52

Bounding of Linear Output Functionals of Parabolic Partial Differential Equations

by

Jeremy Alan Teichman

Submitted to the Department of Mechanical Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Mechanical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1998

Author .....
Department of Mechanical Engineering
April 2, 1998

Certified by .....
Anthony Patera
Professor
Thesis Supervisor

Certified by .....
Jaime Peraire
Associate Professor
Thesis Supervisor

Accepted by .....
Anthony Patera
Acting Graduate Officer

Bounding of Linear Output Functionals of Parabolic Partial Differential Equations

by

Jeremy Alan Teichman

Submitted to the Department of Mechanical Engineering
on April 2, 1998, in partial fulfillment of the
requirements for the degree of
Master of Science in Mechanical Engineering

Abstract

This thesis describes methods for calculating bounds for linear functional outputs which represent scalar metrics of systems described by parabolic partial differential equations. The methods reduce the cost of estimating the outputs by using constrained minimization principles to generate the bounds while bypassing full solution of the original differential equations. The method operates by formulating a Lagrangian with a saddle point at which the value of the Lagrangian equals the output of interest. By reversing the minimization and maximization and then eliminating the maximization, bounds are generated while the requirement that the original equation be satisfied is relaxed. The method is illustrated through application to the Helmholtz equation with a positive dissipative coefficient, the transient fin equation.

Thesis Supervisor: Anthony Patera
Title: Professor

Thesis Supervisor: Jaime Peraire
Title: Associate Professor

Acknowledgments

A number of people were instrumental in bringing this thesis into being. I would like to take this opportunity to thank them. It goes without saying that my two advisors Tony Patera and Jaime Peraire played a central role in creating the research project presented in this thesis as well as in guiding and advising me throughout my work. The enlightening discussions I had with my cohorts involved in similar work, Marius Paraschivoiu, Luc Machiels, and Serhat Yeşilyurt, were likewise indispensible. I also extend my thanks to all those in the MIT Fluids Lab who supported me during my graduate work and put up with me or lent a hand when I was stuck. Among these I would like to especially mention the other members of my research group, Nicolas Hadjiconstantinou, Miltos Kambourides, Thomas Leurent, and Vincent Colmar and my officemates who had to put up with me more than most, Chris Hartemink, Darryl Overby, and once again Miltos. I would like to thank Einar Rønquist for providing and helping me get started with speclib. My roommate Cory Welt also certainly deserves mention as he has supported me since the beginning of my graduate career, every day and in all conceivable realms of life. My parents have been an unending source of encouragement since my education began, and their zeal has not abated with my departure from home. Most especially, I would like to thank my girlfriend Jessica Zlotogura who has supported and comforted me when things were not going well and shared my excitement when things were. While she spent almost as much time with me as Cory, for the most part, she had a choice in the matter. Thank you Jessica.

This work was supported in part by grants from the Air Force Office of Scientific Research and the Defense Advanced Research Projects Agency.

This material is based upon work supported under a National Science Foundation Graduate Fellowship. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation.

Contents

Chapter 1

Introduction

Real world systems can often be accurately represented by differential equations whose solutions are the states of the system at all points in space and time. Engineers and scientists who are interested in these systems, while sometimes concerned with the behavior of all the components of the system throughout its evolution, are frequently only focused on scalars which act as metrics of particular aspects of the system's characteristics rather than the full field solution.

The research described in this thesis is concerned with the evaluation of these scalar metrics and, in particular, those which can be described as linear functionals of the field solutions to the governing differential equations. Important scalar quantities such as averages, point values, and fluxes can all be expressed as linear functionals of the field solution and are known as linear functional outputs of the differential equation. Some examples of pertinent engineering metrics that fall into this category are drag of a body in a fluid flow, temperature at one point in a thermal system, heat transfer rate through a fin, volumetric flow rate, and mass flow rate.

Many differential equations corresponding to real systems cannot be solved exactly with existing techniques. As a result approximate techniques have been developed to find approximate solutions to differential equations. For a given technique, the accuracy of the solution directly corresponds to the computational effort required to generate it.

The standard method for evaluating a linear functional output of a differential equation requires calculation of the field solution of the differential equation. The output is then calculated by evaluating the linear functional applied to the field solution. As the solution's accuracy corresponds to the work put into the calculation, so does the accuracy of the output generally correspond to the accuracy of the solution and, hence, the effort involved.

Because the quantity of interest is a single scalar representative of some particular aspect of the field solution, presumably there might be a way to garner information about this output without actually determining the full field solution. The goal of the research discussed in this thesis is to develop a method for estimating the value of the scalar output and, in fact, generating upper and lower bounds for the true value of the output, without constructing the full solution to the differential equation. The intent is that such a method will drastically reduce the effort required to gather the

necessary information about the output of interest.

Bounds for the output of interest can be of use in a number of ways. The average of the bounds can serve as an estimate of the output's value with a known maximum possible error equal to half the difference between the upper and lower bounds. In a feasibility study, only very coarse estimates might be required to ensure that a particular quantity has a value within reason. Additionally, sometimes the bounds themselves are useful. In a thresholding situation where a quantity expressible as a linear functional output must fall below or above a maximum or minimum acceptable threshold, respectively, the bounds can supply the requisite information without the actual value of the output being known. In a design problem, if the output is part of the specifications of a design, calculating bounds which fall within the tolerances of the specified value bypasses the need to know the exact output. Thus in many significant engineering scenarios, the bounds do not act as surrogates for the true output but are valuable in their own right.

1.1 Related Work

Other people have done work on calculation of bounds for various metrics of systems described by differential equations. Becker and Rannacher developed a method for bounding linear output functionals, but their bounds contain unknown constants whose presence renders the bounds less useful [4, 5]. Leguillon and Ladeveze, Bank and Weiser, and Ainsworth and Oden have all developed methods for finding bounds consisting solely of known quantities, but their methods bound the energy norm of the solution rather than directly useful engineering metrics [7, 3, 1, 2].

My work follows more directly from other work on calculation of bounds for linear functional outputs. Paraschivoiu, Patera, and Peraire developed a method for bounding linear functional outputs of coercive elliptic partial differential equations [10, 12]. Paraschivoiu also extended this type of method to the steady Stokes problem [9, 11]. Patera, Peraire, and Machiels have recently been developing methods which enable these techniques to be applied to non-coercive and non-linear problems and facilitate the use of the methods as a tool for adaptive gridding of problem domains for discrete solution of the governing equations [8, 14, 15].

The research described in this thesis extends these methods into the realm of time-dependent phenomena described by parabolic partial differential equations. Time-dependent processes exhibit behavior which is highly coupled across time. Systems evolve from one state to the next with each new state directly resulting from the previous one. The coupling adds a degree of complexity to the solution of time-dependent problems. In certain cases the time induced coupling can make calculation of the bounds using the methods of this thesis more difficult than solving the original equations. The challenge of bounding outputs of time-dependent problems is to avoid the pitfalls created by the coupling and generate bounds for linear output functionals without unknown constants.

Chapter 2

Methodology

Most problems design engineers focus on today cannot be solved exactly with current methods. When such problems are encountered, engineers and scientists typically utilize discrete solution methods such as the finite element method and the finite difference method to find approximate solutions. Generally, a solution calculated using a very fine discretization with an acceptable method can be treated as exact. In this thesis I will refer to this fine discretization as a "truth mesh." The output functionals evaluated on truth mesh solutions of the differential equations will be considered exact.

This chapter summarizes in very general terms the methods used to generate upper and lower bounds for linear functional outputs. The rest of the thesis attempts to elucidate the concepts in this chapter through detailed examples of their use. The essence of the idea is contained here; the bulk of the work is the proof of principle that follows.

The bounding technique uses a weak formulation of the differential equation as its starting point. Most problems are first posed in their strong formulations, but since the finite element methods which the bounding utilizes operate on the weak formulation, the first step is to compose a weak formulation of the problem. This is done by multiplying the strong form by an arbitrary test function and integrating over the problem domain. Terms involving derivatives of the field variable are integrated by parts to transfer some of the differentiation from the field variable to the test function. Reducing the maximum order of differentiation reduces the amount of regularity required of the field solution. This method considers the weak formulation the exact problem whose output is the target of the bounds. The weak form can be written as a general operator in the form

$$A(v, \theta) = 0 \quad \forall v \in X_h,$$

where v is the test function, θ is the field variable, and X_h is the finite element space.

The bounding technique aims to form bounds by relaxing a method for actually calculating the bounds, but relaxing it in a controlled fashion so that the direction of relaxation is fixed. A lower bound can thus be formed by relaxing the calculation technique in such a way that its result can only fall from the unrelaxed value. The

result of such a relaxed calculation is guaranteed to be lower than the true value and, thus, a lower bound. The technique is general enough that if the output can be treated, so can the negative of the output. The negative of a lower bound to the negative output is an upper bound to the output itself. Utilizing this fact obviates the need for an independent method for generating the upper bound.

The crux of the technique is finding a way to calculate the output that can be relaxed in this controlled fashion. The most straightforward way to do so is to make the output the result of a maximization:

$$s = \max(Q),$$

where s is the output. If the output is a maximum of some function, then any value of the function is a lower bound to the output, $Q \leq s$.

Any benefit this technique has is contingent on its ability to calculate bounds more easily than the field solution to the differential equation can be calculated. The output is by definition a function or rather a functional of the field solution to the differential equation,

$$\ell(\theta) = s.$$

Ideally the maximization described above is equivalent to calculating the field solution:

$$s = \max_v Q(v)$$

and

$$\theta = \arg\max_v Q(v).$$

In this case, bypassing the maximization is effectively bypassing solution of the differential equation yet calculating a bound in the process.

A classic way of making the solution of an equation a maximization problem is through a Lagrange multiplier. This is usually done in the context of a constrained minimization. A maximization over Lagrange multipliers results in an unbounded value of the Lagrange multiplier term unless the constraint it enforces is satisfied; for a constraint on x, $f(x) = 0$,

$$\max_\lambda \lambda f(x) = \begin{cases} \infty & f(x) \neq 0 \\ 0 & f(x) = 0 \end{cases}.$$

A minimization over the constrained variable of a maximization over Lagrange multipliers is equivalent to a minimization with the variable constrained to satisfy the Lagrange multiplier-enforced condition,

$$\min_x \max_\lambda (g(x) + \lambda f(x)) = \min_{x|f(x)=0} g(x).$$

The weak form of a differential equation is effectively a Lagrange multiplier enforced solution of the equation with the test function acting as a Lagrange multiplier. So, if the calculation of the output can be turned into a constrained minimization problem, where the constraint enforces solution of the original differential equation, the problem

can be written as follows:

$$s = \min_{p} \max_{\mu} (g(p) + A(\mu, \chi)),$$

where s is the output, μ is the test function turned Lagrange multiplier, χ is the original field variable of the differential equation, A is the weak form, p is some as yet unspecified function, and g is an unspecified functional. Thus, a minimum of the maximum over Lagrange multipliers of a unspecified functional plus the weak form of the differential equation must equal the output.

To relax the maximization in the calculation of the output to generate a lower bound, the maximization must be outside of the minimization. Classic duality theory states that in the case of a quadratic minimizable functional with linear constraints, the min-max equals the max-min, and both occur at a saddle point [16]. For a linear differential equation, the constraint, the weak form, is linear. Hence, to switch the maximization and the minimization, g must be a quadratic minimizable functional. Also, since the weak form vanishes when the constraint is satisfied as it is at the saddle point, g must equal the output at the constrained minimum,

$$s = \min_{p} \max_{\mu} (g(p) + A(\mu, \chi)) = g(\arg \min_{p} (\max_{\mu} (g(p) + A(\mu, \chi))))$$

The difficulty of finding a functional, $g(p)$, which has a constrained minimum equal to the output can be bypassed by making the minimization trivial when the constraint is enforced. The minimization cannot be made both trivial and independent of the constraint because such a functional would not be quadratic, it would be constant. The constrained minimization can be made trivial by letting the constraint set the value of the function or variable over which the minimization is performed. This suggests letting $p = \chi$. If this simplification is followed, then g must be a quadratic functional of χ which equals the output when χ satisfies the differential equation embodied in the constraint,

$$s = \min_{\chi} \max_{\mu} (g(\chi) + A(\mu, \chi))$$

and

$$A(v, \chi) = 0 \ \ \forall v \in X_h \Rightarrow \chi = \theta, \ \ g(\chi) = s$$

One way to determine the quadratic minimizable functional is to break the functional into two parts to satisfy the two requirements independently, one linear part whose value equals the output and one quadratic and minimizable part that vanishes when χ satisfies the differential equation. It is important that the first part be linear so that it does not disturb either the quadratic order or the minimizable nature of the other. The second part must vanish when the constraint is satisfied so as not to alter the value of the first part which must be equal to the output. The easiest choice for the first part is the output functional itself, $\ell(\chi)$. This constrains the output functional to be linear. For the second part, the linearity of the differential equation can be utilized in creating a quadratic functional. The weak form operator, $A(v, \chi)$,

can be split into a linear portion, $b(v)$, a bilinear symmetric portion, $C^s(v,\chi)$, and a bilinear skew-symmetric portion, $C^{ss}(v,\chi)$. If the test function in the weak form is chosen to be the field variable of the differential equation, then the resulting special case of the weak form with the skew-symmetric portion removed, $C^s(\chi,\chi)+b(\chi)$, is guarateed to be quadratic. However, to be minimizable, all the quadratic terms resulting from this step must have positive coefficients. If this situation can be brought about, then the output can be calculated from the constrained minimization of the resulting functional,

$$s = \min_{\chi} \max_{\mu} C^s(\chi,\chi) + b(\chi) + A(\mu,\chi) + \ell(\chi)$$

(When there are inhomogeneous boundary conditions on the differential equation, a slightly different formulation is necessary because the test functions are in a different space than the field variable, and $A(\theta,\theta)$ may not be zero.).

When the order of the maximization and minimization is reversed, and the maximization is removed, the resulting equation explicitly generates a lower bound for the output. By the modification described earlier, the upper bound follows in the same fashion. The reversed maximization and minimization describes a problem where the extremized Lagrange multiplier is sought first. In the reversed problem, optimizing the Lagrange multiplier solves the differential equation. When the maximization is removed, the solution of the differential equation is bypassed,

$$s = \min_{\chi} \max_{\mu} \mathcal{L}(\mu,\chi) = \max_{\mu} \min_{\chi} \mathcal{L}(\mu,\chi) \geq \min_{\chi} \mathcal{L}(\mu,\chi),$$

where

$$\mathcal{L} \equiv C^s(\chi,\chi) + b(\chi) + A(\mu,\chi) + \ell(\chi).$$

As the subsequent chapters shall illustrate, if the procedure schematically described in this chapter is exercised on a real problem, bounds can be calculated. The major issue that remains is computational efficiency. If solving the equations that emerge from this method is more costly than solving the original differential equation then nothing has been gained. The objective of efficiency often drives the particular way the technique is applied to real problems. In the case of time-dependent differential equations, the causal relationship between sequential states of a system creates a situation where reaching the threshold efficiency is often highly non-trivial.

Chapter 3

Sample Problem

Because the procedure for generating bounds is difficult to discuss abstractly, here I introduce a system and its governing equations as an example to aid in the concrete illustration of the method. A thin, thermally conductive rod with a uniform thermal conductivity and unit length, well insulated on both ends starts out with an initial temperature distribution cool at the ends and hot in the center. The rod is suddenly immersed in a fluid bath of a different temperature. The thermal energy in the rod will diffuse along the length of the rod by conduction but will not penetrate the insulation at the rod's two ends. Thermal energy will also be lost to the surrounding fluid by convection with a uniform heat transfer coefficient. The fluid bath can be considered a heat reservoir whose temperature remains invariant throughout the process. Figure 3-1 depicts a schematic of the system.

The governing equations of this system are most easily formulated in terms of relative temperature, the temperature difference between the rod and the fluid bath, $\theta(x,t)$, where x represents position along the rod and t represents time starting from the moment of immersion. The "good" Helmholtz equation with a transient term, also known as the transient fin equation, governs the evolution of the relative temperature of the rod over time:

$$\theta_t(x,t) = \alpha\theta_{xx}(x,t) - \beta\theta(x,t), \tag{3.1}$$

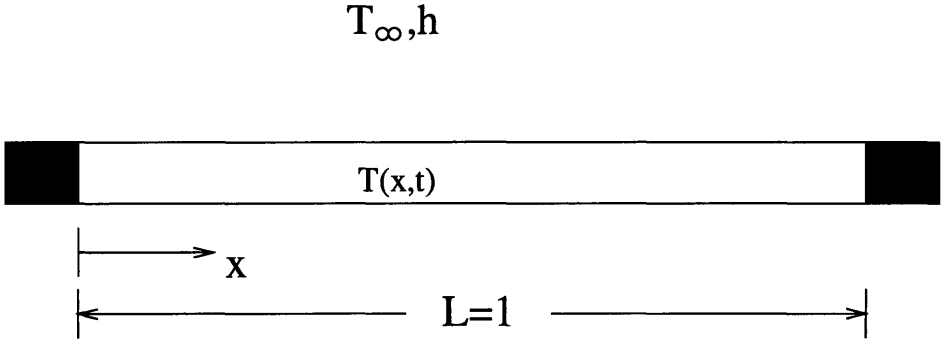

Figure 3-1: System Schematic

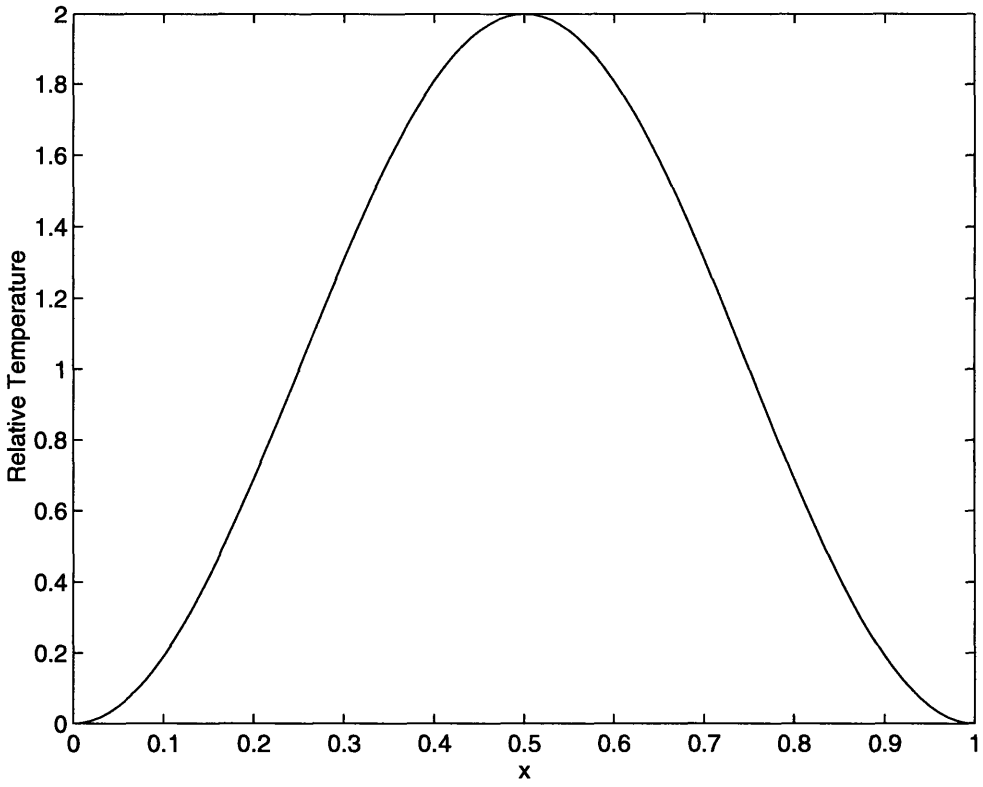

Figure 3-2: $\theta_0(x)$

where subscripts represent partial differentiation, and α and β are positive coefficients. The rod starts out with a given relative temperature profile,

$$\theta(x, 0) = \theta_0(x), \tag{3.2}$$

plotted in Figure 3-2. Since no heat flows into the insulation, Fourier's law dictates that the temperature gradient must be zero at the rod ends,

$$\theta_x(0, t) = \theta_x(1, t) = 0. \tag{3.3}$$

Figure 3-3 shows how the relative temperature profile along the rod evolves over time.

The quantity of interest, s, is the average relative temperature over the length of the rod for a unit time,

$$s = \int_0^1 \int_0^1 \theta(x, t) dx dt. \tag{3.4}$$

Other than being important in its own right, the average is also proportional to the total heat lost by the rod to the fluid.

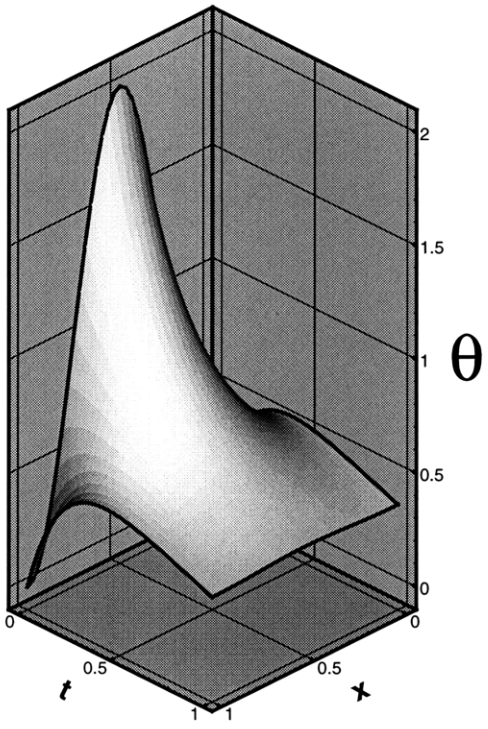

Figure 3-3: $\theta(x,t)$

Chapter 4

First Approach: Finite Elements in Space, Finite Differences in Time

The sample problem described above has both spatial and temporal facets. The discrete solution treats these two aspects independently with different approaches. The finite element method is applied to the spatial problem, while finite differences are used to discretize the temporal problem [13].

To approach the problem with these methods, the strong formulation of the differential equation and boundary conditions stated above needs to be transformed into a weak formulation. To derive the weak form, first the original differential equation is multiplied by an arbitrary function, $v(x,t) \in \mathcal{H}^1$, where $\mathcal{H}^1$ is the Hilbert space of functions which are square integrable and whose first partial derivatives with respect to x are square integrable over the problem domain, $(x,t) \in [0,1] \times [0,1]$. Then the result is integrated over the spatial domain with the conduction term integrated by parts, and the spatial boundary conditions are applied. The resulting weak form states that

$$\int_0^1 v\theta_t dx = -\alpha \int_0^1 v_x \theta_x dx - \beta \int_0^1 v\theta dx \;\; \forall v \in \mathcal{H}^1, \;\; \theta(x,0) = \theta_0(x). \tag{4.1}$$

The spatial domain of the problem is subdivided into N_x equally sized sections of length $\Delta x = 1/N_x$ called elements. The points $x_i = i\Delta x, i = 0, \cdots, N_x$, referred to as nodes, lie on the edges of the domain and on the boundaries separating the elements. The standard linear finite element hat function centered on node i is denoted ϕ_i, where

$$\phi_i(x_j) = \delta_{ij}, \tag{4.2}$$

and δ_{ij} is the Kronecker delta. The set $\{\phi_0, \phi_1, \cdots, \phi_{N_x+1}\}$ forms a basis for all continuous functions which are linear over each element of the problem domain.

Narrowing the definition of the function spaces in the weak formulation generates the equations for the finite element problem statement. Instead of allowing any test function in $\mathcal{H}^1$, the finite element statement of the problem only allows test functions

in the more restrictive space, X_h, defined by

$$X_h = \mathbf{P}_h^1 \cap \mathcal{H}^1 \cap \{w(x,t) | w(x_i, 0) = \theta_0(x_i), i = 0, \cdots, N_t\} \tag{4.3}$$

where $\mathbf{P}_h^1$ is the space of functions which are first order polynomials on each element of the discretized domain. The hat functions, ϕ_i, span this space. The finite element solution, $\theta \in X_h$, satisfies

$$\int_0^1 v\theta_t dx = -\alpha \int_0^1 v_x \theta_x dx - \beta \int_0^1 v\theta dx \;\; \forall v \in X_h. \tag{4.4}$$

The finite element solution can be written as a weighted sum of the basis functions,

$$\theta(x,t) = \sum_{i=0}^{N_x} w_i(t)\phi_i(x) \tag{4.5}$$

where the $w_i(t)$ are the time dependent coefficients of the spatial basis functions. Using a standard Galerkin procedure and setting $v(x,t) = \phi_j(x), \;\; j = 0, \cdots, N_x$, the problem can be written in matrix form as

$$\underline{\underline{\mathbf{M}}}\underline{\theta}_t = -(\alpha \underline{\underline{\mathbf{K}}} + \beta \underline{\underline{\mathbf{M}}})\underline{\theta} \tag{4.6}$$

where $\underline{\underline{\mathbf{M}}}$ is the mass matrix with $\mathbf{M}_{ij} = \int_0^1 \phi_i \phi_j dx$, $\underline{\underline{\mathbf{K}}}$ is the stiffness matrix with $\mathbf{K}_{ij} = \int_0^1 \phi_{i_x} \phi_{j_x} dx$, and $\underline{\theta}$ is the vector of time dependent coefficients of the hat functions $\theta_i = w_i(t)$.

This formulation is discrete in space and continuous in time. To discretize the problem in time, the time period of interest is divided into N_t equally sized intervals. Each finite time interval has a length $\Delta t = 1/N_t$. The backward Euler finite difference scheme makes the discrete approximation $\theta_t(x,t) = \frac{\theta(x,t+\Delta t) - \theta(x,t)}{\Delta t}$. The shorthand notation, t_i, denotes $i\Delta t$, and θ_j^i denotes $w_j(t_i)$. Applying backward Euler to the matrix form of the equation and rearranging yields

$$((1 + \beta\Delta t)\underline{\underline{\mathbf{M}}} + \alpha\Delta t \underline{\underline{\mathbf{K}}})\underline{\theta}^{n+1} = \underline{\underline{\mathbf{M}}}\underline{\theta}^n. \tag{4.7}$$

The discrete version of the initial condition is

$$\theta_i^0 = \theta_0(t_i). \tag{4.8}$$

Using this initial condition, Equation 4.7 can be solved at each time step to generate the full finite element solution at each point in time.

The output of interest, s, the average over space and time, can be written in discrete form in the same fashion. In continuous form, the output appears as

$$s = \frac{\int_0^1 \int_0^1 \theta(x,t) dx dt}{\int_0^1 \int_0^1 dx dt}. \tag{4.9}$$

After substituting the discrete form of θ, the output can be written

$$s = (\sum_{n=1}^{N_t} \underline{\theta}^n)^T \underline{\mathbf{d}} \Delta t, \tag{4.10}$$

where $\underline{\mathbf{d}}$ is the vector $\mathbf{d}_i = \int_0^1 \phi_i(x)dx$.

The standard way to evaluate s is to solve Equation 4.7 on a truth mesh, and then insert the calculated θ into Equation 4.10. In order to avoid solving the full problem on the truth mesh, the system must be transformed into a constrained minimization problem. This transformation results in a lower bound for the value of the output. Since the negative of the linear functional output is also a linear functional, the bounding procedure can be applied to the negative output functional in the same fashion in which it is applied to the positive output functional. The negative of the lower bound thus derived for the negative output functional is, in fact, an upper bound for the original output functional. The functional which will be minimized must be a quadratic minimizable form of the finite element equations that vanishes when the finite element equations are satisfied. This form arises when the finite element equations are multiplied by $\underline{\theta}^{n+1^T}$ and summed over all the time steps. Clearly, if the original equations are valid, so is the new form. From this form a general functional, $E_0(v)$, $v \in X_h$, can be written which vanishes when $v = \theta$,

$$\begin{aligned} E_0(v) &= \frac{1}{2}\sum_{n=0}^{N_t-1} (\underline{\mathbf{v}}^{n+1} - \underline{\mathbf{v}}^n)^T \underline{\underline{\mathbf{M}}} (\underline{\mathbf{v}}^{n+1} - \underline{\mathbf{v}}^n) + \frac{1}{2}\underline{\mathbf{v}}^{N_t T} \underline{\underline{\mathbf{M}}} \underline{\mathbf{v}}^{N_t} - \frac{1}{2}\underline{\mathbf{v}}^{0T} \underline{\underline{\mathbf{M}}} \underline{\mathbf{v}}^0 + \\ &\quad \Delta t \sum_{n=0}^{N_t-1} \underline{\mathbf{v}}^{n+1^T} (\alpha \underline{\underline{\mathbf{K}}} + \beta \underline{\underline{\mathbf{M}}}) \underline{\mathbf{v}}^{n+1}. \end{aligned} \tag{4.11}$$

This functional is known as the energy equality primarily because it is of the same quadratic form as functionals describing the energy contained in a system. Both $\underline{\underline{\mathbf{M}}}$ and $\underline{\underline{\mathbf{K}}}$ are symmetric positive definite matrices. As a result, all terms in $E_0(v)$ are quadratic and positive definite except for the quadratic $\underline{\mathbf{v}}^0$ term. Because the definition of X_h includes the initial condition, the value of $\underline{\mathbf{v}}^0$ is set and the quadratic $\underline{\mathbf{v}}^0$ term behaves as a constant. Hence, $E_0(v)$ is quadratic and minimizable. Also, $E_0(\theta) = 0$ because of the way $E_0(v)$ is constructed.

The first term in the energy equality couples sequential time steps. The high degree of complexity that this coupling induces in the resulting problem makes calculation of bounds far more costly than solution of the original finite element problem. To make the procedure useful, the energy equality must be modified to eliminate the coupling. Subtracting the first term from the energy equality to form an energy inequality, $E(v)$, decouples the problem. It is crucial that the eliminated term is positive definite. For the resulting energy inequality,

$$E(v) = \frac{1}{2}\underline{\mathbf{v}}^{N_t T} \underline{\underline{\mathbf{M}}} \underline{\mathbf{v}}^{N_t} - \frac{1}{2}\underline{\mathbf{v}}^{0T} \underline{\underline{\mathbf{M}}} \underline{\mathbf{v}}^0 + \Delta t \sum_{n=0}^{N_t-1} \underline{\mathbf{v}}^{n+1^T} (\alpha \underline{\underline{\mathbf{K}}} + \beta \underline{\underline{\mathbf{M}}}) \underline{\mathbf{v}}^{n+1}, \tag{4.12}$$

it is true for all v that $E(v) \leq E_0(v)$, and specifically, $E(\theta) \leq 0$.

The output can be written as a general functional, $\ell(v)$, with $s = \ell(\theta)$,

$$\ell(v) = (\sum_{n=1}^{N_t} \underline{\mathbf{v}}^n)^T \underline{\mathbf{d}} \Delta t. \tag{4.13}$$

Augmenting the energy inequality with this output functional forms two new functionals, $\mathcal{S}^+(v)$ and $\mathcal{S}^-(v)$, with the output functional added or subtracted from the energy inequality, respectively. The combined notation, $\pm$, will be used to compactly indicate both simultaneously,

$$\mathcal{S}^{\pm}(v) = E(v) \pm \ell(v). \tag{4.14}$$

This additive operation preserves the inequality so that

$$\mathcal{S}^{\pm}(\theta) \leq \pm s. \tag{4.15}$$

The augmented energy inequality, $\mathcal{S}^{\pm}(v)$, is the core of the constrained minimization problem which produces the bounds. Equation 4.15 can be rewritten as the constrained minimization

$$\pm s \geq \min_{\{v | ((1+\beta\Delta t)\underline{\underline{\mathbf{M}}} + \alpha\Delta t \underline{\underline{\mathbf{K}}})\underline{\mathbf{v}}^{n+1} = \underline{\underline{\mathbf{M}}}\underline{\mathbf{v}}^n,\ n=0,\cdots,N_t-1\}} \mathcal{S}^{\pm}(v). \tag{4.16}$$

The constraint which is recognizable as the original finite element equation, Equation 4.7, forces satisfaction of the finite element equation so that $v = \theta$.

The constraint can be directly incorporated into the functional to be minimized by means of a Lagrange multiplier function, μ. Since the discrete functions only exist at a finite number of points in time, so must the Lagrange multiplier function. Each discrete value of the Langrange multiplier function acts over one time interval and is considered to exist at the earlier end of that time interval. The time superscripts for the finite Lagrange multiplier function are, however, written as if they occur at the midpoint of each time interval to avoid confusion regarding the time interval to which they apply. Incorporating the constraint in this fashion generates the augmented Lagrangian, a functional of the field variable, v, and the Lagrange multiplier function, μ,

$$\begin{aligned} \mathcal{L}^{\pm}(v,\mu) \;=\; & \frac{1}{2}\underline{\mathbf{v}}^{N_t T}\underline{\underline{\mathbf{M}}}\underline{\mathbf{v}}^{N_t} - \frac{1}{2}\underline{\mathbf{v}}^{0T}\underline{\underline{\mathbf{M}}}\underline{\mathbf{v}}^{0} + \Delta t \sum_{n=0}^{N_t-1} \underline{\mathbf{v}}^{n+1T}(\alpha\underline{\underline{\mathbf{K}}} + \beta\underline{\underline{\mathbf{M}}})\underline{\mathbf{v}}^{n+1} \pm \\ & (\sum_{n=1}^{N_t} \underline{\theta}^n)^T \underline{\mathbf{d}} \Delta t + \sum_{n=0}^{N_t-1} \underline{\mu}^{n+\frac{1}{2}T}[((1+\beta\Delta t)\underline{\underline{\mathbf{M}}} + \alpha\Delta t\underline{\underline{\mathbf{K}}})\underline{\mathbf{v}}^{n+1} - \\ & \underline{\underline{\mathbf{M}}}\underline{\mathbf{v}}^n]. \end{aligned} \tag{4.17}$$

The effect of Equation 4.16 can be rewritten as

$$\pm s \geq \min_{v} \max_{\mu} \mathcal{L}^{\pm}(v,\mu), \tag{4.18}$$

where the min-max occurs at the saddle point of the Lagrangian.

The constraint enforced by the Lagrange multiplier is linear. A quadratic functional with a linear constraint has a saddle point at the min-max. At a saddle point, strong duality applies, and duality theory allows the minimization and the maximization to be performed in the reverse order [16]. Thus, Equation 4.18 is equivalent to

$$\pm s \geq \max_{\mu} \min_{v} \mathcal{L}^{\pm}(v, \mu). \tag{4.19}$$

By the nature of maximization,

$$\max_{\mu} \min_{v} \mathcal{L}^{\pm}(v, \mu) \geq \min_{v} \mathcal{L}^{\pm}(v, \mu); \tag{4.20}$$

therefore,

$$\pm s \geq \min_{v} \mathcal{L}^{\pm}(v, \mu) \tag{4.21}$$

or

$$\min_{v} \mathcal{L}^{+}(v, \mu_1) \leq s \leq -\min_{v} \mathcal{L}^{-}(v, \mu_2) \tag{4.22}$$

for any μ_1 and μ_2.

At the saddle point of the Lagrangian, $\mathcal{L}^{\pm}(v, \mu)$, $v = \theta$ and $\mu = \psi^{\pm}$, where $\psi^{\pm}$ is known as the adjoint. If the Lagrangian is smooth, then for μ close to the adjoint, the minimum over v should be close to $\mathcal{L}^{\pm}(\theta, \psi^{\pm})$. The first variation in the Lagrangian with respect to v vanishes at the minimizing v of Equation 4.21, $\hat{\theta}^{\pm}$. Setting to zero the first variation in the Lagrangian with respect to v generates the following relationship between the minimizing v and the corresponding μ, $\hat{\mu}^{\pm}$:

$$((1 + 2\beta\Delta t)\underline{\underline{\mathbf{M}}} + 2\alpha\Delta t\underline{\underline{\mathbf{K}}})\underline{\hat{\theta}}^{\pm N_t} = \mp\underline{\mathbf{d}} - ((1 + \beta\Delta t)\underline{\underline{\mathbf{M}}} + \alpha\Delta t\underline{\underline{\mathbf{K}}})\underline{\hat{\mu}}^{\pm N_t - \frac{1}{2}}, \tag{4.23}$$

$$\begin{aligned} (2\beta\Delta t\underline{\underline{\mathbf{M}}} + 2\alpha\Delta t\underline{\underline{\mathbf{K}}})\underline{\hat{\theta}}^{\pm n} &= \mp\underline{\mathbf{d}} - ((1 + \beta\Delta t)\underline{\underline{\mathbf{M}}} + \alpha\Delta t\underline{\underline{\mathbf{K}}})\underline{\hat{\mu}}^{\pm n - \frac{1}{2}} \\ &\quad + \underline{\underline{\mathbf{M}}}\underline{\hat{\mu}}^{\pm n + \frac{1}{2}}, \\ n &= 1, \cdots, N_t - 1. \end{aligned} \tag{4.24}$$

Evaluating the Lagrangian at the point designated by Equations 4.23 and 4.24 yields the bounds:

$$\mathcal{L}^{+}(\hat{\theta}^{+}, \hat{\mu}^{+}) \leq s \leq -\mathcal{L}^{-}(\hat{\theta}^{-}, \hat{\mu}^{-}). \tag{4.25}$$

These bounds are valid for any choice of approximate adjoint. However, in order for the method to be effective, the estimate must be close enough to the real adjoint to generate acceptably tight bounds. One way to approximate the adjoint is to solve the finite element equation with a very coarse discretization for θ and then solve Equations 4.23 and 4.24 for the corresponding coarse mesh adjoint,

$$((1 + \beta\Delta t)\underline{\underline{\mathbf{M}}} + \alpha\Delta t\underline{\underline{\mathbf{K}}})\underline{\psi}_c^{\pm N_t - \frac{1}{2}} = \mp\underline{\mathbf{d}} - ((1 + 2\beta\Delta t)\underline{\underline{\mathbf{M}}} + 2\alpha\Delta t\underline{\underline{\mathbf{K}}})\underline{\theta}_c^{N_t} \tag{4.26}$$

and

$$\begin{aligned}((1+\beta\Delta t)\underline{\underline{\mathbf{M}}}+\alpha\Delta t\underline{\underline{\mathbf{K}}})\underline{\psi}_c^{\pm n-\frac{1}{2}} &= \mp\underline{\mathbf{d}}-(2\beta\Delta t\underline{\underline{\mathbf{M}}}+2\alpha\Delta t\underline{\underline{\mathbf{K}}})\underline{\theta}_c^n \\ &\quad +\underline{\underline{\mathbf{M}}}\underline{\psi}_c^{\pm n+\frac{1}{2}}, \\ n &= N_{tc}-1,\cdots,1, \end{aligned} \tag{4.27}$$

where the subscript of c denotes quantities on the coarse mesh. The interpolation of the coarse mesh adjoint onto the fine mesh serves as a good approximate adjoint. Progressively finer coarse meshes generate correspondingly better adjoint approximations but require more computational effort.

One issue that arises when this method of approximation is applied is how to determine the boundary conditions for the adjoint. Since the adjoint calculated from the Equations 4.23 and 4.24 is only determined up to the next to last time point, if a bilinear interpolation of the adjoint is performed, there is no obvious way to determine interpolant values between the last two discrete times of the coarse time discretization. After examination of a number of options, including extrapolation of adjoint values into the last time interval and addition of an extra time step in the solution for θ to provide an extra adjoint value at the last time step, it appeared that the most effective method was to use the continuous adjoint boundary condition as the value for the adjoint at the last time step.

If the problem is not discretized, the continuous equations yield a clear initial condition for the adjoint. Since the adjoint propagates backward in time, the initial condition occurs at the final time step. The adjoint initial condition is calculated in depth in Appendix A. The condition is

$$\psi^{\pm N_t} = -\theta^{N_t}. \tag{4.28}$$

The computational cost of this method can be divided into five steps. First, the coarse mesh θ is calculated. This requires that Equation 4.7 be solved once for each coarse time step. Equation 4.7 is a tridiagonal system of linear equations in which the coefficients are not dependent on n; only the right hand side depends on n. With one LU decomposition of the tridiagonal matrix, only forward and backward substitutions are required at each time step. The time steps must be followed sequentially because information from one time step is required for the subsequent step. Second, the coarse mesh adjoint is calculated from Equations 4.26 and 4.27. These tridiagonal equations can also be solved with one LU decomposition and a forward and backward substitution at each time step, but in reverse sequence— the last time step first and the first time step last. The positive and negative adjoints must be calculated separately in this stage. Third, the coarse mesh adjoint is interpolated onto the truth mesh. This operation only requires solution of one linear scalar equation per truth mesh point. Fourth, $\hat{\theta}$ is calculated from Equations 4.23 and 4.24 on the truth mesh. These equations, though coupled when solving for the adjoint, are completely independent at each time step when solving for $\hat{\theta}$ and can be solved in parallel. Each time step requires solution of one tridiagonal matrix system, but since the matrix

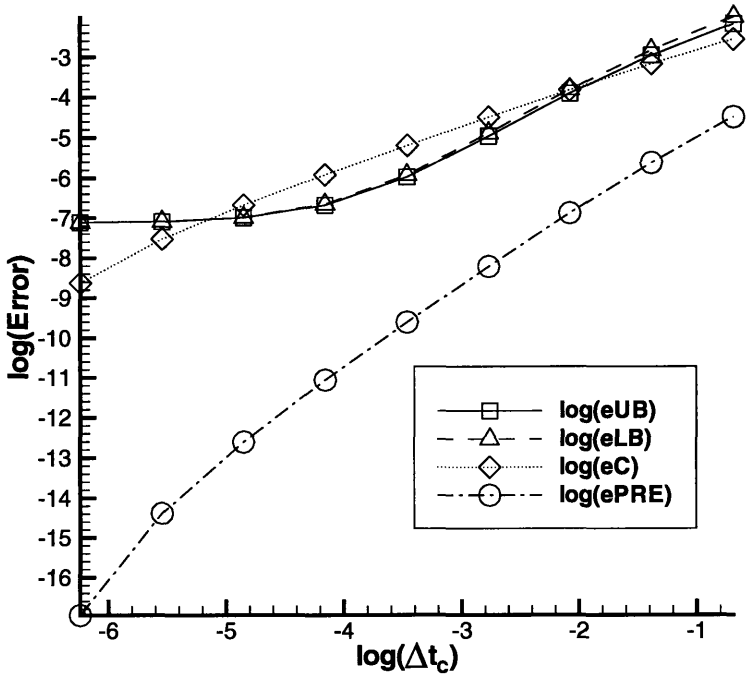

Figure 4-1: Convergence of Bounds with Respect to Coarse Mesh Element Size

is independent of n, one LU decomposition suffices. Fifth, and finally, the results of steps four and five are used to evaluate Equation 4.17 to find the bounds. Step five requires a number of matrix multiplications, but as each time step contributes independently, this step can also be easily parallelized. Steps three, four, and five must each be executed twice, once for the upper bound and once for the lower bound.

Altering the size of the coarse mesh affects the time required for steps one and two of the calculation without affecting the later steps at all. Therefore, it might be most efficient to refine the coarse mesh to a point where the coarse mesh calculation requires perhaps one order of magnitude less time than the truth mesh calculation. The time required for the latter three steps is fixed by the resolution requirements of the truth mesh. If steps three, four, and five require one hour to compute, it might make sense to choose a coarse mesh requiring perhaps five minutes to solve. The difference between a five minute coarse mesh and a thirty second coarse mesh does not greatly impact the total solution time, but the added accuracy of the adjoint approximation can drastically reduce the gap between the bounds.

This approach is convergent as seen in Figure 4-1, where eUB is the upper bound error, eLB is the lower bound error, eC is the coarse mesh output error, $ePRE$ is the error in the average of the upper and lower bounds, and Δt_C is the coarse mesh time step. The bounds grow closer together as the coarse mesh grows closer to the truth mesh. However, the use of the energy inequality instead of the energy equality introduces a small amount of error. As the coarse mesh approaches the

truth mesh, errors due to the quality of the adjoint approximation which dominate the total error for very coarse meshes become negligible compared to errors stemming from the inequality. This phenomenon limits the usefulness of the method because the accuracy needed in the bounds may prove impossible to achieve even with an extremely fine coarse mesh as illustrated by the flattening of the error graph for the bounds as Δt_C shrinks in Figure 4-1. The additional source of error results in the need for more work to achieve the same accuracy as would be present without the inequality.

This method is rather simple to implement but, because of the inequality, not so effective. It does, however, illustrate the general approach to bound formation as well as reveal some of the more serious difficulties inherent in solving time-dependent problems.

Chapter 5

Second Approach: Finite Elements in Space, Spectral Elements in Time

The second approach avoids the problems caused by the energy inequality in the first approach. The primary difference between the two approaches regards the way the temporal aspect of the problem is treated. In the first approach time was discretized using a backward Euler finite difference scheme. The second approach uses spectral elements in time [6]. Appendix B describes the basis functions, quadrature schemes, and other nuances of spectral elements. The use of spectral elements in time is usually a very poor choice for parabolic partial differential equations because of the enormous amount of temporal coupling it tends to introduce. In this particular case using spectral elements is an excellent choice because the unusual circumstances actually lead to total decoupling of the equations in time.

Once again the strong formulation of the problem is the starting point:

$$\theta_t(x,t) = \alpha\theta_{xx}(x,t) - \beta\theta(x,t), \tag{5.1}$$

$$\theta(x,0) = \theta_0(x), \tag{5.2}$$

$$\theta_x(0,t) = \theta_x(1,t) = 0. \tag{5.3}$$

To arrive at the weak form necessary to state the finite element problem, the inner product of an arbitrary function, v, and the strong equation is integrated over the spatial domain. The initial condition is also weakly incorporated to arrive at the following form:

$$\int_0^1 (v,\theta_t)dx - \alpha\int_0^1 (v,\theta_{xx})dx + \beta\int_0^1 (v,\theta)dx + \int_0^1 v(x,0)(\theta(x,0)-\theta_0(x))dx = 0. \tag{5.4}$$

The inner product represents a form of time integration over the domain. For an arbitrary $v \in L^2$, satisfaction of this equation ensures agreement with the strong form.

Part of the benefit of a finite element method is that it allows the constraints on

the solution to be weakened. Some of the relaxation of constraints is done through integration by parts to transfer derivatives from the field variable, θ, to the test function, v. Therefore, to be effective, the inner product must allow integration by parts to proceed in the same fashion as with standard time integration. To that end, the inner product must satisfy

$$(v, \theta_t) = v(x,1)\theta(x,1) - v(x,0)\theta(x,0) - (v_t, \theta) \tag{5.5}$$

and

$$\int_0^1 (v, \theta_{xx})dx = (v(1,t), \theta_x(1,t)) - (v(0,t), \theta_x(0,t)) - \int_0^1 (v_x, \theta_x)dx. \tag{5.6}$$

Using these properties to integrate Equation 5.4 by parts and substituting the homogeneous Neumann boundary conditions of Equation 5.3 into the result yields the final weak form,

$$\begin{aligned} \int_0^1 v(x,1)\theta(x,1)dx &- \int_0^1 (v_t, \theta)dx + \alpha \int_0^1 (v_x, \theta_x)dx + \beta \int_0^1 (v, \theta)dx - \\ \int_0^1 v(x,0)\theta_0(x)dx &= 0. \end{aligned} \tag{5.7}$$

As in the first approach, to form the quadratic minimizable form, the energy equality, from the weak form, v is set equal to θ. Symmetry of the inner product and Equation 5.5 allow the second term in the energy equality to be simplified:

$$-\int_0^1 (\theta_t, \theta)dx = -\frac{1}{2}\theta^2(x,1) + \frac{1}{2}\theta^2(x,0). \tag{5.8}$$

The energy equality can be written as a general functional of an arbitrary function, v, which vanishes when $v = \theta$,

$$\begin{aligned} E(v) = &\frac{1}{2}\int_0^1 v^2(x,1)dx + \frac{1}{2}\int_0^1 v^2(x,0)dx + \alpha \int_0^1 (v_x, v_x)dx + \beta \int_0^1 (v, v)dx - \\ &\int_0^1 v(x,0)\theta_0(x)dx. \end{aligned} \tag{5.9}$$

All of the terms in the energy equality, save the last one, are positive definite and quadratic. The last term is linear. Therefore, the energy equality is a quadratic minimizable functional which vanishes when $v = \theta$, and it can serve as the functional to minimize in the constrained minimization problem that will generate the bounds.

The output functional, the average over space and time, can be written in terms of an inner product,

$$\ell(v) = \int_0^1 (v, 1)dx, \tag{5.10}$$

$$\ell(\theta) = s. \tag{5.11}$$

Augmenting the energy equality with the linear output functional once again results in a quadratic minimizable form which reduces to $\pm s$ when $v = \theta$. As before, the

augmented functional is defined

$$\mathcal{S}^{\pm}(v) = E(v) \pm \ell(v), \tag{5.12}$$

$$\begin{aligned}\mathcal{S}^{\pm}(v) &= \frac{1}{2}\int_0^1 v^2(x,1)dx + \frac{1}{2}\int_0^1 v^2(x,0)dx + \alpha\int_0^1 (v_x, v_x)dx + \beta\int_0^1 (v,v)dx - \\ &\int_0^1 v(x,0)\theta_0(x)dx \pm \int_0^1 (v,1)dx. \end{aligned}\tag{5.13}$$

The weak form itself, Equation 5.7, can be written in the more general form of a Lagrange multiplier constraint enforcing $v = \theta$,

$$\begin{aligned}\int_0^1 \mu(x,1)v(x,1)dx &- \int_0^1 (\mu_t, v)dx + \alpha\int_0^1 (\mu_x, v_x)dx + \beta\int_0^1 (\mu, v)dx - \\ \int_0^1 \mu(x,0)\theta_0(x)dx &= 0. \end{aligned}\tag{5.14}$$

For Equation 5.14 to be satisfied for an arbitrary μ, v must equal θ.

Because $\mathcal{S}^{\pm}(\theta) = \pm s$, trivially

$$\pm s = \min_{v=\theta} \mathcal{S}^{\pm}(v). \tag{5.15}$$

The constraint that $v = \theta$ is equivalent to the constraint that Equation 5.14 be satisfied for an arbitrary Lagrange multiplier functions, μ. Therefore, a Lagrangian can be formed to transform the constrained minimization problem of Equation 5.15 into the min-max problem

$$\begin{aligned}\mathcal{L}^{\pm}(\mu, v) &= \frac{1}{2}\int_0^1 v^2(x,1)dx + \frac{1}{2}\int_0^1 v^2(x,0)dx + \alpha\int_0^1 (v_x, v_x)dx + \beta\int_0^1 (v,v)dx - \\ &\int_0^1 v(x,0)\theta_0(x)dx \pm \int_0^1 (v,1)dx + \int_0^1 \mu(x,1)v(x,1)dx - \\ &\int_0^1 (\mu_t, v)dx + \alpha\int_0^1 (\mu_x, v_x)dx + \beta\int_0^1 (\mu, v)dx - \\ &\int_0^1 \mu(x,0)\theta_0(x)dx, \end{aligned}\tag{5.16}$$

$$\pm s = \min_{v}\max_{\mu} \mathcal{L}^{\pm}(\mu, v). \tag{5.17}$$

In other words, $\mathcal{L}^{\pm}(\mu, v) = \pm s$ at the saddle point of the Lagrangian. At the saddle point, $v = \theta$ and $\mu = \psi^{\pm}$, or $\mathcal{L}^{\pm}(\psi^{\pm}, \theta) = \pm s$; $\psi^{\pm}$ is known as the adjoint.

Satisfaction of the constraint in the Lagrangian produces the exact value of the output. Relaxation of the constraint produces bounds. To relax the constraint, first the minimization and maximization must be swapped as is permissible in the case of a saddle point,

$$\pm s = \max_{\mu}\min_{v} \mathcal{L}^{\pm}(\mu, v). \tag{5.18}$$

The constraint is then relaxed by removing the maximization over the Lagrange

multiplier function. Since the Lagrangian was maximized by $\mu = \psi^{\pm}$, an arbitrary choice of μ produces a value less than the maximum, $\pm s$, a bound, i. e.

$$\pm s \geq \min_{v} \mathcal{L}^{\pm}(\mu, v) \tag{5.19}$$

or

$$\min_{v} \mathcal{L}^{+}(\mu_1, v) \leq s \leq -\min_{v} \mathcal{L}^{-}(\mu_2, v). \tag{5.20}$$

The relaxation of the constraint effectively frees the approximate solution from satisfying the weak form exactly .

The primary calculation involved in finding bounds determines the v that minimizes the Lagrangian for a given μ. This minimizing v is denoted $\hat{\theta}^{\pm}$. The Lagrangian can be minimized with respect to v by setting to zero the first variation of the Lagrangian with respect to v:

$$\begin{aligned}
\delta\mathcal{L}_v^{\pm}(\hat{\mu}^{\pm}, \hat{\theta}^{\pm}) \;=\; & \int_0^1 \hat{\theta}^{\pm}(x,1)\delta v(x,1)dx + \int_0^1 \hat{\theta}^{\pm}(x,0)\delta v(x,0)dx + \\
& 2\alpha \int_0^1 (\hat{\theta}_x^{\pm}, \delta v_x)dx + 2\beta \int_0^1 (\hat{\theta}^{\pm}, \delta v)dx - \int_0^1 \delta v(x,0)\theta_0(x)dx \\
& \pm \int_0^1 (\delta v, 1)dx + \int_0^1 \hat{\mu}^{\pm}(x,1)\delta v(x,1)dx - \int_0^1 (\hat{\mu}_t^{\pm}, \delta v)dx + \\
& \alpha \int_0^1 (\hat{\mu}_x^{\pm}, \delta v_x)dx + \beta \int_0^1 (\hat{\mu}^{\pm}, \delta v)dx,
\end{aligned} \tag{5.21}$$

where $\hat{\mu}$ refers to the approximate guess for the adjoint.

The inner product and finite element spaces have thus far been left unspecified. The time discretization is set by the order of the Gauss-Legendre-Lobatto quadrature scheme used for the spectral elements. If the scheme uses N points, it exactly integrates polynomials of degree $2N-3$ or less. The inner product naturally implied by the quadrature scheme is

$$(v, w) \equiv \sum_{n=1}^{N} \tilde{\rho}_n v(t_n) w(t_n), \tag{5.22}$$

where the t_n are the quadrature points, and the $\tilde{\rho}_n$ are the associated quadrature weights. The nature of the integration scheme points to a polynomial finite element space. The maximum degree of polynomial fully specifiable with one point value at each quadrature point is $N-1$. The set of degree $N-1$ polynomials, $\xi_i(t)$, that satisfy $\xi_i(t_n) = \delta_{in}$ forms a basis for the space. Additionally, these basis functions are orthogonal in the given inner product. If $\hat{\mu}^{\pm}$, $\hat{\theta}^{\pm}$, and δv are restricted to be members of this space, they can be represented as weighted summations of the basis functions,

$$\hat{\mu}^{\pm}(x,t) \;=\; \sum_{n-1}^{N} a_n(x)\xi_n(t) \tag{5.23}$$

$$\hat{\theta}^{\pm}(x,t) \;=\; \sum_{n-1}^{N} b_n(x)\xi_n(t) \tag{5.24}$$

$$\delta v(x,t) \;=\; \sum_{n-1}^{N} d_n(x)\xi_n(t). \tag{5.25}$$

The coefficients of these basis functions for $\hat{\mu}^{\pm}$, δv, and $\hat{\theta}^{\pm}$ vary in space within the prescribed definition of the spatial finite element space. If the finite element space is chosen to be the set of functions over the domain which are piecewise linear on N_x equally sized sections of space, the elements, whose boundaries are known as nodes, then the basis functions are the standard hat functions, $\phi_i(x)$, centered on each node, x_i, where $\phi_i(x_j) = \delta_{ij}$. The spatially varying coefficients in Equations 5.23, 5.24, and 5.25 can be written as weighted summations of the basis functions so that Equations 5.23, 5.24, and 5.25 can be rewritten

$$\hat{\mu}^{\pm} = \sum_{\ell=1}^{N} \sum_{i=1}^{N_x+1} \hat{\mu}^{\pm}_{\ell i}\phi_i(x)\xi_\ell(t), \tag{5.26}$$

$$\hat{\theta}^{\pm} = \sum_{\ell=1}^{N} \sum_{i=1}^{N_x+1} \hat{\theta}^{\pm}_{\ell i}\phi_i(x)\xi_\ell(t), \tag{5.27}$$

$$\delta v = \sum_{\ell=1}^{N} \sum_{i=1}^{N_x+1} \delta v_{\ell i}\phi_i(x)\xi_\ell(t). \tag{5.28}$$

Following a standard Galerkin procedure of setting $\delta v = \phi_j(x)\xi_n(t)$, and substituting the finite element function definitions of Equations 5.26 and 5.27 into Equation 5.21 yields the following relationship between the approximate θ coefficients and the coefficients of the approximate adjoint:

$$\begin{aligned}[(\delta_{n1} + \delta_{nN} + 2\beta\tilde{\rho}_n)\underline{\underline{\mathbf{M}}} + 2\alpha\tilde{\rho}_n\underline{\underline{\mathbf{K}}}]\underline{\hat{\theta}}^{\pm}_n \;=\; & \delta_{n1}\underline{\mathbf{f}} \mp \tilde{\rho}_n\underline{\mathbf{d}} - [(\delta_{nN} + \beta\tilde{\rho}_n)\underline{\underline{\mathbf{M}}} + \alpha\tilde{\rho}_n\underline{\underline{\mathbf{K}}}]\underline{\hat{\mu}}^{\pm}_n + \\ & \tilde{\rho}_n\underline{\underline{\mathbf{M}}}\underline{\mu}^{\pm}\underline{\mathbf{D}}_n,\end{aligned} \tag{5.29}$$

where $\underline{\underline{\mathbf{M}}}$ is the mass matrix of hat functions defined by

$$\mathbf{M}_{ij} = \int_0^1 \phi_i(x)\phi_j(x)dx,$$

$\underline{\underline{\mathbf{K}}}$ is the stiffness matrix of hat functions defined by

$$\mathbf{K}_{ij} = \int_0^1 \phi_{i_x}(x)\phi_{j_x}(x)dx,$$

$\underline{\hat{\theta}}^{\pm}_n$ is the vector of all coefficient values at quadrature point n and similarly for $\underline{\hat{\mu}}^{\pm}_n$, $\underline{\mathbf{f}}$ is the vector defined by

$$\mathbf{f}_i = \int_0^1 \phi_i(x)\theta_0(x)dx,$$

$\underline{\mathbf{d}}$ is the vector defined by

$$\mathbf{d}_i = \int_0^1 \phi_i(x)dx,$$

$\underline{\mathbf{D}}_i$ is the vector of spectral basis function derivatives at quadrature point i such that

$$\mathbf{D}_{ij} = \xi_{j_t}(t_i),$$

and $\underline{\underline{\mu}}^{\pm}$ is the matrix of all the coefficients of $\mu^{\pm}$.

Finding a good approximation for the adjoint is not as trivial as in the first approach. In the above calculation of the approximate field solution, $\hat{\theta}^{\pm}$, from a given approximate adjoint, the benefit of orthogonality of the spectral element basis functions decouples the equations in time. Spectral basis functions are not orthogonal with their time derivatives. In Equation 5.21, the dense vector of derivatives, $\underline{\mathbf{D}}_n$, acts on the known adjoint. An attempt to solve the original system on a coarse grid with the spectral elements in time would result in a fully coupled system across all time because the derivative matrix would operate on the unknown value of the field variable. The solution of the dense system would be a highly inefficient way to arrive at an approximate adjoint.

A far more efficient way to calculate an approximate adjoint is by discretizing Equation 5.21 with the linear finite element and backward Euler finite difference scheme of the first approach, choosing the inner product to be true time integration and applying the continuous adjoint boundary condition of Appendix A. The subscript of c denotes quantities calculated on the coarse mesh. The coarse mesh field solution, θ_c, is calculated as in the first approach. The boundary condition from Appendix A is

$$\underline{\psi}_c^{\pm N_t} = -\underline{\theta}_c^{N_t}. \tag{5.30}$$

The backward evolution equation for the adjoint is then

$$\begin{aligned}((1+\beta\Delta t)\underline{\underline{\mathbf{M}}} + \alpha\Delta t\underline{\underline{\mathbf{K}}})\underline{\psi}_c^{\pm n} &= \underline{\underline{\mathbf{M}}}\underline{\psi}_c^{\pm n+1} \mp \Delta t\underline{\mathbf{d}} - (2\alpha\Delta t\underline{\underline{\mathbf{K}}} + 2\beta\Delta t\underline{\underline{\mathbf{M}}})\underline{\theta}_c^n \\ & \quad n = N-1, \cdots, 0. \end{aligned} \tag{5.31}$$

The bilinear interpolant of $\psi_c^{\pm}$ on the truth mesh can be used as an approximate adjoint, $\hat{\mu}^{\pm}$.

With the values for the approximate field solution and adjoint, the bounds, η^+ and η^-, can be constructed by substituting these values into the Lagrangian. The Lagrangian can be written as a function of the finite element and spectral element coefficients as follows:

$$\begin{aligned}\eta^{\pm} &= \frac{1}{2}\hat{\underline{\theta}}_N^{\pm T}\underline{\underline{\mathbf{M}}}\hat{\underline{\theta}}_N^{\pm} + \frac{1}{2}\hat{\underline{\theta}}_1^{\pm T}\underline{\underline{\mathbf{M}}}\hat{\underline{\theta}}_1^{\pm} + \sum_{n=1}^{N}\tilde{\rho}_n\hat{\underline{\theta}}_n^{\pm T}(\alpha\underline{\underline{\mathbf{K}}} + \beta\underline{\underline{\mathbf{M}}})\hat{\underline{\theta}}_n^{\pm} - \hat{\underline{\theta}}_1^{\pm T}\underline{\mathbf{f}} \pm \tilde{\underline{\rho}}^T\hat{\underline{\underline{\theta}}}^{\pm}\underline{\mathbf{d}} + \\ & \hat{\underline{\mu}}_N^{\pm T}\underline{\underline{\mathbf{M}}}\hat{\underline{\theta}}_N^{\pm} - \sum_{n=1}^{N}\tilde{\rho}_n\hat{\underline{\theta}}_n^{\pm T}\underline{\underline{\mathbf{M}}}\hat{\underline{\underline{\mu}}}^{\pm T}\underline{\mathbf{D}}_n + \\ & \sum_{n=1}^{N}\tilde{\rho}_n\hat{\underline{\mu}}_n^{\pm T}(\alpha\underline{\underline{\mathbf{K}}} + \beta\underline{\underline{\mathbf{M}}})\hat{\underline{\theta}}_n^{\pm} - \hat{\underline{\mu}}_1^{\pm T}\underline{\mathbf{f}}, \end{aligned} \tag{5.32}$$

where $\tilde{\underline{\rho}}$ is the vector of quadrature weights, and $\hat{\underline{\underline{\theta}}}^{\pm}$ is the matrix of basis function

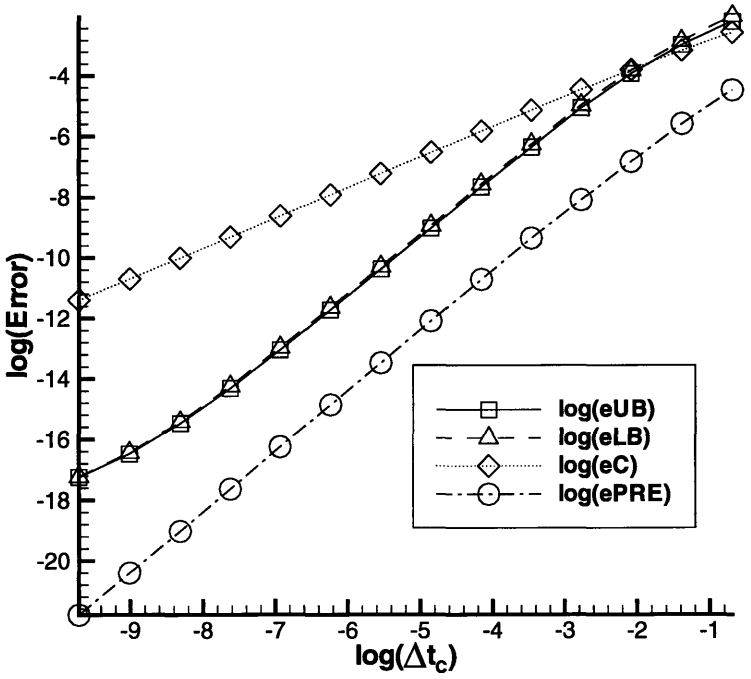

Figure 5-1: Convergence of Bounds with Respect to Coarse Mesh Element Size

coefficients in $\hat{\theta}^{\pm}$, $\hat{\theta}^{\pm}_{ij}$. From Equation 5.20

$$\eta^{+} \leq s \leq -\eta^{-}. \tag{5.33}$$

Thus are the bounds formed.

The second method has some distinct computational advantages over the first method. Equation 5.29 illustrates one advantage of using spectral elements. Though spectral elements provide much better accuracy than backward Euler for the same number of temporal points, Equation 5.29 is decoupled in time in the same fashion as Equation 4.23 and requires the same amount of effort to solve. As before, the computational steps required to solve Equation 5.29 are eminently suitable for parallel computation. Second, the use of the energy equality instead of an energy inequality removes a significant source of error from the calculations. Also, since the same method is used in both approaches to calculate an approximate adjoint including the calculation of the coarse mesh solution, the spectral element method requires no more work than backward Euler to construct an approximate adjoint of similar quality.

Figure 5-1 illustrates the convergence of the method. In the figure, as Δt_C, the size of the coarse mesh time step, decreases, the resolution of the coarse mesh increases, and the errors drop; eUB is the error in the upper bound, eLB is the error in the lower bound, eC is the error in the output calculated by applying the output functional to the coarse mesh solution, and $ePRE$ is the error in the predicted value of the output constructed by averaging the upper and lower bounds. With increasing coarse mesh

resolution, the error in the bounds drops as the upper and lower bound converge toward the true output. The graph also illustrates the utility of the method as a predictive tool. The coarse mesh error represents the error in the output calculated by traditional methods on the given coarse mesh. For a given required error threshold, the graph demonstrates that traditional methods demand a coarse mesh with orders of magnitude more resolution than that required by the constrained minimization approach. The method does not guarantee bounds that are more accurate than the coarse mesh output, but, as the graph indicates, such bounds are possible.

The combined savings of the added accuracy of spectral elements and the decoupling in time generated by the spectral elements ensure the viability of the bounding techniques described here as a rapid alternative to exact output calculation.

Chapter 6

Domain Decomposition

The previous two chapters demonstrate the viability of constrained minimization approaches for output bounding. A real implementation of such techniques would include a spatial decoupling technique called domain decomposition. Because the primary goal of this thesis is to extend constrained minimization techniques to time-dependent problems, domain decomposition is a tangential element. However, as real world application of these techniques relies upon the use of domain decomposition for tractable computation of bounds, this chapter is devoted to a discussion of how domain decomposition is incorporated into constrained minimization techniques for temporal problems using spectral elements in time.

Domain decomposition divides a problem spatially into subdomains and, thus, partially decouples the solution process and produces a strong computational advantage. Much of the computational advantage does not come to bear until a second space dimension is added to the problem, but the simpler problem described here acts as proof of principle for the method. The use of spectral elements and the use of domain decomposition are independent; one can be used without the other. Here they are illustrated working in concert.

The standard constrained minimization technique uses a Lagrange multiplier function to enforce the original finite element equations. In order to utilize domain decomposition, a second constraint must be added to the system. If the problem domain is subdivided into M sections, the problem statement can be relaxed by allowing the field solution to have jump discontinuities across the subdomain boundaries. In order to retrieve the original continuous solution, a Lagrange multiplier function can be used to enforce continuity across the subdomain boundaries. This constraint takes the form

$$\sum_{m=1}^{M-1} (\zeta_m(t), v(x_{m+1_1}, t) - v(x_{m_2}, t)) = 0, \tag{6.1}$$

where the Lagrange multiplier function, ζ, is continuous in time but only exists at the spatial interfaces between subdomains; ζ_m is the time-dependent Lagrange multiplier at the right-hand boundary of subdomain m; and x_{m1} and x_{m2} are the limiting values of x at the left and right ends of subdomain m, respectively. Equation 6.1 is only satisfied for an arbitrary ζ if v is continuous across all subdomain boundaries.

Adding this constraint to the original Lagrangian of Equation 5.16 yields a new Lagrangian that allows θ to have jump discontinuities across subdomain boundaries but then enforces continuity across the same subdomain boundaries by means of a Lagrange multiplier function,

$$\begin{aligned}\mathcal{L}^{\pm}(\mu,\zeta,v) &= \frac{1}{2}\int_0^1 v^2(x,1)dx + \frac{1}{2}\int_0^1 v^2(x,0)dx + \alpha\int_0^1 (v_x,v_x)dx + \beta\int_0^1 (v,v)dx - \\ &\int_0^1 v(x,0)\theta_0(x)dx \pm \int_0^1 (v,1)dx + \int_0^1 \mu(x,1)v(x,1)dx - \\ &\int_0^1 (\mu_t,v)dx + \alpha\int_0^1 (\mu_x,v_x)dx + \beta\int_0^1 (\mu,v)dx - \int_0^1 \mu(x,0)\theta_0(x)dx + \\ &\sum_{m=1}^{M-1} (\zeta_m(t), v(x_{m+1_1},t) - v(x_{m2},t)), \end{aligned} \tag{6.2}$$

$$\pm s = \min_v \max_{\mu,\zeta} \mathcal{L}^{\pm}(\mu,\zeta,v). \tag{6.3}$$

As before, at the saddle point of the Lagrangian, $v = \theta$, the field solution and $\mu = \psi^{\pm}$, the adjoint, and now $\zeta = \omega^{\pm}$, known as the hybrid flux. The bounds are formed by relaxing the constraints and solving the minimization problem for arbritrary Lagrange multiplier functions,

$$\min_v \mathcal{L}^{+}(\mu_1,\zeta_1,v) \le s \le -\min_v \mathcal{L}^{-}(\mu_2,\zeta_2,v). \tag{6.4}$$

The relaxation of constraints frees the approximate solution from satisfying the weak form exactly and allows it to have jump discontinuities across subdomain boundaries.

For given approximate Lagrange multiplier functions, $\hat{\mu}$ and $\hat{\zeta}$, the corresponding minimizing v, $\hat{\theta}$, is found by setting to zero the first variation in the Lagrangian with respect to the field variable,

$$\begin{aligned}\delta\mathcal{L}_v^{\pm}(\hat{\mu}^{\pm},\hat{\zeta}^{\pm},\hat{\theta}^{\pm}) &= \int_0^1 \hat{\theta}^{\pm}(x,1)\delta v(x,1)dx + \int_0^1 \hat{\theta}^{\pm}(x,0)\delta v(x,0)dx + \\ &2\alpha\int_0^1 (\hat{\theta}_x^{\pm},\delta v_x)dx + 2\beta\int_0^1 (\hat{\theta}^{\pm},\delta v)dx - \int_0^1 \delta v(x,0)\theta_0(x)dx \\ &\pm\int_0^1 (\delta v,1)dx + \int_0^1 \hat{\mu}^{\pm}(x,1)\delta v(x,1)dx - \int_0^1 (\hat{\mu}_t^{\pm},\delta v)dx + \\ &\alpha\int_0^1 (\hat{\mu}_x^{\pm},\delta v_x)dx + \beta\int_0^1 (\hat{\mu}^{\pm},\delta v)dx + \\ &\sum_{m=1}^{M-1} (\hat{\zeta}_m^{\pm}(t), \delta v(x_{m+1_1},t) - \delta v(x_{m2},t)) = 0. \end{aligned} \tag{6.5}$$

The hybrid flux is a continuous function in space but only exists at discrete points in space, the subdomain boundaries. Like $\hat{\mu}$ and $\hat{\theta}$, $\hat{\zeta}$ can be written as a linear

combination of spectral basis functions:

$$\hat{\zeta}_m^{\pm}(t) = \sum_{n-1}^{N} \zeta_{nm}\xi_n(t), \quad m = 1, \cdots, M-1. \tag{6.6}$$

In order to enable the finite element space to contain spatially discontinuous functions, the nodes at subdomain boundaries need to be split into double nodes with the hat function centered on the left node truncated to the right of the node and the hat function centered on the right node truncated to the left of the node as is typically done with the hat functions on the edges of the domain. The approximate adjoint and the corresponding field solution, $\hat{\theta}$, can be written in terms of the discontinuous basis functions as

$$\hat{\mu}^{\pm} = \sum_{\ell=1}^{N} \sum_{i=1}^{N_x+M} \hat{\mu}_{\ell i}^{\pm}\phi_i(x)\xi_\ell(t) \tag{6.7}$$

and

$$\hat{\theta}^{\pm} = \sum_{\ell=1}^{N} \sum_{i=1}^{N_x+M} \hat{\theta}_{\ell i}^{\pm}\phi_i(x)\xi_\ell(t). \tag{6.8}$$

Substituting these representations of the functions into Equation 6.5 and following the Galerkin procedure of setting $\delta v = \phi_j(x)\xi_n(t)$ yields a relationship between the coefficients of the approximate θ and those of the corresponding approximate adjoint and hybrid flux,

$$\begin{aligned}[(\delta_{n1} + \delta_{nN} + 2\beta\tilde{\rho}_n)\underline{\underline{\mathbf{M}}} + 2\alpha\tilde{\rho}_n\underline{\underline{\mathbf{K}}}]\underline{\hat{\theta}}_n^{\pm} \;=\; & \delta_{n1}\underline{\mathbf{f}} \mp \tilde{\rho}_n\underline{\mathbf{d}} - [(\delta_{nN} + \beta\tilde{\rho}_n)\underline{\underline{\mathbf{M}}} + \alpha\tilde{\rho}_n\underline{\underline{\mathbf{K}}}]\underline{\hat{\mu}}_n^{\pm} + \\ & \tilde{\rho}_n\underline{\underline{\mathbf{M}}}\underline{\mu}^{\pm}\underline{\mathbf{D}}_n - \tilde{\rho}_n\underline{\underline{\mathbf{V}}}\underline{\hat{\zeta}}_n^{\pm},\end{aligned} \tag{6.9}$$

where $\underline{\hat{\zeta}}_n^{\pm}$ is the vector of all coefficient values of the approximate hybrid flux at quadrature point n, and $\underline{\underline{\mathbf{V}}}$ is the matrix defined by

$$\mathbf{V}_{ij} = \delta_{i,j+1_1} - \delta_{i,j_2},$$

where i is the number of a node, $j + 1_1$ is the number of the node on the left side of subdomain $j + 1$, and j_2 is the number of the node on the right side of subdomain j.

The approximate adjoint is calculated as in the previous chapter; however, continuity of the adjoint is enforced, so the hybrid flux contributions cancel out, thereby reducing Equation 6.9 to Equation 5.29. To calculate an approximate hybrid flux on the same coarse mesh, the discontinuous formulation must be retained. This is done by following the same procedure as for the approximate adjoint but choosing the test functions, δv, to be ramp functions, $\sigma_j(x)$, piecewise linear functions by subdomain which are zero at the left edge of subdomain j, one at the right edge of subdomain j, and zero in the remaining subdomains. The matrices $\underline{\underline{\mathbf{Q}}}$ and $\underline{\underline{\mathbf{R}}}$ and the vector $\underline{\mathbf{g}}$ are defined by

$$\mathbf{Q}_{ij} = \int_0^1 \sigma_i(x)\phi_j(x)dx,$$

$$\mathbf{R}_{ij} = \int_0^1 \sigma_{ix}(x)\phi_{j_x}(x)dx,$$

and

$$\mathbf{d}_i = \int_0^1 \sigma_i(x)dx.$$

Solving the new discrete equations yields

$$\begin{aligned} \underline{\omega}_c^{\pm n} &= (2\alpha\underline{\underline{\mathbf{R}}} + 2\beta\underline{\underline{\mathbf{Q}}})\underline{\theta}_c^n \pm \underline{\mathbf{g}} - \frac{1}{\Delta t}\underline{\underline{\mathbf{Q}}}(\underline{\psi}_c^{\pm n+1} - \underline{\psi}_c^{\pm n}) + (\alpha\underline{\underline{\mathbf{R}}} + \beta\underline{\underline{\mathbf{Q}}})\underline{\psi}_c^{\pm n} \\ & n = 0, \cdots, N-1. \end{aligned} \tag{6.10}$$

Because the time differentiation requires a value for $\underline{\psi}_c^{\pm N+1}$, an additional time step for $\underline{\psi}_c^{\pm}$ must be calculated using a rearranged Equation 5.31,

$$\underline{\underline{\mathbf{M}}}\underline{\psi}_c^{\pm N+1} = ((1+\beta\Delta t)\underline{\underline{\mathbf{M}}} + \alpha\Delta t\underline{\underline{\mathbf{K}}})\underline{\psi}_c^{\pm N} \pm \Delta t\underline{\mathbf{d}} + (2\alpha\Delta t\underline{\underline{\mathbf{K}}} + 2\beta\Delta t\underline{\underline{\mathbf{M}}})\underline{\theta}_c^N. \tag{6.11}$$

The coarse mesh values for the hybrid flux can be linearly interpolated in time on the truth mesh to produce the approximate hybrid flux, $\zeta^{\pm}$, for Equation 6.9.

The bounds are constructed by evaluating the Lagrangian using the calculated values of the approximate field solution, adjoint, and hybrid flux;

$$\begin{aligned} \eta^{\pm} &= \frac{1}{2}\hat{\underline{\theta}}_N^{\pm T}\underline{\underline{\mathbf{M}}}\hat{\underline{\theta}}_N^{\pm} + \frac{1}{2}\hat{\underline{\theta}}_1^{\pm T}\underline{\underline{\mathbf{M}}}\hat{\underline{\theta}}_1^{\pm} + \sum_{n=1}^{N}\tilde{\rho}_n\hat{\underline{\theta}}_n^{\pm T}(\alpha\underline{\underline{\mathbf{K}}} + \beta\underline{\underline{\mathbf{M}}})\hat{\underline{\theta}}_n^{\pm} - \hat{\underline{\theta}}_1^{\pm T}\underline{\mathbf{f}} \pm \tilde{\underline{\rho}}^T\hat{\underline{\theta}}^{\pm}\underline{\mathbf{d}} + \\ & \hat{\underline{\mu}}_N^{\pm T}\underline{\underline{\mathbf{M}}}\hat{\underline{\theta}}_N^{\pm} - \sum_{n=1}^{N}\tilde{\rho}_n\hat{\underline{\theta}}_n^{\pm T}\underline{\underline{\mathbf{M}}}\hat{\underline{\mu}}^{\pm T}\underline{\mathbf{D}}_n + \sum_{n=1}^{N}\tilde{\rho}_n\hat{\underline{\mu}}_n^{\pm T}(\alpha\underline{\underline{\mathbf{K}}} + \beta\underline{\underline{\mathbf{M}}})\hat{\underline{\theta}}_n^{\pm} - \hat{\underline{\mu}}_1^{\pm T}\underline{\mathbf{f}} + \\ & \sum_{n=1}^{N}\tilde{\rho}_n\hat{\underline{\theta}}_n^{\pm T}\underline{\underline{\mathbf{V}}}\hat{\underline{\zeta}}_n^{\pm}, \end{aligned} \tag{6.12}$$

$$\eta^+ \leq s \leq -\eta^-. \tag{6.13}$$

The domain decomposition provides a great source of computational savings. Splitting the nodes at subdomain boundaries decouples the spatial problems in each subdomain. The decomposition effectively turns each subdomain into an independent local Neumann problem. The halved hat functions at the subdomain interfaces mean that the mass and stiffness matrices have no contribution from products of basis functions in different subdomains. This information can be used to solve Equation 6.9 separately in each subdomain. The problem illustrated in this section is one dimensional in space and results in a tridiagonal system, but if the same method is applied to a problem with two spatial dimensions the savings could add up to a factor of M^2 (M is the number of subdomains) in the computational effort required to solve Equation 6.9. Additional effort is required to calculate an approximate hybrid flux and interpolate it on the truth mesh. This computational work is about on par with that required for the approximate adjoint, so that if the guidelines mentioned earlier for choosing the coarse mesh resolution are followed, the added computational cost will not seriously affect the total work. The savings far outweigh the additional com-

putational cost incurred in calculating and interpolating the coarse mesh hybrid flux. The only limitation is that the number of subdomains is limited by the resolution of the coarse mesh because subdomain boundaries cannot fall within coarse mesh elements. The last step in the computation, the actual calculation of the bounds, requires a large number of matrix multiplications. The spatial decoupling generated by domain decomposition facilitates this calculation as well. The contribution to the bounds of each subdomain can be calculated independently and in parallel with the other subdomains. Thus, with minor modifications to the procedure, the valuable tool of domain decomposition can be incorporated into the constrained minimization techniques for time-dependent problems.

Chapter 7

Conclusion

The methods described in this thesis lead to a formulation for calculating bounds for a linear functional output of a parabolic partial differential equation without fully solving the differential equation itself. Converting the differential equation and the calculation of the output of interest into a constrained minimization problem provides the foundation. Relaxing the constraints of the minimization problem leads to the bounds. Allowing an inexact adjoint amounts to relaxing the constraint that forces the solution to exactly satisfy the differential equation. In this manner solution of the original differential equation is bypassed. Permitting an inexact hybrid flux circumvents the requirement that the solution be continuous and decouples the problem in space to allow faster solution.

It is true that poor choices for the approximate adjoint and the approximate hybrid flux will generate poor and possibly useless bounds, but there are inexpensive ways of finding excellent approximations to these quantities and generating very tight bounds. The original differential equation is prohibitively expensive to solve on a truth mesh, but solving the differential equation on a coarse mesh is by no means intractable. The cheaply available coarse mesh solution provides a straightforward means of calculating sufficiently accurate approximations for the adjoint and hybrid flux. In the two approaches illustrated in this thesis, this method of constructing the approximations is shown to provide bounds far tighter than the resolution present in the coarse formulation might suggest is probable.

In the particular cases shown here, the tridiagonal character of the matrices involved in the discrete equations makes the solution of the problem relatively cheap, even compared to the methods shown here. In multiple space dimensions though, the same problems become exceedingly expensive to solve on a truth mesh, while they remain tractable when subjected to the methods of this thesis. The examples shown here demonstrate that this process is possible and that in cases where bounds are useful, either in their own right or as tools to estimate the true value of the output of interest, they can be constructed with far less effort than would be required to find the value of the output by classical approaches.

These methods make computational solution of problems into a viable design tool for optimization processes where high numbers of frequent appeals to the method make speed a necessity instead of a convenience. Even in cases where the method

is exercised only a small number of times, design cycles can be sped up through its use, especially in early stages where high precision may be unnecessary. Certainly, there are large classes of problems which cannot be addressed by these methods as they stand now. However, many significant engineering problems do fall within the domain of these methods, and new methods are under development to deal with the problems that are beyond the realm of current techniques. The community of people interested in outputs of the class of problems addressed by this thesis stands to benefit significantly from the approaches discussed here.

Appendix A

Adjoint Initial Condition

The derivation of the continuous adjoint initial condition takes as its starting point the variation of the continuous Lagrangian with respect to the field solution, Equation 5.21, written in terms of the exact quantities rather than the approximate quantities,

$$\begin{aligned}\delta\mathcal{L}_v^{\pm}(\psi^{\pm},\omega^{\pm},\theta) \;=\; & \int_0^1 \theta(x,1)\delta v(x,1)dx + \int_0^1 \theta(x,0)\delta v(x,0)dx + \\ & 2\alpha\int_0^1 (\theta_x,\delta v_x)dx + 2\beta\int_0^1 (\theta,\delta v)dx - \int_0^1 \delta v(x,0)\theta_0(x)dx \\ & \pm\int_0^1 (\delta v,1)dx + \int_0^1 \psi^{\pm}(x,1)\delta v(x,1)dx - \int_0^1 (\psi_t^{\pm},\delta v)dx + \\ & \alpha\int_0^1 (\psi_x^{\pm},\delta v_x)dx + \beta\int_0^1 (\psi^{\pm},\delta v)dx + \\ & \sum_{m=1}^{M-1} (\omega_m^{\pm}(t),\delta v(x_{m+1_1},t) - \delta v(x_{m2},t)) = 0, \end{aligned} \tag{A.1}$$

where ψ, ω, and θ are the adjoint, hybrid flux, and field solution, respectively, and δv is the variation in the field solution.

The adjoint is continuous, so only continuous test functions, δv, are admissible, and the hybrid flux term cancels out. Grouping the terms commonly dependent on $\delta v(x,1)$ yields

$$\int_0^1 \delta v(x,1)(\theta(x,1) + \psi^{\pm}(x,1))dx. \tag{A.2}$$

Since the field variable variation is arbitrary, it must be true that

$$\psi^{\pm}(x,1) = -\theta(x,1). \tag{A.3}$$

This is the continuous adjoint initial condition. Because the adjoint evolves backward in time, the initial condition occurs at the final time.

Appendix B

Spectral Elements

The spectral elements finite element scheme utilizes the beneficial properties of Gauss-Legendre-Lobatto quadrature to facilitate the implementation of a finite element scheme whose basis functions are orthogonal polynomials [6]. Gauss-Legendre-Lobatto (GLL) quadrature estimates an integral as the weighted sum of samples of the integrand at N points in the integral's domain including the two end points:

$$\int_0^1 f(t)dt \approx \sum_{n=1}^{N} \tilde{\rho}_n f(t_n), \tag{B.1}$$

where t_n are the quadrature points, and $\tilde{\rho}_n$ are the corresponding quadrature weights. GLL quadrature exactly integrates polynomials of degree less than or equal to $2N-3$.

In the case illustrated in the second approach in this thesis, the field variables are expressed as polynomials of degree $N-1$. The inner product used is selected to be time integration by GLL quadrature of the product of the two operands. There are two requirements placed on the inner product:

$$(v, \theta_t) = v\theta|_0^1 - (v_t, \theta) \tag{B.2}$$

and

$$\int_0^1 (v, \theta_{xx})dx = (v(1,t), \theta_x(1,t)) - (v(0,t), \theta_x(0,t)) - \int_0^1 (v_x, \theta_x)dx. \tag{B.3}$$

For the first requirement, v is a polynomial of degree $N-1$, and θ is a polynomial of degree $N-1$ which means that θ_t is a polynomial of degree $N-2$. The product, $v\theta_t$, is a polynomial of degree $2N-3$; therefore, its integral by GLL quadrature is exact. For an exact integral, integration by parts can be performed,

$$(v, \theta_t) = \int_0^1 v\theta_t dt = v\theta|_0^1 - \int_0^1 v_t\theta dt. \tag{B.4}$$

In Equation B.4, v_t is a polynomial of degree $N-2$, and θ is a polynomial of degree $N-1$, so the exact integral of their product in the last term can be replaced with GLL quadrature expressed as their inner product to yield the first condition on the

inner product.

It can be demonstrated that the spectral element inner product satisfies the second condition by expanding the left hand side of Equation B.3 and integrating by parts:

$$\begin{aligned}
\int_0^1 (v, \theta_{xx}) dx &= \int_0^1 \sum_{n=1}^{N} \tilde{\rho}_n v(x, t_n) \theta_{xx}(x, t_n) dx \\
&= \sum_{n=1}^{N} \tilde{\rho}_n \int_0^1 v(x, t_n) \theta_{xx}(x, t_n) dx \\
&= \sum_{n=1}^{N} \tilde{\rho}_n (v(1, t_n)\theta_x(1, t_n) - v(0, t_n)\theta_x(0, t_n) - \int_0^1 v_x(x, t_n)\theta_x(x, t_n) dx \\
&= (v(1, t), \theta_x(1, t)) - (v(0, t), \theta_x(0, t)) - \int_0^1 (v_x(x, t), \theta_x(x, t)) dx. \quad \text{(B.5)}
\end{aligned}$$

In the spectral element method, the basis functions of the finite element space are expressed as $N-1$ degree polynomials which are orthogonal in the inner product. The basis polynomials are selected so that they evaluate to one at a single quadrature point and zero at all of the others. The value of any one of the basis functions evaluated at a quadrature point can be written

$$\xi_i(t_j) = \delta_{ij}. \quad \text{(B.6)}$$

The orthogonality of the basis function leads to a high degree of simplification in the evaluation of inner products of function in the finite element space. Any function in the space can be written as

$$v(t) = \sum_{n=1}^{N} v_n \xi_n(t). \quad \text{(B.7)}$$

The inner product of two such functions, v and w, simplifies as follows:

$$\begin{aligned}
(v, w) &= \sum_{n=1}^{N} \tilde{\rho}_n (\sum_{k=1}^{N} v_k \xi_k(t_n))(\sum_{j=1}^{N} w_j \xi_j(t_n)) \\
&= \sum_{n=1}^{N} \tilde{\rho}_n (\sum_{k=1}^{N} v_k \delta_{nk})(\sum_{j=1}^{N} w_j \delta_{nj}) \\
&= \sum_{n=1}^{N} \tilde{\rho}_n v_n w_n. \quad \text{(B.8)}
\end{aligned}$$

In the computational work described in this thesis, the calculation of spectral basis functions and their derivatives is done with the help of Einar Rønquist's `speclib`.

Bibliography

[1] M. Ainsworth and J.T. Oden, A posteriori *error estimation in finite element analysis,* TICAM Report 96-19, 1996.

[2] M. Ainsworth and J.T. Oden, *A unified approach to* a posteriori *error estimation using element residual methods,* Numer. Math., **65** (1993), pp. 23-50.

[3] R.E. Bank and A. Weiser, *Some* a posteriori *error estimators for elliptic partial differential equations,* Math. Comp., **44**:170 (1985), pp. 1597-1615.

[4] R. Becker and R. Rannacher, *A feedback approach to error control in finite element methods: basic analysis and examples,* IWR Preprint 96-52 (SFB 359), Heidelberg, 1996.

[5] R. Becker and R. Rannacher, *Weighted* a posteriori *error control in finite element methods,* IWR Preprint 96-1 (SFB 359), Heidelberg, 1996.

[6] C. Bernardi and Y. Maday, *Approximations spectrales de problèmes aux limites elliptiques,* Springer-Verlag, Paris (1992).

[7] P. Ladeveze and D. Leguillon, *Error estimation procedures in the finite element method and applications,* SIAM J. Numer. Anal., **20** (1983), pp. 485-509.

[8] L. Machiels, A posteriori *finite element bounds for linear and nonlinear functional outputs of Burgers equation*, in preparation.

[9] M. Paraschivoiu, A Posteriori *Finite Element Bounds for Linear-Functional Outputs of Coercive Partial Differential Equations and of the Stokes Problam,* Ph.D. Thesis, Department of Mechanical Engineering, M.I.T., 1997.

[10] M. Paraschivoiu and A.T. Patera, *A hierarchical duality approach to bounds for the outputs of partial differential equations,* Comp. Meth. Appl. Mech. Engrg., to appear.

[11] M. Paraschivoiu and A.T. Patera, A Posteriori *Bounds for Linear-Functional Outputs of Crouzeix-Raviart Finite Element Discretizations of the Incompressible Stokes Problem* International Journal for Numerical Methods in Fluids, in review.

[12] M. Paraschivoiu, J. Peraire, and A.T. Patera, A posteriori *finite element bounds for linear-functional outputs of elliptic partial differential equations,* Comp. Meth. Appl. Mech. Engrg., **150** (1997), pp. 289-312.

[13] A.T. Patera and E.M. Rønquist, *Introduction to Finite Element Methods*, in preparation.

[14] J. Peraire and A.T. Patera, *Asymptotic* a posteriori *finite element bounds for the outputs of noncoercive problems: the Helmholtz and Burgers equations,* Comp. Meth. Appl. Mech. Engrg., in review.

[15] J. Peraire and A.T. Patera, *Bounds for linear-functional outputs of coercive partial differential equations: local indicators and adaptive refinement,* in Proceedings of the Workshop on New Advances in Adaptive Computational Methods in Mechanics, Cachan, France, eds. P. Ladeveze and J.T. Oden, Elsevier, 1997.

[16] G. Strang, *Introduction to Applied Mathematics*, Wellesley-Cambridge Press, Wellesley (1986).

About the Author

I grew up in Silver Spring, Maryland in a home with my parents, Bob and Lois, and my brother, Daniel. I attended the Hebrew Academy of Greater Washington from elementary school until the end of eleventh grade when I left without graduating to attend Yale University in New Haven, Connecticut in September of 1992. During my four years at Yale, I double-majored in Mechanical Engineering and Computer Science, and I dabbled in philosophy. In my junior year, I joined Tau Beta Pi, the engineering honor society. When Yale graduated me in May of 1996 with a Bachelor of Science, I received the American Society of Mechanical Engineers award for the most outstanding graduate in Yale's Mechanical Engineering Department and the Becton Prize. During the summers of 1991, 1992, and 1993, I worked as a computer programmer for the U.S. Navy at the Naval Surface Warfare Center in White Oak, Maryland. As I converted from a computer programmer to a mechanical engineer, I worked at the Institute for Systems Research at the University of Maryland, College Park during the summer of 1994. The following two summers, as I began to blend my computer experience with my engineering training, I worked for Enig Associates, Incorporated, a small military contractor, doing computational fluid dynamics as part of a research and development team. In the fall of 1996 I relocated to Cambridge, Massachusetts where I began my graduate work in the Fluid Mechanics Laboratory at the Massachusetts Institute of Technology sponsored by a National Science Foundation Graduate Fellowship and a Mechanical Engineering Department research assistantship. When my Master's degree is complete, I will commence research for a Doctor of Philosophy degree in the same laboratory extending the work covered by my Master's thesis.