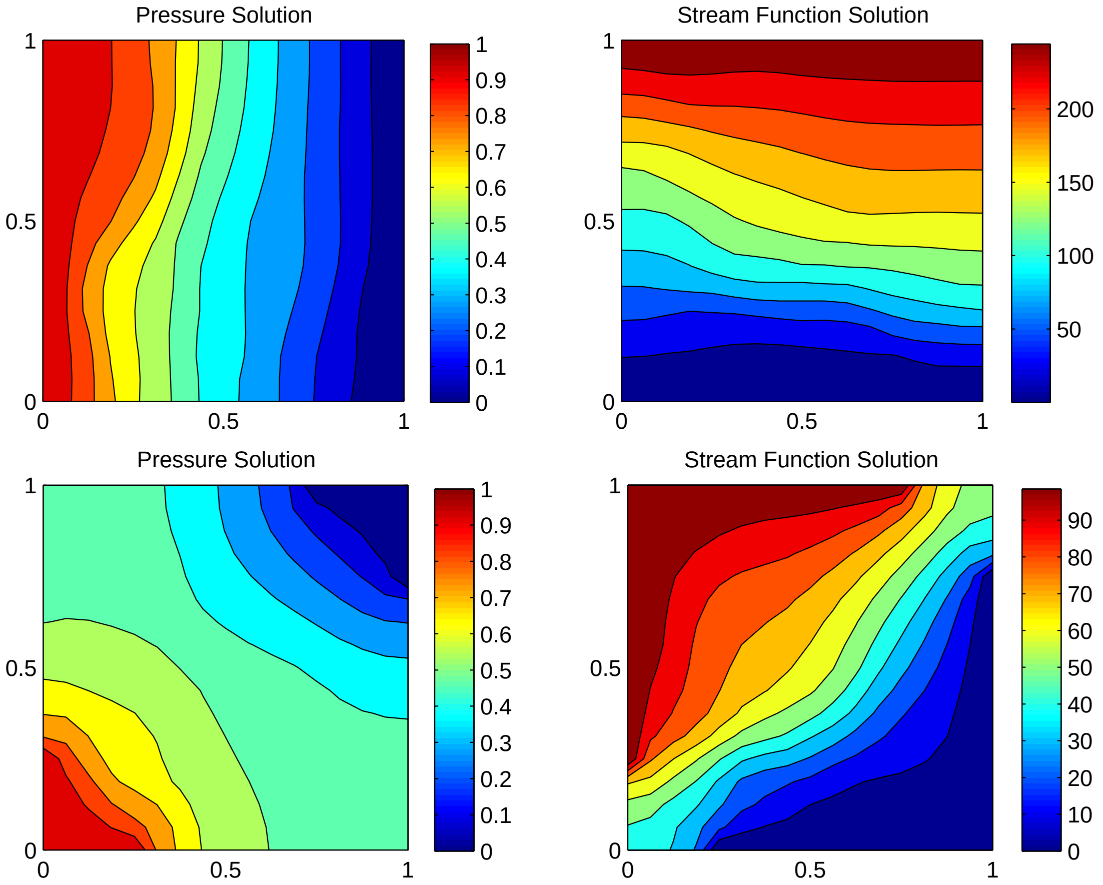

Figure 3.5: Solutions to Model A (above) and Model B (below)

3.3.1 Consistency Error

To investigate the degree of consistency error present in some of the conventional upscaling methods discussed in section 3.1.1 the permeability domain was upscaled using various methods to a 2 × 2 coarse representation. The maximum and minimum principles were then used to obtain bounds on the stationary value over progressively finer grids. All the methods were implemented using regular basis functions. The results are shown in figures 3.6. The consistency error can then be determined since the analytic stationary value of the original problem lies between the bounds calculated without an upscaling method applied, and the analytic stationary value of the upscaled problems lie between the respective bounds. The boundary conditions imposed on Model B enforce

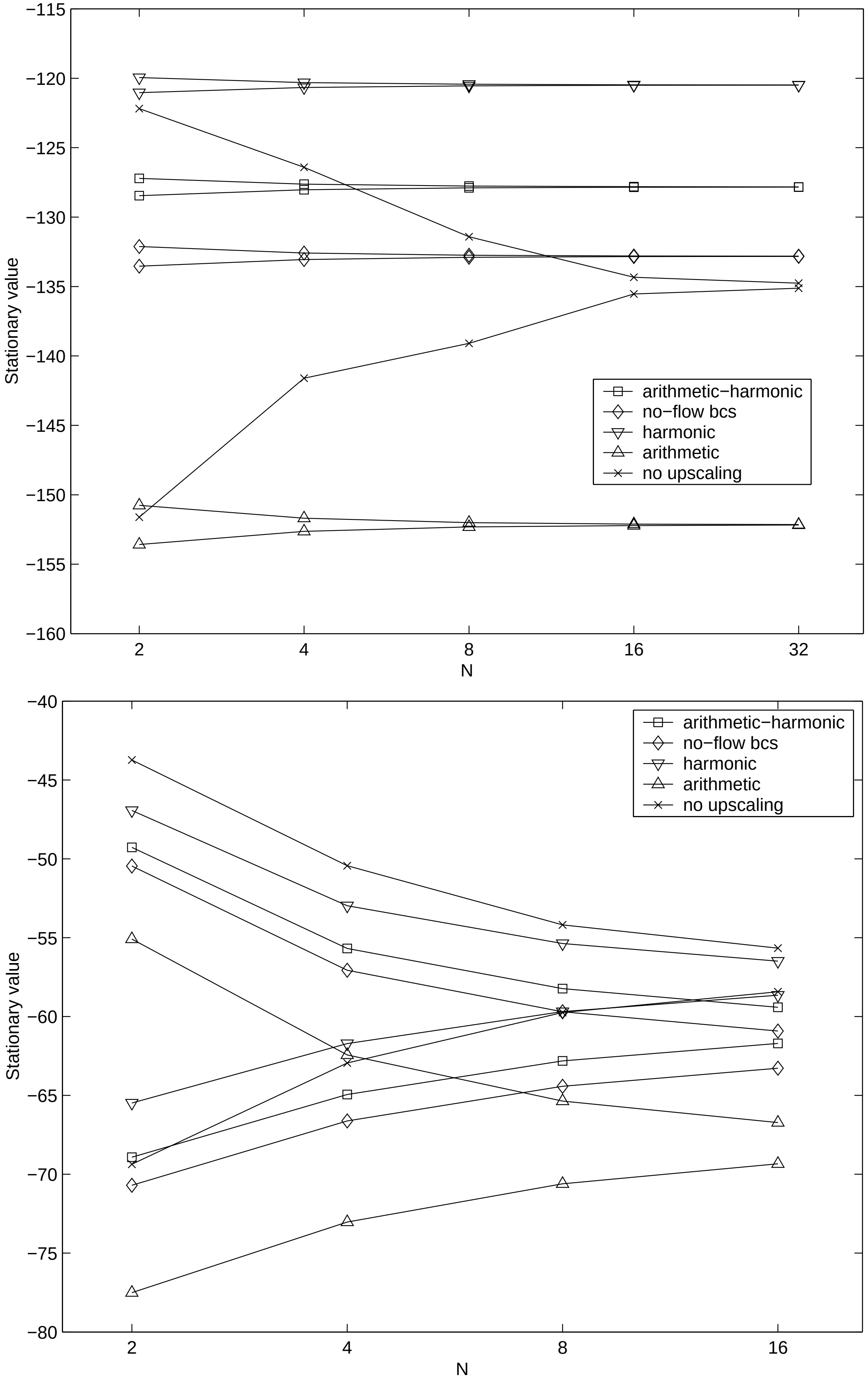

Figure 3.6: Consistency Errors, Model A (above) and Model B (below)

a regular grid of at least 4 × 4 elements, as opposed to Model A which can be modelled using 2 × 2 elements and still satisfy the boundary conditions.

Figure 3.6 indicates the consistency errors incurred by the different conventional upscaling methods, illustrated by the different values that the bounds converge to. The consistency error is naturally quantifiable by considering the maximum difference between the bounds of the upscaled solution and the non-upscaled solution. Similarly the discretisation error is represented by the difference between the upper and lower bounds of a particular method and is therefore also quantifiable. From figure 3.6 it can be seen that in this example the conventional upscaling methods based on local flow solutions, with no-flow boundary conditions, and the arithmetic-harmonic method, produce the better approximations to the stationary value. The discretisation error of the upscaled solutions in the Model A simulation is considerably less than that in the Model B simulation. This suggests that the upscaling Model A to a 2 × 2 coarse mesh is a justifiable procedure.

In contrast the discretisation error associated with the upscaled Model B simulations would suggest that the coarse mesh should be further refined. The ability to measure the discretisation error is a useful aid in diagnosing how well the simulation has been resolved and indicates that the required coarse mesh resolution depends not only on the fine permeability data, but also the flow imposed by the boundary conditions. From figure 3.6 it is also apparent that the solutions constructed over the original permeability field have a relatively high degree of discretisation error. The higher errors associated with the solution over the fine permeability data are due to the interfacial conditions holding particularly weakly, due to the low number of degrees of freedom in the solution relative to the high number of permeability discontinuities. On the other hand, a coarse approximation constructed over a coarse permeability field can be made to coincide with the permeability discontinuities and the errors associated with satisfying the interfacial conditions will then be reduced. This observation may be viewed as a motivation to upscale: construct a coarse permeability field such that the errors in coarse solutions incurred at the interfaces are minimized. In consistent upscaling this is not a possibility as the method can only ever be as accurate as solving over the fine permeability data using the same basis functions. This limiting accuracy is due to the error, in the form of (3.10) and (3.29), being composed of the sum of two squares of the same sign. The first square represents the discretisation error over the original permeability field and the second square is due to the inconsistency between the numerical solution over the original and upscaled permeability field. The limiting accuracy occurs when a perfectly upscaled permeability field is found such that the second square is zero, leaving only the discretisation error.

The composition of the second square term in (3.10) and (3.29) illuminates the general non-linear nature of upscaling in which Λ^+ is an arithmetic mean weighted by the solution ∇P_h and Λ^- is as a harmonic mean weighted by $\mathbf{Q}_h$. One exception is the piecewise constant case in which upscaling is reduced to a linear problem. This should therefore be considered a powerful method.

The consistency error is highlighted by keeping constant the upscaling resolution and increasing the discretisation resolution. However, in general the error incurred by the numerical solution over the upscaled permeability data is a combination of the consistency error and discretisation error discussed in section 3.2. We term this combination of errors the 'net' error.

3.3.2 Net Error

In the application of an upscaling method the mesh that defines the coarse solution often also defines the upscaled permeability field, and therefore refining the numerical simulation also refines the upscaled permeability field. The convergence of the methods as the mesh that defines both the upscaling and discretisation is refined is shown in figures 3.7 and 3.8, which illustrate the performance of the consistent upscaling methods using both tuned and regular basis functions. The conventional upscaling methods were implemented using regular basis functions.

The performance of the consistent upscaling methods is shown in Figures 3.7 and 3.8. The convergence of the upper bounds to a single value is a result of the upscaling methods being implemented at the same resolution as the fine permeability data. In this situation all upscaling methods should return the fine permeability data and therefore all the methods are equivalent. Similarly for the lower bound. The piecewise constant

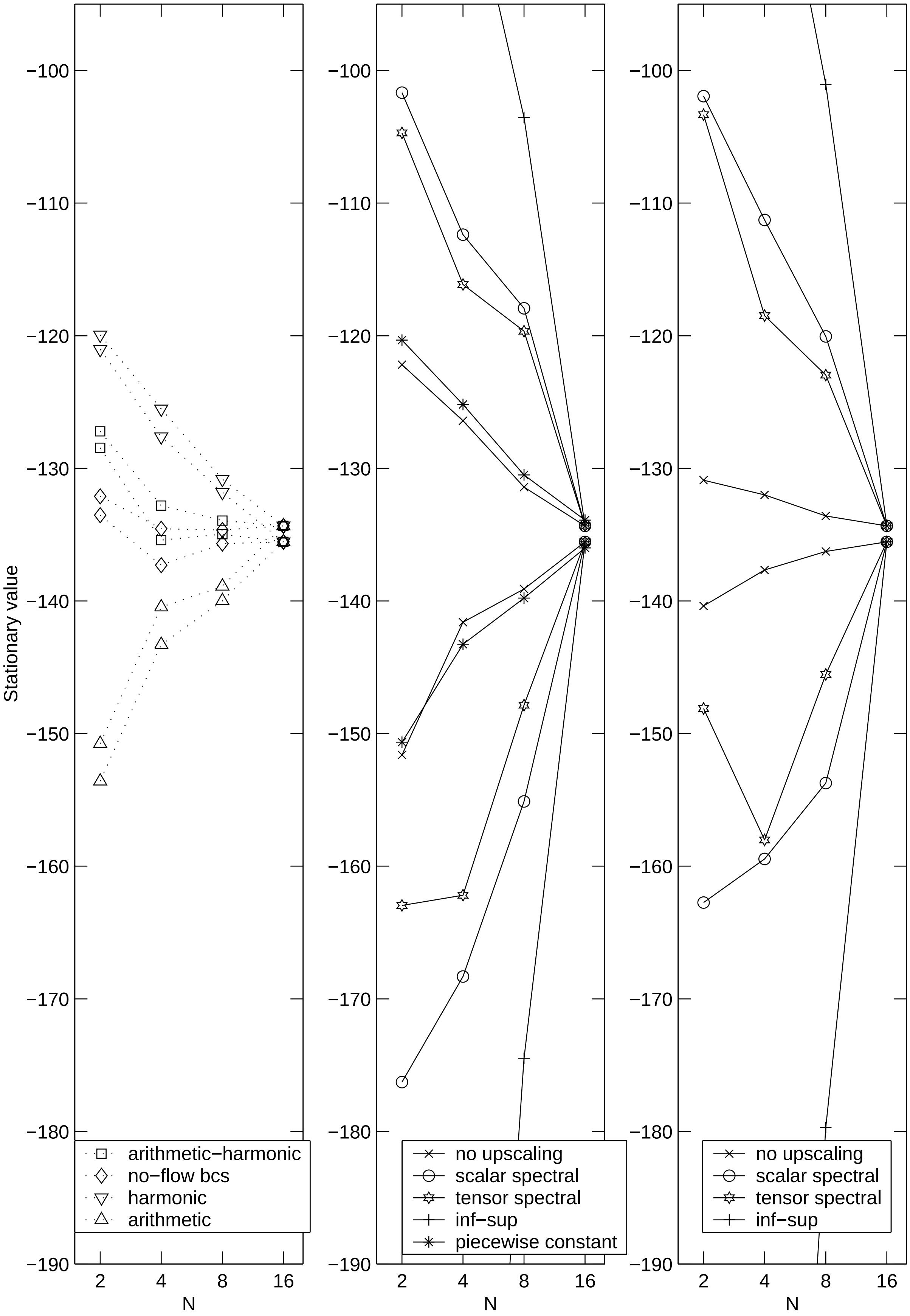

Figure 3.7: Comparison of upscaling methods, Model A

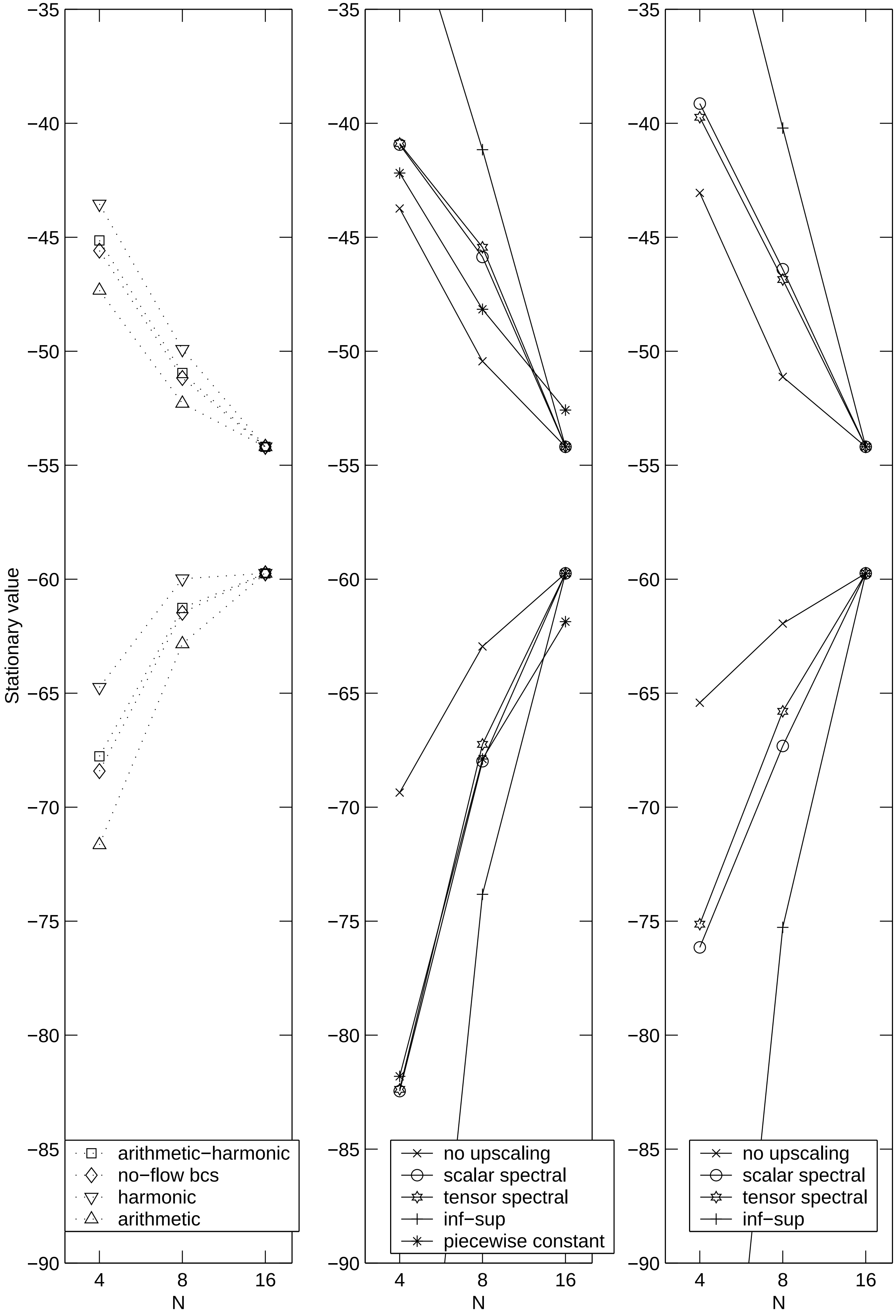

Figure 3.8: Comparison of upscaling methods, Model B

upscaling method converges to a different value since the approximation space spanned by the linear triangular elements differs from that spanned by the bilinear quadrilateral elements. The inf-sup upscaling method was implemented using a piecewise constant expansion for the upscaled permeabilities and from the graphs it can be seen that the method gives very weak bounds. Implementing the method using a piecewise constant expansion is unlikely to achieve adequate performance unless the variations in λ are small. The accuracy of the method would be improved if a higher order expansion , fitting the λ data more closely, were used. The spectral methods are more competitive, with the extra degrees of freedom in the symmetric tensor case helping to tighten the bounds. The restriction of P_h and $\mathbf{Q}_h$ to a known space prior to upscaling enables the spectral method to approach the non-linear cases $\Lambda^{-}(\lambda, P_h)$ and $\Lambda^{+}(\lambda, \mathbf{Q}_h)$ whilst still enabling the upscaling stage to be calculated in advance. In contrast to the inf-sup method the spectral method permits $\Lambda^{-} < \max(\lambda)$ within each coarse element Ω_i, which helps to tighten the bounds. Again, higher order expansions for Λ^{-} and Λ^{+} could also be considered.

The use of the tuned basis functions with the consistent upscaling methods also helps to tighten the bounds, and solving over the original permeability field with the tuned basis functions is particularly effective. Doubling the mesh resolution of the tuned basis functions however does not guarantee an improvement in the solution, as the approximation spaces may not be nested. This may account for the stationary value converging non-monotonically under refinement for the tensor spectral upscaling method shown in figure 3.7. The piecewise constant upscaling method produces good results in comparison with the other consistent methods and is also computationally cheap.

In comparison with the consistent methods the conventional methods obtain greater accuracy in general. In the Model A simulation the methods which were found to have a low consistency error were naturally the better methods. However the flow in the domain is well aligned with x axis, due to the no flow boundary conditions, and in this flow regime the assumptions made in the conventional upscaling methods are reasonably accurate. The consistent method that performed most favourably is the piecewise constant upscaling method, although solving over the original problem with tuned basis functions was effective at producing tight bounds with a low number of degrees of freedom. The flow regime in the Model B simulation has greater complexity and the consistent methods appear more competitive. Interestingly, the conventional methods display 'consistent' characteristics under these flow conditions and retain bounds on the original stationary value. Of the consistent methods the scalar and tensor spectral upscaling using tuned basis functions were the most effective, with the piecewise constant upscaling method also performing well.

In both the Model A and B simulations the no-flow boundary condition $\mathbf{q} \cdot \mathbf{n} = 0$, by definition, aligns portions of the flow with the coordinate system and favours the performance of conventional upscaling methods based on local flow calculations. To remove this trait we chose to consider a domain with periodic boundary conditions along the edges $y = 0, 1$. In addition the permeability range was increased to span a factor of 1×10^3 between the maximum and minimum values.

3.4 A Periodic Test Case

The permeability field and boundary conditions for the periodic test case are shown in figure 3.9. The permeability configuration was constructed to encourage diagonal flow across the domain. The large permeability jumps were intended to test the upscaling methods. The pressure and stream function solutions obtained by solving the periodic

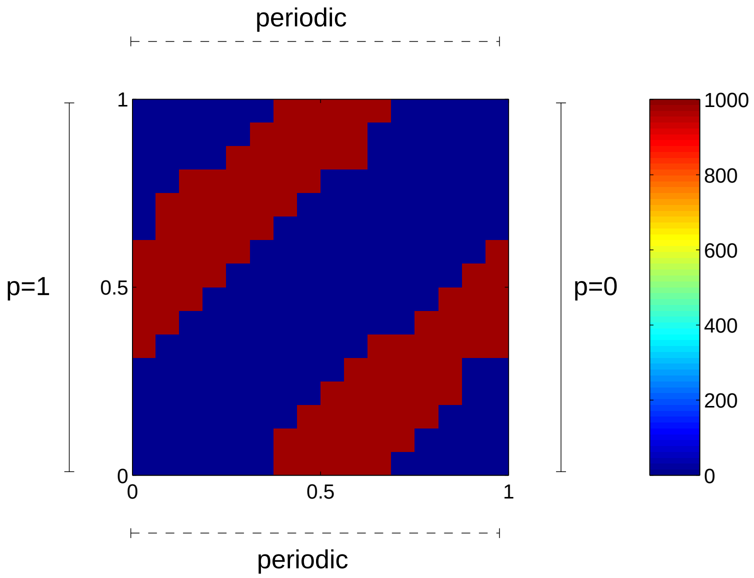

Figure 3.9: The periodic test case

test case with a resolution of 16 × 16 elements is shown in figure 3.10. The consistent

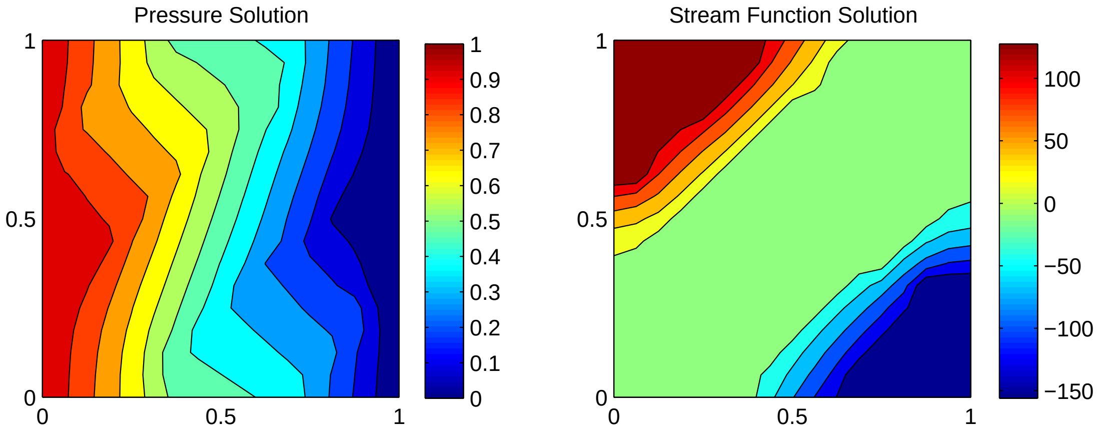

Figure 3.10: Solutions obtained for the periodic test case

upscaling methods were applied to the periodic test case. However it was found that the large jumps in the permeability data distorted the tuned basis functions to an extent that rendered them unusable. Figure 3.11 illustrates the underlying permeability field and a tuned basis function generated by the method described in section 3.2.4. To reduce these distortions prescribed values were assigned along the lines $\xi = 0$ and $\eta = 0$. The prescribed values were calculated as the $1 - D$ solutions of the equations

$$\frac{d}{dl}\left(\bar{\lambda}\frac{d\tilde{\phi}}{dl}\right) = 0 \qquad \text{maximum principle,} \tag{3.37}$$

$$\frac{d}{dl}\left((\bar{\lambda})^{-1}\frac{d\tilde{\phi}}{dl}\right) = 0 \qquad \text{minimum principle,} \tag{3.38}$$

with boundary conditions 1 at the centre and zero on the perimeter. Prescribing the boundary conditions in this manner is similar to the construction on the basis elements employed by Hou and Wu [23]. As the permeability is discontinuous along these lines the arithmetic average of the two values either side of the line was used and is denoted by $\bar{\lambda}$. The one dimensional solutions of (3.37) and (3.38) are monotonic and help to regulate the shape of the basis functions. The one dimensional solutions then act as boundary conditions over each quadrant and the solutions within the quadrants were determined by solving

$$\iint_{\Omega} \nabla\sigma_k \cdot \lambda\nabla\sum_j a_{ij}\sigma_j \, d\Omega = 0 \qquad \forall k \qquad \text{maximum principle,} \tag{3.39}$$

$$\iint_{\Omega} \nabla\sigma_k \cdot \lambda^{-1}\nabla\sum_j a_{ij}\sigma_j \, d\Omega = 0 \qquad \forall k \qquad \text{minimum principle,} \tag{3.40}$$

for the remaining unknown coefficients j. As a result of the classical maximum principle that states that solution extrema exist only on the boundary of the domain of the homogeneous equations (3.39) and (3.40), the basis functions are regularised with the maximum height occurring at (0,0).

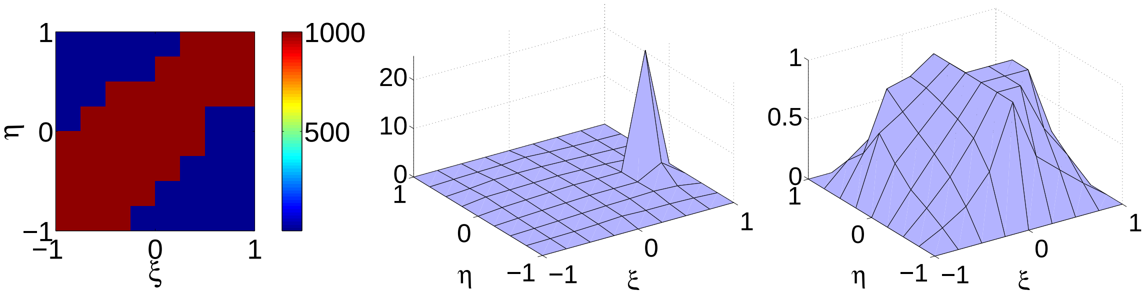

Figure 3.11: The underlying λ data, distorted basis functions and modified basis function

The consistency error and the comparative performance of the methods were again analysed using the same methodology as before. Figure 3.12 illustrates the consistency error incurred by the conventional methods when the fine scale permeability data was upscaled to a 2×2 representation. The magnitude of the consistency error is much higher than those generated by the Model A and Model B simulations. The higher consistency errors reflect the large difference in the arithmetic and harmonic means calculated over the permeability values. The harmonic average is particularly sensitive to any low values which dramatically reduce the average value. Because of this the bounds are very weak unless the upscaling resolves the features of the permeability field. This phenomenon is also observed in the net error, figure 3.13, in which the upper bound is slower to converge than the lower. The piecewise constant upscaling method is again the most competitive consistent method although the lower bound on the no-flow conventional method produces good results. The large discretisation error observed between the bounds in all but the finest simulations indicates that the problem is not ideally suited to upscaling. This is due to the presence of flow features critical to the stationary value that cannot be resolved on a coarse grid. These features are visible in the solution plots, figure 3.10, in the form of flow along the high permeability paths.

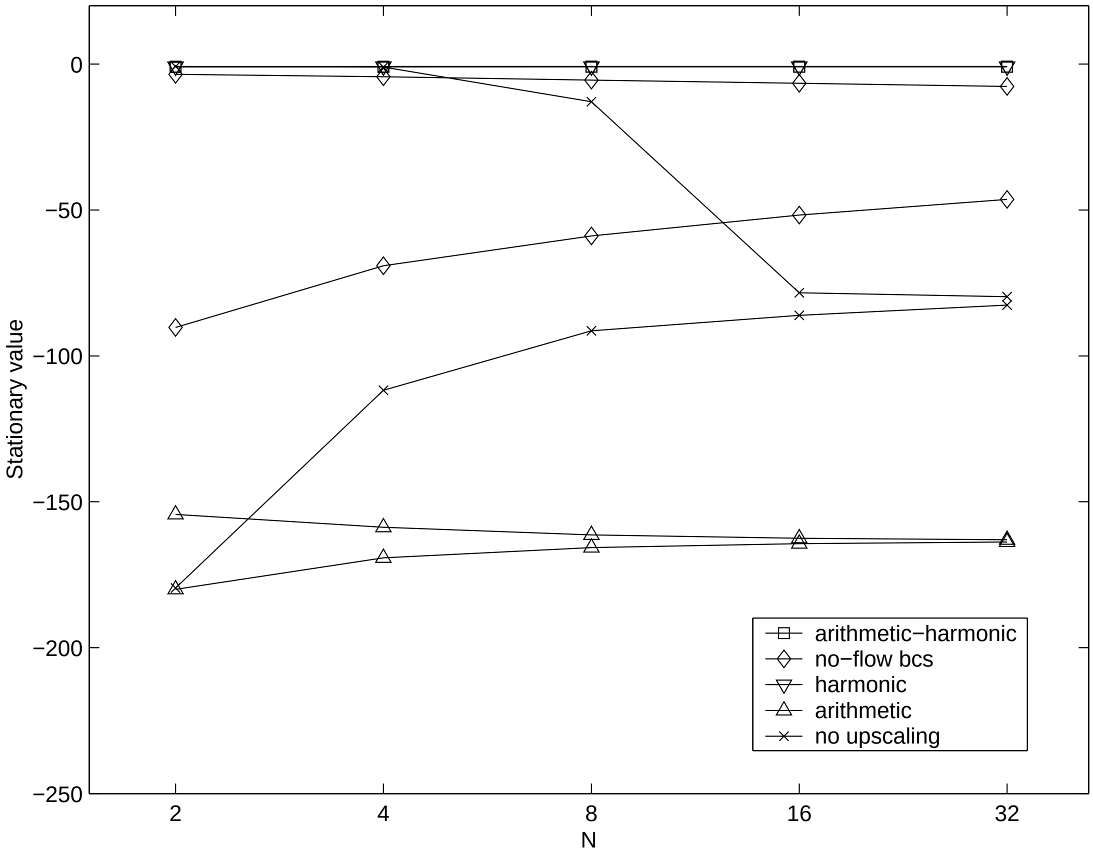

Figure 3.12: Consistency errors, periodic test case

3.5 Conclusions

A consistent upscaling methodology has been constructed based on retaining bounds on the quantity of interest associated with the simulation. The bounds obtained via the method enable the discretisation and consistency errors in the upscaled solution to be quantified and bounded. The availability of these bounds enable the degree of success of an upscaling method to be evaluated. In contrast, error analysis associated with conventional upscaling methods is intractable. The performance of the consistent methods in comparison with the conventional method is not always favourable, but the consistent methods do ensure the retention of the bounds whereas the predictions from conventional methods may drift in complicated flow structures. In addition the tightness of the bounds obtained indicates the degree to which the solution has been resolved. This information is useful in determining the degree of upscaling that can be justified and is not normally obtainable from conventional upscaling methods. In terms of performance and computational cost the piecewise constant upscaling method

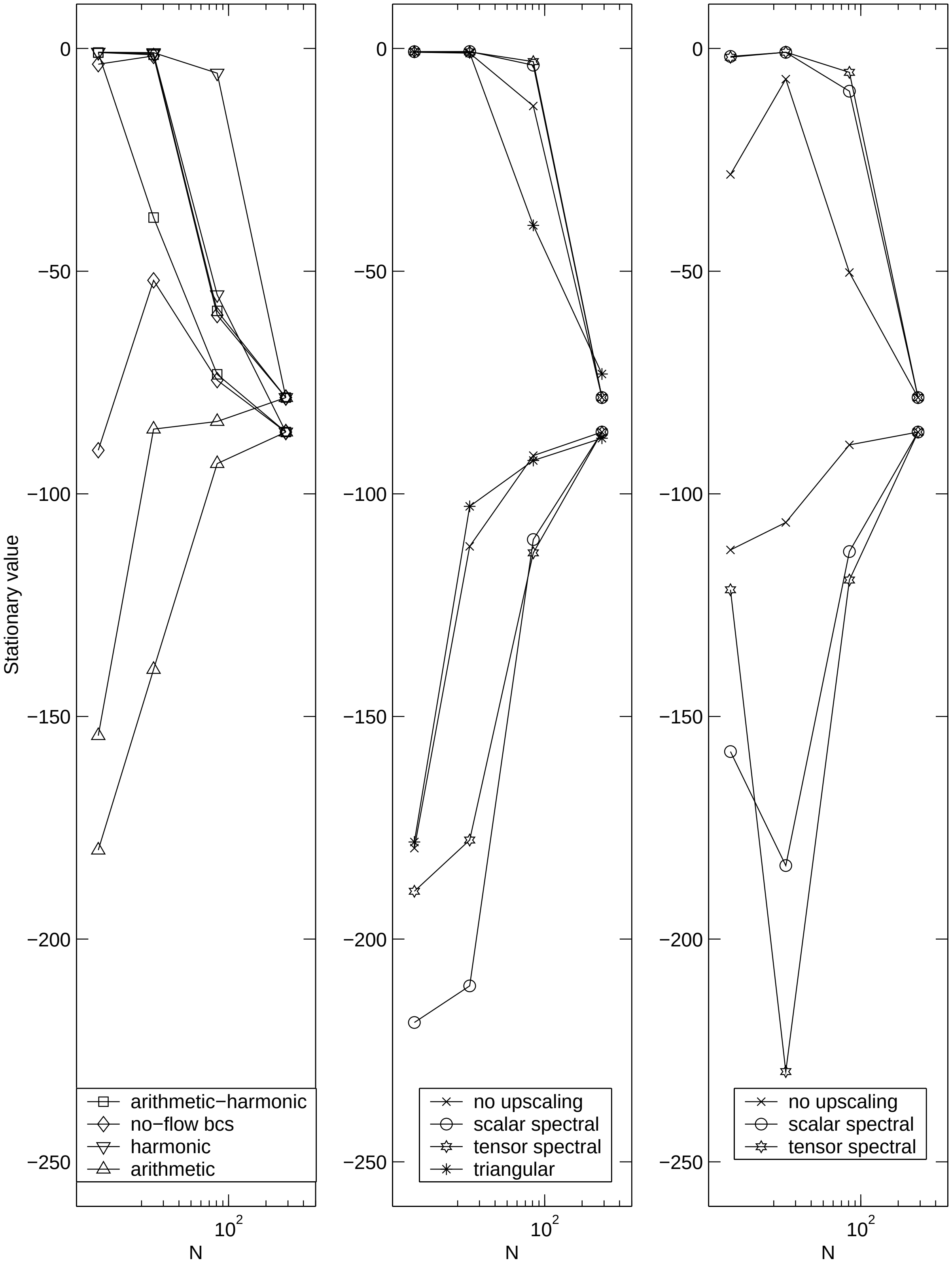

Figure 3.13: Comparison of upscaling methods, periodic test case

is probably the most successful consistent method examined.

A similar description of the consistent upscaling method can be found in [47], published as part of the proceedings of the ICFD 2001 conference held in Oxford.

3.6 Extensions

3.6.1 Permeability Uncertainties

The fine scale permeability is generated largely from statistical data, seismic surveying and core samples. As a result there is a degree of uncertainty in the permeability data which results in uncertainty in the flow predictions. If the uncertainties associated with the finescale permeability data are bounded then the dual extremum principles can be extended by defining λ^- and λ^+ as the lower and upper functions bounding λ respectively. The minimum principle is then constructed using λ^+ and the maximum principle with λ^-. Similarly if a consistent upscaling method is required Λ^+ would be generated using λ^+ as the fine scale data, and correspondingly Λ^- from λ^-. Adopting this approach the upper and lower bounds on the quantity of interest are retained.

Having extended the dual extremum principles to the include uncertainties in the permeability data, the effect of the magnitude and location of the uncertainty can then be investigated. We consider the Model A problem with the area of the domain $0.25 \leq (x, y) \leq 0.75$ subject to varying degrees of uncertainty.

The results obtained by solving directly over λ^+ and λ^- are shown in figure 3.14. For the test case considered, doubling the uncertainty in the permeability data from 10% to 20% also relaxed the bounds by a factor of approximately two. It is expected that results obtained from realisations of the permeability data would lie well within the respective upper and lower bounds. This is due to the natural averaging process in which some values of the realised permeability data would be greater and some less than the original values.

Although the bounds on the quantity of interest are weakened by the degree of uncertainty in the permeability data the variational approach still enables these bounds to

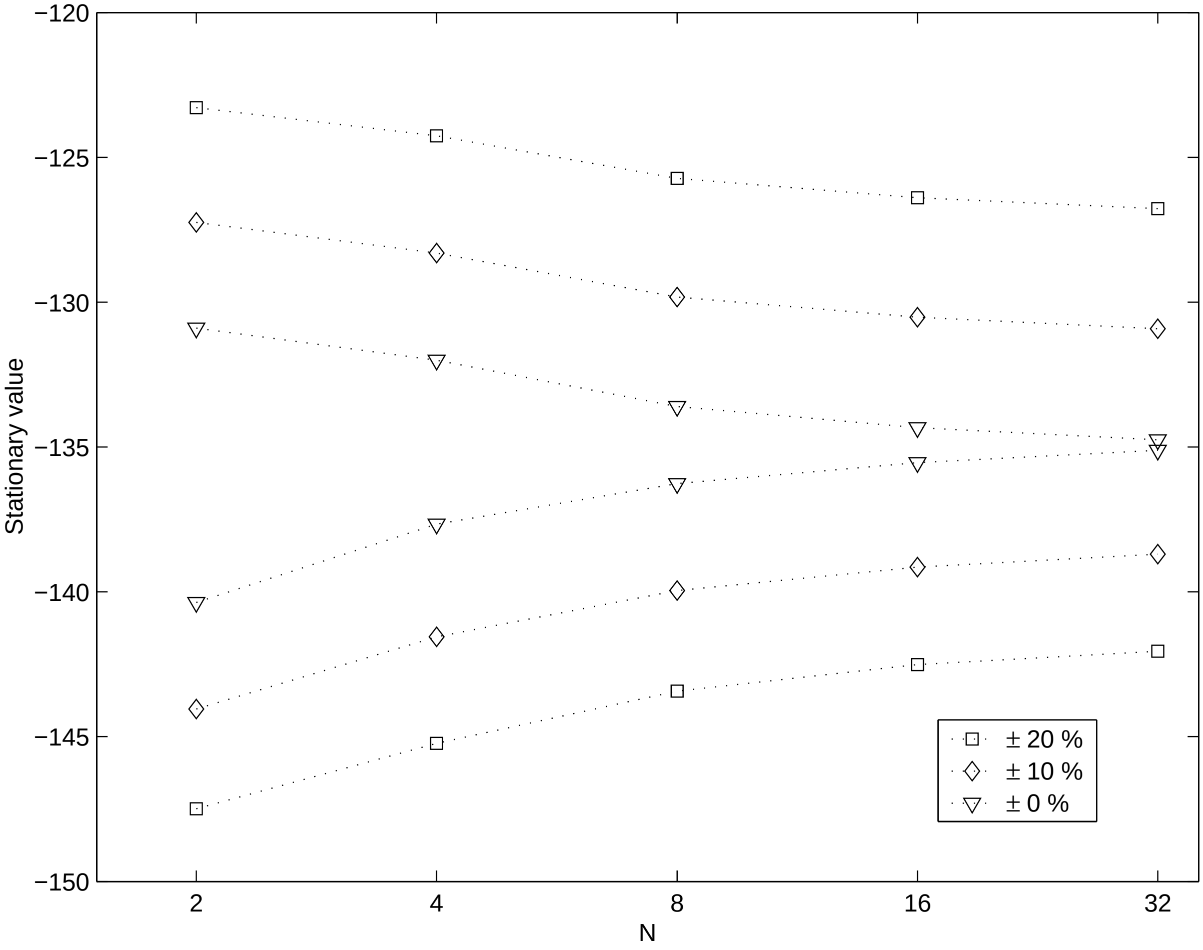

Figure 3.14: Bounds obtained on uncertain permeability data

be found. Alternatively approximate upper and lower bounds on the quantity of interest could be generate by averaging the bounds obtained from many realisations of the permeability data.

3.6.2 Iterated Upscaling Methods

If the conventional sequence of

1. upscale the permeability data

2. solve the flow equations

can be broken, the possibility of iterating the upscaled permeability fields can be addressed. The aim of iterating is to drive the terms

$$\iint_{\Omega} (\Lambda^{-} - \lambda) \nabla P_h \cdot \nabla P_h \, d\Omega \tag{3.41}$$

and

$$\iint_D \left((\Lambda^+)^{-1} - \lambda^{-1}\right) \mathbf{Q}_h \cdot \mathbf{Q}_h \, d\Omega \tag{3.42}$$

to zero. The strategy for the maximum principle is then:

1. Initialise $(\Lambda^-)^n$ using a consistent method, inf-sup for simplicity, $n = 0$.
2. Solve the stationary equations using $(\Lambda^-)^n$ to obtain $P_h^{(n+1)}$
3. Calculate $(\Lambda^-)^{(n+1)}$ from

$$\iint_\Omega ((\Lambda^-)^{(n+1)} - \lambda)\nabla P_h^{(n+1)} \cdot \nabla P_h^{(n+1)} \, d\Omega = 0 \tag{3.43}$$

4. $n = n + 1$, Iterate around loop 2 to 4 until convergence.

A similar algorithm exists for the minimum principle. The results obtained using the algorithm on the Model A problem using regular basis functions is shown in figure 3.15. The convergence of the algorithm is fast and the upper and lower bounds converge, within two iterations, to the results achieved by solving over the original fine permeability data using regular basis functions. Although the method involves iterating, the cost of the method could be efficiently distributed on a parallel machine since stage 3 of the algorithm involves local calculations only and would be well suited to such computations. Solving the stationary equations at stage 2 of the algorithm is a coarse problem and hence should not be excessively expensive.

The iterative method has no application in the piecewise constant upscaling method as the respective terms are already zero. However over the regular grid the triangles can be orientated in one of two ways as shown in figure 3.16. Edge swapping to reduce the error in the solution is therefore a possibility. The aim of the maximum principle is to maximise the functional

$$\mathcal{G}^-(P_h; \Lambda^-) = -\frac{1}{2} \iint_\Omega \Lambda^- \nabla P_h \cdot \nabla P_h \, d\Omega - \int_{\Gamma^+} g P_h \, d\Gamma \tag{3.44}$$

from which we obtain a vector of basis coefficients $\mathbf{P_h}$. Corresponding to the two triangle configurations there are two 4 × 4 stiffness matrices for the square. These matrices are composed of the stiffness matrices for the individual triangles, namely

$$\begin{aligned} K_i^\alpha &= (\Lambda^-)^{\alpha_1} K^{\alpha_1} + (\Lambda^-)^{\alpha_2} K^{\alpha_2}, & (3.45)\\ K_i^\beta &= (\Lambda^-)^{\beta_1} K^{\beta_1} + (\Lambda^-)^{\beta_2} K^{\beta_2}. & (3.46) \end{aligned}$$

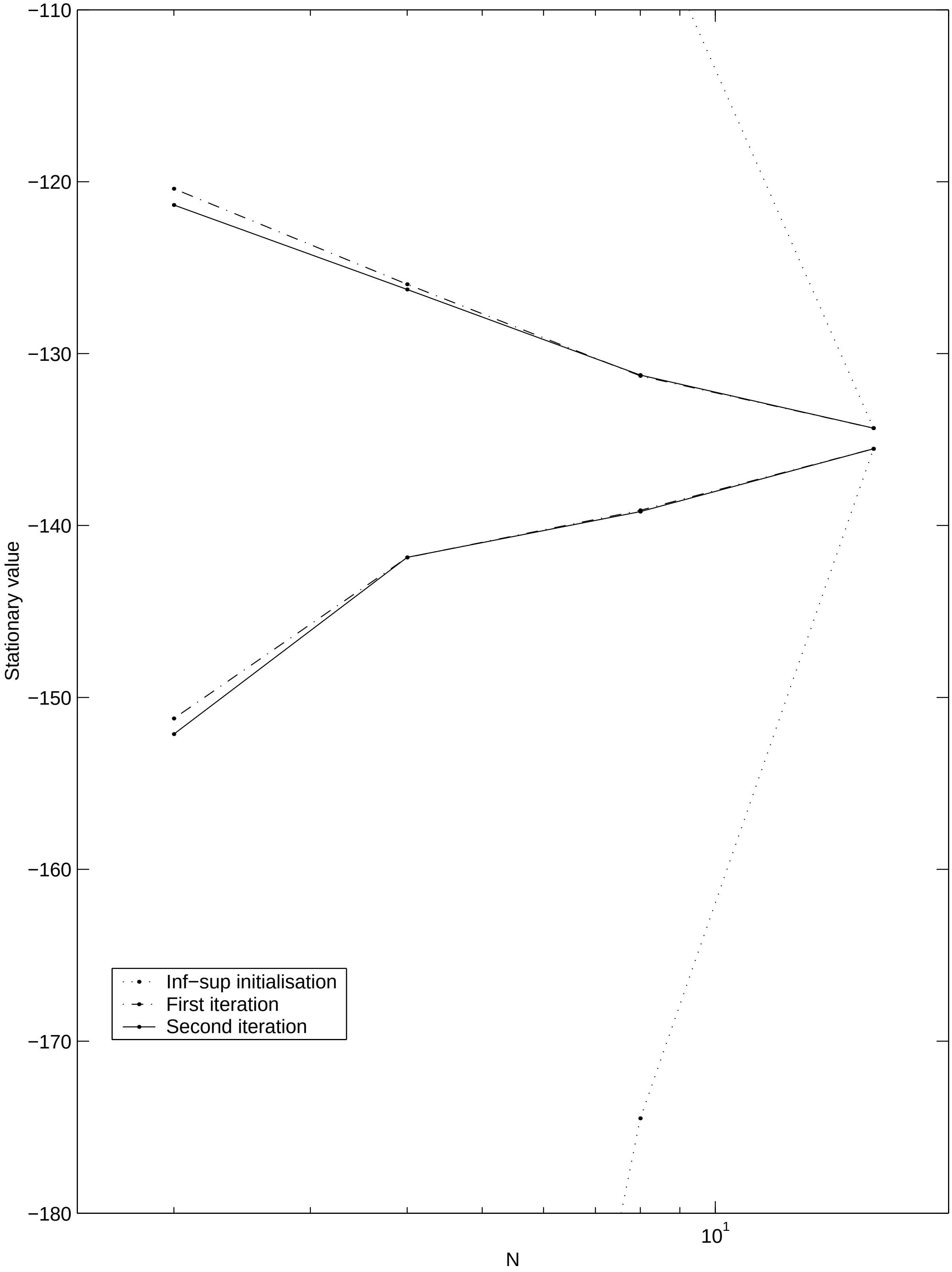

Figure 3.15: Convergence of the iterated upscaling method

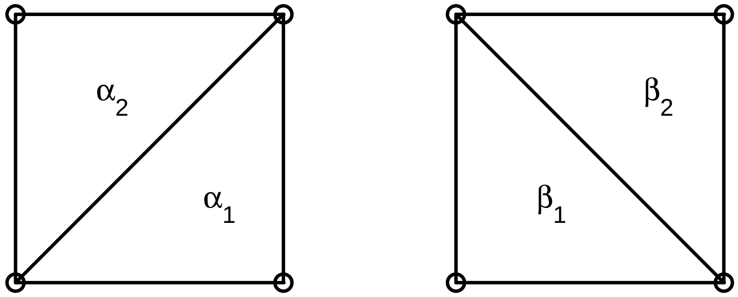

Figure 3.16: The α and β orientations

The procedure is then: solve for the vector $\mathbf{P_h}$ and select the configuration according to the inequalities

$$(\mathbf{P_h})_i^T K_i^\alpha (\mathbf{P_h})_i \quad > \quad (\mathbf{P_h})_i^T K_i^\beta (\mathbf{P_h})_i \qquad \text{Configuration } \alpha \tag{3.47}$$

$$(\mathbf{P_h})_i^T K_i^\beta (\mathbf{P_h})_i \quad \geq \quad (\mathbf{P_h})_i^T K_i^\alpha (\mathbf{P_h})_i \qquad \text{Configuration } \beta \tag{3.48}$$

over each square i. Having re-orientated the triangles the stationary equations are resolved to obtain the lower bound but further iterations were not found necessary. The direct analogue for the minimum principle was also implemented and the results obtained solving Model A are shown in figure 3.17.

3.6.3 Multiphase Upscaling

In industrial reservoir modelling the simulations often involve multiple fluid phases and saturation dependent permeability data. In these circumstances it is unlikely that consistent upscaling methods with upper and lower bounds will exist. However, the guiding philosophy of defining upscaling methods that focus on the quantity of interest could still be experimented with.

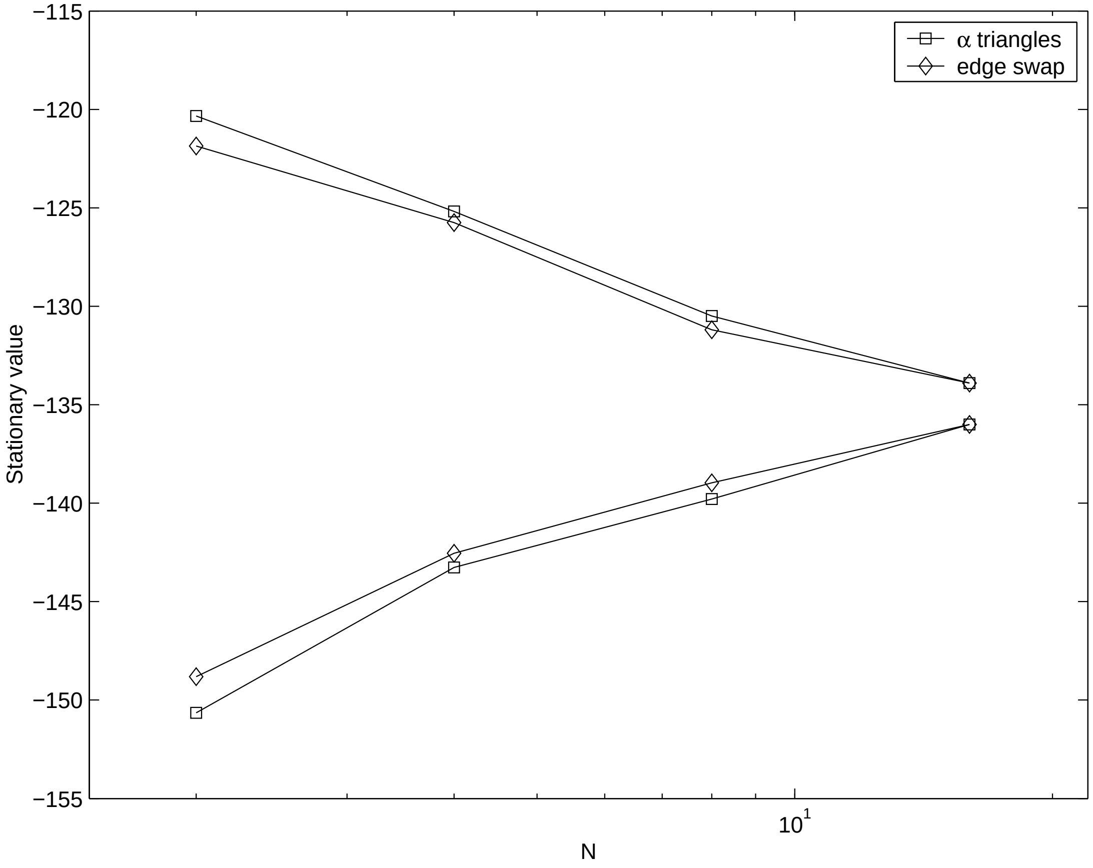

Figure 3.17: Bounds obtained using edge swapping

Chapter 4

A Non-Self-Adjoint Case using Time Discretisation

In the previous chapters bounds were obtained on quantities of interest in which the operator involved is self-adjoint and, although many physical models involve a self adjoint operator, this restriction is limiting. In the following two chapters techniques aimed at obtaining upper and lower bounds on quantities of interest governed by non-self-adjoint operators are explored. To achieve these bounds two approaches are considered.

- Careful discretisation of the non-self adjoint components of the governing equations so as to obtain a self-adjoint semi-discrete system for which dual-extremum principles are applicable.
- Modification of the governing equations so as to introduce self-adjointness into the continuous problem.

This chapter investigates the first method. During the course of this chapter semi-discrete systems are generated for which variational principles provide bounds on integrals of the semi-discrete solution. However, the translation of these bounds from the semi-discrete system to bounds on the analytic solution of the continuous problem does not occur naturally and, in general, the bounds obtained serve merely as approximations to the analytic quantity of interest.

The motivation to consider non-self-adjoint problems is prompted by the advection-diffusion equation, used as a prototype model to simulate the movement of oil in a simplified reservoir model. The advection-diffusion equation governing the evolution of a dissolved solute in a single phase incompressible flow, subject to boundary and initial conditions, is

$$\phi\frac{\partial\widehat{c}}{\partial t} - \nabla\cdot[\mathbf{D}(\widehat{\mathbf{q}})\nabla\widehat{c} - \widehat{\mathbf{q}}\widehat{c}] = e_c, \tag{4.1}$$

where $\widehat{c}(\mathbf{x},t)$ is the concentration of the solute, $\widehat{\mathbf{q}}(\mathbf{x},t)$ is the velocity of the fluid transporting the species, $\phi(\mathbf{x},t)$ is the porosity of the reservoir, $D(\widehat{\mathbf{q}})$ is the diffusion-dispersion tensor and $e_c(\mathbf{x},t)$ represents sources of the species. Again, the hatted functions denote the analytic solutions of the equation.

For simplicity, the evolution of a single chemical species is considered and a value of unity is assigned to the porosity function. Therefore, the full governing equations are,

$$\begin{aligned}
\frac{\partial\widehat{c}(\mathbf{x},t)}{\partial t} - \nabla\cdot[\mathbf{D}(\widehat{\mathbf{q}})\nabla\widehat{c}(\mathbf{x},t) - \widehat{\mathbf{q}}(\mathbf{x})\widehat{c}(\mathbf{x},t)] &= e_c(\mathbf{x},t) \quad \mathbf{x}\in\Omega, && (4.2)\\
\widehat{c}(\mathbf{x},0) &= c_0(\mathbf{x}) \quad \mathbf{x}\in\Omega, && (4.3)\\
\widehat{c}(\mathbf{x},t) &= f_c(\mathbf{x},t) \quad \mathbf{x}\in\Gamma^-, && (4.4)\\
\nabla\widehat{c}(\mathbf{x},t)\cdot\mathbf{n}(\mathbf{x},t) &= g_c(\mathbf{x},t) \quad \mathbf{x}\in\Gamma^+, && (4.5)
\end{aligned}$$

in the spatial domain Ω bounded by the union of the disjoint segments Γ^+ and Γ^-, and where $\mathbf{n}$ is the unit outward normal on this curve. In the oil reservoir context the fluid flux $\widehat{\mathbf{q}}$ is a Darcy velocity obtained from the flow equations

$$\begin{aligned}
\widehat{\mathbf{q}}(\mathbf{x}) &= -\lambda(\mathbf{x})\nabla\widehat{p}(\mathbf{x}) \quad \text{in } \Omega, && (4.6)\\
\nabla\cdot\widehat{\mathbf{q}}(\mathbf{x}) &= 0 \quad \text{in } \Omega, && (4.7)\\
\widehat{p}(\mathbf{x}) &= f(\mathbf{x}) \quad \text{on } \Gamma^-, && (4.8)\\
\widehat{\mathbf{q}}(\mathbf{x})\cdot\mathbf{n}(\mathbf{x}) &= g(\mathbf{x}) \quad \text{on } \Gamma^+, && (4.9)
\end{aligned}$$

considered in the previous chapter, where $\widehat{p}$ is the pressure field associated with the flux $\widehat{\mathbf{q}}$. Provided the permeability λ is independent of the solute concentration, then $\widehat{\mathbf{q}}$ is also independent of the concentration field and can be calculated in advance. Having obtained the solution $\widehat{\mathbf{q}}$, or a suitable approximation to it, the diffusion-dispersion tensor can be calculated and the concentration field, effectively time dependent diffusion of the species along the streamlines, found. If the solution of the flow equations is coupled with the advection-diffusion equation then an iterative scheme may be required. This situation would occur in a multiphase reservoir simulation in which the permeability tensor λ would be dependent on the concentration of each phase. The relationship between the concentration of each species and the permeability function attempts to model the physics governing the displacement of one fluid by another, at a macroscopic level. An explicit alternative to iterating models of this nature is to lag the solution of the advection-diffusion equation behind the solution of the flow equations. By either lagging the advection-diffusion equation or by iterating between the flow and advection-diffusion equation, the direct dependence between the two systems can be removed. Therefore, the case with no interdependencies between the solutions $\widehat{\mathbf{q}}$ and $\widehat{c}$ will be explored.

Although the motivation to consider models governed by non-self-adjoint operators has stemmed from the investigation of novel numerical methods applicable to oil reservoir modelling, the set of equations (4.2)-(4.9) can also be used to simulate the distribution and evolution of a contaminant in an aquifer. Equations of this nature are therefore also of interest to the hydrology community.

The loss of self-adjointness in the equation set (4.2)-(4.5) is due to the presence of the first derivative of the concentration with respect to time and space. A simple progression towards a problem of this nature is to consider time-dependent diffusion in which a single first (time) derivative features. Having developed methods based on the time-dependent diffusion, extensions can be generated to include the advection term. Obtaining bounds on integrals of the solution remains the focus of this chapter. The quantities of interest associated with the prototype reservoir model are integrals of the concentrations over the interior or boundary of the physical domain at a given time. These integrals are sought as they relate directly to the quantity of oil present at these locations and in a commercial context will influence decisions concerning extraction strategies.

4.1 Time Dependent Diffusion

The time-dependent diffusion equation is composed of the self-adjoint diffusion term and the non-self-adjoint first-order time derivative. The strategy of the method will be to retain the properties associated with the self-adjoint term whilst discretising the first-order derivative. To enable comparisons with the previous chapters the same notation will be employed, where $\widehat{p}$ is the analytic solution of the of the problem posed and $\widehat{\mathbf{q}}$ is an intermediate function. The equation governing time-dependent diffusion is then

$$\begin{aligned}
\frac{\partial \widehat{p}(\mathbf{x},t)}{\partial t} - \nabla \cdot D \nabla \widehat{p}(\mathbf{x},t) &= e_c \quad (\mathbf{x},t) \in \Omega, &(4.10)\\
\widehat{p}(\mathbf{x},0) &= p_0(\mathbf{x}) \quad \mathbf{x} \in \Omega, &(4.11)\\
\widehat{p}(\mathbf{x},t) &= f_c(\mathbf{x},t) \quad \mathbf{x} \in \Gamma^-, &(4.12)\\
\nabla \widehat{p}(\mathbf{x},t) \cdot \mathbf{n}(\mathbf{x},t) &= g_c(\mathbf{x},t) \quad \mathbf{x} \in \Gamma^+. &(4.13)
\end{aligned}$$

Continuing the use of the notation introduced in chapter 2, the diffusion operator is again denoted T^*T, and the tensor D is required to be symmetric, positive and considered independent of the solution p in order that this splitting can be found. Re-writing the governing equations in the general notation in which the operator T^*T and the function s include boundary terms, the problem

$$\begin{aligned}
\frac{\partial \widehat{p}(\mathbf{x},t)}{\partial t} + T^*T\widehat{p}(\mathbf{x},t) &= s(\mathbf{x},t), &(4.14)\\
\widehat{p}(\mathbf{x},t) &= p_0(\mathbf{x}), &(4.15)
\end{aligned}$$

is obtained. The aim is to construct a numerical method to solve (4.14-4.15). However, the operator T^*T has been shown to have useful properties and discretising this component is initially deferred. Instead, an implicit discretisation in time is made and a theta method is applied to the continuous spatial terms. Applying this discretisation, the Rothe method

$$\frac{\widehat{p}(\mathbf{x})^{t+\Delta t} - \widehat{p}(\mathbf{x})^t}{\Delta t} = -\theta\, T^*T\widehat{p}(\mathbf{x})^{t+\Delta t} - (1-\theta)\, T^*T\widehat{p}(\mathbf{x})^t \qquad 0 \le \theta \le 1 \qquad (4.16)$$

is obtained. The numerical method (4.16) can be implemented to advance the solution forward in time from the initial condition p_0. Rothe methods are the converse to the Method of Lines. In the Method of Lines a spatial discretisation is initially applied to the problem and the resulting discrete system is integrated in time. In the Rothe Method a temporal discretisation is first implemented and the semi-discrete system is then solved for each time step. A description and an application of the Rothe method can be found in [43].

Interestingly (4.16) can be re-arranged to obtain the Helmholtz equation governing the solution at the next time step, $t + \Delta t$,

$$\begin{aligned}
r &= T^*T\widehat{p}^{t+\Delta t} + \kappa \widehat{p}^{t+\Delta t}, &(4.17)\\
&= (T^*T + \kappa I)\widehat{p}^{t+\Delta t}, &(4.18)
\end{aligned}$$

by introducing the following substitutions,

$$\kappa = \frac{1}{\theta\,\Delta t}, \tag{4.19}$$

$$r = \kappa\widehat{p}^t - \frac{1-\theta}{\theta}T^*T\widehat{p}^t. \tag{4.20}$$

Crucially, the Helmholtz equation has a variational formulation and associated extremum principles. In addition, the appearance of the identity in the Helmholtz operator will be found to greatly simplify the problem of obtaining an upper bound on the stationary value of the functional.

4.1.1 The Helmholtz Operator in General Notation

To illuminate the mechanics of the Helmholtz functional, and contrast it with the diffusion functional, the general operator notation of the previous chapters will again be employed. In the general operator notation the solution $\widehat{p}$ of the Helmholtz equation,

$$T^*T\widehat{p} + \kappa\widehat{p} = r, \tag{4.21}$$

is found to coincide with the stationary point of the functional

$$\mathcal{G}(p,q) = \frac{1}{2}\langle\langle q,q\rangle\rangle - \langle\langle Tp,q\rangle\rangle - \frac{\kappa}{2}\langle p,p\rangle + \langle p,r\rangle. \tag{4.22}$$

The coincidence of the solution $\widehat{p}$ and the stationary point of the $\mathcal{G}(p,q)$ is again demonstrated by considering the first order variation of the functional. This is

$$\delta\mathcal{G}(p,q) = \langle\langle \delta q, q - Tp\rangle\rangle - \langle \delta p, T^*q + \kappa p - r\rangle, \tag{4.23}$$

and therefore the functional is stationary at the point $(p,q) = (\widehat{p},\widehat{q})$, with the natural conditions

$$T\widehat{p} = \widehat{q} \qquad (H^-), \tag{4.24}$$

$$T^*\widehat{q} + \kappa\widehat{p} = r \qquad (H^+), \tag{4.25}$$

equivalent to (4.21). The stationary value of the functional is found by substituting the stationary conditions into the functional to obtain

$$\begin{aligned}
\mathcal{G}(\widehat{p},\widehat{q}) &= \frac{1}{2}\langle\langle \widehat{q},\widehat{q}\rangle\rangle - \langle\langle T\widehat{p},\widehat{q}\rangle\rangle - \frac{\kappa}{2}\langle \widehat{p},\widehat{p}\rangle + \langle \widehat{p},r\rangle, && (4.26)\\
&= \frac{1}{2}\langle\langle \widehat{q},\widehat{q}\rangle\rangle - \langle\langle \widehat{p},r-\kappa\widehat{p}\rangle\rangle - \frac{\kappa}{2}\langle \widehat{p},\widehat{p}\rangle + \langle \widehat{p},r\rangle, && (4.27)\\
&= \frac{1}{2}\langle\langle \widehat{q},\widehat{q}\rangle\rangle + \frac{\kappa}{2}\langle \widehat{p},\widehat{p}\rangle, && (4.28)\\
&= \frac{1}{2}\langle \widehat{p},r\rangle. && (4.29)
\end{aligned}$$

The stationary value of the functional, as expected, is the inner product of the solution with the forcing function. Although the form of stationary value is similar to that of the diffusion functional the appearance of the identity in the Helmholtz operator enables both natural conditions, (4.24) and (4.25), to be used as constraints directly. Essentially the occurrence of the identity enables

$$T^{*}q + \kappa p = r \tag{4.30}$$

to be solved for p in terms of q through the rearrangement

$$p = \frac{1}{\kappa}\left(r - T^{*}q\right). \tag{4.31}$$

Similarly the function q can be expressed in term of p through (4.24). The ability to express p in terms of q, and vice versa, enables both natural conditions to be substituted into the functional $\mathcal{G}(p, q)$ in turn. The two hyper-lines on which the constrained functionals lie are shown in figure 4.1 and they are found to be sufficient to generate upper and lower bounds on the stationary value. In contrast with the constraints required for the diffusion functional, shown in figure 2.1, both of the Helmholtz constraints are relatively easy to satisfy and the difficulties associated with inverting the comparatively degenerate H^{+} constraint, $T^{*}q = r$, are avoided.

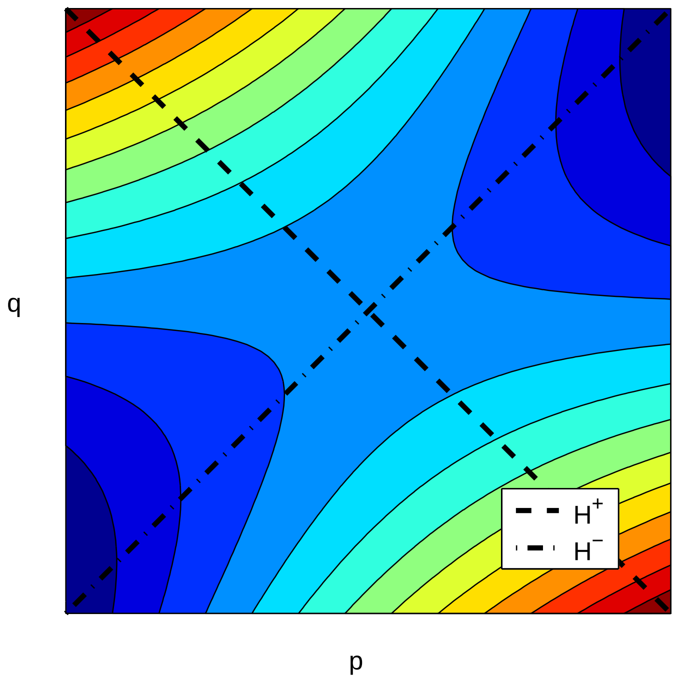

Figure 4.1: The Helmholtz functional

4.1.2 Dual Extremum Principles for the Helmholtz Equation

The dual extremum principles associated with obtaining bounds on the stationary value of the Helmholtz functional $\mathcal{G}(\widehat{p},\widehat{q})$ can be found in the books of Arthurs [3] and Sewell [44]. Generating the constrained functionals and demonstrating the required convexity properties is achieved in a similar manner to those of the diffusion functional described in section 2.1.3. In the general operator notation, substituting for q using the natural condition (4.24) the functional is constrained to lie on the hyperline H^-, and the functional $\mathcal{G}^-(p)$, where

$$\mathcal{G}^-(p) = -\frac{1}{2}\langle\langle Tp, Tp\rangle\rangle - \frac{\kappa}{2}\langle p,p\rangle + \langle p,r\rangle \tag{4.32}$$

is generated. $\mathcal{G}^-(p)$ is a lower bound on the stationary value of the functional since

$$\begin{aligned}
\mathcal{G}^-(\widehat{p}) - \mathcal{G}^-(p) &= -\frac{1}{2}\langle\langle T\widehat{p}, T\widehat{p}\rangle\rangle - \frac{\kappa}{2}\langle \widehat{p},\widehat{p}\rangle + \langle \widehat{p},r\rangle \\
&\quad + \frac{1}{2}\langle\langle Tp, Tp\rangle\rangle + \frac{\kappa}{2}\langle p,p\rangle - \langle p,r\rangle, \\
&= \frac{1}{2}\langle\langle T\widehat{p}, T\widehat{p}\rangle\rangle + \frac{\kappa}{2}\langle \widehat{p},\widehat{p}\rangle + \frac{1}{2}\langle\langle Tp, Tp\rangle\rangle \\
&\quad \frac{\kappa}{2}\langle p,p\rangle - \frac{1}{2}\langle\langle T\widehat{p}, Tp\rangle\rangle - \frac{\kappa}{2}\langle \widehat{p},p\rangle, \\
&= \frac{1}{2}\langle\langle\, T(\widehat{p}-p), T(\widehat{p}-p)\,\rangle\rangle + \frac{\kappa}{2}\langle \widehat{p}-p, \widehat{p}-p\rangle, \\
&= \frac{1}{2}\|T(\widehat{p}-p)\|^2_{\langle\langle\rangle\rangle} + \frac{\kappa}{2}\|\widehat{p}-p\|^2_{\langle\rangle}, \\
&\geq 0.
\end{aligned} \tag{4.33}$$

Similarly, substituting for p using the natural condition (4.31) the functional is constrained to lie on the hyperline H^+ and the functional $\mathcal{G}^+(p)$, where

$$\mathcal{G}^+(q) = \frac{1}{2}\langle\langle q,q\rangle\rangle + \frac{1}{2\kappa}\langle T^*q, T^*q\rangle + \frac{1}{2\kappa}\langle r,r\rangle - \frac{1}{\kappa}\langle r, T^*q\rangle, \tag{4.34}$$

is generated. $\mathcal{G}^+(q)$ is an upper bound on the stationary value of the functional, since

$$\begin{aligned}
\mathcal{G}^+(q) - \mathcal{G}^+(\widehat{q}) &= \frac{1}{2}\langle\langle q,q\rangle\rangle + \frac{1}{2\kappa}\langle T^*q, T^*q\rangle - \frac{1}{\kappa}\langle r, T^*q\rangle, \\
&\quad - \frac{1}{2}\langle\langle \widehat{q},\widehat{q}\rangle - \frac{1}{2\kappa}\langle T^*\widehat{q}, T^*\widehat{q}\rangle + \frac{1}{\kappa}\langle r, T^*q\rangle, \qquad (4.35) \\
&= \frac{1}{2}\langle\langle q,q\rangle\rangle + \frac{1}{2\kappa}\langle\langle T^*q, T^*q\rangle\rangle - \frac{1}{\kappa}\langle T^*\widehat{q} + \kappa\widehat{p}, T^*q\rangle, \\
&\quad - \frac{1}{2}\langle\langle \widehat{q},\widehat{q}\rangle\rangle - \frac{1}{2\kappa}\langle T^*\widehat{q}, T^*\widehat{q}\rangle + \frac{1}{\kappa}\langle T^*\widehat{q} + \kappa\widehat{p}, T^*\widehat{q}\rangle, \qquad (4.36) \\
&= \frac{1}{2}\langle\langle q,q\rangle\rangle + \frac{1}{2\kappa}\langle T^*q, T^*q\rangle - \langle\langle \widehat{q},q\rangle\rangle,
\end{aligned}$$

$$
\begin{aligned}
& -\frac{1}{\kappa}\langle T^*\widehat{q}, T^*q\rangle + \frac{1}{2}\langle\!\langle \widehat{q}, \widehat{q}\rangle\!\rangle + \frac{1}{2\kappa}\langle T^*\widehat{q}, T^*\widehat{q}\rangle, && (4.37)\\
= \;& \frac{1}{2\kappa}\langle T^*(\widehat{q}-q), T^*(\widehat{q}-q)\rangle + \frac{1}{2}\langle\!\langle \widehat{q}-q, \widehat{q}-q\rangle\!\rangle, && (4.38)\\
= \;& \frac{1}{2\kappa}\|T^*(\widehat{q}-q)\|^2_{\langle\rangle} + \frac{1}{2}\|\widehat{q}-q\|^2_{\langle\!\langle\rangle\!\rangle}, && (4.39)\\
\geq \;& 0. && (4.40)
\end{aligned}
$$

The beneficial nature of the upper bound associated with the Helmholtz operator, in the sense that the constraint H^+ is explicitly satisfied, encourages the limit $\kappa \to 0$ to be considered as a means of obtaining an alternative upper bound on the stationary value of the diffusion functional. This limit $\kappa \to 0$ is found by considering either increasingly large time steps, or an increasingly explicit scheme. However, regardless of how well these extremes resolve the solution of the time-dependent diffusion problem, the constraint (4.31) and functional $\mathcal{G}^+(q)$ become ill posed as κ diminishes. In contrast the lower bound remains valid as $\kappa \to 0$ due to the straightforward convergence of the constrained Helmholtz functional $\mathcal{G}^-(p)$

$$\mathcal{G}^-(p) = -\frac{1}{2}\langle\!\langle Tp, Tp\rangle\!\rangle - \frac{\kappa}{2}\langle p, p\rangle + \langle p, r\rangle \tag{4.41}$$

to that of the constrained diffusion functional

$$\mathcal{G}^-(p) = -\frac{1}{2}\langle\!\langle Tp, Tp\rangle\!\rangle + \langle p, r\rangle. \tag{4.42}$$

The behaviour of the Helmholtz functional with decreasing κ is illustrated in the example

$$
\begin{aligned}
-\frac{d^2\widehat{p}}{dx^2} + \kappa\widehat{p} = 1, && (4.43)\\
\widehat{p}(0) = 0, && (4.44)\\
\frac{d\widehat{p}(1)}{dx} = 0. && (4.45)
\end{aligned}
$$

The analytic solution $\widehat{p}$ of this example is

$$\widehat{p} = c_1 e^{\sqrt{k}x} + c_2 e^{-\sqrt{k}x} + \kappa^{-1}x \tag{4.46}$$

where

$$
\begin{aligned}
c_1 &= -\frac{\kappa^{-3/2}}{e^{\sqrt{k}} + e^{-\sqrt{k}}}, && (4.47)\\
c_2 &= \frac{\kappa^{-3/2}}{e^{\sqrt{k}} + e^{-\sqrt{k}}}, && (4.48)\\
& && (4.49)
\end{aligned}
$$

and therefore the analytic stationary value of the functional, corresponding to $\frac{1}{2}\langle \widehat{p}, r\rangle$, is computable and expressed as

$$\mathcal{G}(\widehat{p},\widehat{q}) = \frac{1}{2}\int_0^1 \widehat{p}x\,dx \tag{4.50}$$

$$= \frac{1}{2}\int_0^1 (c_1 e^{\sqrt{k}x} + c_2 e^{-\sqrt{k}x} + \kappa^{-1}x)x\,dx. \tag{4.51}$$

The upper and lower bounds obtained by maximising and minimising $\mathcal{G}^-(p)$ and $\mathcal{G}^+(q)$ respectively, using two linear elements, are shown in figure 4.2. The example and discretisation were chosen in order that neither $\widehat{p}$ or $\widehat{q}$ lie in the approximation space and therefore upper and lower bounds sensitive to κ can be found. The predicted ill-conditioning of the functional $\mathcal{G}^+(q)$ as κ nears zero is apparent in the increasing error observed in the upper bound in this region of the graph. Conversely, the lower bound retains integrity as κ diminishes, as suggested by the convergence of the Helmholtz functional to that of the diffusion functional.

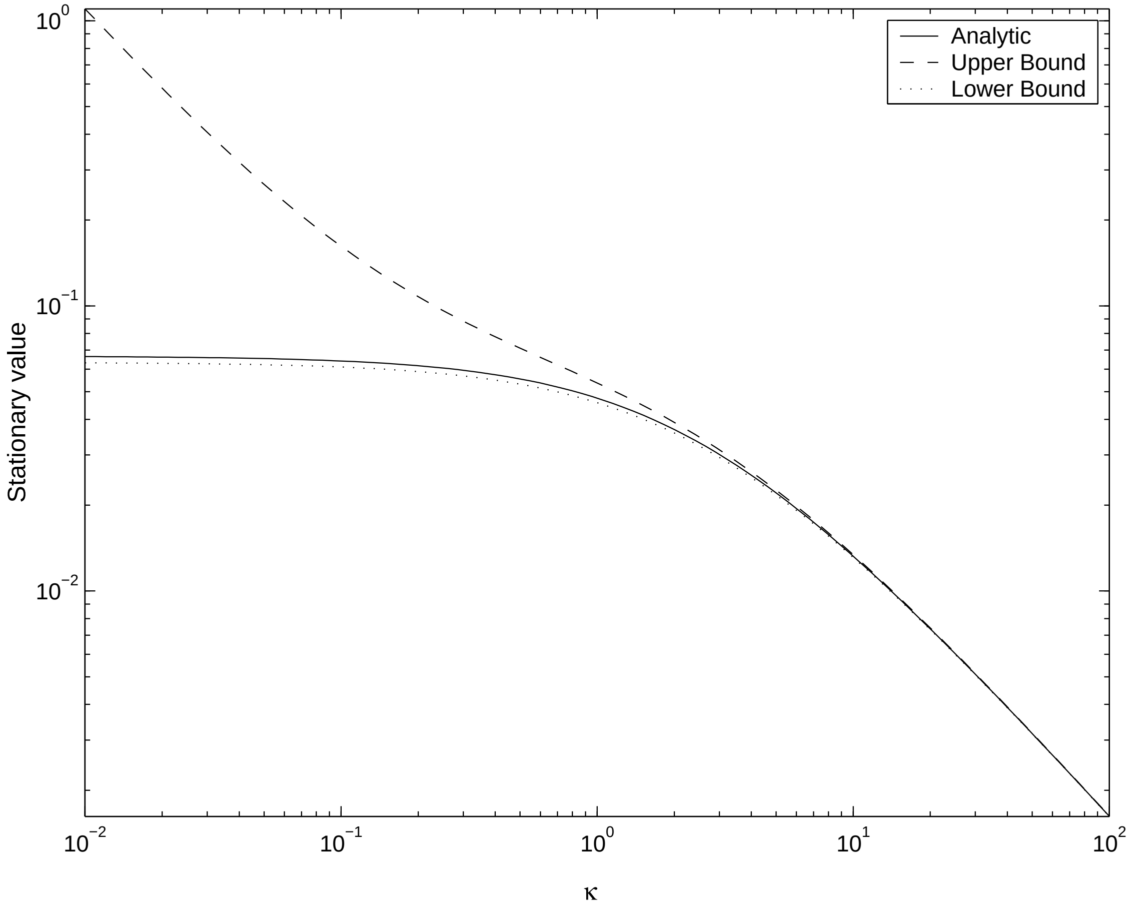

Figure 4.2: Behaviour of the bounds on the stationary value as $\kappa \to 0$

The example illustrates the impracticalities of obtaining an upper bound on the stationary value of the diffusion functional, $\kappa = 0$, by considering the Helmholtz functional in the limit $\kappa \to 0$.

Having demonstrated the Helmholtz functional in the general notation a description in terms of the divergence and gradient operators is now given. The functional will enable comparisons to be made between the bounds obtained by using the Helmholtz equation to solve approximations to the governing equations at each time step and an analytic solution of the time-dependent diffusion problem.

4.1.3 The Helmholtz Functional

In terms of the divergence and gradient operators, the functional that is stationary at the solution of the Helmholtz equation,

$$\begin{aligned}
\widehat{\mathbf{q}}(\mathbf{x}) &= -D(\mathbf{x})\nabla\widehat{p}(\mathbf{x}) \quad \text{in } \Omega, && (4.52)\\
\nabla\cdot\widehat{\mathbf{q}}(\mathbf{x}) + \kappa\widehat{p}(\mathbf{x}) &= e \quad \text{in } \Omega, && (4.53)\\
\widehat{p}(\mathbf{x}) &= f(\mathbf{x}) \quad \text{on } \Gamma^-, && (4.54)\\
\widehat{\mathbf{q}}(\mathbf{x})\cdot\mathbf{n}(\mathbf{x}) &= g(\mathbf{x}) \quad \text{on } \Gamma^+, && (4.55)
\end{aligned}$$

is

$$\mathcal{G}(p,\mathbf{q}) = \iint_D \left\{\frac{D^{-1}}{2}\mathbf{q}\cdot\mathbf{q} + \mathbf{q}\cdot\nabla p - \frac{\kappa}{2}p^2 + pe\right\} d\Omega - \int_{C_p}(p-f)\mathbf{q}\cdot\mathbf{n}\,d\Sigma - \int_{C_q} pg\,d\Sigma, \tag{4.56}$$

where D is a given symmetric positive definite tensor and κ is known. In this chapter a comparison problem between differing diffusion-dispersion functions is not constructed and therefore the semicolon notation of the previous chapter is omitted. The coincidence of the stationary point of the functional $\mathcal{G}(p,\mathbf{q})$ and the set of equations (4.52)-(4.55), is demonstrated by considering the first variation of $\mathcal{G}(p,\mathbf{q})$ which is found to be

$$\begin{aligned}
\delta\mathcal{G}(p,\mathbf{q}) &= \iint_D \left\{\delta p(-\nabla\cdot\mathbf{q} + e - \kappa p) + \delta\mathbf{q}\cdot\left(D^{-1}\mathbf{q} + \nabla p\right)\right\} d\Omega\\
&\quad - \int_{C_p}(p-f)\delta\mathbf{q}\cdot\mathbf{n}\,d\Sigma + \int_{C_q}\delta p(\mathbf{q}\cdot\mathbf{n} - g)\,d\Sigma. && (4.57)
\end{aligned}$$

The stationary value of the functional, as with the general notation, is found by substituting the stationary conditions into the functional to give the integral of the analytic solution weighted by the boundary and source terms, namely

$$\begin{aligned}
\mathcal{G}(\widehat{p},\widehat{\mathbf{q}}) &= \iint_{\Omega} \left\{ \frac{D^{-1}}{2}\widehat{\mathbf{q}}\cdot\widehat{\mathbf{q}} + \widehat{\mathbf{q}}\cdot\nabla\widehat{p} - \frac{\kappa}{2}\widehat{p}^2 + \widehat{p}e \right\} d\Omega - \int_{\Gamma^-} (\widehat{p}-f)\widehat{\mathbf{q}}\cdot\mathbf{n}\,d\Gamma - \int_{\Gamma^+} \widehat{p}g\,d\Gamma, \\
&= \iint_{\Omega} \left\{ \frac{1}{2}\widehat{\mathbf{q}}\cdot\nabla\widehat{p} - \frac{\kappa}{2}\widehat{p}^2 + \widehat{p}e \right\} d\Omega - \int_{\Gamma^+} \widehat{p}g\,d\Gamma, \\
&= \iint_{\Omega} \left\{ \frac{1}{2}\widehat{p}(-\nabla\cdot\widehat{\mathbf{q}} - \kappa\widehat{p} + e) + \frac{1}{2}\widehat{p}e \right\} d\Omega - \int_{\Gamma^+} \widehat{p}g\,d\Gamma + \frac{1}{2}\int_{\Gamma} \widehat{p}\widehat{\mathbf{q}}\cdot\mathbf{n}\,d\Gamma, \\
&= \frac{1}{2}\iint_{\Omega} \widehat{p}e\,d\Omega + \frac{1}{2}\int_{\Gamma^-} f\widehat{\mathbf{q}}\cdot\mathbf{n}\,d\Gamma - \frac{1}{2}\int_{\Gamma^+} \widehat{p}g\,d\Gamma. \qquad (4.58)
\end{aligned}$$

The upper and lower bounds are found by constraining the functional $\mathcal{G}(p,\mathbf{q})$ by the corresponding H^+ and H^- natural conditions and, as illustrated by the derivation in the general notation, both sets of constraints can be directly substituted into the functional. To obtain a lower bound on the stationary value of the functional, the natural conditions

$$\left.\begin{aligned}
-D\nabla p &= \mathbf{q} \qquad \text{in } \Omega \\
p &= f \qquad \text{on } \Gamma^-
\end{aligned}\right\} H^- \qquad (4.59)$$

are applied as constraints. This is achieved by making the substitution (4.59) for q, and using an approximation space satisfying the Dirichlet boundary condition $p = f$. The functional $\mathcal{G}(p,\mathbf{q})$ then reduces to $\mathcal{G}^-(p)$, where

$$\mathcal{G}^-(p) = \iint_{\Omega} \left\{ -\frac{1}{2}D\nabla p\cdot\nabla p - \frac{\kappa}{2}p^2 + pe \right\} d\Omega - \int_{\Gamma^+} pg\,d\Gamma, \qquad (4.60)$$

which is a lower bound on the stationary value of the functional, since

$$\begin{aligned}
\mathcal{G}^-(\widehat{p}) - \mathcal{G}^-(p) &= \iint_{\Omega} \left\{ \frac{D}{2}\nabla p\cdot\nabla p + \frac{\kappa}{2}p^2 - ep - \frac{D}{2}\nabla\widehat{p}\cdot\nabla\widehat{p} - \frac{\kappa}{2}\widehat{p}^2 + e\widehat{p} \right\} d\Omega \\
&\quad + \int_{\Gamma^+} g(p-\widehat{p})\,d\Gamma, \\
&= \iint_{\Omega} \left\{ \frac{D}{2}\nabla p\cdot\nabla p + \frac{\kappa}{2}p^2 - (-\nabla\cdot D\nabla\widehat{p} + \kappa\widehat{p})p - \frac{D}{2}\nabla\widehat{p}\cdot\nabla\widehat{p} \right. \\
&\quad \left. -\frac{\kappa}{2}\widehat{p}^2 + (-\nabla\cdot D\nabla\widehat{p} + \kappa\widehat{p})\widehat{p} \right\} d\Omega - \int_{\Gamma} \mathbf{q}(p-\widehat{p})\cdot\mathbf{n}\,d\Gamma, \\
&= \iint_{\Omega} \left\{ \frac{D}{2}\nabla p\cdot\nabla p + \frac{\kappa}{2}p^2 - D\nabla\widehat{p}\cdot\nabla p - \kappa\widehat{p}p \right. \\
&\quad \left. +\frac{D}{2}\nabla\widehat{p}\cdot\nabla\widehat{p} + \frac{\kappa}{2}\widehat{p}^2 \right\} d\Omega \\
&= \frac{1}{2}\iint_{\Omega} \left\{ (D^{\frac{1}{2}}\nabla p - D^{\frac{1}{2}}\nabla\widehat{p})^2, +\kappa(p-\widehat{p})^2 \right\} \Omega, \\
&\geq 0. \qquad (4.61)
\end{aligned}$$

The upper bound is obtained by applying the natural conditions

$$\left.\begin{aligned} p &= \tfrac{1}{\kappa}(e - \nabla\cdot\mathbf{q}) \qquad \text{in } \Omega \\ \mathbf{q}\cdot\mathbf{n} &= g \qquad \text{on } \Gamma^+ \end{aligned}\right\} H^+ \tag{4.62}$$

as constraints. Again, this is achieved by direct substitution of the relationship (4.62) for p, and employing an approximation space satisfying the Neumann boundary condition $\mathbf{q}\cdot\mathbf{n} = g$. Enforcing the H^+ constraints in this manner generates the functional $\mathcal{G}^+(\mathbf{q})$ given by

$$\mathcal{G}^+(\mathbf{q}) = \iint_\Omega \left\{ \frac{D^{-1}}{2}\mathbf{q}\cdot\mathbf{q} - \frac{e}{\kappa}\nabla\cdot\mathbf{q} + \frac{1}{2\kappa}(\nabla\cdot\mathbf{q})^2 + \frac{1}{2\kappa}e^2 \right\} d\Omega + \int_{\Gamma^-} f\mathbf{q}\cdot\mathbf{n}\, d\Gamma, \tag{4.63}$$

which is an upper bound on the stationary value of the functional for

$$\begin{aligned} \mathcal{G}^+(\widehat{\mathbf{q}}) - \mathcal{G}^+(\mathbf{q}) &= \iint_\Omega \left\{ \frac{D^{-1}}{2}\widehat{\mathbf{q}}\cdot\widehat{\mathbf{q}} - \frac{e}{\kappa}\nabla\cdot\widehat{\mathbf{q}} + \frac{1}{2\kappa}(\nabla\cdot\widehat{\mathbf{q}})^2 - \frac{D^{-1}}{2}\mathbf{q}\cdot\mathbf{q} \right. \\ &\quad \left. + \frac{e}{\kappa}\nabla\cdot\mathbf{q} - \frac{1}{2\kappa}(\nabla\cdot\mathbf{q})^2 \right\} d\Omega + \int_{\Gamma^-} f(\widehat{\mathbf{q}} - \mathbf{q})\cdot\mathbf{n}\, d\Gamma, & (4.64) \\ &= \iint_\Omega \left\{ \frac{D^{-1}}{2}\widehat{\mathbf{q}}\cdot\widehat{\mathbf{q}} - \frac{1}{\kappa}(\nabla\cdot\widehat{\mathbf{q}} + \kappa\widehat{p})\nabla\cdot\widehat{\mathbf{q}} + \frac{1}{2\kappa}(\nabla\cdot\widehat{\mathbf{q}})^2 \right. \\ &\quad \left. - \frac{D^{-1}}{2}\mathbf{q}\cdot\mathbf{q} + \frac{1}{\kappa}(\nabla\cdot\widehat{\mathbf{q}} + \kappa\widehat{p})\nabla\cdot\mathbf{q} - \frac{1}{2\kappa}(\nabla\cdot\mathbf{q})^2 \right\} d\Omega \\ &\quad + \int_{\Gamma^-} f(\widehat{\mathbf{q}} - \mathbf{q})\cdot\mathbf{n}\, d\Gamma, & (4.65) \\ &= \iint_\Omega \left\{ -\frac{D^{-1}}{2}\widehat{\mathbf{q}}\cdot\widehat{\mathbf{q}} - \frac{1}{2\kappa}(\nabla\cdot\widehat{\mathbf{q}})^2 - \frac{D^{-1}}{2}\mathbf{q}\cdot\mathbf{q} \right. \\ &\quad \left. + \frac{1}{\kappa}(\nabla\cdot\widehat{\mathbf{q}})(\nabla\cdot\mathbf{q}) + D^{-1}\widehat{\mathbf{q}}\cdot\mathbf{q} - \frac{1}{2\kappa}(\nabla\cdot\mathbf{q})^2 \right\} d\Omega, & (4.66) \\ &= -\frac{1}{2}\iint_\Omega \left\{ (D^{-\frac{1}{2}}\widehat{\mathbf{q}} - D^{-\frac{1}{2}}\mathbf{q}) + \frac{1}{\kappa}(\nabla\cdot\widehat{\mathbf{q}} - \nabla\cdot\mathbf{q}) \right\} d\Omega, & (4.67) \\ &\leq 0. & (4.68) \end{aligned}$$

The maximum and minimum principles can be used in conjunction with the Rothe method to estimate the required solution integrals of the time-dependent diffusion equation. However, the maximum and minimum principles do not construct upper and lower bounds on the *analytic* value of these integrals as the time derivative in the Rothe method is only approximate. The error in the estimates obtained from the upper and lower bounds is investigated in the following example. In order that approximations to the required quantity of interest can be made, the twinning technique described in section 2.3.2 is employed.

4.1.4 A Time-Dependent Diffusion Example Employing Twinning

As a means of validating the method developed, the time dependent diffusion equation,

$$\frac{\partial \widehat{u}}{\partial t} - \frac{\partial^2 \widehat{u}}{\partial x^2} = 0, \tag{4.69}$$

in the domain $0 \leq x \leq 1$, $0 \leq t \leq \infty$, with boundary and initial conditions

$$\begin{aligned} \widehat{u}(0,t) &= 0, & (4.70)\\ -\widehat{u}_x(1,t)\cdot \mathbf{n} &= 0, & (4.71)\\ \widehat{u}(x,0) &= x(2-x), & (4.72) \end{aligned}$$

is considered. The equation set (4.69)-(4.72) has the analytic solution

$$\widehat{u}(x,t) = \sum_{n=1}^{\infty} \frac{16}{n^3\pi^3}\left(1-(-1)^n\right)\sin\left(\frac{n\pi x}{2}\right) e^{-\frac{1}{4}n^2\pi^2 t}, \tag{4.73}$$

found as a Fourier sine series with symmetry assumed along the axis $x = 1$. For the purposes of the example the quantity of interest is chosen to be the integral of the solution over the spatial domain

$$\Theta(t) = \int_0^1 \widehat{u}(x,t)\, dx, \tag{4.74}$$

at a given time. The analytic value of $\Theta(t)$ can be calculated from solution (4.73) and is found to be

$$\Theta(t) = \sum_{n=1}^{\infty} \frac{32}{n^4\pi^4}\left(1-(-1)^n\right) e^{-\frac{1}{4}n^2\pi^2 t}. \tag{4.75}$$

The evolution of $\Theta(t)$ is shown in figure 4.3 with the exponential dependence of the solution in time ensuring that the solution decays to zero as time tends to infinity.

To obtain approximations to the quantity of interest the Rothe method is implemented by applying a Crank-Nicolson type discretisation, $\theta = \frac{1}{2}$, to the time derivative. The resulting semi-discrete system

$$-\frac{\partial^2 \widehat{u}^{t+\Delta t}}{\partial x^2} + \kappa \widehat{u}^{t+\Delta t} = e = \left(\kappa \widehat{u}^t + \frac{\partial^2 \widehat{u}^t}{\partial x^2}\right), \tag{4.76}$$

with boundary conditions

$$\begin{aligned} \widehat{u}(0)^{t+\Delta t} &= f = 0, & (4.77)\\ -\widehat{u}_x(1)^{t+\Delta t} &= g = 0, & (4.78) \end{aligned}$$

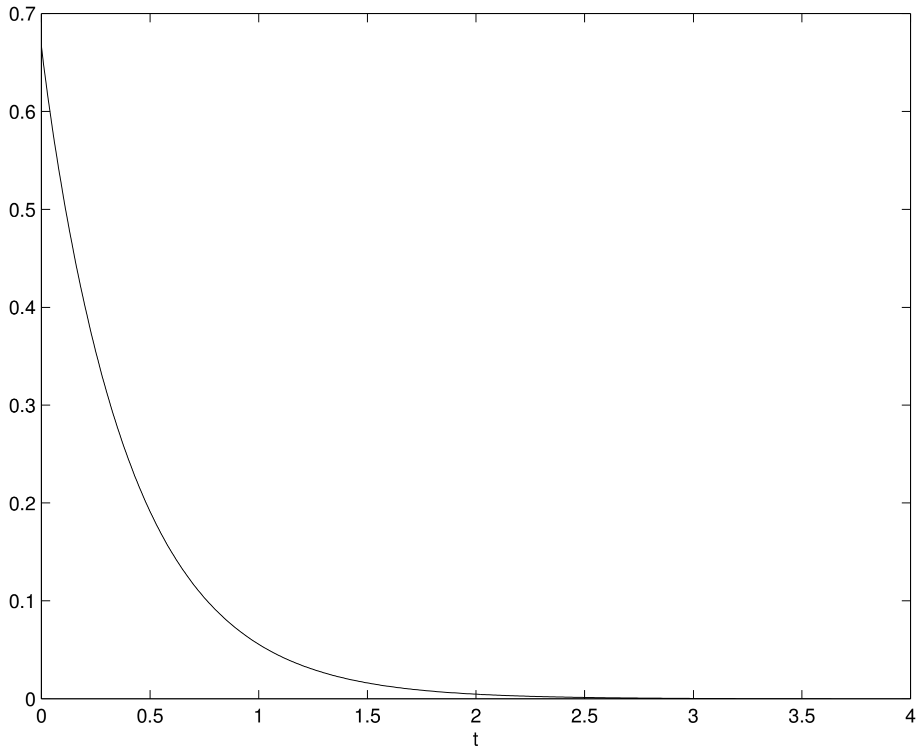

Figure 4.3: Evolution of the analytic quantity of interest

is obtained where

$$\kappa = \frac{2}{\Delta t}. \tag{4.79}$$

Note, $\widehat{u}(x)^{t+\Delta t}$ is the analytic 1D solution of (4.76). In general whilst it may be a good approximation,

$$\widehat{u}(x)^{t+\Delta t} \neq \widehat{u}(x, t + \Delta t) \tag{4.80}$$

due to the discretisation made in time. The scheme is initialised using

$$u^0 = \widehat{u}(x, 0) = x(2 - x), \tag{4.81}$$

and $u(x)^{t+\Delta t}$, an approximation to $\widehat{u}(x)^{t+\Delta t}$, is calculated at successive time-steps using the variational principles associated with the Helmholtz equation. Therefore the approximation

$$u(x)^{t+\Delta t} \approx \widehat{u}(x)^{t+\Delta t} \approx \widehat{u}(x, t + \Delta t) \tag{4.82}$$

is found.

Similarly the quantity of interest associated with the semi-discrete system, $\Theta_h(\widehat{u}(x)^{t+\Delta t})$, will only approximate the quantity of interest associated with the analytic solution,

$\Theta(\widehat{u}(x,t+\Delta t))$, due again to the approximate time discretisations made in the Rothe Method. As a consequence, bounds found on $\Theta_h(\widehat{u}(x)^{t+\Delta t})$ via the extremum principles will not imply bounds on $\Theta(\widehat{u}(x,t))$.

Bounds on $\Theta_h(\widehat{u}(x)^{t+\Delta t})$ are found from the stationary value of the Helmholtz functional. However, specification of a dual problem is required in order that the stationary value and the quantity of interest are linked. The dual solution is required because with the quantity of interest chosen as the integral of the solution, the primal problem is not self-dual. The lack of self-duality can be demonstrated by making the substitutions for the forcing and the boundary conditions of the problem set (4.76)-(4.78), into the stationary value of the functional $\mathcal{G}(\widehat{u}^{t+\Delta t}, -\widehat{u}_x^{t+\Delta t})$. The stationary value of the Helmholtz functional is then found to be equivalent to the integral

$$\begin{aligned}\mathcal{G}(\widehat{u}^{t+\Delta t}, -\widehat{u}_x^{t+\Delta t}) &= \frac{1}{2}\iint_{\Omega} \widehat{u}e \, d\Omega - \frac{1}{2}\int_{\Gamma^-} f\widehat{u}_x^{t+\Delta t}\cdot\mathbf{n}\, d\Gamma - \frac{1}{2}\int_{\Gamma^+} \widehat{u}^{t+\Delta t} g \, d\Gamma, \\ &= \frac{1}{2}\int_0^1 \widehat{u}^{t+\Delta t}\left(\kappa\widehat{u}^t + \nabla\cdot\nabla\widehat{u}^t\right)\, dx \end{aligned} \tag{4.83}$$

as opposed to

$$\Theta_h(\widehat{u}^{t+\Delta t}) = \int_0^1 \widehat{u}^{t+\Delta t}\, dx \tag{4.84}$$

as required. In order to obtain bounds on the correct approximation to the quantity of interest (4.84), the dual problem

$$-\nabla\cdot\nabla\widehat{v} + \kappa\widehat{v} = a_2 = 1, \tag{4.85}$$

$$\widehat{v}(0) = b_2 = 0, \tag{4.86}$$

$$-\nabla\widehat{v}(1)\cdot\mathbf{n} = c_2 = 0, \tag{4.87}$$

is introduced. Then, using the twinning transformations

$$\begin{aligned} p_1^{t+\Delta t} &= u^{t+\Delta t} + v, & u^{t+\Delta t} &= \tfrac{1}{2}(p_1^{t+\Delta t} + p_2^{t+\Delta t}), \\ p_2^{t+\Delta t} &= u^{t+\Delta t} - v, & v &= \tfrac{1}{2}(p_1^{t+\Delta t} - p_2^{t+\Delta t}), \end{aligned} \tag{4.88}$$

the forcing terms

$$\begin{aligned} e_1 &= a_1 + a_2, & a_1 &= \tfrac{1}{2}(e_1 + e_2,) \\ e_2 &= a_1 - a_2, & a_2 &= \tfrac{1}{2}(e_1 - e_2), \\ \\ f_1 &= b_1 + b_2, & b_1 &= \tfrac{1}{2}(f_1 + f_2), \\ f_2 &= b_1 - b_2, & b_2 &= \tfrac{1}{2}(f_1 - f_2), \\ \\ g_1 &= c_1 + c_2, & c_1 &= \tfrac{1}{2}(g_1 + g_2), \\ g_2 &= c_1 - c_2, & c_2 &= \tfrac{1}{2}(g_1 - g_2), \end{aligned} \tag{4.89}$$

and pair of self-dual problems

$$-\nabla \cdot \nabla \widehat{p}_1^{t+\Delta t} + \kappa \widehat{p}_1^{t+\Delta t} = e_1 = \left(\kappa \widehat{u}^t + \nabla \cdot \nabla \widehat{u}^t + 1\right), \tag{4.90}$$

$$\widehat{p}_1(0)^{t+\Delta t} = f_1 = 0, \tag{4.91}$$

$$-\nabla \widehat{p}_1(1)^{t+\Delta t} \cdot \mathbf{n} = g_1 = 0, \tag{4.92}$$

and

$$-\nabla \cdot \nabla \widehat{p}_2^{t+\Delta t} + \kappa \widehat{p}_2^{t+\Delta t} = e_2 = \left(\kappa \widehat{u}^t + \nabla \cdot \nabla \widehat{u}^t - 1\right), \tag{4.93}$$

$$\widehat{p}_2(0)^{t+\Delta t} = f_2 = 0, \tag{4.94}$$

$$-\nabla \widehat{p}_2(1)^{t+\Delta t} \cdot \mathbf{n} = g_2 = 0, \tag{4.95}$$

are defined. Approximations to the solution $\widehat{p}_1^{t+\Delta t}$ and $\widehat{p}_2^{t+\Delta t}$ can then be found using the minimum and maximum principles and importantly the bounds

$$\mu_i^- \leq \mathcal{G}(\widehat{p}_i^{t+\Delta t}, \widehat{\mathbf{q}}_i^{t+\Delta t}) \leq \mu_i^+ \tag{4.96}$$

are obtained, where

$$\widehat{\mathbf{q}}_i^{t+\Delta t} = -\nabla \widehat{p}_i^{t+\Delta t}. \tag{4.97}$$

The approximate quantity of interest is then

$$\begin{aligned} \Theta_h(\widehat{u}^{t+\Delta t}) &= \frac{1}{2}\mathcal{G}(\widehat{p}_1^{t+\Delta t}, \widehat{\mathbf{q}}_1^{t+\Delta t}) - \frac{1}{2}\mathcal{G}(\widehat{p}_2^{t+\Delta t}, \widehat{\mathbf{q}}_2^{t+\Delta t}), \\ &= \frac{1}{4}\iint_\Omega \left\{\widehat{p}_1^{t+\Delta t} e_1 - \widehat{p}_2^{t+\Delta t} e_2\right\} d\Omega + \frac{1}{4}\int_{\Gamma^-} \left\{f_1 \widehat{\mathbf{q}}_1^{t+\Delta t} \cdot \mathbf{n} - f_2 \widehat{\mathbf{q}}_2^{t+\Delta t} \cdot \mathbf{n}\right\} d\Gamma, \\ &\quad - \frac{1}{4}\int_{\Gamma^+} \left\{\widehat{p}_1^{t+\Delta t} g_1 - \widehat{p}_2^{t+\Delta t} g_2\right\} d\Gamma \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{4}\iint_{\Omega}\left\{(\widehat{p}_1^{t+\Delta t}+\widehat{p}_2^{t+\Delta t})(e_1-e_2)+\widehat{p}_1^{t+\Delta t}e_2-\widehat{p}_2^{t+\Delta t}e_1\right\}d\Omega,\\
&\quad+\frac{1}{4}\int_{\Gamma-}\left\{f_1\widehat{\mathbf{q}}_1^{t+\Delta t}\cdot\mathbf{n}-f_2\widehat{\mathbf{q}}_2^{t+\Delta t}\cdot\mathbf{n}\right\}d\Gamma-\frac{1}{4}\int_{\Gamma+}\left\{\widehat{p}_1^{t+\Delta t}g_1-\widehat{p}_2^{t+\Delta t}g_2\right\}d\Gamma,\\
&= \frac{1}{4}\iint_{\Omega}\left\{(\widehat{p}_1^{t+\Delta t}+\widehat{p}_2^{t+\Delta t})(e_1-e_2)+\widehat{p}_1(\nabla\cdot\widehat{\mathbf{q}}_2^{t+\Delta t}+\kappa\widehat{p}_2^{t+\Delta t})\right.\\
&\quad\left.-\widehat{p}_2^{t+\Delta t}(\nabla\cdot\widehat{\mathbf{q}}_1^{t+\Delta t}+\kappa\widehat{p}_1^{t+\Delta t})\right\}d\Omega+\frac{1}{4}\int_{\Gamma-}\left\{f_1\widehat{\mathbf{q}}_1^{t+\Delta t}\cdot\mathbf{n}-f_2\widehat{\mathbf{q}}_2^{t+\Delta t}\cdot\mathbf{n}\right\}d\Gamma\\
&\quad-\frac{1}{4}\int_{\Gamma+}\left\{\widehat{p}_1^{t+\Delta t}g_1-\widehat{p}_2^{t+\Delta t}g_2\right\}d\Gamma,\\
&= \frac{1}{4}\iint_{\Omega}(\widehat{p}_1^{t+\Delta t}+\widehat{p}_2^{t+\Delta t})(e_1-e_2)d\Omega+\frac{1}{4}\int_{\Gamma-}\left\{f_1\widehat{\mathbf{q}}_1^{t+\Delta t}\cdot\mathbf{n}-f_2\widehat{\mathbf{q}}_2^{t+\Delta t}\cdot\mathbf{n}\right\}d\Gamma\\
&\quad+\frac{1}{4}\int_{\Gamma}\left\{\widehat{p}_1^{t+\Delta t}\widehat{\mathbf{q}}_2^{t+\Delta t}\cdot\mathbf{n}-\widehat{p}_2^{t+\Delta t}\widehat{\mathbf{q}}_1^{t+\Delta t}\cdot\mathbf{n}\right\}d\Gamma\\
&\quad-\frac{1}{4}\int_{\Gamma+}\left\{\widehat{p}_1^{t+\Delta t}g_1-\widehat{p}_2^{t+\Delta t}g_2\right\}d\Gamma,\\
&= \frac{1}{4}\iint_{\Omega}(\widehat{p}_1^{t+\Delta t}+\widehat{p}_2^{t+\Delta t})(e_1-e_2)d\Omega-\frac{1}{4}\int_{\Gamma+}(\widehat{p}_1^{t+\Delta t}+\widehat{p}_2^{t+\Delta t})(g_1-g_2)\,d\Gamma\\
&\quad+\frac{1}{4}\int_{\Gamma-}(f_1-f_2)(\widehat{\mathbf{q}}_1^{t+\Delta t}+\widehat{\mathbf{q}}_2^{t+\Delta t})\cdot\mathbf{n}\,d\Gamma,\\
&\quad-\frac{1}{4}\int_{\Gamma+}(\widehat{p}_1^{t+\Delta t}+\widehat{p}_2^{t+\Delta t})(g_1-g_2)\,d\Gamma,\\
&= \iint_{\Omega}a_2\widehat{u}^{t+\Delta t}d\Omega-\int_{\Gamma-}b_2\nabla\widehat{u}^{t+\Delta t}\cdot\mathbf{n}\,d\Gamma-\int_{\Gamma+}c_2\widehat{u}^{t+\Delta t}\,d\Gamma, \qquad (4.98)
\end{aligned}
$$

which, with the choice of boundary and forcing functions, reduces to

$$
\begin{aligned}
\Theta_h(\widehat{u}^{t+\Delta t}) &= \iint_{\Omega}\widehat{u}^{t+\Delta t}d\Omega \qquad (4.99)\\
&= \int_0^1\widehat{u}^{t+\Delta t}dx \qquad (4.100)
\end{aligned}
$$

as required. Bounds on the approximate quantity of interest are then

$$\frac{1}{2}(\mu_1^- - \mu_2^+) \leq \Theta_h(\widehat{u}^{t+\Delta t}) \leq \frac{1}{2}(\mu_1^+ - \mu_2^-). \qquad (4.101)$$

The computational cost of the method is dominated by the four matrix inversions associated with obtaining the solutions p_1, p_2, $\mathbf{q}_1$ and $\mathbf{q}_2$. However, provided the twinned problems are solved on the same computational mesh the cost of assembling these problems is effectively split. Initially, the possibility of further computational saving of the type discussed in section 1.3, in which the dual problem is inverted once as oppose to the primal problem being inverted many times, may seem realisable. However the forcing differs between the pair of problems $(p_1^{t+\Delta t}, \mathbf{q}_1^{t+\Delta t})$ and $(p_2^{t+\Delta t}, \mathbf{q}_2^{t+\Delta t})$ and therefore a common dual solution, through which the saving would be made, does not exist.

4.1.5 Results

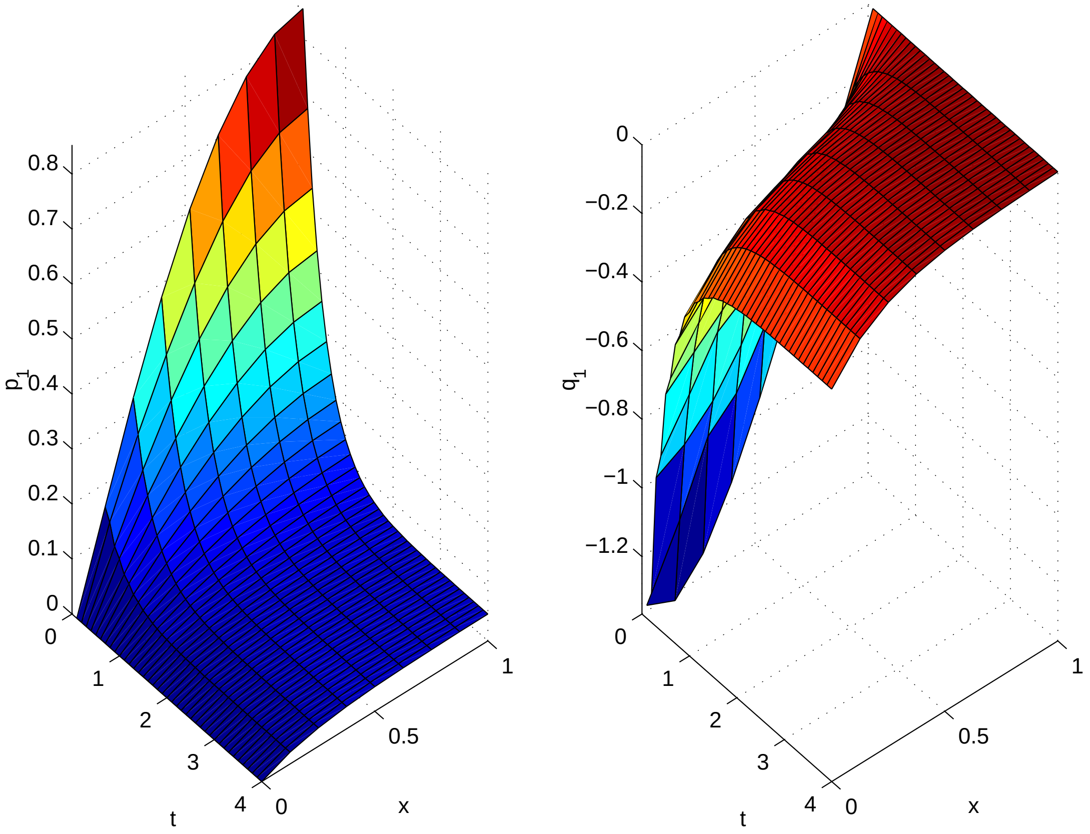

Figure 4.4: Solutions $p(x)_1^{t+\Delta t}$ and $q(x)_1^{t+\Delta t}$ obtained over a sequence of time steps

Examples of the numerical solutions obtained for the pair of problems $(p(x)_1^{t+\Delta t}, q(x)_1^{t+\Delta t})$ and $(p(x)_2^{t+\Delta t}, q(x)_2^{t+\Delta t})$ over a sequence of time steps are shown in figures 4.4 and 4.5. The results were obtained using a timestep of 0.1 and 4 quadratic conforming elements. The primal and dual solutions were found by inverting the twinning transformations, effectively taking the sum and difference of the solutions p_1 and p_2, and are plotted in figure 4.6. Although the dual solution is independent of time, as a consequence of twinning method, it is found as the difference of two time dependent problems in order that bounds on the approximate quantity of interest can be computed.

The approximations to the quantity of interest are plotted in figure 4.7. Figure 4.7 shows the error ϵ_1, the difference between the approximate and the analytic quantity of interest. The upper and lower bounds are generated from bounds on the stationary value of the functional associated with the Helmholtz equation, using the twinning method.

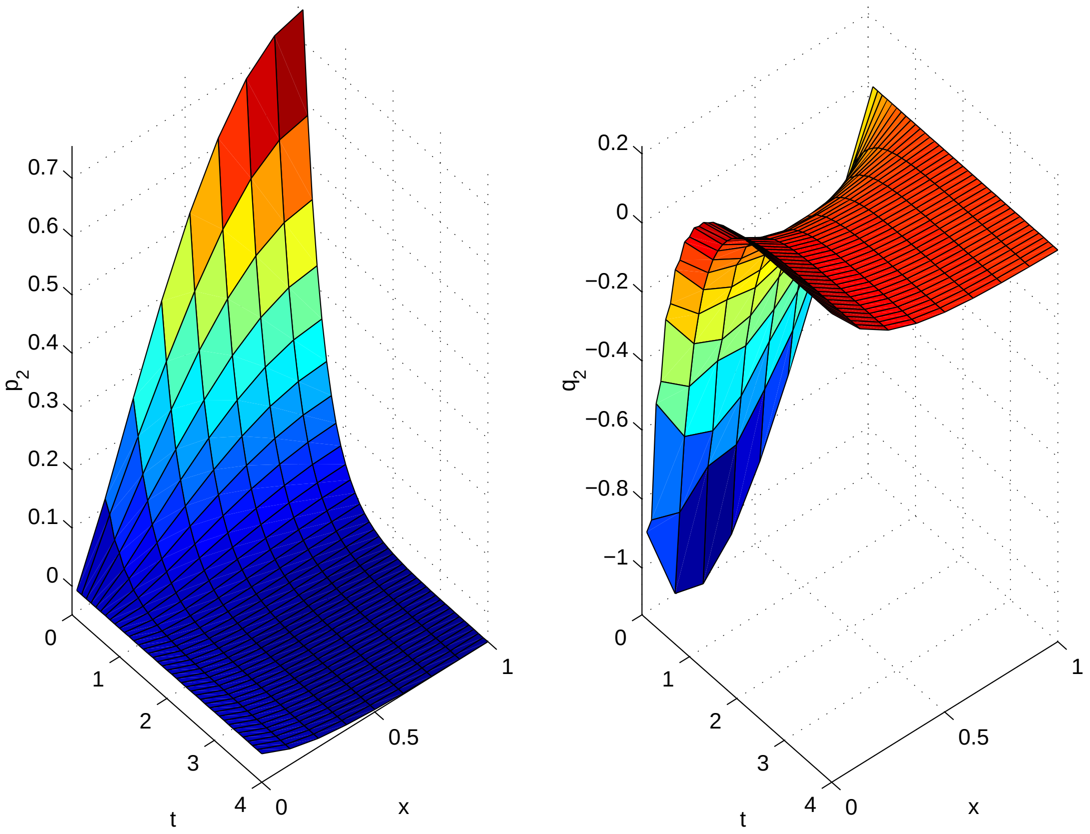

Figure 4.5: Solutions $p(x)_2^{t+\Delta t}$ and $q(x)_2^{t+\Delta t}$ obtained over a sequence of time steps

As a result, the bounds are generated by approximations to the semi-discrete system of equations at each time step and not by the continuous governing equations and therefore the bounds do not necessarily enclose the analytic quantity of interest. The deviation of the bounds from the analytic value is indicated by the pairs of curves both lying above $\epsilon_1 = 0$. The sensitivity of the method to the number of elements and the timestep employed is summarised in figure 4.7. From figure 4.7, and over the range considered, the method appears to be more sensitive to the fineness of the discretisation rather than the size of the timestep. In particular increasing the number of elements both reduces the error in the approximations and the difference between the bounds. The ability of the method to correct initial increases in the error and perform well at large time values is due to both the analytic and numerical solution decaying with time. The decay of both solutions is governed by the following shared geometric property.

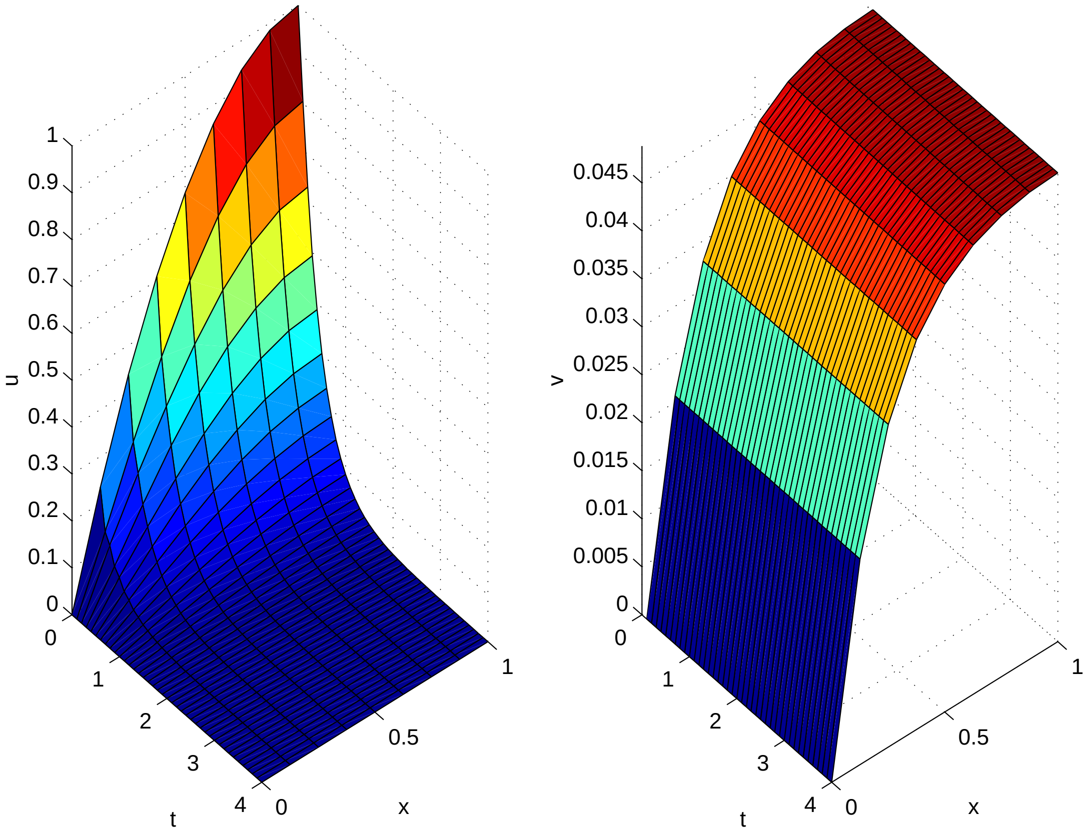

Figure 4.6: Primal and dual solutions, $u(x,t)$ and $v(x)$

4.1.6 A Geometric Property

The decay of the analytic solution $\widehat{u}$ is demonstrated by multiplying the governing equation

$$\widehat{u}_t - \nabla \cdot \nabla \widehat{u} = 0, \tag{4.102}$$

with homogenous boundary conditions

$$\begin{aligned} \widehat{u} &= 0 \qquad \text{on } \Gamma^-, & (4.103)\\ -\nabla \widehat{u} \cdot \mathbf{n} &= 0 \qquad \text{on } \Gamma^+, & (4.104)\end{aligned}$$

by the solution $\widehat{u}(\mathbf{x},t)$ and integrating over the spatial coordinates $\mathbf{x}$ to obtain

$$\begin{aligned} 0 &= \iint_\Omega \{\widehat{u}\widehat{u}_t - \widehat{u}\nabla \cdot \nabla \widehat{u}\}\, d\Omega, & (4.105)\\ &= \iint_\Omega \{\widehat{u}\widehat{u}_t + \nabla \widehat{u} \cdot \nabla \widehat{u}\}\, d\Omega. & (4.106)\end{aligned}$$

The positivity of the term $\nabla \widehat{u} \cdot \nabla \widehat{u}$ then implies

$$0 \;>\; \iint_\Omega \widehat{u}\widehat{u}_t \, d\Omega, \tag{4.107}$$

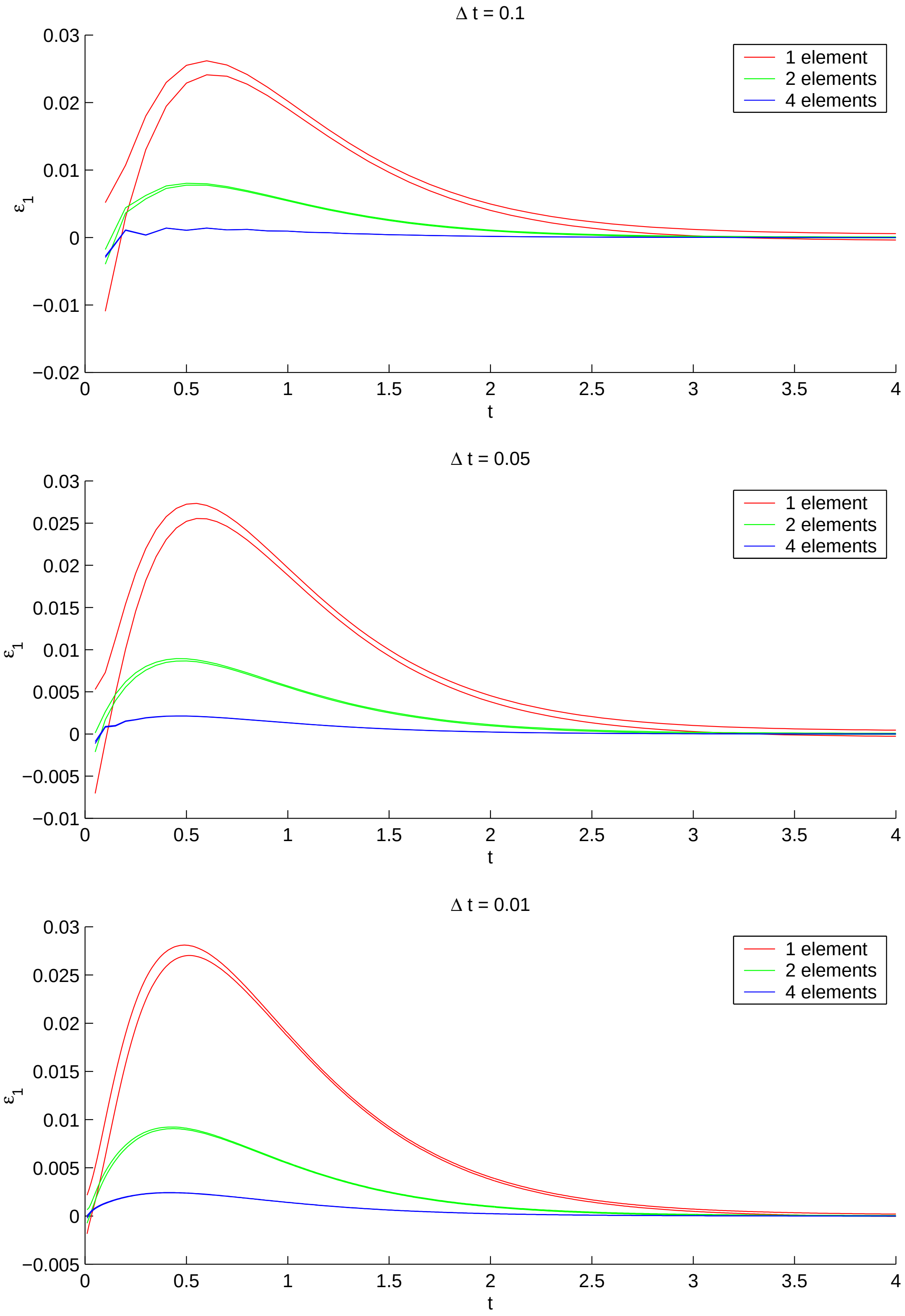

Figure 4.7: Error in the bounds on Θ_h

$$= \frac{1}{2}\frac{\partial}{\partial t}\|\widehat{u}\|_{\Omega}^{2}, \tag{4.108}$$

for non-trivial $\widehat{u}$, and therefore the integral of the solution is tending to a constant. From the governing equation it is observed that in the steady-state limit the solution satisfies Laplace's equation with zero Dirichlet and Neumann boundary conditions and therefore $\widehat{u}(\mathbf{x},t) \to 0$ as $t \to \infty$, assuming sufficient smoothness of solution.

Similarly the semi-discrete system exhibits the same geometric property, where the functions $\widehat{u}(\mathbf{x})^{t}$ and $\widehat{u}(\mathbf{x})^{t+\Delta t}$ are the analytic solutions of the semi-discrete equations and therefore functions of the spatial coordinates only. Multiplying (4.16) by $\widehat{u}^{t+\Delta t} + u^{t}$ and integrating over the spatial coordinates the equation

$$\begin{aligned} 0 &= \iint_{\Omega} (\widehat{u}^{t+\Delta t} + \widehat{u}^{t}) \frac{\widehat{u}^{t+\Delta t} - \widehat{u}^{t}}{\Delta t}\, d\Omega - \frac{1}{2}\iint_{\Omega} (\widehat{u}^{t+\Delta t} + \widehat{u}^{t}) \nabla \cdot \nabla (\widehat{u}^{t+\Delta t} + \widehat{u}^{t})\, d\Omega, \\ &= \iint_{\Omega} \left\{ \frac{(\widehat{u}^{t+\Delta t})^{2}}{\Delta t} - \frac{(\widehat{u}^{t})^{2}}{\Delta t} \right\} d\Omega + \frac{1}{2}\iint_{\Omega} \nabla (\widehat{u}^{t+\Delta t} + \widehat{u}^{t}) \cdot \nabla (\widehat{u}^{t+\Delta t} + \widehat{u}^{t})\, d\Omega, \end{aligned} \tag{4.109}$$

is obtained. Similarly the positivity of the second integral in (4.109) implies that for every time step

$$\iint_{\Omega} (\widehat{u}^{t+\Delta t})^{2}\, d\Omega - \iint_{\Omega} (\widehat{u}^{t})^{2}\, d\Omega < 0, \tag{4.110}$$

or equivalently

$$\|\widehat{u}^{t+\Delta t}\|_{\Omega}^{2} < \|\widehat{u}^{t}\|_{\Omega}^{2}. \tag{4.111}$$

for non-trivial functions $\widehat{u}^{t+\Delta t}$ and $\widehat{u}^{t}$. Hence through a similar argument, but considering the uniqueness of solutions governed by the discrete version of the Laplace equation $\widehat{u}^{t} \to 0$ as $t \to \infty$.

The geometric property ensures that the numerical approximations have the same characteristics as the analytic solution. As a result of the decaying solution, obtaining accurate relative error estimates at large time values is difficult. Figure (4.8) shows the relative error ϵ_2, defined as ϵ_1 divided by the analytic value of the quantity of interest. The graph also reinforces the findings, that over the range considered, the relative error incurred by the method is more sensitive to the number of elements employed than the size of the timestep.

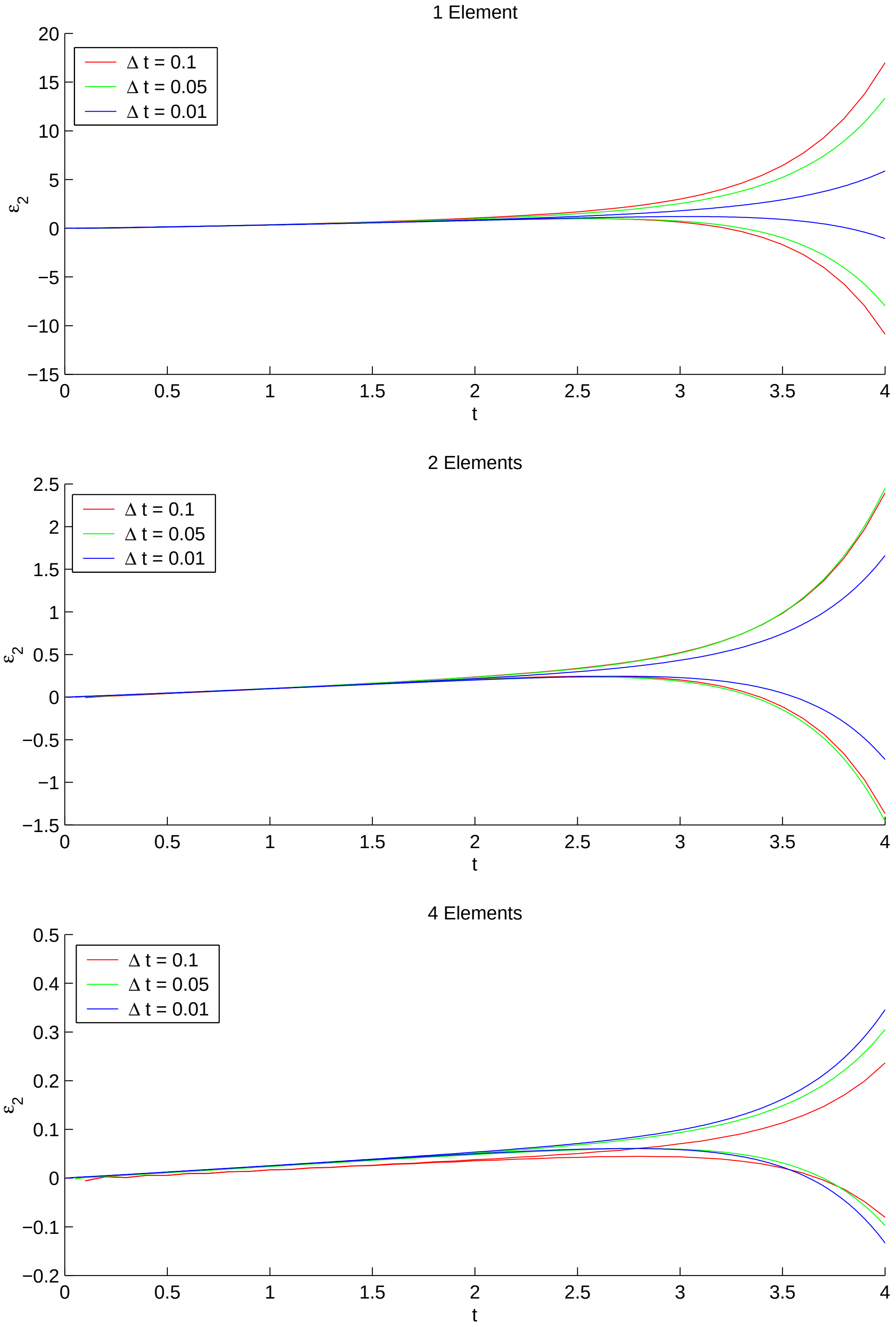

Figure 4.8: Relative error in the bounds on Θ_h

Having developed a numerical method for the time-dependent diffusion equation which replicates the decaying property of the analytic solution, and as a result produces accurate approximations to the quantity of interest, the advection-diffusion equation is now considered.

4.2 The Advection-Diffusion Equation

The original motivation to consider the time dependent diffusion equation was as a means of developing methods to model an advection-diffusion process governed by the equations

$$\begin{aligned} \frac{\partial \widehat{u}(\mathbf{x},t)}{\partial t} - \nabla \cdot [\mathbf{D}\nabla \widehat{u}(\mathbf{x},t) - \widehat{\mathbf{w}}\widehat{u}(\mathbf{x},t)] &= 0 \qquad \mathbf{x} \in \Omega, & (4.112)\\ \widehat{u}(\mathbf{x},t) &= 0 \qquad \mathbf{x} \in \Gamma^-, & (4.113)\\ -\nabla \widehat{u}(\mathbf{x},t) \cdot \mathbf{n} &= 0 \qquad \mathbf{x} \in \Gamma^+, & (4.114)\\ \widehat{u}(\mathbf{x},0) &= u_0 \qquad \mathbf{x} \in \Omega, & (4.115) \end{aligned}$$

in a flow field $\widehat{\mathbf{w}}$ where

$$\nabla \cdot \widehat{\mathbf{w}} = 0. \qquad (4.116)$$

Employing a similar philosophy to that used in the time-dependent diffusion model, the discretisation of the diffusion term is deferred. Instead, the equation is considered from a Lagrangian perspective and the remaining temporal and spatial derivatives are combined to form the Lagrangian derivative of the concentration, so that

$$\begin{aligned} 0 &= \frac{\partial \widehat{u}}{\partial t} + \nabla \cdot (\widehat{\mathbf{w}}\widehat{u}) - \nabla \cdot \mathbf{D}\nabla \widehat{u}, & (4.117)\\ &= \frac{\partial \widehat{u}}{\partial t} + \widehat{u}\nabla \cdot \widehat{\mathbf{w}} + \widehat{\mathbf{w}} \cdot \nabla \widehat{u} - \nabla \cdot \mathbf{D}\nabla \widehat{u}, & (4.118)\\ &= \frac{\partial \widehat{u}}{\partial t} + \widehat{\mathbf{w}} \cdot \nabla \widehat{u} - \nabla \cdot \mathbf{D}\nabla \widehat{u}, & (4.119)\\ &= \frac{D\widehat{u}}{Dt} - \nabla \cdot \mathbf{D}\nabla \widehat{u}. & (4.120) \end{aligned}$$

A discretisation in time can then be made using the Crank-Nicolson method to obtain the semi-discrete system

$$\begin{aligned} \frac{\widehat{u}(\mathbf{x})^{t+\Delta t} - \widehat{u}(\mathbf{x})^t_f}{\Delta t} &= \frac{1}{2}(\nabla \cdot \lambda \nabla \widehat{u}(\mathbf{x})^{t+\Delta t} + \nabla \cdot \lambda \nabla \widehat{u}(\mathbf{x})^t) \qquad \mathbf{x} \in \Omega, & (4.121)\\ \widehat{u}(\mathbf{x})^{t+\Delta t} &= 0 \qquad \mathbf{x} \in \Gamma^-, & (4.122)\\ -\nabla \widehat{u}(\mathbf{x})^{t+\Delta t} \cdot \mathbf{n} &= 0 \qquad \mathbf{x} \in \Gamma^+, & (4.123) \end{aligned}$$

where $\widehat{u}_f^t$ is the value of the function $\widehat{u}(\mathbf{x})^t$ at the foot of the characteristic running through the node at time $t+\Delta t$. The method is therefore a semi-Lagrangian scheme as at each time step a different set of characteristics is traced backwards in time from the new node positions in (x,t) space. This concept is schematically shown in figure 4.9. Crucially the semi-discrete equations now once more resemble the Helmholtz equation

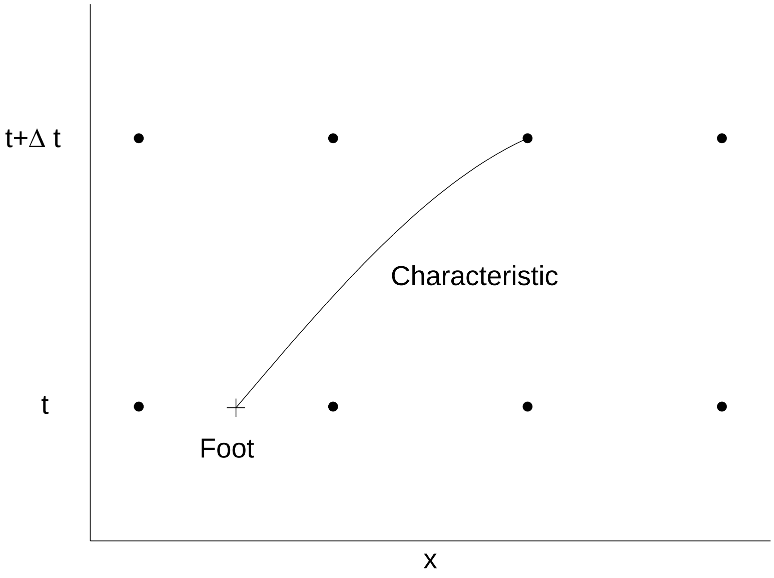

Figure 4.9: The semi-Lagrangian method

and re-arranging the set (4.121)-(4.123) we obtain

$$
\begin{aligned}
-\nabla \cdot \mathbf{D}\nabla\widehat{u}(\mathbf{x})^{t+\Delta t} + \kappa\widehat{u}(\mathbf{x})^{t+\Delta t} &= a_1 = \kappa u_f^t + \nabla \cdot \mathbf{D}\nabla u^t \qquad \mathbf{x} \in \Omega, && (4.124)\\
\widehat{u}(\mathbf{x})^{t+\Delta t} &= b_1 = 0 \qquad \mathbf{x} \in \Gamma^-, && (4.125)\\
-\nabla\widehat{u}(\mathbf{x})^{t+\Delta t} \cdot \mathbf{n} &= c_1 = 0 \qquad \mathbf{x} \in \Gamma^+, && (4.126)
\end{aligned}
$$

where

$$\kappa = \frac{2}{\Delta t}. \tag{4.127}$$

The solution of the equation set (4.124) - (4.126) coincides with the stationary point of the Helmholtz functional (4.56), where $\widehat{p} = \widehat{u}(\mathbf{x})^{t+\Delta t}$, and therefore enables bounds on weighted integral of the solutions $\widehat{u}(\mathbf{x})^{t+\Delta t}$ to be obtained.

In a similar manner to the time-dependent diffusion problem, the bounds obtained are not on integrals of the analytic solution $\widehat{u}(\mathbf{x},t)$ but rather bounds on a series of discrete problems and approximations to the quantity of interest. The ability of the numerical solution to accurately model the analytic solution of the time-dependent diffusion equations was due in part to both solutions sharing a geometric property that ensures the respective solutions decay in time. A similar property exists for the analytic solution of the advection-diffusion model under certain conditions, the easiest of which to demonstrate is when the velocity field satisfies

$$\nabla \cdot \widehat{\mathbf{w}} \;=\; 0 \qquad in \quad \Omega. \tag{4.128}$$

$$\widehat{\mathbf{w}} \cdot \mathbf{n} \;=\; 0 \qquad on \quad \Gamma^{+},\, \Gamma^{-}. \tag{4.129}$$

Enforcing the condition (4.129) ensures that data is not advected over the boundary whilst the condition (4.128) prevents data being generated or destroyed by sources and sinks in the velocity field.

4.2.1 The Advection-Diffusion Geometric Property

For a velocity field satisfying (4.129) and (4.128) the following property is observed,

$$\begin{aligned}\iint_{\Omega} \rho\nabla \cdot (\widehat{\mathbf{w}}\rho) &= \frac{1}{2}\iint_{\Omega} \{\rho\nabla \cdot (\widehat{\mathbf{w}}\rho) + \rho(\rho\nabla \cdot \widehat{\mathbf{w}}) + \rho(\widehat{\mathbf{w}} \cdot \nabla\rho)\}\, d\Omega \\ &= \frac{1}{2}\iint_{\Omega} \rho(\rho\nabla \cdot \widehat{\mathbf{w}})\, d\Omega + \frac{1}{2}\int_{\Gamma^{+}\cup\Gamma^{-}} \rho^{2}\widehat{\mathbf{w}} \cdot \widehat{\mathbf{n}} d\Gamma \\ &= 0 \end{aligned} \tag{4.130}$$

for a sufficiently smooth functions $\rho(\mathbf{x})$. Therefore the solution u is ever decreasing since by multiplying the governing equation by the solution $\widehat{u}(\mathbf{x}, t)$, integrating over the spatial domain, employing the boundary conditions on $\widehat{u}$ and making use of (4.130)

$$0 \;=\; \iint_{\Omega} \{\widehat{u}\widehat{u}_t + \widehat{u}\nabla \cdot (\widehat{\mathbf{w}}\widehat{u}) - \widehat{u}\nabla \cdot \nabla\widehat{u}\}\, d\Omega, \tag{4.131}$$

$$\;=\; \iint_{\Omega} \{\widehat{u}\widehat{u}_t + \nabla\widehat{u} \cdot \nabla\widehat{u}\}\, d\Omega. \tag{4.132}$$

Again, positivity of the term $\nabla\widehat{u} \cdot \nabla\widehat{u}$ then implies

$$0 \;>\; \iint_{\Omega} \widehat{u}\widehat{u}_t\, d\Omega, \tag{4.133}$$

$$\;=\; \frac{1}{2}\frac{\partial}{\partial t}\|\widehat{u}\|_{\Omega}^{2}, \tag{4.134}$$

for non-trivial $\widehat{u}$.

However, a corresponding geometric property for the semi-discrete solution is not self-evident due to the occurrence of the term $\widehat{u}_f^t$. The behaviour of this solution will partly depend on how accurately the foot of the characteristic is found and in turn how well the Lagrangian derivative is approximated. The semi-discrete solution will be investigated with numerical solutions.

4.2.2 An Advection-Diffusion Example Employing Twinning

The approximations to the quantity of interest obtained using the Helmholtz functional have been found to be accurate for the time dependent diffusion problem. However, the addition of an advection term can add considerable complexity to the flow structure and introduces the notion of diffusion or advection dominated flows. In order to investigate the performance of the semi-Lagrangian method described we consider solving the equation set

$$\begin{aligned}
\frac{\partial \widehat{u}(\mathbf{x},t)}{\partial t} - \nabla \cdot [\mathbf{D}\nabla\widehat{u}(\mathbf{x},t) - \widehat{\mathbf{w}}\widehat{u}(\mathbf{x},t)] &= 0, && (4.135)\\
\widehat{u}(\mathbf{x},t) &= 0 \qquad x = 0, 1, && (4.136)\\
-\nabla\widehat{u}(\mathbf{x},t)\cdot\mathbf{n} &= 0 \qquad y = 0, 1, && (4.137)\\
\widehat{u}(\mathbf{x},0) &= 4x(1-x), && (4.138)
\end{aligned}$$

in the domain $0 \leq (x, y) \leq 1$, $0 \leq t \leq 4$. The flow field was chosen to be

$$\widehat{\mathbf{w}} = \alpha\pi\sin(\pi x)\cos(\pi y)\mathbf{i} - \alpha\pi\cos(\pi x)\sin(\pi y)\mathbf{j}, \tag{4.139}$$

satisfying the conditions (4.128) and (4.129) and introducing a rotational structure into the solution from which the ability of the method to retain symmetries can be observed. The value of the constant α governing the magnitude of the velocity was varied and numerical results obtained. The structure of the flow field is shown in figure 4.10.

For simplicity the quantity of interest is again chosen to be

$$\Theta(\,\widehat{u}(\mathbf{x},t)\,) = \iint_\Omega \widehat{u}(\mathbf{x},t)d\Omega. \tag{4.140}$$

Once again, bounds on the analytic continuous quantity of interest are not available and instead the approximation

$$\begin{aligned}
\Theta_h(\widehat{u}(\mathbf{x})^{t+\Delta t}) &= \iint_\Omega \widehat{u}^{t+\Delta t}d\Omega, && (4.141)\\
&= \int_0^1\int_0^1 \widehat{u}^{t+\Delta t}dx\,dy, && (4.142)
\end{aligned}$$

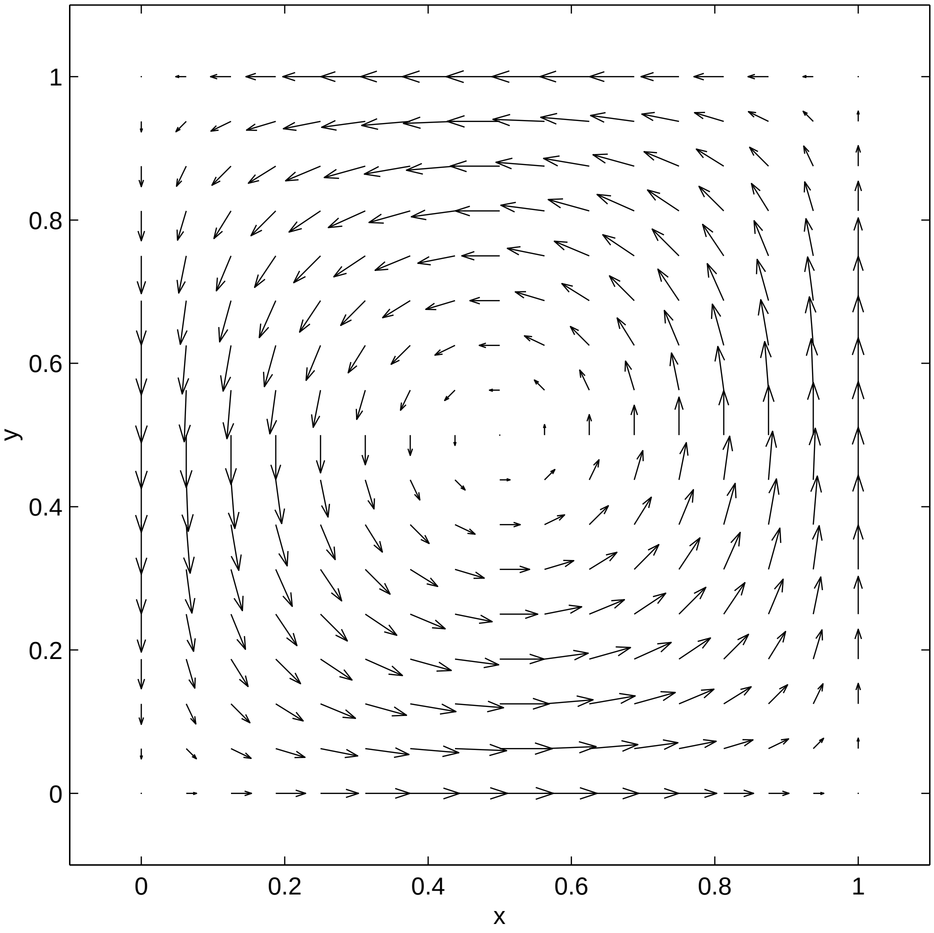

Figure 4.10: Structure of flow field $\widehat{\mathbf{w}}$

is bounded where $\widehat{u}(\mathbf{x})^{t+\Delta t}$ is the analytic solution of the semi-discrete problem. The use of the same quantity of interest enables the definitions of the primal and dual problems from section 4.1.4 to be recycled. Explicitly the primal problem is governed by the equations

$$-\nabla \cdot \mathbf{D}\nabla\widehat{u}(\mathbf{x})^{t+\Delta t} + \kappa\widehat{u}(\mathbf{x})^{t+\Delta t} = a_1 = \kappa u_f^t + \nabla \cdot \mathbf{D}\nabla u^t \qquad \mathbf{x} \in \Omega, \quad (4.143)$$

$$\widehat{u}(\mathbf{x})^{t+\Delta t} = b_1 = 0 \qquad \mathbf{x} \in \Gamma^-, \quad (4.144)$$

$$-\nabla\widehat{u}(\mathbf{x})^{t+\Delta t} \cdot \mathbf{n} = c_1 = 0, \qquad \mathbf{x} \in \Gamma^+, \quad (4.145)$$

and the dual problem is the solution of the set

$$-\nabla \cdot \mathbf{D}\nabla\widehat{v}(\mathbf{x})^{t+\Delta t} + \kappa\widehat{v}(\mathbf{x})^{t+\Delta t} = a_2 = 1 \qquad \mathbf{x} \in \Omega, \quad (4.146)$$

$$\widehat{v}(\mathbf{x})^{t+\Delta t} = b_2 = 0 \qquad \mathbf{x} \in \Gamma^-, \quad (4.147)$$

$$-\nabla\widehat{v}(\mathbf{x})^{t+\Delta t} \cdot \mathbf{n} = c_2 = 0 \qquad \mathbf{x} \in \Gamma^+, \quad (4.148)$$

where

$$\kappa = \frac{2}{\Delta t}. \tag{4.149}$$

In parallel with the time-dependent diffusion example the primal and dual problems are twinned to obtain the pair of self-dual problems

$$\begin{aligned} -\nabla \cdot \nabla \widehat{p}_1 + \kappa \widehat{p}_1 &= e_1 = \left(\kappa u_f^t + \nabla \cdot \nabla u^t + 1\right) && (4.150)\\ \widehat{p}_1 &= f_1 = 0 \qquad x = 0, 1, && (4.151)\\ -\nabla \widehat{p}_1 \cdot \mathbf{n} &= g_1 = 0 \qquad y = 0, 1, && (4.152) \end{aligned}$$

and

$$\begin{aligned} -\nabla \cdot \nabla \widehat{p}_2 + \kappa \widehat{p}_2 &= e_2 = \left(\kappa u_f^t + \nabla \cdot \nabla u^t - 1\right) && (4.153)\\ \widehat{p}_2 &= f_2 = 0 \qquad x = 0, 1, && (4.154)\\ -\nabla \widehat{p}_2 \cdot \mathbf{n} &= g_2 = 0 \qquad y = 0, 1, && (4.155) \end{aligned}$$

which are used to generate the upper and lower bounds on $\Theta_h(\widehat{u}(x)^{t+\Delta t})$ using the inequality (4.101). Again, due to the errors introduced by the discretisation of the first order derivatives, the semi-discrete system only approximates the continuous governing equations. Correspondingly the bounds obtained from the semi-discrete system, by way of the Helmholtz functional, are only approximations to the quantity of interest $\Theta(\widehat{u})$.

4.2.3 Results

Approximations to the analytic quantity of interest were generated by solving the extremum principles associated with the pair of self-dual problems, (4.150)-(4.152) and (4.153)-(4.155), for a range of values of α. The numerical approximations to the solutions of the self-dual problems $(p_1, \mathbf{q}_1)$ and $(p_2, \mathbf{q}_2)$ were constructed using 8×8 quadratic quadrilateral finite elements, and a four-stage Runge-Kutta method was used to integrate the velocity backwards in time and obtain the position of the foot of the characteristic.

The bounds on Θ_h obtained for varying values of α, and a time step of 0.0002, are shown in figure 4.11. From figure 4.11 it is observed that the bounds on the approximate quantity of interest are tight when the advection term is small, inseparable on the graph, but start to weaken when $\alpha = 1000$. A time series of the solutions $u(x)^{t+\Delta t}$ obtained in an advection dominated flow, with $\alpha = 1000$ is shown in figure 4.12. From this series of plots it can be seen that the rotational structure in the solution emerges after an initial period of time. This initial period coincides with that in which the bounds on the approximate quantity of interest are further apart, figure 4.11. During this initial period the solution is aligning itself with the flow $\widehat{\mathbf{w}}$, and therefore in the following phase the solution evolves mainly through diffusion. During the diffusion driven phase of the solution the bounds tend to converge again as the solution decays.

In contrast the time series of the solutions $u(x)^{t+\Delta t}$ obtained in the diffusion dominated flow, $\alpha = 10$, shown in figure 4.13 indicates that the rotational structure never develops in this regime and the solution behaviour more closely resembles a perturbation of the time-dependent diffusion equation.

From the time series the ability of the method to retain symmetries in the solution can be evaluated. The symmetry present in these solution is rotational, 180^o about $(0.5, 0.5)$, and from the contour plots in figures 4.12 and 4.13 it can be seen that this behaviour is largely apparent in the solution, even after considerable time. The retention of these symmetries in the numerical solution may be in part due to the regular grid employed, and the coincidence of a mesh node with the point around which the solution is symmetrical. A more stringent test of the numerical method would be to implement the method on an irregular grid and monitor the same symmetries.

From figure 4.11 the evolution of $\Theta_h(\widehat{u}(\mathbf{x})^{t+\Delta t})$ indicates that the integral of the numerical solutions are ever decreasing. Although Θ_h is not the same measure as $\|\widehat{u}(\mathbf{x})^{t+\Delta t}\|_\Omega$ the decay of the numerical approximations to the quantity of interest, without oscillations, indicates that the numerical solutions are replicating the geometric property of the continuous problem described in section 4.2.1. Although the numerical method appears to effectively capture the solution attributes in advection dominated flows, the approximations to the quantity of interest appears significantly more accurate in diffusion dominant flows. The accuracy of the approximations to the quantity of interest is indicated by how tight the two bounds are, a natural indication of how well the semi-discrete solution has been resolved at each time step.

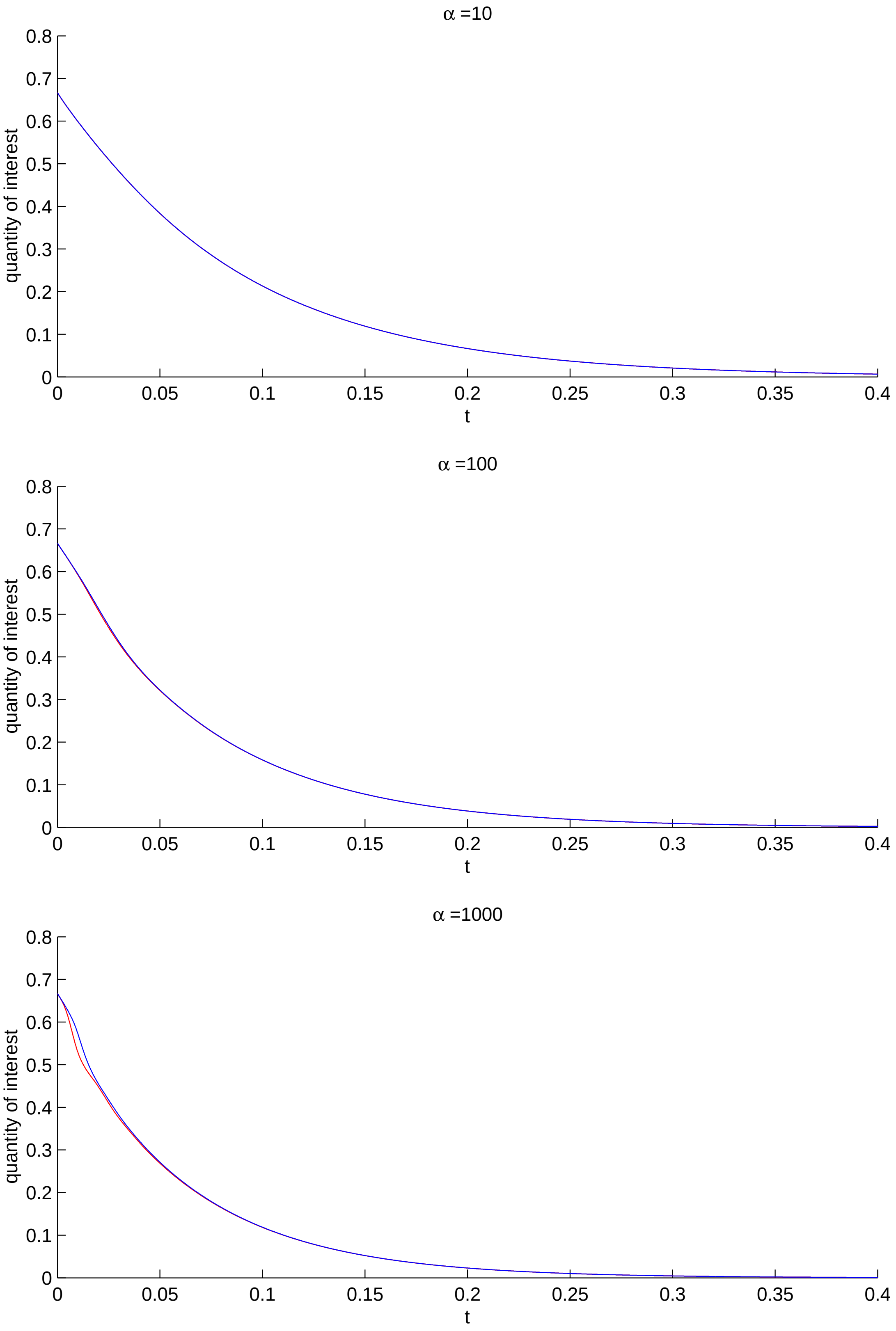

Figure 4.11: Bounds on $\Theta_h(\widehat{u}(\mathbf{x})^{t+\Delta t})$ with increasing α

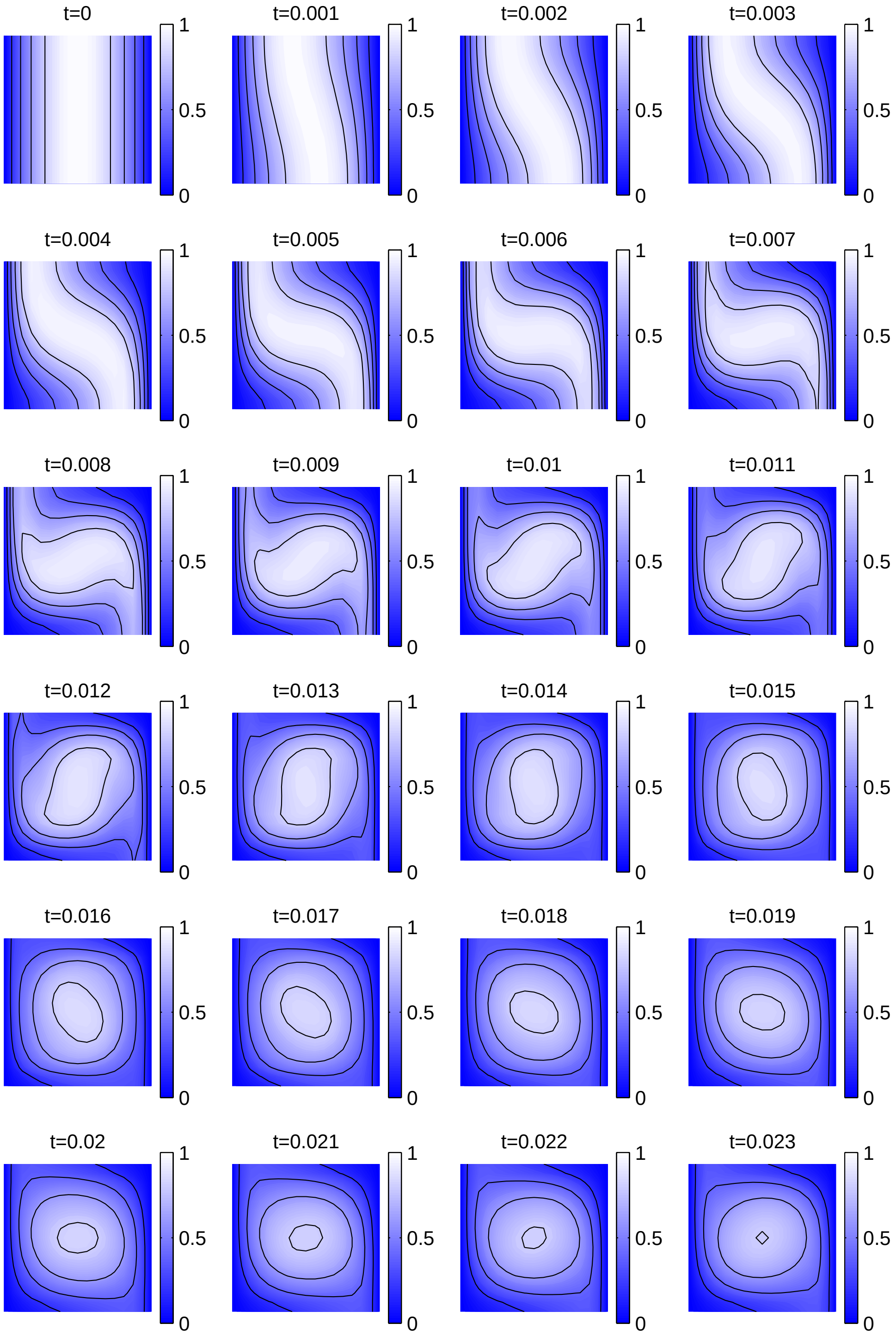

Figure 4.12: Time series of the numerical solution $u(x)^{t+\Delta t}$, $\alpha = 1000$

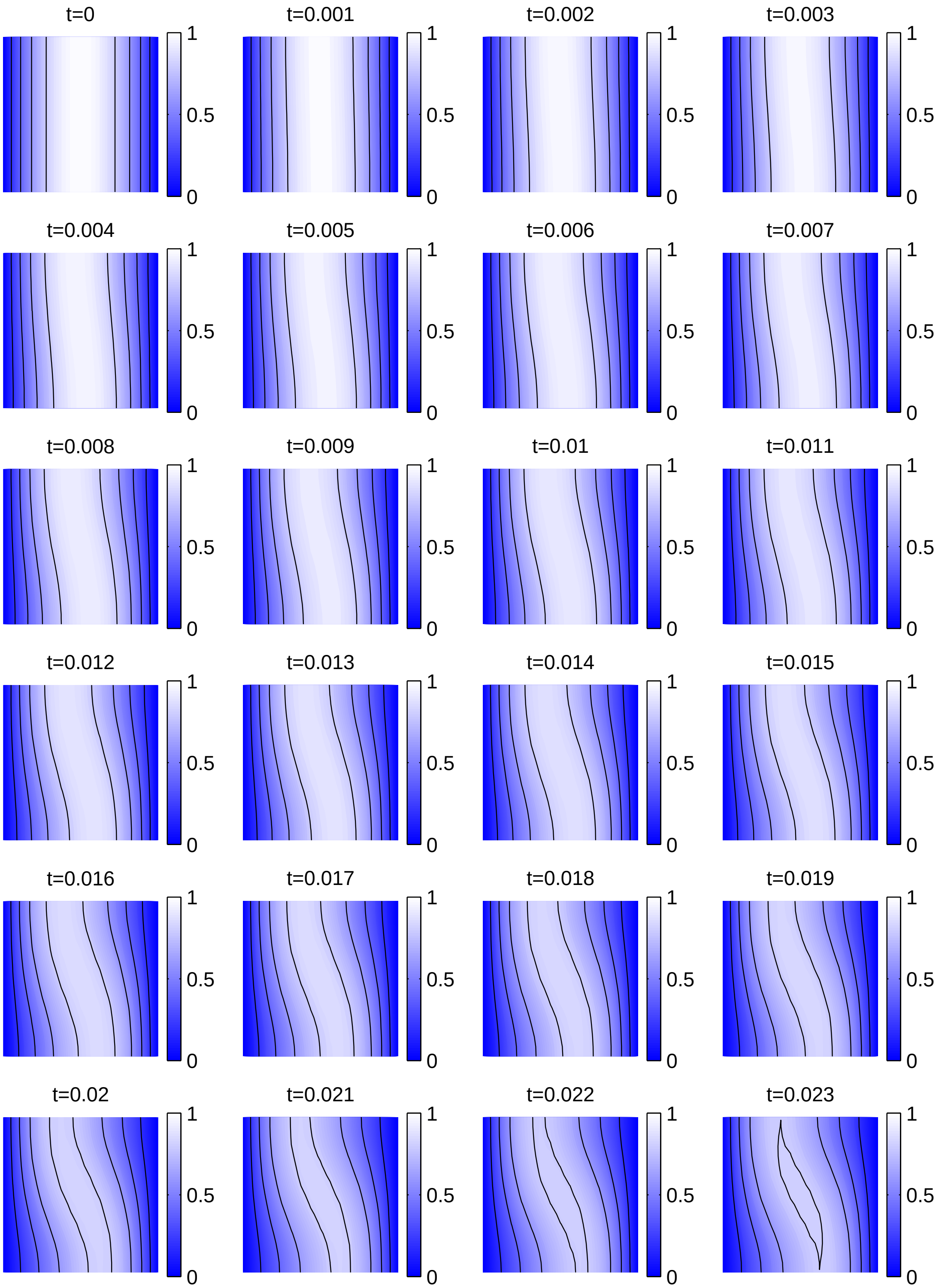

Figure 4.13: Time series of the numerical solution $u(x)^{t+\Delta t}$, $\alpha = 10$

4.3 Conclusions and Extensions

In this chapter semi-discrete methods have been constructed for which existing dual extremum principles can be used to obtain numerical solutions. The crux of the method is choosing appropriate discretisations in order that the semi-discrete problem is defined appropriately. In the examples considered a discretisation in time was required to reduce the continuous problem to a series of semi-discrete problems governed by the Helmholtz operator. Discretising the continuous problem is also the weakness of the method, as contact with the analytic solution is then lost. As a result of the initial discretisations the bounds obtained on the semi-discrete problem serve purely as approximations to the analytic quantity of interest. The approximations to the quantity of interest were found to be reasonably accurate, especially in the diffusion dominated solutions, but the confidence associated with the bounds obtained in previous chapters is absent. It may be possible to increase the accuracy of the approximations by employing a high order discretisation for the time and space derivatives, but bounds on the analytic quantity of interest can only be obtained if the error incurred in the approximations to the quantity of interest can be bounded at each time step. One such method of bounding the error introduced by the discretisation of the continuous problem is illustrated in figure 4.14.

In essence bounds are obtained on the analytic quantity of interest by attempting to construct discretisations that result in over and under approximations to the derivatives, and from which an envelope containing the quantity of interest of the continuous problem can be constructed. The obvious manner to achieve this is to attempt to simulate the gradient of the analytic quantity of interest at either the start or end point of the time step and apply forward or backward discretisations appropriately. This technique corresponds to the bounds constructed in figure 4.14 and the effectiveness of methods based on this theme are explored in the following example.

4.3.1 Bounds Constructed from Forward and Backward Discretisations

An investigation into the practicalities of generating a numerical method based on the techniques embodied in figure 4.14 is made using the one dimensional time-dependent

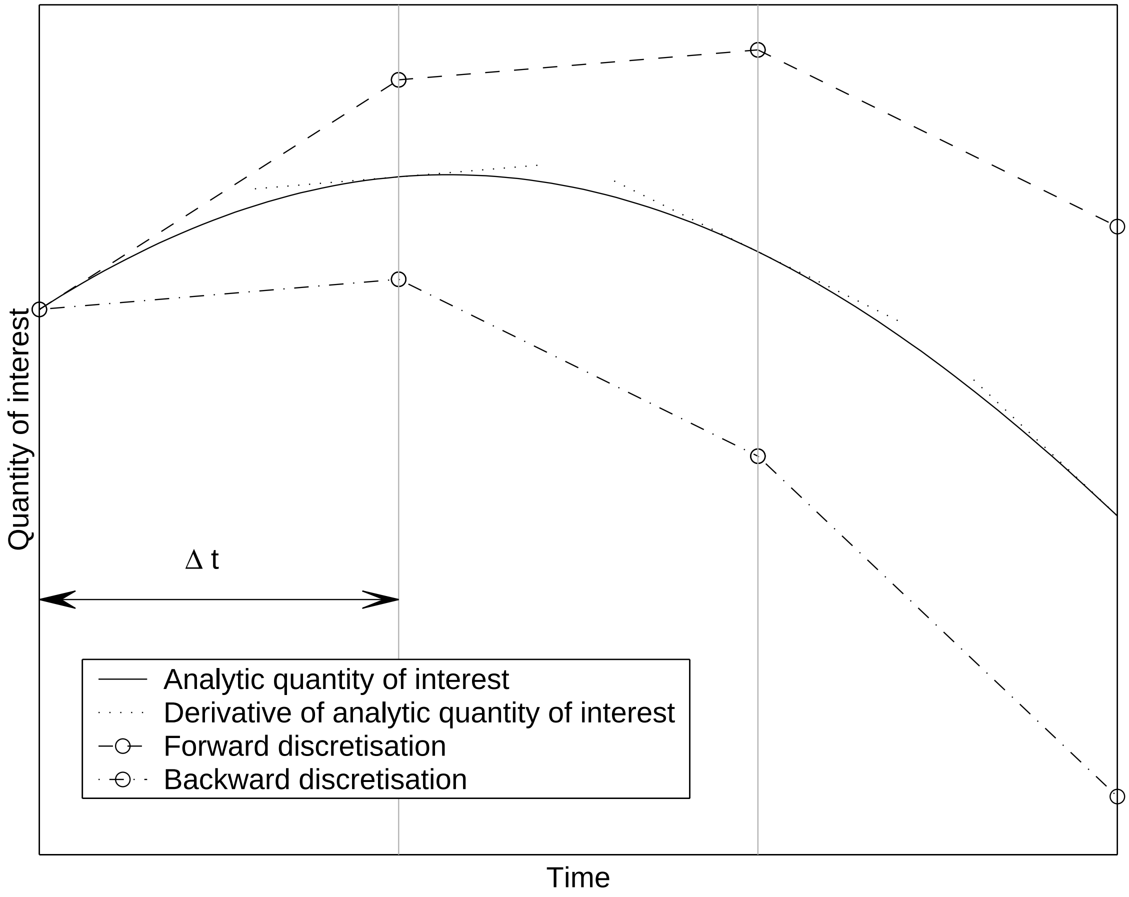

Figure 4.14: Schematic bounds obtained using forward and backward derivatives

diffusion example considered in section 4.1.4, namely

$$\frac{\partial \widehat{u}}{\partial t} - \frac{\partial^2 \widehat{u}}{\partial x^2} = 0, \tag{4.156}$$

in the domain $0 \leq x \leq 1$, $0 \leq t \leq \infty$, with boundary and initial conditions

$$\begin{aligned} \widehat{u}(0,t) &= 0, & (4.157) \\ -\widehat{u}(1,t)_x &= 0, & (4.158) \\ \widehat{u}(x,0) &= x(2-x). & (4.159) \end{aligned}$$

The quantity of interest was again chosen to be

$$\Theta(\widehat{u}) = \int_0^1 \widehat{u}(x,t)\,dx. \tag{4.160}$$

The known analytic solution (4.73) and quantity of interest (4.75) enable the performance of the results obtained using the different discretisation methods to be evaluated.

Initially the first time step of the method is considered. The two approximations for the quantity of interest at $t = \Delta t$ are then

$$\begin{aligned}\Theta_{fd} &= \Theta(\widehat{u}(x,0)) + \Theta'(\widehat{u}(x,0))\,\Delta t \qquad \text{forward} && (4.161)\\ \Theta_{bd} &= \Theta(\widehat{u}(x,0)) + \Theta'(\widehat{u}(x,\Delta t))\,\Delta t \qquad \text{backward} && (4.162)\end{aligned}$$

which, given a small enough time step, satisfy either the bounds

$$\Theta_{fd} \leq \Theta(\widehat{u}(x,\Delta t)) \leq \Theta_{bd} \tag{4.163}$$

or conversely

$$\Theta_{bd} \leq \Theta(\widehat{u}(x,\Delta t)) \leq \Theta_{fd} \tag{4.164}$$

depending on the evolution of the quantity of interest. Although the forward and backward discretisations contain the derivative of the analytic quantity of interest, and therefore cannot be implemented directly, numerical approximations to these schemes can be constructed. Considering the quantity of interest to be defined as

$$\Theta = \langle \widehat{u}, r\rangle, \tag{4.165}$$

the backwards discretisation method suggests the following semi-discrete approximation $\Theta^{bd}_h \approx \Theta^{bd}$,

$$\begin{aligned}\Theta^{bd}_h(\widehat{u}^{t+\Delta t}_{bd}) &= \langle \widehat{u}^{t}_{bd}, r\rangle + \Delta t\langle \nabla\cdot\nabla\widehat{u}^{t+\Delta t}_{bd}, r\rangle, && (4.166)\\ &= \langle \widehat{u}^{t+\Delta t}_{bd}, r\rangle, && (4.167)\end{aligned}$$

where $u^{t+\Delta t}_{bd}$ satisfies the analytic solution of the implicit numerical method

$$\frac{\widehat{u}^{t+\Delta t}_{bd} - \widehat{u}^{t}_{bd}}{\Delta t} = \nabla\cdot\nabla\widehat{u}^{t+\Delta t}_{bd}. \tag{4.168}$$

Similarly, the forward discretisation suggests the semi-discrete approximation $\Theta^{fd}_h \approx \Theta^{fd}$,

$$\begin{aligned}\Theta^{fd}_h(\widehat{u}^{t+\Delta t}_{fd}) &= \langle \widehat{u}^{t}_{fd}, r\rangle + \Delta t\langle \nabla\cdot\nabla\widehat{u}^{t}_{fd}, r\rangle, && (4.169)\\ &= \langle \widehat{u}^{t+\Delta t}_{fd}, r\rangle, && (4.170)\end{aligned}$$

and the explicit numerical scheme

$$\frac{\widehat{u}^{t+\Delta t}_{fd} - \widehat{u}^{t}_{fd}}{\Delta t} = \nabla\cdot\nabla\widehat{u}^{t}_{fd}. \tag{4.171}$$

In fact, both of the numerical methods derived, (4.168) and (4.170), are specific examples of the general theta method described in section 4.1 with $\theta = 1$ and $\theta = 0$ respectively. In section 4.1 it was found that, as a result of applying the theta method, approximations to the solution of the time-dependent diffusion problem at each time-step are governed by the Helmholtz equation. Moreover, using the dual extremum principles to construct these approximations enables bounds on quantities of interest, at each time step, to be found.

However, in order to apply the Helmholtz functional a degree of implicitness is required in the numerical scheme and therefore the modification

$$\frac{\widehat{u}_{fd}^{t+\Delta t} - \widehat{u}_{fd}^{t}}{\Delta t} = 0.9\,\nabla \cdot \nabla \widehat{u}_{fd}^{t} + 0.1\,\nabla \cdot \nabla \widehat{u}_{fd}^{t+\Delta t}, \tag{4.172}$$

to (4.170) is suggested. Similarly, for balance, the modification

$$\frac{\widehat{u}_{bd}^{t+\Delta t} - \widehat{u}_{bd}^{t}}{\Delta t} = 0.1\,\nabla \cdot \nabla \widehat{u}_{bd}^{t} + 0.9\,\nabla \cdot \nabla \widehat{u}_{fd}^{t+\Delta t}, \tag{4.173}$$

is made to (4.168). With the first time-step complete the method is advanced in time. The key to maintaining the upper and lower bounds after multiple time-steps depends on the ability of the numerical solutions u_{fd}^{t}, $u_{fd}^{t+\Delta t}$, u_{bd}^{t} and $u_{bd}^{t+\Delta t}$ to accurately approximate the derivatives of the quantity of interest.

Implementing the technique is a straightforward procedure because employing the same quantity of interest as in section 4.1.4 enables the pair of twinned problems derived in that section to again be re-cycled. From the pair of twinned problems upper and lower bounds on the quantity of interest in terms of the analytic solution of the semi-discrete problem can be found for both forward and backwards discretisations. Then, due to the concave nature of the analytic quantity of interest in time, figure 4.3, a proposed upper bound is found by using the backward discretisation at each time step. A proposed lower bound is found using the forward discretisation at each time step. The results obtained using a time-step of 0.01 and 1 quadratic element are shown in figure 4.15.

From figure 4.15, it can be seen that after a relatively short period of time the proposed lower bound no longer lies below the analytic quantity of interest. The failure of the lower bound was observed for a range of time-steps and grid densities. The failure can be attributed to errors propagated in the solutions $u_{fd}^{t+\Delta t}$ which over time render

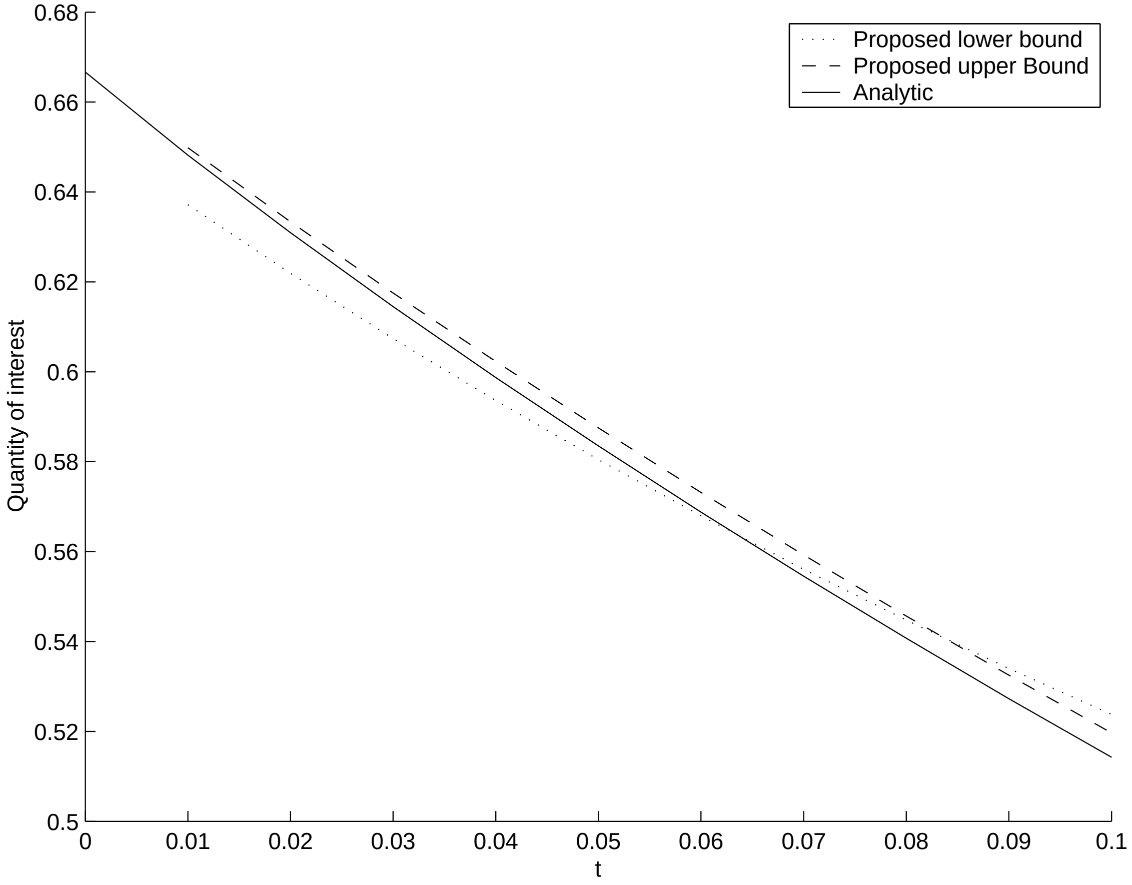

Figure 4.15: Bounds on the continuous problem using forward and backward derivatives

the numerical solution a poor candidate from which to calculate approximations to the gradient of the quantity of interest. In addition, the errors in the solution $u_{fd}^{t+\Delta t}$ tend to overestimate the gradient of the quantity of interest, destroying the lower bound. The possibility that the overestimate of the gradient of the quantity of interest is linked to the convexity of the functional $\mathcal{G}^-$ effectively employed in constructing the semi-discrete schemes (4.168) and (4.168), requires further research. However if this proves a plausible explanation for the poor lower bound, generating semi-discrete methods based on the dual functional $\mathcal{G}^+$ could be explored.

The failure to construct effective upper and lower bounds on an evolving quantity of interest from a series of discrete approximations suggests that methods in which the solution is considered over the complete space-time domain simultaneously, should be considered. Two such alternatives are now discussed.

4.3.2 Laplace Transforms

The first alternative to computing a series of time-stepped solutions is to apply a Laplace transform to the governing equation in order to transform the time derivative into an algebraic relationship. This method is considered by Gurtin [19] and enables the minimum and maximum principles to be applied in the transformed space. For practical problems numerical inversion of the Laplace transform would probably then be favoured in order to recover approximations to the quantity of interest. Numerical methods to invert the Laplace transform are reviewed in [14] and provided a sufficient degree of accuracy could be obtained, bounds on the analytic quantity of interest would be established.

The application of the Laplace transform is obvious in the case of the time-dependent diffusion case but in the advection-diffusion case the first order spatial derivative remains. Despite the possibility of removing the remaining spatial derivative with a further transform implementation of such techniques is likely to hinder the application of the extremum principles in general.

4.3.3 Continuous Time

The second alternative to a time-stepping method is to discretise the governing equation over the complete spatial and temporal domain simultaneously. Associated with this technique is a greater computational cost, especially as the end time increases and the domain grows. However, the accumulation of errors experienced as a solution is marched forward in time is avoided. One technique to solve over the complete space-time domain is to effectively square the operator in the governing equation. The primal problem to solve is then (in a general framework)

$$A^*(A\widehat{\phi} - s) = 0, \tag{4.174}$$

as opposed to the original non-self-adjoint equation

$$A\widehat{\phi} = s. \tag{4.175}$$

The effect of squaring the operator renders the originally non-self-adjoint problem self-adjoint, and generates the additional boundary conditions required around the domain. Methods based on solving over the complete space-time domain will be developed in chapter 5.

In addition to governing the solution of the semi-discrete equations considered in this chapter, the Helmholtz functional motivates two alternative approaches to obtaining the relatively difficult upper bound associated with the diffusion operator. The first suggestion is a new constraint for the diffusion functional.

4.3.4 A Helmholtz Inspired Constraint

The H^+ constraint for the Helmholtz functional enables the function p to be expressed in terms of the function q without having to invert the operator T^*. This is a desirable attribute that enables the constraint to be satisfied through direct substitution. The H^+ constraint for the diffusion functional is lacking this property. However a new hyper-line, H^{++} passing through the stationary point $(\widehat{p}, \widehat{q})$, and on which the functional is convex, can be found by inspection. The equation for the new hyper-line,

$$p = \widehat{p} + r - T^* q, \qquad (H^{++}) \tag{4.176}$$

enables the function p to be expressed in terms of the function q, but also contains the analytic solution $\widehat{p}$. The inclusion of the analytic solution prevents the constraint from being used directly, but approximations to the constraint of the form

$$p = p^f + r - T^* q, \tag{4.177}$$

where p^f is a fine scale solution of suitable accuracy, are explored in chapter 5. A schematic representation of the hyperline H^{++} is plotted in figure 4.16 from which the similarities with the Helmholtz constraints, figure 4.1, can be identified.

The second suggestion prompted by the Helmholtz functional is time-stepping the upper and lower bounds to the required steady-state solution.

4.3.5 Time Stepping to a Steady-State Solution

For problems in which a time-dependent solution with homogenous boundary conditions can be shown to decay to zero a steady-state solution exists corresponding to the inhomogeneous boundary condition. Consider the problem

$$\begin{aligned} r &= \frac{\partial u}{\partial t} + T^* T u, && (4.178) \\ &= \frac{\partial u_{TD}}{\partial t} + T^* T (u_{TD} + u_{SS}), && (4.179) \end{aligned}$$

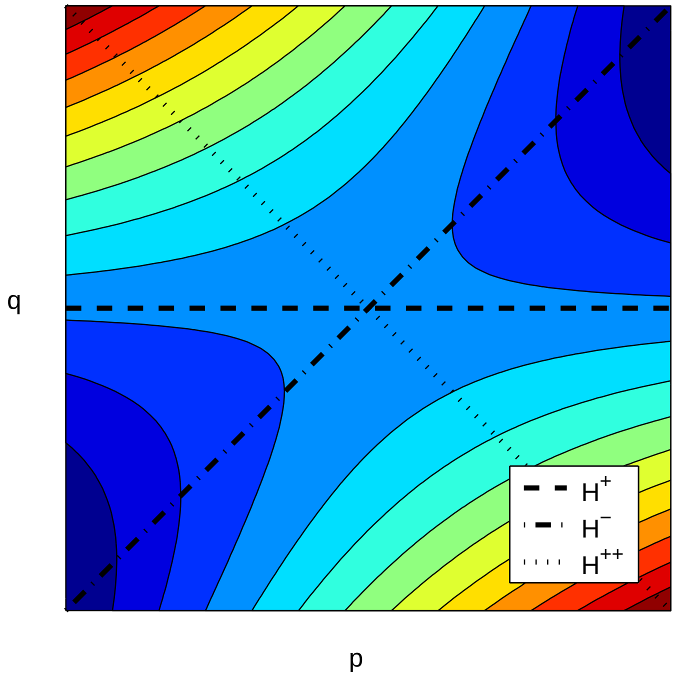

Figure 4.16: The H^{++} constraint

where

$$\frac{\partial u_{TD}}{\partial t} + T^*Tu_{TD} = 0, \tag{4.180}$$

and as a result of a geometric property similar to that described in section (4.1.6), $u_{TD} \to 0$ as $t \to \infty$. The steady-state solution then satisfies the equation

$$T^*Tu_{SS} = r \tag{4.181}$$

and hence as $t \to \infty$, $u \to u_{SS}$. The convergence of the solution of the time-dependent problem to that of a steady-state problem motivates considering the convergence of the bounds obtained from the time-dependent problem to that of the steady-state problem. The advantage of such a method would be the simplicity of the constraints associated with the Helmholtz functional governing the time-dependent solution, in comparison to those of the diffusion functional governing the steady-state solution. The possibility of obtaining bounds on the quantity of interest by time-stepping to the steady-state solution is also explored in chapter 5.

Chapter 5

The Non-Self-Adjoint Case - Continuous Time

In chapters 2 and 3 self-adjoint operators have been considered for which bounds on the stationary value of the governing functional are known to exist. The nature of these bounds all arise form the positivity inherent in some self-adjoint operator and which translates to the required convexity in the associated functional. When non-self-adjoint operators are considered this underlying framework is removed and upper and lower bounds of the type sought are not naturally forthcoming. In chapter 4, attempts to simulate these bounds were developed by careful discretisation of the non-self-adjoint components. However, the resulting bounds only approximated the quantity of interest and the long term behaviour of the numerical method could only be justified by mimicking additional properties of the analytic solution.

In this chapter a more general approach is taken in which the original non-self-adjoint equations are modified in order to obtain new self-adjoint problems. Bounding the quantity of interest is still the primary goal of the method and retaining the required quantity of interest whilst modifying the governing equations is crucial. This is achieved by adjusting the dual problem appropriately.

5.1 Construction of the Method

For completeness the method will be described from first principles although many of these techniques previously used will again be exercised. The techniques include twinning a pair of self-dual problems to obtain bounds on a non-self-dual problem and demonstrating extremum principles. The aim of the method is to bound the quantity of interest

$$\Theta(\widehat{\phi}) = \langle \widehat{\phi}, t \rangle, \tag{5.1}$$

where $\widehat{\phi}$ is the analytic solution of the primal problem

$$A\widehat{\phi} = s, \tag{5.2}$$

and A is a non-self-adjoint operator. The dual problem is governed by the adjoint operator A^* and is forced by the function t,

$$A^*\widehat{\sigma} = t. \tag{5.3}$$

The definition of the primal and dual problems coincides with the stationary point of the functional $\mathcal{G}(\phi, \sigma)$

$$\mathcal{G}(\phi, \sigma) = -\langle\!\langle \sigma, A\phi \rangle\!\rangle + \langle\!\langle \sigma, s \rangle\!\rangle + \langle t, \phi \rangle, \tag{5.4}$$

for the first variation of the functional $\mathcal{G}(\phi, \sigma)$ is

$$\delta\mathcal{G}(\phi, \sigma) = \langle\!\langle \delta\sigma, s - A\phi \rangle\!\rangle + \langle t - A^*\sigma, \delta\phi \rangle. \tag{5.5}$$

The stationary value of the functional

$$\mathcal{G}(\widehat{\phi}, \widehat{\sigma}) = \langle t, \widehat{\phi} \rangle = \langle\!\langle \widehat{\sigma}, A\widehat{\phi} \rangle\!\rangle \tag{5.6}$$

coincides with the quantity of interest sought but, due to the lack of self-adjointness, the factorisation $A = T^*T$ is not available and the required positivity of the inner product is absent. With the aim of introducing positivity into the stationary value the alternative functional

$$\mathcal{J}(u, v) = -\langle v, A^*Au \rangle + \langle v, A^*s \rangle + \langle t, u \rangle \tag{5.7}$$

is considered. The first order variation of the functional $\mathcal{J}(u, v)$ is

$$\delta\mathcal{J}(u, v) = \langle \delta v, A^*(s - Au) \rangle + \langle t - A^*Av, \delta u \rangle, \tag{5.8}$$

and hence the stationary conditions of the functional are

$$\begin{aligned} A^*(A\widehat{u} - s) &= 0 \qquad \text{Primal Problem,} & (5.9) \\ A^*A\widehat{v} &= t \qquad \text{Dual Problem.} & (5.10) \end{aligned}$$

In this arrangement of the problem it is assumed that the null spaces of A and A^* are empty, or equivalently that the solutions $\widehat{\phi}$ and $\widehat{\sigma}$ are unique. Under these assumptions the two solutions $\widehat{\phi}$ and $\widehat{u}$ of (5.2) and (5.9) are equal. In contrast the dual solutions $\widehat{\sigma}$ and $\widehat{v}$ satisfy different equations, (5.3) and (5.10), and have a different character from each other. Critically the stationary value of the functional $\mathcal{J}(u, v)$,

$$\mathcal{J}(\widehat{u}, \widehat{v}) = \langle t, \widehat{u}\rangle = \langle\langle A\widehat{v}, A\widehat{u}\rangle\rangle \tag{5.11}$$

is identical to the quantity of interest $\Theta(\widehat{\phi})$ through the coincidence of $\widehat{\phi}$ and $\widehat{u}$. In addition, the stationary value $\mathcal{J}(\widehat{u}, \widehat{v})$ can be evaluated as the difference of two positive inner products, using the twinning method, and therefore the possibility of bounding the quantity of interest is present.

The modifications to the operator have succeeded in producing self-adjoint primal and dual problems, and the required positivity in the stationary value inner product. In relationship to the existing literature if a discretisation using finite elements is employed the primal problem (5.9) is termed a least-squares finite element method (LSFEM), for which extensive analysis exists including [25]. One advantageous characteristic of the method is that the self-adjointness of the operator A^*A translates to a symmetric, positive-definite matrix in a finite-dimensional discretisation, for which efficient methods such as conjugate gradients can be used.

5.1.1 De-Twinning

The operator A^*A is self-adjoint and positive, and in order to obtain the quantity of interest as the difference between two positive quantities the twinning transformations are again used,

$$p_1 = u + v, \qquad u = \frac{1}{2}(p_1 + p_2), \qquad r_1 = A^*s + t, \tag{5.12}$$

$$p_2 = u - v, \qquad v = \frac{1}{2}(p_1 - p_2), \qquad r_2 = A^*s - t. \tag{5.13}$$

The twinning method transforms the functional $\mathcal{J}(u, v)$ of (5.7) into

$$\mathcal{J}(p_1, p_2) = \left[-\frac{1}{4}\langle p_1, A^*Ap_1\rangle + \frac{1}{2}\langle p_1, r_1\rangle\right] + \left[\frac{1}{4}\langle p_2, A^*Ap_2\rangle - \frac{1}{2}\langle p_2, r_2\rangle\right]. \tag{5.14}$$

This functional can then naturally be decoupled into the pair of ‘twinned’ problems corresponding to the squared brackets in (5.14). Finally we introduce the intermediate variables q_1, q_2 where

$$q_1 = Ap_1, \tag{5.15}$$

$$q_2 = Ap_2, \tag{5.16}$$

which enables $\mathcal{J}(u, v)$ to be represented as

$$\begin{aligned}\mathcal{J}(p_1, q_1, p_2, q_2) &= \left[-\frac{1}{2}\langle\langle q_1, Ap_1\rangle\rangle + \frac{1}{4}\langle\langle q_1, q_1\rangle\rangle + \frac{1}{2}\langle p_1, r_1\rangle\right] \\ &\quad + \left[\frac{1}{2}\langle\langle q_2, Ap_2\rangle\rangle - \frac{1}{4}\langle\langle q_2, q_2\rangle\rangle - \frac{1}{2}\langle p_2, r_2\rangle\right], \qquad (5.17) \\ &= \frac{1}{2}\mathcal{T}_1(p_1, q_1) - \frac{1}{2}\mathcal{T}_2(p_2, q_2), \qquad (5.18)\end{aligned}$$

where

$$\mathcal{T}_i(p_i, q_i) = \frac{1}{2}\langle\langle q_i, q_i\rangle\rangle - \langle\langle Ap_i, q_i\rangle\rangle + \langle p_i, r_i\rangle \qquad i = 1, 2 \tag{5.19}$$

is recognised to be of the same form as that obtained in the diffusion functional. The diffusion functional has been shown to be saddle shaped and as a result of this topology, and provided the appropriate constraints can be satisfied, bounds can be constructed on the stationary value $\mathcal{T}_i(\widehat{p}_i, \widehat{q}_i)$, of the form

$$\mu_i^- \leq \mathcal{T}_i(\widehat{p}_i, \widehat{q}_i) \leq \mu_i^+ \qquad i = 1, 2. \tag{5.20}$$

In particular, bounds on the quantity of interest $\Theta(\widehat{\phi})$ can then be calculated as

$$\Theta(\widehat{\phi}) = \mathcal{J}(\widehat{u}, \widehat{v}) = \frac{1}{2}\mathcal{T}_1(\widehat{p}_1, \widehat{q}_1) - \frac{1}{2}\mathcal{T}_2(\widehat{p}_2, \widehat{q}_2), \tag{5.21}$$

and hence the quantity of interest is bounded above and below by the computable bounds

$$\frac{1}{2}(\mu_1^- - \mu_2^+) \leq \Theta(\widehat{\phi}) \leq \frac{1}{2}(\mu_1^+ - \mu_2^-). \tag{5.22}$$

The ability to obtain the bounds (5.20) is solely dependent on being able to satisfy the appropriate constraints. In section 2.1.3 the lower bound on the stationary value of the functional was found by constraining the functional to satisfy

$$q = Ap. \tag{5.23}$$

This constraint, as with the diffusion functional, is simply satisfied by direct substitution to obtain the constrained functional $\mathcal{T}^-(p)$,

$$\mathcal{T}^-(p) = -\frac{1}{2}\langle\langle Ap, Ap\rangle\rangle + \langle p, r\rangle, \tag{5.24}$$

which is maximised at $\widehat{p}$ since

$$\begin{aligned}
\mathcal{T}^{-}(p) - \mathcal{T}^{-}(\widehat{p}) &= -\frac{1}{2}\langle\!\langle Ap, Ap\rangle\!\rangle + \langle p, r\rangle + \frac{1}{2}\langle\!\langle A\widehat{p}, A\widehat{p}\rangle\!\rangle - \langle \widehat{p}, r\rangle, && (5.25)\\
&= -\frac{1}{2}\langle\!\langle Ap, Ap\rangle\!\rangle + \langle\!\langle p, A^*A\widehat{p}\rangle\!\rangle - \frac{1}{2}\langle\!\langle A\widehat{p}, A\widehat{p}\rangle\!\rangle, && (5.26)\\
&= -\frac{1}{2}\langle\!\langle Ap - A\widehat{p}, Ap - A\widehat{p}\rangle\!\rangle, && (5.27)\\
&\leq 0. && (5.28)
\end{aligned}$$

A lower bound on the stationary value is therefore found by maximising the functional $\mathcal{T}^{-}(p)$ in a finite dimensional subspace.

An upper bound on the stationary value of the diffusion functional was found by constraining the functional such that

$$A^*q = r. \qquad (5.29)$$

However in this case, as a consequence of the uniqueness assumptions on the inversion of the operator A^*, satisfying the constraint $A^*q = r$ strongly is no easier than solving the original dual problem. As a result, an alternative method of obtaining an upper bound on the stationary value $\mathcal{T}(\widehat{p}, \widehat{q})$ is required. Two alternative methods were inspired by the Helmholtz functional, the alternative constraint in section 4.3.4, and time-stepping to a steady-state solution in section 4.3.5. The validity and practicality of these two methods are now considered.

5.1.2 Alternative Upper Bounds on $\mathcal{T}(\widehat{p}, \widehat{q})$

One approach to obtain an upper bound on the stationary value $\mathcal{T}(\widehat{p}, \widehat{q})$ is to build the constraint $A^*q = r$ into the functional. A simplistic method of achieving this is to simply add the residual squared, multiplied by a user defined constant α to obtain

$$\mathcal{T}^{AL}(q) = \frac{1}{2}\langle\!\langle q, q\rangle\!\rangle + \alpha\langle A^*q - r, A^*q - r\rangle. \qquad (5.30)$$

Functionals of this nature are called augmented Lagrangians and heuristically, minimisation of the functional should both result in the constraint being closely satisfied and the minimisation of $\langle\!\langle q, q\rangle\!\rangle$. In order to achieve this balance α must be chosen carefully and possibly adjusted during the minimisation process. Beside the complexities of determining α to ensure a useful minimum is attained, choosing α such that

$$\mathcal{T}^{AL}(q) - \mathcal{T}^{AL}(\widehat{q}) = \frac{1}{2}\langle\!\langle q, q\rangle\!\rangle - \frac{1}{2}\langle\!\langle \widehat{q}, \widehat{q}\rangle\!\rangle + \alpha\langle A^*q - r, A^*q - r\rangle, \qquad (5.31)$$

is greater than zero is non-trivial. A superior method to 'augment' the functional is to consider the H^{++} constraint inspired by the Helmholtz functional. The H^{++} constraint is

$$p = \widehat{p} + r - A^* q. \qquad (H^{++}) \tag{5.32}$$

Substituting the constraint into $\mathcal{T}(p,q)$ the constrained functional $\mathcal{T}^{++}(q)$ is obtained

$$\begin{aligned} \mathcal{T}^{++}(q) &= \frac{1}{2}\langle\!\langle q,q\rangle\!\rangle - \langle\!\langle A\widehat{p}, q\rangle\!\rangle - \langle r - A^*q, A^*q\rangle + \langle \widehat{p} + r - A^*q, r\rangle, \\ &= \frac{1}{2}\langle\!\langle q,q\rangle\!\rangle - \langle\!\langle \widehat{q}, q - \widehat{q}\rangle\!\rangle + \langle A^*(\widehat{q} - q), A^*(\widehat{q} - q)\rangle, \end{aligned} \tag{5.33}$$

which is an upper bound on the stationary value of the functional, for

$$\begin{aligned} \mathcal{T}^{++}(q) - \mathcal{T}^{++}(\widehat{q}) &= \frac{1}{2}\langle\!\langle q,q\rangle\!\rangle - \langle\!\langle \widehat{q}, q - \widehat{q}\rangle\!\rangle + \langle A^*(\widehat{q} - q, A^*(\widehat{q} - q)\rangle - \frac{1}{2}\langle\!\langle \widehat{q}, \widehat{q}\rangle\!\rangle, \\ &= \frac{1}{2}\langle\!\langle q - \widehat{q}, q - \widehat{q}\rangle\!\rangle + \langle A^*(\widehat{q} - q), A^*(\widehat{q} - q)\rangle, \\ &\geq 0. \end{aligned} \tag{5.34}$$

In addition, the competition between minimising the two inner products of the functional $\mathcal{T}^{AL}(q)$ is removed. Taking the first variation of the functional $\mathcal{T}^{++}(q)$

$$\delta\mathcal{T}^{++}(q) = \langle\!\langle \delta q, q - \widehat{q} + 2AA^*(\widehat{q} - q)\rangle\!\rangle, \tag{5.35}$$

the eigenvalue problem

$$(AA^* - \left(-\frac{1}{2}\right) I)(\widehat{q} - q) = 0 \tag{5.36}$$

is obtained at the stationary point, and as the eigenvalues of the positive operator AA^* are all greater than zero (5.36) implies $q = \widehat{q}$. Therefore at the stationary point of the functional both the minimum of the inner product $\langle\!\langle q, q\rangle\!\rangle$ and the constraint $A^*q = r$ are satisfied.

As remarked in section 4.3.4 the constraint (5.32) is not practically implementable due to the appearance of the unknown analytic solution $\widehat{p}$. Instead, using an accurate fine-scale solution p^f is considered. The constrained functional $\mathcal{T}^{++}(q)$ is then

$$\mathcal{T}^{++}(q) = \frac{1}{2}\langle\!\langle q,q\rangle\!\rangle - \langle\!\langle Ap^f, q - \widehat{q}\rangle\!\rangle + \langle A^*(\widehat{q} - q), A^*(\widehat{q} - q)\rangle, \tag{5.37}$$

and an upper bound exists if,

$$\begin{aligned} \mathcal{T}^{++}(q) - \mathcal{T}^{++}(\widehat{q}) &= \frac{1}{2}\langle\!\langle q,q\rangle\!\rangle - \langle\!\langle Ap^f, q - \widehat{q}\rangle\!\rangle + \langle A^*(\widehat{q} - q, A^*(\widehat{q} - q)\rangle - \frac{1}{2}\langle\!\langle \widehat{q}, \widehat{q}, \rangle \\ &= \frac{1}{2}\langle\!\langle q - \widehat{q}, q - \widehat{q}\rangle\!\rangle - \langle\!\langle Ap^f - \widehat{q}, q - \widehat{q}\rangle\!\rangle + \langle A^*(\widehat{q} - q), A^*(\widehat{q} - q)\rangle, \end{aligned} \tag{5.38}$$

is greater than zero. The difference between the functionals can be re-written by completing the square to give

$$\begin{aligned}
\mathcal{T}^{++}(q)-\mathcal{T}^{++}(\widehat{q}) &= \frac{1}{2}\langle\!\langle q-\widehat{q},q-\widehat{q}\rangle\!\rangle-\langle\!\langle Ap^f-\widehat{q},q-\widehat{q}\rangle\!\rangle+\langle A^*(\widehat{q}-q),A^*(\widehat{q}-q)\rangle,\\
&= \frac{1}{2}\langle\!\langle\,(q-\widehat{q})-(Ap^f-\widehat{q}),(q-\widehat{q})-(Ap^f-\widehat{q})\,\rangle\!\rangle\\
&\quad -\frac{1}{2}\langle\!\langle Ap^f-\widehat{q},Ap^f-\widehat{q}\rangle\!\rangle+\langle A^*(\widehat{q}-q),A^*(\widehat{q}-q)\rangle,\\
&= \frac{1}{2}\langle\!\langle Ap^f-q,Ap^f-q\rangle\!\rangle-\frac{1}{2}\langle\!\langle Ap^f-\widehat{q},Ap^f-\widehat{q}\rangle\!\rangle\\
&\quad +\langle A^*(\widehat{q}-q),A^*(\widehat{q}-q)\rangle,\\
&= \frac{1}{2}\|Ap^f-q\|^2_{\langle\!\langle\rangle\!\rangle}-\frac{1}{2}\|Ap^f-\widehat{q}\|^2_{\langle\!\langle\rangle\!\rangle}+\|A^*(\widehat{q}-q)\|^2_{\langle\rangle},
\end{aligned}\tag{5.39}$$

and therefore provided

$$\frac{1}{2}\|Ap^f-q\|^2_{\langle\!\langle\rangle\!\rangle}+\|A^*(\widehat{q}-q)\|^2_{\langle\rangle}\geq\frac{1}{2}\|Ap^f-\widehat{q}\|^2_{\langle\!\langle\rangle\!\rangle},\tag{5.40}$$

$\mathcal{T}^{++}(q)$ is an upper bound on the stationary value.

A slightly weaker inequality is found by directly comparing the functionals $\mathcal{T}(p^f,q)$ and $\mathcal{T}(\widehat{p},\widehat{q})$ with p^f and q unconstrained,

$$\begin{aligned}
\mathcal{T}(p^f,q)-\mathcal{T}(\widehat{p},\widehat{q}) &= \frac{1}{2}\langle\!\langle q,q\rangle\!\rangle-\langle p^f,A^*q\rangle+\langle p,s\rangle-\frac{1}{2}\langle\!\langle\widehat{q},\widehat{q}\rangle\!\rangle+\langle\widehat{p},A^*\widehat{q}\rangle-\langle\widehat{p},s\rangle,\\
&= \frac{1}{2}\langle\!\langle q,q\rangle\!\rangle-\langle p^f,A^*q\rangle+\langle p^f,A^*\widehat{q}\rangle-\frac{1}{2}\langle\!\langle\widehat{q},\widehat{q}\rangle\!\rangle,\\
&= \frac{1}{2}\langle\!\langle q,q\rangle\!\rangle-\langle\!\langle Ap^f,q-\widehat{q}\rangle\!\rangle-\frac{1}{2}\langle\!\langle\widehat{q},\widehat{q}\rangle\!\rangle,\\
&= \frac{1}{2}\langle\!\langle q,q\rangle\!\rangle-\langle\!\langle\widehat{q},q-\widehat{q}\rangle\!\rangle-\frac{1}{2}\langle\!\langle\widehat{q},\widehat{q}\rangle\!\rangle-\langle\!\langle Ap^f,q-\widehat{q}\rangle\!\rangle+\langle\!\langle\widehat{q},q-\widehat{q}\rangle\!\rangle,\\
&= \frac{1}{2}\langle\!\langle q-\widehat{q},q-\widehat{q}\rangle\!\rangle-\langle\!\langle Ap^f-\widehat{q},q-\widehat{q}\rangle\!\rangle.
\end{aligned}\tag{5.41}$$

Again, the difference between the functionals can be re-written by completing the square to give

$$\begin{aligned}
\mathcal{T}(p^f,q)-\mathcal{T}(\widehat{p},\widehat{q}) &= \frac{1}{2}\langle\!\langle q-\widehat{q},q-\widehat{q}\rangle\!\rangle-\langle\!\langle Ap^f-\widehat{q},q-\widehat{q}\rangle\!\rangle,\\
&= \frac{1}{2}\langle\!\langle\,(q-\widehat{q})-(Ap^f-\widehat{q}),(q-\widehat{q})-(Ap^f-\widehat{q})\,\rangle\!\rangle\\
&\quad -\frac{1}{2}\langle\!\langle Ap^f-\widehat{q},Ap^f-\widehat{q}\rangle\!\rangle,\\
&= \frac{1}{2}\langle\!\langle Ap^f-q,Ap^f-q\rangle\!\rangle-\frac{1}{2}\langle\!\langle Ap^f-\widehat{q},Ap^f-\widehat{q}\rangle\!\rangle,\\
&= \frac{1}{2}\|Ap^f-q\|^2_{\langle\!\langle\rangle\!\rangle}-\frac{1}{2}\|Ap^f-\widehat{q}\|^2_{\langle\!\langle\rangle\!\rangle},
\end{aligned}\tag{5.42}$$

and therefore provided

$$\|Ap^f - q\|^2_{\langle\langle\rangle\rangle} \geq \|Ap^f - \widehat{q}\|^2_{\langle\langle\rangle\rangle}, \tag{5.43}$$

$\mathcal{T}(p^f, q)$ is an upper bound on the stationary value. The inequality (5.43) also illustrates the lower bound available if the constraint $q = Ap^f$ is applied. One strategy would therefore be to obtain a lower bound using the standard constraint $q = Ap^f$, and obtain an upper bound by ensuring that the relative errors $\|Ap^f - q\|_{\langle\langle\rangle\rangle}$ and $\|Ap^f - \widehat{q}\|_{\langle\langle\rangle\rangle}$ satisfy the inequality (5.43). In this manner by evaluating either the functional $\mathcal{T}(p^f, q)$ or $\mathcal{T}^{++}(q)$ an upper bound can be found. Although not a strict argument, the additional residual squared term in the inequality (5.40) may act as a safety net if the inequality (5.43) is proving hard to satisfy.

A second method of obtaining an upper bound on the stationary value of the diffusion functional by time-stepping the Helmholtz functional to a steady state was outlined in section 4.3.5. Unfortunately a conflict in the numerical schemes required to guarantee convergence to a steady-state solution whilst iteratively satisfying the required constraint prevents this method from being viable. This conflict is demonstrated by considering the equation

$$A^*Ap = r, \tag{5.44}$$

notionally solved by taking the limit as $t \to \infty$ of the equation

$$p_t + A^*Ap = r, \tag{5.45}$$

where t may be a pseudo time as oppose to physical time variable. To simulate this limit a numerical scheme is constructed and in order to preserve the geometric property that ensures that (5.44) is the correct limit of (5.45) the Crank-Nicolson type discretisation, as used in chapter 4, is employed. The resulting numerical scheme

$$\left(A^*A + \frac{2}{\Delta t}\right) p^{t+\Delta t} = \frac{2}{\Delta t} p^t - T^*Tp^t + 2r \tag{5.46}$$

is solved using the variational principle for the Helmholtz equation. In particular the constraint imposed to obtain an upper bound on the Helmholtz functional is

$$A^*q^{t+\Delta t} + \frac{2}{\Delta t} p^{t+\Delta t} = \frac{2}{\Delta t} p^t - T^*Tp^t + 2r. \tag{5.47}$$

However, in the limit as both $p^{t+\Delta t} \to p_{SS}$ and $q^{t+\Delta t} \to q_{SS}$ this constraint tends to

$$A^*q_{SS} = 2r - T^*Tp_{SS}, \tag{5.48}$$

rather than the required constraint

$$A^* q_{SS} = r. \tag{5.49}$$

To obtain the constraint (5.49) the fully implicit scheme

$$\left(A^* A + \frac{1}{\Delta t} \right) p^{t+\Delta t} = \frac{1}{\Delta t} p^t + r \tag{5.50}$$

is required. The constraint to obtain the upper bound on the Helmholtz functional for the fully implicit scheme is then

$$A^* q^{t+\Delta t} + \frac{1}{\Delta t} p^{t+\Delta t} = \frac{1}{\Delta t} p^t + r. \tag{5.51}$$

which if $p^{t+\Delta t}$ and p^t tended to p_{SS} would satisfy (5.49). However a convergence proof similar to the geometric property described in section 4.1.6 is not available for the fully implicit scheme and therefore there is no guarantee that a steady-state solution will be found, and the constraint (5.49) satisfied. The inability of the method to satisfy the constraint through an iterative procedure is not too surprising as there is no guarantee that the solutions of the constraint will lie in the finite-dimensional approximation space. The existence of an element in the approximation space satisfying the constraint is even less likely in the current framework as the constraint has a unique solution.

The uniqueness properties associated with the constraint H^+ in effect renders the saddle-shaped topology of the functional degenerate. The constraint is excessively strong and instead of producing a set of functions q, over which the functional can be minimised the constraint generates the analytic solution $\widehat{q}$ solely. By employing the constraint H^{++} the uniqueness condition is weakened and the functional can again be minimised over a set of functions q. As with the lower bound, optimisation problems of this form reduce to problems in linear algebra when considered in a finite dimensional approximation space and therefore upper bounds can easily be computed in this manner. To guarantee a valid upper bound the inequality (5.43) must be satisfied, and to ensure that this occurs the following procedure is adopted.

5.1.3 The Procedure

The following procedure attempts to satisfy the inequality (5.43) by considering two solutions of differing accuracies defined on a fine and a coarse mesh. The procedure is then to:

1. maximize the quadratic functional $\mathcal{T}^-(p_h^f)$ by solving the associated stationary equations in a large finite dimensional subspace obtaining a relatively accurate approximation p_h^f to $\widehat{p}$, and an optimal lower bound μ^- for the subspace.

2. obtain a relatively inaccurate approximation q_h^c to q in a smaller approximation space such that the inequality

$$\|Ap_h^f - q_h^c\|_{\langle\langle\rangle\rangle} \geq \|Ap_h^f - \widehat{q}\|_{\langle\langle\rangle\rangle} \tag{5.52}$$

   holds. This can be achieved in one of two ways.

   (a) maximize the quadratic functional $\mathcal{T}^-(p_h^c)$ by solving the associated stationary equations in a small finite dimensional subspace to obtain a relatively coarse approximation p_h^c to $\widehat{p}$, and generate q_h^c as

$$q_h^c = Ap_h^c. \tag{5.53}$$

   (b) Construct a weakly equivalent function $q = Ap_h^c$ by solving the stationary equation

$$0 = \langle\langle A\delta p_h^c, Ap_h^c - Ap_h^f\rangle\rangle \tag{5.54}$$

   using the solution p_h^f obtained from maximising $\mathcal{T}^-(p_h^f)$. This method defines a projection from the fine to coarse grid denoted $\mathcal{P}_c^f$ where

$$q_h^c = A(\mathcal{P}_c^f p_h^f). \tag{5.55}$$

   (c) Construct q_h^c by applying the operator A to an interpolant of the solution p_h^f in the coarse approximation space,

$$q_h^c = A(\mathcal{I}_c^f p_h^f) \tag{5.56}$$

   where $\mathcal{I}_c^f$ is the interpolation operator between fine and coarse spaces defined by transferring the value of the nodes where the coarse and fine grids coincide.

To enable the three methods to be compared the projection and interpolation operators must honour the Dirichlet boundary conditions on the solution p. This requirement ensures that the three elements p_h^c, $\mathcal{P}_c^f p_h^f$ and $\mathcal{I}_c^f p_h^f$ belong to the same approximation space.

3. evaluate the functional $\mathcal{T}^{++}(q_h^c)$ or $\mathcal{T}(p_h^f, q_h^c)$ to obtain an upper bound.

Control of the relative magnitudes of the errors $\|Ap_h^f - q_h^c\|_{\langle\!\langle\rangle\!\rangle}$ and $\|Ap_h^f - \widehat{q}\|_{\langle\!\langle\rangle\!\rangle}$ is achieved through the choice and size of the spaces that the functions lie in. Consider two spaces S^f and S^c such that $S^c \subset S^f$ with $p^f \in S^f$ and $p^c \in S^c$. With a view to obtaining the lower bounds using continuous finite elements, the spaces S^f and S^c are the result of a fine and coarse discretisations of the domain respectively, and include the Dirichlet boundary conditions. We choose $q = Ap_h^c$ so that the inequality to satisfy is then

$$\|Ap_h^f - Ap_h^c\|_{\langle\!\langle\rangle\!\rangle} \geq \|Ap_h^f - A\widehat{p}\|_{\langle\!\langle\rangle\!\rangle}, \tag{5.57}$$

as $\widehat{q} = A\widehat{p}$. Provided that the spaces are nested in such a way that $S^c \subset S^f$ the inequality (5.57) can always be satisfied by making S^f sufficiently large. The exact fineness required for the discretisations depends on the rate of convergence of the solution in the 'energy norm', $\|Ap_h - A\widehat{p}\|_{\langle\!\langle\rangle\!\rangle}$, with respect to the grid size.

Convergence of the solutions in the energy norm is studied via Cea's Lemma (cf [8]). This states that for any $\xi \in S^h$

$$\begin{aligned}
\|Ap_h - A\widehat{p}\|^2_{\langle\!\langle\rangle\!\rangle} &= \langle\!\langle Ap_h - A\widehat{p}, Ap_h - A\widehat{p}\rangle\!\rangle, && (5.58)\\
&= \langle\!\langle Ap_h - A\widehat{p}, Ap_h - A\widehat{p}\rangle\!\rangle + \langle\!\langle A\xi, Ap_h - A\widehat{p}\rangle\!\rangle, && (5.59)\\
&= \langle\!\langle Ap_h - A\widehat{p}, A(p_h + \xi) - A\widehat{p}\rangle\!\rangle, && (5.60)\\
&\leq \|Ap_h - A\widehat{p}\|_{\langle\!\langle\rangle\!\rangle}\, \|A\, I_h(\widehat{p}) - A\widehat{p}\|_{\langle\!\langle\rangle\!\rangle}, && (5.61)
\end{aligned}$$

where $I_h(\widehat{p})$ is an interpolant of $\widehat{p}$ constructed by choosing the appropriate ξ. Hence the rate of convergence of the approximate solution in the energy norm is the same as the rate of convergence of the interpolation induced by the basis functions employed. Bounds on the rate of convergence can be found in [8] and [26], and will be of the form

$$\|Ap_h - A\widehat{p}\|_{\langle\!\langle\rangle\!\rangle} \leq Ch^s. \tag{5.62}$$

A conservative estimate of the order s of the scheme, is then required in order that an inequality of the form

$$\|Ap_h^f - A\widehat{p}\|_{\langle\!\langle\rangle\!\rangle} \leq \frac{1}{n^s}\|Ap_h^c - A\widehat{p}\|_{\langle\!\langle\rangle\!\rangle}, \tag{5.63}$$

can be asserted. Obtaining a practical value for the order of scheme is a non-trivial task as in general the analytic solution will not be available. However, it is envisaged that if a degree of agreement is found on the order of a particular finite element discretisation over multiple test problems then the use of the same value on further problems could be justified. The obvious case in which greater care would be required is to solutions containing singularities. Such solutions are not considered here. Defining S^f to be a n-fold mesh interleaving of S^c such that $n^s \geq 2$ then

$$\begin{aligned} \|Ap_h^f - A\widehat{p}\|_{\langle\langle\rangle\rangle} &\leq \frac{1}{n^s}\|Ap_h^c - A\widehat{p}\|_{\langle\langle\rangle\rangle}, & (5.64) \\ &\leq \frac{1}{2}\|Ap_h^c - A\widehat{p}\|_{\langle\langle\rangle\rangle}. & (5.65) \end{aligned}$$

From the triangle inequality

$$\begin{aligned} \|Ap_h^f - Ap_h^c\|_{\langle\langle\rangle\rangle} &\geq \|Ap_h^c - A\widehat{p}\|_{\langle\langle\rangle\rangle} - \|Ap_h^f - A\widehat{p}\|_{\langle\langle\rangle\rangle}, & (5.66) \\ &\geq \frac{1}{2}\|Ap_h^c - A\widehat{p}\|_{\langle\langle\rangle\rangle}, & (5.67) \end{aligned}$$

and therefore combining (5.65) and (5.67) the required inequality

$$\|Ap_h^f - Ap_h^c\|_{\langle\langle\rangle\rangle} \geq \|Ap_h^f - A\widehat{p}\|_{\langle\langle\rangle\rangle} \qquad (5.68)$$

is satisfied.

The mesh condition $n^s \geq 2$ is not very restrictive, since the requirement that $S^c \subset S^f$ restricts n to the integers greater than 1. Hence for a mesh doubling, $n = 2$, the finite element method is only required to have a first order rate of convergence in the energy norm with respect to the mesh size.

Having identified that the correct space for q, the function is constructed by one of the methods 2a,2b or 2c. Method 2a generates the solution p_h^c directly from the variational principles and therefore minimises the energy norm. Methods 2b and 2c construct the coarse solution directly from the fine solution and as a result they are likely to be non-optimal in the sense that

$$\begin{aligned} \|Ap_h^c - A\widehat{p}\|_{\langle\langle\rangle\rangle} &\leq \|A\,\mathcal{P}_c^f(p_h^f) - A\widehat{p}\|_{\langle\langle\rangle\rangle} \qquad \text{Method 2b}, & (5.69) \\ \|Ap_h^c - A\widehat{p}\|_{\langle\langle\rangle\rangle} &\leq \|A\,\mathcal{I}_c^f(p_h^f) - A\widehat{p}\|_{\langle\langle\rangle\rangle} \qquad \text{Method 2c}, & (5.70) \end{aligned}$$

but may offer computational savings. In particular an interpolation operator defined by transferring the value of the nodes from the fine to the coarse grid at the locations where the nodes coincide requires no further matrix inversions. Choosing to construct

q_c^h using a non-optimal projection or interpolation does not alter the procedure required to obtain an upper bound. For example choosing to use method 2c, the inequality to satisfy is

$$\|Ap_h^f - A\mathcal{I}_c^f p_h^f\|_{\langle\!\langle\rangle\!\rangle} \geq \|Ap_h^f - A\widehat{p}\|_{\langle\!\langle\rangle\!\rangle}. \tag{5.71}$$

From the rate of convergence of the finite element method a mesh interleaving is chosen such that

$$\begin{aligned} \|Ap_h^f - A\widehat{p}\|_{\langle\!\langle\rangle\!\rangle} &\leq \frac{1}{2}\|Ap_h^c - A\widehat{p}\|_{\langle\!\langle\rangle\!\rangle}, & (5.72)\\ &\leq \frac{1}{2}\|A\mathcal{I}_c^f p_h^f - A\widehat{p}\|_{\langle\!\langle\rangle\!\rangle}. & (5.73) \end{aligned}$$

Then, again using the triangle inequality,

$$\begin{aligned} \|Ap_h^f - A\mathcal{I}_c^f p_h^f\|_{\langle\!\langle\rangle\!\rangle} &\geq \|A\mathcal{I}_c^f p_h^f - A\widehat{p}\|_{\langle\!\langle\rangle\!\rangle} - \|Ap_h^f - A\widehat{p}\|_{\langle\!\langle\rangle\!\rangle}, & (5.74)\\ &\geq \frac{1}{2}\|A\mathcal{I}_c^f p_h^f - A\widehat{p}\|_{\langle\!\langle\rangle\!\rangle}, & (5.75) \end{aligned}$$

and combining (5.73) and (5.75) the inequality

$$\|Ap_h^f - Ap_h^c\|_{\langle\!\langle\rangle\!\rangle} \geq \|Ap_h^f - A\widehat{p}\|_{\langle\!\langle\rangle\!\rangle} \tag{5.76}$$

is satisfied. A similar argument holds with the projection operator.

5.2 An Example - Simple Advection

In chapter 4 the addition of the advective term led to a deterioration in the quality of the results obtained. Therefore to test the new method the advection equation in the (x, t) domain $\Omega = [0, 1] \times [0, T]$, periodic in the sense that $x(1, t) = x(0, t)$, is considered. Although the advection operator has additional structure including skew symmetry, this has intentionally not been exploited in order to maintain generality. The primal problem is then advection of the initial data ϕ_0, that is,

$$\begin{aligned} \widehat{\phi}_t + \widehat{\phi}_x &= 0 \qquad \text{in } \Omega, & (5.77)\\ \widehat{\phi} &= \phi_0 \qquad t = 0, & (5.78) \end{aligned}$$

corresponding to $A\widehat{\phi} = s$. The solution of the primal problem coincides with the stationary point of the functional

$$\mathcal{G}(\phi, \sigma) = \iint_\Omega \left\{\phi b - \left(\frac{\partial\phi}{\partial t} + \frac{\partial\phi}{\partial x}\right)\sigma\right\} dx\, dt - \int_0^1 [\sigma(\phi - \phi_0)]_{t=0}\, dx + \int_0^1 [\phi\sigma_T]_{t=T}\, dx, \tag{5.79}$$

for the fist variation of the functional $\delta\mathcal{G}(\phi,\sigma)$ is

$$\begin{aligned}\mathcal{G}(\phi,\sigma) &= \iint_\Omega \left\{\delta\phi\left(b+\frac{\partial\sigma}{\partial t}+\frac{\partial\sigma}{\partial x}\right)-\left(\frac{\partial\phi}{\partial t}+\frac{\partial\phi}{\partial x}\right)\delta\sigma\right\}dx\,dt \\ &\quad -\int_0^1 [\delta\sigma(\phi-\phi_0)]_{t=0}\,dx-\int_0^1 [\delta\phi(\sigma-\sigma_T)]_{t=T}\,dx. \end{aligned} \tag{5.80}$$

In addition the stationary point of the functional defines the dual problem to be

$$-\widehat{\sigma}_t-\widehat{\sigma}_x = b \qquad \text{in } \Omega, \tag{5.81}$$

$$\widehat{\sigma} = \sigma_T \qquad t=T, \tag{5.82}$$

corresponding to $A^*\widehat{\sigma}=t$. Comparing the primal and dual problems confirms that the advection equation is non-self-adjoint, and characteristically the dual problem runs in reverse from the final condition σ_T. The stationary value of the functional is, as expected, the solution of the primal problem weighted by the forcing terms of the dual problem

$$\mathcal{G}(\widehat{\phi},\widehat{\sigma})=\iint_\Omega \widehat{\phi}b\;dx\,dt+\int_0^1\left[\widehat{\phi}\sigma_T\right]_{t=T}dx. \tag{5.83}$$

The physical significance of the stationary value can be chosen by the appropriate selection of the forcing functions b and σ_T.

Following the framework constructed in the general notation the operator is modified so as to obtain a self-adjoint problem set, the solution of which coincides with the stationary point of the functional $\mathcal{J}(u,v)$. Initially the additional constraints $u=\phi_0$ on $t=0$ and $v=0$ on $t=0$, are placed on the functions u and v. In due course these constraints will be included in the functional using the intermediate functions q_1 and q_2 as Lagrangian multipliers. The functional $\mathcal{J}(u,v)$ is then

$$\mathcal{J}(u,v)=\iint_\Omega\left\{ub-\left(\frac{\partial u}{\partial t}+\frac{\partial u}{\partial x}\right)\left(\frac{\partial u}{\partial t}+\frac{\partial u}{\partial x}\right)\right\}dx\,dt+\int_0^1[uv_T]_{t=T}\,dx, \tag{5.84}$$

and the primal problem is defined to be

$$\left.\begin{aligned}-\widehat{u}_{tt}-2\widehat{u}_{xt}-\widehat{u}_{xx} &= 0 \quad \text{in } \Omega,\\ \widehat{u} &= \phi_0 \quad t=0,\\ \widehat{u}_t+\widehat{u}_x &= 0 \quad t=T,\end{aligned}\right\}\text{corresponding to } A^*(A\widehat{u}-s)=0. \tag{5.85}$$

The natural boundary condition arising at $t=T$ ensures that the boundary conditions on $\widehat{u}$ coincide with the solution $\widehat{\phi}$ and hence enforces equality of the solutions of the primal problems, $\widehat{u} = \widehat{\phi}$. The dual problem is defined to be

$$\left.\begin{array}{rcll} -\widehat{v}_{tt} - 2\widehat{v}_{xt} - \widehat{v}_{xx} & = & b & \text{in } \Omega, \\ \widehat{v} & = & 0 & t = 0, \\ \widehat{v}_t + \widehat{v}_x & = & \sigma_T & t = T, \end{array}\right\} \text{corresponding to } A^*A\widehat{v} = b, \tag{5.86}$$

with the original Dirichlet condition $\sigma = \sigma_T$ at $t = T$ being modified to a Neumann type condition $\widehat{v}_t + \widehat{v}_x = \sigma_T$ at $t = T$. As required the stationary value $\mathcal{J}(\widehat{u}, \widehat{v})_s$ is identical to $\mathcal{G}(\widehat{\phi}, \widehat{\sigma})_s$ through the relationship $\widehat{u} = \widehat{\phi}$,

$$\mathcal{J}(\widehat{u}, \widehat{v})_s = \iint_\Omega \widehat{u}b \; dx\, dt + \int_0^1 [\widehat{u}\sigma_T]_{t=T}\, dx. \tag{5.87}$$

To obtain the pair of self-dual problems the twinning transformations $u, v \rightarrow p_1, p_2$,

$$\begin{aligned} u &= \frac{1}{2}(p_1 + p_2), && (5.88)\\ v &= \frac{1}{2}(p_1 - p_2), && (5.89) \end{aligned}$$

are applied to the functions to obtain

$$\begin{aligned} \mathcal{J}(p_1, p_2) = & \iint_\Omega \left\{ \frac{1}{2}(p_1 + p_2)b - \frac{1}{4}\left(\frac{\partial p_1}{\partial t} + \frac{\partial p_1}{\partial x}\right)\left(\frac{\partial p_1}{\partial t} + \frac{\partial p_1}{\partial x}\right) \right. \\ & \left. + \frac{1}{4}\left(\frac{\partial p_2}{\partial t} + \frac{\partial p_2}{\partial x}\right)\left(\frac{\partial p_2}{\partial t} + \frac{\partial p_2}{\partial x}\right) \right\} dx\, dt + \frac{1}{2}\int_0^1 [(p_1 + p_2)\sigma_T]_{t=T}\, dx. \end{aligned} \tag{5.90}$$

The intermediate variables q_1 and q_2 are now introduced as Lagrangian multipliers for the Dirichlet conditions on p_1 and p_2 respectively. The complete functional $\mathcal{J}(p_1, q_1, p_2, q_2)$ is then

$$\begin{aligned} \mathcal{J}(p_1, q_1, p_2, q_2) = & \iint_\Omega \left\{ \frac{1}{2}(p_1 + p_2)b - \frac{1}{2}\left(\frac{\partial p_1}{\partial t} + \frac{\partial p_1}{\partial x}\right) q_1 + \frac{1}{4}q_1^2 - \frac{1}{4}q_2^2 \right. \\ & \left. + \frac{1}{2}\left(\frac{\partial p_2}{\partial t} + \frac{\partial p_2}{\partial x}\right) q_2 \right\} dx\, dt + \frac{1}{2}\int_0^1 [(p_1 + p_2)\sigma_T]_{t=T}\, dx \\ & - \frac{1}{2}\int_0^1 [(p_1 - \phi_0)q_1 - (p_2 - \phi_0)q_2)]_{t=0}\, dx, && (5.91) \\ = & \frac{1}{2}\mathcal{T}_1(p_1, q_1) - \frac{1}{2}\mathcal{T}_2(p_2, q_2). && (5.92) \end{aligned}$$

From (5.92) the pair of functionals $\mathcal{T}_i(p_i, q_i)$ can be identified as

$$\begin{aligned} \mathcal{T}_i(p_i, q_i) = & \iint_\Omega \left\{ p_i e_i - \left(\frac{\partial p_i}{\partial t} + \frac{\partial p_i}{\partial x}\right) q_i + \frac{1}{2}q_i^2 \right\} dx\, dt - \int_0^1 [(p_i - f_i)q_i]_{t=0}\, dx \\ & + \int_0^1 [p_i g_i]_{t=T}\, dx, \end{aligned} \tag{5.93}$$

with the forcing terms

$$\left.\begin{aligned} e_1 &= b, \\ f_1 &= \phi_0, \\ g_1 &= \sigma_T, \end{aligned}\right\}\text{corresponding to } r_1, \tag{5.94}$$

and

$$\left.\begin{aligned} e_2 &= -b, \\ f_2 &= \phi_0, \\ g_2 &= -\sigma_T, \end{aligned}\right\}\text{corresponding to } r_2. \tag{5.95}$$

The stationary value of the functional is obtained by substituting the analytic solutions $p_i = \widehat{p}_i$ and $q_i = \widehat{q}_i$ and is found to be

$$\begin{aligned} \mathcal{T}_i(\widehat{p}_i, \widehat{q}_i) &= \iint_\Omega \left\{ \widehat{p}_i e_i - \frac{1}{2}\left(\frac{\partial \widehat{p}_i}{\partial t} + \frac{\partial \widehat{p}_i}{\partial x} \right) \widehat{q}_i \right\} dx\,dt + \int_0^1 \left[\widehat{p}_i g_i \right]_{t=T} dx, && (5.96) \\ &= \frac{1}{2} \iint_\Omega \widehat{p}_i e_i dx\,dt + \frac{1}{2} \int_0^1 \left[\widehat{p}_i g_i \right]_{t=T} dx + \frac{1}{2} \int_0^1 \left[f_i \widehat{q}_i \right]_{t=0} dx. && (5.97) \end{aligned}$$

Having isolated the pair of self-adjoint and self-dual problems the saddle shaped topology of the governing functionals can then be demonstrated. Again, dropping the index for clarity, and considering the general functional $\mathcal{T}(p, q)$ the constraint corresponding to the hyper-line H^-,

$$p_t + p_x = q \qquad \text{in } \Omega, \tag{5.98}$$

$$p = f \qquad t = 0, \tag{5.99}$$

can be applied to the functional by direct substitution. On substitution of the constraints (5.98) and (5.99) the functional

$$\mathcal{T}^-(p) = \iint_\Omega \left\{ pe - \frac{1}{2}(p_t + p_x)(p_t + p_x) \right\} dx\,dt + \int_0^1 \left[pg \right]_{t=T} dx, \tag{5.100}$$

is obtained. It satisfies the maximum principle since

$$\begin{aligned} \mathcal{T}^-(p) - \mathcal{T}^-(\widehat{p}) &= \iint_\Omega \Big\{ (p - \widehat{p})(-\widehat{p}_{tt} - 2\widehat{p}_{xt} - \widehat{p}_{xx}) - \frac{1}{2}(p_t + p_x)(p_t + p_x) \\ &\qquad + \frac{1}{2}(\widehat{p}_t + \widehat{p}_x)(\widehat{p}_t + \widehat{p}_x) \Big\} dx\,dt + \int \left[(p - \widehat{p})(p_t + p_x) \right]_{t=T} dx, \\ &= \iint_\Omega \Big\{ (p_t + p_x)(\widehat{p}_t + \widehat{p}_x) - \frac{1}{2}(p_t + p_x)(p_t + p_x) \\ &\qquad - \frac{1}{2}(\widehat{p}_t + \widehat{p}_x)(\widehat{p}_t + \widehat{p}_x) \Big\} dx\,dt + \int \left[(p - \widehat{p})(p_t + p_x) \right]_{t=0} dx, \\ &= -\frac{1}{2} \iint_\Omega \left(\, (p_t + p_x) - (\widehat{p}_t + \widehat{p}_x) \, \right)^2 dx\,dt, \\ &\leq 0. \end{aligned} \tag{5.101}$$

Alternatively, applying the constraints corresponding to the hyper-line H^+,

$$-q_t - q_x = e \qquad \text{in } \Omega, \tag{5.102}$$

$$q = g \qquad t = T, \tag{5.103}$$

the functional

$$\mathcal{T}^+(q) = \iint_\Omega \frac{1}{2} q^2 dx\, dt + \int [fq]_{t=0}\, dx, \tag{5.104}$$

is obtained. Conversely, this functional satisfies the minimum principle since

$$\begin{aligned} \mathcal{T}^+(q) - \mathcal{T}^+(\widehat{q}) &= \iint_\Omega \left\{ \frac{1}{2}q^2 - \frac{1}{2}\widehat{q}^2 + \widehat{p}(q_t + q_x - \widehat{q}_t - \widehat{q}_x) \right\} dx\, dt + \int [\widehat{p}(q-\widehat{q})]_{t=0}\, dx, \\ &= \iint_\Omega \left\{ \frac{1}{2}q^2 - \frac{1}{2}\widehat{q}^2 - (\widehat{p}_t + \widehat{p}_x)(q - \widehat{q}) \right\} dx\, dt + \int [\widehat{p}(q-\widehat{q})]_{t=T}\, dx, \\ &= \frac{1}{2} \iint_\Omega (q - \widehat{q})^2 dx\, dt, \\ &\geq 0. \end{aligned} \tag{5.105}$$

As anticipated, applying the constraint (5.102) strongly is as difficult as solving the original primal problem and an upper bound is calculated by applying the procedure outlined in section 5.1.3.

5.2.1 Results

The quantity of interest is chosen as the weighted integral of the solution at the final time using the functions $b = 0$, and ϕ_0 and σ_T constructed from quadratic sections. The functions ϕ_0 and σ_T are plotted in figure 5.1 The analytic solutions to the pair of twinned problems are then

$$\widehat{p}_1 = \phi_0(x-t) + t\sigma(x-t+1), \tag{5.106}$$

$$\widehat{p}_2 = \phi_0(x-t) - t\sigma(x-t+1). \tag{5.107}$$

The knowledge of the analytic solutions of the twinned problems enables the required inequality expressed as either

$$\frac{1}{2}\|Ap_h^c - A\widehat{p}\|_{\langle\!\langle\rangle\!\rangle} \geq \|Ap_h^f - A\widehat{p}\|_{\langle\!\langle\rangle\!\rangle}, \tag{5.108}$$

or

$$\|Ap_h^f - Ap_h^c\|_{\langle\!\langle\rangle\!\rangle} \geq \|Ap_h^f - A\widehat{p}\|_{\langle\!\langle\rangle\!\rangle}, \tag{5.109}$$

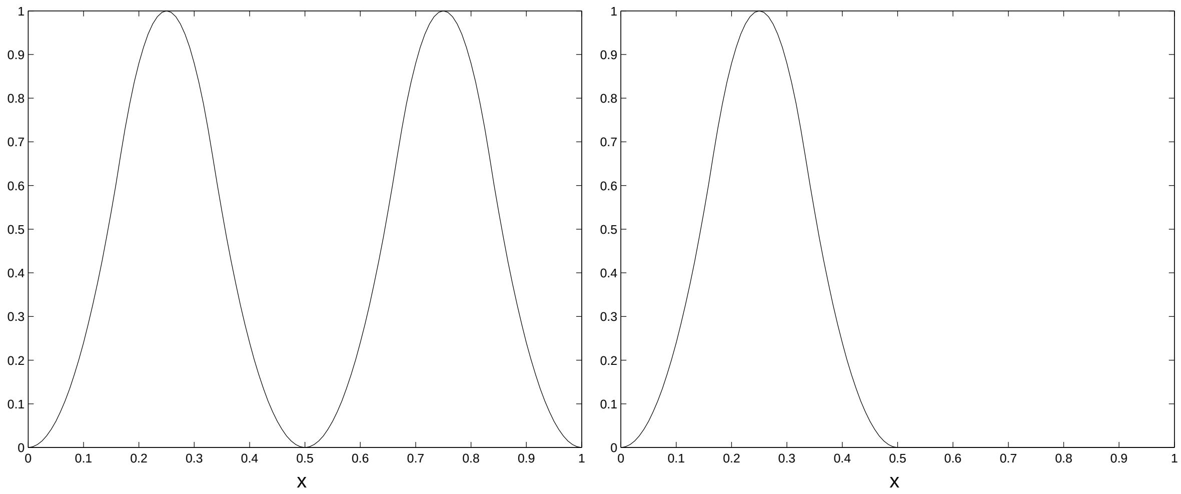

Figure 5.1: The boundary conditions ϕ_0 and σ_T

to be verified for the method employed. The twinned problems where discretised using piecewise continuous quadratic finite elements. In accordance with the procedure outlined in section 5.1.3 the lower bound was calculated by maximising the functional $\mathcal{T}^-(p_h^f)$. From the lower bound the rate of convergence of the method can be calculated as the error in the energy norm is directly related to the error in the lower bound,

$$\begin{aligned} \|Ap_h - A\widehat{p}\|^2_{\langle\!\langle\rangle\!\rangle} &= \langle\!\langle Ap_h - A\widehat{p}, Ap_h - A\widehat{p}\rangle\!\rangle & (5.110)\\ &= 2\mathcal{T}^-(p_h) - 2\mathcal{T}^-(\widehat{p}) & (5.111)\end{aligned}$$

from (5.28). The convergence in the energy norm of the solutions of the twinned problems with $T = 1$ is shown in figure 5.2. From figure 5.2 a conservative estimate for the order of the scheme is one and therefore a simple mesh doubling between the coarse and fine grids is sufficient to obtain an upper bound on the stationary values of the functional. The upper bound was found by constructing q_h^c as

$$q_h^c = A\mathcal{I}_c^f p_h^f, \tag{5.112}$$

where the interpolation operator is defined by transferring the value of the nodes at the points where the meshes coincide. The upper bounds for the pair of twinned problems were evaluated as $\mathcal{T}(p_h^f, q_h^c)$.

As a result of the chosen weighting function the quantity of interest fluctuates as the final time is varied, and is sensitive to errors in the wave speed and diffusion of the

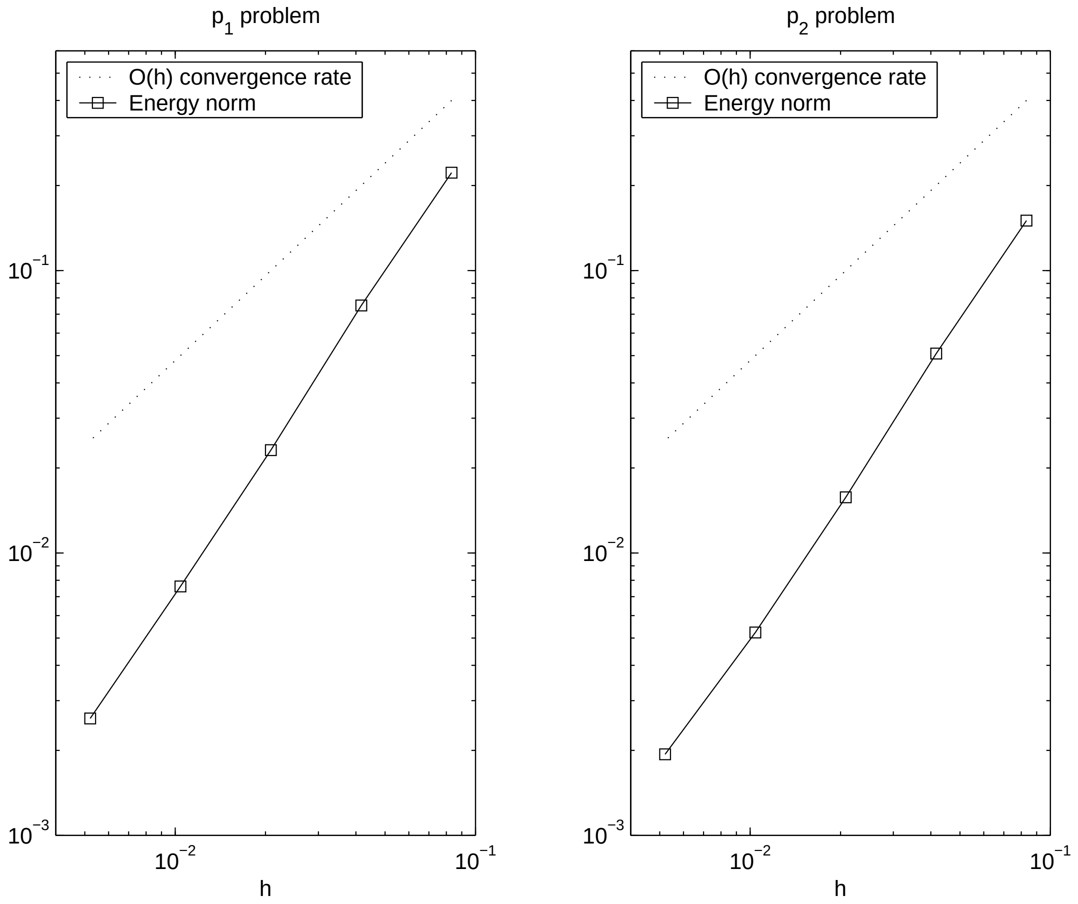

Figure 5.2: Convergence of the finite element solutions in the energy norm, $\|Ap_h - A\widehat{p}\|$, with respect to element size, h, $T = 1$

numerical solution. The analytic value of the quantity of interest can be calculated from the analytic solution of the primal problem

$$\widehat{\phi}(x,t) = \phi_0(x-t), \tag{5.113}$$

and therefore

$$\Theta(\widehat{\phi}) = \int_0^1 \left[\phi_0(x-T)\sigma_T(x)\right]_{t=T} \, dx. \tag{5.114}$$

The numerical solutions of the pair of problems governing p_1 and p_2 with $T = 1$ are plotted in figure 5.3. The solutions were obtained using 24 × 24 elements for the fine grid. The approximate solutions of the primal and dual problems, (5.85) and (5.86), are obtained from the solutions p_1 and p_2 by inverting the twinning transformation, taking the sum and difference respectively. These solutions are shown in figure 5.4. From this it is noted that while the primal solution u is an approximation to the advection of the initial data, the dual solution v is not simply advection of the function σ_T backwards in time. The change in the characteristic of the dual solution is a result of introducing self-adjointness into the governing equations. In doing so the quantity of interest is evaluated as the boundary integral of solution of the primal problem, weighted by the Neumann boundary condition of the dual solution. The Neumann type boundary condition at $t = T$ is responsible for the growth in the solution and the deviation of the solution from $\widehat{\sigma}$.

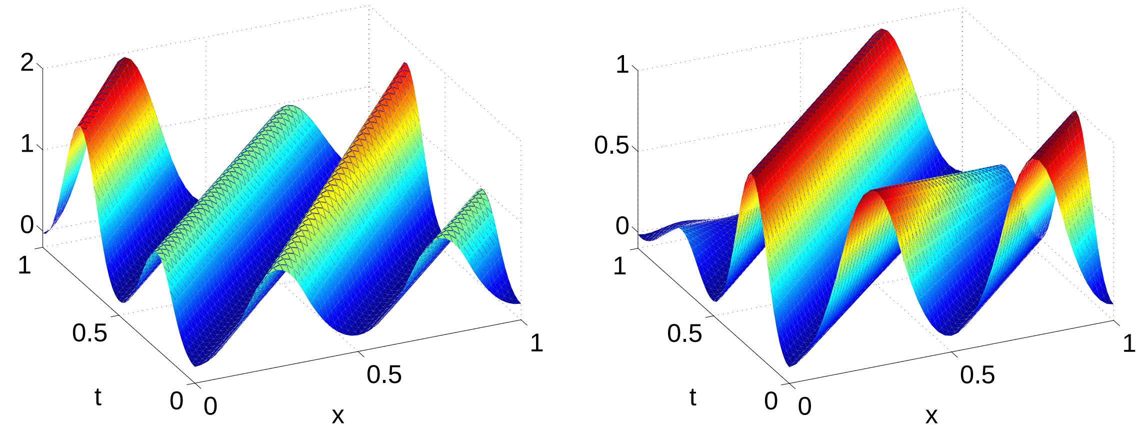

Figure 5.3: Solutions of the p_1 and p_2 problems

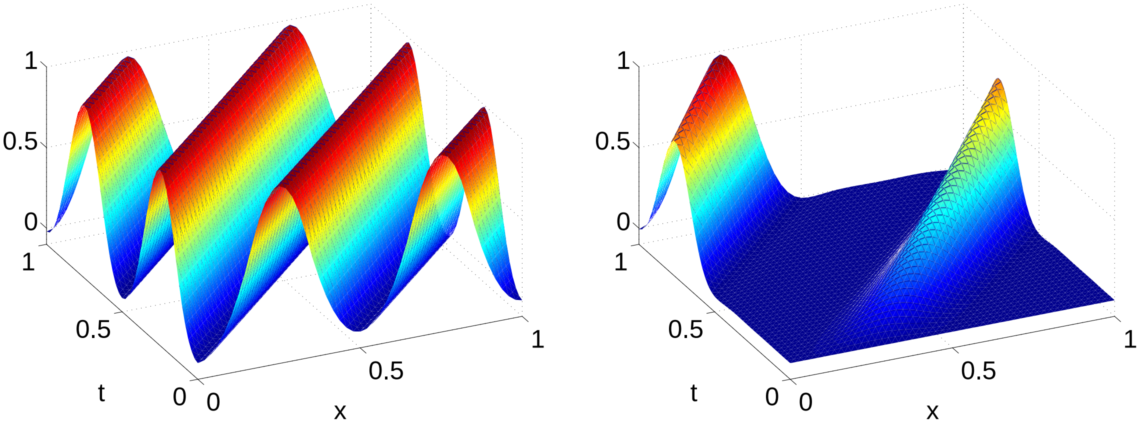

Figure 5.4: Solutions u and v of the primal and dual problems

Figure 5.5 shows the upper and lower bounds obtained on the $\Theta(\widehat{\phi})$ for varying end times. In solving over domains in which $T \neq 1$ it is assumed that the mesh doubling employed is sufficient to ensure valid upper and lower bounds. Although the upper bounds μ_i^+ are potentially weaker than the lower bounds the construction of the bounds on the quantity of interest as

$$\frac{1}{2}(\mu_1^- - \mu_2^+) \leq \Theta(\widehat{\phi}) \leq \frac{1}{2}(\mu_1^+ - \mu_2^-) \tag{5.115}$$

distributes this imbalance evenly. In figure 5.5 each pair of bounds is generated by solving the pair of twinned problems in the domain $(0 \leq x \leq 1, 0 \leq t \leq T)$ and in order that the divergence of the bounds with time can be investigated, the ratio of elements in the time direction to the end time was kept constant. From figure 5.5 a gradual drift in the accuracy of the bounds is experienced as the end time is increased. This drift in is due to the gradual diffusion of the advected data and as the quantity of interest is a particular weighted integral of the solution the shape of the solution at the final time is important. In addition, the finite element method is unlikely to be conservative and therefore the solution is likely to experience gradual decay. In order to try and tighten the bounds obtained a higher grid density can be employed. Figure 5.5 indicates a significant improvement in the accuracy of the simulations when the grid density is doubled.

A characteristic of this method is that the complete (x, t) domain is required to be discretised which for long time simulations yields costly simulations. In order to try and maximise the efficiency of long-time simulations grid adaption techniques can be implemented.

5.3 Grid Adaption

The advantages of defining a local error indicator based on the quantity of interest has been discussed in section 1.3.1. One method of obtaining an error indicator based on the quantity of interest is by localising the contribution each element makes to the difference between the bounds, a quantity directly related to the accuracy to which the quantity of interest has been resolved. In [45] Suli and Houston equidistribute the contribution to the error representation formula (1.6) made by each element. In a similar manner a grid adaption method can be formulated within the present framework to equidistribute the contribution to the difference between the upper and lower bound made by each element. In this way it is possible to ensure that the error in the quantity of interest,

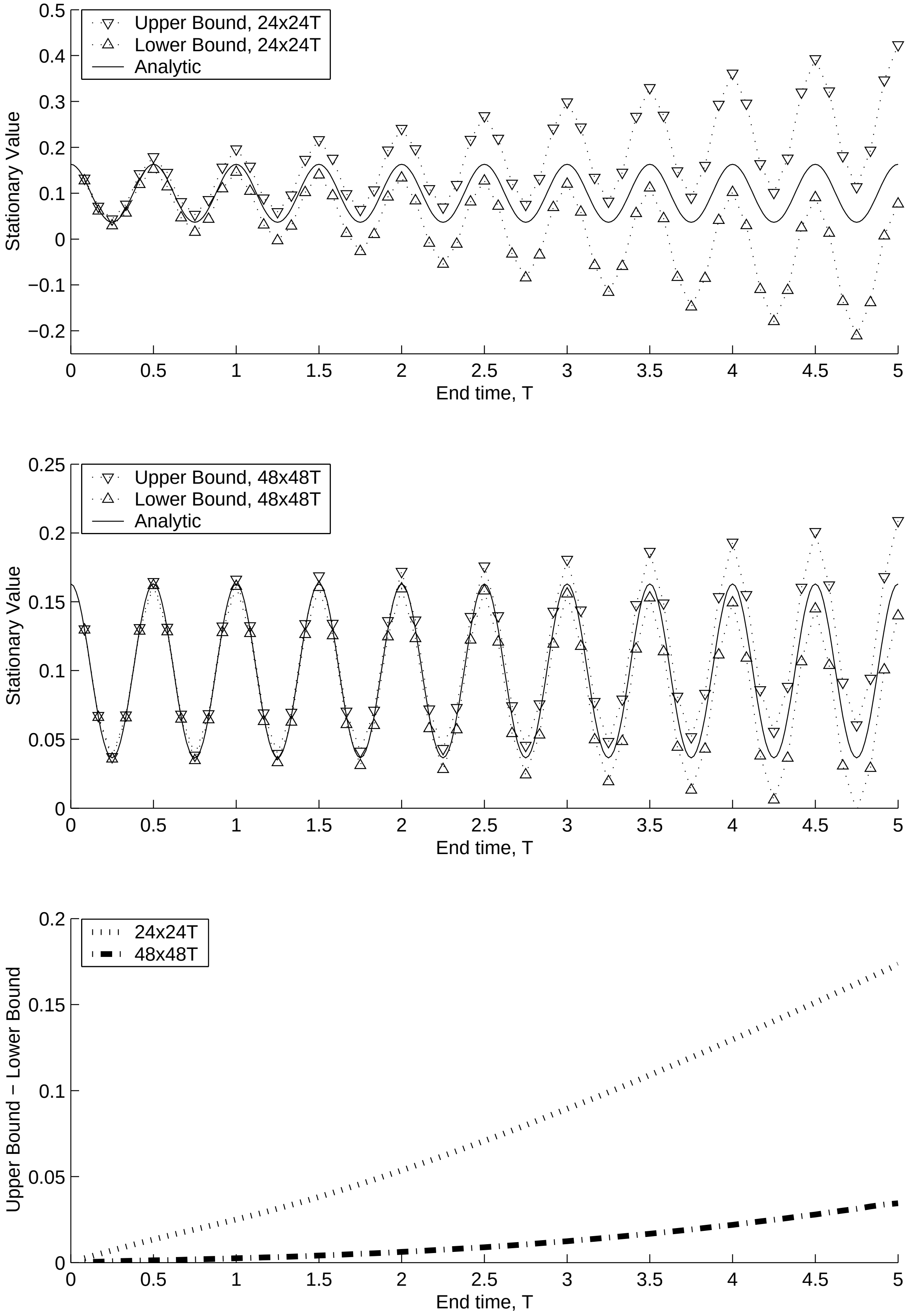

Figure 5.5: Bounds on $\Theta(\widehat{\phi})$

essentially the difference in the bounds, is below a user defined tolerance. Defining our approximation Θ_h to be the average of the upper and lower bounds the following bounds on the quantity of interest can be formed

$$\begin{aligned} \left|\Theta_h - \Theta(\widehat{\phi})\right| &\leq \frac{1}{4}\left|\mu_1^+ - \mu_2^- - \mu_1^- + \mu_2^+\right|, && (5.116)\\ &= \frac{1}{4}\left|\sum_{el=1}^{N}\left[\mu_1^+ - \mu_2^- - \mu_1^- + \mu_2^+\right]_{el}\right|, && (5.117)\\ &\leq \frac{1}{4}\sum_{el=1}^{N}\left|\left[\mu_1^+ - \mu_2^- - \mu_1^- + \mu_2^+\right]_{el}\right|. && (5.118)\end{aligned}$$

where $\left[\mu_1^+ - \mu_2^- - \mu_1^- + \mu_2^+\right]_{el}$ is the contribution from the element el to the error in the quantity of interest, and N is the total number of elements in the discretisation. The desired accuracy in the quantity of interest is found by enforcing

$$\left|\left[\mu_1^+ - \mu_2^- - \mu_1^- + \mu_2^+\right]_{el}\right| \leq \frac{4}{N}tol \qquad (5.119)$$

and iterating to obtain

$$\left|\Theta_h - \Theta(\widehat{\phi})\right| \leq tol. \qquad (5.120)$$

The grid adaption method described will now be applied to the advection equation considered in the previous example. For the purposes of this example the end time $T = 1$ is chosen.

5.3.1 Grid Adaption Applied to the Advection Example

Grid adaption can be naturally applied to the advection example as the contribution from each element to the error in the quantity of interest is computable. The local contributions are found from the bounds μ_i^- and μ_i^+ which are expressed as integrals over the whole computation domain. For example

$$\begin{aligned} \mu^- &= \mathcal{T}^-(p_h^f),\\ &= \iint_{\Omega}\left\{p_h^f e_1 - \frac{1}{2}((p_h^f)_t + (p_h^f)_x)((p_h^f)_t + (p_h^f)_x)\right\}dx\,dt + \int_0^1 \left[p_h^f g\right]_{t=T} dx,\\ &= \sum_{el=1}^{N}\iint_{el}\left\{p_h^f e_1 - \frac{1}{2}((p_h^f)_t + (p_h^f)_x)((p_h^f)_t + (p_h^f)_x)\right\}dx\,dt\\ &\quad + \sum_{el=1}^{N}\int_{el}\left[p_h^f g\right]_{t=T} dx, && (5.121)\end{aligned}$$

where it is understood that the second integral can only have non-zero contributions from elements with a boundary segment along the line $t = T$. Therefore

$$\mu_{el}^{-} = \iint_{el} \left\{ p_h^f e_1 - \frac{1}{2}((p_h^f)_t + (p_h^f)_x)((p_h^f)_t + (p_h^f)_x) \right\} dx\, dt + \int_{el} \left[p_h^f g \right]_{t=T} dx, \quad (5.122)$$

and the local contributions to the upper bound can be found in a similar fashion. The desired accuracy in the quantity of interest is again found by enforcing

$$\left| \left[\mu_1^+ - \mu_2^- - \mu_1^- + \mu_2^+ \right]_{el} \right| \leq \frac{4}{N} tol \quad (5.123)$$

and iterating to obtain

$$\left| \Theta_h - \Theta(\widehat{\phi}) \right| \leq tol. \quad (5.124)$$

In constructing a grid refinement method based upon the upper and lower bounds we are assuming that the procedure outlined in section 5.1.3 extends to unstructured meshes. The validity of this assumption will be investigated with numerical values obtained from the refinement procedure. The upper bound will again be found by using the interpolation operator to define the coarse function p_h^c on a grid half the density of that on which p_h^f is defined. The refinement algorithm will be to simply double the mesh density in any element failing to satisfy the refinement criterion 5.123. In addition elements will always be grouped in sets of four to enable the coarse mesh, with half the mesh density, to be easily defined. Any resulting hanging nodes are removed by interpolating the value of the solution from the neighbouring nodes, using the same basis functions that are being used to represent the solution. For optimal bounds on $\Theta(\widehat{\phi})$ separate grids should be implemented for the two solutions p_1 and p_2, however, the grid requirements of the two problems appear very similar and the two problems were solved on the same grid for reasons of economy. The fine grids in figure 5.6, on which p_h^f is defined, were produced using the refinement procedure with the tolerance set at 0.045. The refinement of the grids displays the structure expected and observed by Suli and Houston [45]. Namely, the procedure refines the grid along the characteristics transporting components of the solution pertinent to the accuracy of the quantity of interest. In this example the quantity of interest is effectively an integral evaluated in the left half of the domain at the final time. The characteristics of the solution are the straight lines

$$x = t + \text{constant} \quad (5.125)$$

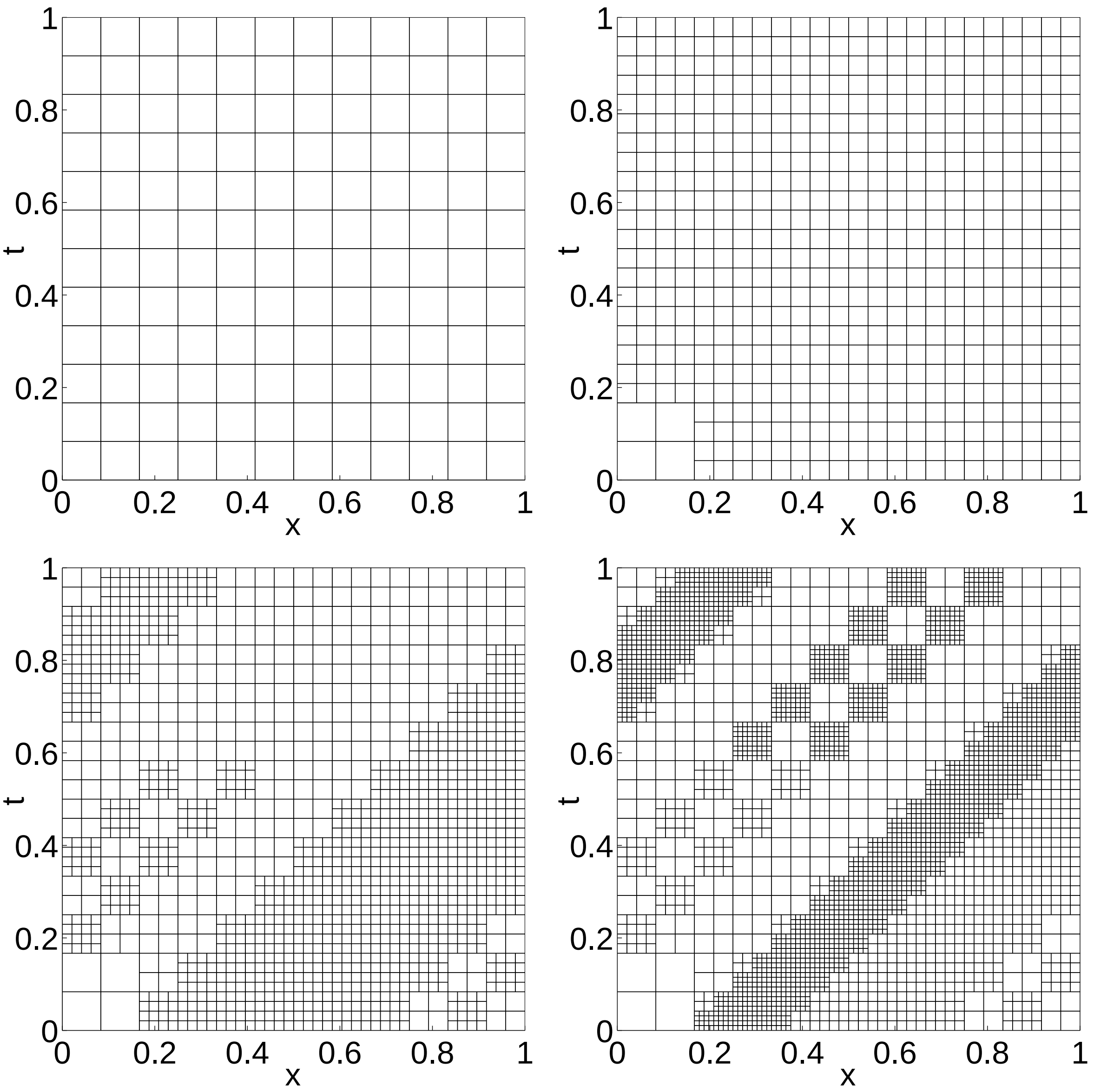

Figure 5.6: The fine computational grids, refinement levels 1 - 4

and from figure 5.6 it is observed that the areas of the grid that are most significantly refined correspond to the characteristics that pass through the left half of the domain at the final time. In particular the central strip of these characteristics are refined to the highest degree and this portion relates to the greater weighting in the quantity of interest stimulated by the function σ_T around the region $x = 0.25$, $(T = 1)$. In theory the solution can be arbitrarily inaccurate away from the key characteristics and therefore the grid can remain relatively coarse. However, a certain degree of resolution is required across the solution in order that large errors incurred in irrelevant sections of the solution do not pollute areas in which a greater degree of accuracy is necessary.

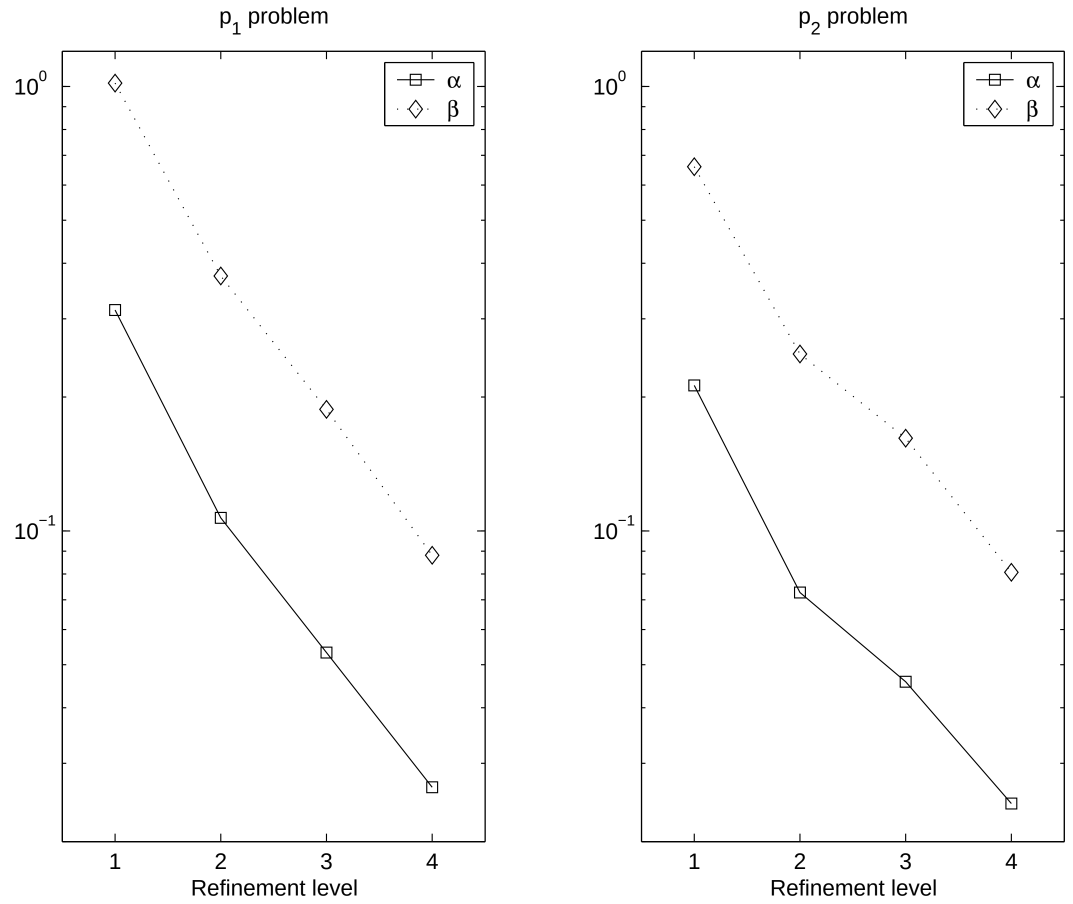

Figure 5.7: Convergence of the norms $\alpha = \|Ap_h^f - A\widehat{p}\|_{\langle\langle\rangle\rangle}$ and $\beta = \|Ap_h^f - Ap_h^c\|_{\langle\langle\rangle\rangle}$ for the pair of twined problems with respect to refinement level

The balance between the grid requirements of the various parts of domain is quite subtle and the basic grid refinement procedure employed is probably far from optimal. For example the method did not include a grid coarsening step. However, it was successful in driving the error in the quantity of interest down and the numerical values obtained, corresponding to the grids in figure 5.6, are show in table 5.1 where,

$$|\Theta_h - \Theta(\widehat{\phi})| \leq \epsilon_1 = \frac{1}{4}\left|\mu_1^+ - \mu_2^- - \mu_1^- + \mu_2^+\right|, \tag{5.126}$$

$$|\Theta_h - \Theta(\widehat{\phi})| \leq \epsilon_1 \leq \epsilon_2 = \frac{1}{4}\sum_{el=1}^{N}\left|\left[\mu_1^+ - \mu_2^- - \mu_1^- + \mu_2^+\right]_{el}\right|. \tag{5.127}$$

From table 5.1 it can be seen that the bound on the error in quantity of interest is much less than the bound ϵ_2. This difference is due to the evaluation of ϵ_2 preventing the cancellation of many local error terms. However both bounds can be calculated during the refinement procedure and therefore although the method maybe driven by forcing $\epsilon_2 \leq tol$, the user has the liberty of stopping the refinement procedure when $\epsilon_1 \leq tol$.

Level	No Nodes	ϵ_1	ϵ_2
1	600	0.18382	0.23810
2	2296	0.02534	0.06753
3	5424	0.00766	0.04888
4	13200	0.00179	0.04262

Table 5.1: Error and error indicators corresponding to grid refinement levels 1 - 4

Again, a tacit assumption is made that the difference in the errors between the coarse, fine and analytic solutions of the pair of twinned problems satisfy the required inequality

$$\|Ap_h^f - Ap_h^c\|_{\langle\langle\rangle\rangle} \geq \|Ap_h^f - A\widehat{p}\|_{\langle\langle\rangle\rangle}. \tag{5.128}$$

Inspection of the convergence data of the numerical solutions, figure (5.7), confirms this assumption at all refinement levels.

Using the upper and lower bounds as an error indicator compares favourably with methods based on the error representation formula (1.6). An error indicator is required to reliably represent the magnitude of the local contribution to the error in the quantity sought. However, the inability to evaluate the error representation formula due to the presence of the analytic solution results in the set of inequalities (5.116)-(5.118) existing only as approximations to the error in the quantity of interest, although the method performs effectively over the catalogue of examples implemented by Suli and Houston [45]. In contrast the ability to evaluate the error in the quantity of interest using the upper and lower bounds produces strong results in the form of the inequalities (5.116)-(5.118) and which are confirmed by the results displayed in table 5.1.

5.4 Conclusions and Extensions

The procedure outlined provides an effective means of bounding quantities of interest governed by non-self-adjoint operators. The ability to obtain these bounds hinges on satisfying the inequality (5.43). The example considered illustrated that satisfying the inequality (5.43) is feasible in practice although the method relies on obtaining a conservative estimate of the convergence rate of the finite element method used. In practice determining the convergence rate would require a preliminary convergence study and in general add an additional cost to the method. If a mesh adaption routine is to be used the convergence study would also have to consider unstructured grids. Further research is required in this area to explore fully the limitations that this places on the method. However, in contrast to the dual extremum principles constructed by Collins for the heat equation [12] the method described in this chapter is implementable and the required constraints to obtain the upper and lower bounds can be satisfied in practice.

A description of this method for obtaining bounds on a quantity of interest governed by a non-self-adjoint operator has been submitted to *Computers and Fluids* as part of the proceedings of the AMIF 2002 conference held in Lisbon.

Having generated the method using regular quadrilateral elements the method could also be extended to unstructured triangular meshes. An unstructured triangular mesh has greater flexibility and node movement, as well as mesh refinement could be considered. However, unstructured meshes also introduce additional complexities and, in particular, obtaining the coarse grid with the required accuracy properties could be challenging. Obtaining progressively coarser grids is a component of the multigrid solving routine and this body of literature would provide a practical starting point.

The ability to obtain bounds on the quantity of interest to a given tolerance is a useful attribute of the method, although the cost of the refinement procedure will depend on the efficiency of the algorithm. The technique of refining the approximation space depending on the magnitude of the local contribution to the error in the quantity of interest can be employed to any problem where the upper and lower bounds are evaluated as integrals over the computational domain. Therefore, a similar method could also be defined to refine the grids used to obtain numerical solutions in chapter 3.

5.4.1 Generating Numerical Methods

In the example considered standard finite elements were used to obtain the numerical solutions. However the numerical solutions do not retain any of the properties of the solution of the advection equation, namely conservation of the advected quantity. Provided that the requirements to generate upper and lower bounds are satisfied other numerical methods can also be considered. Ideally a numerical method could be found for which the conservation of the primal solution u could also be shown. A possible method to generate conservative numerical methods may stem from choosing the quantity of interest to be the integral of the conserved quantity and then applying discretisations to the primal and dual problems that preserve this quantity. Further research into this area is required before the feasibility of this approach can be evaluated.

5.4.2 Splitting the Computational Domain

The key draw back to the method is the requirement to discretise over the complete spatial and temporal domain considered. At first sight it seems that this limitation may be removed as the quantity of interest is evaluated as an integrals over the complete domain and therefore it may be possible to split the domain, for example along the line $t = T/2$, and sum the pairs of integrals calculated over each sub-domain. The procedure would then be to, calculate approximations to the solutions (p_1, q_1) and (p_2, q_2) in the first sub-domain, $0 \leq t \leq T/2$, and then use the solution obtained along the boundary $t = T/2$, as boundary conditions for the solution in the second sub-domain $T/2 \leq t \leq T$. Unfortunately this procedure in not feasible, due to the lack of information on the internal boundary $t = T/2$. To solve over the fist sub-domain, information about the dual solution is required to form a boundary condition on $t = T/2$. However, this information has yet to be propagated backwards in time from the end condition and is therefore unavailable. The advection of the dual solution backwards in time prevent methods based on stepping forward in time from being constructed. The only circumstance in which such a method could be formulated is if the quantity of interest did not require data from 'upstream'. A quantity of interest of this form might be the integral of the solution over space and time up to the latest sub-domain. The quantity of interest considered in section 5.2 is not of this form as the integral of the solution at the final time was sought.

Chapter 6

Focusing on the Stationary Value Directly

During the course of this research the goal has been to bound physical quantities evaluated as weighted integrals of the solution of the governing equations. Using the method of twinning such quantities of interest can be expressed as the difference between the stationary value of two self-dual variational problems and therefore only self dual problems will be considered in this section. The stationary value of the functional sought in the context of a self-dual problems is

$$\Theta(\widehat{p}) = \frac{1}{2}\langle \widehat{p}, r \rangle, \tag{6.1}$$

which for this section will be considered the quantity of interest. The solution $\widehat{p}$ satisfies

$$A\widehat{p} = r, \tag{6.2}$$

and where in addition to the self-duality of the problem, the operator is considered to be self-adjoint, either naturally or by modifications to the original system through the method outlined in chapter 5. Therefore the governing equation can be written as

$$T^*T\widehat{p} = r. \tag{6.3}$$

for some operator T.

In the preceding chapter, approximations to $\Theta(\widehat{p})$ have been found by projecting the stationary equations associated with the variational principle into a finite dimensional subspace and solving weakly to find approximations $p_h \approx \widehat{p}$. Having obtained an approximation to the stationary point of the functional an approximation to the stationary value can be evaluated and, due to the convexity properties of the constrained functionals considered, these approximations have been one-sided.

The natural extension to the philosophy outlined in the abstract is to consider whether an approximation to the stationary value can be obtained directly, without first calculating the stationary point. After all, the prime objective of the simulation is solely to obtain an approximation to the stationary value. This chapter investigates the possibility of directly obtaining approximations to the stationary value of the functional, the quantity of interest sought.

Methods to evaluate the quantity of interest can be constructed in either the continuous case or through finite dimensional discretisations of the variational principles. For simplicity, the methods considered will be constructed in the finite dimensional context with reference made to the continuous analogues as appropriate.

The intuitive response to the feasibility of determining the stationary value of a functional without effectively determining the stationary point is negative. In a continuous formulation, seeking to evaluate the stationary value of a functional without knowledge of the stationary point is equivalent to evaluating a weighted integral of a function without explicit knowledge of the values the function takes over the required interval. In this context the unknown function is $\widehat{p}$ and the weighting function is r. The paradoxical nature of this task is highlighted if the weighting function r can be chosen to be a delta function, implying that the integral sought is equal to the value of the unknown function at a specified location. Of course, although the function $\widehat{p}$ is considered unknown additional information about the solution exists. For example, $\widehat{p}$ is related to the weighting function r through the governing equation (6.3), and dual extremum principles may be available on the stationary value (6.1). However, regardless of this additional information, the algorithms considered in this chapter which are capable of producing the stationary value have had in some sense to have inverted an operator of the form T^*T.

6.1 The Finite Dimensional Model

Discretisation of the stationary equations of the constrained functionals $\mathcal{G}^-(p)$ and $\mathcal{G}^+(q)$ result in linear matrix equations of the form

$$\begin{aligned} \mathsf{K}^-\mathbf{p} &= \mathbf{b}^-, & (6.4)\\ \mathsf{K}^+\mathbf{q} &= \mathbf{b}^+, & (6.5) \end{aligned}$$

where K^- and K^+ are symmetric positive definite matrices, as a consequence of the self-adjointness of the operator T^*T. The construction of K^- and K^+ in the case of the diffusion functional is given in section (2.4.3). The bounds on the stationary value of the functional Θ are then

$$\frac{1}{2}\mathbf{p}^T\mathbf{b}^- \leq \Theta(\widehat{p}) \leq \frac{1}{2}\mathbf{q}^T\mathbf{b}^+ + c_1, \tag{6.6}$$

where c_1 is a constant required in the case of the Helmholtz functional, section 4.1.2.

For generality we consider the matrix equation

$$\mathsf{K}\mathbf{x} = \mathbf{b}, \tag{6.7}$$

and wish to evaluate the scalar

$$\Theta_h = \Theta^h = \frac{1}{2}\mathbf{x}^T\mathbf{b}, \tag{6.8}$$

where Θ_h will be a one-sided bound on Θ due to the properties inherited from the variational principle. The notation Θ^h will be preferred when the subscript is used as the iteration index in the following algorithms. As with many numerical solution procedures there are direct and iterative alternatives, both of which will be considered. Moreover the formulation of the problem from which Θ_h will be obtained can be derived in a multiplicity of ways, including descent methods and eigenvalue problems.

6.2 Descent Methods

Of the many descent, or gradient, methods formulated for solving a system of linear equations the conjugate gradient algorithm is particularly efficient. Equivalently, it is used in finding the stationary point of a quadratic functional when the matrix K

is symmetric and positive definite. Therefore, implementing the conjugate gradient algorithm to obtain the quantity of interest is a natural approach since the self-adjoint positive nature of the governing operator translates to a symmetric positive definite matrix.

6.2.1 A Conjugate Gradient Approach

Descent methods for obtaining the stationary point of a function are based on the observation that that the solution vector $\mathbf{x}$ of (6.7) coincides with the stationary point, a maximum say, of the function

$$g(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^T \mathsf{K}\mathbf{x} + \mathbf{b}^T\mathbf{x} \tag{6.9}$$

over all $\mathbf{x}$. Of greater interest is that the stationary value $g(\widehat{\mathbf{x}})$ coincides with the quantity sought, Θ_h. It is worth noting that this is just the finite dimensional analogue of the variational principles discussed in the previous chapters.

The efficiency of the conjugate gradient method in obtaining the stationary point is due to the iterative construction of a set of conjugate basis vectors spanning the solution space, and in which $\widehat{\mathbf{x}}$ can be expanded. In exact arithmetic and for N unknowns it is known that this construction is completed in N iterations. However, for large problems a sufficient degree of accuracy may be attained earlier. The efficiency of the method to find the solution $\mathbf{x}$ extends to efficiently finding Θ_h. In addition a small storage saving is made as the vector $\mathbf{x}$ is not stored in memory.

The conjugate gradient method as stated in Braess [8] is: for an initial vector $\mathbf{x}_0$ and search direction $\mathbf{d}_0 = -\mathbf{g}_0 = \mathbf{b} - \mathsf{K}\mathbf{x}_0$, iterate $k = 0, 1, 2, ...$

$$\begin{aligned}
\alpha_k &= \frac{\mathbf{g}_k^T \mathbf{g}_k}{\mathbf{d}_k^T \mathsf{K}\mathbf{d}_k^T}, && (6.10)\\
\mathbf{x}_{k+1} &= \mathbf{x}_k + \alpha_k \mathbf{d}_k, && (6.11)\\
\mathbf{g}_{k+1} &= \mathbf{g}_k + \alpha_k \mathsf{K}\mathbf{d}_k, && (6.12)\\
\beta_k &= \frac{\mathbf{g}_{k+1}^T \mathbf{g}_{k+1}}{\mathbf{g}_k^T \mathbf{g}_k}, && (6.13)\\
\mathbf{d}_{k+1} &= -\mathbf{g}_{k+1} + \beta_k \mathbf{d}_k, && (6.14)
\end{aligned}$$

until the residual $\mathbf{g}_k = 0$, or a convergence criteria is satisfied.

The conjugate gradient method applied directly to $\Theta^h = \Theta_h$ is found simply by updating Θ^h instead of $\mathbf{x}$. Therefore, with an initial vector $\mathbf{x}_0$, an initial value for the quantity of interest is calculated,

$$\Theta_0^h = \frac{1}{2}\mathbf{x}_0^T\mathbf{b} \tag{6.15}$$

and the same search direction $\mathbf{d}_0 = -\mathbf{g}_0 = \mathbf{b} - \mathsf{K}\mathbf{x}_0$ is employed. The modified conjugate gradients algorithm is then: iterate $k = 0, 1, 2, ...,$

$$\alpha_k = \frac{\mathbf{g}_k^T\mathbf{g}_k}{\mathbf{d}_k^T\mathsf{K}\mathbf{d}_k^T}, \tag{6.16}$$

$$\Theta_{k+1}^h = \Theta_k^h + \frac{\alpha_k}{2}\mathbf{d}_k^T\mathbf{b}, \tag{6.17}$$

$$\mathbf{g}_{k+1} = \mathbf{g}_k + \alpha_k\mathsf{K}\mathbf{d}_k, \tag{6.18}$$

$$\beta_k = \frac{\mathbf{g}_{k+1}^T\mathbf{g}_{k+1}}{\mathbf{g}_k^T\mathbf{g}_k}, \tag{6.19}$$

$$\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \beta_k\mathbf{d}_k, \tag{6.20}$$

until the residual $\mathbf{g}_k = 0$ or a convergence criteria is satisfied.

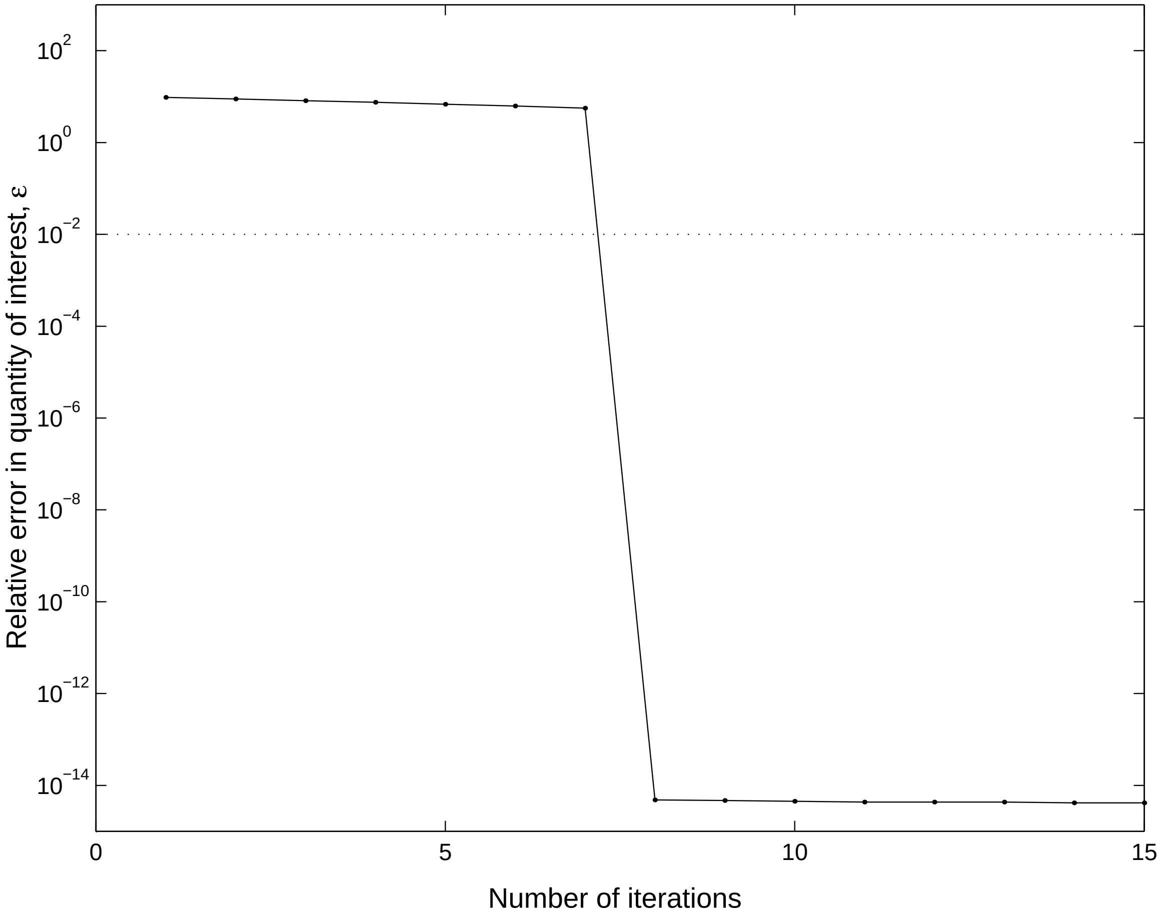

Figure 6.1: Convergence of Θ_h computed using the conjugate gradient method, $N = 15$

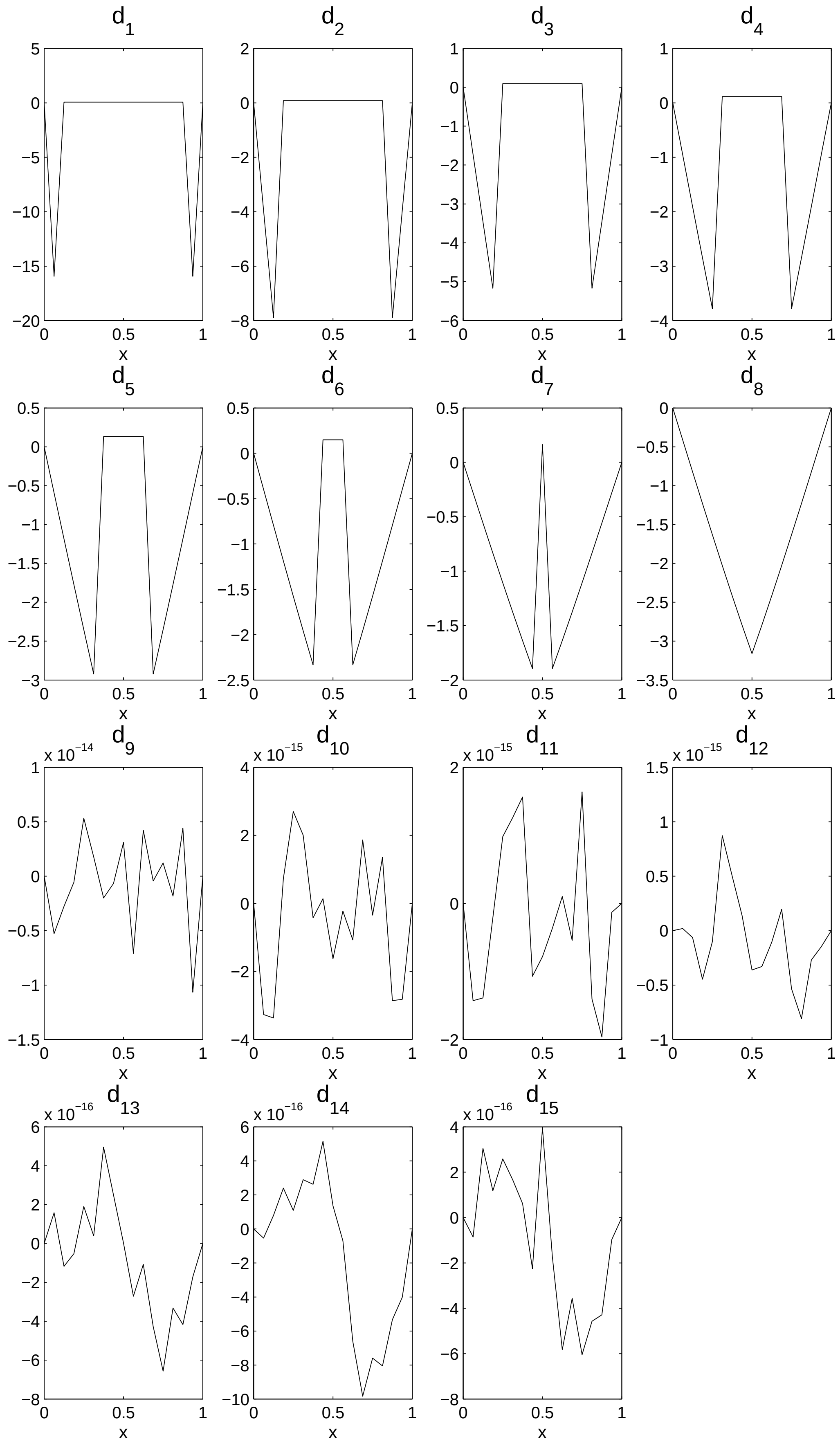

Figure 6.2: Basis vectors generated by the conjugate gradient method, $N = 15$

The performance of the conjugate gradient method is evaluated by obtaining a lower bound on the stationary value of the functional governing the solution of the Poisson equation

$$-\frac{d^2\widehat{p}(x)}{dx^2} = 1 \qquad 0 \leq x \leq 1, \tag{6.21}$$

$$\widehat{p}(0) = \widehat{p}(1) = 0. \tag{6.22}$$

The diffusion functional governs the solution of the equation set (6.21)-(6.22), and in particular a lower bound on the stationary value of the functional is given by

$$\mathcal{G}^{-}(p) = \int_0^1 \left\{ p - \frac{1}{2}\frac{dp}{dx}\frac{dp}{dx} \right\} dx, \tag{6.23}$$

where the function p satisfies the boundary conditions (6.22). Discretising the stationary equations using a finite element expansion for p with basis functions ϕ_i also satisfying the boundary conditions (6.22),

$$p = \sum_{i=1}^{N} x_i \phi_i, \tag{6.24}$$

$$\delta p = \phi_i \qquad i = 1, ..., N \tag{6.25}$$

results in the matrix equation

$$\mathsf{K}\mathbf{x} = \mathbf{b}, \tag{6.26}$$

governing the coefficients x_i. The elements of the $N \times N$ matrix K and the $N \times 1$ vector $\mathbf{b}$ in this case are

$$\mathsf{K}_{ij} = \int_0^1 \frac{d\phi_i}{dx}\frac{d\phi_j}{dx} dx, \tag{6.27}$$

$$\mathbf{b}_i = \int_0^1 \phi_i dx. \tag{6.28}$$

The convergence of the stationary value calculated using the conjugate gradient method, with respect to the number of iterations, is shown in figure 6.1. The method was implemented with differing levels of discretisation fineness and the number of iterations required to achieve a one percentage relative error in the stationary value is graphed in figure 6.6. In general, the required degree of accuracy in the quantity of interest was obtained within approximately $\frac{N}{2}$ iterations. This characteristic is displayed in the convergence history of the method with $N = 15$ shown in figure 6.1. The sudden reduction in the error can be explained by considering the basis vectors generated by the algorithm. On inspection of the basis vectors, shown in figure 6.2, it is apparent that the solution, which is symmetric about the axis $x = 0.5$, is efficiently represented by the first eight vectors alone, with the remaining vectors contributing negligibly.

In general the convergence of the quantity of interest may not exhibit the characteristics displayed by this example. However, the conjugate gradient method offers a robust descent method for which the convergence of the quantity of interest could always be accelerated by the use of an appropriate pre-conditioner if required.

6.3 An Eigenvalue Formulation

The descent method implemented in the previous sections replicates the structure of the continuous problem at a discrete level. Alternatively, the problem of obtaining the quantity of interest can be re-formulated as an eigenvalue equation. This finite dimensional eigenvalue problem will be examined with the aim being to construct numerical methods to evaluate the quantity of interest. The corresponding continuous eigenvalue problem is considered by Levine and Schwinger [28] to obtain approximations to scale invariant quantities in electromagnetic theory.

The eigenvalue problem is formulated by rendering the solution scale invariant by rewriting it as

$$\mathsf{K}\mathbf{x} = \frac{(\mathbf{x}^T\mathbf{b})}{2\Theta_h}\mathbf{b}. \tag{6.29}$$

Re-arranging (6.29), the generalised eigenvalue problem

$$\begin{aligned} \Theta_h \mathsf{K}\mathbf{x} &= \frac{1}{2}\mathbf{b}\mathbf{b}^T\mathbf{x}, & (6.30)\\ &= \mathsf{B}\mathbf{x}, & (6.31) \end{aligned}$$

or equivalently

$$(\mathsf{B} - \Theta_h \mathsf{K})\mathbf{x} = 0 \tag{6.32}$$

is obtained, where

$$\mathsf{B} = \frac{1}{2}\mathbf{b}\mathbf{b}^T \tag{6.33}$$

is a rank 1 operator, Θ_h is the eigenvalue parameter and $\mathbf{x}$ the corresponding eigenvector. The degenerate nature of the operator B strongly affects the distribution of the eigenvalues of this problem. In particular, the eigenvalues of (6.32) are the same as those of the standard eigenvalue problem

$$(\mathsf{K}^{-1}\mathsf{B} - \Theta_h \mathsf{I})\mathbf{x} = 0 \tag{6.34}$$

as K is of full rank. The eigenvalues of the operator $\mathsf{K}^{-1}\mathsf{B}$ can be investigated by decomposing the symmetric positive definite matrix K such that

$$\mathsf{K}{=}\mathsf{L}^{-1}\mathsf{L}^{-T}. \tag{6.35}$$

The standard eigenvalue problem is then of the form

$$\begin{aligned}\Theta_h \mathbf{x} &= \frac{1}{2}\left(\mathsf{L}\mathbf{b}\right)\left(\mathbf{b}^T\mathsf{L}^T\right)\mathbf{x}, && (6.36)\\ &= \frac{1}{2}\mathbf{z}\mathbf{z}^T\mathbf{x}, && (6.37)\end{aligned}$$

where

$$\mathbf{z} = \mathsf{L}\mathbf{b}, \tag{6.38}$$

and the hence the product of the operators $\mathsf{K}^{-1}\mathsf{B}$ is shown to be a the rank 1 outer product of the vector $\mathbf{z}$. The rank one nature of the operator $\mathsf{K}^{-1}\mathsf{B}$ implies that there exists a single non-zero eigenvalue which corresponds to the quantity of interest sought.

Ideally the single non-zero eigenvalue of (6.32) could be found directly. One obvious approach is to invert the full rank matrix K and evaluate Θ_h as the trace of the matrix $\mathsf{K}^{-1}\mathsf{B}$. However, inverting the matrix K is costly and little different from solving the original matrix problem (6.7). An alternative method to calculate this eigenvalue using determinants can be found in Porter and Stirling [40].

6.3.1 A Direct Method

The ability to obtain an explicit expression for the quantity of interest, the non-zero eigenvalue, is due to the degenerate nature of the operator B. To obtain the result of Porter and Stirling it is observed that for a non-trivial eigenvector to exist the operator must satisfy

$$\begin{aligned}0 &= \det(\mathsf{K} - \frac{1}{2\Theta_h}\mathbf{b}\mathbf{b}^T), && (6.39)\\ &= \left|\mathsf{K} - \frac{1}{2\Theta_h}\mathbf{b}\mathbf{b}^T\right|. && (6.40)\end{aligned}$$

Denoting the governing operator as

$$\mathsf{G} = \Theta_h^{-1}\mathbf{b}\mathbf{b}^T - 2\mathsf{K}, \tag{6.41}$$

where $\Theta_h \neq 0$ the elements of G are

$$\mathsf{G}_{ij} = \Theta_h^{-1}\mathbf{b}_i\mathbf{b}_j^T - 2\mathsf{K}_{ij}. \tag{6.42}$$

Elementary row operations can then be applied to G to obtain $\bar{\mathsf{G}}$ where

$$\bar{\mathsf{G}}_{ij} = \begin{cases} \Theta_h^{-1}\mathbf{b}_1\mathbf{b}_j^T - 2\mathsf{K}_{1j} & i \;=\; 1, \\ 2\alpha_i\mathsf{K}_{1j} - 2\mathsf{K}_{ij} & i \;\neq\; 1, \end{cases} \tag{6.43}$$

and

$$\alpha_i = \frac{\mathbf{b}_i}{\mathbf{b}_1}. \tag{6.44}$$

The ability to eliminate Θ_h^{-1} from all but the first row is due to the rank 1 property of the product $\mathbf{b}\mathbf{b}^T$. In addition, the structure of $\bar{\mathsf{G}}$ enables the determinant of the operator to be written as the difference of two closely related determinants. Therefore,

$$\begin{aligned} 0 &= \left|\mathsf{G}\right|, & (6.45) \\ &= \left|\bar{\mathsf{G}}\right|, & (6.46) \\ &= \sum_{j=1}^{N} \left(\Theta_h^{-1}\mathbf{b}_1\mathbf{b}_j^T - 2\mathsf{K}_{1j}\right)\mathsf{C}_{ij}, & (6.47) \\ &= \sum_{j=1}^{N} \Theta_h^{-1}\mathbf{b}_1\mathbf{b}_j^T\mathsf{C}_{ij} - 2\mathsf{K}_{1j}\mathsf{C}_{ij}, & (6.48) \\ &= \Theta_h^{-1}\mathbf{b}_1\left|\mathsf{X}\right| - 2\left|Y\right|, & (6.49) \end{aligned}$$

where C_{ij} are the cofactors of the $(N-1\times N)$ matrix A with elements

$$\mathsf{A}_{ij} = 2\alpha_{i+1}\mathsf{K}_{1,j} - 2\mathsf{K}_{i+1,j}, \tag{6.50}$$

such that

$$\mathsf{X} = \left[\begin{array}{c} \mathbf{b}^T \\ \hline \mathsf{A} \end{array}\right], \qquad \mathsf{Y} = \left[\begin{array}{c} \mathbf{k}^T \\ \hline \mathsf{A} \end{array}\right], \tag{6.51}$$

and $\mathbf{k}^T$ is the first row of K

$$\mathbf{k}_i^T = \mathsf{K}_{1,i}. \tag{6.52}$$

An explicit formula for the quantity of interest is then

$$\Theta_h = \frac{\mathbf{b}_1}{2}\frac{\left|\mathsf{X}\right|}{\left|\mathsf{Y}\right|}. \tag{6.53}$$

In adopting this method it is assumed that $\mathbf{b}_1 \neq 0$. If this is an invalid assumption then the equations can be re-numbered or the determinant expanded around a different row. The existence of at least one non-zero element of $\mathbf{b}$ is guaranteed in order that the solution $\mathbf{x}$ of

$$\mathsf{K}\mathbf{x} = \mathbf{b} \tag{6.54}$$

is non-trivial. Although the formula (6.53) provides an explicit means of obtaining the quantity of interest the direct numerical evaluation of determinants is generally avoided due to the high number of multiplication operations involved. Practically, an LU factorisation of the matrices X and Y would be implemented and the determinant evaluated from the diagonal elements. However, the computational cost of calculating the determinants is unlikely to be less than computing the stationary value of the functional using the conjugate gradient algorithm.

An exceptional case in which the evaluation of the determinants may prove the cheaper option is when the computation is executed in parallel and the explicit nature of the determinant exhibits greater scalability. The relative savings of parallel computing have not been considered in this research and it is assumed that all computations are carried out in a serial fashion.

The quotient of two determinants appearing in the computation of quantity of interest is reminiscent of Cramer's Rule and suggests the existence of a matrix system in which Θ_h is evaluated directly from a single component of the solution vector. The corresponding matrix system can be found through further simplification of the result of Porter and Stirling. The key to the simplification is the high degree of similarity between the operators X and Y enabling X to be written as

$$\mathsf{X} = (\mathsf{I} + \mathsf{H})\,\mathsf{Y}, \tag{6.55}$$

where

$$\mathsf{H} = \left[\begin{array}{c} \mathbf{h}^T \\ \hline \mathsf{0} \end{array}\right]. \tag{6.56}$$

Therefore, the quantity of interest can be expressed in the following manner

$$\Theta_h \;=\; \frac{\mathbf{b}_1}{2}\frac{|\mathsf{X}|}{|\mathsf{Y}|}, \tag{6.57}$$

$$= \frac{\mathbf{b}_1}{2}\frac{|\mathsf{I}+\mathsf{H}|\,|\mathsf{Y}|}{|\mathsf{Y}|}, \tag{6.58}$$

$$= \frac{\mathbf{b}_1}{2}\left(1+\mathbf{h}_1\right), \tag{6.59}$$

where the vector $\mathbf{h}$ is obtained by solving the system

$$\mathsf{Y}^T\mathbf{h} = \mathbf{b} - \mathbf{k}. \tag{6.60}$$

Again, obtaining $\mathbf{h}_1$ from (6.60) is unlikely to be as cost efficient as computing the quantity of interest using the conjugate gradient algorithm since Y is no longer symmetric.

The matrix system (6.60) illustrates that although quantity of interest appears to have been computed in isolation from an N dimensional stationary point, the eigenvalue formulation embeds the quantity of interest within a complementary N dimensional problem. Therefore, in either approach the core cost of inverting a $N \times N$ operator is incurred in obtaining the quantity of interest.

The iterative descent method discussed in section 6.2.1 was found to be an effective method to obtain the quantity of interest and therefore it is natural to consider whether an effective iterative algorithm to compute the single non-zero eigenvalue exists.

6.3.2 An Iterative Eigenvalue Approach

An iterative procedure to evaluate $\Theta^h = \Theta_h$ is constructed by considering the largest positive eigenvalue, β, of the operator

$$\mathsf{F} = \mathsf{B} - \Theta^h_k(\mathsf{K}-\mathsf{I}), \tag{6.61}$$

and tracing a convergence path such that in the limit as the iteration count k gets large, $\beta_k \to \Theta^h_k \to \Theta$. Under such conditions the eigenvalue problem

$$[\mathsf{B} - \Theta^h_k(\mathsf{K}-\mathsf{I})]\mathbf{x} = \beta_k\mathbf{x} \tag{6.62}$$

converges to

$$[\mathsf{B} - \Theta^h\mathsf{K}]\mathbf{x} = 0. \tag{6.63}$$

The power method is used to obtain approximations to the largest eigenvalue of the operator F and, to ensure that the power method converges to the largest eigenvalue, an additional shift is added and the operator

$$\mathsf{F}(\Theta_k^h, \gamma) = \mathsf{B} - \Theta_k^h(\mathsf{K} - \mathsf{I}) + \gamma \mathsf{I} \tag{6.64}$$

is considered. A suitable convergence path is found to be attained by a simple relaxation of the Θ_k. The algorithm is then

- Initialise: An initial value for Θ^h, Θ_0^h is required, which may be obtained from a coarse solution, an estimation, or arbitrary value. An estimation for the magnitude of the shift required can be obtained by applying a few iterations of the power method to the operator $\mathsf{F}(\Theta_0^h, 0)$. If the results of the power method, α, is negative and reasonably converged then an effective shift is $\gamma = -1.5\alpha$, which helps ensure that all the eigenvalues of F are positive and therefore β is the dominant eigenvalue identified by the power method. Due to the additional shift $\beta_0 = \alpha + \gamma$. An initial vector $\mathbf{x}_0$ is also required. In the limit $\Theta_k^h \to \Theta^h$, $\mathbf{x}$ converges to the eigenvector with largest eigenvalue and hence this will be the scale invariant solution of the matrix equation (6.3). Therefore, a vector containing any coarse attributes of the solution is a good initialisation point, although an arbitrary vector is also acceptable.

- Iterate: The iterations proceed according to the algorithm

$$\begin{aligned}
\Theta_k^h &= \theta^h(\beta_{k-1} - \gamma) + (1 - \theta)\Theta_{k-1} && (6.65)\\
\mathbf{x}_k &= \frac{\mathbf{x}_k}{\mathbf{x}_k^T \mathbf{x}_k} && (6.66)\\
\mathbf{y} &= \mathsf{F}(\Theta_k^h, \gamma)\mathbf{x}_k && (6.67)\\
\beta_k &= \mathbf{x}_k^T \mathbf{y} && (6.68)\\
\mathbf{x}_{k+1} &= \mathbf{y} && (6.69)
\end{aligned}$$

The results obtained by applying this iterated power method to the example (6.5) are shown in figure 6.6. In comparison with the results obtained using the conjugate gradient algorithm the convergence of the power method is considerably slower. In contrast to the conjugate gradient method the power method does not systematically construct a basis in which to expand the solution. Instead, the power method converges to the dominant eigenvector in the limit of a sequence in which the contributions from the remaining eigenvectors become negligible. Therefore, using the power method the limit of a sequence is sought and acceleration techniques can be employed in an attempt to reach the limit point more efficiently. The use of acceleration methods is discussed in the next section.

Alternatively, the reciprocal eigenvalue problem

$$\mathsf{K}\mathbf{x} = \mu\mathsf{B}\mathbf{x} \tag{6.70}$$

can be considered where

$$\Theta_h = \frac{1}{\mu}. \tag{6.71}$$

The eigenvalues of (6.70) are now all infinite except one corresponding to the reciprocal of the quantity of interest. An iteration method of the form

$$[\mathsf{K} - \mu(\mathsf{B} - \mathsf{I})]\mathbf{x} = \beta_k\mathbf{x} \tag{6.72}$$

where the convergence $\beta_k \rightarrow \mu_k \rightarrow \Theta_h^{-1}$ is then desired. The large interval between the eigenvalue of interest and the remaining eigenvalues suggest that iterative methods will converge rapidly. However, to obtain approximations to the smallest eigenvalue the inverse power method is required and this involves the inversion of the governing operator. The cost of inverting the operator precludes serious consideration of this method.

6.3.3 An Accelerated Power Method

The convergence rate of the power method is the main detraction from the algorithm and therefore attempts to accelerate this should be investigated. Ideally the convergence of the sequence of eigenvector approximations $\mathbf{x}_k$ tending to $\mathbf{x}$ would be accelerated as the eigenvector, as opposed to the eigenvalue, is the driving variable in the algorithm. The accelerated algorithm is then

- Initialise as before

- Iterate with every n^{th} eigenvalue update accelerated

$$\Theta_k^h = \theta(\beta_{k-1} - \gamma) + (1 - \theta)\Theta_{k-1}^h \tag{6.73}$$

$$\mathbf{x}_k = \frac{\mathbf{x}_k}{\mathbf{x}_k^T\mathbf{x}_k} \tag{6.74}$$

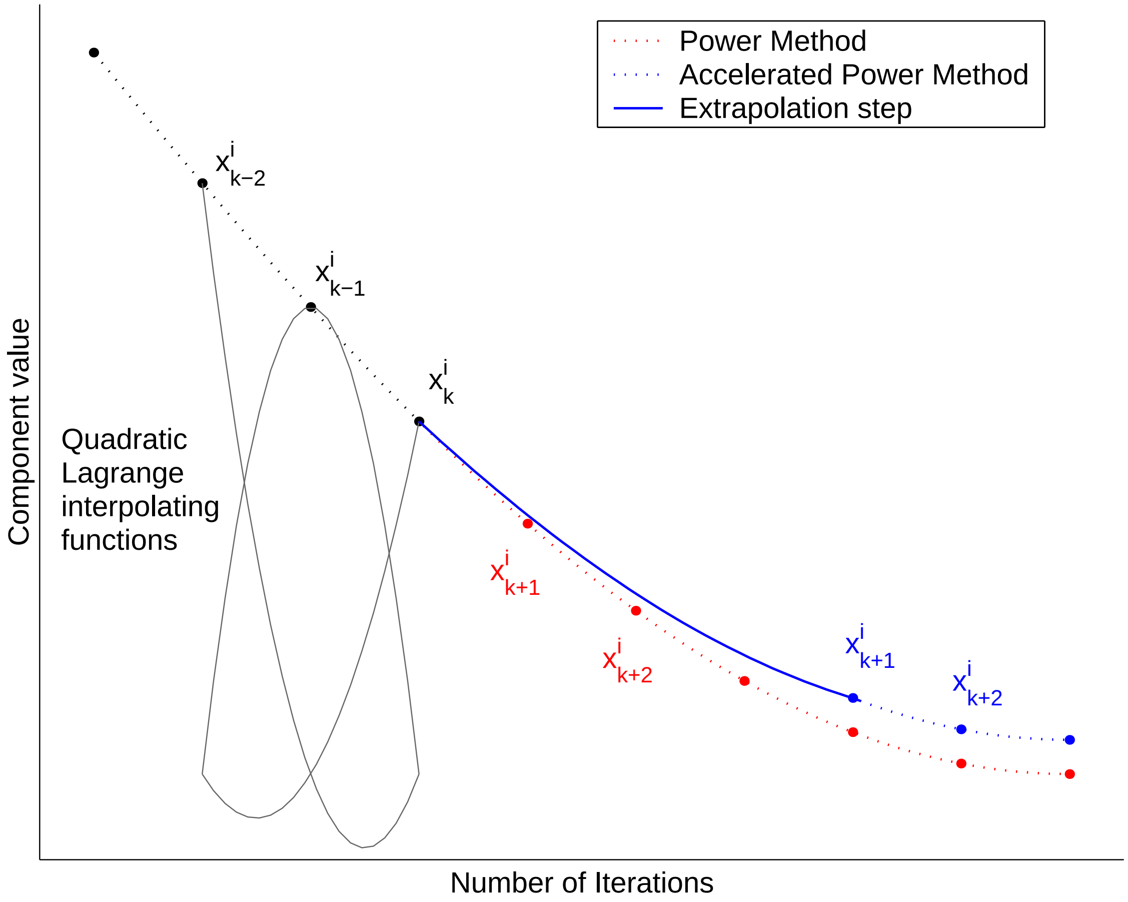

Figure 6.3: Schematic diagram of the interpolation-extrapolation method applied to the i^{th} component of $\mathbf{x}$

$$\mathbf{y}_k = \mathsf{F}(\Theta^h_k, \gamma)\mathbf{x}_k \tag{6.75}$$

$$\beta_k = \mathbf{x}^T_k \mathbf{y}_k \tag{6.76}$$

$$\text{every } n^{th} \text{ iteration} \tag{6.77}$$

$$\quad \mathbf{x}_{k+1} = A(\mathbf{y}_k, \mathbf{y}_{k-1}, ..., \mathbf{y}_{k-m}) \tag{6.78}$$

$$\text{else} \tag{6.79}$$

$$\quad \mathbf{x}_{k+1} = \mathbf{y}_k \tag{6.80}$$

where $A(\mathbf{y}_k, \mathbf{y}_{k-1}, ...)$ is an acceleration scheme depending on the previous m iterations.

The standard Aitken acceleration method was found to be inefficient in this application due to fluctuations in the convergence path of the eigenvector components. Instead, an interpolation-extrapolation acceleration method is considered. The interpolation-

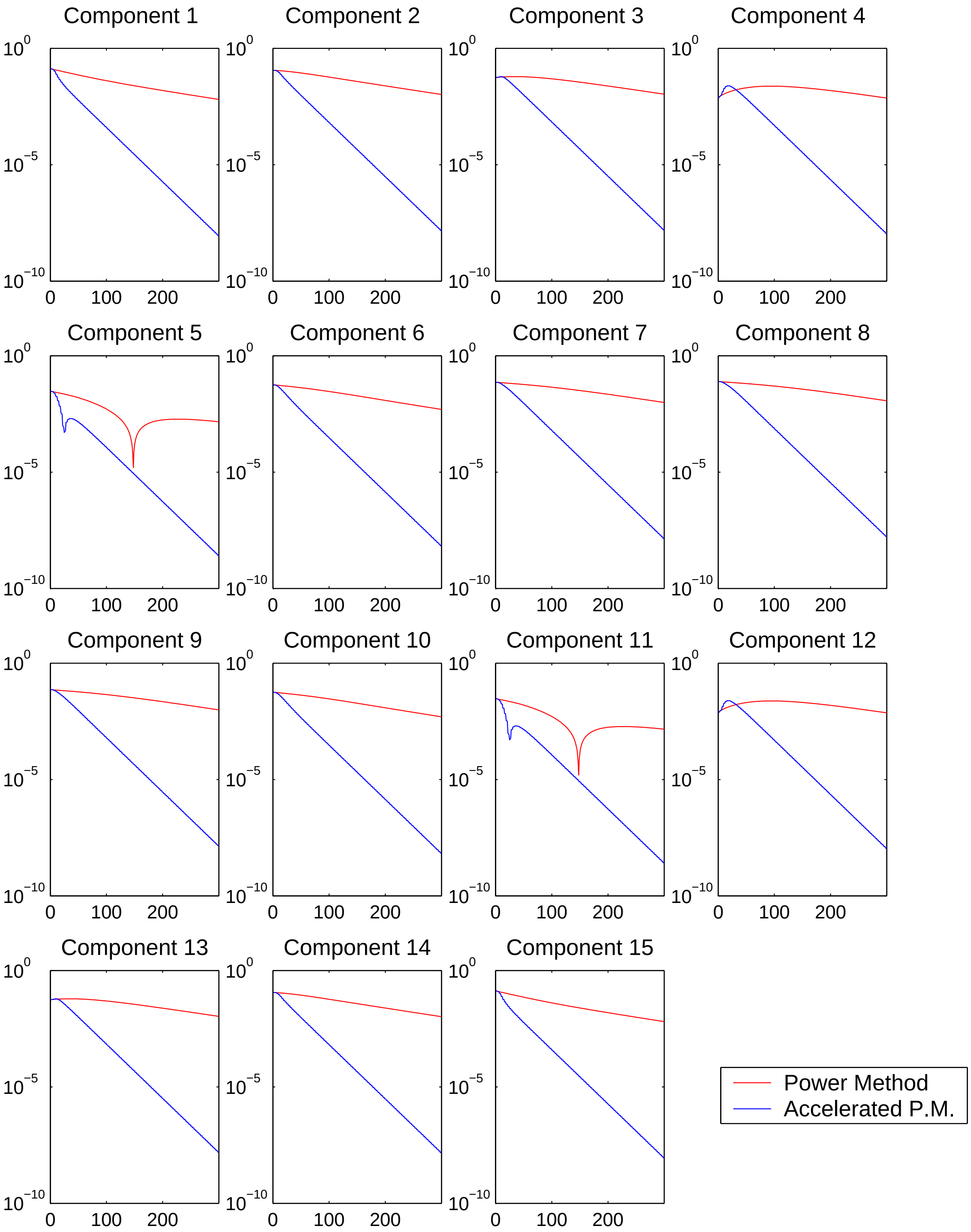

Figure 6.4: Convergence of the eigenvector components with number of iterations

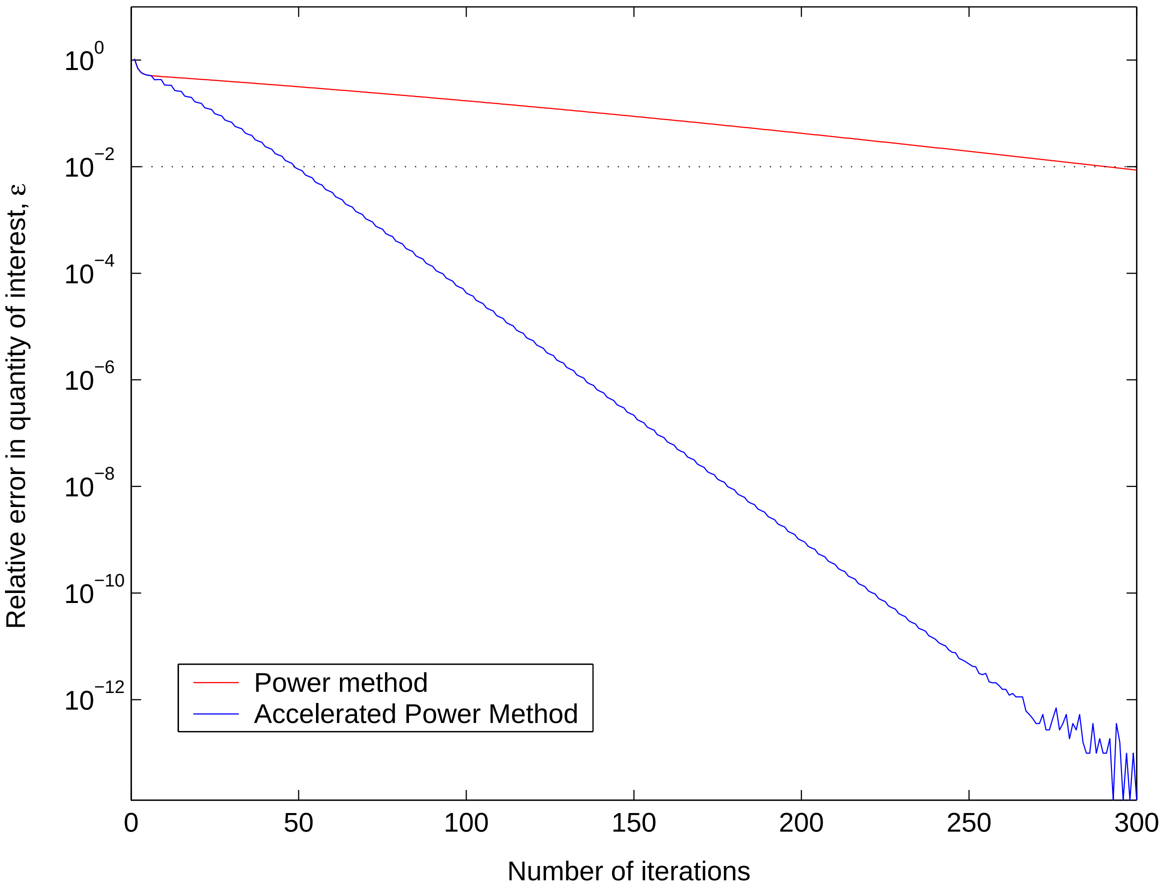

Figure 6.5: Convergence of Θ_h using the standard and accelerated power methods

extrapolation convergence method aims to construct a local approximation to the components of the eigenvector by interpolating the value of the components over previous iterations using the Lagrangian functions commonly used as basis functions in finite element methods. In the implementation considered, the individual components of the last three normalised vectors $\mathbf{y}$ were interpolated using quadratic polynomials. From this interpolation the first and second derivatives of the components with respect to the number of iterations can be approximated and then the path of each component extrapolated forward using a truncated Taylor series. This approach is schematically illustrated in figure 6.3.

The step size over which to extrapolate is calculated automatically by balancing the terms in the Taylor series. Therefore

$$\text{step} = \min\left(\left|\frac{2f'}{f''}\right|, \text{max step}\right) \tag{6.81}$$

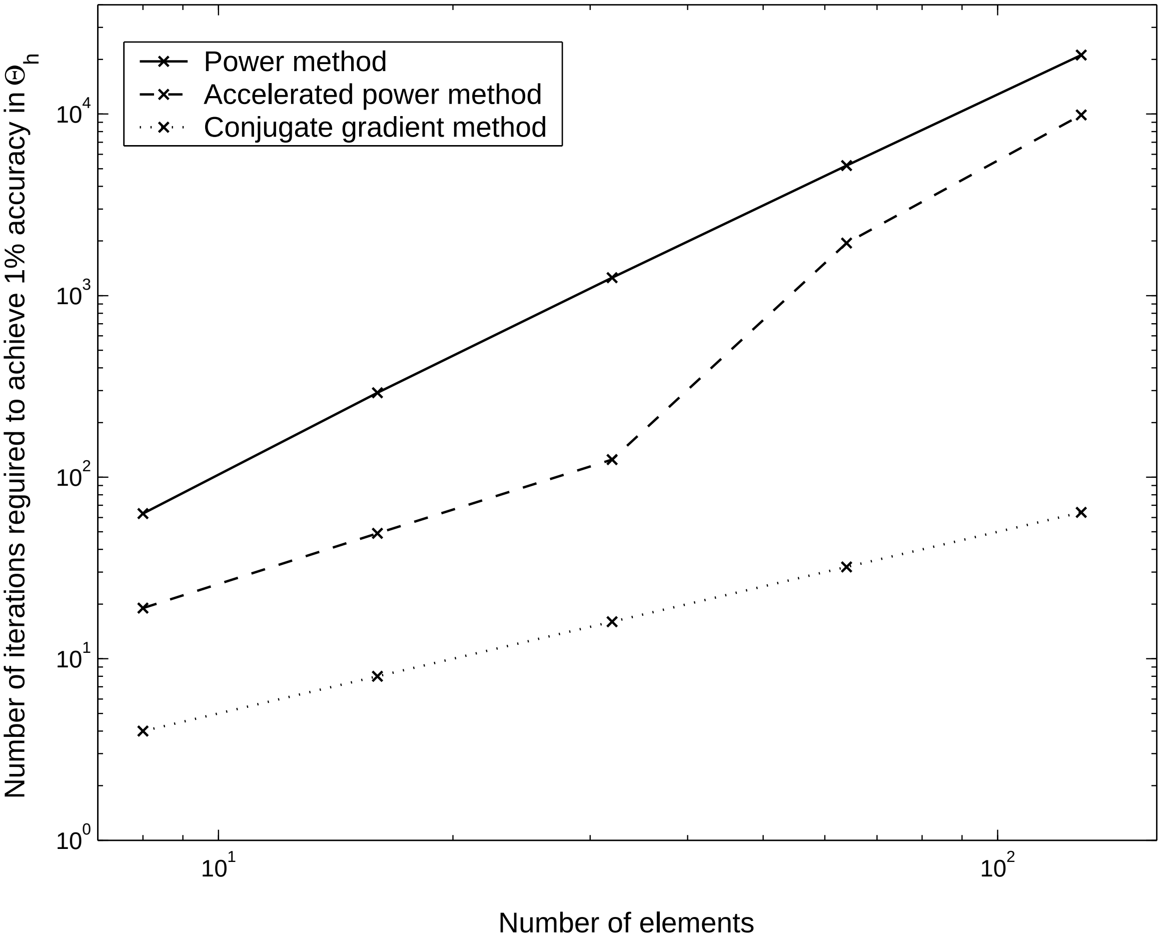

Figure 6.6: Comparison of descent and power methods

where f' and f'' are the approximations to the derivatives of each component of the eigenvector with respect to number of iterations. An upper limit is placed on the step size to prevent large step sizes occurring in regions where there is little curvature present. A different step size is then calculated for each component of the eigenvector in order to speed convergence. A suitable value for the maximum step size would seem to be N, the number of unknowns, and therefore the number of components in the eigenvector. An explanation as to why N is a suitable maximum step size is due to the observation that in larger problems the convergence rate of the eigenvector is slower and therefore a larger maximum step length can be afforded.

The interpolation-extrapolation method improves the speed at which the components of the eigenvector converge. This improvement is registered in figure 6.4 and is also reflected in the speed in which the accelerated power method converges to the required eigenvalue, figure 6.5. However, in comparison with the conjugate gradient method the accelerated power method still proves costly. The relative cost of the methods are shown in figure 6.6. From figure 6.6 the relationship between computational cost and size of problem can be investigated. The cost of an iteration of the power method and the conjugate gradient method is considered comparable as both are dominated by a single matrix vector multiplication. Figure 6.6 demonstrates that not only is the conjugate gradient method cheaper but the cost of the method grows less quickly with problem size than the power method. In addition, the graph suggests that although the accelerated power method is faster than the non-accelerated version both methods share the same dependence on the size of the problem. This is due to both methods sharing similar convergence paths, but with the accelerated version effectively traversing these paths more swiftly.

6.4 Conclusions and Extensions

In this chapter algorithms to compute the quantity of interest have been investigated. The conclusions that can be drawn from this investigation are that, regardless of the method, a base cost exists equivalent to solving

$$\mathsf{K}\mathbf{x} = \mathbf{b}. \tag{6.82}$$

Moreover, whilst re-formulating the original system as an eigenvalue problem enables a succinct representation for the quantity of interest to be found, the standard computational cost of evaluating the expression detracts from the method. In general it is therefore recommended that the solution method should mirror the original continuous structure and the conjugate gradient approach be adopted.

The methods considered in sections 6.2 and 6.3 have been formulated on the basis of continuous counterparts. In addition, the quantity of interest can appear in the solution vector through augmenting the resulting set of stationary equations or careful selection of the basis functions for the solution.

6.4.1 The Quantity of Interest as a Component of the Solution Vector

The simplest method to obtain the quantity of interest as a component of the solution vector is to augment the set of finite dimensional stationary equations

$$\mathsf{K}\mathbf{x} = \mathbf{b} \tag{6.83}$$

with the equation for the quantity of interest

$$\Theta_h = \frac{1}{2}\mathbf{x}^T\mathbf{b}. \tag{6.84}$$

The resulting matrix system is then

$$\left[\begin{array}{c|c} 2 & \mathbf{b}^T \\ \hline 0 & \mathsf{K} \end{array}\right] \left[\begin{array}{c} \Theta_h \\ \mathbf{x} \end{array}\right] = \left[\begin{array}{c} 0 \\ \mathbf{b} \end{array}\right] \tag{6.85}$$

which could be solved for the first unknown. However, the reformulation of the problem in this form generates little computational advantage as the matrix is now larger and lacks symmetry. Of greater interest is the discretisation in which the original forcing function r is included in the approximations space. A lower bound would then be found by considering the expansion

$$p_h = \sum_{i=1}^{N} x_i\phi_i \tag{6.86}$$

where the first basis function is chosen as

$$\phi_1 = \frac{1}{\langle r, r\rangle} r \tag{6.87}$$

and the remaining satisfy

$$\langle \phi_1, \phi_i\rangle = 0 \qquad\qquad i = 2, \cdots, N. \tag{6.88}$$

The basis functions generate the matrix

$$\mathsf{K}_{ij} = \langle\!\langle T\phi_i, T\phi_j\rangle\!\rangle \tag{6.89}$$

and vector

$$\mathbf{b}_i = \langle \phi_i, r\rangle = \left[\begin{array}{c} 1 \\ \hline 0 \end{array}\right] \tag{6.90}$$

respectively. The quantity of interest is then the first unknown in the matrix equation

$$\mathsf{K}\mathbf{x} = \mathbf{b} \tag{6.91}$$

where K has now retained symmetry.

In the construction of this example it has been assumed that the boundary conditions on p_h do not imply that $x_1 = 0$. Considerations of this nature would have to be examined for each particular application and if appropriate, the boundary conditions could be applied through the use of Lagrange multipliers. In general, it is expected that including the function onto which the projection of the solution is sought within the approximation space might be advantageous. Further research is required in this area to investigate this hypothesis. In this example the forcing function has been used explicitly to generate the approximation space, however, providing the function is representable by the basis functions similar results could be expected, although the quantity of interest would no longer be identified with the solution component x_1.

Chapter 7

Conclusions and Further Work

The thesis has explored the applications of dual variational principles to a range of practical problems. In the self-adjoint setting the existing theory is elegant and effective at generating sharp bounds on integral quantities for self-dual problems. Incorporating the twin transformation enables these bounds to be generated on non-self-dual problems. The geometric interpretation of the dual variational formulation has proved a useful tool from which many of the extensions of method have been generated.

The construction of the novel consistent upscaling methods is an example in which the saddle-shaped interpretation of the governing functional is exploited. The comparison problem at the heart of the consistent upscaling method ensures that the substitute functionals, defined over the coarse permeability data, lie above and below the convex and concave portions of the original functional respectively.

A considerable part of the research undertaken was involved in extending and applying dual extremum principles to non-self adjoint problems. Extending the theory to include non-self-adjoint problems was motivated by the desire to apply the dual extremum techniques to time dependent problems. The first extension considered involved constructing methods that are discrete in time but continuous in space. These semi-discrete methods (constructed in chapter 4) provide an efficient means of approximating the quantity of interest via the dual extremum principles associated with the Helmholtz equation. The extremum principles for the Helmholtz equation enables two approximations to the quantity of interest to be made and therefore provide an estimate of how well the problem has been resolved.

The semi-discrete methods provide a means of approximating the quantity of interest but in order to obtain strict upper and lower bounds on the quantity of interest associated with a non-self-adjoint problems consideration of the complete space time domain was found to be necessary. The requirement to consider the complete domain is due to the embedding of the original non-self-adjoint problem in a larger self-adjoint problem. Embedding the original problem in this manner is not normally considered advantageous, for example greater differentiability is demanded from the basis functions. However, in this context it is assumed that the construction of upper and lower bounds on the quantity of interest offsets the difficulties associated with the embedding. To obtain the upper and lower bounds in the continuous non-self-adjoint case the alternative bound of section 5.1.2 was implemented. The validity of the alternative bound depends on the convergence rate of the finite element method employed. Further work is required in this area to determine situations in which this bound may fail. In conclusion, the extension of dual extremum principles to non-self-adjoint problems is possible provided that a variational formulation exists for the self-adjoint alternative containing the original problem. The attractiveness of implementing the method will depend on the importance of obtaining bounds on the quantity of interest relative to the cost of the computation. However the bounds are computable, determined purely from easily obtainable weak solution, e.g. finite elements. This is in direct contrast to the bounds proposed for non-self-adjoint problems by Collins [12] and Gurtin [19] which in practice are hard to implement.

The final chapter pursues efforts to obtain the quantity of interest without explicitly determining the solution of the governing equation. In a finite dimensional discretisation of the continuous problem the conjugate gradient method is found to be particularly effective at calculating the stationary point of the functional and therefore indirectly the quantity of interest. In addition, consideration of the scale invariant form of the variational principle transforms the approximation of the quantity of interest into an eigenvalue problem. However, approximating the required eigenvalue is found to be more expensive than implementing the conjugate gradient minimization itself, and therefore, the conclusion is that conjugate gradient algorithm provides the most effective means of approximating the quantity of interest directly.

7.1 Further Work

The further work associated with this research can be divided into two categories: research directly concerned with the contents of the thesis and research inspired by the thesis. The research directly concerned with the thesis involves continuing the development of dual extremum principles for non-self-adjoint problems. The initial focus of this work would be on further analysis of the alternative bound in section 5.1.2 and attempts to construct a comparison principle from which the bounds on the quantity of interest could be constructed from the semi-discrete methods described in chapter 4.

The research inspired by the thesis includes computing physical quantities of interest for a wider range of problems. In order that a wider range of range of problems can be addressed the determination of strict upper and lower bounds may have to be relaxed. Efficient techniques to calculate functionals of this nature have been developed through the adjoint partial differential equation methods [7, 17, 38, 41, 45] described in the literature review. However, such methods do not necessarily aim to preserve qualitative features of the partial differential equation solutions. The preservation of qualitative features of the analytic solution at a discrete level is the essence of geometric methods [9]. An interesting direction for further research would be to investigate the advantages that adopting a geometric solution method that preserves global qualitative features of the solution has upon the accuracy of the quantity of interest.

Although the dual extremum principles have not been considered from a geometric approach in this thesis, the constraints imposed on the solution can certainly be view as qualitative features of the solution. For example, considering the simple case of Laplace's equation for a function $p(\mathbf{x})$, the pair of constraints, or geometric properties, in the domain are

$$-\mathbf{q}(\mathbf{x}) = \nabla p(\mathbf{x}) \qquad \text{invariance to datum,} \tag{7.1}$$

$$\nabla \cdot \mathbf{q}(\mathbf{x}) = 0 \qquad \text{conservation of flux.} \tag{7.2}$$

In the case of Laplace's equation constraining the solution to satisfy either of the geometric properties introduces the required convexity into the governing functional and bounds can be constructed on the stationary value of the functional. If convexity was not induced into the governing functional through the application of these constraints the natural question is which of these properties, for a given output functional, should be satisfied in order to obtain a greater degree of accuracy in the output functional. More generally, the partial differential equation considered may have many symmetries and invariances and it is conjectured that a subset of the geometric quantities can be identified which have a dominant effect on the output functional, depending on its nature.

The goal of the research would then be to identify, for a given boundary value problem or initial boundary value problem and a quantity of interest, the qualitative properties of the solution affecting the accuracy of numerical approximations to the quantity of interest. Having identified the qualitative properties pertinent to the accuracy of the numerical approximation, appropriate discrete analogues could be constructed.

Considering methods in the spirit of the discrete variational techniques of Marsden *et al* [31, 32] would form a good starting point and, although strict bounds on the quantity of interest would no longer be obtained, a wider range of problems could be addressed. These problems include

- Multiphase and oil recovery, governed by the black-oil models, with the output functional defined as the net well production. The symmetries to consider initially include conservation of the mass and momentum of the individual fluid phases.

- Aspects of atmospheric modelling governed by the semi-geostrophic equations, including frontogenesis, and climate modelling governed by a balanced energy model. The output functionals considered in these contexts are mean wind velocities and integrals of the energy flux to the poles respectively. The symmetries associated with these quantities include the conservation of potential vorticity, momentum and energy.

In addition, adopting a geometric approach complements the current philosophy in which the quantity of interest is also defined globally. Moreover, it is expected that by incorporating the dominant qualitative properties of an analytic solution into a discrete method, significant gains in the effectiveness and efficiency of the numerical approximation to the output functional will be achieved.

Bibliography

[1] J. W. Amyx, D. M. Bass, and R. L. Whiting. *Petroleum Reservoir Engineering, Physical Properties*. McGraw-Hill Book Company, Inc., London, 1960.

[2] T. Arbogast, S. E. Minkhoff, and P. T. Keenan. An operator-based approach to upscaling the pressure equation. In *Computational Methods in Water Resources XII*, pages 405–412. Computational Mechanics Publications, 1998.

[3] A. M. Arthurs. *Complementary Variational Principles*. Clarendon Press, Oxford, second edition, 1980.

[4] K. Aziz and A. Settari. *Petroleum Reservoir Simulation*. Applied Science Publishers Ltd, London, 1979.

[5] H. Bateman. Notes on a differential equation which occurs in the two-dimensional motion of a compressible fluid and the associated variational problems. *Proceedings of the Royal Society of London, Series A*, 125:598–618, 1929.

[6] R. P. Batycky, M. J. Blunt, and M. R. Thiele. A 3d field-scale streamline-based reservoir simulator. *SPE Reservoir Engineering*, 12:246–254, November 1997.

[7] R. Becker and R. Rannacher. An optimal control approach to a-posteriori error estimation in finite element methods. *Acta Numerica*, pages 1–102, 2001.

[8] D. Braess. *Finite Elements*. Cambridge University Press, 1997.

[9] C.J. Budd and M.D. Piggott. The geometric integration of scale invariant ordinary and partial differential equations. *J. Comp. Appl. Math*, 128:399–342, 2001.

[10] Z. Chen. Formulations and numerical methods of the black oil model in porous media. *SIAM Journal of Numerical Analysis*, 38(2):489–514, 2000.

[11] Z. Chen and R. E. Ewing. Mathematical analysis for reservoir engineering. *SIAM Journal of Numerical Analysis*, 30(2):431–453, 1999.

[12] W.D. Collins. Dual extremum principles for the heat equation. *Proceedings of the Royal Society of Edingburgh*, 77A:273–292, 1977.

[13] R. Courant and D. Hilbert. *Methods of Mathematical Physics.* Interscience Publishers, Inc., New York, 1966.

[14] B. Davies and B. Martin. Numerical inversion of the laplace transform: a survey and comparison of methods. *Journal of Computational Physics*, 33:1–32, 1979.

[15] C. L. Farmer. Upscaling: a review. *International Journal for Numerical Methods in Fluids*, 40:63–78, 2002.

[16] M. B. Giles. Aerodynamic design optimisation for complex geometries using unstructured grids. Technical Report 97/08, Oxford University Computing Laboratory, 1997.

[17] M. B. Giles and N. A. Pierce. Improved lift and drag estimates using adjoint euler equations. *AIAA Paper*, (99-3293), 1999.

[18] M. B. Giles and N. A. Pierce. Adjoint error correction for integal outputs. Technical Report 01/18, Oxford University Computing Laboratory, 2001.

[19] M. E. Gurtin. Variational principles for linear initial-value problems. *Quaterly of Applied Mathematics*, 12(3):252–256, 1964.

[20] K. Gustafson and R. Hartman. Divergence-free bases for finite element schemes in hydrodynamics. *SIAM Journal of Numerical Analysis*, 20:697–721, 1983.

[21] R. Hartmann and P. Houston. Adaptive discontinuous galerkin finite element methods for nonlinear hyperbolic conservation laws. *SIAM Journal on Scientific Computing*, to appear.

[22] U. Hornung, editor. *Homogenization and Porous Media.* Springer, 1997.

[23] T. Y. Hou and X. Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *Journal of Computational Physics*, 134:169–189, 1997.

[24] L. Ingebrigsten, F. Bratvedt, and J. Berge. A streamline based approach to solution of three-phase flow, SPE 51904. Presented at the 1999 SPE Reservoir Simulation Symposium held in Houston, Texas, 14-17 February 1999.

[25] B. Jiang. *The Least-Squares Finite Element Method.* Springer, 1998.

[26] C. Johnson. *Numerical Solution of Partial Differential Equations by the Finite Element Method.* Cambridge University Press, 1987.

[27] M. J. King, P. R. King, C. A. McGill, and J. K. Williams. Effective properties for flow calculations. *Transport in Porous Media*, 20:169–196, 1995.

[28] H. Levine and J. Schwinger. On the theory of electromagnetic wave diffraction by an aperture in an infinite plane conducting screen. *Comm. Pure and Appl. Math.*, pages 355–391, 1950.

[29] J. C. Luke. A variational principle for a fluid with a free surface. *J. Fluid Mech.*, 27(2):395–397, 1967.

[30] J. E. Marsden and T. S. Ratiu. *Introduction to Mechanics and Symmetry.* Springer-Verlag, 1994.

[31] J.E. Marsden, G.W. Patrick, and S. Shkoller. Multisymplectic geometry, variational integrators, and nonlinear pdes. *Commun. Math. Phys.*, 199:351–395, 1998.

[32] J.E. Marsden and M. West. Discrete mechanics and variational integrators. *Acta Numerica*, pages 357–514, 2001.

[33] A. Mayer-Gürr. *Petroleum Engineering*, volume 3 of *Geology of Petroleum.* Pitman Publishing, 1976.

[34] J. D. Moulton, J .E. Dendy, and J. M. Hyman. The black box multigrid numerical homogenization algorithm. *Journal of Computational Physics*, 142:80–108, 1998.

[35] M. Muskat. *Physical Principles Of Oil Production.* McGraw-Hill Book Company, Inc., 1949.

[36] B. F. Nielsen and A. Tveito. An upscaling method for one-phase flow in heterogeneous reservoirs. a weighted output least squares (wols) approach. *Computational Geosciences*, 2:93–123, 1998.

[37] M. Panfilov. *Macroscale Models of Flow through Highly Hetrogeneous Porous Media.* Kluwer Academic, 2000.

[38] N. A. Pierce and M. B. Giles. Adjoint recovery of superconvergent functionals from pde approximations. *SIAM Rev*, 42(2):247–264, 2000.

[39] D. K. Ponting. Hybrid streamline methods, SPE 39756. Presented at the 1998 SPE Asia Pacific Conference on Integrated Modelling for Asset Management held in Kuala Lumpar, Malysia 23-24 March 1998.

[40] D. Porter and D. S. G. Stirling. *Integral Equations : a practical treatment, from spectral theory to applications.* Cambridge University Press, 1990.

[41] R. Rannacher. Error control in finite element computations. Technical report, Universitat Heidlberg, 1998.

[42] P. A. Raviart and J. M. Thomas. Primal hybrid finite element methods for 2nd order elliptic equations. *Mathematics of Computation*, 31:391–413, 1977.

[43] M. Schemann and F.A. Bornemann. An adaptive Rothe method for the wave equation. *Comput. Visual Sci.*, 1:137–144, 1998.

[44] M. J. Sewell. *Maximum and minimum principles : a unified approach with applications.* Cambridge University Press, 1987.

[45] E. Suli and P. Houston. Adaptive finite element approximation of hyperbolic problems. Technical Report 2002/08, University of Leicester, 2002.

[46] J. L. Synge. *The Hypercircle in Mathematical Physics.* Cambridge University Press, 1957.

[47] M. A. Wakefield. A variational approach to consistent single phase upscaling. *International Journal for Numerical Methods in Fluids*, 40:539–549, 2002.

[48] T. C. Wallstrom, M. A. Christie, L. J. Durlofsky, and D. H. Sharp. Effective flux boundary conditions for upscaling porous media equations. Technical Report LA-UR-2001-235, Los Alamos National Laboratory, 2001.

[49] J. E. Warren and H. S. Price. Flow in heterogeneous porous media. *Society of Petroleum Engineers Journal*, pages 153–169, 1961.