REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	29 May 1997	Final Technical Report 1 May 1993 - 30 April 1997

4. TITLE AND SUBTITLE

The Air For Center for Optimal Design and Control

5. FUNDING NUMBERS

AFOSR-TR-97 0630

(G) F49620-93-1-0280

6. AUTHOR(S)

J.A. Burns

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Interdisciplinary Center for Applied Mathematics
Wright House, West Campus Drive
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061-0531

8. PERFORMING ORGANIZATION REPORT NUMBER

ICAM Report 97-06-01

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Office of Scientific Research
Code NM
110 Duncan Avenue, Suite B115
Bolling AFB, DC 20332

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This report contains a summary and highlights of the research funded by the Air Force under AFOSR URI Grant F49620-93-1-0280, titled "Center for Optimal Design and Control of Distributed Parameter Systems" (CODAC), for the period 1 May 1993 to 30 April 1997. The Center conducts a wide range of research and educational programs, and promotes linkages between Air Force Laboratories, industry and university scientists.

During this four year period, CODAC researchers produced more than 150 scientific papers, made 160 presentations at conferences and colloquium and directed more than 35 graduate students. This research effort has produced several new mathematical algorithms for optimal design and control of fluid systems and a new solvability result for nonlinear hyperbolic systems. The effort in optimal design produced a revolutionary new approach for optimal design that combines continuous sensitivity equations with computational mathematics to greatly reduce design cycle time. This Sensitivity Equation Method has been transitioned into several commercial software packages and is the basis for continuing joint projects with industries throughout the United States. The effort in control of fluids and structures has produced two fundamental breakthroughs in the areas of distributed parameter actuator/sensor placement and in reduced basis methods for design of low order dynamic controllers for fluid/structure interactions. In addition, this report contains a summary of the interactions between Air Force facilities and industrial partners.

14. SUBJECT TERMS

Optimal Design and Control of Fluids, Scientific Computation

15. NUMBER OF PAGES

226

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

The Air Force Center for Optimal Design and Control

Final Technical Report
May 1, 1993 - April 30, 1997

FINAL TECHNICAL REPORT ON AFOSR GRANT

F49620-93-1-0280

THE AIR FORCE CENTER FOR OPTIMAL DESIGN AND CONTROL

for the period

1 May 1993 - 30 April 1997

by

John A. Burns, Director
Air Force Center for Optimal Design and Control

Wright House / West Campus Drive
Virginia Polytechnic Institute & State University
Blacksburg, VA 24061- 0531

29 May 1997

Prepared for the : Air Force Office of Scientific Research
Mathematical & Computer Sciences
110 Duncan Avenue Suite B115
Bolling AFB, DC 20332-0001

Contents

Chapter 1

Introduction and Overview

This final technical report contains a summary of the activities supported under the Air Force AFOSR URI Grant F49620-93-1-0280 during the four year period 1 May 1993 through 30 April 1997. This project is concerned with an interdisciplinary research program for the modeling, design, analysis, control and testing of new approaches to distributed parameter systems. The principal investigator is Dr. John A. Burns, Hatcher Professor of Mathematics at Virginia Tech.

1.1 Introduction

The *Air Force Center for Optimal Design and Control* (**CODAC**) was established in May 1993 under the Air Force AFOSR URI Grant F49620-93-1-0280. **CODAC** is an interdisciplinary research center with core academic participants at Virginia Tech and North Carolina State University. The Center is built on a highly integrated interdisciplinary program with four major components:

- **Long Term Multi-disciplinary Research**

  Activities of **CODAC** are organized around specific research topics. Each topic focuses on a particular Air Force or industrial problem area and includes joint research efforts involving the two universities, the appropriate Air Force Laboratory, and industry. This research has theoretical, computational, and experimental components.

- **Experimental Testing and Validation**

  The direction of the research programs is formed in part by experimental results. Moreover, we use both numerical experiments and laboratory experiments to test and validate the theoretical research. Laboratory experiments and testing and development of software take place at Virginia Tech, at Air Force Laboratories and, when appropriate, at an industrial site.

- **Air Force Laboratory Interaction and Industrial Links**

  There is a constant interaction between scientists and engineers at Air Force facilities, Virginia Tech, and the North Carolina Sate University. We have several ongoing projects involving core participants and Air Force scientists. We have expanded the current efforts to actively pursue new laboratory interactions, and to use these ties to facilitate the transition of basic research results to the Air Force. In addition, we have initiated several new projects with both large and small companies in order to facilitate the transition of research into dual usage.

- **Interdisciplinary Educational Program**

  The Center offers an unparalleled potential for strengthening the educational and scientific infrastructure by training students and post-doctoral researchers in an interdisciplinary team approach to scientific and engineering research. The Center provides unique opportunities

for theoretical, computational, and experimental research. Through the interactions with Air Force laboratories and industrial partners, students are exposed to real problems. The combined theoretical, computational, and experimental approach provides a meaningful interdisciplinary research experience.

1.2 Center Organization and Facilities

Dr. John A. Burns is the *Director* of **CODAC** and is responsible for the day-to-day operation of the Center and for implementing and coordinating the research, laboratory/industry interactions, and educational programs. *Dr. Eugene M. Cliff* is the *Director for Engineering.* The **CODAC** is located within the Interdisciplinary Center for Applied Mathematics **(ICAM)** at Virginia Tech. *Dr. Terry Herdman* is the *Director* of **ICAM**. The *Executive Advisory Committee* consists of the Center Director, the Director for Engineering and the Director of the Interdisciplinary Center for Applied Mathematics. The Executive Committee is responsible for program planning and for advising the Director on the allocation of resources and on ways to make **CODAC** more effective as an Air Force resource. The research conducted at **CODAC** requires computational, as well as more traditional laboratory facilities. Although much of the large scale computing is done on supercomputers at Air Force facilities, pre-processing and post-processing must be done locally.

ICAM Computing Facilities
ICAM houses a heterogeneous Unix system with file-sharing under a Network File System (NFS). The Unix system currently consists of the following platforms:

- Our main file server (`sun.icam.vt.edu`) is four processor SUN-1000, with 128MB internal memory, 1GB internal hard-drive, an external SCSI hard-drive system with 27GB of available storage.
- Our graphics workstations include a Silicon Graphics IRIS 4D/310 graphics workstation (`sgi.icam.vt.edu`). This features an R4000 processor, 64MB of memory and supports a 650MB read-write optical disk drive. A second 'public' SGI-Indigo2 features an R8000 processor, 128MB of memory and Extreme graphics (`sgi2.icam.vt.edu`). Similar machines (`sgi3.icam.vt.edu`) are reserved for use by the Post-Doctoral researchers and by Dr. Burns (`burns.icam.vt.edu`).
- Our main compute engines are two DEC Alpha 3000/600 computers with 256MB of memory (`alpha1.icam.vt.edu` and `alpha2.icam.vt.edu`). A third DEC AlphaServer 2100 is a dual-processor machine with 512MB of memory (`alpha3.icam.vt.edu`).
- In addition to the dedicated monitor/keyboard for each platform the computers can be accessed from the Ethernet. The Center currently has (10) Pentium–based personal computers (two in faculty offices and eight for public use) and (7) Power-Macintosh machines (one in a faculty office). One of these PowerMacs is principally dedicated to producing high–quality video graphics.
- Standard print output is directed to an HP LaserJet 5SiMX printer while color output can be produced on any of three HP DeskJet 1600CM printers.
- The system is connected via a (10baseT) Ethernet (10Mb/s).

ESM Mechanical Systems Laboratory - MSL
Experiments on integrated actuator/sensor/materials are carried out in the Mechanical Systems Laboratory (MSL) of the Engineering Science and Mechanics Department. This laboratory consists of 2 laboratories with more than 1500 square feet of floor space. The laboratory contains a variety

of actuators, sensors and piezoelectric devices. To provide control signals to the various actuators, the laboratory has an EAI 2000 analog computer, capable of handling up to 60 state variables, plus several smaller analog computers. Digital control is accomplished by a 16 channel Systolic Systems' Optima 3, capable of real time digital control. Motion sensing is accomplished by a variety of encoders, tachometers, 20 Kistler Instrument Piezobeam accelerometers, rotational accelerometers, strain gauges, a laser vibrometer (DISA), proximity probes, and piezoelectric films. These items are complete with the required electronics and data acquisition systems. There are a variety of piezoelectric devices - both films and ceramics, 6 self-sensing actuator circuits with layered PZTs and an embedded self-sensing actuator beam system.

ME Adaptive Structures Laboratory - ASL

The experiments on structural/fluid control are carried out in the Adaptive Structures Laboratory of the Mechanical Engineering Department. Currently, the facilities of the (ASL) consist of several vibration measurement systems, smart structures test beds (beams, plates, and cylinders), and a variety of digital control systems. Modal sensors and piezoelectric actuators are fabricated and attached to structures routinely. The laboratory has the complete capability to perform structural dynamic testing, including amplifiers, accelerometers, distributed piezoelectric sensors, non-contacting transducers, force transducers, and shakers, together with amplifiers and signal conditioning. A Tektronix 4-channel expandable frequency analyzer with integrated software facilitates frequency analysis of electromechanical systems. Together with a complete machine shop and electrical shop located within the department, fabrication of supports and testing of structures is readily performed. The laboratory has access to a scanning laser velocimeter manufactured by Ometron. This scanning laser has made dynamic verification of distributed models of structures actuated by smart materials possible to an extent previously not possible. The scanning laser vibrometer allows extremely high density velocity measurements of distributed structures (the density of points is much higher than the mesh found in finite element models). The use of this laser measurement system has been made easy by the in-house development and implementation of instrument software implemented on a Macintosh platform. Control of structures is enabled with the following instruments and capability. Digital control is provided by a Spectrum 32 input/16 output channel floating point digital signal processor system. Analog control is provided by a 10-channel Comdyna analog computer, and access to a 64-channel EAI 2000 analog computer is also available. Furthermore, small controllers and signal conditioning units are developed in the laboratory using a breadboard development system. This has allowed additional channels of control to be custom designed and built very cheaply. The laboratory includes a variety of PC's and is currently being upgraded with workstations (IBM Risc 6000's) which will run the software currently being run on a VAX-cluster. This software includes modal analysis software (SMS,LMTS,I-DEAS), finite element software (ANSYS,I- DEAS,NASTRAN), and acoustic analysis software (SYSNOISE).

Chapter 2

Objectives

The goal of **CODAC** is to become a resource enabling the realization of the tremendous opportunities that new approaches to interdisciplinary research offer as they are bought to bear on the problems that face the Air Force. Problems of optimal design and control of complex flows, fluid-structure interactions, distributed sensor and actuators, and smart structures are approached by multi-disciplinary research teams (theoretical, computational and experimental). In addition, this research effort is part of a laboratory/academic/industrial program that ensures the transition of the fruits of the research.

2.1 Research Objectives

The Center and its research programs have the following five long term scientific objectives:

- To apply new techniques in modeling, control, and optimal design to the nonlinear interactions between aerodynamic loading and structural deformations, focusing on reduced order modeling, novel approaches to computing sensitivities, adjoint methods, shape optimization, and acceleration algorithms that take advantage of parallel architectures.
- To develop new adaptive distributed sensors and actuators for controlling structural components in a manner which will control the surrounding fluid medium. This effort focuses on the use of segmented smart materials in sensor/actuator design.
- To demonstrate the use of active control in turbulent flows. The fundamental components of this interdisciplinary effort include low order mathematical modeling, and nonlinear adaptive control.
- To provide a scientific investigation into modeling and control of dynamically adaptive, multi-functional smart structural systems for use in the control of nonlinear fluid/structure interactions. This research combines applied mathematics, modeling, control, structural dynamics, computations, and experimental techniques and focuses on smart/self-sensing actuators.
- To develop local and global approaches to non-linear feedback control that incorporate new breakthroughs in nonlinear observers and that are applicable to reduced order models of complex flow and fluid/structure control problems. These new approaches will be tested in flow control and structural control experiments.

2.2 Educational Objectives and Interactions

In addition to the research program, **CODAC** has expanded the interactions between academic researchers, Air Force laboratories and industry. Moreover, we have developed an educational program

to train applied mathematicians, engineers, and scientists to work on interdisciplinary problems. In this regard, we have the following objectives:

- To develop an active program directed at educating interdisciplinary scientists and encouraging these graduates to work at Air Force facilities and laboratories. Special efforts are being made to recruit women and minority students into the educational programs.
- To develop an organizational framework designed to promote the transition of basic research to Air Force laboratories and industry. This framework includes a program of regular short-term visits to Air Force laboratories and industries by academic researchers, post-doctoral associates, and graduate students and workshops devoted entirely to joint research.

Chapter 3

Status and Highlights

During the past four years **CODAC** has been extremely active in all aspects of its mission. This final report contains a summary of several ongoing research projects (Chapter 4) and the laboratory and industrial interactions (Chapter 7). In addition, we provide a list of visitors and a list of faculty and students supported by **CODAC** during this period. The following items are particularly noteworthy.

3.1 Status of Effort

During this four year period, **CODAC** researchers have been extremely productive in terms of scientific papers, presentations at conferences and students. In particular, during the period from 1 May 1993 to 30 April 1997 **CODAC** researchers have :

- **produced more than 125 scientific papers,**
- **made more than 150 presentations at conferences and colloquium,**
- **produced 17 Ph.D. students,**
- **produced 6 M.S. students,**
- **provided partial support for more than 35 graduate students.**

During the past four years this research effort has produced several new mathematical algorithms for optimal design and control of fluid flow systems and fluid/structure interactions. The effort was divided into two basic tasks: **Optimal Design** and **Feedback Control**. This effort has produced revolutionary new methodologies in both areas. In particular, a hybrid sensitivity equation method for optimal design was developed, refined and applied to a wide variety of flow optimization problems. This method has been transitioned into several commercial software packages and is now the basis for many continuing university-industry projects. In the area of feedback control, large scale computational mathematics and reduced basis methods were combined to produce a solid mathematical framework for constructing low order dynamic controllers for non-linear fluid/structure interactions. These ideas also produced a new methodology for attacking the fundamental problem of optimal sensor-actuator location for distributed parameter feedback control.

3.2 Highlights of the Research Activities

During the period 1 May 1993 to 30 April 1997 the DOD URI *Center For Optimal Design And Control* (**CODAC**) at Virginia Tech made significant and fundamental advances in the development and application of completely new approaches to optimal design and control of distributed parameter

aerospace systems. The following projects provide a sample of the breath of the research and were particularly noteworthy:

CONTROL OF FLUIDS AND STRUCTURES

- **Computational Approach to Sensor/Actuator Location for Feedback Control of Fluid Flow Systems**

  The problem of selecting and placing sensor/actuator pairs for optimal performance and robustness of feedback controllers is extremely complex so that designers often resort to trial and error methods. This process can be costly and increases the design cycle time. A fundamental issue in applications may be stated as the following three part question: Given that one desires to implement the most robust feedback control law, (i) what needs to be sensed, (ii) what types of sensors and actuators should be used, and (iii) where should these devices be located in order to maximize control effectiveness?

  Professor John Burns at Virginia Tech and Professor Belinda King at Oregon State have developed an entirely new approach to the problem of optimal sensor placement for controlling systems governed by hyperbolic and parabolic partial differential equations. These models have applications to control of fluid flows, control of structure/fluid interactions, flutter suppression and control of thermal processes. Burns and King have constructed a mathematical framework to address problems of this nature and used this framework to solve the problem of optimal sensor placement for various nonlinear control problems involving fluid dynamics and hybrid structural systems. Distributed parameter theory is used to determine the important physical quantities to be sensed and thus to provide insight into the types of sensors needed. The innovation was to assume a general form of the actuator, obtain an integral representation of the corresponding robust control law and then use rigorous approximation theory to compute finite dimensional sub-optimal controllers. These sub-optimal controllers are then used to attack the problem of optimal sensor placement. One of the main breakthroughs was the development of a new mathematical theory that provided explicit integral representations needed in the implementation of this approach. This research provides a key step toward the development of a rigorous and rational methodology for attacking the problems of sensor selection and placement in the design of aerospace control systems.

- **Reduced Basis Approach to Design of Low Order Feedback Controllers**

  The development of practical feedback controllers for nonlinear partial differential equations is one of the most basic problems that must be faced before one can address active control problems for fluid flows and nonlinear structures. One of the main difficulties is the construction of the low-order nonlinear observer needed to estimate those states that can not be sensed. In fluid flows the Navier Stokes equations serve as the fundamental model. It is common practice to first discretize the distributed parameter model and then to use the corresponding lumped parameter model in the controller design. Although this approximate-then-design method can work for simple problems, when applied to complex flow control and nonlinear vibration problems it produces large- order, and hence, non-practical observers. It is also well known that this approach can lead to erroneous results and one must exercise care to ensure that the resulting design is robust.

  Professor John Burns of Virginia Tech, along with several other AFOSR-supported researchers such as Gal Berkooz (Beam Engineering) and Belinda King (Oregon State) have achieved major advances in the development of an approach to overcome these difficulties and applied this method to structural vibration problems. By introducing numerical approximations at the last stage of the design, they were able to increase robustness, enhance performance and achieve a practical design. The new method proceeds in steps. First, distributed parameter control theory is used to compute robust control laws for a linearized model. The spatial behavior of these laws is analyzed to determine those regions in space where the state of the

dynamical system is most needed in the implementation of the feedback control law. Local low dimensional dynamical systems are then constructed by either certain projection schemes or reduced- basis finite element techniques. The resulting nonlinear system is then used as an observer for the nonlinear system to be controlled. This method has been applied to several nonlinear systems. In a recent application to vibration suppression, this reduced basis approach not only produced a low-dimensional controller, the controller was shown to be more robust than the full-order controller. This new design methodology generates very simple control laws that, combined with low order nonlinear observers, have proven to be quite effective.

- **Global Solvability for Damped Nonlinear Hyperbolic Systems**

  The development of practical feedback controllers for nonlinear partial differential equations is one of the most basic problems that must be faced before one can address active control problems for smart materials and nonlinear structures. One of the main difficulties is the construction of the practical computational schemes that are appropriate for control design and optimization. For new rubber-based materials, the constitutive relationships are highly nonlinear in both material and geometry. Standard mathematical theories do not cover such cases and there is a need to develop the basic existence and well-posedness of mathematical models. This foundation is used in the development of rigorous approximation algorithms. H. T. Banks and co-workers have established the well-posedness of these systems and used these results to obtain the regularity needed to establish convergence of Galerkin approximations. This effort provides the foundation that was needed to build a framework for attacking the complex distributed parameter control problems in smart structures.

- **An Experimentally Validated Damage Detection Theory in Smart Structures**

  A method for non-destructive detection and location of damage using parameterized partial differential equations and Galerkin approximation techniques is presented. Damages in a structure cause changes in the physical coefficients of mass density, elastic modulus and damping coefficients. This section examines the use of beam like structures with piezoceramic sensors and actuators to perform identification of those physical parameters, and hence to detect the damage. The method casts the inverse problem as an optimization problem. The iterative method is based on enhanced least-square error minimization. Experimental results are presented from tests on cantilevered aluminum beams damaged at different locations and with damage of different dimensions. It is demonstrated that the method can sense the presence of damage, and locate and characterize the damage to a satisfactory precision.

- **Feedback Control of a 2D Thermal Fluid**

  This accomplishment illustrates the application of modern distrubuted parameter control to the development of practical controllers for Navier-Stokes flows. Dr. D. Rubio developed both linear and non-linear feedback control laws for boundary control of a heated fluid in a thermal convection loop. These laws were tested on models of the flow ranging from a simple one mode model to a high fidelity finite element model. The method made use of optimal feedback theory for partial differential equations and combined computational mathematics with careful modeling to produce effective control laws. This effort is the first to show that distributed parameter control can produce simple practical feedback laws for complex non-linear flows. In addition, it was shown that making use of the PDE to guide controller design yields valuable information concerning the importance of the non-linearity in design. In particular, if one can accurately resolve the flow, then non-linearities can be used to assist in the controller design. This work is being extended to more complex fluid flow problems and is providing the foundations for a better understanding of the role that non-linearity plays in flow control.

OPTIMAL DESIGN

- **Optimal Design of a Forebody Simulator**

  The aerodynamic performance of an aerospace system or its components can often be enhanced by careful tailoring of shapes and fairings. Computational tools to aid in the design of these objects can enhance performance and reduce development costs. Such tools combine computational fluid dynamics (CFD) and optimization methods. Efficient optimization schemes require information about both the flow and also its sensitivity to design changes. Moreover, standard finite-difference approximations are far too costly in computational resources. As an alternative, we use the basic continuum model to derive sensitivity equations (SE) - linear partial differential equations describing the flow sensitivity - which are then solved in discrete approximation.

  We have applied this method to several problems. These problems were motivated by the design of the Free-Jet Test Facility at the Arnold Engineering Development Center. The combined flow/sensitivity code is a modification of the PARC analysis code used at AEDC. This was run on the C-90 at USAE Waterways Experiment Station, Vicksburg MS. A 2D problem was used to test the method. We used a Quasi-Newton/Trust Region algorithm and gradient information from the SE method. Timing test show that for a problem of modest size the SE method produces 50% savings in computer time. More importantly, the method lends itself to parallel implementation and for large problems the savings could be 90%. This approach is now the basis for several joint ventures with Air Force Laboratories and industrial partners. In particular, we are in the process of transitioning this into software products.

- **Asymptotically Consistent Gradients in Optimal Design**

  This accomplishment is based on a research effort that started with a project to develop a new computational algorithm for aerodynamic design. Researchers at **CODAC** have developed what is now known as the Sensitivity Equation Method (SEM) for optimal design. The fundamental idea is to use the partial differential equations that describe the flow sensitivities (to given parameters) as a basis for the development of approximate gradients in the optimization loop. However, there were several theoretical and practical issues that needed to be addressed before this method could be placed on a sound theoretical footing. Dr. Jeff Borggaard, working with other researchers at **CODAC**, obtained the first theoretical convergence results for this method. In addition, these results are now being transitioned into major CFD codes around the world.

- **Reduced-Hessian Methods for Optimal Design**

  Optimal design problems incorporating compressible flow models present some interesting challenges for optimization algorithms. Since the velocity-field can be discontinuous, it happens that the flow solution can be a non-smooth function of design parameters. To overcome this difficulty we employ a shock-fitting formulation so that the shock location is explicit. We formulate the design-optimization problem in a modern SQP-setting (Sequential-Quadratic Programming). This means that the flow variables, the geometric design variables and the shock location are all treated as independent variables and that the Euler flow equations (including explicit shock-jump conditions) are imposed as constraints. The *natural* problem structure is exploited in the *null-space* representation of the linearized constraint operator. This leads to a separation of the variables into *primary* and *secondary* categories; the linearized constraints can be uniquely solved for the *secondary* variables given values of the *primary* variables. Following this idea we construct a reduced Hessian involving the *primary* variables only. Numerical results have shown considerable efficiencies. While there is a modest cost in linear algebra, we obtain optimized designs with a 90% reduction in Euler flux evaluations.

- **Airfoil Design by an All-At-Once Method**

Abstractly, the optimum airfoil design problem can be formulated as a constrained optimization problem and many techniques have been applied to its solution. Most of the recent approaches combine optimization techniques with computational fluid dynamics. In many treatments the flow variables q are viewed as functions of the design parameters w. This function $q(w)$ is implicitly defined by the governing flow equations ($R(q, w) = 0$). In the present case of steady, compressible these are the 2-D Euler equations of fluid dynamics. This is the view in the so-called black-box approach. The Euler equations are, in a sense, hidden from the optimizer by eliminating the flow variables.

The all-at-once formulation provides an alternative approach in which one views the flow variables q and the design variables w as independent variables in the optimization problem. The Euler equations, which couple these together, are included in the optimization problem as a constraint, possibly along with other constraints. The optimizer now iteratively computes a sequence of points that move toward optimality and feasibility at the same time (*i.e.* all-at-once).

Professors Heinkenschloss and Cliff along with A. Shenoy (a recent Ph.D. student) have developed a formulation and related software for this class of problems. The software combines an existing research code for 2-D compressible flows with Heinkenschloss' trust-region optimization algorithm TRICE. The optimization algorithm requires that the flow code produce an increment in the flow variables (δq) corresponding to a step in the design variables (δw) so that the pair will satisfy a *linearized* version of the Euler equation. A related adjoint system must also be solved. The TRICE algorithm finds search directions to improve feasiblity and optimality. Trust-region ideas are used to adaptively compute a trust-region-radius and this provides a framework for establishing global convergence of the algorithms. The combined algorithm is quite efficient and leads to an optimized airfoil for approximately the computational effort required for five – six nonlinear flow solutions.

- **Optimal Shape Design in Forced Convection Using Adaptive Finite Elements**

  A common strategy in developing a method to solve these optimal design problems is to cascade an existing numerical scheme for the PDEs into a gradient-based optimization algorithm. This so-called black-box strategy allows one to retain trusted numerical software, while using a gradient-based optimization algorithm helps reduce the number of objective function evaluations. These evaluations require the solution of the PDEs and are typically expensive. We have considered a forced convection design problem which involves the two-dimensional steady Navier-Stokes and energy equations. An adaptive finite element solver was used for the approximation of these equations. Adaptive remeshing strategies are natural in the context of optimal design as the mesh used for the initial design is unlikely to be appropriate as the design changes. This methodology allows for constructing optimal design methods which require minimal intervention at the intermediate designs.

3.3 Highlights of Laboratory and Industrial Interactions

In addition to the development of a strong basic research program, **CODAC** developed and expanded its interactions with Air Force Laboratories and worked to transition the new methods into industry. A complete description of these projects may be found in Chapter 7. The following list indicates the breadth of this effort for the past four years.

3.3.1 Air Force Laboratories

- **Arnold Engineering and Development Center** Tullahoma, TN
  J. Benek, M. Briski, P. Hoffman, S. Keeling, K. Kneile, S. Tennent and D. Todd

- **Phillips Laboratory** Albuquerque, NM
  A. Das, J. Mason, D. Sciulli and A. Weston

- **Wright Laboratory** Wright Patterson PAFB, OH
  S. Banda and J. Malas

3.3.2 Industrial Partners

- **AeroSoft Inc.** Blacksburg, VA
  A. Godfrey and R. Walters

- **Analytical Mechanics Associates** Hampton, VA
  R. Kumar and H. Seywald

- **Aurora Flight Sciences Inc.** Manasas, VA
  M. Hutchinson

- **BEAM Engineering and Applied Research** Ithaca, NY
  G. Berkooz and Richard Newsome

- **Sverdrup** Tullahoma, TN
  J. Benek, P. Hoffman and S. Keeling

- **Tektronix Graphics, Printing and Imaging Division** Wilsonville, OR
  S. Berger, R. Burr and P. Gilmore

3.4 Other Activities

- **Interactions and Educational Program**

  During this four year period, **CODAC** researchers have given more than 150 lectures at national and international conferences. Over 35 graduate students have worked on **CODAC** projects under AFOSR funding. Many of these students have visited Air Force laboratories and industrial sites. In addition, **CODAC** was asked to make a presentation on how AFOSR funded research supports the national defense. This presentation took place on May 9, 1996 in Washington, DC and was sponsored by the Association of American Universities.

- **Visitors Program**

  During this period more than 160 visitors, representing 13 different countries, have visited **CODAC**. A partial list of visitors is given in Chapter 10.

- **Students and Post-Doctoral Researchers**

  During this period **CODAC** provided partial support for 4 undergraduate, more than 35 graduate students and 5 post-doctoral scientists. Chapter 5 contains the list of people supported under this contract.

- **Workshop on Optimal Design and Control**

  In April 1993 we organized a workshop on optimal design and control. The workshop was attended by more than 90 scientists from four countries and included more than 20 attendees from government laboratories. The proceedings of the workshop were published as *Optimal Design and Control* in the Birkhäuser series *Progress in Systems and Control Theory.*

Chapter 4

Accomplishments and New Findings

This chapter contains a detailed summary of the research highlights discussed in Chapter 3. These projects were selected because they are indicative of the innovative new ideas produced under this grant and they illustrate the breadth of the accomplishments. Since this final technical report covers a four year period, we present summaries of projects from each of the three previous years and three new projects for the fourth and final year.

Control of Fluids and Structures

The five projects described below provide samples of the research conducted in the areas of feedback control and sensor/actuator location for distrubuted parameter control systems.

4.1 Computational Approach to Sensor/Actuator Location for Feedback Control of Fluid Flow Systems

In this section, we discuss a computational approach to sensor/estimator design for feedback control of fluid dynamic systems. This approach is based on combining minmax compensator design with piecewise constant approximations of "optimal feedback gains". A driven cavity flow control problem is presented to illustrate the idea and to demonstrate the feasibility of this approach, which leads to design of low order practical controllers for the Navier-Stokes equations. A finite difference Galerkin scheme with divergence free basis is applied to the infinite dimensional system to illustrate the feasibility of this approach. Active flow control, specifically, control of fluids based on the Navier-Stokes equations is an important area which has received much attention in recent years. Research has been devoted to all aspects of this problem, including theoretical, computational and experimental. Of considerable importance is the development of practical, low order controllers. We consider a specific approach based on combining linear state feedback with non-linear (low order) observers. Consequently, questions about types and locations of sensors arise.

To answer these questions in the context of the driven cavity flow problem, we make use of the functional gains arising from minmax design. The gains motivate not only what type of sensors are useful, but also the "best" place to locate these sensors to gain the most information. To show the viability of the scheme, we implement it computationally.

4.1.1 Problem Formulation

The driven cavity flow problem is posed as follows. Let Ω denote an open bounded set in $\mathbf{R}^2$ with boundary Γ. In this summary, we shall assume that the cavity is the square as shown in Figure 4.1. On the interior, the fluid flow is governed by the Navier-Stokes equations

$$\begin{cases} \dfrac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} = -\nabla p + \nu\nabla^2\mathbf{u} \qquad \text{in } \Omega\times[0,T]; \\ \nabla\cdot\mathbf{u} = 0 \quad \text{in } \Omega\times[0,T], \end{cases} \tag{4.1}$$

where $\mathbf{u} = (u, v)$ represents the velocity of the fluid in the x and y directions, respectively, and p denotes the pressure. Also, ν is the kinematic viscosity of the fluid. The boundary conditions are given on the side walls by

$$u|_{\Gamma_1} = u|_{\Gamma_3} = v|_{\Gamma_1} = v|_{\Gamma_3} = 0.$$

We assume that the top boundary, Γ_2, is moved horizontally by an "unknown disturbance". That is,

$$u|_{\Gamma_2} = \eta(t), \quad v|_{\Gamma_2} = 0.$$

Control is applied on the bottom boundary, Γ_4, leading to the boundary conditions

$$u|_{\Gamma_4} = g(t), \quad v|_{\Gamma_4} = 0.$$

The goal is to design a feedback control (using state estimation) to attenuate the disturbance $\eta(t)$. Measurements are provided by four wall shearing stress sensors placed on the vertical walls as shown in Figure 4.1. Thus, we have four outputs

$$\zeta(t) = \begin{bmatrix} \zeta_1(t) \\ \zeta_2(t) \\ \zeta_3(t) \\ \zeta_4(t) \end{bmatrix} = \rho\nu \begin{bmatrix} \dfrac{\partial v}{\partial x}(t,0,y_1) \\ \dfrac{\partial v}{\partial x}(t,0,y_2) \\ \dfrac{\partial v}{\partial x}(t,1,y_3) \\ \dfrac{\partial v}{\partial x}(t,1,y_4) \end{bmatrix},$$

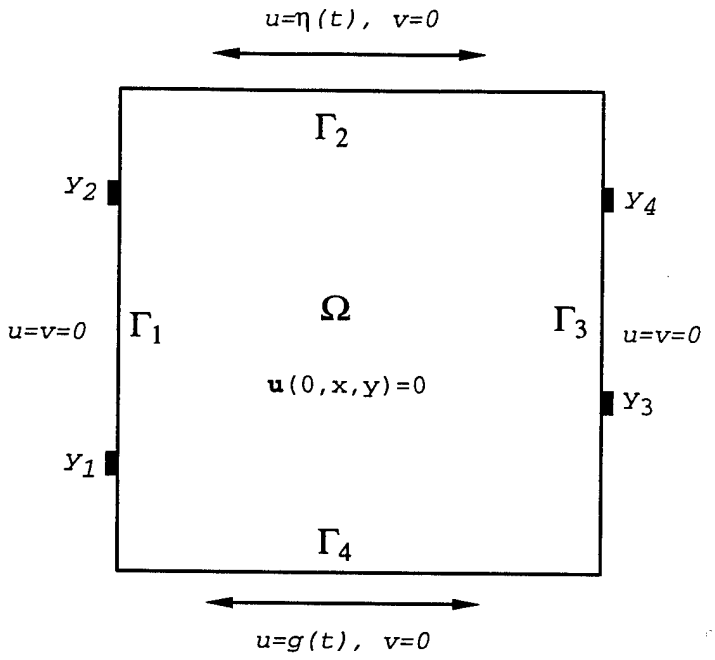


Figure 4.1: The driven cavity with a moving bottom wall.

where y_1, y_2 are the sensor locations on Γ_1 and y_3, y_4 are the locations on Γ_3. The placement of sensors need not be symmetric. We use a minmax design and low order finite dimensional approximations to construct a dynamic compensator for the full nonlinear Navier-Stokes system.

4.1.2 The Minmax Compensator

One can formulate the above system as an abstract distributed parameter model of the form

$$\frac{\partial \mathbf{u}(t)}{\partial t} = A\mathbf{u}(t) + N(\mathbf{u}(t)) + Bg(t) + D\eta(t), \quad \mathbf{u}(0) = \mathbf{u}_0, \tag{4.2}$$

with measured output

$$\zeta(t) = C\mathbf{u}(t) + E\eta(t), \tag{4.3}$$

in the state space of divergence free vector fields. We assume the linearized control system has the form

$$\frac{\partial z}{\partial t} = Az(t) + Bg(t) + D\eta(t), \quad z(0) = z_0 \tag{4.4}$$

with measured output

$$\zeta(t) = Cz(t) + E\eta(t). \tag{4.5}$$

The idea is to design a nonlinear compensator of the form

$$\begin{cases} \dfrac{\partial \mathbf{u}_c}{\partial t} = A_c\mathbf{u}_c(t) + N_c(\mathbf{u}_c(t)) + F_c\zeta(t) \\ g(t) = -K_c\mathbf{u}_c(t), \end{cases}$$

and then construct low order approximations to $\mathbf{u}_c(t)$. The simplest idea is to set $N_c(\mathbf{u}) = N(\mathbf{u})$ and use a linear design method to compute A_c, F_c and K_c. Here, we use a "minmax control" design

for infinite dimensional systems. The idea is to first solve the Riccati equations

$$A^T P + PA - P\left[BR^{-1}B^T - \theta^2 M\right] P + Q = 0 \tag{4.6}$$

$$A\Pi + \Pi A^T - \Pi\left[C^T N^{-1} C - \theta^2 Q\right] \Pi + M = 0. \tag{4.7}$$

Assuming (for small θ) that (4.6) has a minimal solution $P_\theta \geq 0$, (4.7) has a minimal solution $\Pi_\theta > 0$ and

$$\left[\Pi_\theta^{-1} - \theta^2 P_\theta\right] > 0,$$

one defines

$$\begin{cases} K_\theta = R^{-1} B^* P_\theta, \\ F_\theta = \left[I - \theta^2 P_\theta \Pi_\theta\right]^{-1} \Pi_\theta C^* N^{-1}, \\ A_\theta = A - BK_\theta - F_\theta C + \theta^2 M \Pi_\theta. \end{cases}$$

The corresponding "minmax control" can be written

$$g_{opt}(t) = -B^T P_\theta \mathbf{u}_c(t) = -K_\theta \mathbf{u}_c(t).$$

The nonlinear closed loop system formed by the above design can be written as

$$\frac{d}{dt}\begin{bmatrix} \mathbf{u}(t) \\ \mathbf{u}_c(t) \end{bmatrix} = \begin{bmatrix} A & -BK_\theta \\ F_\theta C & A_\theta \end{bmatrix} \begin{bmatrix} \mathbf{u}(t) \\ \mathbf{u}_c(t) \end{bmatrix} + \begin{bmatrix} N(\mathbf{u}(t)) \\ N(\mathbf{u}_c(t)) \end{bmatrix} + \begin{bmatrix} D \\ F_\theta E \end{bmatrix} \eta(t). \tag{4.8}$$

4.1.3 Numerical Results and Discussions

Using a finite difference-Galerkin scheme, we discretized the variational form of the system to obtain a sequence of finite dimensional systems suitable for computation. This approach is a discrete analog of the Galerkin technique used in finite element methods.

There are two controller "gains" for this model corresponding to horizontal and vertical velocities. Thus, we assume that the feedback control for the system can be represented by the integrals

$$g(t) = -\int_\Omega K_u(x,y) u_c(t,x,y) dxdy - \int_\Omega K_v(x,y) v_c(t,x,y) dxdy, \tag{4.9}$$

where $K_u(x,y)$ and $K_v(x,y)$ are the so-called "functional gains" and $\mathbf{u}_c(t) = (u_c, v_c)$ is the estimate of the state.

The parameters chosen for the computations are $Re = 100$ and $\theta = 0.001$. Figure 4.2 shows the computed feedback gains $K_u(x,y)$ and $K_v(x,y)$ with variation of the grid size. The computational results demonstrate the "convergence" of both components of the feedback gain functions as the mesh size decreases.

Notice that in certain region of the domain the gains are negligibly small. Thus, we can approximate the control law (4.9) by taking an M-th order projection of the gains onto characteristic functions and making use of the fact that the functions are nearly zero except in small regions near the boundary. In particular, let

$$K_u(x,y) \approx K_u^M(x,y) = \sum_{i,j=1}^{M} (K_u)_{ij}^M \chi_{ij}(x,y),$$

and

$$K_v(x,y) \approx K_v^M(x,y) = \sum_{i,j=1}^{M} (K_v)_{ij}^M \chi_{ij}(x,y),$$

where $\chi_{ij}(x,y)$ is the characteristic function of the i,j grid element. With these piecewise constant approximations of the functional gains, the control law (4.9) is approximated by

$$g_M(t) = -\sum_{i,j=1}^{M} (K_u)_{ij}^M \int_{\Omega_{ij}} u_c(t,x,y)dxdy - \sum_{i,j=1}^{M} (K_v)_{ij}^M \int_{\Omega_{ij}} v_c(t,x,y)dxdy. \tag{4.10}$$

Observe that in (4.10) one needs only the spatial averages of the velocity fields over subdomains. Using the same finite difference-Galerkin scheme for $\mathbf{u}_c$, one obtains another level of approximation

$$g_M^N(t) = -\sum_{i,j=1}^{M} (K_u)_{ij}^M \int_{\Omega_{ij}} u_c^N(t,x,y)dxdy - \sum_{i,j=1}^{M} (K_v)_{ij}^M \int_{\Omega_{ij}} v_c^N(t,x,y)dxdy, \tag{4.11}$$

where (u_c^N, v_c^N) is the "low order" approximation of the distributed parameter observer equations. Observe that it is not necessary for the computational mesh to be "super fine" in order that the integral averages

$$\int_{\Omega_{ij}} u_c^N(t,x,y)dxdy \quad \text{and} \quad \int_{\Omega_{ij}} v_c^N(t,x,y)dxdy$$

be "close" to the integrals

$$\int_{\Omega_{ij}} u_c(t,x,y)dxdy \quad \text{and} \quad \int_{\Omega_{ij}} v_c(t,x,y)dxdy$$

needed in (4.10). This provides the possibility of constructing robust finite dimensional (low order) controllers.

There are two observer "functional gains" for each sensor. In particular, the observer gain operator has the form

$$F_\theta\, \zeta(t) = \sum_{i=1}^{4} \zeta_i(t) \begin{bmatrix} f_{iu}(x,y) \\ f_{iv}(x,y) \end{bmatrix}.$$

Figures 4.3-4.5 show these "functional gains" (f_{iu}) and (f_{iv}) for various sensor locations. Observe that the observer gains are supported on neighborhoods of the sensor location.

4.1.4 Conclusion

Although we have described, in loose terms, an algorithm for low order control design, much work remains to be done before a sound theoretical framework can be completed. However, preliminary numerical results indicate that this approach offers considerable promise for practical design.

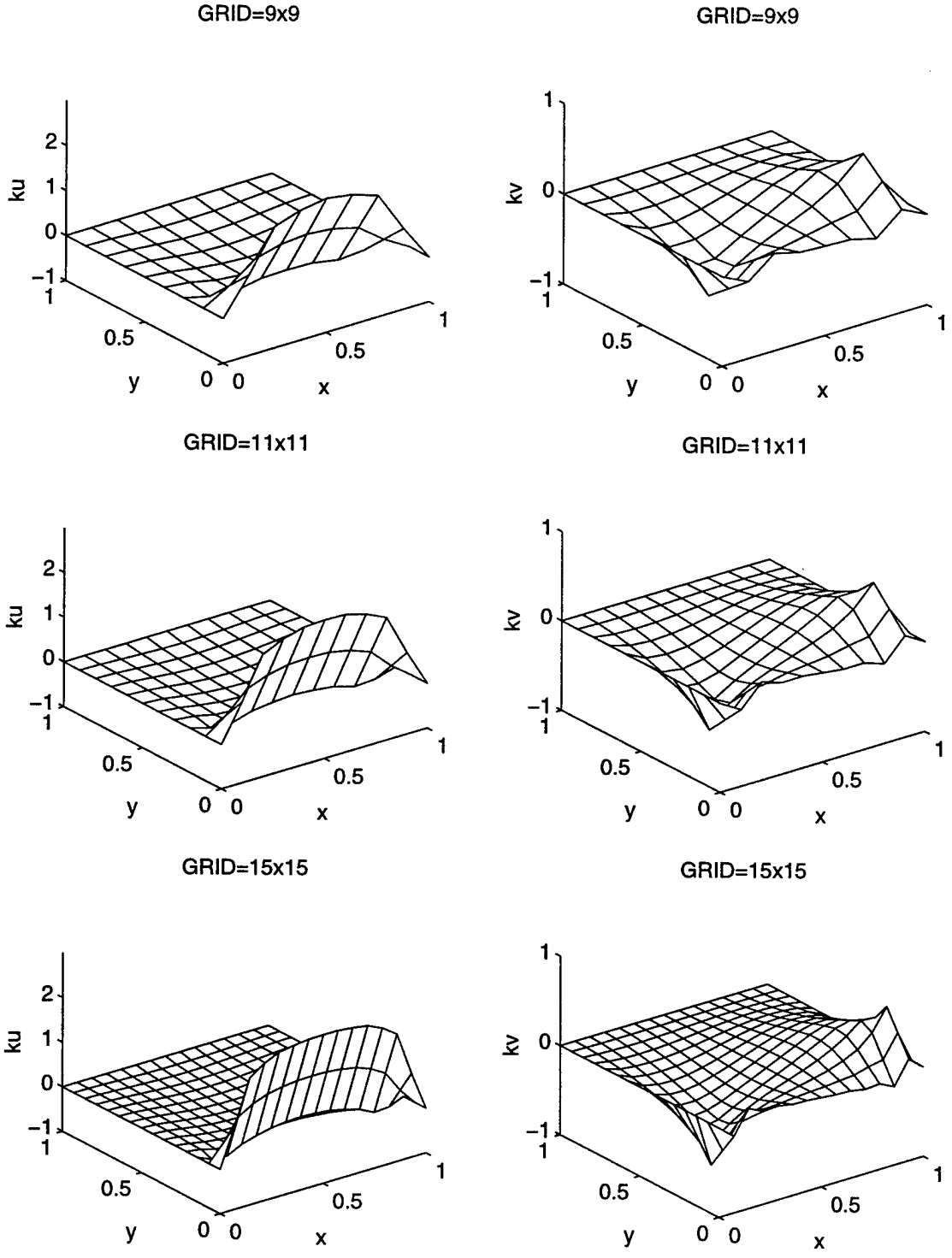


Figure 4.2: The feedback functional gains

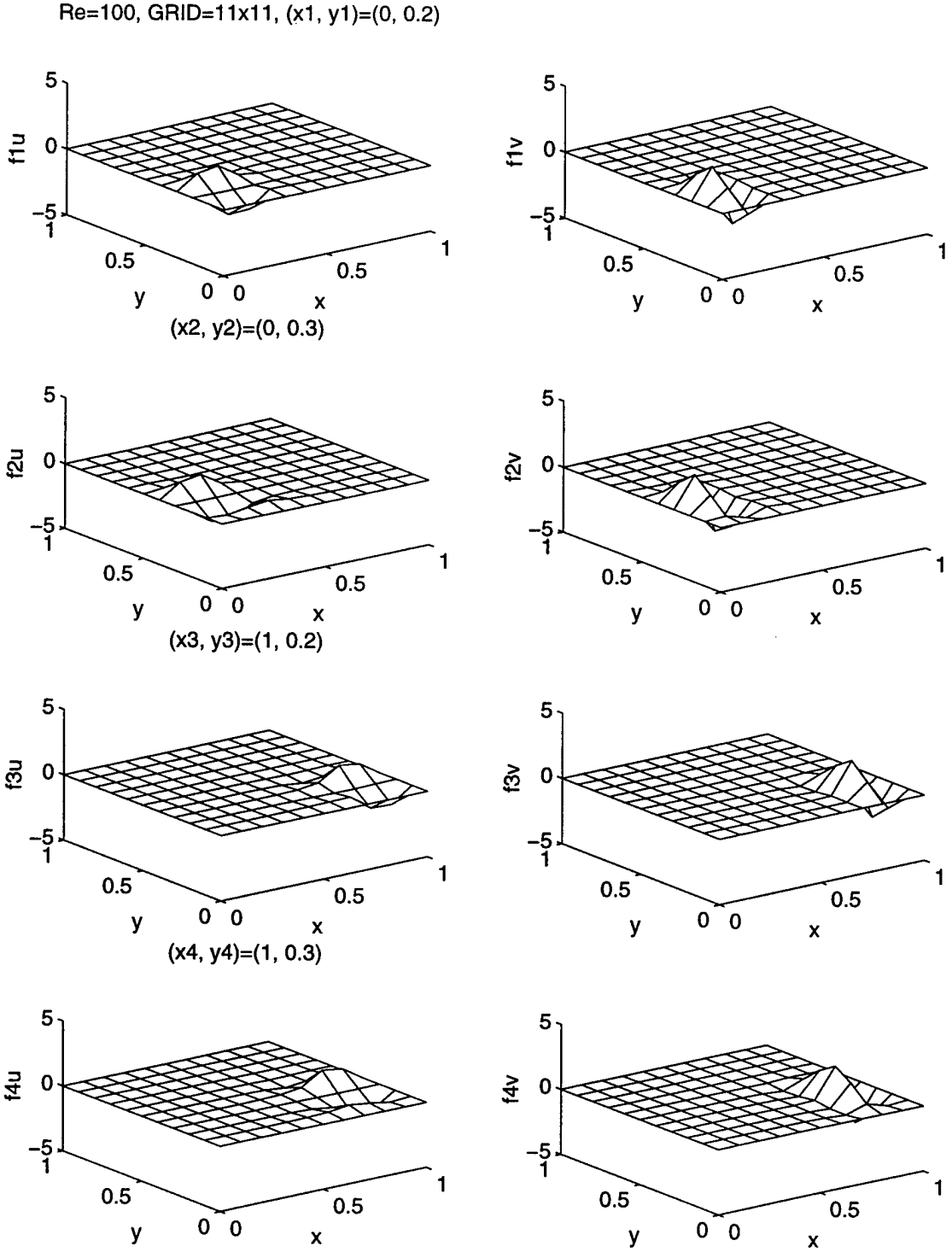


Figure 4.3: The observer functional gains, $(y_1, y_2, y_3, y_4) = (0.2, 0.3, 0.2, 0.3)$

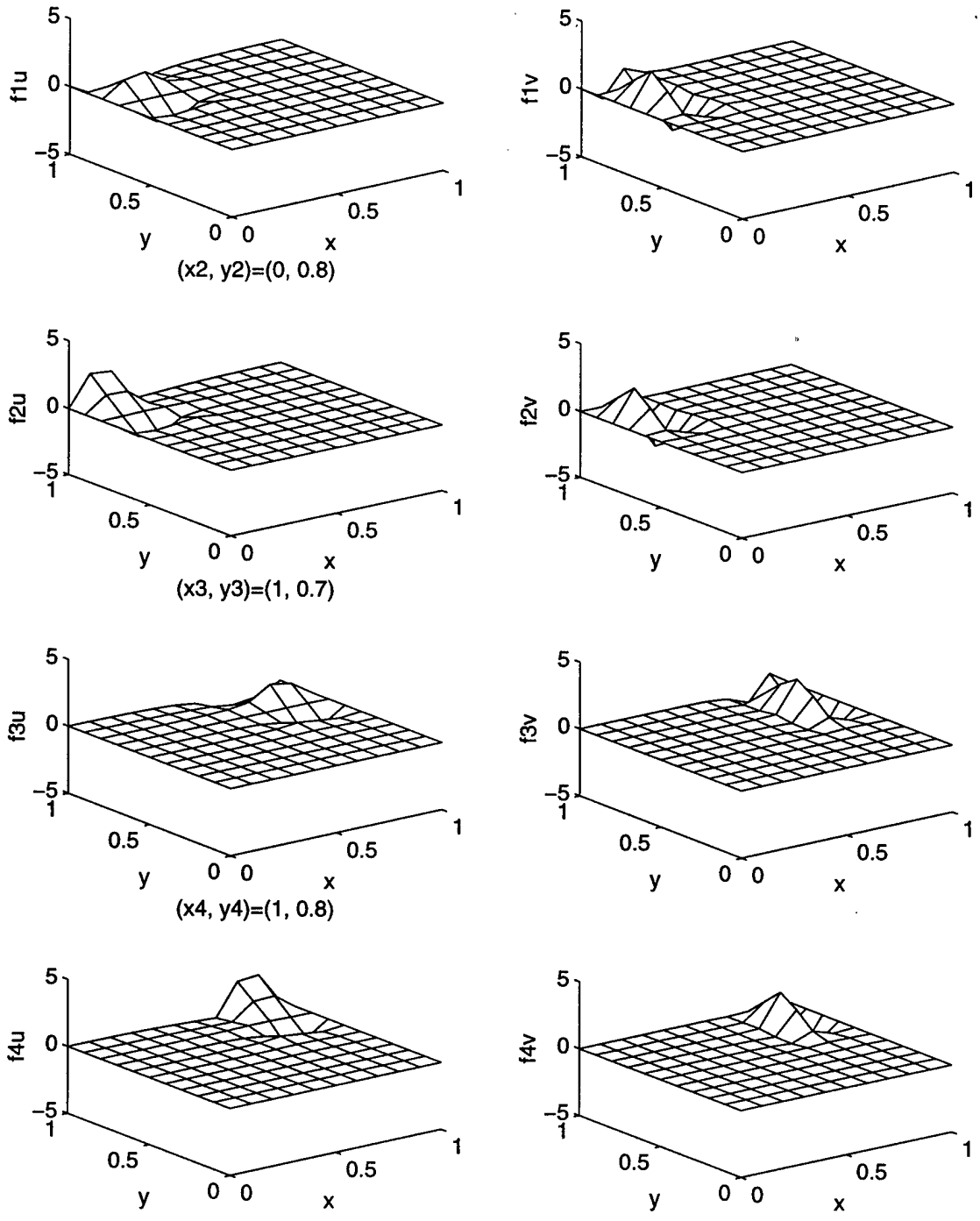


Figure 4.4: The observer functional gains, $(y_1, y_2, y_3, y_4) = (0.7, 0.8, 0.7, 0.8)$

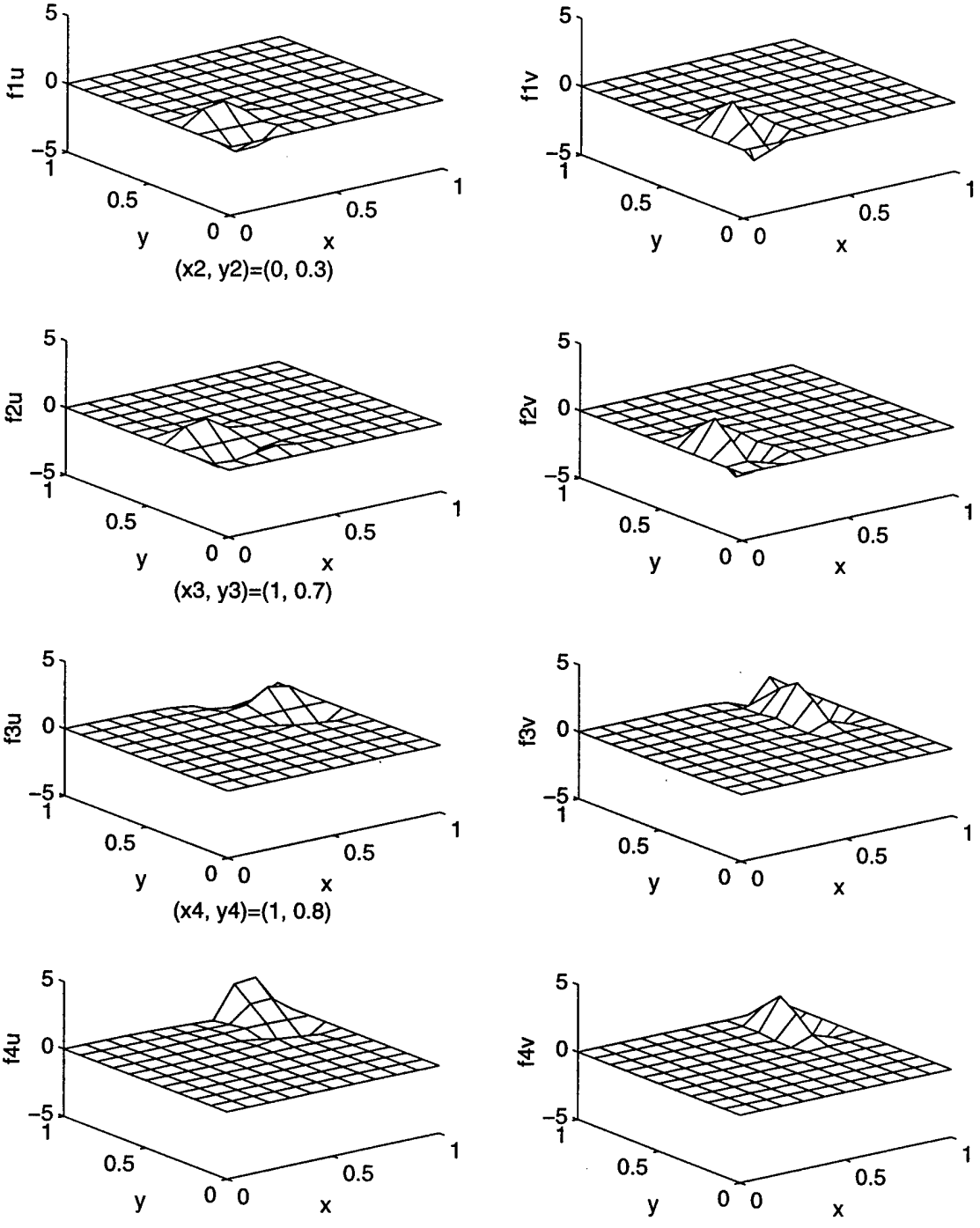


Figure 4.5: The observer functional gains, $(y_1, y_2, y_3, y_4) = (0.2, 0.3, 0.7, 0.8)$

4.2 Reduced Basis Approach to Design of Low Order Feedback Controllers

In this section, we discuss an approach to the development of low order nonlinear feedback controllers for hybrid distributed parameter systems. This approach involves the use of distributed parameter control theory to design "optimal" infinite dimensional feedback control laws and approximation theory to design and compute low order finite dimensional compensators. The resulting finite dimensional controller combines a nonlinear observer with a linear feedback law to produce a practical design. We concentrate on a weakly nonlinear distributed parameter system to illustrate the ideas.

The development of feedback controllers for nonlinear partial differential equations (PDEs) is one of the most basic problems that must be faced before one can address active control problems for fluid flows and nonlinear structures. In fluid flows the Navier-Stokes equations serve as the fundamental model. It is common practice to first discretize the distributed model and then to use the corresponding lumped parameter model in the controller design. For example, a three mode Ritz-Galerkin discretization (one in velocity and two in temperature) applied to a thermal convection loop, reduces the Boussinesq equations for fluid convection to a hybrid system of nonlinear ordinary differential equations (i.e., the Lorenz equations) coupled to an infinite dimensional "linear" system. It can be shown that introducing truncated Bessel function approximations produces reasonable open-loop models for simulation. Several papers have now appeared that discuss the feasibility of using feedback to control and stabilize these finite dimensional chaotic systems. Although this "approximate-then-design" method often works well, it is also well known that this approach can lead to erroneous results and one must exercise care to ensure that the resulting design is robust. Moreover, the information lost in modal truncation can be used to enhance performance (even when one "truncates only linear terms") and to ensure robust feedback design.

Using the "approximate-then-design" approach to feedback control of distributed parameter systems, a simple example of a nonlinear system can be constructed with the property that many standard discretized lumped models failed to capture the essential nonlinear behavior of the dynamic system governed by the partial differential equation. We shall use this example to illustrate a "design-then-approximate" approach to the control nonlinear distributed parameter systems. The basic idea is to use infinite dimensional control theory to design a feedback law and then use rigorous numerical approximations to construct practical finite dimensional controllers. The goal is to show how the method can be applied to PDE systems and to illustrate the benefits of using distributed parameter control theory to help guide the design of practical controllers. For this reason we do not attempt to design the "best" nonlinear controller. In particular, we limit our discussion to nonlinear controllers defined by linear feedback laws combined with low order nonlinear observers. We compare a distributed parameter LQG-type design (extended Kalman filter), for which the theory is complete, with a distributed parameter MinMax-type design, even though the MinMax approximation theory is not complete. The MinMax approach provides a "robust state feedback control law" which is less sensitive to disturbances and certain unmodeled dynamics than is the LQG design.

The idea is to obtain a representation of the linear control law and then use approximation theory to compute finite dimensional suboptimal controllers. These suboptimal controllers can then be used in conjunction with reduced basis ideas to design low order nonlinear state estimators. It is important to note that any practical feedback controller designed for a distributed parameter system must incorporate some type of "state estimator" and, regardless of the approach, one must introduce approximations at some point in the analysis. We introduce approximations at the last stage of the design. It is also important to note that the resulting feedback controller is nonlinear and finite dimensional.

The Model

We consider a hybrid distributed parameter system. This model represents the nonlinear dynamic response of a relief valve used to protect a pneumatic system from overpressure. The valve mechanics

consist of a ball pressed by a uniform helical spring against a valve seat having nonlinear elastic characteristics. The helical spring is considered to be a distributed parameter system, and its motion is governed by the wave equation subject to appropriate boundary conditions. Specifically, the spring is fixed at one end. In addition to the forces exerted by the spring and the valve seat, the ball is subjected to a pressure where it comes in contact with the fluid. The pressure consists of a static component and a sinusoidally varying dynamic component. This sinusoidal term results from small vibrations in the pneumatic transmission lines. The equations for this hybrid system are

$$\rho \frac{\partial^2}{\partial t^2} w(t,s) = \frac{\partial}{\partial s}\left[\tau \frac{\partial}{\partial s} w(t,s) + \gamma \frac{\partial^2}{\partial t \partial s} w(t,s)\right], \quad 0 < s < \ell,\ t > 0, \tag{4.12}$$

$$m \frac{\partial^2}{\partial t^2} w(t,\ell) = -\left[\tau \frac{\partial}{\partial s} w(t,\ell) + \gamma \frac{\partial^2}{\partial t \partial s} w(t,\ell)\right] - \alpha_1 w(t,\ell) - \alpha_3 [w(t,\ell)]^3 + \eta(t) + u(t), \tag{4.13}$$

with boundary condition

$$w(t,0) = 0. \tag{4.14}$$

To obtain a solution to the system, initial conditions are chosen of the form

$$w(0,s) = w_0(s), \quad \frac{\partial}{\partial t} w(0,s) = w_1(s). \tag{4.15}$$

Here, $w(t,s)$ represents the displacement of the spring at time t, position s, $w(t,\ell)$ represents the position of the mass at time t, ρ and m are the densities of the spring and mass respectively, τ is Young's modulus for the spring, and γ is a damping coefficient. The α_i's are the coefficients describing the nonlinear effects of the valve seat. The term $\eta(t)$ is viewed as a disturbance and $u(t)$ is a control input.

Another way of viewing this model is to imagine an elastic cable which is fixed at one end and attached to a mass at the other. The mass is suspended by a spring which has nonlinear stiffening terms and is forced by a sinusoidal disturbance (see Figure 4.6). In this case, $w(t,s)$ represents the displacement of the cable at time t, and position s, $w(t,\ell)$ represents the position of the mass at time t, ρ and m are the densities of the cable and mass respectively, τ is the tension in the cable, and γ is a damping coefficient. The α_1 and α_3 are spring stiffness constants with the latter describing the nonlinear effects of the spring.

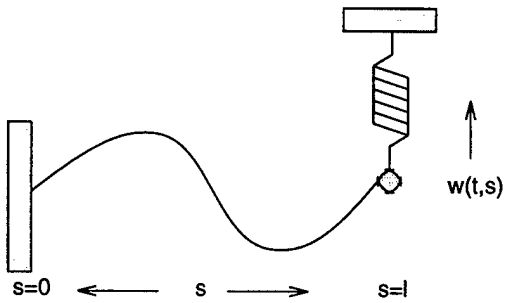


Figure 4.6: Cable-mass system

This view of the problem allows for a simple description of the control problem and clearly illustrates the hybrid nature of the system. In particular, the system is described by a linear partial differential equation (the wave equation) coupled, through the boundary condition, to a low order nonlinear ordinary differential equation (Duffing's equation). We are interested in using sensed information to design a feedback controller that attenuates the disturbance $\eta(t)$. In particular, we assume that the control is allowed to act exclusively on the mass and the only measured information

available to the controller is the position and velocity at the mass, i.e., there are two observations $y_1(t)$ and $y_2(t)$ where

$$y_1(t) = w(t,\ell) + \xi_1(t), \quad y_2(t) = \frac{\partial}{\partial t} w(t,\ell) + \xi_2(t), \tag{4.16}$$

and $\xi_1(t)$, $\xi_2(t)$ represent sensor noise.

4.2.1 Abstract Formulation

Herein, we apply what have now become "standard" techniques from distributed parameter control theory. In order to do this properly, care must be taken to formulate the second order (hyperbolic system) as a first order system in the proper state space. As will be shown later, the precise form of the state space provides essential information about practical issues concerning the placement of sensors and the design of the nonlinear observer.

This model is often first written as a second order system in a Hilbert space H of the form

$$\ddot{\zeta}(t) + D_0\dot{\zeta}(t) + A_0\zeta(t) + F_0(\zeta(t)) = B_0u(t) + G_0\eta(t) \tag{4.17}$$

where we use the notation $\dot{\zeta} = \frac{\partial}{\partial t}\zeta$. For the cable-mass problem considered here, $H = L_2(0,1) \times \mathbb{R}^1$ and $\zeta(t)$ is given by $\zeta(t) = [w(t,\cdot), w(t,\ell)]^T$. In addition, as we consider Kelvin-Voigt damping in our system, $D_0 = A_0$.

This formal system has the advantage that it has the same appearance as the finite dimensional case and in order to address viscous and "structural" damping one merely replaces $D_0 = A_0$ with $D_0 = I$ and $D_0 = [A_0]^{1/2}$, respectively. We note however, that it is more consistent with physics to write the system in the form

$$\ddot{\zeta}(t) + S^*(S\zeta(t) + \gamma T\dot{\zeta}(t)) + F_0(\zeta(t)) = B_0u(t) + G_0\eta(t) \tag{4.18}$$

where $S = T = [A_0]^{1/2}$. Observe that $A_0 = A_0^* > 0$ and so $S^* = S$ and $S^*S = S^*T = A_0$. Hence, (4.18) is formally obtained by factoring $[A_0]^{1/2}$ out of the expression $D_0\dot{\zeta}(t) + A_0\zeta(t)$ in (4.17). Note also that (4.18) is of a form that allows for structural damping where $S = [A_0]^{1/2}$ and $T = I$, as well as for viscous damping where $S = [A_0]^{1/2}$ and $T = S^{-1} = [A_0]^{-1/2}$. In addition, by writing the system in the second order form (4.18), one captures a form that comes from balance laws and at the same time sets the stage for a simple formulation of the problem in first order state space form.

The system governed by equations (4.12) - (4.16) can be written as a dynamical system in an appropriate (infinite dimensional) state space. Although there are several equivalent formulations for this problem, we shall write the governing equations as the first order system

$$\begin{aligned} \dot{x}(t) &= Ax(t) + F(x(t)) + Bu(t) + G\eta(t), \qquad x(0) = x_0 \\ y(t) &= Cx(t) + E\xi(t) \end{aligned} \tag{4.19}$$

where at time t the state $x(t) = [\zeta(t), \dot{\zeta}(t)]^T$ lies in the Hilbert space $X = H_L^1 \times \mathbb{R} \times L_2 \times \mathbb{R}$. Here, H_L^1 is the subspace of the Sobolev space $H^1 = H^1(0,\ell)$ defined by $H_L^1 = H_L^1(0,\ell) = \{w \in H^1 : w(0) = 0\}$, and $L_2 = L_2(0,\ell)$ is the standard Lebesgue space of square integrable functions. The control $u(t)$ lies in the control space $U = \mathbb{R}$; the observation (or state measurement) $y(t) = (y_1(t), y_2(t))^T$ belongs to $Y = \mathbb{R}^2$. Denote $w(t,\ell) = w_\ell$ and $\dot{w} = v$. Then the inner product in X is

$$\begin{aligned} \langle [w(\cdot), w_\ell, v(\cdot), v_\ell]^T, [\hat{w}(\cdot), \hat{w}_\ell, \hat{v}(\cdot), \hat{v}_\ell]^T \rangle = \tau &\int_0^\ell \frac{\partial}{\partial s} w(s) \frac{\partial}{\partial s} \hat{w}(s) ds + \alpha_1 w_\ell \hat{w}_\ell \\ &+ \rho \int_0^\ell v(s)\hat{v}(s) ds + m\hat{v}_\ell \hat{v}_\ell. \end{aligned} \tag{4.20}$$

It is important to precisely define the system operators and their domains in order to obtain correct representations of the feedback operators that will be used to control the system. Let δ_ℓ denote the "evaluation operator" defined on $H^1(0,\ell)$ by $\delta_\ell(\phi(\cdot)) = \phi(\ell)$ and define the linear operator A on the domain $\mathcal{D}(A) \subseteq X$ by

$$\mathcal{D}(A) = \left\{ x = [w, w_\ell, v, v_\ell]^T \in X : w, v \in H^1_L, \left\{ \frac{\tau}{\rho}\frac{d}{ds}w + \frac{\gamma}{\rho}\frac{d}{ds}v \right\} \in H^1, \right. \\ \left. w(\ell) = w_\ell, v(\ell) = v_\ell \right\}, \tag{4.21}$$

and

$$Ax = \left[v, v_\ell, \frac{d}{ds}\left\{ \frac{\tau}{\rho}\frac{d}{ds}w + \frac{\gamma}{\rho}\frac{d}{ds}v \right\}, -\delta_\ell\left\{ \frac{\tau}{m}\frac{d}{ds}w + \frac{\gamma}{m}\frac{d}{ds}v \right\} - \frac{\alpha_1}{m}w_\ell \right]^T. \tag{4.22}$$

The control input operator B, the disturbance operator G and the output operator C are defined by

$$B = \left[0, 0, 0, \frac{1}{m} \right]^T = G \text{ and } C = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{4.23}$$

respectively. The operator E is the 2×2 identity and the nonlinear operator F is defined on X by

$$F(x) = F([w(\cdot), w_\ell, v(\cdot), v_\ell]^T) = \left[0, 0, 0, -\frac{\alpha_3}{m}[w_\ell]^3 \right]^T = [0, 0, 0, F_0(w_\ell)]^T \tag{4.24}$$

where $F_0 : R \to R$ is continuous.

Before discussing the control problem we present the basic properties of the dynamical system defined by (4.21)-(4.22). First we note that the linearized problem is well posed and analytic. The important issue here is that A generates a exponentially stable analytic semigroup $S(t) = e^{At}$ on X.

Theorem 1 *The operator A in defined by (4.21)-(4.22) is the infinitesimal generator of an analytic semigroup $S(t)$ on X satisfying $\|S(t)\| \leq Me^{-\mu t}$ for some $\mu > 0$.*

Although we consider only Kelvin-Voigt damping here, the above theorem holds for structural damping where $D_0 = [A_0]^{1/2}$ or, in fact, for any damping operator of the form $D_0 = [A_0]^\alpha$ for $\frac{1}{2} \leq \alpha \leq 1$. Moreover, the following theorem shows that the nonlinear system also has global solutions for each $x_0 \in X$, and that the zero solution is an exponentially equilibrium.

Theorem 2 *For each $x_0 \in X$, the unforced nonlinear system (4.19) has a unique global mild solution $x(t)$ and there exists a constant $N(x_0)$ such that $\|x(t)\| \leq N(x_0)$ for all $t \geq 0$. If $x_0 \in \mathcal{D}(A)$, then the mild solution is a classical solution. Moreover, there exist $\lambda > 0, \beta > 0$ and $M > 0$ such that if $\|x_0\| < \lambda/2M$, then $\|x(t)\| < 2Me^{-\beta t}\|x_0\|$.*

Proof. Given $M > 0$, let c be a constant such that

$$\frac{M}{c}\int_0^\infty e^{-\mu t}dt = \frac{M}{c\mu} < 1/4$$

and select $\lambda > 0$ such that if $\|x\| < \lambda$, then

$$\|F(x)\| = |F_0(w_\ell)| = \frac{\alpha_3}{m}|w_\ell|^3 < \frac{\|x\|}{c}.$$

If $x_0 \in X$ and $\|x_0\| < \frac{\lambda}{2M}$, then the mild solution given by

$$x(t) = S(t)x_0 + \int_0^t S(t-s)F(x(s))ds$$

exists on a finite interval $0 \leq t < t_1$ for some $t_1 > 0$ and satisfies

$$\|x(t)\| \leq \lambda, \quad 0 \leq t < t_1.$$

If t_1 is chosen as large as possible, then either $t_1 = +\infty$ or $\|x(t_1)\| = \lambda$. If $t_1 < +\infty$ and $t < t_1$, then $\|x(t)\| \leq \lambda$ and

$$\begin{aligned} \|x(t)\| &\leq \|S(t)x_0\| + \int_0^t \|S(t-s)F(x(s))\| \, ds \\ &\leq Me^{-\mu t}\|x_0\| + M\int_0^t e^{-\mu(t-s)}\|F(x(s))\| \, ds. \end{aligned}$$

Since

$$\begin{aligned} M\int_0^t e^{-\mu(t-s)}\|F(x(s))\| \, ds &\leq \frac{M}{c}\int_0^t e^{-\mu(t-s)}\|x(s)\| \, ds \\ &\leq \frac{\lambda}{2} + \frac{M\lambda}{c}\int_0^\infty e^{-\mu s}ds, \end{aligned}$$

it follows that for all $0 \leq t < t_1$

$$\begin{aligned} \|x(t)\| &\leq Me^{-\mu t}\|x_0\| + \frac{M}{c}\int_0^t e^{-\mu(t-s)}\|x(s)\| \, ds \\ &\leq \frac{\lambda}{2} + \frac{M\lambda}{c}\int_0^\infty e^{-\mu s}ds \\ &\leq \frac{\lambda}{2} + \frac{\lambda}{4}. \end{aligned}$$

Hence $\|x(t_1)\| \leq \frac{3\lambda}{4} < \lambda$. Thus, $t_1 = +\infty$ and the solution exists for all positive time. If $x_0 \in \mathcal{D}(A)$, then it can be shown that the mild solution is a classical solution.

In order to obtain exponential stability, let $h(t) = e^{\mu t}\|x(t)\|$ and note that $0 < h(t)$ and

$$\begin{aligned} h(t) = \|x(t)\|e^{\mu t} &\leq M\|x_0\| + \frac{M}{c}\int_0^t e^{\mu s}\|(x(s))\| \, ds \\ &= M\|x_0\| + \frac{M}{c}\int_0^t h(s)ds. \end{aligned}$$

Gronwall's inequality implies that

$$h(t) = \|x(t)\|e^{\mu t} \leq M\|x_0\|e^{\frac{M}{c}t},$$

or equivalently, that

$$\|x(t)\| \leq M\|x_0\|e^{-(\mu - \frac{M}{c})t} = Me^{-\beta t}\|x_0\|,$$

where $\beta = (\mu - \frac{M}{c}) > 0$. This completes the proof.

The Nonlinear Control Problem

Since $x_e = 0$ is the only equilibrium for the unforced ($\eta(t) = 0$) nonlinear system we concentrate on the stability of this equilibrium. Given the nonlinear control system

$$\dot{x}(t) = Ax(t) + F(x(t)) + Bu(t) + G\eta(t), \qquad x(0) = x_0 \tag{4.25}$$

with sensed output

$$y(t) = Cx(t) + E\xi(t), \tag{4.26}$$

the goal is to find a controller $u(t)$ that enhances the stability of $x_e = 0$ for the unforced nonlinear plant and attenuates the disturbance. We shall design a controller of the form

$$u(t) = -K^M x_c^M(t), \tag{4.27}$$

where $x_c^M(t)$ satisfies a finite dimensional nonlinear compensator equation driven by the sensed output $y(t)$ defined by (4.26) and K^M is a bounded linear operator. This is done in steps. First we use linearization and infinite dimensional LQG and MinMax control theory to find an infinite dimensional linear feedback law based on dynamic compensation. The representation of the feedback gain operator is then used to construct approximate feedback operators and this information can then be applied to construct low order nonlinear state estimators. It is important to note that even though the feedback law is linear, the controller is nonlinear due to the nonlinearity of the dynamic compensator. Although the basic idea is rather standard for finite dimensional systems (extended Kalman filtering), we shall see how working directly with the distributed parameter model can lead to new insights and practical controllers based on rigorous approximation theory. Also, we show how the use of *reduced basis techniques* from finite element theory can greatly enhance the design of low order controllers without loss in performance and robustness.

MinMax and LQG Design for the Linearized Control Problem

Linearizing (4.25) about $x_e = 0$, one obtains the linear distributed parameter control system defined on X by

$$\dot{z}(t) = Az(t) + Bu(t) + G\eta(t), \quad z(0) = z_0 \tag{4.28}$$

with sensed output

$$y(t) = Cz(t) + E\xi(t). \tag{4.29}$$

As mentioned above, rather than using full state feedback we design a state estimator, $z_c(t)$, satisfying the linear system on X

$$\dot{z}_c(t) = A_c z_c(t) + F_c y(t), \quad z_c(0) = z_{c_0} \tag{4.30}$$

and use the linear feedback law

$$u(t) = -Kz_c(t), \tag{4.31}$$

where A_c, F_c and K are operators to be determined. If (4.31) is inserted into the linear system (4.28), then one has the closed-loop system defined by

$$\frac{d}{dt}\begin{bmatrix} z(t) \\ z_c(t) \end{bmatrix} = \begin{bmatrix} A & -BK \\ F_cC & A_c \end{bmatrix}\begin{bmatrix} z(t) \\ z_c(t) \end{bmatrix} + \begin{bmatrix} G & 0 \\ 0 & F_cE \end{bmatrix}\begin{bmatrix} \eta(t) \\ \xi(t) \end{bmatrix}. \tag{4.32}$$

The infinite dimensional controller defined above is completely determined by the three operators A_c, F_c and K. To determine these operators, we will use both LQG and MinMax design. It is well known that this approach is equivalent to finding the optimal solution to a quadratic differential game. However, we do not wish to devote time here to the discussion of this aspect of the problem. Our interest lies only in the fact that the MinMax controller stabilizes the system and attenuates the disturbance to controlled output map. We shall see that it performs this task much better

than the LQG controller and, when combined with the reduced basis compensator, has reasonable robustness properties. Thus, we shall present only the items essential to the construction of the MinMax controller. In particular, assume that $Q = Q^* > 0$ and $M = M^* = GG^* \geq 0$ are bounded linear operators on X and $r > 0$. Also, since $E = I_{2\times 2}$ it follows that $N = EE^* = I > 0$. For each $\theta \geq 0$ consider the Riccati equations

$$A^*\Pi + \Pi A - \Pi \left[Br^{-1}B^* - \theta^2 M\right]\Pi + Q = 0 \tag{4.33}$$

and

$$AP + PA^* - P\left[C^*C - \theta^2 Q\right]P + M = 0. \tag{4.34}$$

Since B and C are bounded linear operators and the linearized system is exponentially stable, the theory implies that, for sufficiently small θ, the Riccati equations (4.33) and (4.34) have minimal solutions $\Pi_\theta \geq 0$ and $P_\theta \geq 0$, respectively. In addition, the operator $\left[I - \theta^2 P_\theta \Pi_\theta\right]$ is positive definite, i.e.,

$$\left[I - \theta^2 P_\theta \Pi_\theta\right] > 0. \tag{4.35}$$

If one defines

$$\begin{aligned} K_\theta &= r^{-1}B^*\Pi_\theta, \\ F_\theta &= [I - \theta^2 P_\theta \Pi_\theta]^{-1} P_\theta C^*, \\ A_\theta &= A - BK_\theta - F_\theta C + \theta^2 M\Pi_\theta, \end{aligned} \tag{4.36}$$

then the corresponding control can be written in the form

$$u_\theta(t) = -r^{-1}B_\theta^*\Pi z_c(t) = -K_\theta z_c(t). \tag{4.37}$$

Observe that for $\theta = 0$ the resulting controller is the LQG (i.e., Kalman Filter) controller.

The Infinite Dimensional Nonlinear Controller

If we now substitute the K_θ, F_θ, and A_θ as computed above for the linear system into the nonlinear system, then the resulting nonlinear observer becomes

$$\dot{x}_c(t) = A_\theta x_c(t) + F(x_c(t)) + F_\theta y(t), \quad x_c(0) = x_{c_0}. \tag{4.38}$$

We shall assume that the feedback law is of the form

$$u_{nl}(t) = -K_\theta x_c(t), \tag{4.39}$$

where K_θ is a bounded linear operator.

The resulting closed loop nonlinear system formed by the above design can be written

$$\begin{aligned} \frac{d}{dt}\begin{bmatrix} x(t) \\ x_c(t) \end{bmatrix} = &\begin{bmatrix} A & -BK_\theta \\ F_\theta C & A_\theta \end{bmatrix}\begin{bmatrix} x(t) \\ x_c(t) \end{bmatrix} + \begin{bmatrix} F(x(t)) \\ F(x_c(t)) \end{bmatrix} \\ &+ \begin{bmatrix} G & 0 \\ 0 & F_\theta E \end{bmatrix}\begin{bmatrix} \eta(t) \\ \xi(t) \end{bmatrix}. \end{aligned} \tag{4.40}$$

In the case of a linear system, results exist which guarantee exponential stability of this closed loop system under the above control/compensator design. Though we have no such theory for the linear control law applied to the nonlinear problem, we know from the above theorem that the uncontrolled problem is exponentially stable. Therefore, we expect the above controller to further stabilize the system and to attenuate the disturbance. The nonlinear controller defined by (4.38)-(4.39) is infinite

dimensional and, before one can make use of this structure, approximations must be introduced. However, considerable practical information can be extracted from this infinite dimensional controller that greatly enhances the development of "good" low order approximations. To illustrate this idea we first make use of representation theory to obtain explicit forms for the linear feedback law.

The nonlinear control law defined by (4.39) has the representation

$$u_{nl}(t) = -K_\theta x_c(t) = -\int_0^\ell \rho k_{v_s}(s)\frac{\partial}{\partial t}w_c(t,s)ds - \int_0^\ell \tau k_s(s)\frac{\partial}{\partial s}w_c(t,s)ds \\ -mk_{v_m}\frac{\partial}{\partial t}w_c(t,\ell) - \alpha_1 k_d w_c(t,\ell) \tag{4.41}$$

where $k_{v_s}(s)$ and k_{v_m} are the velocity gains for the string and mass, respectively, $k_s(s)$ is the strain gain for the string, k_d is the displacement gain for the mass, and $w_c(t,s)$ is the solution to the nonlinear partial differential equation corresponding to the estimated "state", i.e.,

$$x_c(t) = [w_c(t,\cdot), w_c(t,\ell), \frac{\partial}{\partial t}w_c(t,\cdot), \frac{\partial}{\partial t}w_c(t,\ell)]^T.$$

The form of the control law in (4.41) is obtained using the integral representation of the feedback gain operator $K_\theta : X \to \mathbb{R}$. The "functional gains" $k_{v_s}(s)$ and $k_s(s)$ are known to belong to $L_2(0,\ell)$. Moreover, during the past ten years convergent numerical algorithms have been developed to compute these functional gains. Most of these schemes are based on finite element methods which use splines to compute numerical approximations of the functional gains $k_{v_s}(s)$ and $k_s(s)$ and the gains k_{v_m} and k_d. We shall use such schemes not only to compute the infinite dimensional controller, but to construct practical finite dimensional controllers of low order.

The observer gain operator $F_\theta : \mathbb{R}^2 \to X$ is continuous and has range in $\mathcal{D}(A) \subseteq X$. Therefore, there exist functions $g_1(s), g_2(s), h_1(s)$ and $h_2(s)$ in $H_L^1(0,\ell)$ such that

$$F_\theta \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} g_1(s) & h_1(s) \\ g_1(\ell) & h_1(\ell) \\ g_2(s) & h_2(s) \\ g_2(\ell) & h_2(\ell) \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \in X. \tag{4.42}$$

The functions $g_1(s), g_2(s), h_1(s)$ and $h_2(s)$ are called the observer functional gains and will be approximated by a reduced basis finite element scheme.

The design process proceeds in two steps:

1: The infinite dimensional controller is "computed" by using finite element approximations (of high order) of the Riccati equations (4.33)-(4.34). This provides accurate approximations of the functional gains $k_{v_s}(s)$ and $k_s(s)$. By introducing "optimal" nodal placement, one approximates $k_{v_s}(s)$ and $k_s(s)$ by low order splines (or even step functions). The spatial structure of the functional gains is used to guide the selection of the optimal projection.

2: A reduced basis-finite element approximation of the nonlinear observer equation is constructed and combined with the low order approximation of the observer functional gains to produce a low order nonlinear controller.

This approach makes use of the structure of the full infinite dimensional nonlinear control law and hence often leads to a control design that is "better" than one might obtain by blindly introducing approximations at the outset. This process is illustrated in the next section where numerical experiments are provided to illustrate the ideas and benefits of this approach.

4.2.2 Approximations

We shall apply standard finite element approximation techniques to the system described above. Although it is not essential to use finite elements, we restrict our attention to this approach because it is easy to implement and there is a well established convergence theory.

Observe that $X = E \times H$ where $E = H_L^1 \times \mathbb{R}$ with energy inner product

$$\langle [w(\cdot), w_\ell]^T, [\hat{w}(\cdot), \hat{w}_\ell]^T \rangle_E = \tau \int_0^\ell \frac{\partial}{\partial s} w(s) \frac{\partial}{\partial s} \hat{w}(s) ds + \alpha_1 w_\ell \hat{w}_\ell \tag{4.43}$$

and $H = L_2 \times \mathbb{R}$ with inner product

$$\langle [v(\cdot), v_\ell]^T, [\hat{v}(\cdot), \hat{v}_\ell]^T \rangle_H = \rho \int_0^\ell v(s) \hat{v}(s) ds + m v_\ell \hat{v}_\ell. \tag{4.44}$$

In order to write this problem in a form that is conducive to numerical computation, we use the variational form of the problem. We take the state as defined above. To provide a class of test functions for the variational form, we define another Hilbert space $V \subset E$ by

$$V = \left\{ \varphi = [\varphi_1(\cdot), \varphi_2]^T \in E : \varphi_1(\ell) = \varphi_2 \right\}.$$

The weak form of (4.12)-(4.15) is given by the variational equation for each $\varphi = [\varphi_1(\cdot), \varphi_2]^T \in V$;

$$\begin{aligned} &\int_0^\ell \rho \frac{\partial}{\partial t} v(t,s) \varphi_1(s) ds + \int_0^\ell \left[\tau \frac{\partial}{\partial s} w(t,s) + \gamma \frac{\partial}{\partial s} v(t,s) \right] \frac{d}{ds} \varphi_1(s) ds \\ &+ m \frac{\partial}{\partial t} v_\ell \varphi_2 + \left[\alpha_1 w_\ell + \alpha_3 [w_\ell]^3 \right] \varphi_2 = [u(t) + \eta(t)] \varphi_2, \end{aligned} \tag{4.45}$$

with initial condition

$$w(0,s) = w_0, \quad v(0,s) = w_1. \tag{4.46}$$

Finite Dimensional Approximations

A Galerkin-based finite element approximation scheme is applied to the variational form of the problem to obtain a sequence of finite dimensional systems suitable for computation. We choose a basis $\{e_i\}_{i=1}^N$ for the approximating space $V^N \subseteq V$ so that the state is approximated by a linear combination of the basis vectors. We use linear B-splines to approximate the position of the spring; after matching the fixed end boundary condition, there are N splines along the interval $[0, \ell]$ which we denote as $\{b_i(s)\}_{i=1}^N$. Thus, the basis vectors will be of the form

$$e_i^N = \begin{bmatrix} b_i^N(s) \\ b_i^N(\ell) \end{bmatrix}, \quad i = 1, \cdots, N. \tag{4.47}$$

Since the N^{th} linear-B spline is the only spline which is nonzero at $s = \ell$, the second entry of the basis vectors is zero for $i = 1, \cdots, N-1$, and is one for $i = N$. Thus we approximate the state as

$$\begin{bmatrix} w(t,s) \\ w(t,\ell) \end{bmatrix} \approx \begin{bmatrix} w^N(t,s) \\ w^N(t,\ell) \end{bmatrix} = \sum_{i=1}^N \zeta_i^N(t) e_i^N(s) = \begin{bmatrix} \sum_{i=1}^N \zeta_i^N(t) b_i^N(s) \\ \zeta_N^N(t) \end{bmatrix}. \tag{4.48}$$

Substituting this approximation to the state into the variational form in equations (4.45) and (4.46) letting the test functions φ range over the basis vectors, we obtain the finite dimensional system

$$\begin{aligned} M^N \frac{d^2}{dt^2} \zeta^N(t) + D_0^N \frac{d}{dt} \zeta^N(t) + A_0^N \zeta^N(t) + F_0^N(\zeta^N(t)) = B_0^N u(t) + G_0^N \eta(t), \\ \zeta^N(0) = \zeta_0^N, \quad \frac{d}{dt} \zeta^N(0) = \zeta_1^N. \end{aligned} \tag{4.49}$$

In (4.49), $\zeta^N(t) = [\zeta_1^N(t), \zeta_2^N(t), \ldots, \zeta_N^N(t)]^T$, M^N is the mass matrix, D_0^N is the damping matrix, A_0^N is the linear stiffness matrix, $F_0^N(\zeta^N(t))$ contains the nonlinear terms, B_0^N is the input matrix, and G^N is the disturbance matrix given by

$$
\begin{aligned}
[M^N]_{i,j} &= \int_0^\ell \rho b_i^N(s) b_j^N(s) ds + m b_i^N(\ell) b_j^N(\ell), \quad i,j = 1, \cdots, N \\
[D_0^N]_{i,j} &= \int_0^\ell \gamma \frac{d}{ds} b_i^N(s) \frac{d}{ds} b_j^N(s), \qquad i,j = 1, \cdots, N \\
[A_0^N]_{i,j} &= \int_0^\ell \tau \frac{d}{ds} b_i^N(s) \frac{d}{ds} b_j^N(s) ds + \alpha_1 b_i^N(\ell) b_j^N(\ell), \quad i,j = 1, \cdots, N \\
F_0^N(\zeta^N) &= \alpha_3 \left[w_N^N \right]^3 \\
B_0^N &= \frac{1}{m} [b_1^N(\ell), b_2^N(\ell), ..., b_N^N(\ell)]^T \\
G_0^N &= \frac{1}{m} [b_1^N(\ell), b_2^N(\ell), ..., b_N^N(\ell)]^T
\end{aligned}
\tag{4.50}
$$

We write this finite element equation as the first order system

$$
\begin{aligned}
\frac{d}{dt} x^N(t) &= A^N x^N(t) + F^N(x^N(t)) + B^N u(t) + G^N \eta(t) \\
x^N(0) &= x_0^N,
\end{aligned}
\tag{4.51}
$$

with output

$$
y^N(t) = C^N x^N(t) + E\xi(t) \in \mathbb{R}^2, \tag{4.52}
$$

where $x^N(t) = [\zeta^N(t), \frac{d}{dt}\zeta^N(t)]^T$, $x_0^N = [\zeta_0^N, \zeta_1^N]^T$ and

$$
\begin{aligned}
A^N &= \begin{bmatrix} 0 & I \\ -M^{-N} A_0^N & -M^{-N} D_0^N \end{bmatrix}, \quad B^N = \begin{bmatrix} 0 \\ -M^{-N} B_0^N \end{bmatrix}, \\
F^N(x^N(t)) &= \begin{bmatrix} 0 \\ -M^{-N} F_0^N(w^N(t)) \end{bmatrix}, \quad G^N = \begin{bmatrix} 0 \\ -M^{-N} G_0^N \end{bmatrix}, \\
\text{and } C^N &= \begin{bmatrix} 0_{1\times N-1} & 1 & 0_{1\times N-1} & 0 \\ 0_{1\times N-1} & 0 & 0_{1\times N-1} & 1 \end{bmatrix}.
\end{aligned}
\tag{4.53}
$$

Remark 1 It is important to note that (4.51) provides the approximation $w^N(t,s)$ of $w(t,s)$ in the finite element space $V^N \subseteq V$. In particular,

$$
e_i^N = \begin{bmatrix} b_i^N(s) \\ b_i^N(\ell) \end{bmatrix}, i = 1, 2, \ldots, N
$$

is a basis for V^N. More importantly, this approximation leads to a numerical method for computing the functional gains and for constructing finite dimensional observers. We shall use high order approximations to compute the functional gains and then "project" these gains onto a space of appropriate simple functions X^M. To complete the design, a low order dynamic compensator is constructed by using a reduced basis/finite element approximation of the infinite dimensional compensator equation. This process produces a practical finite dimensional (nonlinear) controller and at the same time makes use of the structure of the infinite dimensional feedback law. Although there are several possible "reduced basis" spaces, we shall restrict our attention to the simplest case where the reduced basis space is a finite element space satisfying $V^M \subseteq V^N \subseteq V$ with $M << N$. By using the space X^M to approximate the functional gains, we do make use of the spatial structure of the feedback law. However, there is even the possibility of constructing local (in space) nonlinear observers.

The Reduced Basis Approach to Low Order Controller Design

By considering the representation of the nonlinear control law given in (4.41), there are two obvious places where one can introduce approximations to obtain an approximate controller. First, the operator K_θ can be replaced by an approximate operator K_θ^N. This results in an approximation of the feedback gains and produces the approximate feedback control

$$u_{nl}^N(t) = -K_\theta^N x_c(t) = -\int_0^\ell \rho k_{v_s}^N(s)\frac{\partial}{\partial t}w_c(t,s)ds - \int_0^\ell \tau k_s^N(s)\frac{\partial}{\partial s}w_c(t,s)ds \\ -mk_{v_m}^N\frac{\partial}{\partial t}w_c(t,\ell) - \alpha_1 k_d^N w_c(t,\ell) \tag{4.54}$$

where $k_{v_s}^N(s)$ and $k_s^N(s)$ are approximations of the functional gains. Observe that this controller is still infinite dimensional. However, if one computes K_θ^N by solving the approximate Riccati equation

$$[A^N]^*\Pi^N + \Pi^N A^N - \Pi^N\left[B^N r^{-1}[B^N]^* - \theta^2 M^N\right]\Pi^N + Q^N = 0 \tag{4.55}$$

and defining

$$K_\theta^N = r^{-1}[B^N]^*\Pi_\theta^N$$

then it is known that $K_\theta^N \to K_\theta$, or equivalently, that $k_{v_s}^N(s) \to k_{v_s}(s)$ and $k_s^N(s) \to k_s(s)$. Unless one approximates the observer $w_c(t,s)$ then the controller is still infinite dimensional. Thus, we must also replace the infinite dimensional compensator equation by a finite dimensional compensator of the form

$$\dot{x}_c^M(t) = A_\theta^M x_c^M(t) + F^M(x_c(t)) + F_\theta^M y(t), \quad x_c^M(0) = x_{c_0}^M \tag{4.56}$$

where $M \leq N$.

This second approximation is often accomplished by simply using the same $M = N$ finite element model (4.51)-(4.52) and solving the corresponding Riccati equations

$$A^N P^N + P^N[A^N]^* - P^N\left[[C^N]^*C^N - \theta^2 Q^N\right]P^N + M^N = 0. \tag{4.57}$$

If this approach is used, then one can show that, for $N = M$ sufficiently large, the resulting finite dimensional controller will have performance and robustness properties similar to the infinite dimensional controller. However, this approach can lead to large (non practical) observers and does not make full use of the information contained in structure of the functional gains. As one example of this idea, simply note that if there is a region in space, say $0 < s < b < \ell$, where the functional gain $k_s(s)$ is zero (or small), then there is no reason to approximate the strain $\frac{\partial}{\partial s}w(t,s)$ in the interval $(0,b)$. This is illustrated in the numerical section below and often occurs in boundary control. Moreover, the control law (4.54) requires only that the weighted integral averages be computed and hence it is not always necessary to have high fidelity (global in space) models of the nonlinear observer. These observations indicate that reduced basis schemes may offer some improvement over straight forward approximate-then-design approaches. As we show below, this is the case for our model problem.

Although the following approach can be made very general, we shall restrict our attention to one particular scheme in order to keep this work focused. The design is accomplished in stages:

Stage 1: Select a fine mesh (large N) finite element approximation and compute the functional gains $k_{v_s}^N(s)$ and $k_s^N(s)$ (which, for all practical purposes will be $k_{v_s}(s)$ and $k_s(s)$ using established convergence results).

Stage 2: Select a low order numerical scheme that makes use of the structure of $k_{v_s}^N(s)$ and $k_s^N(s)$ and "project" these functional gains onto this space. This leads to a simplified control law by approximating the "converged" gains. We do this by taking an P^{th} order projection of the gains onto a space of step functions. The spatial domain $[0,\ell]$ is subdivided by a mesh of the form $0 = s_0^N < s_1^N < \cdots < s_{P+1}^N = \ell$ where the nodes s_i^N are clustered in neighborhoods in

which $k_{v_s}^N(s)$ and $k_s^N(s)$ are large and no nodes are placed in neighborhoods where $k_{v_s}^N(s)$ and $k_s^N(s)$ are essentially zero. Specifically,

$$\begin{aligned} k_s^N(s) &\approx k_s^{N,P}(s) = \sum_{i=1}^{P} k_{s,i}^{N,P} \chi_i^{N,P}(s) \\ k_{v_s}^N(s) &\approx k_{v_s}^{N,P}(s) = \sum_{i=1}^{P} k_{v_s,i}^{N,P} \chi_i^{N,P}(s) \end{aligned} \tag{4.58}$$

where

$$\begin{aligned} \chi_i^{N,P}(s) &= \begin{cases} 1 & s_{i-1}^N \le s \le s_i^N \\ 0 & \text{otherwise} \end{cases}, \\ 0 = s_0^N &< s_1^N < \cdots < s_{P+1}^N = \ell. \end{aligned} \tag{4.59}$$

In this case,

$$k_{s,i}^{N,P}(s) = \frac{\int_0^\ell k_s^N(s)\chi_i^{N,P}(s)ds}{\int_0^\ell \chi_i^{N,P}(s)ds}, \quad k_{v_s,i}^{N,M}(s) = \frac{\int_0^\ell k_{v_s}^N(s)\chi_i^{N,P}(s)ds}{\int_0^\ell \chi_i^{N,P}(s)ds}. \tag{4.60}$$

With these piecewise constant approximations of the functional gains, the control law (4.41) is replaced by

$$\begin{aligned} u_{nl,P}^N(t) = &-\sum_{i=1}^{P} k_{v_s,i}^{N,P} \int_{s_i^N}^{s_{i+1}^N} \rho \frac{\partial}{\partial t} w_c(t,s)ds \\ &-\sum_{i=1}^{P} k_{s,i}^{N,P} \int_{s_i^N}^{s_{i+1}^N} \tau \frac{\partial}{\partial s} w_c(t,s)ds \\ &-mk_{v_m}^N \frac{\partial}{\partial t} w_c(t,\ell) - \alpha_1 k_d^N w_c(t,\ell). \end{aligned} \tag{4.61}$$

Observe that in (4.61) one now needs only the spatial averages of velocity and strain over subdomains where the averaged functional gains are "large".

Stage 3: Select a reduced basis subspace of order $M << N$, $V^M \subset V^N$, and build the MinMax observer defined by (4.56) on the space V^M. This produces approximations $g_1^M(s), g_2^M(s), h_1^M(s)$ and $h_2^M(s)$ of the observer functional gains $g_1(s), g_2(s), h_1(s)$ and $h_2(s)$, respectively and hence the approximate observer operator $F_\theta^M : \mathbb{R}^2 \to V^M \subseteq X$ has the representation,

$$F_\theta^M \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} g_1^M(s) & h_1^M(s) \\ g_1^M(\ell) & h_1^M(\ell) \\ g_2^M(s) & h_2^M(s) \\ g_2^M(\ell) & h_2^M(\ell) \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \in V^M. \tag{4.62}$$

Moreover, one has the natural "low order" estimate $w_c^M(t,s)$ of $w_c(t,s)$, which can now be

substituted into (4.61) to yield the approximate controller

$$\begin{aligned} u_{nl,P}^{N,M}(t) = &-\sum_{i=1}^{P} k_{v_s,i}^{N,P} \int_{s_i^N}^{s_{i+1}^N} \rho \frac{\partial}{\partial t} w_c^M(t,s)ds \\ &-\sum_{i=1}^{P} k_{s,i}^{N,P} \int_{s_i^N}^{s_{i+1}^N} \tau \frac{\partial}{\partial s} w_c^M(t,s)ds \\ &-mk_{v_m}^N \frac{\partial}{\partial t} w_c^M(t,\ell) - \alpha_1 k_d^N w_c^M(t,\ell) \\ = &-K_\theta^{N,M} x_c^M(t). \end{aligned} \tag{4.63}$$

The resulting closed loop nonlinear system formed by the above design can be written

$$\begin{aligned} \frac{d}{dt}\begin{bmatrix} x(t) \\ x_c^M(t) \end{bmatrix} = &\begin{bmatrix} A & -BK_\theta^{N,M} \\ F_\theta^M C^M & A_\theta^M \end{bmatrix}\begin{bmatrix} x(t) \\ x_c^M(t) \end{bmatrix} + \begin{bmatrix} F(x(t)) \\ F^M(x_c(t)) \end{bmatrix} \\ &+ \begin{bmatrix} G & 0 \\ 0 & F_\theta^M E \end{bmatrix}\begin{bmatrix} \eta(t) \\ \xi(t) \end{bmatrix}. \end{aligned} \tag{4.64}$$

Remark 2 Observe that it is not necessary for the finite element mesh to be "super fine" (i.e., M large) in order that the integral averages

$$\int_{s_i^N}^{s_{i+1}^N} \frac{\partial}{\partial t} w_c^M(t,s)ds \text{ and } \int_{s_i^N}^{s_{i+1}^N} \frac{\partial}{\partial s} w_c^M(t,s)ds \tag{4.65}$$

be "close" to the integrals

$$\int_{s_i^N}^{s_{i+1}^N} \frac{\partial}{\partial t} w_c(t,s)ds \text{ and } \int_{s_i^N}^{s_{i+1}^N} \frac{\partial}{\partial s} w_c(t,s)ds, \tag{4.66}$$

respectively. We note that there are several theoretical issues that remain to be addressed in the nonlinear case. However, the numerical results below indicate that this approach holds considerable promise as a design approach.

Remark 3 In order to have confidence in the overall design, it is essential that the control design be robust. This is one reason we chose the MinMax design described above. However, the MinMax design only provides robustness in a limited sense and at this point there is no theoretical assurance that the reduced basis design will be robustly stable. However, we shall provide numerical evidence that show that the reduced basis controller does retain some robust stabilization properties. As numerical approximations must always be introduced in the design of distributed parameter systems, one may view the resulting numerical errors as (unstructured) parameter uncertainties and design for robust stability.

4.2.3 Numerical Experiments and Conclusions

Here we present results from two sets of numerical experiments for the cable mass system. They differ in the forcing disturbance which was applied to the system; all other system parameters and initial conditions were the same and the parameters are presented in the table below. It can be seen from the small choice for γ that the spring is extremely lightly damped. Initial conditions were chosen to be $[w(t,s), v(t,s)] = [s, -2]$, $[w_c(0,s), v_c(0,s)] = [0.5s, 0]$.

Table 4.1: System Parameters

ρ	τ	γ	m	ℓ	α_1	α_3	θ
1	1	.005	1.5	2	.01	3	1.21

For Stage 1 described above, we chose the $N = 32$ finite dimensional approximation to the infinite dimensional system as convergence of the functional gains was clear at this level of mesh refinement. These "converged" gains are shown by the dashed lines in Figures 4.7 and 4.8. For Stage 2, we chose $P = 2$. In Figures 4.7 and 4.8, the functional gains for strain, k_s, and velocity, k_v, for the LQG and MinMax compensators are plotted against the optimal second order projections ($P = 2$) described in the previous section. The number of and particular subintervals chosen on which to approximate the gains was based upon gain shape. Both strain and velocity gains for both designs showed little variation over the interval $[0, 1]$ and significant variation over $[1, 2]$ so these two subintervals were used to compute the approximate (reduced order) gains. The structure of the LQG and MinMax gains for $N = 32$ are similar in overall shape, but differences in numerical size lead to significant differences in the projected gains. Note that for the strain gain, the first value of the reduced order gain for LQG is nearly zero and the second value is about two-thirds of the MinMax gain. The velocity gains show significant differences in magnitude; the $N = 32$ MinMax gain attains a minimum value of about -3 and a maximum of nearly 8, while the LQG gain reaches only -1 and 1.5. These differences in magnitude produce marked differences in the reduced order velocity gains.

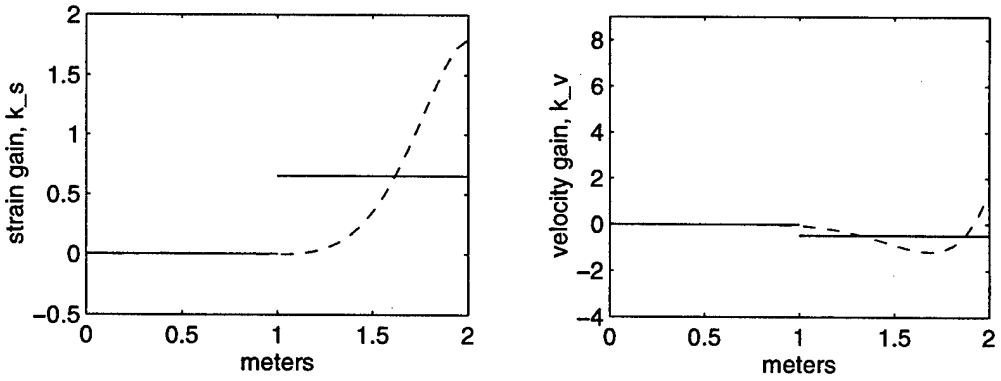


Figure 4.7: Controller functional gains for LQG compensator.

As described in (4.62), there are four functional observer gains, $g_1(s)$, $g_2(s)$, $h_1(s)$, and $h_2(s)$. As with the controller gains, the observer gains for $N = 32$ are shown in Figure 4.9 by dashed lines and the gains for the reduced order observer, $g_1^M(s)$, $g_2^M(s)$, $h_1^M(s)$, $h_2^M(s)$, (for $M = 2$) are shown by solid lines. The lumped gains $(g_1(\ell), \ldots, h_2(\ell))$ can be obtained from the value of the functional gains at $s = \ell$.

For both of the simulation experiments discussed below, a value of $N = 8$ was used as $N = 32$ yielded too fine a mesh in terms of computational time. A reduced order basis was chosen with $M = 2$. As we shall show, the MinMax compensator performs better than the LQG compensator. More surprisingly perhaps, the MinMax second order compensator performs as well as, and in some respects better than the full order MinMax compensator.

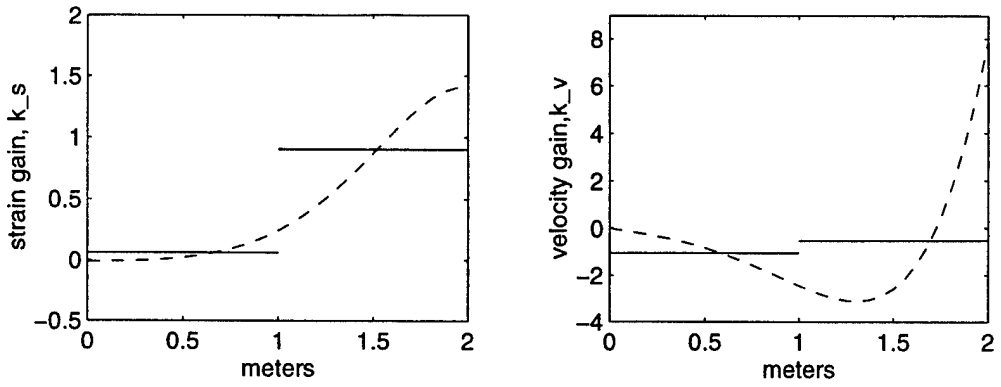


Figure 4.8: Controller functional gains for MinMax compensator.

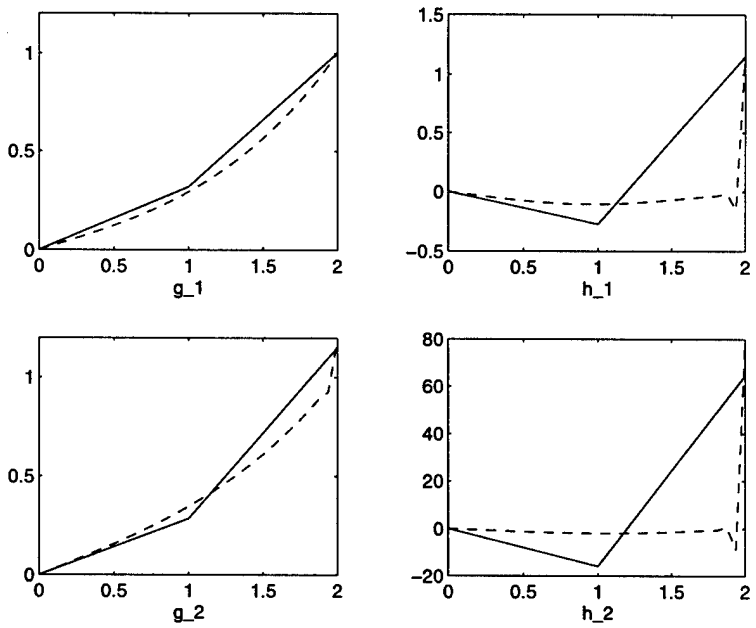


Figure 4.9: Observer functional gains for MinMax reduced order compensator.

Disturbance $\eta(t) = 5\cos(.4803t) + 10\cos(1.7644t)$

The sensor disturbance $\xi(t)$ is assumed to be 0 and the system disturbance was chosen as $\eta(t) = 5\cos(.4803t) + 10\cos(1.7644t)$. This particular forcing function was chosen because .4803 and 1.7644 are the lowest two frequencies of the uncontrolled (open loop) system (i.e., the first two eigenvalues of A^N).

Technically, to qualify as chaotic behavior, the system should be forced at a single frequency, so one must be careful in applying concepts of chaos to this example. However, as shown in Figure 4.10 by the dotted line, after 12000 seconds (i.e., 12 hours) the mass has never adopted a periodic orbit nor shown evidence of an attractor. Specifically, Figure 4.10 shows the phase plots of the open loop behavior of the mass (between 11800 and 12000 seconds) by the dotted line as compared to the behavior of the mass (between 0 and 200 seconds) when LQG and MinMax compensators are used to control the system. Both drive the mass to a periodic (yet still complicated) orbit,

However, MinMax better attenuates the mass. The time histories for the mass position under LQG and MinMax design are shown in Figure 4.11.

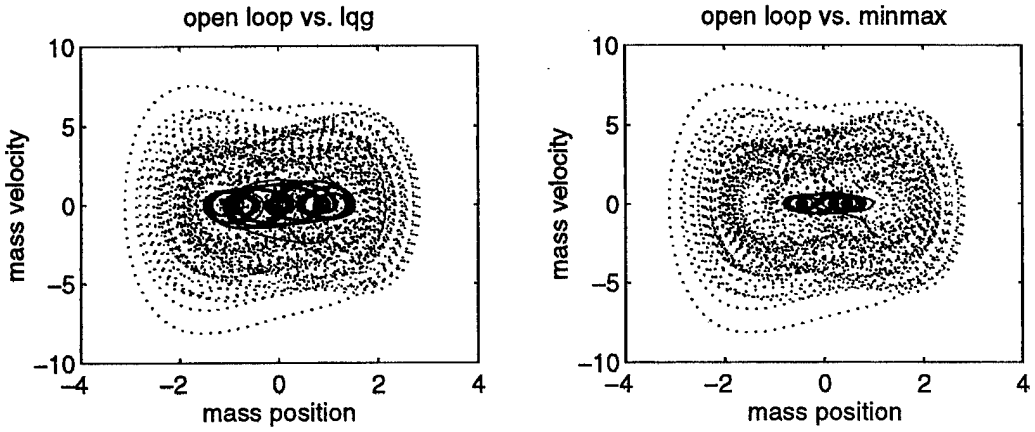


Figure 4.10: Open loop vs. LQG and MinMax compensator phase portraits for the mass.

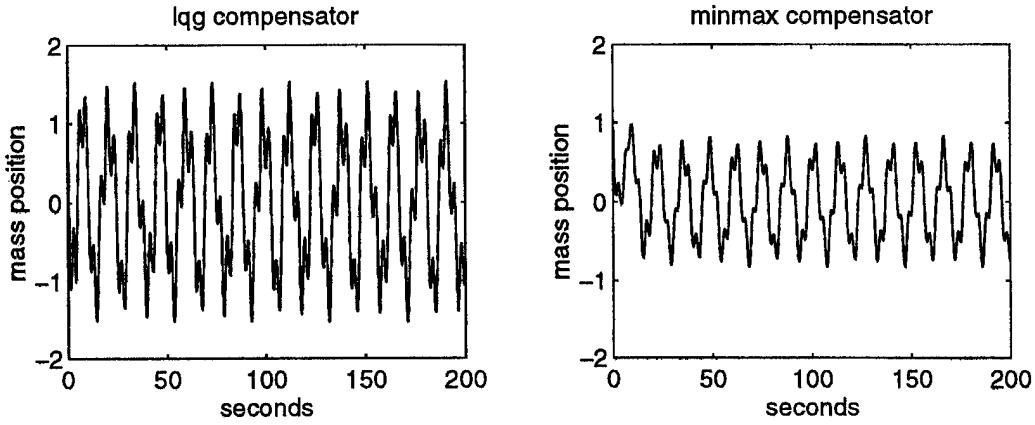


Figure 4.11: Time history of mass position with LQG and MinMax compensators.

In Figure 4.12, the phase portraits for the mass are shown for the LQG compensator and the reduced (second) order LQG compensator. With the reduced order compensator, both attenuation and robustness are lost. However, as shown in Figure 4.13, there is no significant loss in either robustness or attenuation when one compares the mass phase portrait for the MinMax compensator with the phase portrait for the reduced (second) order compensator.

The behavior of the mass represents only part of the story; the cable (the distributed part of the system) shows similar effects. In Figure 4.14, the phase portrait for mid-cable ($s = 1$) is plotted for the LQG and MinMax compensators. Once again, MinMax shows better attenuation of the cable. A comparison of the full and reduced order MinMax compensator phase plots and time histories for mid-cable are shown in Figure 4.15; as one can see, the mid-point of the cable adopts a periodic motion more quickly in the case of the reduced order compensator. Plots for other points along the cable showed similar results.

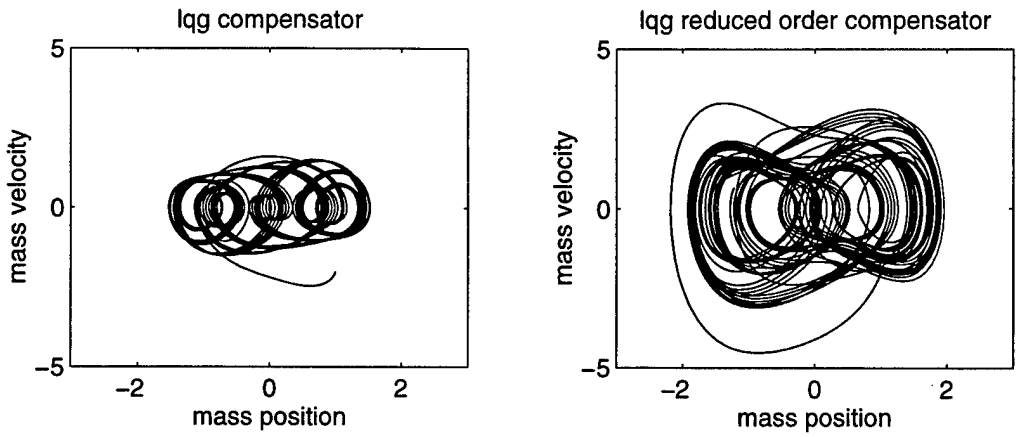


Figure 4.12: LQG full order vs. LQG reduced (second order) compensator phase portraits for the mass.

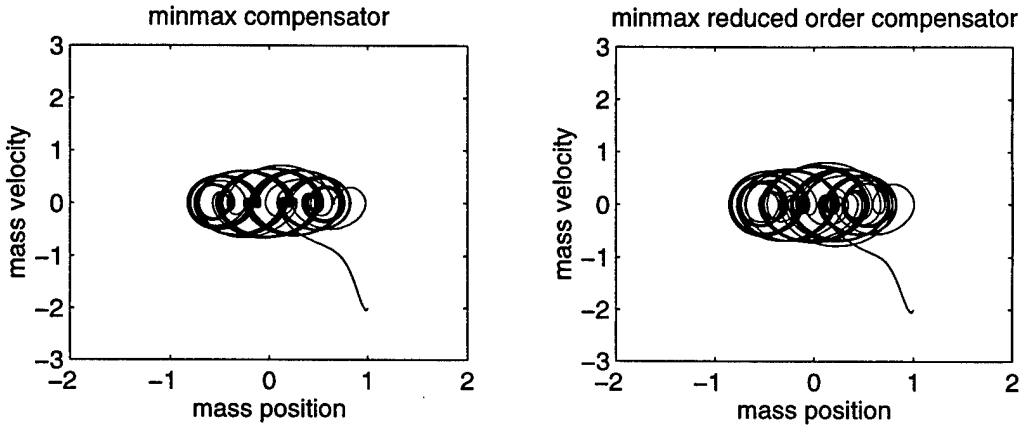


Figure 4.13: Minmax full order vs. Minmax reduced (second order) compensator phase portraits for the mass.

Disturbance $\eta(t) = 5\cos(2.2447t)$

As before, the sensor disturbance $\xi(t)$ is assumed to be 0 but the system disturbance was chosen as $\eta(t) = 5\cos(2.2447t)$. This frequency for the forcing function is the sum of .4803 and 1.7644 (the frequencies in the previous example). This choice was made achieve "combination resonance".

After 900 seconds (i.e., 15 minutes) the mass adopts a periodic orbit as shown in Figure 4.16 by the solid line; the dotted line is the phase portrait or the mass between 0 and 500 seconds. In Figure 4.17, the phase portrait for the mass under LQG and MinMax compensation as compared to the open loop behavior is shown. Again, MinMax shows a marked improvement in attenuation; see also Figure 4.18 for the time histories.

In Figure 4.19, the phase portraits for the mass are shown for the full and reduced order LQG compensators. There is a significant loss of attenuation with the reduced order compensator. Again, as seen in Figure 4.20, there is no significant loss in either robustness or attenuation when one compares the mass phase portrait for the MinMax compensator with the phase portrait for the reduced order compensator.

For the mid-cable behavior, LQG compared unfavorably with MinMax as in the previous experiment. However, the MinMax reduced order compensator shows marked attenuation capabilities as

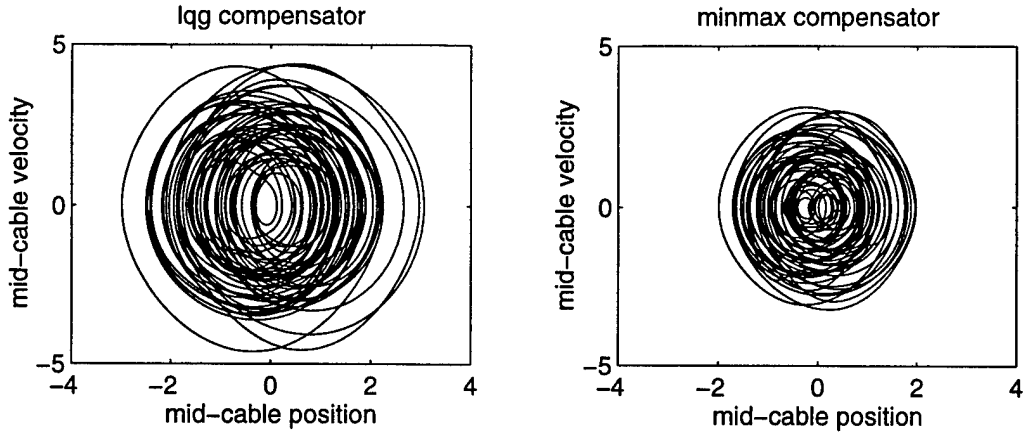


Figure 4.14: LQG vs. MinMax compensator phase portraits for mid-cable.

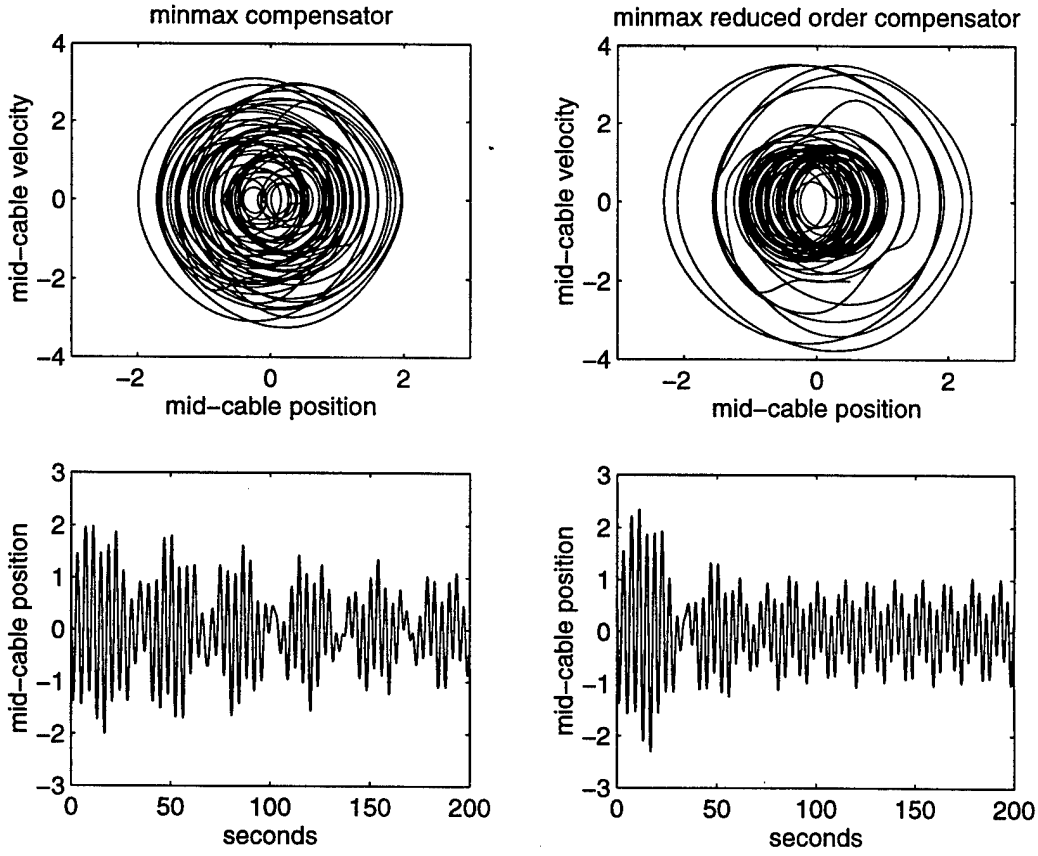


Figure 4.15: MinMax full order vs. MinMax reduced (second) order compensator phase portraits and time histories for mid-cable.

compared with the full order MinMax compensator (see Figure 4.21).

Insight into the performance of the reduced order MinMax compensator can be obtained by computing stability radii for the system. The stability radius of a system gives a measure of the distance to the nearest unstable system, i.e., robustness. The stability radii for the open and closed

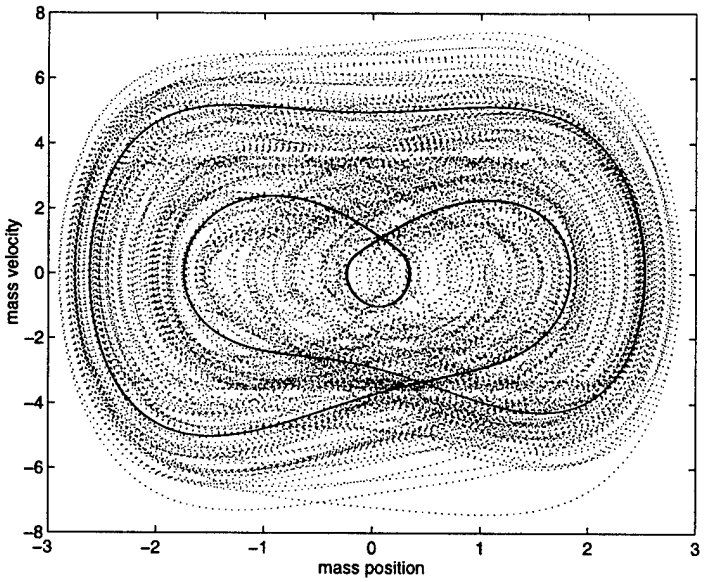


Figure 4.16: Phase portrait of open loop behavior of mass.

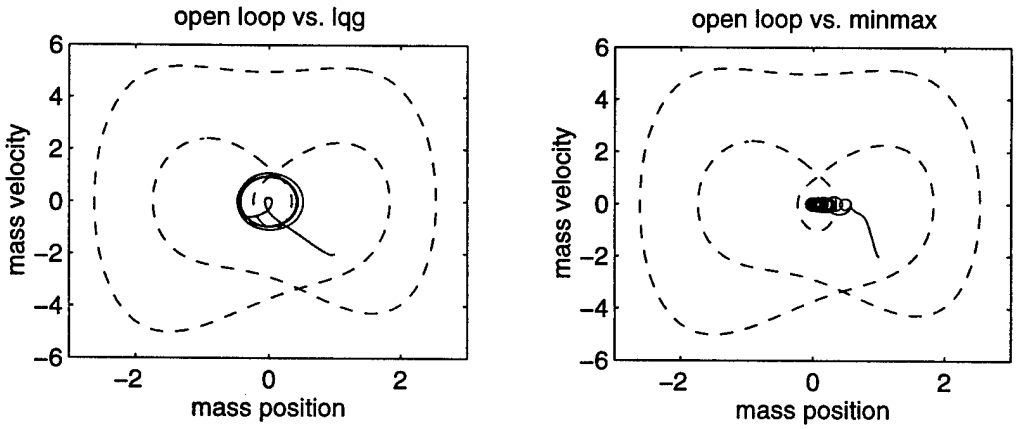


Figure 4.17: Open loop vs. LQG and MinMax compensator phase portraits for the mass.

loop systems formed by various control designs are given below. The full order LQR controller (which we have not used in this discussion), is known to have good robustness results, and so its stability radius is included as a basis for comparison. While MinMax and LQG have similar radii which are an order of magnitude smaller than LQR, the reduced order MinMax compensator has a stability radius which is nearly as large as that of the full state LQR feedback control design.

4.2.4 Conclusions

We can use the results for this model for the MinMax controller to suggest sensor location and to construct a reduced order observer. The basic idea is illustrated in the finite dimensional case by letting k_i denote the i^{th} column of $-B^T P = -K$. Then one can write u_{opt} as the sum of a product

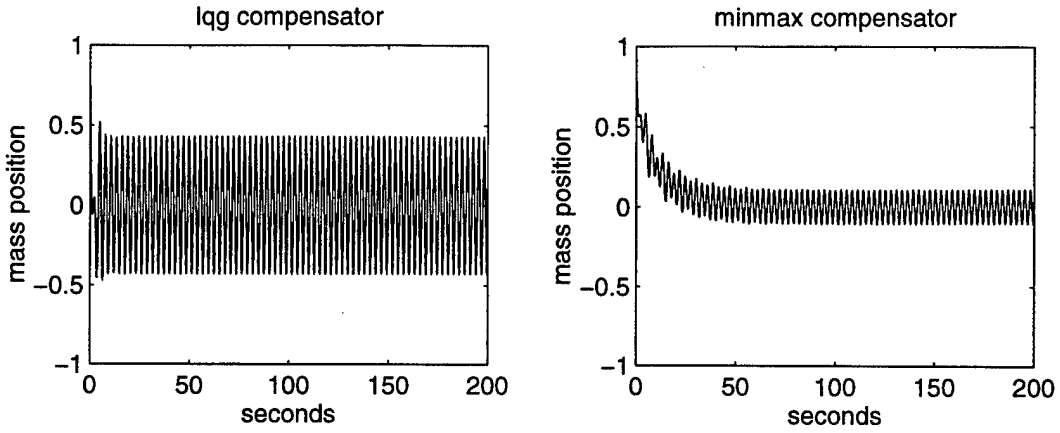


Figure 4.18: Time history of mass position with LQG and MinMax compensators.

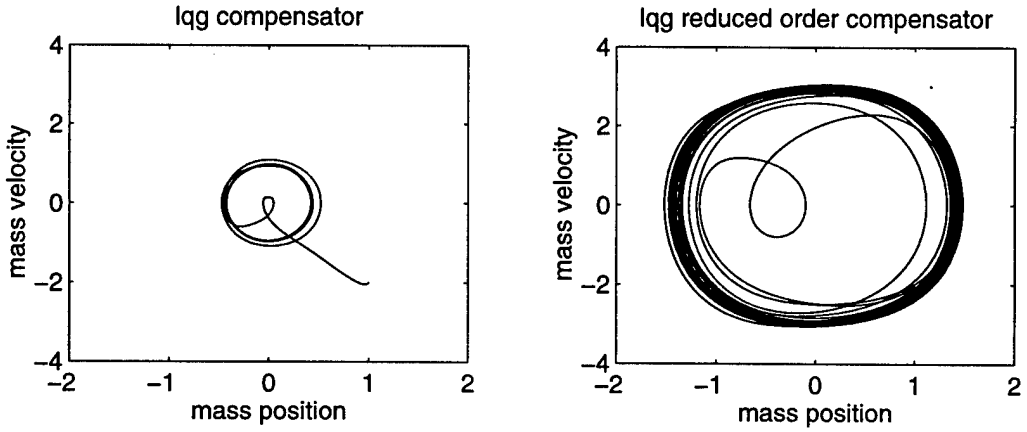


Figure 4.19: LQG full order vs. LQG reduced (second order) compensator phase portraits for the mass.

of the feedback gains, k_i, and the states, $x_{i_{opt}}$,

$$u_{opt}(t) = \sum_{i=1}^{N} k_i x_{i_{opt}}(t). \tag{4.67}$$

An initial approach to sensor placement would be place no sensors at states corresponding to gains which are zero. Conversely, if some gain is large, then the corresponding state should be

Table 4.2: Stability Radii

Open loop	.0003167
LQR Full state	.040609
LQG Full order	.00526176
LQG Reduced order ($N = 2$)	.01704709
MinMax Full order	.00416306
MinMax Reduced order ($N = 2$)	.02601083

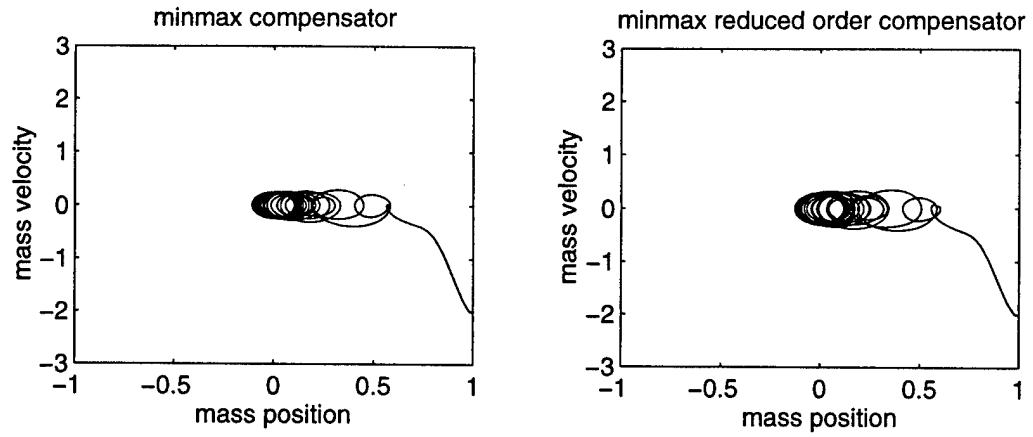


Figure 4.20: Minmax full order vs. Minmax reduced (second order) compensator phase portraits for the mass.

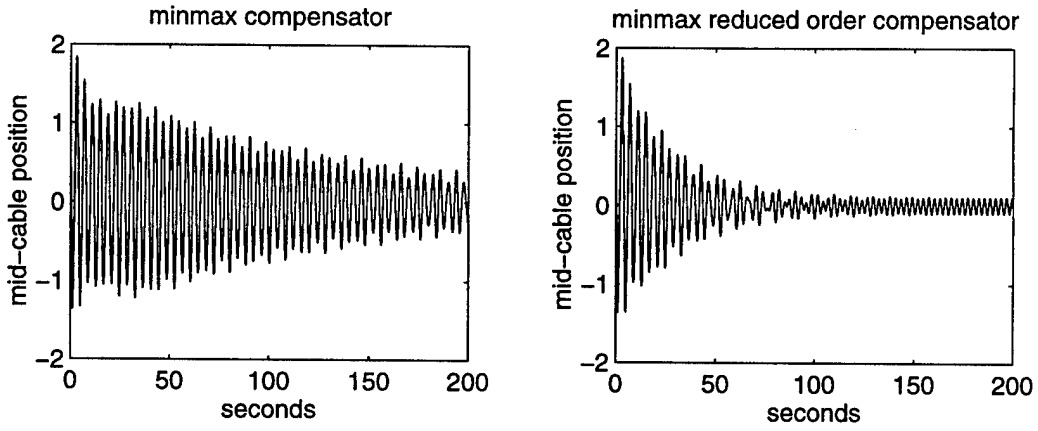


Figure 4.21: Time history of mid-cable position with MinMax full order and MinMax reduced (second) order compensators.

sensed or estimated. However, it is important to note that this simple "analysis" does not take into account such issues as loss of robustness, performance and fault tolerance. Our use of stepwise approximations to the functional gains is a first step toward a reduced order compensator and practical design for implementation.

In conclusion, our numerical experiments show the combination of the MinMax design and the reduced basis state estimator can lead to an effective reduced order nonlinear control law for a nonlinear system of partial and ordinary differential equations. In addition to the errors introduced by numerical approximations, the error that results from the reduced basis approximation of the nonlinear PDE can be viewed as unmodeled dynamics. These observations along with the numerical experiments above suggest that it is worthwhile to investigate the remaining theoretical questions and to investigate nonlinear robust control approaches to these distributed parameter systems.

4.3 Global Solvability for Damped Abstract Nonlinear Hyperbolic Systems

We consider abstract nonlinear second order in time systems with damping. The nonlinearity is assumed to satisfy a monotonicity condition as well as certain smoothness conditions. Well posedness of solutions is established and several examples of interest are discussed. A nonlinear variation-of-parameters representation for solutions in terms of an associated linear semigroup is also given.

In this section, we present new well-posedness results for a class of nonlinear distributed parameter models that arise in a number of applications. Our efforts are a continuation of our earlier endeavors on systems arising in so-called "smart" materials. Indeed, our efforts are basic to our eventual goal of development of computational methodologies for the identification and control of smart material composites undergoing large deformations and/or deformations that fall within the regime of nonlinear stress-strain laws. It is well known in engineering applications that large deformations can occur even when strain levels remain relatively small. More important to certain emerging applications involving composites and certain types of elastomers is that one encounters a nonlinear stress-strain relationship even in the case of small deformations. We describe one such application as a motivating example for the theoretical well-posedness discussions to follow in subsequent sections.

A problem of fundamental interest and great importance in modern material sciences is the development of both passive and active ("smart") vibration devices constructed from polymer (long molecular chains of covalently bonded atoms often having cross-linking chains) composites such as elastomers filled with carbon black and/or silica or with active elements (i.e., piezoelectric, electrostrictive and magnetic or conductive particles). These rubber based products (even without active elements) involve very complex viscoelastic materials that are not at all like metals (where large deformations lead to permanent material changes) and do not satisfy the usual, well-developed linear theory of (infinitesimal) elasticity for deformable bodies.

In considering macroscopic elastic behavior, one finds that the usual constitutive relationships (e.g., Hooke's law) or rheological equations of state for pure elastics are not applicable. Indeed, one observes nonlinearities in both material and geometric behavior - in general, there is a nonlinear relationship between stress and strain even for small strains. Moreover, deformations in the range of practical interest are large and infinitesimal based theories break down.

In spite of these difficulties, there is a substantial literature on modeling of rubber-like elastomers, predominantly based on one of the two rather distinct approaches: (i) molecular (polymer chain) statistical thermodynamic formulations (ii) phenomenological (usually continuum) formulations involving stored energy or strain energy functions (SEF) and/or finite strain (FS) theories. In the phenomenological approach (which will be the basis of our motivating example here) most investigators begin with an isotropic material under homogeneous strain.

Strain energy function theories typically embody only elastic properties of elastomers or rubbers and hence are mostly used in static (equilibrium) finite element analysis of materials (e.g. natural gum rubbers) that exhibit little or no hysteretic behavior. SEF material models, such as those of Mooney-Rivlin, Ogden, Treloar and numerous others, are based on strain invariants I_i, where $I_1 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2$, $I_2 = \lambda_2^2\lambda_3^2 + \lambda_1^2\lambda_3^2 + \lambda_1^2\lambda_2^2$ and $I_3 = \lambda_1^2\lambda_2^2\lambda_3^2$ and the λ_i are the principal extension ratios (deformed length of unit vectors along directions parallel to the principal axes i.e. the axes of zero shear strain).

The finite strain elastic theory of Rivlin is developed with a generalized Hooke's law in an analogy to infinitesimal strain elasticity but makes no "small deformation" assumption and includes higher order exact terms in its formulation. Moreover, finite stresses are defined relative to the deformed body and hence are the "true stresses" as opposed to the "nominal" or "engineering" stresses (relative to the undeformed body) one usually encounters in the infinitesimal linear elasticity used with metals. This Eulerian measure of strain (relative to a coordinate system convected with the deformations) - as opposed to the usual Lagrangian measure (relative to a fixed coordinate system for the undeformed body) - is an important feature of any development of models for use in analytical/computation/experimental investigations of rubber-like material bodies.

Whether one begins with a choice of the SEF or with Rivlin's finite strain formulation, one can use these along with standard material independent force and moment balance derivations as the basis of dynamical models. To illustrate this we take the simplest example: an isotropic, incompressible ($\lambda_1\lambda_2\lambda_3 = 1$) rubber-like rod under simple elongation with a finite applied stress in the principal axis direction x_1. The finite stress theory leads to a true stress $\sigma = \frac{E}{3}(\lambda_1^2 - \frac{1}{\lambda_1})$ for $|\lambda_1| < 1$ or an engineering or nominal stress for what are termed neo-Hookean materials

$$\sigma_{\text{eng}} = \frac{\sigma}{\lambda_1} = \frac{E}{3}\{\lambda_1 - \frac{1}{\lambda_1^2}\} \tag{4.68}$$

where in terms of deformation w in the $x_1 = x$ direction we have (since deformations in the y and z directions are negligible)

$$\lambda_1^2 = \left(1 + \frac{\partial w}{\partial x}\right)^2 \tag{4.69}$$

Here E is a generalized modulus of elasticity.

This can be used in the Timoshenko theory for longitudinal vibrations of a rubber bar to obtain (ρ is the mass density, F is an applied external force)

$$\rho A\frac{\partial^2 w}{\partial t^2} - \frac{\partial S}{\partial x} = F \tag{4.70}$$

where S, the internal (engineering) stress resultant, is given by

$$S = \frac{AE}{3}\{\lambda_1 - \frac{1}{\lambda_1^2}\} = \frac{AE}{3}s\left(\frac{paw}{\partial x}\right) \tag{4.71}$$

with $s(\xi) = 1 - \xi - (1+\xi)^{-2}$ for $|\xi| < 1$ and A is the cross sectional area. This leads to the nonlinear partial differential equation

$$\rho A\frac{\partial^2 w}{\partial t^2} - \frac{\partial}{\partial x}\left(\frac{EA}{3}s\left(\frac{\partial w}{\partial x}\right)\right) = F \tag{4.72}$$

for dynamic longitudinal displacements of a neo-Hookean material rod in extension. Since a series expansion of s yields $s(\xi) = 3\xi - 3\xi^2 + 4\xi^3 - \ldots$, this is readily seen, in the case of small displacements, to reduce to the usual longitudinal deformation equation for Hookean materials. For our subsequent discussions, it is convenient to write (4.72) in the form

$$\rho A\frac{\partial^2 w}{\partial t^2} - \frac{\partial}{\partial x}\left(\frac{EA}{3}\frac{\partial w}{\partial x}\right) - \frac{\partial}{\partial x}\left(\frac{EA}{3}\tilde{g}\left(\frac{\partial w}{\partial x}\right)\right) = F \tag{4.73}$$

where $\tilde{g}(\xi) = 1 - \frac{1}{(1+\xi)^2}$ for $-1 < \xi < 1$. This can be written in a generalized or variational form for a given set of boundary conditions. To be specific, suppose we have a slender rod of length ℓ that satisfies $w(t,0) = w(t,\ell) = 0$. Then defining $\mathcal{V} = H_0^1(0,\ell)$ and $\mathcal{H} = L^2(0,\ell)$ we obtain the usual Gelfand triple $\mathcal{V} \hookrightarrow \mathcal{H} \approx \mathcal{H}^* \hookrightarrow \mathcal{V}^*$ where $\mathcal{V}^* = H^{-1}(0,\ell)$. Then equation (4.73) along with the specified boundary conditions can be written in variational form

$$\rho A w_{tt} + \mathcal{A}_1 w + D^*\tilde{g}(Dw) = F \quad \text{in } \mathcal{V}^* \tag{4.74}$$

where $\mathcal{A}_1 \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$ is given by

$$\langle \mathcal{A}_1\varphi, \psi\rangle_{\mathcal{V}^*,\mathcal{V}} = \left\langle \frac{EA}{3}D\varphi, D\psi \right\rangle_{\mathcal{H}}$$

and $D = \frac{\partial}{\partial x} \in \mathcal{L}(\mathcal{V}, \mathcal{H})$ is the spatial differentiation operator. This model is unrealistic in that it does not include material damping which is known to be present in typical elastomers. If one

assumes an internal damping of the form $\mathcal{A}_2 w_t$ (the exact form of the internal dynamic damping mechanisms in elastomers is a subject of current research - almost nothing is found in the research literature on this even though it is a very important material property that is critical to design of "smart" elastomers), then the model in variational form for the neo-Hookean elastomer rod is given by

$$\rho A w_{tt} + \mathcal{A}_1 w + \mathcal{A}_2 w_t + D^* \tilde{g}(Dw) = F \quad \text{in } \mathcal{V}^* \tag{4.75}$$

4.3.1 Formulation of the Problem

The remainder of this discussion is concerned with establishing global existence of weak solutions for a class of abstract nonlinear damped hyperbolic systems evolving in a complex separable Hilbert space $\mathcal{H}$ (actually holding in the sense of $\mathcal{V}^*$ as explained below):

$$w_{tt} + \mathcal{A}_1 w + \mathcal{A}_2 w_t + \mathcal{N}^* g(\mathcal{N} w) = f(t) \tag{4.76}$$

$$w(0) = \varphi_0 \tag{4.77}$$

$$w_t(0) = \varphi_1 \tag{4.78}$$

Throughout this work we assume there is a sequence of separable Hilbert spaces $\mathcal{V}$, $\mathcal{V}_2$, $\mathcal{H}$, $\mathcal{V}^*$, $\mathcal{V}_2^*$ forming a Gelfand quintuple satisfying

$$\mathcal{V} \hookrightarrow \mathcal{V}_2 \hookrightarrow \mathcal{H} \hookrightarrow \mathcal{V}_2^* \hookrightarrow \mathcal{V}^*$$

where we assume that the embedding $\mathcal{V} \hookrightarrow \mathcal{V}_2$ is dense and continuous with $\|\varphi\|_{\mathcal{V}_2} \leq c\|\varphi\|_{\mathcal{V}}$ for $\varphi \in \mathcal{V}$ and $\mathcal{V}_2 \hookrightarrow \mathcal{H}$ is a dense compact embedding. We denote by $\langle\ ,\ \rangle_{\mathcal{V}^*,\mathcal{V}}$, etc., the usual duality products. These duality products are the extensions by continuity of the inner product in $\mathcal{H}$, denoted by $\langle\ ,\ \rangle$ throughout. The norm in $\mathcal{H}$ will be denoted by $\|\cdot\|$ while those in $\mathcal{V}$, $\mathcal{V}_2$ etc. will carry an appropriate subscript. The operators $\mathcal{A}_1$ and $\mathcal{A}_2$ are defined (under the assumptions below) as usual in terms of their sesquilinear forms $\sigma_1 : \mathcal{V} \times \mathcal{V} \to \mathbb{C}$ and $\sigma_2 : \mathcal{V}_2 \times \mathcal{V}_2 \to \mathbb{C}$. That is, $\mathcal{A}_1 \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*), \mathcal{A}_2 \in \mathcal{L}(\mathcal{V}_2, \mathcal{V}_2^*)$ and $\langle \mathcal{A}_1\varphi, \psi\rangle_{\mathcal{V}^*,\mathcal{V}} = \sigma_1(\varphi,\psi), \langle \mathcal{A}_2\varphi, \psi\rangle_{\mathcal{V}_2^*,\mathcal{V}_2} = \sigma_2(\varphi,\psi)$.

In addition, we make the following assumptions.

A1) The form σ_1 is a Hermitian sesquilinear form: for $\varphi, \psi \in \mathcal{V}$

$$\sigma_1(\varphi,\psi) = \overline{\sigma_1(\psi,\varphi)}. \tag{4.79}$$

A2) The form σ_1 is $\mathcal{V}$ bounded: for $\varphi, \psi \in \mathcal{V}$

$$|\sigma_1(\varphi,\psi)| \leq c_1\|\varphi\|_{\mathcal{V}}\|\psi\|_{\mathcal{V}}. \tag{4.80}$$

A3) The form σ_1 is strictly coercive on $\mathcal{V}$: for $\varphi \in \mathcal{V}$

$$\operatorname{Re}\sigma_1(\varphi,\varphi) = \sigma_1(\varphi,\varphi) \geq k_1\|\varphi\|_{\mathcal{V}}^2, \quad k_1 > 0. \tag{4.81}$$

A4) The form σ_2 is bounded on $\mathcal{V}_2$: for $\varphi, \psi \in \mathcal{V}_2$

$$|\sigma_2(\varphi,\psi)| \leq c_2\|\varphi\|_{\mathcal{V}_2}\|\psi\|_{\mathcal{V}_2}. \tag{4.82}$$

A5) The real part of σ_2 is coercive and is symmetric on $\mathcal{V}_2$:

$$\operatorname{Re}\sigma_2(\varphi,\varphi) + \lambda_0\|\varphi\|^2 \geq k_2\|\varphi\|_{\mathcal{V}_2}^2 \quad k_2 > 0, \lambda_0 \geq 0 \tag{4.83}$$

$$\operatorname{Re}\sigma_2(\varphi,\psi) = \operatorname{Re}\sigma_2(\psi,\varphi), \text{ for any } \varphi, \psi \in \mathcal{V}_2. \tag{4.84}$$

We note that the condition in 4.84 is weaker than requiring that σ_2 be Hermitian.

A6) The forcing term f satisfies

$$f \in L^2([0,T], \mathcal{V}_2^*). \tag{4.85}$$

A7) The operator $\mathcal{N}$ in the nonlinear term satisfies

$$\mathcal{N} \in \mathcal{L}(\mathcal{V}, \mathcal{H}) \text{ with } \|\mathcal{N}\varphi\| \leq \sqrt{k}\,\|\varphi\|_{\mathcal{V}}. \tag{4.86}$$

To prove that weak solutions are unique we need to replace A7) by the strengthened condition:

A7a) The operator $\mathcal{N}$ satisfies

$$\mathcal{N} \in \mathcal{L}(\mathcal{V}_2, \mathcal{H}) \text{ with } \|\mathcal{N}\varphi\| \leq \sqrt{\tilde{k}}\,\|\varphi\|_{\mathcal{V}_2} \tag{4.87}$$

and the range of $\mathcal{N}$ on $\mathcal{V}$ is dense in $\mathcal{H}$.

Note that (4.87) implies (4.86) with $k = c^2\tilde{k}$.

A8) The nonlinear function $g : \mathcal{H} \to \mathcal{H}$ is a continuous nonlinear mapping of real gradient (or potential) type. This means that there exists a continuous Frechet-differentiable nonlinear functional $G : \mathcal{H} \to R^1$, whose Frechet derivative $G'(\varphi) \in \mathcal{L}(\mathcal{H}, R^1)$ at any $\varphi \in \mathcal{H}$ can be represented in the form

$$G'(\varphi)\psi = \mathrm{Re}\langle g(\varphi), \psi\rangle \quad \text{for any } \psi \in \mathcal{H}. \tag{4.88}$$

We also require that there are constants C_1, C_2, C_3 and $\varepsilon > 0$ such that

$$-\frac{1}{2}k^{-1}(k_1 - \varepsilon)\|\varphi\|^2 - C_1 \leq G(\varphi) \leq C_2\|\varphi\|^2 + C_3, \tag{4.89}$$

where k is from 4.86 and k_1 from 4.81.

In the case $\mathcal{V} = \mathcal{V}_2$ it is possible to take $\varepsilon = 0$.

A9) The nonlinear function g also satisfies

$$\|g(\varphi)\| \leq \tilde{C}_1\|\varphi\| + \tilde{C}_2, \quad \varphi \in \mathcal{H}, \tag{4.90}$$

for some constants $\tilde{C}_1$, $\tilde{C}_2$.

An additional condition is necessary for uniqueness of solutions as well as for the integral equation semigroup formulation of the problem discussed in Section 4.3.6 below.

A10) For any $\varphi \in \mathcal{H}$ the Frechet derivative of g exists and satisfies

$$g'(\varphi) \in \mathcal{L}(\mathcal{H}, \mathcal{H}) \text{ with } \|g'(\varphi)\|_{\mathcal{L}(\mathcal{H},\mathcal{H})} \leq \tilde{C}_3. \tag{4.91}$$

Let $\mathcal{L}_T$ denote the space of functions $w : [0,T] \to \mathcal{H}$ such that

$$w \in C_W([0,T], \mathcal{V}_2) \cap L^\infty([0,T], \mathcal{V})$$

(W means weak continuity), and

$$w_t \in C_W([0,T], \mathcal{H}) \bigcap L^2([0,T], \mathcal{V}_2),$$

where the time derivative w_t is understood in the sense of distributions with values in a Hilbert Space. The space $\mathcal{L}_T$ is equipped with the norm

$$\|w\|_{\mathcal{L}_T} = \operatorname*{ess\,sup}_{t\in[0,T]} (\|w_t(t)\| + \|w(t)\|_{\mathcal{V}}) + \left(\int_0^T \|w_t(t)\|_{\mathcal{V}_2}^2 dt\right)^{1/2}. \tag{4.92}$$

A11) We assume that for any $u, v \in \mathcal{L}_T$, the following inequality is satisfied for any $t \in [0,T]$:

$$\begin{aligned}&\int_0^t \Big\{ \mathrm{Re}\langle g(\mathcal{N}u(\tau)) - g(\mathcal{N}v(\tau)), \mathcal{N}u(\tau) - \mathcal{N}v(\tau)\rangle \\ &+k_1 k^{-1} \|\mathcal{N}u(\tau) - \mathcal{N}v(\tau)\|^2 \Big\}\, dt \\ &+a\left(\left(\int_0^t \|u(\tau) - v(\tau)\|^2\, dt\right)^{1/2}\right) \geq 0,\end{aligned} \tag{4.93}$$

where $a(\xi) \geq 0$ is a continuous function in $\xi \geq 0$ such that

i) $a(0) = 0$,

ii) there exists a first derivative such that $a'(0) = 0$.

Note that 4.93 is satisfied if, for example,

$$\mathrm{Re}\langle g(\varphi) - g(\psi), \varphi - \psi\rangle + k_1 k^{-1} \|\varphi - \psi\|^2 \geq 0 \tag{4.94}$$

for any $\varphi, \psi \in \mathcal{H}$, where k and k_1 are the constants in 4.86 and 4.102.

We say that $w \in \mathcal{L}_T$ is a *weak solution* of the problem 4.76– 4.78 with $\varphi_0 \in \mathcal{V}$ and $\varphi_1 \in \mathcal{H}$ if it satisfies the equation:

$$\begin{aligned}&\int_0^t \Big[-\langle w_\tau(\tau), \eta_\tau(\tau)\rangle + \sigma_1(w(\tau), \eta(\tau)) + \sigma_2(w_\tau(\tau), \eta(\tau)) + \\ &+\langle g(\mathcal{N}w(\tau)), \mathcal{N}\eta(\tau)\rangle \Big] d\tau + \langle w_t(t), \eta(t)\rangle = \\ &= \langle \varphi_1, \eta(0)\rangle + \int_0^t \langle f(\tau), \eta(\tau)\rangle_{\mathcal{V}_2^*, \mathcal{V}_2} d\tau,\end{aligned} \tag{4.95}$$

for any $t \in [0,T]$ and any $\eta \in \mathcal{L}_T$, as well as the initial condition

$$w(0) = \varphi_0. \tag{4.96}$$

We note that this notion of weak solution of (4.76)–(4.78) agrees with the usual one in that it yields $w_{tt} \in L^2([0,T], \mathcal{V}^*) = L^2([0,T], \mathcal{V})^*$ with (4.76) holding in the sense of $L^2([0,T], \mathcal{V}^*)$. Also, the class of test functions η used in this definition are somewhat smoother than necessary (e.g., see the remarks following (4.141) in the existence theorem below).

4.3.2 The Main *a priori* Estimate

Under our standing assumption A1)-A9) and the formulations of the previous section, Equation (4.76) or equivalently (4.95) can be written

$$\langle w_{tt}, \eta\rangle_{\mathcal{V}^*, \mathcal{V}} + \sigma_1(w, \eta) + \sigma_2(w_t, \eta) + \langle g(\mathcal{N}w), \mathcal{N}\eta\rangle = \langle f, \eta\rangle_{\mathcal{V}_2^*, \mathcal{V}_2} \tag{4.97}$$

for all $\eta \in \mathcal{L}_T$ and for almost all $t \in [0,T]$.

Treating this equation formally for the present, choosing $\eta = w_t$ and taking the real part we find that if a solution exists, it must satisfy:

$$\frac{d}{dt}\left\{\frac{1}{2}\|w_t\|^2 + \frac{1}{2}\sigma_1(w,w) + G(\mathcal{N}w)\right\} + \mathrm{Re}\,\sigma_2(w_t, w_t) = \mathrm{Re}\langle f, w_t\rangle_{\mathcal{V}_2^*, \mathcal{V}_2} \tag{4.98}$$

Here we have used the fact that due to 4.88 we have

$$\frac{d}{dt}G(\mathcal{N}w) = \mathrm{Re}\langle g(\mathcal{N}w), \mathcal{N}w_t\rangle.$$

Using the conditions 4.79 – 4.83, 4.85, 4.86, (4.89) we obtain from 4.98:

$$\|w_t\|^2 + \varepsilon\|w\|_{\mathcal{V}}^2 + k_2\int_0^t \|w_\tau(\tau)\|_{\mathcal{V}_2}^2 d\tau \le \|\varphi_1\|^2 + \tilde{c}_1\|\varphi_0\|_{\mathcal{V}}^2 +$$

$$+\frac{1}{k_2}\int_0^t \|f(\tau)\|_{\mathcal{V}_2^*}^2 d\tau + 2\lambda_0\int_0^t \|w_\tau(\tau)\|^2 d\tau + 2C_1 + 2C_3, \tag{4.99}$$

where $\tilde{c}_1 = c_1 + 2kC_2$.

To obtain 4.99 we first use 4.83 in 4.98 to obtain

$$\frac{d}{dt}\left\{\frac{1}{2}\|w_t\|^2 + \frac{1}{2}\sigma_1(w,w) + G(\mathcal{N}w)\right\} + k_2\|w_t\|_{\mathcal{V}_2}^2 \le$$

$$\delta\|w_t\|_{\mathcal{V}_2}^2 + \frac{1}{4\delta}\|f\|_{\mathcal{V}_2^*}^2 + \lambda_0\|w_t\|^2 \tag{4.100}$$

for any $\delta > 0$. We next choose $\delta = k_2/2$, integrate the terms in 4.100 from 0 to t and use

$$\sigma_1(w,w) + 2G(\mathcal{N}w) \ge k_1\|w\|_{\mathcal{V}}^2 - k^{-1}(k_1-\varepsilon)\|\mathcal{N}w\|^2 - 2C_1$$

$$\ge k_1\|w\|_{\mathcal{V}}^2 - (k_1-\varepsilon)\|w\|_{\mathcal{V}}^2 - 2C_1 = \varepsilon\|w\|_{\mathcal{V}}^2 - 2C_1,$$

$$G(\mathcal{N}\varphi_0) \le kC_2\|\varphi_0\|_{\mathcal{V}}^2 + C_3$$

and

$$|\sigma_1(\varphi_0,\varphi_0)| \le c_1\|\varphi_0\|_{\mathcal{V}}^2.$$

Having established 4.99, by ignoring the 2nd and 3rd terms on the left in 4.99 and applying Gronwall's lemma, we obtain

$$\|w_t(t)\|^2 \le \left(\|\varphi_1\|^2 + c_1\|\varphi_0\|_{\mathcal{V}}^2 + \frac{1}{k_2}\int_0^T \|f(\tau)\|_{\mathcal{V}_2^*}^2 d\tau + 2C_1 + 2C_3\right)e^{2\lambda_0 t}. \tag{4.101}$$

Substituting 4.101 back into 4.99 we have

$$\|w_t\|^2 + \varepsilon\|w\|_{\mathcal{V}}^2 + k_2\int_0^t \|w_\tau(\tau)\|_{\mathcal{V}_2}^2 d\tau \le C, \tag{4.102}$$

where the constant $C = C(\|\varphi_1\|, \|\varphi_0\|_{\mathcal{V}}, \|f\|_{L^2([0,T],\mathcal{V}_2^*)})$ is easily computable.

4.3.3 Galerkin Approximations

Let $\{\psi_k\}_{k=1}^{\infty} \subset \mathcal{V}$ be any total linearly independent system in $\mathcal{V}$. We assume without loss of generality that the elements ψ_j have been normalized in $\mathcal{V}$ and hence are uniformly bounded in $\mathcal{H}$ and $\mathcal{V}$.

We define the "Galerkin" approximations for (4.76) by

$$w^N(t) = \sum_{k=1}^{N} c_k^N(t)\psi_k, \tag{4.103}$$

where the $\{c_k^N(t)\}_{k=1}^N$ are chosen so that $w^N(t)$ is the unique solution of

$$\frac{d^2}{dt^2}(w^N(t),\psi_j) + \sigma_1(w^N(t),\psi_j) + \frac{d}{dt}\sigma_2(w^N(t),\psi_j) +$$
$$(g(\mathcal{N}w^N(t)),\mathcal{N}\psi_j) = (f,\psi_j)_{\mathcal{V}_2',\mathcal{V}_2} \tag{4.104}$$
$$\text{for } j = 1,\ldots,N, \text{ with initial conditions}$$
$$c_k^N(0) = c_{0k}^N, \quad \frac{d}{dt}c_k^N(0) = c_{1k}^N,$$

where $\{c_{0k}^N\}, \{c_{1k}^N\}$ are chosen so that $\varphi_0 = \lim_{N\to\infty}\sum_1^N c_{0k}^N\psi_k, \varphi_1 = \lim_{N\to\infty}\sum_1^N c_{1k}^N\psi_k$ where the limits are in the $\mathcal{V}$ and $\mathcal{H}$ sense, respectively.

Multiplying 4.104 by $\frac{d}{dt}\overline{c_j^N(t)}$ and summing over $j = 1,\ldots,N$, we obtain 4.98 with w replaced by w^N. Repeating the above arguments, we then obtain

$$\|w_t^N\|^2 + \varepsilon\|w^N\|_{\mathcal{V}}^2 + k_2\int_0^t \|w_\tau^N(\tau)\|_{\mathcal{V}_2}^2 d\tau \le \widetilde{C}, \tag{4.105}$$

where the constant $\tilde{C}$ is independent of N, depending only on φ_0, φ_1 and f as in the constant C of 4.102. (We note that the convergences $\varphi_0^N \to \varphi_0$ in $\mathcal{V}$, $\varphi_1^N \to \varphi_1$ in $\mathcal{H}$ guarantee uniform boundedness of $\|\varphi_0^N\|_{\mathcal{V}}$ and $\|\varphi_1^N\|$.)

4.3.4 Convergence of the Galerkin Approximations

To establish existence of solutions to (4.76)-(4.78), we shall use the bounds of (4.105) to extract successive subsequences of the Galerkin approximations and argue that the final subsequence converges to a solution for the problem. In these arguments, we shall not distinguish subsequences but shall denote by the same symbol $\{w^N\}$ the subsequences of $\{w^N\}_{N=1}^\infty$ selected at each step.

It follows from 4.105 that the set $\{w^N\}$ is bounded in $C([0,T],\mathcal{V}) \subset L^2([0,T],\mathcal{V})$ and $\{w_t^N\}$ is bounded in $C([0,T],\mathcal{H})$ and in $L^2([0,T],\mathcal{V}_2)$. This allows us to conclude that there exist a subsequence such that

$$w^N \longrightarrow w \text{ weakly in } L^2([0,T],\mathcal{V}) \tag{4.106}$$
$$w_t^N \longrightarrow \widehat{w} \text{ weakly in } L^2([0,T]),\mathcal{V}_2). \tag{4.107}$$

We can readily show that $w_t(t)$ exists in the $\mathcal{V}_2$ sense and $w_t(t) = \widehat{w}(t)$ a.e. in $[0,T]$. These considerations are sufficient to treat the linear problem. The nonlinear case requires some additional effort.

The main lemma needed to carry out the proof of existence is the following.

Lemma 1 *There exists a subsequence $\{w^N\}$ of the original sequence of Galerkin approximations and $w \in \mathcal{L}_T$ such that the following statements hold.*

a)
$$w^N \to w \text{ weakly in } L^2([0,T],\mathcal{V}); \tag{4.108}$$

b) The set $\{w^N\}$ is an equicontinuous and bounded subset of $C([0,T],\mathcal{V}_2)$; moreover,
$$w^N(t) \to w(t) \text{ weakly in } \mathcal{V}_2 \tag{4.109}$$
uniformly in $t \in [0,T]$, i.e., $w^N \to w$ in $C_W([0,T],\mathcal{V}_2)$;

c)
$$w_t^N \to w_t \text{ weakly in } L^2([0,T],\mathcal{V}_2); \tag{4.110}$$

d) The set $\{w_t^N\}$ is bounded in $C([0,T],\mathcal{H})$ and equicontinuous in $C_W([0,T],\mathcal{H})$; moreover

$$w_t^N(t) \to w_t(t) \text{ weakly in } \mathcal{H} \tag{4.111}$$

uniformly in $t \in [0,T]$;

e)

$$w_t^N \to w_t \text{ strongly in } L^2([0,T],\mathcal{H}); \tag{4.112}$$

f) There exists $h \in L^2([0,T],\mathcal{H})$ such that

$$g(\mathcal{N}w^N) \to h \text{ weakly in } L^2([0,T],\mathcal{H}). \tag{4.113}$$

Remark 1 i) The statements 4.108 and 4.110 are just repetitions of (4.106) and (4.107). Statement f) follows immediately from (4.105), A7) and A9).

ii) We shall make use of the following version of the Arzela-Ascoli theorem: If Y is a complete metric space and $\mathcal{F} \subset C([0,T],Y)$, then $\mathcal{F}$ is relatively compact if and only if $\mathcal{F}$ is equicontinuous and $\{f(t) : f \in \mathcal{F}\}$ is relatively compact in Y for each $t \in [0,T]$.

iii) Note that b) along with the compactness of the embedding $\mathcal{V}_2 \subset \mathcal{H}$ implies that

$$w^N \to w \text{ strongly in } C([0,T],\mathcal{H}). \tag{4.114}$$

iv) Statement 4.110 does not imply 4.112, since the embedding $L^2([0,T],\mathcal{V}_2) \subset L^2([0,T],\mathcal{H})$ is not compact even though $\mathcal{V}_2$ embeds compactly in $\mathcal{H}$.

For this lemma, we thus only need to prove the statements b), d) and e). We consider b) first. From the main *a priori* estimate (4.105) we see that $\{w^N\}$ is bounded in $C([0,T],\mathcal{V})$:

$$\max_{t\in[0,T]} \|w^N(t)\|_{\mathcal{V}}^2 \le \varepsilon^{-1}\tilde{C}. \tag{4.115}$$

Since $\mathcal{V} \hookrightarrow \mathcal{V}_2$, the set is also bounded in $C([0,T],\mathcal{V}_2)$. Furthermore, we have

$$\begin{aligned} \|w^N(t+\triangle t) - w^N(t)\|_{\mathcal{V}_2}^2 &= \left\| \int_t^{t+\triangle t} w_\tau^N(\tau) d\tau \right\|_{\mathcal{V}_2}^2 \\ \le \left(\int_t^{t+\triangle t} \|w_\tau^N(\tau)\|_{\mathcal{V}_2} d\tau \right)^2 &\le \triangle t \int_t^{t+\triangle t} \|w_\tau^N(\tau)\|_{\mathcal{V}_2}^2 d\tau \\ &\le k_2^{-1}\tilde{C} \triangle t, \end{aligned} \tag{4.116}$$

and the desired equicontinuity follows. The convergence statement (4.109) then results from use of the Arzela-Ascoli theorem (see ii) of Remark 1) with Y chosen as the appropriate closed bounded subset in $\mathcal{V}_2$ taken with the weak topology. (Recall that Y is then a compact metric space, and the equicontinuity in the sense of (4.116) implies the equicontinuity in the sense of the metric of Y.)

Next, we turn to e) and use a lemma of Aubin which can be stated succinctly as follows: Let $X_0 \hookrightarrow X \hookrightarrow X_1$ where the embedding $X_0 \hookrightarrow X$ is compact. Define the space
$Y = \{y \in L^2([0,T],X_0) : y_t \in L^2([0,T],X_1)\}$ with norm

$$\|y\|_Y = \|y\|_{L^2([0,T],X_0)} + \|y_t\|_{L^2([0,T],X_1)}.$$

Then the embedding $Y \hookrightarrow L^2([0,T],X)$ is compact.

It is actually a corollary of the lemma which we use; it can be stated as follows for our configuration of spaces: Suppose $\{w_t^N\}$ is bounded in $L^2([0,T],\mathcal{V}_2)$ and $\{w_{tt}^N\}$ is bounded in $L^2([0,T],\mathcal{V}^*)$ then $\{w_t^N\}$ is relatively compact in $L^2([0,T],\mathcal{H})$. To obtain this corollary, choose

$X_0 = \mathcal{V}_2$, $X = \mathcal{H}$ and $X_1 = \mathcal{V}^*$ in Aubin's lemma.

Since we have $\{w_t^N\}$ bounded in $L^2([0,T],\mathcal{V}_2)$ by the *a priori* estimate (4.105), it suffices for e) of Lemma 1 to argue that $\{w_{tt}^N\}$ is bounded in $L^2([0,T],\mathcal{V}^*)$. Since $L^2([0,T],\mathcal{V}^*) = L^2([0,T],\mathcal{V})^*$, it is suffices to show that

$$\begin{aligned} |w_{tt}^N(\Phi)| &= |\int_0^T \langle w_{tt}^N(\tau), \Phi(\tau)\rangle_{\mathcal{V}^*,\mathcal{V}}\, d\tau| \\ &\leq K\|\Phi\|_{L^2([0,T],\mathcal{V})} \quad \text{for any } \Phi \in L^2([0,T],\mathcal{V}). \end{aligned}$$

For fixed M, let Φ_M be of the form $\Phi_M(t) = \sum_{k=1}^M a_k(t)\psi_k$ where $a_k \in C^1[0,T]$. From the equation for w^N we find that for $N \geq M$ we must have

$$\begin{aligned} |w_{tt}^N(\Phi_M)| &= |\int_0^T \langle w_{tt}^N(\tau), \Phi_M(\tau)\rangle_{\mathcal{V}^*,\mathcal{V}}\, d\tau| \\ &\leq c_1\int_0^T \|w^N(\tau)\|_{\mathcal{V}}\|\Phi_M(\tau)\|_{\mathcal{V}}\, d\tau + c_2\int_0^T \|w_t^N(\tau)\|_{\mathcal{V}_2}\|\Phi_M(\tau)\|_{\mathcal{V}_2}\, d\tau \\ &+ \sqrt{k}\int_0^T \|g(\mathcal{N}w^N(\tau))\|\,\|\Phi_M(\tau)\|_{\mathcal{V}}\, d\tau + \int_0^T \|f(\tau)\|_{\mathcal{V}_2^*}\|\Phi_M(\tau)\|_{\mathcal{V}_2}\, d\tau. \end{aligned}$$

From the *a priori* bounds, A9), A6) and standard inequalities, we find that this estimate leads to

$$|w_{tt}^N(\Phi_M)| \leq K\|\Phi_M\|_{L^2([0,T],\mathcal{V})} \tag{4.117}$$

where the constant K depends on $c_1, c_2, k, \tilde{C}, \tilde{C}_1, \tilde{C}_2$ but not N or Φ_M. Since elements of the form $\{\Phi_M\}_{M=1}^\infty$ form a dense subset of $L^2([0,T],\mathcal{V})$, the desired boundedness is readily inferred and thus e) is established.

Finally, we consider d). The boundedness statement follows from (4.105) and the convergence statement will once again follow from an application of Arzela-Ascoli in $C_W([0,T],\mathcal{H})$ once we establish the equicontinuity. To do this we first note that

$$|\langle w_t^N(t+\Delta t) - w_t^N(t), v\rangle| \leq \hat{k}\|v\|_{\mathcal{V}}\sqrt{|\Delta t|} \tag{4.118}$$

for $v \in \mathcal{V}$ which is obtained using arguments similar to those employed in obtaining (4.116) and (4.117). Assume now that $\varphi \in \mathcal{H}$ and fix $\epsilon > 0$. For $v \in \mathcal{V}$ (and $t, t+\Delta t \in [0,T]$) we have

$$\begin{aligned} &|\langle w_t^N(t+\Delta t) - w_t^N(t), \varphi\rangle| \\ &\leq |\langle w_t^N(t+\Delta t) - w_t^N(t), v\rangle| + |\langle w_t^N(t+\Delta t) - w_t^N(t), \varphi - v\rangle| \\ &\leq \hat{k}\|v\|_{\mathcal{V}}\sqrt{|\Delta t|} + 2\tilde{C}\|\varphi - v\|, \end{aligned} \tag{4.119}$$

where $\tilde{C}$ is the constant from (4.105). Selecting v so that $2\tilde{C}\|\varphi - v\| \leq \epsilon/2$ we can conclude that the right hand side of (4.119) does not exceed ϵ for any N if $|\Delta t| \leq \delta = (\epsilon/2\hat{k}\|v\|_{\mathcal{V}})^2$, and the desired equicontinuity follows.

Remark 2 We note that the statement b) of Lemma 1 can be strengthened. Namely, the set $\{w^N\}$ is equicontinuous in $C_W([0,T],\mathcal{V})$. Since, due to (4.105), it is also bounded in $C([0,T],\mathcal{V})$ we can conclude that

$$w^N \to w \text{ in } C_W([0,T],\mathcal{V}). \tag{4.120}$$

The convergence (4.120) will not be used in the next section, where we show that w satisfies (4.95) and (4.96). However, (4.120) is important to verify that $w \in \mathcal{L}_T$. Indeed, from Lemma 1 we can only conclude that $w \in C_W([0,T],\mathcal{V}_2) \cap L^2([0,T],\mathcal{V})$ and

$w_t \in C_W([0,T],\mathcal{H}) \cap L^2([0,T],\mathcal{V}_2)$, which is sufficient for (4.95) to make sense, but we cannot conclude that $w \in L^\infty([0,T],\mathcal{V})$.

We sketch the proof of (4.120) which establishes, in fact, that $w \in C_W([0,T],\mathcal{V})$. Note that, due to our assumptions A1)-A3) on σ_1, the mapping $\mathcal{A}_1 : \mathcal{V} \to \mathcal{V}^*$ is a topological isomorphism. Define $\mathrm{dom}\mathcal{A}_1 = \{v \in \mathcal{V} : \mathcal{A}_1 v \in \mathcal{H}\} = \mathcal{A}_1^{-1}(\mathcal{H})$. Since $\mathcal{H}$ is dense in $\mathcal{V}^*$, we see that $\mathrm{dom}\mathcal{A}_1$ is dense in $\mathcal{V}$. We prove the equicontinuity of $\{w^N\}$ in $C_W([0,T],\mathcal{V})$ assuming that $\mathcal{V}$ is equipped with the inner product $\sigma_1(\cdot,\cdot)$, which is equivalent to the original inner product in $\mathcal{V}$. Let $v \in \mathrm{dom}\mathcal{A}_1$. Then we have

$$\begin{aligned} |\sigma_1(v, w^N(t+\Delta t) - w^N(t))| &= |\langle \mathcal{A}_1 v, w^N(t+\Delta t) - w^N(t)\rangle| \\ &\le \|\mathcal{A}_1 v\| \int_t^{t+\Delta t} \|w_\tau^N(\tau)\| d\tau \le \sqrt{\tilde{C}} \|\mathcal{A}_1 v\| \, |\Delta t|, \end{aligned} \tag{4.121}$$

where $\tilde{C}$ is the constant from (4.105). Since $\mathrm{dom}\mathcal{A}_1$ is dense in $\mathcal{V}$, the desired equicontinuity can be deduced from (4.121) by an argument similar to the one used in deriving d) from (4.118).

4.3.5 Existence of Weak Solutions

In this section we verify, based on Lemma 1, that it is possible to pass to the limit in an integral identity (see 4.124 below) for the Galerkin approximations. We thus obtain the fundamental existence results.

Theorem 3 *Under assumptions A1) - A9) and A11), there exists a weak solution of (4.76)-(4.78) (or equivalently (4.97), (4.77), (4.78)). If in addition, A7a) and A10) hold, then the solution is unique.*

To give the arguments for this theorem, we denote by $\mathcal{P}_M$ $(M = 1, 2, \cdots)$ the class of functions $\eta \in \mathcal{L}_T$, which can be represented in the form

$$\eta(t) = \sum_{k=1}^{M} a_k(t)\psi_k, \tag{4.122}$$

where $a_k \in C^1([0,T])$. Let

$$\mathcal{P} = \bigcup_{M=1}^{\infty} \mathcal{P}_M. \tag{4.123}$$

It is obvious that $\mathcal{P}$ is dense in $\mathcal{L}_T$. Recalling the definition of the Galerkin approximation in 4.103, 4.104, we multiply the jth equation in 4.104 by $a_j(t)$, take the sum from 1 to M and integrate over $[0,t]$ to obtain

$$\begin{aligned} &\int_0^t [-\langle w_\tau^N(\tau), \eta_\tau(\tau)\rangle + \sigma_1(w^N(\tau), \eta(\tau)) + \sigma_2(w_\tau^N(\tau), \eta(\tau)) + \\ &\langle g(\mathcal{N}w^N(\tau)), \mathcal{N}\eta(\tau))\rangle] \, d\tau + \langle w_t^N(t), \eta(t)\rangle - \langle w_t^N(0), \eta(0)\rangle \\ &= \int_0^t \langle f(\tau), \eta(\tau)\rangle_{\mathcal{V}_2^*, \mathcal{V}_2} \, d\tau \end{aligned} \tag{4.124}$$

which is satisfied for all $\eta \in \mathcal{P}_M$, for $M \le N$.

Now, fix $\eta \in \mathcal{P}_M$ with $M \le N$. Using 4.108-4.113, we can pass to the limit $N \to \infty$ in 4.124 and obtain

$$\begin{aligned} &\int_0^t [-\langle w_\tau(\tau), \eta_\tau(\tau)\rangle + \sigma_1(w(\tau), \eta(\tau)) + \sigma_2(w_\tau(\tau), \eta(\tau)) + \\ &\langle h(\tau), \mathcal{N}\eta(\tau))\rangle] \, d\tau + \langle w_t(t), \eta(t)\rangle - \langle \varphi_1, \eta(0)\rangle \\ &= \int_0^t \langle f(\tau), \eta(\tau)\rangle_{\mathcal{V}_2^*, \mathcal{V}_2} \, d\tau. \end{aligned} \tag{4.125}$$

Let us explain the latter statement in a more detailed manner. Note, first of all, that all the statements of Lemma 1 are true for any interval $[0,t]$, $t \le T$. Now, let us examine all four terms under the integral in the left-hand side of (4.124). To pass to the limit in the first term we only need the weak convergence

$w_\tau^N \to w_\tau$ in $L^2([0,t],\mathcal{H})$, which follows, e.g., from (4.110) or from (4.112). To treat the second term we note that for a fixed $\eta \in \mathcal{L}_T$ the mapping $u \to \int_0^t \sigma_1(u(\tau),\eta(\tau))d\tau$ is a bounded linear functional on $L^2([0,t],\mathcal{V})$ due to (4.80). Therefore, this functional is weakly continuous, and we can pass to the limit due to (4.108). A similar argument holds for the third term due to (4.82). In the fourth term we can pass to the limit due to (4.113).

Finally, in the first term outside the integral in the left side of (4.124) we can pass to the limit due to (4.111) and in the second term due to the fact that $w_t^N(0) \to \varphi_1$ in $\mathcal{H}$ as $N \to \infty$.

Equation 4.125 is satisfied for all $\eta \in \mathcal{P}_M$ for all M, and, therefore for all $\eta \in \mathcal{L}_T$ since $\mathcal{P}$ is dense in $\mathcal{L}_T$.

Except for the term involving the limit function h, this is the equation for weak solutions (see (4.95)). The condition (4.96) is clearly satisfied since $w^N(0) \to \varphi_0$ in $\mathcal{V}$ as $N \to \infty$. We argue that the h term is the correct term involving $g(\mathcal{N}w(t))$ to yield that the limit function w is a weak solution.

To prove this we use the Minty-Browder monotonicity method.

Lemma 2 *For any $\eta \in \mathcal{L}_T$ and for $t \in [0,T]$*

$$\int_0^t \langle g(\mathcal{N}w(\tau)), \mathcal{N}\eta(\tau)\rangle \, d\tau = \int_0^t \langle h(\tau), \mathcal{N}\eta(\tau)\rangle \, d\tau \tag{4.126}$$

Proof: The condition 4.93 plays a crucial role in this proof. Combining 4.93 with 4.86, we obtain

$$\int_0^t \Big[\mathrm{Re}\langle g(\mathcal{N}u(\tau)) - g(\mathcal{N}v(\tau)), \mathcal{N}u(\tau) - \mathcal{N}v(\tau)\rangle + \tag{4.127}$$

$$k_1 \|u(\tau) - v(\tau)\|_{\mathcal{V}}^2 \Big] \, d\tau + a\left(\left(\int_0^t \|u(\tau) - v(\tau)\|^2 \, d\tau \right)^{1/2} \right) \ge 0,$$

for any $u, v \in \mathcal{L}_T$.

Now consider 4.127 with $u = w^N \in \mathcal{L}_T$ and any $v \in \mathcal{P}_M \subset \mathcal{L}_T$ with $M \le N$. Taking into account 4.81 we obtain

$$\int_0^t \big[\mathrm{Re}\langle g(\mathcal{N}w^N(\tau)) - g(\mathcal{N}v(\tau)), \mathcal{N}w^N(\tau) - \mathcal{N}v(\tau)\rangle + \tag{4.128}$$

$$\sigma_1(w^N(\tau) - v(\tau), w^N(\tau) - v(\tau))\big] \, d\tau + a\left(\|w^N - v\|_{L^2([0,T],\mathcal{H})} \right) \ge 0.$$

We next return to 4.124 with $\eta = w^N - v$, (notice that this is possible since $w^N \in \mathcal{P}_N$ and $v \in \mathcal{P}_M$ with $M \le N$). Taking the real parts of both sides, we obtain an expression which can be written in the form

$$\mathrm{Re} \int_0^t \langle g(\mathcal{N}w^N(\tau)), \mathcal{N}w^N(\tau) - \mathcal{N}v(\tau)\rangle \, d\tau = \tag{4.129}$$

$$\mathrm{Re} \int_0^t \big[\langle w_\tau^N(\tau), w_\tau^N(\tau) - v_\tau(\tau)\rangle - \sigma_1(w^N(\tau), w^N(\tau) - v(\tau)) -$$

$$- \sigma_2(w_\tau^N(\tau), w^N(\tau) - v(\tau))\big] \, d\tau - \mathrm{Re}\langle w_t^N(t), w^N(t) - v(t)\rangle +$$

$$+ \mathrm{Re}\langle w_t^N(0), w^N(0) - v(0)\rangle + \mathrm{Re} \int_0^t \langle f(\tau), w^N(\tau) - v(\tau)\rangle_{\mathcal{V}_2^*, \mathcal{V}_2} \, d\tau$$

Substituting 4.129 into 4.128 we obtain, after a straightforward simplification: for all $v \in \mathcal{P}_M$ $(M \leq N)$

$$\begin{aligned}
&\|w_t^N\|^2_{L^2([0,t],\mathcal{H})} - \frac{1}{2}\operatorname{Re}\sigma_2(w^N(t),w^N(t)) \qquad (4.130)\\
&+\frac{1}{2}\operatorname{Re}\sigma_2(w^N(0),w^N(0)) + \operatorname{Re}\int_0^t \Big[-\langle w_\tau^N(\tau), v_\tau(\tau)\rangle + \\
&+\sigma_2(w_\tau^N(\tau),v(\tau)) + \langle f(\tau), w^N(\tau)-v(\tau)\rangle_{\mathcal{V}_2^*,\mathcal{V}_2}\\
&-\langle g(\mathcal{N}v(\tau)), \mathcal{N}w^N(\tau)-\mathcal{N}v(\tau)\rangle - \sigma_1(v(\tau), w^N(\tau)-v(\tau))\Big]\,d\tau\\
&-\operatorname{Re}\langle w_t^N(t), w^N(t)-v(t)\rangle + \operatorname{Re}\langle w_t^N(0), w^N(0)-v(0)\rangle\\
&+a\left(\|w^N - v\|_{L^2([0,T],\mathcal{H})}\right) \geq 0.
\end{aligned}$$

Here we have used the fact that, due to the symmetry of the real part of σ_2 it follows that:

$$\operatorname{Re}\sigma_2(w_t^N, w^N) = \frac{1}{2}\frac{d}{dt}\operatorname{Re}\sigma_2(w^N, w^N).$$

Now the most important observation is that we can pass to the limit $N \to \infty$ in 4.130 to obtain

$$\begin{aligned}
&\|w_t\|^2_{L^2([0,T],\mathcal{H})} - \frac{1}{2}\operatorname{Re}\sigma_2(w(t),w(t)) \qquad (4.131)\\
&+\frac{1}{2}\operatorname{Re}\sigma_2(w(0),w(0)) + \operatorname{Re}\int_0^t [-\langle w_\tau(\tau), v_\tau(\tau)\rangle\\
&+\sigma_2(w_\tau(\tau),v(\tau)) + \langle f(\tau), w(\tau)-v(\tau)\rangle_{\mathcal{V}_2^*,\mathcal{V}_2}\\
&-\langle g(\mathcal{N}v(\tau)), \mathcal{N}w(\tau)-\mathcal{N}v(\tau)\rangle - \sigma_1(v(\tau), w(\tau)-v(\tau))]\,d\tau\\
&-\operatorname{Re}\langle w_t(t), w(t)-v(t)\rangle + \operatorname{Re}\langle w_t(0), w(0)-v(0)\rangle\\
&+a\left(\|w - v\|_{L^2([0,T],\mathcal{H})}\right) \geq 0
\end{aligned}$$

The inequality 4.131 requires some discussion. In the first term in 4.130, we can pass to the limit due to the strong convergence in e) of Lemma 1. In the third term we can pass to the limit because $w^N(0) \to w(0) = \varphi_0$ strongly in $\mathcal{V}$ and, therefore, also in $\mathcal{V}_2$. In all the terms under the integral we can pass to the limit due to the weak convergence a), b), c) and d) in Lemma 1. This limit can be justified by precisely the same arguments that were used to justify passing to the limit in (4.124). We only note that in the third term under the integral in (6.9) we use the weak convergence

$$w^N \to w \text{ in } L^2([0,T],\mathcal{V}_2),$$

which follows, from (4.108) or (4.109), and to pass to the limit in the fourth term we observe that (4.108) and (4.86) imply the weak convergence

$$\mathcal{N}w^N \to \mathcal{N}w \text{ in } L^2([0,T],\mathcal{H}).$$

In the next two terms outside the integral we can pass to the limit due to 4.111 and 4.114. Finally, in the last term we use the strong convergence $w^N \to w$ in $L^2([0,T],\mathcal{H})$, which follows from 4.114, and the fact that the function a is continuous. It remains to explain why we can pass to the limit in the second term of (4.130). Here we only have the *weak* convergence

$$w^N(t) \to w(t) \text{ in } \mathcal{V}_2 \qquad (4.132)$$

for any $t \in [0,T]$.

From (4.82), (4.83), (4.84), $\operatorname{Re}\sigma_2(\cdot,\cdot) + \lambda_0\|\cdot\|^2$ is topologically equivalent to the norm inner product on $\mathcal{V}_2$. Since norms are weakly lower semicontinuous in Hilbert spaces, when passing to the limit we have

$$\operatorname{Re}\sigma_2(w(t),w(t)) \le \underline{\lim}_{N\to\infty} = \operatorname{Re}\sigma_2(w^N(t),w^N(t)). \tag{4.133}$$

Taking into account the inequality in (4.130), we can thus obtain the desired inequality (4.131) when passing to the limit. Note that (4.131) is valid for any $v \in \mathcal{P} = \cup_{M=1}^{\infty}\mathcal{P}_M$ and, therefore, for any $v \in \mathcal{L}_T$.

Now we return to (4.125). Observe that in this relation we can set $\eta = w$, since $w \in \mathcal{L}_T$ and (4.125) is valid for $\eta \in \mathcal{L}_T$. Taking the real parts of both sides, we obtain after a straightforward computation

$$-\|w_t\|^2_{L^2([0,t],\mathcal{H})} + \frac{1}{2}\operatorname{Re}\sigma_2(w(t),w(t)) - \frac{1}{2}\operatorname{Re}\sigma_2(w(0),w(0)) + \tag{4.134}$$

$$+\int_0^t \left[\sigma_1(w(\tau),w(\tau)) + \operatorname{Re}\langle h(\tau),\mathcal{N}w(\tau)\rangle - \operatorname{Re}\langle f(\tau),w(\tau)\rangle_{\mathcal{V}_2^*,\mathcal{V}_2}\right] = d\tau +$$

$$+\operatorname{Re}\langle w_t(t),w(t)\rangle - \operatorname{Re}\langle w_t(0),w(0)\rangle = 0.$$

Let us now consider (4.125) with $\eta = -v$ where v is from 4.131. Taking the real parts of both sides we obtain

$$\operatorname{Re}\int_0^t \Big[\langle w_\tau(\tau),v_\tau(\tau)\rangle - \sigma_1(w(\tau),v(\tau)) - \sigma_2(w_\tau(\tau),v(\tau)) - \tag{4.135}$$

$$-\langle h(\tau),\mathcal{N}v(\tau)\rangle + \langle f(\tau),v(\tau)\rangle_{\mathcal{V}_2^*,\mathcal{V}_2}\Big]\,d\tau - \operatorname{Re}\langle w_t(t),v(t)\rangle +$$

$$+\operatorname{Re}\langle w_t(0),v(0)\rangle = 0$$

We next add the inequality (4.131) and the relations (4.134) and (4.135). After considerable cancellation we arrive at

$$\int_0^t \Big[\operatorname{Re}\langle h(\tau) - g(\mathcal{N}v(\tau)),\mathcal{N}w(\tau) - \mathcal{N}v(\tau)\rangle + \tag{4.136}$$

$$+\sigma_1(w(\tau)-v(\tau),w(\tau)-v(\tau))\Big] + a\left(\|w-v\|_{L^2([0,T],\mathcal{H})}\right) \ge 0.$$

Now take any $\theta > 0$ and let $\zeta \in \mathcal{L}_T$. Select

$$v(t) = w(t) - \theta\zeta(t). \tag{4.137}$$

Substituting (4.137) into (4.136) and dividing by $\theta > 0$, we obtain

$$\int_0^t \Big[\operatorname{Re}\langle h(\tau) - g(\mathcal{N}w(\tau) - \theta\mathcal{N}\zeta(\tau)),\mathcal{N}\zeta(\tau)\rangle + \tag{4.138}$$

$$+\theta\sigma_1(\zeta(\tau),\zeta(\tau))\Big]\,d\tau + \theta^{-1}a\Big(\theta\|\zeta\|_{L^2([0,t],\mathcal{H})}\Big) \ge 0,$$

for any $\zeta \in \mathcal{L}_T$, $\theta > 0$. In (4.138) we can pass to the limit $\theta \to 0$ and obtain

$$\operatorname{Re}\int_0^t \langle h(\tau) - g(\mathcal{N}w(\tau)),\mathcal{N}\zeta(\tau)\rangle\,d\tau \ge 0. \tag{4.139}$$

Here we have used the fact that $g : \mathcal{H} \to \mathcal{H}$ is a continuous mapping. We have also used the fact that

$$\lim_{\theta \to 0} \frac{a(\theta\rho)}{\theta} = 0, \text{ for any } \rho \geq 0,$$

which follows from condition ii) following 4.93: $a(0) = a'(0) = 0$.

The inequality 4.139 holds for all $\zeta \in \mathcal{L}_T$ only if it holds for equality. Indeed, suppose that for some ζ we have a strict inequality in 4.139, then replacing ζ by $-\zeta$ we obtain a contradiction to 4.139. Thus we have

$$\text{Re} \int_0^t \langle h(\tau) - g(\mathcal{N}w(\tau)), \mathcal{N}\zeta(\tau)\rangle \, d\tau = 0 \tag{4.140}$$

for all $\zeta \in \mathcal{L}_T$.

It remains to observe, that replacing ζ in 4.140 by $i\zeta$ we obtain that the imaginary part of the integral in 4.140 is also equal to zero. Lemma 2 is thus established and the proof of existence is complete.

We turn next to the uniqueness statement of Theorem 3. Let w and v be two solutions of (4.97) corresponding to the data φ_0, φ_1, f. Then $u \equiv w - v$ satisfies $u(0) = u_t(0) = 0$ and

$$\langle u_{tt}, \eta\rangle_{\mathcal{V}^*,\mathcal{V}} + \sigma_1(u, \eta) + \sigma_2(u_t, \eta) + \langle g(\mathcal{N}w) - g(\mathcal{N}v), \mathcal{N}\eta\rangle = 0 \tag{4.141}$$

for all $\eta \in \mathcal{L}_T$.

At this point we observe that (4.141), as well as (4.97) and (4.95), will still be satisfied if we extend the class of test functions η. Namely, (4.141) holds for $\eta \in \mathcal{M}_T$, where $\mathcal{M}_T$ denotes the space of functions $\eta : [0, T] \to \mathcal{H}$ such that

$$\eta \in C_W([0, T], \mathcal{V}_2) \cap L^\infty([0, T], \mathcal{V}),$$
$$\eta_t \in L^2([0, T], \mathcal{V}_2).$$

For fixed $s \in (0, T)$, let ψ be defined by

$$\psi(t) = \begin{cases} -\int_t^s u(\theta)d\theta & t < s \\ 0 & t \geq s \end{cases}$$

so that $\psi(T) = 0$, $\psi(s) = 0$ and $\psi(t) \in \mathcal{V}$ for each t. Indeed, $\psi \in \mathcal{M}_T$. (Note that $\psi \notin \mathcal{L}_T$, since $\psi_t \notin C_W([0, T], \mathcal{H})$.) The usual arguments reveal that

$$\int_0^s \{\langle u_{tt}(t), \psi(t)\rangle_{\mathcal{V}^*,\mathcal{V}} + \langle u_t(t), u(t)\rangle\} \, dt = \int_0^s \frac{d}{dt}\langle u_t(t), \psi(t)\rangle \, dt = 0.$$

Hence, choosing $\eta = \psi$ in (4.141) and integrating we obtain

$$\int_0^s \{\langle u_t(t), u(t)\rangle - \sigma_1(u(t), \psi(t)) - \sigma_2(u_t(t), \psi(t)) - \langle \Delta g(t), \mathcal{N}\psi(t)\rangle\} \, dt = 0$$

where $\Delta g(t) \equiv g(\mathcal{N}w(t)) - g(\mathcal{N}v(t))$. Since $\psi_t(t) = u(t)$, we can rewrite this as

$$\begin{aligned} &\int_0^s \frac{d}{dt}\{\|u(t)\|^2 - \sigma_1(\psi(t), \psi(t))\} \, dt \\ &\quad = 2Re \int_0^s \{\sigma_2(u_t(t), \psi(t)) + \langle \Delta g(t), \mathcal{N}\psi(t)\rangle\} \, dt. \end{aligned}$$

But since $u(0) = 0, \psi(s) = 0$, this implies

$$\|u(s)\|^2 + \sigma_1(\psi(0), \psi(0)) = 2Re \int_0^s \{\sigma_2(u_t(t), \psi(t)) + \langle \Delta g(t), \mathcal{N}\psi(t)\rangle\} \, dt.$$

However,

$$\begin{aligned}\int_0^s \sigma_2(u_t(t),\psi(t))dt &= \int_0^s \Big\{\frac{d}{dt}\sigma_2(u(t),\psi(t)) - \sigma_2(u(t),u(t))\Big\}dt \\ &= -\int_0^s \sigma_2(u(t),u(t))\,dt\end{aligned}$$

so that we obtain

$$\|u(s)\|^2 + \sigma_1(\psi(0),\psi(0)) + 2Re\int_0^s \sigma_2(u(t),u(t))dt = 2Re\int_0^s \langle \Delta g(t), \mathcal{N}\psi(t)\rangle\,dt. \tag{4.142}$$

Considering the last term in this equality, using A10) and A7a), we find

$$\begin{aligned}&|\int_0^s \langle \Delta g(t), \mathcal{N}\psi(t)\rangle\,dt| \\ &= |\int_0^s \left\langle \int_0^1 g'(\theta\mathcal{N}w(t) + (1-\theta)\mathcal{N}v(t))\,[\mathcal{N}w(t) - \mathcal{N}v(t)]\,d\theta, -\int_t^s \mathcal{N}u(\theta)d\theta \right\rangle dt| \\ &\le \int_0^s \{\tilde{C}_3\|\mathcal{N}w(t) - \mathcal{N}v(t)\| \int_t^s \|\mathcal{N}u(\theta)\|d\theta\}dt \\ &\le \int_0^s \tilde{C}_3\|\mathcal{N}u(t)\|dt \int_0^s \|\mathcal{N}u(\theta)\|d\theta \\ &= \tilde{C}_3(\int_0^s \|\mathcal{N}u(t)\|dt)^2 \le \tilde{C}_3\tilde{k}(\int_0^s \|u(t)\|_{\mathcal{V}_2}dt)^2 \\ &\le \tilde{C}_3\tilde{k}\,s\int_0^s \|u(t)\|^2_{\mathcal{V}_2}dt.\end{aligned}$$

Using this along with A5), A3) in (4.142) we obtain

$$\begin{aligned}&\|u(s)\|^2 + k_1\|\psi(0)\|^2_{\mathcal{V}} + 2k_2\int_0^s \|u(t)\|^2_{\mathcal{V}_2}dt \\ &\le 2\tilde{C}_3\tilde{k}\,s\int_0^s \|u(t)\|^2_{\mathcal{V}_2}dt + 2\lambda_0\int_0^s \|u(t)\|^2dt.\end{aligned}$$

This implies

$$\|u(s)\|^2 + (2k_2 - 2\tilde{C}_3\tilde{k}s)\int_0^s \|u(t)\|^2_{\mathcal{V}_2}dt \le 2\lambda_0\int_0^s \|u(t)\|^2dt.$$

Hence for $s < s_0 \equiv k_2/\tilde{C}_3\tilde{k}$ we have

$$\|u(s)\|^2 \le 2\lambda_0\int_0^s \|u(t)\|^2dt.$$

By Gronwall's lemma, we thus find $u(s) \equiv 0$ on $[0, s_0)$ where s_0 is independent of the solutions w, v. It follows that one must have $u \equiv 0$ on $[s_0, 2s_0)$, etc so that $u \equiv 0$ on any finite interval $[0, T]$.

4.3.6 Semigroup Formulation

In this section we show that the weak solution of our problem 4.76-4.78 satisfies a variation of parameters type integral equation. Let us first formally derive this equation. Eq. 4.76 can be formally rewritten as

$$\dot{z}(t) = \mathbf{A}z(t) + F(t) \tag{4.143}$$

where

$$z(t) = \begin{pmatrix} z_1(t) \\ z_2(t)\end{pmatrix} \equiv \begin{pmatrix} w(t) \\ w_t(t)\end{pmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0 & I \\ -\mathcal{A}_1 & -\mathcal{A}_2\end{bmatrix}, \tag{4.144}$$

and

$$F(t) = \begin{pmatrix} 0 \\ \Phi(t) \end{pmatrix}, \quad \Phi(t) = f(t) - \mathcal{N}^* g(\mathcal{N} z_1(t)). \tag{4.145}$$

The operator $\mathbf{A}$ generates a C_0-semigroup $S(t)$ on the space $\mathcal{Z} \equiv \mathcal{V} \times \mathcal{H}$, where we can, without loss of generality, use the equivalent σ_1 inner product on $\mathcal{V}$. Moreover, the operator $\mathbf{A}$ can be extended to an operator $\hat{\mathbf{A}} : \mathcal{Z} \to \mathcal{W}$ where $\mathcal{W} = \mathcal{Y}^*, \mathcal{Y} = [dom\mathbf{A}^*]$ with inner product $\langle \Phi, \Psi \rangle_{\mathcal{Y}} = \langle (\lambda - \mathbf{A}^*)\Phi, (\lambda - \mathbf{A}^*)\Psi \rangle_{\mathcal{Z}}$ where $\lambda > \lambda_0$ and $\hat{\mathbf{A}}$ is the infinitesimal generator of a C_0-semigroup $\hat{S}(t)$ on the space $\mathcal{W}$. This semigroup $\hat{S}(t)$ is an extension of $S(t)$ from $\mathcal{Z}$ to $\mathcal{W}$. Moreover, $\mathcal{Z}^* \subset \mathcal{W}$ with $\|\Psi\|_{\mathcal{W}} \leq C\|\Psi\|_{\mathcal{Z}^*}$ for $\Psi \in \mathcal{Z}^*$. This semigroup can be used to formally rewrite Eq. 4.143 in the form

$$z(t) = \hat{S}(t)z(0) + \int_0^t \hat{S}(t-\tau)F(\tau)\, d\tau. \tag{4.146}$$

Theorem 4 *In addition to the assumptions A1)-A9), A11) used to prove existence of a weak solution in Theorem 3, we also assume A7a) and A10). Then the weak solution w satisfies the integral equation 4.146.*

Proof: First of all we notice that the statement 4.113 of Lemma 1 can be strengthened if 4.91 is satisfied. Namely,

$$g(\mathcal{N}w^N) \to h, \text{ in } C_W([0,T],\mathcal{H}) \tag{4.147}$$

or, in other words,

$$g(\mathcal{N}w^N(t)) \to h(t), \text{ weakly in } \mathcal{H} \tag{4.148}$$

uniformly with respect to $t \in [0,T]$. This is certainly correct for a subsequence of $\{w^N\}$ (recall our convention at the beginning of Section 4.3.4).

By the Arzela-Ascoli theorem, to prove (4.148) it suffices to show that the set $\{g(\mathcal{N}w^N)\}_{N=1}^{\infty}$ is uniformly bounded and equicontinuous on $[0,T]$ in the $\mathcal{H}$-norm and recall that bounded sets in $\mathcal{H}$ are sequentially compact in the weak topology. We have

$$\begin{aligned} &\|g(\mathcal{N}w^N(t))\| \leq \widetilde{C}_1\|\mathcal{N}w^N(t)\| + \widetilde{C}_2 \\ &\leq \widetilde{C}_1\sqrt{k}\|w^N(t)\|_{\mathcal{V}} + \widetilde{C}_2 \leq \varepsilon^{-1/2}\sqrt{k}\,\widetilde{C}_1\widetilde{C}^{1/2} + \widetilde{C}_2 \end{aligned} \tag{4.149}$$

where we have used (4.90), (4.86) and the main *a priori* estimate (4.105). To show equicontinuity we check that the set $\left\{\frac{d}{dt}g(\mathcal{N}w^N)\right\}_{N=1}^{\infty}$ is bounded in $L^2([0,T],\mathcal{H})$. Here $\frac{d}{dt}$ means the strong $\mathcal{H}$-derivative. We have

$$\begin{aligned} &\|\frac{d}{dt}g(\mathcal{N}w^N(t))\|^2_{L^2([0,T],\mathcal{H})} = \\ &= \int_0^T \|g'(\mathcal{N}w^N(t))\mathcal{N}w_t^N(t)\|^2\, dt \leq \\ &\leq \widetilde{C}_3^2 \int_0^T \|\mathcal{N}w_t^N(t)\|^2\, dt \leq \\ &\leq \widetilde{k}\widetilde{C}_3^2 \int_0^T \|w_t^N(t)\|^2_{\mathcal{V}_2}\, dt \leq \widetilde{k}k_2^{-1}k\widetilde{C}_3\widetilde{C}, \end{aligned} \tag{4.150}$$

where we have used 4.87, 4.91, and 4.105. Note that in 4.150 we have also used the fact that $(\mathcal{N}w^N)_t(t) = \mathcal{N}w_t^N(t)$ for all $t \in [0,T]$, where on the left the derivative is understood in the sense of distributions with values in $\mathcal{H}$ and on the right with values in $\mathcal{V}_2$. Thus (4.148) is established.

Next we use the additional assumption in A7a) that $\mathcal{N}(\mathcal{V})$ is dense in $\mathcal{H}$. From this assumption and the statement 4.126 in Lemma 2 we can conclude that

$$g(\mathcal{N}w(t)) = h(t), \quad \text{for a.e. } t \in [0,T]. \tag{4.151}$$

Comparing 4.151 with 4.148 we conclude that

$$g(\mathcal{N}w^N) \rightharpoonup g(\mathcal{N}w) \text{ in } C_W([0,T],\mathcal{H}). \tag{4.152}$$

Thus we can choose our subsequences and limit function w so that we have

$$g(\mathcal{N}w(t)) \in \mathcal{H} \text{ for all } t \in [0,T]. \tag{4.153}$$

Recall that we have imposed the additional restriction (4.87) on $\mathcal{N}$. Hence we have

$$\mathcal{N}^* g(\mathcal{N}w(t)) \in \mathcal{V}_2^* \text{ for all } t \in [0,T] \tag{4.154}$$

and, moreover,

$$\mathcal{N}^* g(\mathcal{N}w) \in C_W([0,T],\mathcal{V}_2^*). \tag{4.155}$$

From 4.155 we conclude that, in particular,

$$\mathcal{N}^* g(\mathcal{N}w) \in L^2([0,T],\mathcal{V}_2^*). \tag{4.156}$$

From this last conclusion we can consider our original equation 4.76 as a linear equation with right side term

$$\Phi = f - \mathcal{N}^* g(\mathcal{N}w) \in L^2([0,T],\mathcal{V}_2^*). \tag{4.157}$$

Then the statement of the theorem follows from Theorem 4.3 in Banks et. al.

4.3.7 An Explicit Example

In this section we present an example of a system governed by a partial differential equation for which all the assumptions are satisfied. In particular, we consider an m-dimensional, nonlinear damped membrane with fixed boundary.

Let $\Omega \subset R^m$ be a bounded domain with C^1-smooth boundary Γ. We consider the problem

$$w_{tt} + \kappa_1 \Delta^2 w + \kappa_2 \Delta^2 w_t + \Delta g(\Delta w) = f \tag{4.158}$$

$$w|_{\Gamma=0} \tag{4.159}$$

$$\left.\frac{\partial w}{\partial n}\right|_{\Gamma} = 0 \tag{4.160}$$

$$w(x,0) = \varphi_0(x) \in H_0^2(\Omega), \quad w_t(x,0) = \varphi_1(x) \in L^2(\Omega) \tag{4.161}$$

$$x = (x_1, \cdots, x_m) \in \Omega, \ t \in [0,T], (x,t) \in \Omega \times [0,T] \equiv Q_T$$

We assume that $f(\cdot,t) \in H^{-2}(\Omega)$ for almost all $t \in [0,T]$ and

$$\int_0^T \|f(\cdot,t)\|^2_{H^{-2}(\Omega)}\, dt < \infty. \tag{4.162}$$

Assumption 1 *We assume that*

$$G(\xi) = \int_0^\xi g(\tau)\, d\tau, \quad g(\xi) = G'(\xi) \tag{4.163}$$

satisfies

1. *There exist positive constants C_j for $j = 1,2,3$ such that*

$$-\frac{1}{2}(\kappa_1 + \kappa_2 - \epsilon)|\xi|^2 - C_1 \le G(\xi) \le C_2|\xi|^2 + C_3 \tag{4.164}$$

for $\epsilon > 0$.

2. *There are positive constants* $\widetilde{C}_j$, $j = 1, 2$ *such that*

$$|g(\xi)| \leq \widetilde{C}_1|\xi| + \widetilde{C}_2. \tag{4.165}$$

3. *We also assume that*

$$g'(\xi) \geq -k_1. \tag{4.166}$$

Notice that in this problem

$$\mathcal{V} = \mathcal{V}_2 = H_0^2(\Omega) = \left\{\psi \in H^2(\Omega) : \psi|_\Gamma = \left.\frac{\partial \psi}{\partial n}\right|_\Gamma = 0\right\}$$

and

$$\mathcal{A}_1 = \mathcal{A}_2 = \Delta^2, \quad \mathcal{N} = \Delta, \quad k = \widetilde{k} = 1.$$

Let us check that 4.166 implies the monotonicity condition 4.94. We have for $\varphi, \psi \in L^2(\Omega)$

$$\begin{aligned}
&(g(\varphi) - g(\psi), \varphi - \psi) = \int_\Omega [g(\varphi) - g(\psi)]\,(\overline{\varphi(x)} - \overline{\psi(x)})\,dx \\
&= \int_\Omega \left[\int_0^1 ds \frac{d}{ds} g(s\varphi(x) + (1-s)\psi(x)\right]\,(\overline{\varphi(x)} - \overline{\psi(x)})\,dx \\
&= \int_\Omega \left[\int_0^1 ds\; g'(s\varphi(x) + (1-s)\psi(x))\right]\,|\varphi(x) - \psi(x)|^2\,dx \\
&\geq -k_1 \|\varphi - \psi\|^2_{L^2(\Omega)},
\end{aligned}$$

and the result follows. All other conditions (A1)-A11) are also satisfied.

In concluding this section, we note that the motivating example on nonlinear elastomers also falls within the class of examples that can be treated with the theory developed in this section. Of course, the neo-Hookean nonlinearity $\tilde{g}$ of (4.73) (which is only locally defined) must be appropriately extended to a map $\tilde{g} : R^1 \to R^1$. Once this is properly done, the functions $g : \mathcal{H} \to \mathcal{H}, \mathcal{H} = L^2(0, \ell)$, defined by $g(\varphi)(x) = \tilde{g}(\varphi(x))$ and $G(\varphi) = \int_0^1 \tilde{G}(\varphi(x))dx = Re\langle \tilde{G}(\varphi), 1\rangle$ where $\tilde{G}(\xi) = \int_0^\xi \tilde{g}(s)ds$, $\tilde{G}(\varphi)(x) = \tilde{G}(\varphi(x))$ will satisfy the necessary hypotheses for the theory of Sections 4.3.1-4.3.6.

4.3.8 Concluding Remarks

In the previous sections we have presented arguments of existence, uniqueness and regularity for solutions of abstract systems described by (4.76)-(4.78).

These arguments are constructive in the sense that they also establish convergence of certain classes of finite element Galerkin approximations that are the foundation of computational methods.

To be specific, suppose we have a family of approximation spaces

$$\mathcal{H}^N \equiv span\{\psi_1^N, \psi_2^N, \ldots, \psi_N^N\},\ N = 1, 2, \ldots,$$

where the basis elements $\{\psi_j^N\}$ satisfy the standard finite element condition:

(C1) For each N, $\mathcal{H}^N \subset \mathcal{V}$ and for each $\psi \in \mathcal{V}$, we have

$$\|\psi - P^N\psi\|_\mathcal{V} \to 0 \text{ as } N \to \infty$$

where $P^N : \mathcal{H} \to \mathcal{H}^N$ is the orthogonal projection of $\mathcal{H}$ onto $\mathcal{H}^N$.

We can then define the Galerkin approximations in the standard manner:
$w^N(t) = \sum_{k=1}^N c_k^N(t)\psi_k^N$ are chosen to satisfy (4.104) for each test function $\psi_j = \psi_j^N$, $j = 1, 2, \ldots, N$, with initial conditions

$$w^N(0) = P^N\varphi_0,\ w_t^N(0) = P^N\varphi_1.$$

We note that (C1) immediately yields that $w^N(0) \to \varphi_0$ in $\mathcal{V}$ and $w_t^N(0) \to \varphi_1$ in $\mathcal{H}$ for φ_0 and φ_1 in $\mathcal{V}$ and $\mathcal{H}$ respectively. Then under the additional condition

(C2) For all N, $\mathcal{H}^N \subset \mathcal{H}^{N+1}$,

we can prove that $w^N \to w$ in $C([0,T],\mathcal{H})$. The arguments follow almost immediately from those of Section 4.3.5 above. In both (4.117) and (4.122) we choose test functions $\Phi_M(t) = \eta(t) = \sum_{k=1}^{M} a_k^M(t)\psi_k^M$ with the a_k^M arbitrary $C^1([0,T])$ functions. We then have $\eta \in \mathcal{P}_M$ for every $M \leq N$ (we use the condition (C2) here only), so that (4.124) again holds for $\eta \in \mathcal{P}_M$, $M \leq N$. Then the remainder of the arguments of Section 4.3.5 remain unchanged and we thus conclude that beginning with *any* subsequence of the Galerkin sequence $\{w^N\}$, we can obtain a further subsequence which converges to w, the unique solution of (4.76)-(4.78). Hence the original Galerkin sequence itself must converge in $C([0,T],\mathcal{H})$ to w.

The condition (C1) is standard in finite element and spectral family approximation schemes. The condition (C2) is also readily satisfied in certain finite element and spectral approximations. For example, consider the one-dimensional elastomer rod with strong damping (so that $\mathcal{V}_2$ embeds compactly into $\mathcal{H} = L^2(0,\ell)$) and let ψ_j^K be the usual piecewise linear elements corresponding= to the discretizations of $0 < x < \ell$ with $\Delta x = \ell/K$. Define $V^K = span\{\psi_1^K, \ldots, \psi_K^K\}$ and then choose $\mathcal{H}^N = V^{2^N}$ in the Galerkin scheme described= above. It is readily seen that (C1) and (C2) hold where $\mathcal{V} = H_0^1(0,\ell)$.

4.4 An Experimentally Validated Damage Detection Theory in Smart Structures

This work discusses a theoretical, numerical and experimental investigation of the use of smart structures, parameterized partial differential equations and Galerkin approximation techniques to detect and locate damage. Smart structures, as used here refer to structures with embedded and/or surface mounted piezoceramic patches which may be used to sense and actuate vibrations of the host structure. Unlike many competing methods, the approach presented here is independent of modal information from the structure. Rather, changes in damping, mass and stiffness properties of the structure are estimated using time histories of the input and vibration response of the structure, generated and measured by the piezoceramic patches internal to the structure.

The premise of the effort proposed here is that damage to a structure will correspond in some way to changes, albeit small, in the structure's mass, damping and stiffness properties. Such damage might be due to fatigue, delaminations, cracks, or corrosion. However, in the study here all damage consists of holes in the structure which can be created in a systematic fashion allowing for a controlled study. Furthermore it is assumed that the structure of interest can be modeled by using a partial differential equation associated with basic structural elements (i.e., bars, beams, plates, membranes and shells).

Most of the previous efforts in the substantial literature on vibration related damage detection are based on modal methods. The basis for such methods is that damage produces a decrease in dynamic stiffness EI. This decrease in turn produces decreases in natural frequencies for an undamped simple beam (recall the eigenvalues are given by $\lambda \sim \sqrt{EI/\rho A}$ where A is the cross sectional area of the beam and ρ is the mass density). This basic premise has produced a number of results using modal analysis, i.e., frequency measurement to perform diagnostics. While modal based methods may have certain advantages (e.g., they are simple if they do work), modal based methods possess a number of major disadvantages. First of all, some of the modal based method investigations provide a strong argument for including geometry of the damage in any diagnostic testing scheme, something which is not easily done in frequency based methods. Indeed, mode and frequency characterizations are not so simple in variable structure systems; there is ample evidence that one should not use modal methods based on uniform undamped simple beams or plates as is often done in the engineering literature in addressing damage assessment methodologies. Since material parameters are most properly considered as spatially dependent quantities with damage manifested by changes in geometry (and hence in the spatial dependence of these parameters), it is unlikely that any rigorous theoretical basis for modal based methods for variable material structures will emerge. But perhaps the most serious objection to modal based methods resides in the fact that modal based methods have been shown to be highly unreliable for estimation of variable material parameters such as damping in composite material structures.

Adams et al. and Cawley and Adams provided insight into using vibration data for non-destructive evaluation of a structure's integrity. In 1978, Adams et al. analyzed one dimensional structures and showed that single point measurements coupled with a suitable model could be used to indicate both the location and magnitude of a defect. This claim is backed by experimental evidence. Their premise is that detailed models of damage mechanisms are of little value in detection, but rather noted that damage is usually accompanied by a local reduction in stiffness and increase in damping. They used receptance methods (defined as response displacement divided input magnitude at a harmonic steady state), and neglected damping, to model the structure. Local or distributed changes in stiffness produced changes in natural frequencies which effect each mode differently depending on the damage location. Later, Cawley and Adams extended these results to two and three dimensional structures via finite element models. They also performed a sensitivity analysis and showed that the ratio of frequency changes in different modes is only a function of damage location and not the magnitude of the damage.

Sato investigated the free vibration of beams with abrupt changes of cross section. Sato examines vibrating beams with grooves. While this work did not specifically address damage detection, it

does provide strong theoretical and experimental evidence that simple beam theory (i.e. a constant parameter beam equation) may lead to incorrect results as the ratio of the groove width to beam thickness increases. Sato's work uses a combination of beam theory modified by careful finite element analysis near the groove, held together by a transfer matrix method.

Cawley and Ray examined natural frequency changes in a beam due to cracks with changes caused by machined slots in an attempt to correlate theoretical damage detection results based on machined slots with the reality of actual cracks. Their results show conclusively that the width of the machined slots must be accounted for, adding evidence to the mounting case that geometry must be included in modeling and testing for damage. As in Cawley's earlier work, the results are theoretically supported by sound experiments.

More recently, Armon et al. modeled transverse cracks in beams as a simple reduction in beam stiffness. This in turn produces a drop in natural frequency which can then be measured and hence detected. As others before them, they used an undamped uniform beam and developed sensitivity formulas for frequency shifts as a function of small changes in global beam stiffness. They also assume that the mode shapes remain unchanged by the small change in stiffness. They develop a rank-ordering of fractional frequency shifts which are insensitive to the damage magnitude.

The results of the previous literature in damage detection provide evidence that something is gained by including the effects of geometry and hence modeling the local changes in modulus (stiffness). This then raises significant questions as to the validity of using traditional modal analysis (i.e., measurements of natural frequencies, assuming a uniform model) as the foundation of a damage detection methodology.

Can Modal Analysis Be Used To Detect Damage?

As one might expect from our comments in the previous section, the use of measured modal parameters to determine the existence of, extent of, and location of damage has been highly debated in the last few years. Conferences on modal analysis and on smart materials have produced scores of papers on the topic. Indeed the conference literature on using modal methods is extensive. In the journal literature, several manuscripts have demonstrated the feasibility of using measured changes in vibration characteristics to detect damage (in composite structures) by measuring vibration response both before and after a specific composite structure is damaged. As mentioned above these particular methods do not consider damping effects or mode shape changes.

Later, Stubbs, Stubbs and Osegueda and Saunders et al. presented a serious attempt to detect, locate and quantify damage in composite structures from measured modal data (i.e., damping ratios, natural frequencies and mode shapes). Stubbs' work is based on sensitivity equations for mass, damping and stiffness matrices as well as on the internal state variable constitutive theory of damage in composites. The approach makes heavy use of a finite element model of the structure which must produce natural frequencies in strong agreement with the measured natural frequencies of the undamaged structure. The work of Stubbs et al. focuses on changes in stiffness and damping (proportional) only but applied to both isotropic and anisotropic material. The experimental component of this work examines simple cantilevered beam specimens. Stubbs' sensitivity formula depends on knowing the mode shape transformation. In particular, modes of both the damaged (measured) and undamaged (theoretical) system are required. Recently, Tomlinson showed that the mode shapes of a damaged system and that of an undamaged system are nearly the same. In fact, Stubbs shows that they are exactly the same if the damage is uniform. Unfortunately, Stubbs was unable to measure mode shapes during the experimental verification phase of their work even though the damage was clearly nonuniform and hence this issue remains unresolved.

Experimentally, Stubbs measured frequency shifts of the order of 1 to 2% for structure damage by loading them from 60% up to 75% of their ultimate strength. Thus it seems that the issue of using modal parameters to determine the location of a quantitative value of damage is also not settled by these results.

One of the major concerns regarding using modal analysis to detect damage is that damage is a local phenomena and modal information is a reflection of the global system properties. To

further investigate this criticism, a simple single degree of freedom system is examined followed by an examination of a beam with local mass changes. Damping is not considered in this example. However, note that changes in global damping ratios of a material can be related to significant local damage by simply examining a bolted connection. For instance two struts held together with a joint exhibit damping ratios which are dominated by the torque or tightness of the joint. In space frame structures the damping is thought to come largely from joints and connections.

The frequency of a single spring mass model of a structure is related to its physical stiffness k and mass m by $\omega^2 = k/m$. Using a simple derivative we see that the change in frequency with respect to stiffness is

$$\frac{d\omega}{dk} = \frac{1}{2\sqrt{km}} = \frac{\omega}{2k},$$

and the change in frequency with respect to mass is

$$\frac{d\omega}{dm} = -\frac{1}{2}\sqrt{\frac{k}{m^3}} = -\frac{\omega}{2m}.$$

These derivatives represent the sensitivity of the natural frequency to changes in stiffness and mass respectively. In each case, the sensitivity of ω decreases inversely as the value of the stiffness (or mass) increases. For a two meter long airplane wing, the stiffness is about 116 N/m so that changes in frequency near 100 Hz corresponding to 1% changes in stiffness would be expected to be of the order of .4 Hz, or a fraction of a Hertz, while similar changes measured at 1000 Hz would be of the order of 4 Hz, etc. Measurements of frequencies with modern analyzers often have difficulty detecting a 0.4 Hz change. It is also known that small changes in unmodeled parameters such as humidity and temperature can cause such small changes in frequency measurement taken on different days.

Next consider the possibility of detecting a mass change by using frequency information. Again using the single degree of freedom frequency equation, a one percent change in mass of a 10 Hz system will produce a 0.05 Hz change in frequency or a 0.5 Hz change at 100 Hz. Thus, for single degree of freedom system 1% change in mass requires that the measurement scheme be able to discriminate frequencies to within 0.5%. This conclusion does not address simultaneous changes in the system mass and stiffness.

Next consider damage as characterized by a finite element eigenvalue problem. Here damage is simulated by local changes in either mass or stiffness matrix by perturbing just one element in the model. For simple structural models such as a beam, the equations of motion can be used directly without use of a finite element model. The results reported in Weissenburger, examines analytical solutions of the eigenvalue problem for local changes in stiffness and mass at various points on a beam. While Weissenburger's interest was not in diagnostics, his results make a significant and systematic comment on using frequency shifts to detect changes in physical properties. For a mass placed about $\frac{1}{3}$ out from one end along the length of a pinned-pinned beam, the nondimensional frequency shifts (denoted β_n^2 where $\omega_n = \beta_n^2\sqrt{EI/\rho A}$) are given by Table 4.3. The last two

	undamaged beam	10%	50%
β_1^2/π^2	1	0.9397	0.7715
β_2^2/π^2	4	3.7130	3.2218
β_3^2/π^2	9	8.9339	8.8292

Table 4.3: Nondimensional frequency shifts .

columns are the ratio of "added mass" to total beam mass. This result illustrates that small local changes in the mass of a beam can make significant changes in the frequency. For instance, the first frequency changes by 6% for a modification in mass of 10% and 23% for a modification of 50%.

While these numbers are more encouraging than the changes predicted by looking at the single degree of freedom model, a 10% change in mass represents relatively large damage. Note, as in the single degree of freedom case the percent change in frequency is about $\frac{1}{2}$ of the change in a physical parameter. The conclusion here is based entirely on an analytical solution and does not depend on a choice of grids and nodes as does the FEM approach. The same study shows clearly that the mode shapes do not change much as local mass is modified.

Next Weissenburger considered the same analytical technique applied to changes in local stiffness of a pinned-pinned beam. As Table 4.4 reveals, the results found for using frequency measurements to determine changes are even more interesting. The 3rd mode remains unaffected because the local

	undamaged beam	10%	1.45%
β_1^2/π^2	1	2.2995	2.7887
β_2^2/π^2	4	5.0996	9
β_3^2/π^2	9	9	9

Table 4.4: Frequency changes as a function of local stiffness modification .

stiffness change is made at the third modes' first node. The 1.45% damage case is special because it causes the characteristic equation to have a double root. This case illustrates that very small changes in local stiffness can have extremely large effects on frequency if the changes in stiffness occur at certain primary locations on the structure. While these results are obtained by adding stiffness and mass, the same effects should be obtained by subtracting stiffness and mass at local points. Weissenburger's study combined with Sato's result indicates a possible source of why the literature contains conflicting statements on the use of frequency measurements to predict damage. An explanation for the controversy may be as simple as it depends on where the damage is located i.e., it depends on the geometry of the damage. For some types of damage modal analysis may be appropriate while for other configurations it may not be appropriate. This point is also made by Adams et al.

The results we propose here specifically avoid using modal analysis methods because the results based on modal methods seem controversial and their success depends on the unknown relationship between damage location and measurement location. Rather we use direct time domain estimation of spatially varying physical parameters of a structure to determine the structures' health.

In light of the above comments, a question of rather great interest then is: can one develop analytically sound, non-modal based self-excitation/self-sensing methods for detection and characterization (geometrical and quantitative) of damage in smart material structures? Here we address this question in the context of embedded piezoceramic structures.

For an embedded piezoceramics smart material damage detection and characterization methodology, there are several distinct requirements. These include:

(a) One must be able to *estimate reliably* (repeatable across experiments) the *variable structure material parameters* of a piezoceramic loaded structure. This must be done using piezo actuation and sensing with accuracy comparable to that achievable with standard actuating (impulse hammers, solenoidal actuators) and sensing (accelerometers, strain gauges, laser vibrometers) devices in non-smart material testing schemes.

(b) One must be able to use the actuation and sensing properties of the piezoceramics to *excite the structure and analyze the response* (in a single experiment) for a reliable methodology that is the basis of self-excitation/self-sensing.

(c) One must be able to *detect and characterize damage via vibration self-excitation/ self-sensing* that relies only on the input/output signals for the piezoceramics.

The first two of these requirements have been studied analytically, numerically, and experimentally by Banks et al. In this section we address the last requirement in the context of a piezoceramic loaded beam. This particular structure is sufficiently representative to make a compelling case for feasibility of the ideas we propose. As noted above, it is essential to model the micro structure related to the local geometry and elastic characteristics. In Section 4.4.1 below, the partial differential equations describing the dynamics of a beam with surface bonded piezoceramic patches is outlined. The damage detection problem is formulated as an optimization problem and is described in Section 4.4.2. The experimental verification with regard to requirement (c) is detailed in Section 4.4.3.

4.4.1 Model for Damaged Structures

It is clear that having a model including the geometry and material parameters of the structures is important. We devote this section to the modeling of a structure of interest to us.

The test structure is a cantilever beam with two piezoceramic patches attached on the opposite side of the beam. It should be noted that the underlying mathematical theory is neither restricted to two patches configuration nor, as shall become clear later, restricted to beams. This homogeneous material beam, which we shall assume satisfies the Euler-Bernoulli hypothesis for displacements and the Kelvin-Voigt hypothesis (damping proportional to strain rate), is fixed at $x = 0$, free at $x = \ell$. Two piezoceramic patches are bonded to the beam (one on each side) in the region $x_1 < x < x_2$. We denote the length by ℓ, width w, thickness t, mass density ρ, Young's modulus E and damping coefficient c_D, and as usual, A represents the cross sectional area and I is the moment of inertia of the cross sectional area. The different materials will be indicated by subscripts: b for beam and p for piezoceramic. For simplicity here, we ignore the bonding layer material properties and geometry although this is readily included if so desired.

For such a beam, subject only to forces and moments generated by actuating the patches, force and moment balancing lead to the dynamic system of equations

$$\begin{aligned} \rho(x)\frac{\partial^2 u}{\partial t^2} - \frac{\partial N_x}{\partial x} &= -S_{1,2}\frac{\partial [N_x]_p}{\partial x} \\ \rho(x)\frac{\partial^2 y}{\partial t^2} + \frac{\partial^2 M_x}{\partial x^2} &= -\frac{\partial^2 [M_x]_p}{\partial x^2}, \qquad 0 < x < \ell \end{aligned} \tag{4.167}$$

for the axial displacement $u = u(t,x)$ and the transverse displacement $y = y(t,x)$. Here N_x and M_x are the internal force and moment resultants given by

$$\begin{aligned} N_x &= EA(x)\frac{\partial u}{\partial x} + c_D A(x)\frac{\partial^2 u}{\partial x \partial t} \\ M_x &= EI(x)\frac{\partial^2 y}{\partial x^2} + c_D I(x)\frac{\partial^3 y}{\partial x^2 \partial t}, \end{aligned} \tag{4.168}$$

where

$$EA(x) = E_b t_b w_b + 2E_p t_p w_p \chi_p(x), \quad c_D A(x) = c_{Db} t_b w_b + 2c_{Dp} t_p w_p \chi_p(x), \tag{4.169}$$

$$EI(x) = \frac{1}{12}t_b^3 w_b E_b + \frac{2}{3}a w_p E_p \chi_p(x), \quad c_D I(x) = \frac{1}{12}t_b^3 w_b c_{Db} + \frac{2}{3}a w_p c_{Dp} \chi_p(x), \tag{4.170}$$

with $\chi_p(x)$ the characteristic function that is 1 for $x_1 \le x \le x_2$, and zero elsewhere and patch constant $a = (\frac{1}{2}t_b + t_p)^3 - (\frac{1}{2}t_b)^3$. The linear mass density ρ is given by

$$\rho(x) = \rho_b t_b w_b + 2\rho_p t_p w_p \chi_p(x). \tag{4.171}$$

The indicator function $S_{1,2}$ has form

$$S_{1,2} = \begin{cases} 1 & x < (x_1 + x_2)/2 \\ 0 & x = (x_1 + x_2)/2 \\ -1 & x > (x_1 + x_2)/2 \end{cases}.$$

The external forces and moments $[N_x]_p$ and $[M_x]_p$ depend on the voltages supplied to each of the two patches. If the voltages are denoted by v_1 and v_2, respectively, these forces and moments are given by

$$\begin{aligned}[N_x]_p &= \mathcal{K}_A S_{1,2}(x)\chi_p(x)[v_1(t)+v_2(t)] \\ [M_x]_p &= -\mathcal{K}_B\chi_p(x)[v_1(t)-v_2(t)],\end{aligned} \tag{4.172}$$

where $\mathcal{K}_A$, $\mathcal{K}_B$ are constants depending on the piezoceramic material properties. If the patches are excited in-phase, for example, with $v_1(t) = v_2(t) = v(t)$, we find

$$[N_x]_p = 2\mathcal{K}_A S_{1,2}(x)\chi_p(x)v(t)$$

$$[M_x]_p = 0,$$

resulting in axial motion only. If the patches are excited out-of-phase, with $v_1(t) = -v_2(t) = v(t)$, we have pure bending or transverse motion since

$$[N_x]_p = 0,$$

$$[M_x]_p = -2\mathcal{K}_B\chi_p(x)v(t).$$

Next we consider the beam with damages. Since the beam model is based on the Euler-Bernoulli theory, we assume that the center of the damage coincides with the neutral axis (the axis parallel to the longest edges) of the beam and the damage is symmetric with respect to this center line. Furthermore, we assume that the damage located between x_{d_1} and x_{d_2} is characterized by shape functions. The shape functions represent a change in the geometry of the beam, resulting in the thickness and the width of the beam no longer being uniform. With damage, the coefficients in the equations 4.167 and 4.168 become

$$\begin{aligned}EA(x) &= E_b t_b w_b + 2E_p t_p w_p \chi_p(x) - E_b S_A(x)\chi_d(x) \\ c_D A(x) &= c_{Db} t_b w_b + 2c_{Dp} t_p w_p \chi_p(x) - c_{Db} S_A(x)\chi_d(x) \\ EI(x) &= \frac{1}{12}t_b^3 w_b E_b + \frac{2}{3}a w_p E_p \chi_p(x) - E_b S_I(x)\chi_d(x) \\ c_D I(x) &= \frac{1}{12}t_b^3 w_b c_{Db} + \frac{2}{3}a w_p c_{Dp}\chi_p(x) - c_{Db} S_I(x)\chi_d(x) \\ \rho(x) &= \rho_b t_b w_b + 2\rho_p t_p w_p \chi_p(x) - \rho_b S_A(x)\chi_d(x)\end{aligned} \tag{4.173}$$

where the characteristic function $\chi_d(x)$ is 1 for $x_{d_1} \le x \le x_{d_2}$, and zero elsewhere. The shape function $S_A(x)$ is the missing area of the cross section due to the damage, while the function $\frac{1}{12}t_b^3 w_b - S_I(x)$ is the inertia of the cross section of the beam containing damages.

Coupled to the system 4.167–4.168 and 4.172–4.173 are appropriate boundary conditions

$$u(t,0) = 0, \qquad \frac{\partial u}{\partial x}(t,\ell) = 0 \tag{4.174}$$

$$y(t,0) = \frac{\partial y}{\partial x}(t,0) = 0, \qquad M_x(t,\ell) = \frac{\partial}{\partial x}M_x(t,\ell) = 0, \tag{4.175}$$

and initial conditions

$$u(0,x) = \eta(x), \qquad \frac{\partial u}{\partial t}(0,x) = \gamma(x), \tag{4.176}$$

$$y(0,x) = \zeta(x), \qquad \frac{\partial y}{\partial t}(0,x) = \psi(x). \tag{4.177}$$

For a beam containing damages and a pair of identical patches which are bonded symmetrically about the middle surface, the differential equations 4.167–4.168 and 4.172–4.173, under the first order Euler-Bernoulli assumptions, describe vibrations in the axial and transverse directions that are uncoupled. If one has only a single patch bonded to the beam, or if the patches are not identical, then one obtains a set of equations for longitudinal and transverse vibrations that are coupled (which is not surprising since the structure consisting of beam plus patch is no longer symmetric).

Without loss of generality, let us consider the beam described above under out-of-phase excitation, resulting in pure bending (transverse vibrations). Using the abbreviated notation $D = \frac{\partial}{\partial x}$, we then have the model

$$\rho y_{tt} + D^2(EID^2y + c_D I D^2 y_t) + \gamma y_t = -2\mathcal{K}_B D^2 \chi_p(x) v(t) \tag{4.178}$$

coupled with the boundary and initial conditions of 4.175, 4.177. In addition to Kelvin-Voigt damping, the equation 4.178 also incorporates some viscous (air) damping $\gamma\, y_t$. The coefficients ρ, EI and $c_D I$ are discontinuous (see 4.173) while the coefficient involving the input voltages is given by

$$D^2\chi_p(x) = D^2\{H(x-x_1) - H(x-x_2)\} = D\delta(x-x_1) - D\delta(x-x_2) \tag{4.179}$$

where H is the Heaviside function and $D\delta$ is the "derivative" of the Dirac function with mass at $x = 0$. This strong form of the equation 4.178 involves irregularities which can (and have) led to computational difficulties for estimation and control efforts found in the literature. Retention of such irregular terms as the discontinuous coefficients in 4.178 and the impulse derivatives in 4.179 is of great importance (indeed, essential) when using such models with experimental data from actual structures; this has been shown in Banks et al.

For the same configuration, when the beam is under deformation (bending), the generated charges in terms of voltage across the piezoelectric sensors has the expression.

$$\mathcal{K}_s \int_{x_1}^{x_2} \frac{\partial^2}{\partial x^2} yx(t,x)dx = \mathcal{K}_s\Big(\frac{\partial}{\partial x} yx(t,x_2) - \frac{\partial}{\partial x} yx(t,x_1)\Big) \tag{4.180}$$

where $\mathcal{K}_s$ is a sensor constant which is also a constant dependent on material piezoelectric properties and geometry.

Instead of seeking solutions to the partial differential equations as formulated in strong form 4.178-4.179 coupled with the boundary conditions 4.175 and 4.177, we seek the solutions to variational formulation

$$\begin{aligned} \int_0^\ell \{\rho y_{tt}\phi + (EID^2y + c_D I D^2 y_t) D^2\phi + \gamma y_t \phi\}\, dx \\ = \left(\int_{x_1}^{x_2} -2\mathcal{K}_B D^2\phi dx\right) v(t) \end{aligned} \tag{4.181}$$

for some ϕ in an appropriate class of "test" functions. The well posedness and computational techniques will be based on this variational form of the damped second order system.

Applying standard functional analysis techniques, one could establish existence of a unique solution with $y(t,\cdot) \in H_L^2(0,\ell) \equiv \{\psi \in L_2(0,\ell) | \psi, \psi', \psi'' \in L_2(0,\ell), \psi(0) = \psi'(0) = 0\}$ satisfying 4.181 for all test functions ϕ in the Hilbert space $H_L^2(0,\ell)$. In this sense the initial boundary value problem 4.177,4.178 with the boundary conditions 4.175 is well-posed under very mild smoothness assumptions on $EI(x) > 0$, $c_D I(x) > 0$ and $\rho(x) > 0$. The well-posedness result can be established for other types of internal damping such as spatial hysteresis damping which represents one form of energy dissipation mechanism in certain composite materials, and structural (square-root) damping which has been suggested in others. Detailed statements and the nontrivial arguments underlying these mathematical results can be found in Banks et al.

4.4.2 Damage Detection Technique

Damage detection is carried out by determining the shape functions S_A and S_I in 4.181 using observations of the system output response to excitations $\{v_i\}$ to the patches. This estimation

problem is formulated as an enhanced least squares fit to observations, i.e., we seek $\bar{q} \in Q$ which minimizes

$$J(q) = \sum_{i=1}^{M_v} c_i \hat{J}(q; v_i) \tag{4.182}$$

where $\hat{J}(q; v_i)$ is in form of

$$\hat{J}(q; v_i) = \sum_{k=1}^{\bar{N}} \left| K_s \left(\frac{\partial}{\partial x} yx(t_k, x_2; q, v_i) - \frac{\partial}{\partial x} yx(t_k, x_1; q, v_i) \right) - z_k) \right|^2, \tag{4.183}$$

for the piezoelectric elements being located on the beam between x_1 and x_2. In the cost function $\{z_k\}$ are measured voltages across the piezoelectric elements due to bending, and $\{y(t_k, \cdot; q, v_i)\}$ are the parameter dependent weak solutions of 4.181 with zero initial conditions evaluated at each time t_k, $k = 1, 2, \ldots, \bar{N}$ for a given input $v_i(t)$. In 4.183, $|\cdot|$ is an appropriately chosen Euclidian norm. The set Q is some admissible parameter set which arises in parameterization of the desired shape functions.

In 4.182, the coefficients $\{c_i\}$ are chosen so that amplitude of the weighted response $\{c_i\, y(t, x; q, v_i)\}_{i=1}^{M_v}$ are in the same order. Furthermore, the system is excited in a manner so that the dominant mode in the response data corresponding to the input $v_i(t)$ is different from one excited by $v_j(t)$ for $j \neq i$.

The standard least squares cost function 4.183 alone is not adequate for unique solution of the damage detection problem. We illustrate this point with the following example in which the simulated piezoceramic responses corresponding to different damage locations are compared. One response data is generated on a beam with a centered circular hole at 2.413 cm away from the clamped end. The second is generated with the hole at 14.478 cm. In both cases, identical broad band input signals are applied. Studying the responses at different damage locations, it is obvious that vibration data and location of damages do not have a one to one relationship. Hence there may exist more than one set of damage parameters which yield the same value of cost function 4.183.

If one examines closely the very beginning of the response data, the second mode responses do not match even though the first mode responses match quite well. This difference could be observed more clearly if only the second mode were excited. When a bandwidth signal concentrated on the second mode is applied, the responses are quite different as the data clearly demonstrates.

This nonuniqueness was observed implicitly by Armon et al. The rank-ordering method in the article is based on the fact that the amount of phase shift in each mode differs from one type of damage (characterized by location and dimensions) to another, and the amount of shifts yield different order in response to different damages.

The above remarks raise the question as to whether it is possible to use one response data in which multiple modes are excited. Since the amount of change in the response due to the damage is very small, recorded data in a short time period would not provide enough information in the sense of displaying the difference caused by damages. On the other hand, the higher vibrational modes are damped out in the response over a longer time period. The example above demonstrated what could happen in damage detection by using only one vibration response data. This motivates the use of the cost functional 4.182. The basic idea underlying this enhanced cost function is to use time responses over a long period while at the same time taking into account the information in several modes so that the cost function is sensitive to changes in any mode.

In the enhanced cost function, a time domain cost function is adopted. As was pointed out in the Introduction, frequency data (such as embodied in transfer functions) does not provide adequate information to detect damage. In two experimental data sets, one recorded from an undamaged beam and one from a damaged beam, with limited resolution of data acquisition equipment, the frequency response does not provide damage knowledge, while the time response data indicates the change in the structure. A detailed description of the experiment is presented in Section 4.4.3.

The minimization in the above setting involves an infinite dimensional state space and without further parameterization of the shape functions, an infinite dimensional (functions) admissible

parameter space. We consider Galerkin type approximation in the context of the variational formulation of Section 4.4.1. The beam displacement $y(t,x)$ is approximated by

$$y^N(t,x) = \sum_{i=1}^{N} w_i(t)\,\phi_i^N(x) \tag{4.184}$$

in an appropriate Nth-order finite dimensional space. The basis elements $\{\phi_i^N\}$ are chosen to be in the same space as those test functions in 4.181 with modifications such that the boundary conditions are satisfied. The generalized Fourier coefficients $\{w_i(t)\}$ represent the state relative to this basis.

Replacing $y(t,x)$ and ϕ in the variational formulation 4.181 by 4.184 and $\{\phi_i^N\}_{i=1}^N$ respectively, 4.181 yields an N vector ordinary differential equation system for the N vectors $w(t) = [w_1(t), w_2(t), \cdots, w_N(t)]^T$:

$$M^N\,\ddot{w}(t) + C^N\,\dot{w}(t) + K^N\,w(t) = F^N\,v(t) \tag{4.185}$$

The usual Galerkin form coefficient matrices are derived by integrating each term in 4.181:

$$[M^N]_{i,j} = \int_0^\ell \rho(x)\,\phi_i^N(x)\,\phi_j^N(x)\,dx$$

$$[C^N]_{i,j} = \int_0^\ell c_D I(x)\,\frac{d^2\phi_i^N(x)}{dx^2}\,\frac{d^2\phi_j^N(x)}{dx^2}\,dx + \int_0^\ell \gamma\,\phi_i^N(x)\,\phi_j^N(x)\,dx$$

$$[K^N]_{i,j} = \int_0^\ell EI(x)\,\frac{d^2\phi_i^N(x)}{dx^2}\,\frac{d^2\phi_j^N(x)}{dx^2}\,dx$$

$$[F^N]_i = -2\mathcal{K}_B \int_0^\ell \frac{d^2\phi_i^N(x)}{dx^2}\,dx.$$

The dimension and location of the damage are unknown and must be determined. The functions $S_A(x)$ and $S_I(x)$ which characterize the damage are elements to be chosen from an infinite class of functions. Rather than attempting to reconstruct $S_A(x)$ and $S_I(x)$, we search for the projections of $S_A(x)$ and $S_I(x)$ on the linear span of finite dimensional sets $\{\Phi_i\}_{i=1}^{M_A}$ and $\{\Psi_j\}_{j=1}^{M_I}$ respectively. Thus we parameterize S_A and S_I by

$$S_A(x) = \sum_{i=1}^{M_A} \alpha_i\,\Phi_i(x) \qquad 0 < x_{d_1} \le x \le x_{d_2} < \ell \tag{4.186}$$

$$S_I(x) = \sum_{j=1}^{M_I} \beta_j\,\Psi_j(x) \qquad 0 < x_{d_1} \le x \le x_{d_2} < \ell. \tag{4.187}$$

Then a family of approximating estimation problems with finite dimensional state spaces and parameters can be formulated by seeking a vector parameter

$$q^M = (\alpha_1, \alpha_2, \cdots, \alpha_{M_A}, \beta_1, \beta_2, \cdots, \beta_{M_I}, x_{d_1}, x_{d_2})$$

which minimizes

$$J^N(q^M) = \sum_{i=1}^{M_v} c_i\,\hat{J^N}(q^M; v_i) \tag{4.188}$$

where $\hat{J^N}(q^M; v_i)$ is given by 4.183 with $y(t,x)$ replaced by $y^N(t,x)$, and the dimension of the parameter space is $M = M_A + M_I + 2$. The coefficients $\{w_i(t)\}$ in 4.184 are the solutions to 4.185.

Based on our previous computational experiences in similar estimation problems, we chose cubic splines for the basis of our approximation scheme. The finite state space is defined as the span of $\{b_i\}_{i=1}^{\tilde{N}+1}$ with $\{b_i\}$ being a standard cubic B-spline basis set with the elements modified to satisfy

the essential (or geometric) boundary conditions $b_i(0) = Db_i(0) = 0$. That is, if $\{\hat{b}_i\}_{i=-1}^{\tilde{N}+1}$ are the standard cubic B-splines, then $b_1 = \hat{b}_0 - 2\hat{b}_1 - 2\hat{b}_{-1}$ and $b_i = \hat{b}_i$ for $i = 2, \cdots, N(= \tilde{N} + 1)$. For the parameter space linear splines are selected as the basis elements Φ, Ψ_j in 4.186,4.187.

Solving the approximation problems, we obtain a sequence of estimates $\{\bar{q}^{N,M}\}$. The sequence admits a convergent subsequence which converges to some $\bar{q} \in Q$ under the assumptions that Q is a compact set and parameter functions $S_A(x)$ and $S_I(x)$ are bounded above. Furthermore the limit parameter $\bar{q}$ is a solution to the original infinite dimensional optimization problem. The relevant parameter estimate convergence and continuous dependence with respect to the observations results can be found in Banks el at.

4.4.3 Experimental and Numerical Results

To demonstrate the capability of damage detection with the algorithm outlined in the previous section, six different experiments were carried out. Each of them has a distinct damage location and size which allows us to conduct a sensitivity study. The detailed experimental procedure is described in the following section and damage detection results are reported afterwards.

Test specimens and procedures

The test articles for our non-destructive damage detection studies are cantilever aluminum beams which are 48.26 cm long, 2.032 cm wide and 0.15875 cm thick. One pair of piezoceramics is surface bonded opposite one other and they are electrically coupled to create one sensor/actuator. The piezoceramics have the following geometric dimension: the length is 6.35 cm, the width is 2.032 cm and the thickness is 0.0254 cm. An accelerometer is also affixed to each beam near the midspan (this can be used to corroborate our finding using the accumulated strain data e.g. 4.180.

We focused on vibration tests in which a maximum of the first three vibrational modes would be excited simultaneously. Since the Euler-Bernoulli theory with Kelvin-Voigt damping was employed, our beam model is inadequate for use in analyzing vibration data containing high frequencies. For the chosen beam dimension and material, the first three natural modes are under 100 Hz and the Euler-Bernoulli model is appropriate.

The piezoceramic pair located at 5.08 cm from the clamped end is used as both a sensor and actuator for the damage detection experiments. A switching mechanism allows the ceramic to be used in this manner. The switch is engaged by applying 12 VDC during the excitation period and it is disengaged afterward, thus allowing the piezoceramic to be used as a sensor. When used as an actuator, the ceramic is connected in series to an amplifier ($\times 10$) with a voltage range of ± 50 V. When used as a sensor, the piezoceramic is connected in series to a high impedance low-pass filter. The device has a 10 MΩ impedance and a corner frequency of 10,000 Hz, i.e. any frequencies higher than 10,000 Hz are cut off to avoid aliasing. These parameters are chosen such that the output voltage of the ceramic is proportional to strain over the frequency range of interest (between 1 Hz and 100 Hz).

If small frequency changes are to be measured reliably, it is essential to have the same experimental conditions such as input and boundary clamp for both the undamaged and damaged structures. In the vibration tests, an impulse hammer is commonly used to excite multiple modes. The disadvantage in using an impulse hammer is the associated poor repeatability of input. Moreover, it is difficult to simulate the resulting input force accurately. For the same reason, a random force input is also not suitable. To have repeatable input with frequency content over a desired bandwidth, a Schroeder-phased signal is adopted. The excitation voltage to the piezoceramic patches is generated by a DSpace computer where MATLAB Simulab Toolbox is utilized for implementing the Schroeder-phased signal. Each signal contains 1024 points and excites the beam for 4 seconds. The frequency content of each signal is chosen to excite certain modes of the beam. Signal 1 is broadband; signals 2 and 3 are narrowband (see Table 4.5). Schroeder-phased inputs exhibit a flat power spectrum over the specified frequency range, i.e. the power in the input signal is equally distributed over the range. This characteristic permits very repeatable input time histories for all of the cases studied.

	Frequency Range (Hz)		
	0-100	0-20	20-50
Signal 1	x	-	-
Signal 2	-	x	-
Signal 3	-	-	x

Table 4.5: Frequency content of the Schroeder-phased inputs.

The data acquisition system is a Fourier analyzer, Tektronix Analyzer 2600, with a PC (IBM AT) connected to it to display and record the data. The sampling rate of the analyzer is set at 512 Hz, and a total of 4096 data points (8 seconds) are recorded for each vibration response. The analyzer is set to start recording at time 3.6 seconds. Hence 0.4 seconds of the forced vibration is recorded since the excitation lasts for 4 seconds. The rest of the data contains free decay vibration information.

Two separate series of experiments are performed. The location of the damage differs for each series. An accelerometer was located at 25.4 cm in both experiments for monitoring the vibration and double checking the data collected from piezoceramic patches. To conduct the studies on the sensitivity of the method to the distances between the damage, the clamped end, and the sensor location, one defect is introduced near the clamped end and one near midspan. Table 4.6 lists the damage type and location for each experiment. All damage geometries are symmetric to the neutral axis of the beams.

	Damage Cases			
Exp. #	a	b	c	d
1	No damage	DT1-1	DT2-1	DT3-1
2	No damage	DT1-9	DT2-9	DT3-9

DT#-* refers to damage type (#) and location(*(inch)).

Table 4.6: Damage types and locations for the two experiments.

The procedure for each of the two series of experiments is identical. First, the undamaged beam is tested by exciting it with the Schroeder-phased inputs and measuring the response of the piezoceramics and the accelerometer. Responses are measured separately for each excitation signal. After the undamaged beam is tested, the beam is damaged by drilling a hole (type 1) through the structure at a specified location. After testing, the beam is again drilled to produce a larger hole (type 2). Again, after testing this beam is further drilled to produce damage type 3. The holes are drilled without removing the beam from the clamp. This eliminates variation in the response due to changes in the boundary condition. Hence, each series of experiments is carried out on this same beam with the level of damage varying from no damage to the most severe damage, type 3. The excitation pattern is repeated for each damage case.

Identification results

For computations the dimension of the approximation space N was set to 14 since the eigenvalues of the approximate finite dimensional system became stable at $N \geq 14$ in the sense that the eigenvalues

do not change significantly as N increases beyond 14. A comparison with the selection of $N = 10$ in Banks et al. where approximation was done for an undamaged beam reveals that more finite elements are required to capture the small changes in the structure due to the damage than are needed to study the undamaged beam.

As might be expected from our earlier discussions, we readily observed in the experimental data that the frequency information is not sufficient to use as damage detection information. In Table 4.7, we list the frequency change in percentage for beams with different damage type and location versus undamaged beam. Note that in each case the change in frequency was at most 2.22%.

		Damaged Cases					
	Undamaged	DT1		DT2		DT3	
Mode	f (Hz)	f (Hz)	Δf/f (%)	f (Hz)	Δf/f (%)	f (Hz)	Δf/f (%)
1	6.0000	6.0000	0	5.8667	2.22	5.8667	2.22
2	33.8667	33.8667	0	33.6000	0.79	33.2000	1.97

(i) Experiment 1

		Damaged Cases					
	Undamaged	DT1		DT2		DT3	
Mode	f (Hz)	f (Hz)	Δf/f (%)	f (Hz)	Δf/f (%)	f (Hz)	Δf/f (%)
1	6.1333	6.1333	0	6.1333	0	6.1333	0
2	34.4000	34.4000	0	34.2667	0.39	34.0000	1.16

(ii) Experiment 2

Table 4.7: Frequency shift for $48.26 \times 2.032 \times 0.1651\,\text{cm}^3$ beams with holes at (i) 2.54 cm and (ii) 22.86 cm from clamp.

The piezoceramic parameters (K_B, E_p, ρ_p), undamaged beam parameters (E_b, ρ_b) and damping parameters are estimated first from data on the undamaged beam vibrations. The results are then employed as fixed values in the damage detection.

To be consistent with the formulation for an Euler-Bernoulli beam, the geometry of the damage is assumed to be symmetric about the neutral axis of the beam. And for the same reason, we did not attempt to identify the shape of the damage since we use a 1-D equation to describe a 3-D structure, and many different 3-D shapes could be represented by a 1-D function (the level of nonuniqueness would be too high in such an endeavor). Instead, the possibility of using a fixed damage shape in the mathematical model is investigated. A reasonable shape would be a circular function

$$S(x) = 2\sqrt{r^2 - (x - x_c)^2}, \tag{4.189}$$

in which r is the radius, x_c is the center of the circle, and 2 in the equation is due to the symmetry property. In this case, the shape functions $S_A(x)$ and $S_I(x)$ become

$$S_A(x) = t_b\, S(x)$$
$$S_I(x) = \frac{1}{12} t_b^3\, S(x)\,.$$

To have a good initial guess for the shape function parameters in searching for the optimal size and location of the hole, a series of simulated responses was computed for different sizes and locations. The parameter sets (r, x_c)'s which yield smaller residual (comparing the numerical solutions with the experimental data) among all the integration runs were selected as initial guesses.

An IMSL routine of the Levenberg-Marquardt algorithm with a finite difference Jacobian algorithm (ZXSSQ in IMSL9) is used to solve the approximating finite dimensional least squares minimization problems.

A summary of the estimated parameters for the undamaged beam is given in Table 4.8. The data fit result for Beam II is not presented here since it is similar to one for Beam I.

A summary of damage detection results is listed in Table 4.9. The radius for damage case d is an equivalent radius, i.e. the removed area of a circle with the radius is the same as the actual removed area due to three holes.

		given	Beam I	Beam II
E	beam	7.3000×10^{10}	7.3413×10^{10}	6.9709×10^{10}
($N \cdot m^2$)	PZT	6.3000×10^{10}	7.1510×10^{10}	6.8285×10^{10}
ρ	beam	2.7659×10^3	3.0010×10^3	2.7659×10^3
(kg/m^3)	PZT	7.6000×10^3	9.1713×10^3	8.9869×10^3
c_D	beam	—	1.3391×10^6	1.0521×10^6
($N \cdot m^2 \cdot s$)	PZT	—	1.2188×10^6	1.2188×10^6
γ ($N \cdot m^2 \cdot s$)		—	0.95648×10^{-2}	1.1717×10^{-2}

Table 4.8: Given and estimated structural parameters for undamaged beam I & II.

	Damage Center	Damage Cases					
		b		c		d	
Beam	& Radius	Actual	Est.	Actual	Est.	Actual	Est.
I	$\bar{x}_c$ (cm)	2.5400	2.8217	2.5400	2.7798	2.5400	2.5943
	$\bar{r}$ (cm)	0.1588	0.2303	0.3175	0.4420	0.3550	0.6745
II	$\bar{x}_c$ (cm)	22.8659	23.8167	22.8659	24.1751	22.8659	24.2546
	$\bar{r}$ (cm)	0.1588	0.4764	0.3175	0.6350	0.3550	0.8263

Table 4.9: Given and estimated damage for Beam I & II

Discussion

We first point out that the parameter identification results of Table 4.8 for the undamaged beams demonstrate again the consistency of our method for structures without damage. Two sets of

estimated structural parameters obtained by using response data recorded from two identically made beams are extremely close in the values. This difference is caused in part by experimental limitations, e.g. it is very difficult to obtain consistent clamping in several experiments. To partially alleviate these limitations, the optimization procedure yields different stiffness and mass density parameters for the two experiments. As one might expect, the ratio of stiffness to mass density for Beam I is lower than the one for Beam II since the damped natural frequencies for Beam I are smaller than those for Beam II.

Good agreement is obtained between the estimated damage location and the actual ones. The parameter estimation method is sensitive to small changes due to the damage - the smallest hole is 3.175 mm in diameter which is only 0.08% damage (ratio of removed mass to undamaged beam mass). However, the size of the damages are all over estimated. We suspect that beams' characteristics were slightly changed during the drilling process in a manner in which the change was not modeled in our equations (e.g., changes in mass density around the holes due to shearing and stress, etc.). Simulations of the numerical solutions with actual damage in the model yield less frequency response change than is present in the experimental data. The estimation of damage location for damage type (d) is as good as the other two even though the assumed damaged shape (one circle) in the mathematical model is very different from actual shape.

Comparing the damaged beams DT#-1 to DT#-9, we find that the experimental frequency response changes are less for a damage location further away from the clamped end. Even so, our results demonstrate that the method is sensitive to the different locations of the damage. Even though the estimated locations for experiment II are not as good as those obtained from experiment I, they are within 6% error from the actual locations.

Our attempt at using a fixed damage shape is successful. The estimation of damage location for Damage Type 3 is as good as for Damage Type 1 and 2 although we matched the experimental data for DT3 with a numerical solution to a mathematical model with a singular circular hole. In many situations, it is more important to find whether there are damages and where they are than what shape they possess. In these cases, we can proceed to estimate the locations and the approximate sizes characterized by radii which significantly reduces computation time since we only estimate one parameter, the radius - as opposed to a function characterized by many unknown coefficients.

4.4.4 Conclusion

As we have already noted, the idea of using vibration testing as a base for damage detection in structures is not a new one. However, most methods to date are based on modal information. In this report we have presented a theoretically sound computational, PDE based non-modal framework for the identification of spatially dependent dynamic parameters in piezoceramic embedded structures using nondestructive vibration tests. Using data from beam experiments, we demonstrated the feasibility of our approach in obtaining reliable physically meaningful dynamic parameters such as stiffness, damping, and mass density, and hence identifying damages from the changes in those physical coefficients. Furthermore, this rigorous systematic approach permits one to use the piezoceramics to both excite and sense vibrations in a self-analysis framework that is the defining feature of smart material structures.

Although results have been obtained only for aluminum beams, the framework can be readily apply to plate, shell and beam like structures. However in case of the crack damage and delaminations, the PDE model developed here can not be applied directly since the physical parameters ρ, EI and $c_D I$ can, of course, no longer be described by equations 4.173 in section 4.4.1. Nevertheless, once a proper model describing the dynamics of a particular structure with cracks is developed, we suggest that the same framework and convergence arguments could be applied with some modifications.

The framework developed in this report is also valid for composite material structures since our theoretical and computational methods can include weaker (than Kelvin-Voigt) and more complex damping operators.

Our efforts with variable geometry and variable elasticity parameters promise to alleviate some difficulties encountered by modal methods. This, among other of our findings, strongly supports our

own belief that geometry based partial differential equation techniques can play a very important role in modeling for the emerging technology of adaptive or smart material structures. Our conclusions are based on the methodology developed here which contains theory, computational and experimental tests, all of which are consistent.

4.5 Feedback Control of a 2D Thermal Fluid

The problem of controlling a viscous fluid flow in a convection loop is considered. We present different approaches using feedback controllers for a convective flow in a circular pipe standing in a vertical plane (see Figure 4.22).

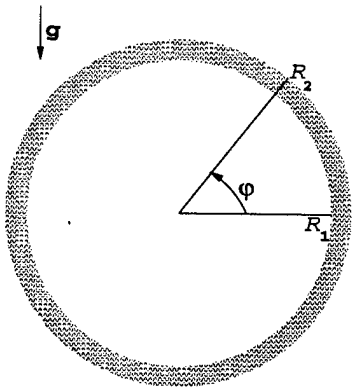


Figure 4.22: Description of the thermal convection loop

The inner radius of the pipe is R_1 and the outer radius is R_2. The radial coordinate r is measured from the inner wall to the outer wall of the pipe and the angular position φ is measured counterclockwise. The velocity $\vec{v}(t,r,\varphi)$ of the flow is given by $\vec{v} = v(t,r)\hat{\varphi}$ where $\hat{\varphi}$ is the unit vector along the tube and the fluid's temperature $T(t,r,\varphi)$ depends on the time and the position. We assume that the fluid temperature at the wall equals the wall temperature plus a control source. In particular, we consider the temperature on the inner and outer walls

$$T(t,R_1,\varphi) = T_{W1}(t,\varphi) + w_1(t,\varphi)$$

and

$$T(t,R_2,\varphi) = T_{W2}(t,\varphi) + w_2(t,\varphi),$$

respectively, as Dirichlet boundary controls. Here $T_{W1}(t,\varphi)$ and $T_{W2}(t,\varphi)$ are given wall temperatures.

Boussinesq's approximation assumes that the fluid is assumed incompressible even when the temperature is not uniform. Therefore, properties of the fluid are assumed constant except the density in the buoyancy term, since in free convection the nonuniformity in specific weight is the motive force. The body force $\vec{F}$ is due to the gravitational acceleration and the buoyancy force per unit mass, that is,

$$\vec{F} = \vec{g} + \beta\,(T(t,r,\varphi) - T_0)(-\vec{g}),$$

where $\vec{g}$ is the gravity acceleration, β is the thermal expansion coefficient and T_0 is the bulk fluid temperature. The equation is written in polar coordinates and we introduce the operator ∇_r^2 defined by $\nabla_r^2 = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r}$.
The pressure term is eliminated from the Navier-Stokes equation by integrating along a circular path at a fixed radius r. Under these assumptions it can be shown that the dynamics of the flow can be described by the nonlinear system

$$\frac{\partial v}{\partial t}(t,r) = \frac{g\beta}{2\pi}\int_0^{2\pi} T(t,r,\varphi)\,\cos\varphi\,d\varphi + \nu\nabla_r^2 v(t,r), \tag{4.190}$$

$$\frac{\partial T}{\partial t}(t,r,\varphi) = -\frac{v(t,r)}{r}\,\frac{\partial T}{\partial \varphi}(t,r,\varphi) + \chi\,\nabla_r^2 T(t,r,\varphi) + \frac{\chi}{r^2}\,\frac{\partial^2 T}{\partial \varphi^2}(t,r,\varphi), \tag{4.191}$$

with boundary conditions

$$v(t,R_1) = v(t,R_2) = 0,$$
$$T(t,R_1,\varphi) = u_1(t,\varphi), \qquad T(t,R_2,\varphi) = u_2(t,\varphi).$$

Here, ν is the kinematic viscosity, χ is the fluid's thermal conductivity and u_1 and u_2 are the applied temperature controls. Therefore, we have a quasilinear infinite-dimensional distributed parameter system.

Let $\Omega_1 = [R_1, R_2]$, $\Omega_2 = [0, 2\pi]$, $\Omega = \Omega_1 \times \Omega_2$ and $\Gamma = \{R_1, R_2\} \times \Omega_2$. The system can be written in the abstract form

$$\dot{z}(t) = Az(t) + f(z(t)) + Bu(t), \qquad t > 0, \tag{4.192}$$

on the state space $X = L^2(\Omega_1) \times L^2(\Omega)$. Here, $z(t) = (v(t,.), T(t,.,.))^T$ is the state. The linear operator $A = A_0 + A_1$ is defined on

$$Dom(A) = Dom(A_0) = [H^2(\Omega_1) \cap H_0^1(\Omega_1)] \times [H^2(\Omega) \cap H_0^1(\Omega)]$$

by

$$A_0 = \begin{pmatrix} \frac{\mu}{\rho} \nabla_r^2 & 0 \\ 0 & \chi \nabla_r^2 + \frac{\chi}{r^2} \frac{\partial^2}{\partial \varphi^2} \end{pmatrix}, \qquad A_1 = \begin{pmatrix} 0 & \mathcal{I} \\ 0 & 0 \end{pmatrix}, \tag{4.193}$$

where $\mathcal{I} : L^2(\Omega) \to L^2(\Omega_1)$ is the bounded linear operator

$$[\mathcal{I}\omega](r) = \frac{g\beta}{2\pi} \int_0^{2\pi} \cos\varphi \; \omega(t,r,\varphi) d\varphi. \tag{4.194}$$

The nonlinear operator $f : H_0^1(\Omega_1) \times H_0^1(\Omega) \to X$ is defined by

$$[f(v(.), T(.,.))](r,\varphi) = \left(0, -\frac{v(r)}{r} \frac{\partial T}{\partial \varphi}(r,\varphi)\right)^T. \tag{4.195}$$

The input operator B is the unbounded linear operator $B = -\hat{A}H$ where $\hat{A}$ is the lifting of A from X to $Dom(A^*)'$ and $H : L^2(\Gamma) \to X$ is given by $Hg = (0, Dg)^T$ where D is the Dirichlet map for the Laplacian ∇^2 on Ω. Thus $D : L^2(\Gamma) \to L^2(\Omega)$ is the bounded linear operator satisfying

$$Dg = \omega \qquad \text{where} \qquad \nabla^2 \omega = 0 \qquad \text{and} \qquad \omega|_\Gamma = g. \tag{4.196}$$

4.5.1 The Linear Quadratic Regulator Problem

We linearize the system about the equilibrium point $v = 0$, $T = 0$. Since $f(0,0) = (0,0)^T$, the linearized system becomes

$$\dot{z}(t) = Az(t) + Bu(t), \quad t > 0, \qquad z(0) = z_0, \tag{4.197}$$

where $z_0 \in H$, $u \in U = L^2(\Gamma)$.

The LQR problem is to minimize the quadratic cost defined by

$$J(z_0, u) = \int_0^\infty [\langle Qz(t), z(t) \rangle_H + \langle Ru(t), u(t) \rangle_U] \, dt \tag{4.198}$$

over all controls $u \in L_2((0,\infty); U)$, subject to the linear system (4.197). The state weighting operator for the LQR problem is

$$Q = \begin{pmatrix} Q_v & 0 \\ 0 & Q_T \end{pmatrix}$$

with $Q_v = q_v I_{L_2(\Omega_1)}, Q_T = q_T I_{L_2(\Omega)}$ and q_v, q_T positive constants. The operators $I_{L_2(\Omega_1)}, I_{L_2(\Omega)}$ denote the identity operators in $L_2(\Omega_1)$ and $L_2(\Omega)$, respectively. The control weighting operator is given by $R = q_u I_U$, I_U denotes the identity operator on U and q_u is a positive constant.

Recall that H and U are Hilbert spaces. Also, it can be shown that the operator $A : \mathcal{D}(A) \to H$ generates an analytic semigroup on H. The dense continuous injection $\mathcal{D}(A) \hookrightarrow H$ allowed us to lift A to H. The input operator B defined by $-\hat{A}\hat{D}$ is not bounded as an operator into H. However, we have the following result.

Lemma 3 *The operator $B : U \to W' = (\mathcal{D}(A))'$ satisfies*

$$\hat{A}^{-(3/4+\epsilon)} B \in \mathcal{L}(L^2(\Gamma); L^2(\Omega_1) \times L^2(\Omega)). \tag{4.199}$$

Proof: The input operator B is given by

$$\begin{aligned} \langle Bu, \omega \rangle &= \langle u, B^* \omega \rangle \\ &= -\frac{g\beta R_2}{2\pi} \int_{R_1}^{R_2} \int_0^{2\pi} u(\varphi) \cos\varphi \; \omega_1(r) \; dr d\varphi - \int_\Gamma u(\varphi) \; \frac{\partial \omega_2}{\partial \eta}(r, \varphi) \; r \; d\varphi. \end{aligned} \tag{4.200}$$

We define the operators B_v and B_T as follows

$$\langle B_v u, \omega_1 \rangle = \langle u, B_v^* \omega_1 \rangle = -\frac{g\beta R_2}{2\pi} \int_{R_1}^{R_2} \int_0^{2\pi} u(\varphi) \cos\varphi \; \omega_1(r) \; dr d\varphi \tag{4.201}$$

$$\langle B_T u, \omega_2 \rangle = \langle u, B_T^* \omega_2 \rangle = -\int_\Gamma u(\varphi) \; \frac{\partial \omega_2}{\partial \eta}(r, \varphi) \; r \; d\varphi. \tag{4.202}$$

For each $\epsilon > 0$, the trace operator

$$\omega_2 \to \frac{\partial \omega_2}{\partial \eta} : H^{3/2+2\epsilon}(\Omega) \to H^{2\epsilon}(\Gamma) \subset L^2(\Gamma)$$

is continuous. Thus, we have

$$B_T \in \mathcal{L}\left(U, \left[H^{3/2+2\epsilon}(\Omega)\right]'\right) \quad \forall \epsilon > 0.$$

Also, if $\omega_1 \in H^{3/2+2\epsilon}(\Omega_1)$ we have $B_v^* \omega_1 \in L^2(\Gamma)$. Thus, for $\omega \in H^{3/2+2\epsilon}(\Omega_1) \times H^{3/2+2\epsilon}(\Omega)$, we have $B^* \omega \in L^2(\Gamma)$. Hence,

$$B \in \mathcal{L}\left(U, \left[H^{3/2+2\epsilon}(\Omega_1) \times H^{3/2+2\epsilon}(\Omega)\right]'\right).$$

On the other hand,

$$H^{3/2+2\epsilon}(\Omega_1) \times H^{3/2+2\epsilon}(\Omega) = \mathcal{D}\left(\widehat{A}^{3/4+\epsilon}\right),$$

thus,

$$B \in \mathcal{L}\left(U, \mathcal{D}\left(\widehat{A}^{-(3/4+\epsilon)}\right)\right),$$

or equivalently,

$$\widehat{A}^{-(3/4+\epsilon)} B \in \mathcal{L}(U; H). \tag{4.203}$$

Δ

In order to obtain existence of the LQR problem we recall that A generates an exponentially stable semigroup and hence we have the finite cost condition.

Finite Cost Condition: For every $z_0 \in H$, there exists $u \in U$ such that the cost function defined in (4.198) is finite.

Theorem 1 *There exist a self-adjoint, non-negative definite operator $P \in \mathcal{L}(H)$ that satisfies the Algebraic Riccati Equation (ARE)*

$$\langle Pz, A\omega\rangle_H + \langle Az, P\omega\rangle_H - \langle R^{-1}B^*Pz, B^*P\omega\rangle_U + \langle Qz, \omega\rangle_H = 0. \tag{4.204}$$

Moreover,

1. $(\mathcal{A}^*)^{(1-\epsilon)}P \in \mathcal{L}(H), \quad \forall \epsilon > 0,$
2. $R^{-1}B^*P \in \mathcal{L}(H, U),$
3. $J(z_0, u_{opt}) = \langle Pz_0, z_0\rangle_X.$

The LQR problem has a solution of the form

$$u_{opt}(t, z_0) = -R^{-1}B^*Pz_{opt}(t, z_0), \tag{4.205}$$

where z_{opt} is the corresponding solution to (4.197) *with $u = u_{opt}$.*

For a given initial condition $z_0 \in H$, we solve the LQR problem (4.197)-(4.198) for the optimal control u_{opt}. Note that the nonlinear operator $f(z)$ is not taken into account in the LQR problem. However, in order to see how well this linear feedback performs we feed it into the full non-linear system (4.192) to obtain the closed-loop nonlinear system

$$\begin{aligned} \dot{z} &= (A - BK)z + f(z), \qquad t > 0, \\ z(0) &= z_0. \end{aligned} \tag{4.206}$$

Here K is the bounded linear operator defined by $K = R^{-1}B^*P$.

In practice, we must use some type of approximation. We consider finite element and one-mode approximations and use these models to construct the feedback controllers. Therefore, we can use existing finite dimensional algorithms. We now summarize an algorithm due to A. Krener.

4.5.2 Krener's Algorithm

Once the PDE has been approximated we can apply finite dimensional design. Here we review an approach due to Krener which consists of choosing a transformation and a state feedback that linearize the system. We emphasize that Krener's algorithm applies only to finite dimensional systems and has not yet been extended to infinite dimensional systems. We shall apply Krener algorithm to finite dimensional approximations of the Boussinesq equations. The Boussinesq equations lead to approximating systems with a special form so we limit our discussion to these systems.

Consider an autonomous nonlinear system

$$\dot{\mathbf{x}} = f(\mathbf{x}) + Bu \tag{4.207}$$

where $\mathbf{x} \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $f(\mathbf{x})$ is a nonlinear vector function in $\mathbf{x}$, with equilibrium point $\mathbf{x} = 0$ and $B \in \mathbb{R}^{nxm}$.

In order to apply the method, the system is linearized and written as follows

$$\dot{\mathbf{x}} = \mathrm{A}\mathbf{x} + \mathrm{f}^{(k)}(\mathbf{x}) + \mathrm{B}\omega + O(\mathbf{x}^{k+1}), \tag{4.208}$$

where A and B are matrices, $\mathrm{A} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathrm{B} : \mathbb{R}^m \to \mathbb{R}^n$, and $\mathrm{f}^{(k)}(\mathbf{x})$ is a nonlinear vector of dimension n, having entries of degree $k \geq 2$ in $\mathbf{x}$, for example, if $\mathbf{x} = (x_1,\ x_2) \in \mathbb{R}^2$, $\mathrm{f}^{(3)}(\mathbf{x})$ contains terms which are linear combinations of x_1^3, $x_1^2x_2$, $x_1x_2^2$, x_2^3. Terms of order greater than k are contained in $O(\mathbf{x}^{k+1})$. Thus, the system is assumed to have nonlinear terms of order greater or equal k.

This can be accomplished by taking A as the Jacobian of f at 0. Thus $\mathrm{f}^{(k)}(\mathbf{x})$ represents the difference.

One then seeks a change of coordinates and a nonlinear control of order k in $\mathbf{x}$,

$$\mathbf{z}_k = \mathbf{x} - \phi^{(k)}(\mathbf{x}) \tag{4.209}$$
$$\omega_k = \alpha^{(k)}(\mathbf{x}) + \beta^{(k-1)}(\mathbf{x})v + v, \tag{4.210}$$

where $\phi^{(k)}(\mathbf{x})$ is a vector of degree k in $\mathbf{x}$ and $\alpha^{(k)}$ and $\beta^{(k-1)}$ are polynomials in $\mathbf{x}$ of degree k and $k-1$, respectively, such that the terms of order k cancel.

Let $[.,.]$ denote the Lie bracket defined for two vector fields f, g by

$$[f,g] = \frac{\partial g}{\partial x} f - \frac{\partial f}{\partial x} g.$$

The transformation $\phi^{(k)}(\mathbf{x})$ and the nonlinear vectors $\alpha^{(k)}$ and $\beta^{(k-1)}$ have to satisfy the homological equations

$$\mathrm{f}^{(k)}(\mathbf{z}_k) = -\mathrm{B}\alpha^{(k)}(\mathbf{z}_k) + [\mathrm{A}\mathbf{z}_k, \phi^{(k)}(\mathbf{z}_k)] \tag{4.211}$$
$$\mathrm{B}\beta^{(k-1)}(\mathbf{z}_k)v = [\mathrm{B}v, \phi^{(k)}(\mathbf{z}_k)]. \tag{4.212}$$

Assuming that a solution to (4.211)-(4.212) can be found, the resulting system is given by

$$\dot{\mathbf{z}}_k = \mathrm{A}\mathbf{z}_k + \mathrm{B}v_k + O(\mathbf{z}_k^{k+1}), \tag{4.213}$$

and can be written as

$$\dot{\mathbf{z}}_k = \mathrm{A}\mathbf{z}_k + \mathrm{f}^{(k+1)}(\mathbf{z}_k) + \mathrm{B}v_k + O(\mathbf{z}_k^{k+2}). \tag{4.214}$$

This produces a system of the form (4.208). The procedure may be applied repeatedly to obtain a system of the desired order of linearization, given by

$$\dot{\mathbf{z}} = \mathrm{A}\mathbf{z} + \mathrm{B}v + O(\mathbf{z}^{m+1}), \tag{4.215}$$

for any $m > 0$. Here v is the control, which is calculated by solving an LQR problem. In particular, we minimize a given cost function $J(\mathbf{z}_0, v)$ of the form (4.198) subject to the linear system

$$\dot{\mathbf{z}} = \mathrm{A}\mathbf{z} + \mathrm{B}v. \tag{4.216}$$

If the system is controllable, the LQR problem has a unique solution

$$v_{opt} = -K\mathbf{z} = R^{-1}B^T P\mathbf{z},$$

where P is the unique, symmetric, non-negative matrix satisfying the algebraic Riccati equation

$$\mathrm{A}^T P + P\mathrm{A}^T - P\mathrm{B}R^{-1}\mathrm{B}^T P + Q = 0.$$

The control v is linear in $\mathbf{z}$ but is nonlinear in the original state $\mathbf{x}$. The transformations $\phi^{(j)}$ and the polynomials $\alpha^{(j)}$ and $\beta^{(j-1)}$, $j = 2, \ldots, m$ are used to rewrite the control v in terms of $\mathbf{x}$. In Section 4.5.3 we illustrate all the steps in this algorithm and apply the results to the Lorenz equations.

4.5.3 Application to a One Mode Approximation

Lorenz equations are typically used to describe chaotic systems. In particular, the Lorenz equations may be viewed as a one mode approximation to the full Boussinesq system for the thermal convection loop. Wang, Singer and Bau present experimental and numerical results by applying a nonlinear feedback law based on Lorenz model for the problem under consideration. We shall use this model to test and compare the LQR/linearization based control with the non-linear control generated by Krener's method.

Derivation of the Equations

Assume a Fourier series expansion for the difference between the fluid temperature and the temperature at the wall, as follows

$$T(t,r,\varphi) - T_W(t,\varphi) = \sum_{n=0}^{\infty} c_n(t,r)\cos(n\varphi) + s_n(t,r)\sin(n\varphi). \tag{4.217}$$

This expansion is then substituted into the partial differential equations (4.190)-(4.191) obtaining differential equations for $v(t,r)$ and the coefficients of the Fourier series. The equations for $v(t,r)$, $c_1(t,r)$, and $s_1(t,r)$ decoupled from the rest of the system.

Wang, Singer and Bau assumed that the velocity and the temperature are independent of r. York, York and Mallet-Paret expanded $v(t,r), c_1(t,r)$ and $s_1(t,r)$ in a series of Bessel functions of order zero. Both approaches produce an infinite set of ordinary differential equations. The first three equations, which correspond to the first mode, decoupled from the others. They are similar to the Lorenz equations and are given by

$$\begin{aligned} \dot{x}_1(t) &= P(-x_1(t) + x_2(t)) \\ \dot{x}_2(t) &= -x_2(t) - x_1(t)x_3(t) \\ \dot{x}_3(t) &= x_1(t)x_2(t) - x_3(t) - R\tilde{u}, \end{aligned} \tag{4.218}$$

where, in both cases, P and R are related to the loop's Prandtl number and the loop's Rayleigh number, respectively. Also , R and $\tilde{u}$ are related to the temperature at the wall. These relationships depend on the approximation used to derive the equations and are obtained after considering the dimensionless numbers P and R.

The Controlled System

Setting $\tilde{u}(t) = 1 - \frac{1}{R}u(t)$ leads to the lumped parameter control system

$$\begin{aligned} \dot{x}_1(t) &= P(-x_1(t) + x_2(t)) \\ \dot{x}_2(t) &= -x_2(t) - x_1(t)x_3(t) \\ \dot{x}_3(t) &= x_1(t)x_2(t) - x_3(t) - R + u(t). \end{aligned} \tag{4.219}$$

For $\tilde{u} = 1$ (i.e. $u(t) = 0$), the three equilibrium points of the system (4.219) are given by

$$\overline{\mathbf{x}}^1 = (0,0,-R)^T, \quad \overline{\mathbf{x}}^2 = (\sqrt{R-1}, \sqrt{R-1}, -1)^T$$

$$\text{and} \qquad \overline{\mathbf{x}}^3 = (-\sqrt{R-1}, -\sqrt{R-1}, -1)^T.$$

The stability of each fixed point depends on the parameter values.

LQR Optimal Control

We consider now the LQR problem associated to equations (4.219). First, we linearize the system about an equilibrium $\mathbf{x}_e = \overline{\mathbf{x}} = (\overline{x}_1, \overline{x}_2, \overline{x}_3)^T$. Then, we have

$$\dot{\tilde{\mathbf{x}}} = A_L\tilde{\mathbf{x}} + B_L u \tag{4.220}$$

where $\tilde{\mathbf{x}} = \mathbf{x} - \overline{\mathbf{x}}$,

$$A_L = \begin{pmatrix} -P & P & 0 \\ -\overline{x}_3 & -1 & -\overline{x}_1 \\ \overline{x}_2 & \overline{x}_1 & -1 \end{pmatrix} \tag{4.221}$$

and

$$B_L = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{4.222}$$

The system (4.218) is equivalent to the system in the new variable $\tilde{\mathbf{x}}$, given by

$$\dot{\tilde{\mathbf{x}}} = A_L\tilde{\mathbf{x}} + f_L\tilde{\mathbf{x}} + B_L u \tag{4.223}$$

where f_L is defined by

$$f_L(\tilde{\mathbf{x}}) = \begin{pmatrix} 0 \\ -\tilde{x}_1\tilde{x}_3 \\ \tilde{x}_1\tilde{x}_2 \end{pmatrix}. \tag{4.224}$$

The LQR problem is to minimizes the quadratic functional

$$J = \int_0^\infty (\tilde{\mathbf{x}}^T(t)Q_L\tilde{\mathbf{x}}(t) + R_L[u(t)]^2 dt, \tag{4.225}$$

where $Q_L = Q_L^T \geq 0$ and $R_L > 0$, subject to the linear system (4.220)-(4.222).

The optimal control

$$\tilde{\mathbf{u}}_{opt}(t) = -K_L\tilde{\mathbf{x}}(t) \tag{4.226}$$

is calculated solving the above stationary linear quadratic regulator problem.

We then feed the system (4.223) to get the nonlinear closed-loop system

$$\dot{\tilde{\mathbf{x}}} = (A_L - B_L K_L)\tilde{\mathbf{x}} + f_L\tilde{\mathbf{x}}, \tag{4.227}$$

Krener's Nonlinear Control

Here we show how the nonlinear control proposed by Krener is applied to (4.219). The system (4.219) is written in the form

$$\dot{\tilde{\mathbf{x}}} = A_L\tilde{\mathbf{x}} + f_L(\tilde{\mathbf{x}}) + B_L u, \tag{4.228}$$

where A_L,B_L and f_L are defined in (4.221), (4.222) and (4.224), respectively. Introduce now a quadratic change of coordinates

$$\mathbf{z} = \tilde{\mathbf{x}} - \phi^{(2)}(\tilde{\mathbf{x}}) \tag{4.229}$$

and a quadratic control of the form

$$u = \alpha^{(2)}(\tilde{\mathbf{x}}) + \beta(\tilde{\mathbf{x}})v + v, \tag{4.230}$$

in order to obtain a higher order approximate system. We used a Matlab package provided by Krener in order to accomplish this numerically.

Observe that the given system has a nonlinearity of degree 2, thus it is completely linearized in one step. The resulting linear system, in the new variable $\mathbf{z}$, is

$$\dot{\mathbf{z}} = A_L\mathbf{z} + B_L v, \tag{4.231}$$

where A_L and B_L are given by (4.221) and (4.222), respectively. The matrices A_L and B_L are the same as in the LQR problem considered previously (see Section 4.5.3). Thus, the optimal control is $v = -K_L\mathbf{z}$, where K_L is the same gain matrix as in equation (4.226).

The input u to the equation (4.228) in the $\tilde{\mathbf{x}}$-coordinates can be obtained form (4.229)-(4.230) as follows

$$\begin{aligned} u(\tilde{\mathbf{x}}) &= \alpha^{(2)}(\tilde{\mathbf{x}}) + (1 + \beta(\tilde{\mathbf{x}}))v \\ &= \alpha^{(2)}(\tilde{\mathbf{x}}) + (1 + \beta(\tilde{\mathbf{x}}))(-K_L\tilde{\mathbf{x}} + K_L\phi^{(2)}(\tilde{\mathbf{x}})) \\ &= -K_L\tilde{\mathbf{x}} + \left[\alpha^{(2)}(\tilde{\mathbf{x}}) - K_L\beta(x)\tilde{\mathbf{x}} + K_L\phi^{(2)}(\tilde{\mathbf{x}})\right] \\ &\quad + K_L\phi^{(2)}(\tilde{\mathbf{x}})\beta(\tilde{\mathbf{x}}). \end{aligned} \tag{4.232}$$

The resulting nonlinear closed-loop system in $\tilde{\mathbf{x}}$ is given by

$$\dot{\tilde{\mathbf{x}}} = (A_L - B_L K_L)\tilde{\mathbf{x}} + f_L(\tilde{\mathbf{x}}). \tag{4.233}$$

Although the controller (4.232) has a cubic term, this term is relative small and is neglected. Consequently, we are lead to the non-linear feedback law

$$u(\tilde{\mathbf{x}}) = -K_L\tilde{\mathbf{x}} + q(\tilde{\mathbf{x}}) \tag{4.234}$$

where $q(x)$ is the quadratic function given by

$$q(\tilde{\mathbf{x}}) = \alpha^{(2)}(\tilde{\mathbf{x}}) - K_L\beta(\tilde{\mathbf{x}})\tilde{\mathbf{x}} + K_L\phi^{(2)}(\tilde{\mathbf{x}}). \tag{4.235}$$

A Modification of Krener's Controller

While experimenting with the control (4.234) we observed that we could improve performance by scaling the non-linear term. In particular, we weighted the quadratic term by an scalar κ so that a modification of nonlinear control (4.234) becomes

$$u(\tilde{x}) = -K_L\tilde{x} + \kappa q(\tilde{x}). \tag{4.236}$$

An Ad Hoc Controller

In an experimental study, Wang, Singer and Bau propose a nonlinear control of the form

$$\begin{aligned} u(x) &= C(\text{sgn}(x_2)x_2 - \overline{x}_2), \\ &= C(\text{abs}(x_2) - \overline{x}_2), \end{aligned}$$

where C is a constant and $\overline{x}_2$ is the second component of the fixed point to be stabilized. Although this controller is nonlinear, it is based on "engineering insight" and not on any given algorithm.

A Comparison of the Four Controllers

Numerical experiments were conducted to test the controllers. For $P = 4$ the critical value of R is $R_c = 16$. Here, $P = 4$ and $R = 50$ were chosen. Hence all three equilibrium points,

$$\{(0, 0, -50), (7, 7, -1), (-7, -7, -1)\}$$

are unstable. We set $Q = I$, $N = 1$ and controlled to the equilibrium $\mathbf{x}_e = (7, 7, -1)$.

The Lorenz attractor obtained by setting $P = 4$ and $R = 50$ is shown in Figure 4.23. The marks '$*$' in the figure indicate the location of the equilibrium points for the parameter values $\{(0, 0, -50); (7, 7, -1); (-7, -7, -1)\}$.

In Figure 4.24 we show the trajectories of the closed loop system for each of the control laws described above. Here $Q = I$, $N = 1$, the initial point is $(-3, 9, -8)$ and the equilibrium point to be stabilized is $\mathbf{x}_e = (7, 7, -1)$.

In Figure 4.25 the components of each of the controls used are plotted. Similar results are obtained for different initial points. Here we present a particular case as an example. We conducted several simulations with different initial data and similar results were observed. From the numerical results we see that all of the feedback controllers stabilize the chaotic system. The linear LQR optimal control and the nonlinear controller suggest by Wang, Singer and Bau have slow response. Similar results are obtained for the nonlinear control (4.234). However, the scaled controller (4.236) improved performance as illustrated in Figure 4.24.

Observe that the non-linear controller (4.234) shows some improvement over the LQR controller and greatly reduces the oscillations found in the nonlinear controller proposed by Wang, Singer and Bau. Hence, it is worthwhile to investigate the effectiveness of these approach on a more detailed finite element model.

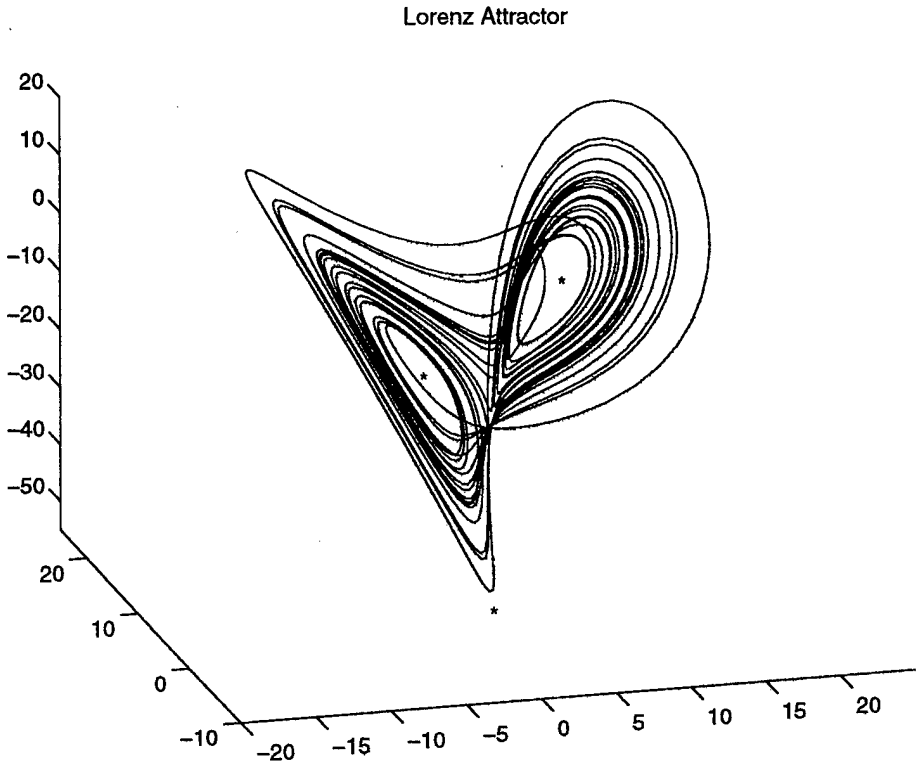


Figure 4.23: Lorenz Attractor. Parameters: $P = 4, R = 50$. Equilibrium points $(7, 7, -1)$, $(-7, -7, -1)$ and $(0, 0, -50)$. Initial Point $\mathbf{x}_0 = (-3, 9, -8)$.

4.5.4 Weak Formulation and Finite Element Model

The abstract form of the PDE model in W' is

$$\begin{aligned} \dot{z} &= Az + f(z) + Bu, \qquad t > 0, \\ z(0) &= z_0 \in H. \end{aligned}$$

Therefore, for $\omega \in W = \mathcal{D}(A_0)$, it follows that

$$\begin{aligned} \langle \dot{z}, \omega \rangle &= \langle Az, \omega \rangle + \langle f(z), \omega \rangle + \langle Bu, \omega \rangle \\ &= \langle A_0 z, \omega \rangle + \langle A_1 z, \omega \rangle + \langle f(z), \omega \rangle + \langle Bu, \omega \rangle \end{aligned}$$

Note that if a finite element scheme is based on the above weak problem, then one needs test functions $\omega \in \mathcal{D}(A_0) = \left[H^2(\Omega_1) \cap H_0^1(\Omega_1)\right] \times \left[H^2(\Omega) \cap H_0^1(\Omega)\right]$.

Thus, for $\omega \in W$, we obtain

$$\langle \dot{z}, \omega \rangle = \left\langle (-A_0)^{1/2} z, (-A_0)^{1/2} \omega \right\rangle + \langle A_1 z, \omega \rangle + \langle f(z), \omega \rangle + \langle u, B^* \omega \rangle . \tag{4.237}$$

The last term in this equation is the only difficult term. In order to relax the smoothness on ω one must define $< u, B^* \omega >$ for $\omega \in H_0^1(\Omega_1) \times H_0^1(\Omega)$. First observe that

$$\langle u, B^* \omega \rangle = -\frac{g \beta R_2}{2\pi} \int_{R_1}^{R_2} \int_0^{2\pi} u(\varphi) \cos\varphi \; \omega_1(r) \; dr d\varphi - \chi \int_\Gamma u(\varphi) \; \frac{\partial \omega_2}{\partial \eta}(r, \varphi) \; r \; d\varphi,$$

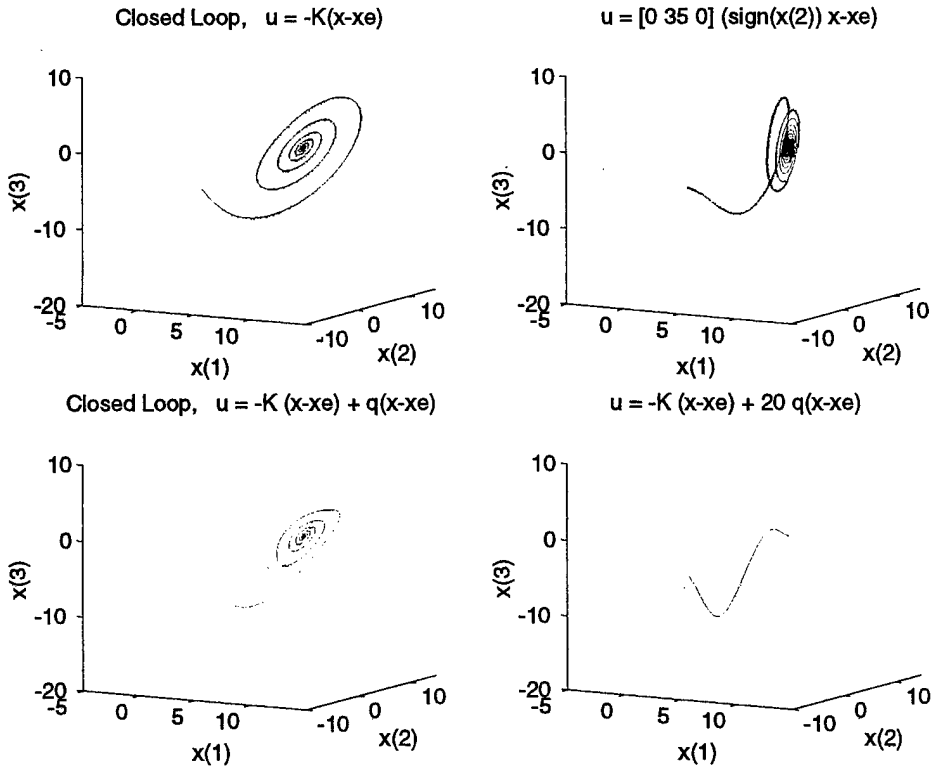


Figure 4.24: Solutions of the nonlinear closed-loop system feeding by four feedback controllers. Parameter values: $P = 4, R = 50$. Equilibrium Point $\mathbf{x}_e = (7, 7, -1)$. Initial Point $\mathbf{x}_0 = (-3, 9, -8)$.

holds for $\omega = (\omega_1,\ \omega_2)^T$ with $\omega_1 \in H_0^1(\Omega_1), \omega_2 \in H^2(\Omega) \cap H_0^1(\Omega)$. Moreover, if $u \in H^{1/2}(\Gamma)$ and $\omega_2 \in H^1(\Omega)$, then the trace theorem implies that, $\frac{\partial \omega_2}{\partial \eta} \in H^{-1/2}(\Gamma)$ and

$$\int_\Gamma u(\varphi)\ \frac{\partial \omega_2}{\partial \eta}(r,\varphi)\ r\ d\varphi \equiv \left\langle u, \frac{\partial \omega_2}{\partial \eta} \right\rangle_{H^{1/2}(\Gamma)\times H^{-1/2}(\Gamma)}$$

is well defined. Thus, if $u \in H^{1/2}(\Gamma)$ then $Bu \in V'$ and

$$\langle Bu, \omega \rangle = \langle u, B^*\omega \rangle = -\frac{g\beta R_2}{2\pi} \int_{R_1}^{R_2} \int_0^{2\pi} u(\varphi) \cos\varphi\ \omega_1(r)\ dr d\varphi - \chi \int_\Gamma u(\varphi)\ \frac{\partial \omega_2}{\partial \eta}(r,\varphi)\ r\ d\varphi, \quad (4.238)$$

for $\omega \in H_0^1(\Omega_1) \times H_0^1(\Omega)$. Consequently, (4.238) is well defined for any $u \in H^{1/2}(\Gamma)$ and $\omega \in H_0^1(\Omega_1) \times H_0^1(\Omega)$. Thus, in this case, piecewise linear functions that vanish on the boundary may be considered in the finite element scheme.

Finite Element Approximation

A Galerkin-based finite element approximation scheme is applied to variational form of the Boussinesq model of the thermal convection problem. The discretization in space (only) yields a semidiscrete scheme and an approximate solution is obtained by solving a finite dimensional ordinary differential equation.

We consider a uniform triangulation in the polar plane consisting of sector elements. This is a natural partition for the given domain and it provides a perfect fit at the boundary (see Figure 4.26). Moreover, sector elements are regarded as rectangles in the polar coordinate system, yielding easy

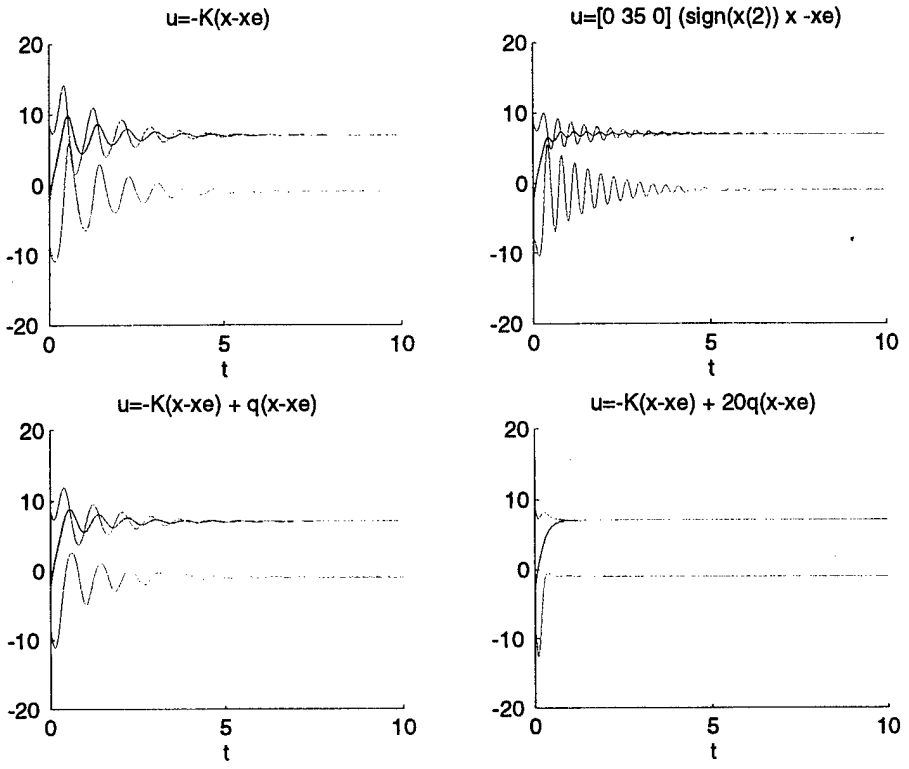


Figure 4.25: Components of the feedback controllers. Parameter values: $P = 4, R = 50$. Equilibrium Point to be Stabilize $\mathbf{x}_e = (7, 7, -1)$. Initial Point $\mathbf{x}_0 = (-3, 9, -8)$.

computations. The finite dimensional space is given by a product space of the form $V^h = V_1^h \times V_2^h$. We take V_1^h to be the space of quadratic splines defined on $\Omega_1 = [R_1, R_2]$ that vanish on the boundary. Then, the approximate velocity is obtained by using Lagrangian quadratic elements in $\mathbb{R}$. We choose a basis $\{\Phi_k^h,\ 1 \le k \le N_v\}$ of $V_1^h \subset H_0^1(\Omega_1)$, the approximate velocity $v^h(t, r)$ can be expressed as a linear combination of $\{\Phi_k^h(r)\}$,

$$v^h(t, r) = \sum_{k=1}^{N_v} v_k^h(t)\Phi_k^h(r) \tag{4.239}$$

with $v_k^h(t) \in \mathbb{R},\ 1 \le k \le N_v$. In order to approximate the temperature, we used the space V_2^h of piecewise bilinear functions defined on Ω that vanish on the boundary Γ. Let $\{\Psi_k^h,\ 1 \le k \le N_T\}$ be a basis of $V_2^h \subset H_0^1(\Omega)$. The approximate temperature $T^h(t, r, \varphi)$ can be expressed as a linear combination of $\{\Psi_k^h(r, \varphi)\}$,

$$T^h(t, r, \varphi) = \sum_{k=1}^{N_T} T_k^h(t)\Psi_k^h(r, \varphi), \tag{4.240}$$

with $T_k^h(t) \in \mathbb{R},\ 1 \le k \le N_T$ and periodic condition $T^h(t, r, 0) = T^h(t, r, 2\pi)$.

Let $N = N_v + N_T$ and consider the set of functions

$$\begin{aligned} \mathcal{B} &= \{\mathcal{B}_i\}_{i=1}^N \\ &= \left\{ \begin{pmatrix} \Phi_1^h \\ 0 \end{pmatrix}, \begin{pmatrix} \Phi_2^h \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} \Phi_{N_v}^h \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ \Psi_0^h \end{pmatrix}, \begin{pmatrix} 0 \\ \Psi_1^h \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ \Psi_{N_T-1}^h \end{pmatrix} \right\}, \end{aligned}$$

where $\{\Phi_k^h,\ 1 \le k \le N_v\}$ is a basis for V_1^h and $\{\Psi_k^h,\ 1 \le k \le N_T\}$ is a basis for $\hat{V}_2^h$. Clearly, $\mathcal{B}$ is a basis for $\hat{V}^h$.

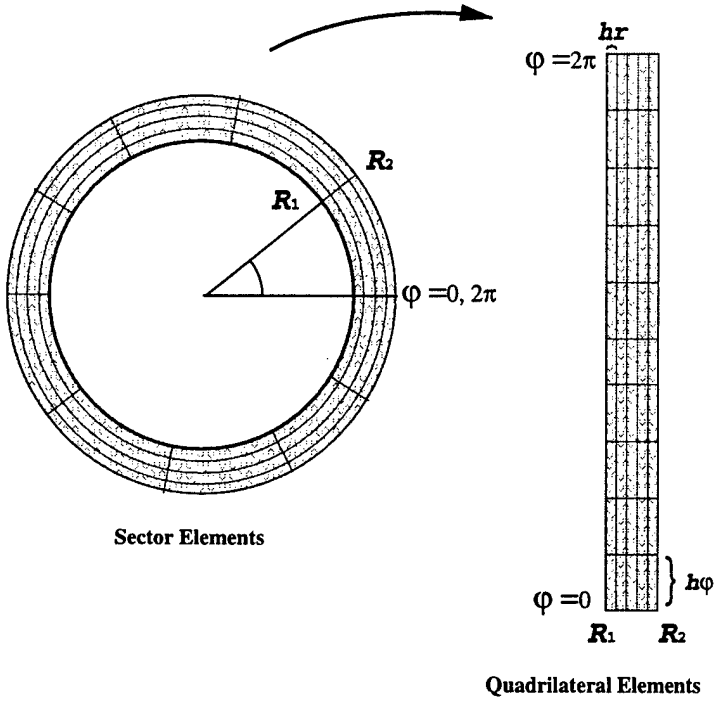


Figure 4.26: Polar Coordinates Triangulation Mapping

Thus, the approximated solution $\mathbf{z}^h(t) = \left(v^h(t),\ T^h(t)\right)^T \in \hat{V}^h$ to the abstract system (4.192)is written as

$$\mathbf{z}^h(t) = \sum_{j=1}^{N} \mathbf{z}_j^h(t)\mathcal{B}_j(r,\varphi),$$

for some $\mathbf{z}_j^h(t) \in \mathbb{R},\ 1 \le j \le N$.

Substituting this approximation to the state into the variational form in equations (4.237) letting the test functions ω range over the basis vectors, we obtain the finite dimensional system

$$\dot{\mathbf{z}}^h(t) = A^h\mathbf{z}^h(t) + f^h(\mathbf{z}^h(t)) + B^h u^h, \tag{4.241}$$

or equivalently,

$$\frac{d}{dt}v^h = A_v^h v^h + A_{1_v}^h T^h + B_v^h u^h, \tag{4.242}$$

$$\frac{d}{dt}T^h = A_T^h T^h + f_T^h(\mathbf{z}^h) + B_T^h u^h \tag{4.243}$$

where

$$\mathbf{z}^h = \begin{pmatrix} v^h \\ T^h \end{pmatrix},$$

$$A^h = \begin{bmatrix} A_v^h & \mathbf{0} \\ \mathbf{0} & A_T^h \end{bmatrix} + \begin{bmatrix} \mathbf{0} & A_{1_v}^h \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

$$f^h(\mathbf{z}^h) = \begin{bmatrix} \mathbf{0} \\ f_T^h(\mathbf{z}^h) \end{bmatrix}$$

$$B^h = \begin{bmatrix} B_v^h \\ B_T^h \end{bmatrix}.$$

Let $M^h = \begin{bmatrix} M_v^h & 0 \\ 0 & M_T^h \end{bmatrix}$ denotes the mass matrix where

$$M_v^h = \left[\left\langle \Phi_i^h, \Phi_j^h \right\rangle_{L^2(\Omega_1)}\right]_{i,j=1,\ldots,N_v} \tag{4.244}$$

and

$$M_T^h = \left[\left\langle \Psi_i^h, \Psi_j^h \right\rangle_{L^2(\Omega)}\right]_{i,j=1,\ldots,N_T}. \tag{4.245}$$

The operators A_v^h, A_T^h, $A_{1_v}^h$, f_T^h, B_v^h and B_T^h are represented by the following matrices

$$A_v^h = \left(M_v^h\right)^{-1}\left[-\nu\left\langle \left(\Phi_i^h\right)', \left(\Phi_j^h\right)'\right\rangle_{L^2(\Omega_1)}\right]_{i,j=1,\ldots,N_v}, \tag{4.246}$$

$$A_T^h = \left(M_T^h\right)^{-1}\left[-\chi\left\langle \nabla\Psi_i^h, \nabla\Psi_j^h\right\rangle_{L^2(\Omega)}\right]_{i,j=1,\ldots,N_T}, \tag{4.247}$$

$$A_{1_v}^h = \left(M_v^h\right)^{-1}\left[\frac{g\beta}{2\pi}\left\langle \Phi_i^h, \int_0^{2\pi}\cos\varphi\Psi_j^h(.,\varphi)\,d\varphi\right\rangle_{L^2(\Omega_1)}\right]_{i=1,\ldots,N_v,1\le j<N_T}, \tag{4.248}$$

$$\mathbf{f}^h(\mathbf{z}^h) = \left(M_v^h\right)^{-1}\left[\left\langle f(\mathbf{z}^h), \Psi_i^h\right\rangle_{L^2(\Omega)},\right]_{i=1,\ldots,N_T}, \tag{4.249}$$

$$B_v^h = \left(M_v^h\right)^{-1}\left[\frac{g\beta}{2\pi}\left\langle \Phi_i^h, \int_0^{2\pi}\cos\varphi u^h(.,\varphi)\,d\varphi\right\rangle_{L^2(\Omega_1)}\right]_{i=1,\ldots,N_v} \tag{4.250}$$

$$B_T^h = \left(M_T^h\right)^{-1}\left[-\chi\left\langle \frac{\partial\Psi_i^h}{\partial\eta}, u^h\right\rangle_{L^2(\Omega)}\right]_{i=1,\ldots,N_T}. \tag{4.251}$$

4.5.5 Numerical Results

In this section we present some numerical results to analyze the performance of a nonlinear control and compare it to the optimal control given in (4.226). Here we consider water flowing in the pipe. The state $T(t,r,\varphi)$ is interpreted as a difference in temperature from the bulk temperature to $60°F$. We consider a pipe with the same dimensions as the one used by Wang, Singer and Bau in their experiments. The system parameters are given in Table I.

Table I. System parameters.

R_1	R_2	ν	β	χ
$1.1975in$	$1.2959in$	$1.22\ 10^{-5}$ ft^2/s	$8.0\ 10^{-5}$ /°F	$1.514\ 10^{-6}$ ft^2/s

The state weighting, Q, and control weighting, R, are given by

$$Q^h = \begin{pmatrix} 1500I_v^h & 0 \\ 0 & 50I_T^h \end{pmatrix},$$

$R^h = 7.5\ 10^{-3}I_U^h$.

In this case, a control is applied on the outer boundary. Figures 4.27-4.29 show the results of a typical run. The number of subdivisions is 3 in radial direction and 5 in angular direction. Figure 4.27 shows the open-loop system response.

The closed-loop response for the LQR controller applied to the full nonlinear system is illustrated in Figure 4.28. Finally, Figure 4.29 shows the closed-loop response for the nonlinear control law.

Although there is much more to do before a theoretical resolution of these issues can be found, the numerical evidence seems to suggest that, for the full Boussinesq equations, nonlinear controllers

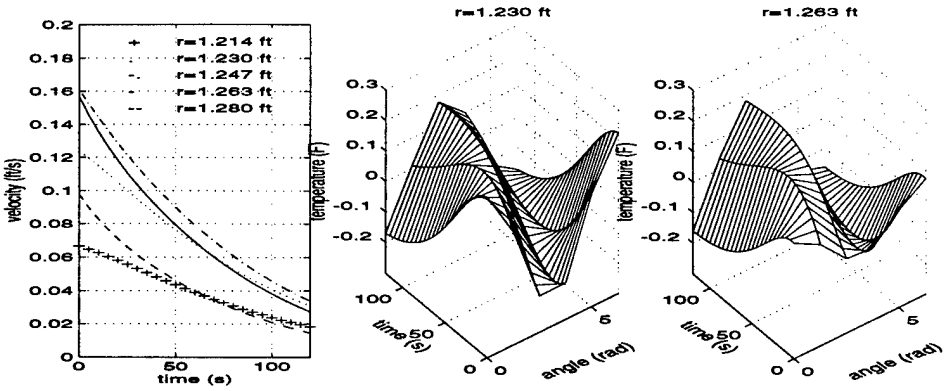


Figure 4.27: Open Loop response.

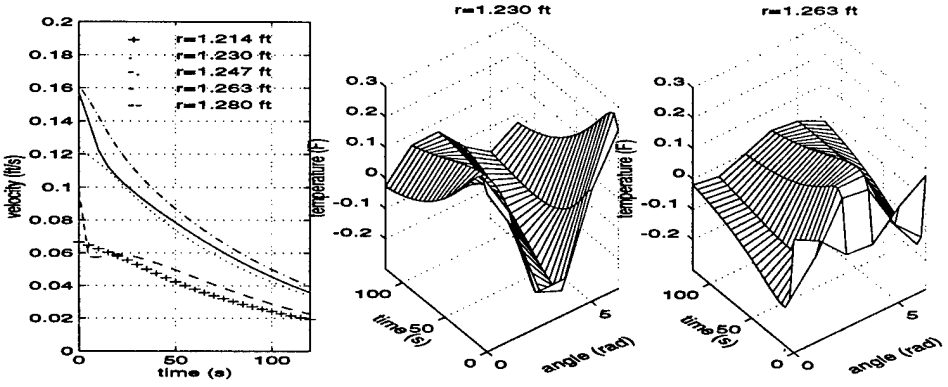


Figure 4.28: Closed Loop response with LQR control.

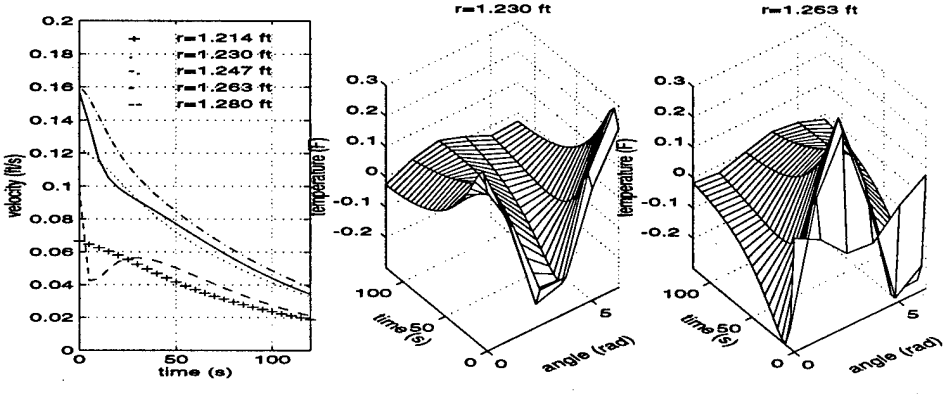


Figure 4.29: Closed Loop response with nonlinear control.

can enhance performance. Although there is some improvement over linear controllers, it is not yet clear that this improvement is significant. One reason for this may be that, if the approximate model sufficiently resolves the weakly excited small-scale (spatial) modes, then the open-loop system can have considerable damping. Hence, feedback controllers (neither linear nor nonlinear) may not significantly enhance the existing natural dissipation.

Optimal Design

Here we present a detailed summary of several projects in the area of optimal design for fluid flow systems. Again, since this report covers the four year period defined by the grant, we present summaries of projects completed during the first three years as well as summaries of two new projects. These two final projects were completed during the period 1 May 1996 and 30 April 1997.

In the area of optimal design, we made significant progress on Sensitivity Equation and Adjoint Methods for fluid flow systems. This area will continue to be a central focus of **CODAC**'s research program. In particular, we plan to enhance these methods and to move the basic research into industry and Air Force applications.

4.6 Optimal Design of a Forebody Simulator

In this section, we review our progress in the area of optimal design. Initially, our projects were motivated by design problems for wind tunnel configurations at the Arnold Engineering Development Center (AEDC). However, this work has expanded to several other application areas. A typical design involves the specification of the internal geometry and tunnel operating conditions to produce a desired flow profile in the test region of the wind tunnel.

The initial problem involved a forebody simulator design. The AEDC is considering a free-jet test facility for full-scale testing of engines in various free flight conditions. Although the test cells are large enough to house the jet engines, they are too small to contain the full airplane forebody and engine. Thus the effect of the forward fuselage on the engine inlet flow conditions must be "simulated" in order to achieve an accurate test. One approach to solving this problem is to replace the actual forebody by a smaller object, called a "forebody simulator" (FBS), see Figure 4.30, and determine the shape of the FBS that produces the best flow match at the engine inlet.

In developing practical computational methods for such optimal design problems, one often relies on cascading simulation software into optimization algorithms. Most optimization algorithms require gradient (sensitivity) information which, in this case, describes how sensitive the cost function (flow variables) is (are) to changes in the shape of the FBS, boundary conditions and wind tunnel operating conditions. In our study below, we use the Sensitivity Equation Method SEM to generate optimum forebody designs for a model 2D forebody design problem.

The following two sections are concerned with the effect of cascading simulation software into optimization algorithms. The example in the second section shows that the numerical approximation of the cost functional can generate artificial local minimum. Two methods are investigated to correct this problem. The first method removes the artificial local minimum by adding dissipation to the flow (which smears out the shock). The second method removes them by adding "dissipation" to the cost functional directly. This is possible here because of the shock capturing scheme used to predict the flow.

The Sensitivity Equation Method (SEM) is an approach that views the simulation scheme as a device to produce approximations of both the function and the sensitivities. The basic idea is to produce approximations of the infinite dimensional sensitivities and to pass these "approximate derivatives" to the optimizer along with the approximate function evaluations. The goal here is to illustrate that a SE based method can be used with standard optimization schemes to produce a practical fast algorithm for optimal design. We concentrate on a particular application (the optimal forebody design problem) and use a specific iterative solver for the flow equations (PARC). Many flow solvers are iterative and for these types of codes, the SE method has perhaps the maximum potential for improving speed and accuracy.

This problem is a two dimensional analogue of the forebody simulator design problem, see Figure 4.30. The geometry we consider is illustrated in Figure 4.31. The goal is to find the shape of the forebody simulator and prescribe the inlet Mach number so that the flow at the inlet reference plane matches a desired flow profile (the operating condition of the aircraft) as closely as possible. To make this problem statement precise, we describe below the model used for computing the flow.

The underlying mathematical model is based on conservation laws for mass, momentum and energy. We will assume a steady, inviscid flow as our model. Thus the flow is governed by the steady Euler equations with appropriate boundary conditions. For now we will consider the unsteady form of the equations which are used in generating a numerical solution to the steady state equations. The equations, in conservation form, are:

$$\frac{\partial}{\partial t}Q + \frac{\partial}{\partial x}F_1 + \frac{\partial}{\partial y}F_2 = 0 \tag{4.252}$$

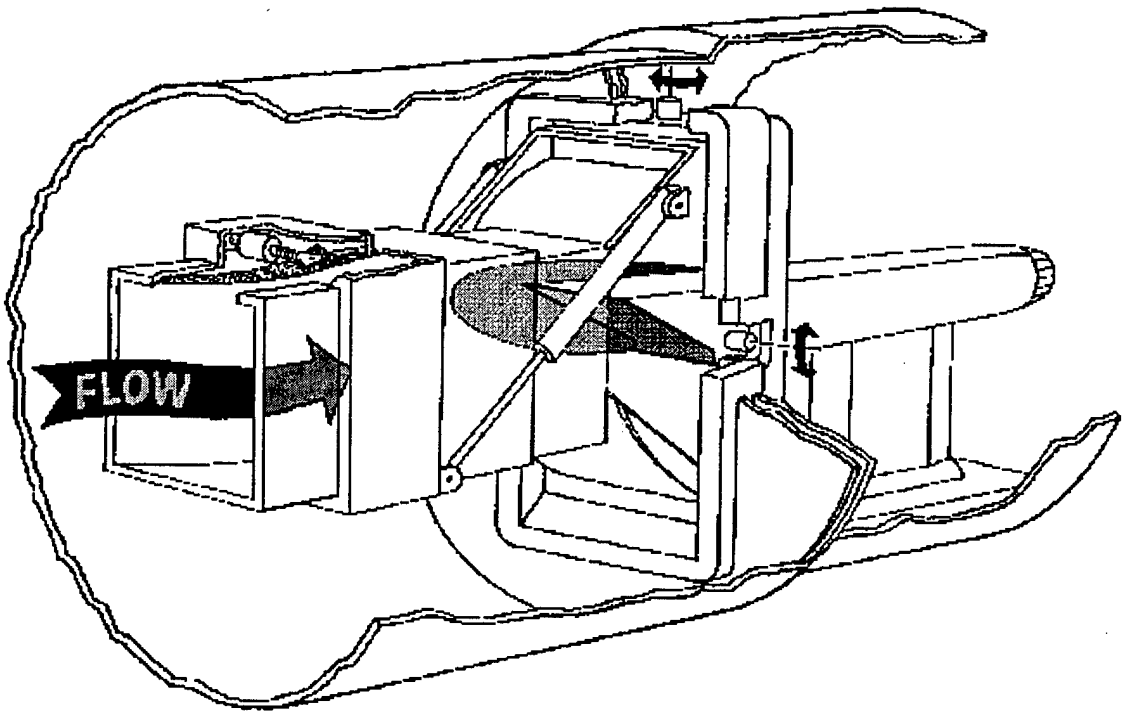


Figure 4.30: Three Dimensional Forebody Simulator Problem

where

$$Q = \begin{pmatrix} \rho \\ m \\ n \\ E \end{pmatrix}, \qquad F_1 = \begin{pmatrix} m \\ mu + P \\ mv \\ (E+P)u \end{pmatrix} \qquad \text{and} \qquad F_2 = \begin{pmatrix} n \\ nu \\ nv + P \\ (E+P)v \end{pmatrix}. \tag{4.253}$$

The velocity components u and v, the pressure P, the temperature T, and the Mach number M are related to the conservation variables, i.e., the components of the vector Q, by

$$u = \frac{m}{\rho}, \qquad v = \frac{n}{\rho}, \qquad P = (\gamma - 1)\left(E - \frac{1}{2}\rho(u^2 + v^2)\right),$$
$$T = \gamma(\gamma - 1)\left(\frac{E}{\rho} - \frac{1}{2}(u^2 + v^2)\right) \qquad \text{and} \qquad M^2 = \frac{u^2 + v^2}{T}. \tag{4.254}$$

At the inflow boundary, we want to simulate a free-jet, so that we specify the total pressure P_0, the total temperature T_0 and the Mach number M_0. We also set $v = 0$ at the inflow boundary. If u_I, P_I and T_I denote the inflow values of the x-component of the velocity, the pressure and the temperature, these may be recovered from P_0, T_0 and M_0 by

$$T_I = \frac{T_0}{(1 + \frac{\gamma-1}{2}M_0^2)}, \qquad P_I = \frac{P_0}{(1 + \frac{\gamma-1}{2}M_0^2)^{\frac{\gamma}{\gamma-1}}} \qquad \text{and} \qquad u_I^2 = M_0^2 T_I = \frac{M_0^2 T_0}{(1 + \frac{\gamma-1}{2}M_0^2)}. \tag{4.255}$$

The components of Q at the inflow may then be determined from (4.255) through the relations

$$\rho_I = \frac{\gamma P_I}{T_I}, \qquad m_I = \rho_I u_I, \qquad n_I = 0 \qquad \text{and} \qquad E_I = \frac{P_I}{\gamma - 1} + \rho_I \frac{u_I^2}{2}. \tag{4.256}$$

The forebody is a solid surface, so that the normal component of the velocity vanishes, i.e.,

$$u\eta_1 + v\eta_2 = 0 \qquad \text{on the forebody,} \tag{4.257}$$

where η_1 and η_2 are the components of the unit normal vector to the boundary. Note that we impose (4.257) on the velocity components u and v, and not on the momentum components m and n. Insofar as the state is concerned, it is clear that it does not make any difference whether (4.257) is imposed on m and n or on u and v, since $m = \rho u$ and $n = \rho v$ and $\rho \neq 0$. It can be shown that it does not make any difference to the sensitivities as well.

Assume that at $x = \beta$ the desired steady state flow $\hat{Q} = \hat{Q}(y)$ is given as data on the line (called the Inlet Reference Plane)

$$IRP = \{(x, y) | x = \beta, \sigma \leq y \leq \delta\}.$$

Also, we assume here that the inflow (total) Mach number M_0 can be used as a design (control) variable along with the shape of the forebody. Let the forebody be determined by the curve $\Gamma = \Gamma(x)$, $\alpha \leq x \leq \beta$ and let $p = (M_0, \Gamma(\cdot))$. The problem can be stated as the following optimization problem:

Problem FBS Given data $\hat{Q} = \hat{Q}(y)$ on the IRP, find the parameters $p^* = (M_0^*, \Gamma^*(\cdot))$ such that the functional

$$\mathcal{J}(p) = \frac{1}{2} \int_\sigma^\delta \parallel Q_\infty(\beta, y) - \hat{Q}(y) \parallel^2 dy \tag{4.258}$$

is minimized, where $Q_\infty(x, y) = Q_\infty(x, y, p)$ is the solution to the steady state Euler equations

$$G(Q, p) = \frac{\partial}{\partial x} F_1 + \frac{\partial}{\partial y} F_2 = 0. \tag{4.259}$$

In the FBS design problem, the data $\hat{Q}$ is generated both experimentally and numerically. In particular, the full airplane forebody (which is longer and larger than the desired FBS) is used to generate the data. Since the FBS is "constrained" to be shorter and smaller, we shall consider the optimization problem illustrated in Figure 4.32 below. The data $\hat{Q}$ is generated by solving (4.252)-(4.257) for the long forebody in Figure 4.32-(a) and the problem is to find p^* to minimize $\mathcal{J}$ where the shortened FBS is constrained to be one half the length of the "real forebody." This problem provides a realistic test of the optimal design algorithm in that the data can not be fitted exactly. Also, we note that we have a problem with shocks in the flow field. We have shown that optimization of flows with shocks can be difficult and requires some understanding of the impact that shocks have on the smoothness of the cost functional.

Clearly the statement of the problem is not complete. For example, one should carefully specify the set of admissible curves $\Gamma(\cdot)$ and questions remain about existence, uniqueness and integrability of "the" solution Q_∞. We will not address these issues here.

Most optimization based design methods require the computation of the derivatives $\frac{\partial}{\partial p} Q_\infty(x, y, p)$. These derivatives are called sensitivities and various schemes have been developed to approximate the sensitivities numerically. A common approach is to use finite differences. In particular, the steady state equation (4.259) is solved for $\tilde{p}$ and again for $\tilde{p} + \Delta p$ and then $\frac{\partial}{\partial p} Q_\infty(x, y, \tilde{p})$ is approximated by $\frac{Q_\infty(x,y,\tilde{p}+\Delta p) - Q_\infty(x,y,\tilde{p})}{\Delta p}$. This method is often costly and can introduce large errors. Another approach is to first derive an equation (the sensitivity equation) for $\frac{\partial}{\partial p} Q_\infty(x, y, p)$ and then numerically solve this equation. We shall illustrate this approach for the forebody design problem.

4.6.1 Sensitivity Equations

First, we consider the design parameter p, either the square of the inlet Mach number (M_o^2) or a forebody shape parameter. We will present the equations for the sensitivity

$$Q' \equiv \frac{\partial Q}{\partial p} \equiv \begin{pmatrix} \rho' \\ m' \\ n' \\ E' \end{pmatrix}, \tag{4.260}$$

where

$$\rho' \equiv \frac{\partial \rho}{\partial p}, \qquad m' \equiv \frac{\partial m}{\partial p}, \qquad n' \equiv \frac{\partial n}{\partial p} \qquad \text{and} \qquad E' \equiv \frac{\partial E}{\partial p}. \tag{4.261}$$

The differential equation system (4.252) has no explicit dependence on the design parameter p, so that equations for the components of Q' are easily determined by formally differentiating (4.252) with respect to p. The result is the system

$$\frac{\partial Q'}{\partial t} + \frac{\partial F_1'}{\partial x} + \frac{\partial F_2'}{\partial y} = 0, \tag{4.262}$$

where

$$F_1' = \begin{pmatrix} m' \\ mu' + m'u + P' \\ mv' + m'v \\ (E+P)u' + (E'+P')u \end{pmatrix} \qquad \text{and} \qquad F_2' = \begin{pmatrix} n' \\ nu' + n'u \\ nv' + n'v + P' \\ (E+P)v' + (E'+P')v \end{pmatrix}, \tag{4.263}$$

and where,

$$u' = \frac{\partial u}{\partial p}, \qquad v' = \frac{\partial v}{\partial p}, \qquad P' = \frac{\partial P}{\partial p} \qquad \text{and} \qquad T' = \frac{\partial T}{\partial p}, \tag{4.264}$$

and where, through (4.254), the sensitivities (4.261) and (4.264) are related by

$$\begin{aligned} u' &= \frac{1}{\rho} m' - \frac{m}{\rho^2}\rho', \qquad P' = (\gamma - 1)\left(E' - \frac{1}{2}\rho'(u^2+v^2) - \rho(uu' + vv')\right), \\ v' &= \frac{1}{\rho} n' - \frac{n}{\rho^2}\rho' \qquad \text{and} \qquad T' = \gamma(\gamma-1)\left(\frac{1}{\rho}E' - \frac{E}{\rho^2}\rho' - (uu' + vv')\right). \end{aligned} \tag{4.265}$$

Note that (4.262) is of the same form as (4.252), with a different flux vector. In particular, (4.262) is in conservation form. As a result of the fact that (4.262) is *linear* in the primed variables, and that by (4.265) u', v' and P' are linear in the components of Q', (4.262) is a linear system in the sensitivity (4.260), i.e., in the components of Q'.

Now, we need to discuss the boundary conditions for Q'. Here we will need to be more explicit about which design parameter we are studying. Except for the inflow conditions, all boundary conditions are independent of the design parameter M_0^2. Thus, the latter may be differentiated with respect to M_0^2 to obtain boundary conditions for the sensitivities. For example, at the forebody where (4.257) holds, we simply would have that

$$u'\eta_1 + v'\eta_2 = 0 \qquad \text{on the forebody.} \tag{4.266}$$

Similar operations yield boundary conditions for the sensitivities along symmetry lines, other solid surfaces and at the outflow boundary.

The inflow boundary conditions for the sensitivities may be determined by differentiating (4.255) and (4.256) with respect to the design parameter M_0^2. Note that this parameter appears explicitly in the right-hand-sides of the equations in (4.255) and (4.256). Without difficulty, one finds from (4.256) that

$$\rho_I' = \frac{\gamma}{T_I} P_I' - \frac{\gamma P_I}{T_I^2} T_I', \qquad m_I' = \rho_I u_I' + u_I \rho_I',$$
$$n_I' = 0 \qquad \text{and} \qquad E_I' = \frac{1}{\gamma - 1} P_I' + \frac{1}{2} u_I^2 \rho_I' + \rho_I u_I u_I', \tag{4.267}$$

where, from (4.255),

$$T_I' = -\left(\frac{\gamma - 1}{2}\right) \frac{T_0}{\left(1 + \frac{\gamma-1}{2} M_0^2\right)^2}, \qquad P_I' = -\left(\frac{\gamma}{2}\right) \frac{P_0}{\left(1 + \frac{\gamma-1}{2} M_0^2\right)^{\frac{2\gamma-1}{\gamma-1}}}$$
$$\text{and} \qquad u_I' = \frac{\sqrt{T_I}}{2M_0} + \frac{M_0}{2\sqrt{T_I}} T_I' = \frac{\sqrt{T_0}}{2M_0 \left(1 + \frac{\gamma-1}{2} M_0^2\right)^{3/2}} \left(1 + (\gamma - 1) M_0^2\right). \tag{4.268}$$

We assume that the forebody is described in terms of a finite number of design parameters which we denote by P_k, $k = 1, \ldots, K$, and that the forebody may be described by the relation

$$y = \Phi(x; P_1, P_2, \ldots, P_K), \qquad \alpha \le x \le \beta. \tag{4.269}$$

All boundary conditions except the one on the forebody also do not depend on the forebody design parameters P_k, $k = 1, \ldots, K$. For example, consider the inflow boundary conditions (4.255)-(4.256). Differentiating these with respect to P_k, $k = 1, \ldots, K$ (and denoting the differentiated variables by a subscript k) yields that

$$\rho_{kI} = m_{kI} = n_{kI} = E_{kI} = T_{kI} = P_{kI} = u_{kI} = v_{kI} = 0 \tag{4.270}$$

at the inflow boundary. Now consider the boundary condition (4.257) on the forebody. We have that on the forebody

$$\frac{\eta_1}{\eta_2} = -\frac{\partial \Phi}{\partial x}. \tag{4.271}$$

Combining (4.257) and (4.271) we have that

$$u \frac{\partial \Phi}{\partial x} - v = 0 \tag{4.272}$$

along the forebody or, displaying the full functional dependence on the coordinates and design parameters, we have at a point (x, y) on the forebody, and at any time t,

$$u\left(t, x, y = \Phi(x; P_1, P_2, \ldots, P_K); M_0^2, P_1, P_2, \ldots, P_K\right) \frac{\partial \Phi}{\partial x}(x; P_1, P_2, \ldots, P_K)$$
$$- v\left(t, x, y = \Phi(x; P_1, P_2, \ldots, P_K); M_0^2, P_1, P_2, \ldots, P_K\right) = 0. \tag{4.273}$$

We can proceed to differentiate (4.273) with respect to any of the forebody design parameters P_k, $k = 1, \ldots, K$. The result is that, along the forebody for $k = 1, \ldots, K$,

$$u_k \frac{\partial \Phi}{\partial x} - v_k = -\left(\frac{\partial u}{\partial y}\right)\left(\frac{\partial \Phi}{\partial P_k}\right)\left(\frac{\partial \Phi}{\partial x}\right) - u \frac{\partial}{\partial x}\left(\frac{\partial \Phi}{\partial P_k}\right) + \left(\frac{\partial v}{\partial y}\right)\left(\frac{\partial \Phi}{\partial P_k}\right), \tag{4.274}$$

where u, v and their derivatives are evaluated at the forebody $(x, y = \Phi(x))$.

If an iterative scheme is used to find a steady state solution of this system ((4.262), (4.270), (4.274)), then we assume that present guesses for the state variables u and v and their derivatives

$\partial u/\partial y$ and $\partial v/\partial y$ and for the design parameters M_0^2 and P_k, $k = 1, \ldots, K$, are known. It follows that the right-hand-side of (4.274) is known as well and equation (4.274), the boundary conditions along the forebody for the sensitivities with respect to the forebody design parameters, is merely an inhomogeneous version of (4.272), the boundary condition along the forebody for the state.

Consider now the problem of minimizing $\mathcal{J}(p)$ as defined above. Most optimization algorithms use gradient information. In particular, if p denotes one of the design parameters, then the derivative

$$\frac{\partial}{\partial p}\mathcal{J}(\tilde{p}) = \int_{\sigma}^{\delta} < \left[\frac{\partial}{\partial p}Q_\infty(\beta, y, \tilde{p})\right], Q_\infty(\beta, y, \tilde{p}) - \hat{Q}(y) > dy \tag{4.275}$$

may be required in the optimization loop. The sensitivity $\frac{\partial}{\partial p}Q_\infty(x, y, \tilde{p})$ satisfies the steady-state version of the sensitivity equations (4.262). In practice one must construct approximations to $\frac{\partial}{\partial p}Q_\infty(x, y, \tilde{p})$ and feed this information into the optimizer.

Assume that one has a particular simulation scheme (finite differences, finite elements, etc.) to approximate the flow $Q_\infty(x, y, \tilde{p})$ on a given grid, i.e.

$$Q_h(x, y, \tilde{p}) \to Q_\infty(x, y, \tilde{p}), \tag{4.276}$$

as the "step size" $h \to 0$. Given the design parameter $\tilde{p}$, one constructs a grid (depending on $\tilde{p}$) and then computes $Q_h(x, y, \tilde{p}) \approx Q_\infty(x, y, \tilde{p})$. This process may require some type of iterative scheme. We will address this issue below. In theory, one could use the same grid and computational scheme to approximate $\frac{\partial}{\partial p}Q_\infty(x, y, \tilde{p})$ so that one generates "approximate sensitivities"

$$\left[\frac{\partial}{\partial p}Q_\infty(x, y, \tilde{p})\right]_h \to \frac{\partial}{\partial p}Q_\infty(x, y, \tilde{p}) \tag{4.277}$$

as $h \to 0$. It is important to note that in general

$$\left[\frac{\partial}{\partial p}Q_\infty(x, y, \tilde{p})\right]_h \neq \frac{\partial}{\partial p}\left[Q_h(x, y, \tilde{p})\right], \tag{4.278}$$

i.e. this approach may not provide "consistent sensitivities". However, some schemes do provide consistent derivatives and even if (4.278) holds, the error

$$ED_h = \left[\frac{\partial}{\partial p}Q_\infty(x, y, \tilde{p})\right]_h - \frac{\partial}{\partial p}\left[Q_h(x, y, \tilde{p})\right] \tag{4.279}$$

may be sufficiently small so that the optimization algorithm converges. Trust region methods are particularly well suited for problems of this type, where derivative information may contain (small) errors. As we shall see below, there are certain cases where $\left[\frac{\partial}{\partial p}Q_\infty(x, y, \tilde{p})\right]_h$ can be computed fast and accurately. Hence, the SE method provides estimates for sensitivities that may prove "good enough" for optimization and yet relatively cheap to compute. A comparison of $\left[\frac{\partial}{\partial p}Q_\infty(x, y, \tilde{p})\right]_h$ and various finite difference approximations of $\frac{\partial}{\partial p}\left[Q_h(x, y, \tilde{p})\right]$ has shown this to be the case.

It is important to note that the details of the computations needed to approximate a sensitivity are not the central issue here. For example, the sensitivity equations (4.262) are viewed as independent partial differential equations that must be solved by "some" numerical scheme. This scheme does not necessarily have to be the same scheme used to solve the flow equation (4.252), although as we shall see below, there are cases where using the same scheme is a useful approach.

Also, note that the sensitivity equations are derived for the problem formulated on the "physical" domain. If one uses a computational method that maps the problem to a computational domain (as does PARC), then the SEM does not require derivatives of this mapping. One simply maps the sensitivity equation (including the necessary boundary conditions), grids the computational domain, solves the resulting transformed equations and then maps back to the physical domain.

If, on the other hand, one mapped the flow equation (4.252) and derived a sensitivity equation in the computational domain, then to obtain the correct sensitivities one would have to compute the mapping sensitivity. Therefore, it is more efficient to derive the sensitivity equations in the physical domain.

Finally, we note that the SEM described here has one additional benefit. To compute a sensitivity, say $\frac{\partial}{\partial p}Q_\infty(x,y,\tilde{p})$, one first selects the parameter value $\tilde{p}$, constructs a computational grid and solves for $\left[\frac{\partial}{\partial p}Q_\infty(x,y,\tilde{p})\right]_h$. There is no need to compute grid sensitivities.

4.6.2 Computing Sensitivities using an Existing Code for the State

Suppose one has available a code to compute the state variables, i.e., to find approximate solutions of (4.252) along with boundary and initial conditions. In principle, it is an easy matter to amend such a code so that it can also compute sensitivities.

First, let us compare (4.252) with (4.262). If one wishes to amend the existing code that can handle (4.252) so that it can treat (4.262) as well, one has to change the definitions of the flux functions from those given in (4.253) to those given in (4.263). Note that the solution for the state is needed in order to evaluate the flux functions of (4.263).

Next, note that (4.262) is the same differential equation for sensitivity with respect to either the inlet Mach number squared or the forebody parameters. Thus, the changes made to the code in order to treat (4.262) will handle both of these cases. In fact, as long as the differential equation and any other part of the problem specification do not explicitly depend on the design parameters, the analogous relation will be the same for all the sensitivities.

The only changes that vary from one sensitivity calculation to another are those that arise from conditions in which the design parameters appear explicitly. In our example, for the sensitivity with respect to M_0^2, one must change the portion of the code that treats the inflow conditions (4.255)-(4.256) so that it can instead treat (4.267)-(4.268). In the problem considered here, the nature (i.e. what variables are specified) of the boundary conditions at the inflow, and everywhere else, is not affected. Note that for the sensitivity with respect to M_0^2, the boundary condition (4.266) on the forebody is the same as that for the state, given by (4.257).

For the sensitivities with respect to the forebody design parameters, the inflow boundary conditions simplify to (4.270), i.e., they become homogeneous. The boundary condition at the forebody is now given by (4.274). Once again, the nature of the boundary conditions is unchanged from that for the state and only the specified data is different. For the inflow boundary conditions, we may still specify the same conditions for the sensitivities, but now they would be homogeneous. The boundary conditions along the forebody change in that they become inhomogeneous, (compare (4.272) and (4.274)).

In summary, to change a code for the state so that it also handles the sensitivities, one must redefine the flux functions in the differential equations, and the data in the boundary conditions. The changes necessary in the code to account for any particular relation that does not explicitly involve the design parameters are independent of which sensitivity one is presently considering.

The previous remarks are concerned only with the changes one must effect in a state code in order to handle the fact that one is discretizing a different problem when one considers the sensitivities. We have seen that these changes are not major in nature. However, there are additional changes that may be needed when one attempts to solve the discrete equations. In the numerical results presented below we use the finite difference code "PARC" (used by engineers at AEDC) to solve the state and sensitivity equations. However, the following comments apply equally well to other CFD codes of this type.

Since we are interested in steady design problems, the time derivative in (4.252) is considered only to provide a means for marching to a steady state. Now, suppose that at any stage of a Gauss-Newton, or other iteration, we have used PARC to find an approximate steady state solution of (4.252) plus boundary conditions. In order to do this, one has to solve a sequence of linear algebraic

systems of the type

$$\left(I + \Delta t A(Q_h^{(n)})\right) Q_h^{(n+1)} = \left(Q_h^{(n)} + \Delta t B(Q_h^{(n)})\right), \qquad n = 0, 1, 2, \ldots, \tag{4.280}$$

where the sequence is terminated when one is satisfied that a steady state has been reached and where $Q_h^{(n)}$ denotes the discrete approximation to the state Q at the time $t = n\Delta t$. We denote this steady state solution for the approximation to the state by Q_h. One problem of the type (4.280) is solved for every time step. In (4.280), the matrix A and vector B arise from the spatial discretization of the fluxes and the boundary conditions. Both of these depend on the state at the previous time level.

Having computed a steady state solution by (4.280), the task at hand is now to compute the sensitivities. We will focus on Q', the sensitivity with respect to the inflow Mach number. Analogous results hold for the sensitivities with respect to the forebody design parameters. Recall that given a state, the sensitivity equations are linear in the sensitivities. Therefore, if one is interested in the steady state sensitivities, instead of (4.262) one may directly treat its stationary version

$$\frac{\partial F_1'}{\partial x} + \frac{\partial F_2'}{\partial y} = 0. \tag{4.281}$$

Since (4.281) is linear in the components of Q', one does not need to consider marching algorithms in order to compute a steady sensitivity. One merely discretizes (4.281) and solves the resultant linear system, which has the form

$$\mathcal{A}(Q_h)Q_h' = \mathcal{B}(Q_h), \tag{4.282}$$

where Q_h' denotes the discrete approximation to the steady sensitivity. The matrix $\mathcal{A}$ and vector $\mathcal{B}$ differ from the A and B of (4.280) because we have discretized different differential equations and boundary conditions. Note that $\mathcal{A}$ and $\mathcal{B}$ in (4.282) depend only on the steady state Q_h and thus (4.282) is a *linear system of algebraic equations* for the discrete sensitivity Q_h'.

The cost of finding a solution of (4.282) is similar to that for finding the solution of (4.280) for a single value of n, i.e. for a single time step. The differences in the assembly of the coefficient matrices and right-hand-sides of (4.280) and (4.282) are minor. Thus, in theory at least, *one can obtain a steady sensitivity in the same computer time it takes to perform one time step in a state calculation.* If one wants to obtain all the sensitivities, e.g., $K + 1$ in our example, one can do so at a cost similar to , e.g., K+1 time steps of the state calculation. This is very cheap compared to the multiple state calculations necessary in order to compute sensitivities through the use of difference quotients.

Although (4.282) is in theory no more complex than one time step in (4.280), we can solve (4.281) by using the same iterative (or another) scheme. The simplest approach (but certainly not the optimal approach) is to use the PARC code to solve (4.281) by time marching. In particular, assume that $Q_h^{(n)}$ is a solution to (4.280), then the system

$$\left[I + \Delta t A'(Q_h^{(n)})\right] (Q')_h^{(n+1)} = \left[(Q')_h^{(n)} + \Delta t B'(Q_h^{(n)})\right] \tag{4.283}$$

can be used to find $(Q')_h^{(n+1)}$ given $(Q')_h^{(n)}$. Thus, one makes an initial guess for $Q_h^{(0)}$ and $(Q_h')^{(0)}$ and then iterates (4.280) and (4.283) simultaneously.

In practice, these "optimal" estimates of speed up are rarely achieved. Moreover, as noted above, it is important to note that finite difference (FD) and sensitivity equation (SE) methods do not necessarily produce the same results. Since the ultimate goal is to find useful and cheap gradients for optimization, the most important issue is whether or not the SE method combined with an optimization algorithm produces a convergent optimal design as fast as possible. We have tested this scheme on the forebody design problem and the next section contains a summary of these results.

4.6.3 An Optimal Design Example

In order to illustrate the SEM and to test its use in an optimization problem, we used the PARC code as described above to compute sensitivities and the used these sensitivities in a BFGS/Trust Region scheme to find an optimal shortened forebody simulator. As shown in Figure 4.32, data was generated by solving the Euler equations over the long forebody at a Mach number of 2.0. The objective is to find a forebody simulator with length one half of the long forebody and such that the resulting flow matches the data as well as possible, i.e., minimizes $\mathcal{J}$ along the outflow boundary.

The shortened forebody was parameterized by a Bezier curve using two parameters. Thus, there are three design parameters $p = (M_0^2, P_1, P_2)$. The algorithm used in this numerical experiment was based on using the PARC code to simultaneously march to the steady state solutions of the flow and sensitivity equations. We made no attempt to optimize the algorithm since the main goal was to test for convergence.

The design algorithm proceeds as follows. First, an initial guess for the optimal design is made, i.e., we select a $p^0 = \left((M_0^2)^0, P_1^0, P_2^0\right)$. A good selection of initial parameters can be made knowing the operating conditions of the aircraft and some rough guess of the shape from the aircraft forebody. In our example, we chose M_0^2 as the inlet Mach number from the computation which generated our data. The initial guess for the parameters were those used to generate the long forebody (although corresponding to different x-locations). These parameters, p^0, are used to generate a grid, the inflow and forebody boundary conditions for both the flow (4.252) and sensitivity equations ((4.262) and an initial guess for both $Q_h^{(0)}$ and $\left(\frac{\partial}{\partial p}Q\right)_h^{(0)}$. In our example, a rough guess for the flow field $Q_h^{(0)}$ uses the constant inflow boundary condition throughout the flow domain. Likewise, the initial guess for the inlet Mach number squared sensitivity is taken as the inflow boundary conditions (given in equation (4.267)) throughout the flow domain. The initial guess for the forebody parameter sensitivities is initially taken as zero (except on the forebody). The systems (4.280) and (4.283) are then solved simultaneously (in our case the left hand side matrix is the same for (4.280) as for the sensitivity equations (4.283), i.e. $A = A'$) for the updated $Q_h^{(1)}$ and $(Q')_h^{(1)}$. The updated $Q_h^{(n)}$ is then used to formulate (4.280) and (4.283) and solve for $(Q_h)^{(n+1)}$ and $\left(\frac{\partial}{\partial p}Q\right)_h^{(n+1)}$. Then one iterates until the desired convergence is achieved. In our example, the residuals, $\Delta Q_h = \left[Q_h^{(n+1)} - Q_h^{(n)}\right]$ were converged to approximately 10^{-15} (in 800 time steps). The outflow data Q_h and $\left(\frac{\partial}{\partial p}Q\right)_h$ are then used to compute $\mathcal{J}(p^0)$ and $\nabla\mathcal{J}(p^0)$.

The optimization algorithm consisted of a BFGS secant method coupled with a "hook" step model trust region method. The initial Hessian was obtained by finite differences on $\nabla\mathcal{J}(\tilde{p})$. The function and gradient information needed by the optimization algorithm is obtained by calling the modified PARC code with $p = \tilde{p}$.

This algorithm was tested for the case where the forebody simulator was allowed to have the full length of the body generating the data. In this case the optimization algorithm produced exact data fits, i.e. $\mathcal{J}(p^*) = 0$ and it recovered the parameters used to generate the data. However, the more realistic test (constraining the length of the forebody simulator) also produced a convergent design and reduced the cost functional significantly.

Figure 4.33 shows the flow field over the long forebody. Observe, that there is a shock in the flow. As noted in the next section, shocks can cause difficulties if one is not careful in the selection of an appropriate numerical scheme. High order schemes can produce (numerically generated) local minima that can cause the optimization loop to fail. This problem is avoided here because the numerical viscosity in PARC (required for stability) is sufficient to "smooth" the cost functional (see the next section for details).

Figure 4.34 shows the shape and flow field of the optimal shortened forebody. This design was obtained after 12 iterations of the optimization loop. Figure 4.35 shows the 1^{st}, 2^{nd}, 3^{rd}, 5^{th} and

12^{th} iterations for the x-component of momentum. The initial guess for the parameters were

$$p^0 = \left(\left(M_0^2\right)^0, P_1, P_2\right) = (2.0, 0.10, 0.15)$$

and

$$\mathcal{J}(p^0) = 3.2339.$$

The "converged" optimal parameters are

$$p^* = p^{12} = (2.020, 0.294, 0.156)$$

with

$$\mathcal{J}(p^*) = 0.2229.$$

Observe that the cost function was decreased by more than 93%. The optimization loops converged rapidly. For example, $\mathcal{J}(p^3) = 0.2334$ and $\mathcal{J}(p^5) = 0.2289$. This is due to the fact that the shock location was found quickly.

Note that although the flows are close, there is a significant error near the forebody. This can also be seen in the plots in Figure 4.36. It is worthwhile to note that the match is good considering the fact that the shortened forebody is constrained to be one half the length of the "real" forebody and that only two Bezier parameters are used to model $\Gamma(\cdot)$. It is also important to note that the shock is captured by the optimal design. In particular, observe in Figure 4.35 how the optimization algorithm "shapes" the shortened forebody so that the optimal shape has a blunt nose. This is necessary in order to generate the correct shock location at the outflow.

4.6.4 Conclusions

The numerical experiment above illustrates that the SEM can produce sensitivities suitable for optimization based design. There are a number of interesting theoretical issues that need to be addressed in order to analyze the convergence of this approach. Moreover, one should investigate "fast solvers" for the sensitivity equations (multi-grid, etc.) as well as develop numerical schemes that are not only fast, but produce consistent derivatives when possible.

Finally, we note that we have conducted a number of timing tests which compute sensitivities to compare the SE method with the finite difference method. In particular, we observed that for the problem above (with three design parameters), the SE method needed only 58% of the CPU time required by finite differencing. When twenty design parameters were used, the SE method produced these sensitivities in about 38% of the time required by finite differencing. These early numerical results indicate that considerable computational savings may be possible if one extends and refines the basic SE method presented here.

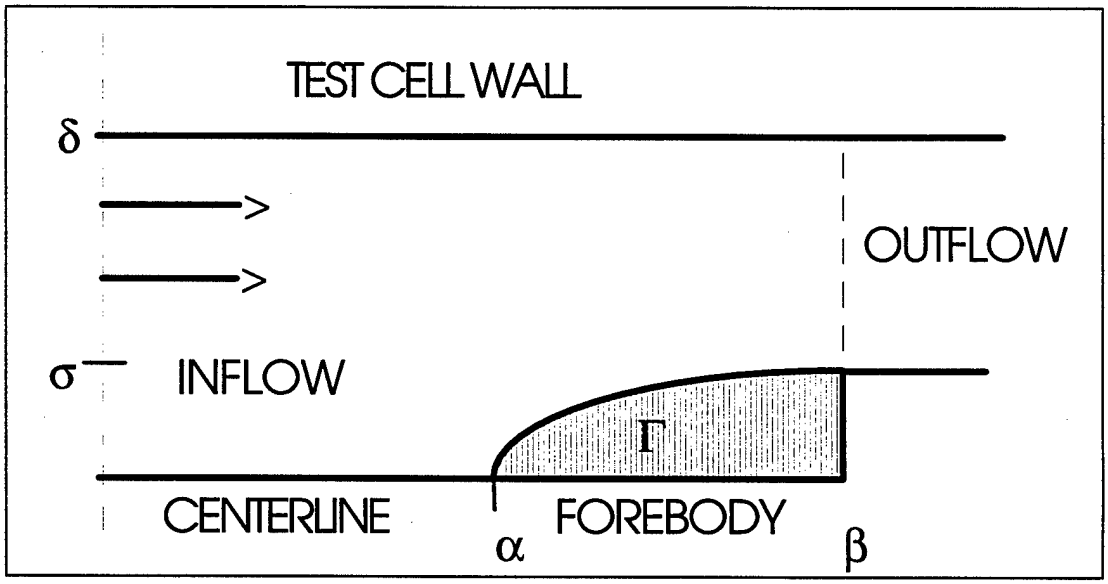


Figure 4.31: Two Dimensional Forebody Simulator Problem

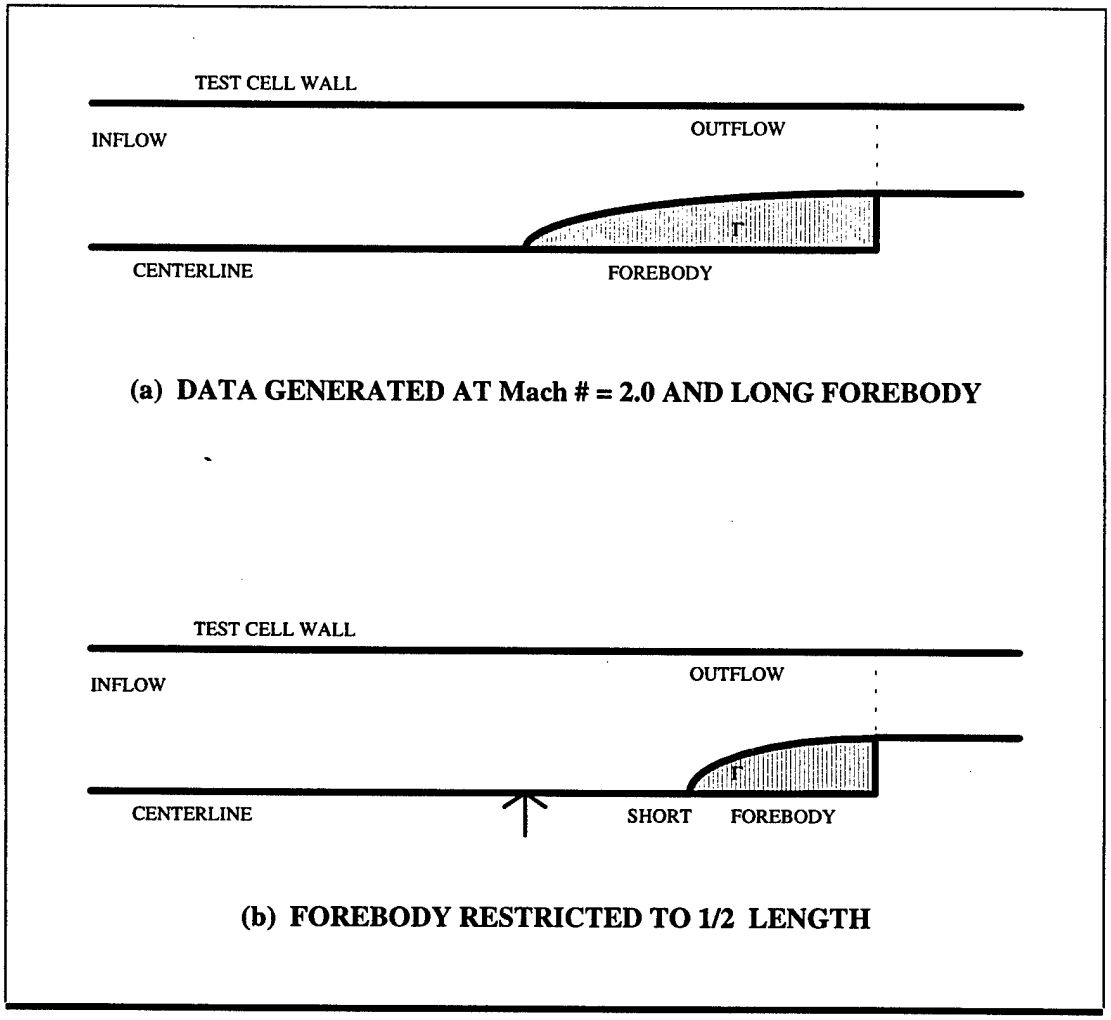


Figure 4.32: Model Forebody Simulator Design Problem

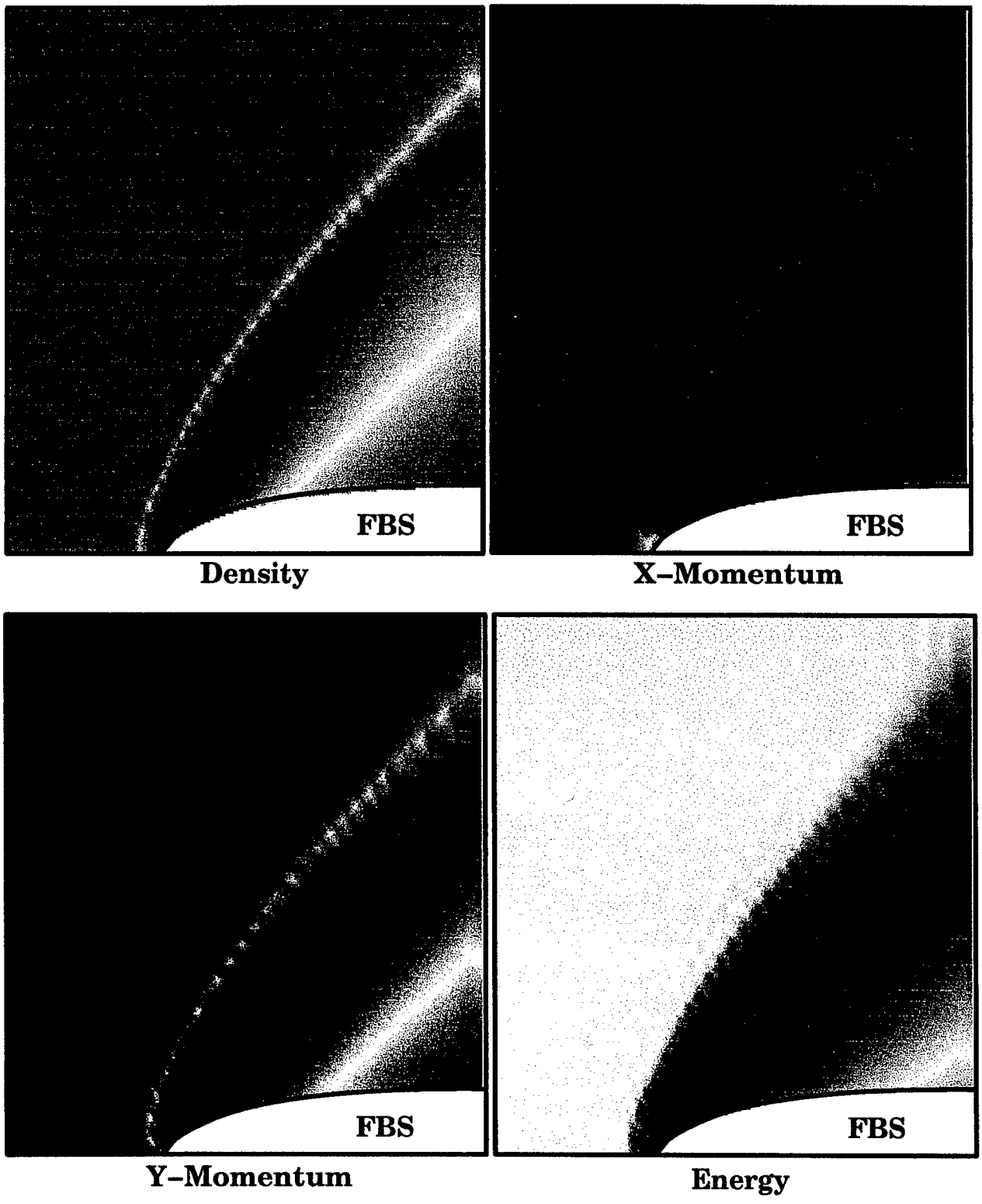


Figure 4.33: Long Forebody Flow Data

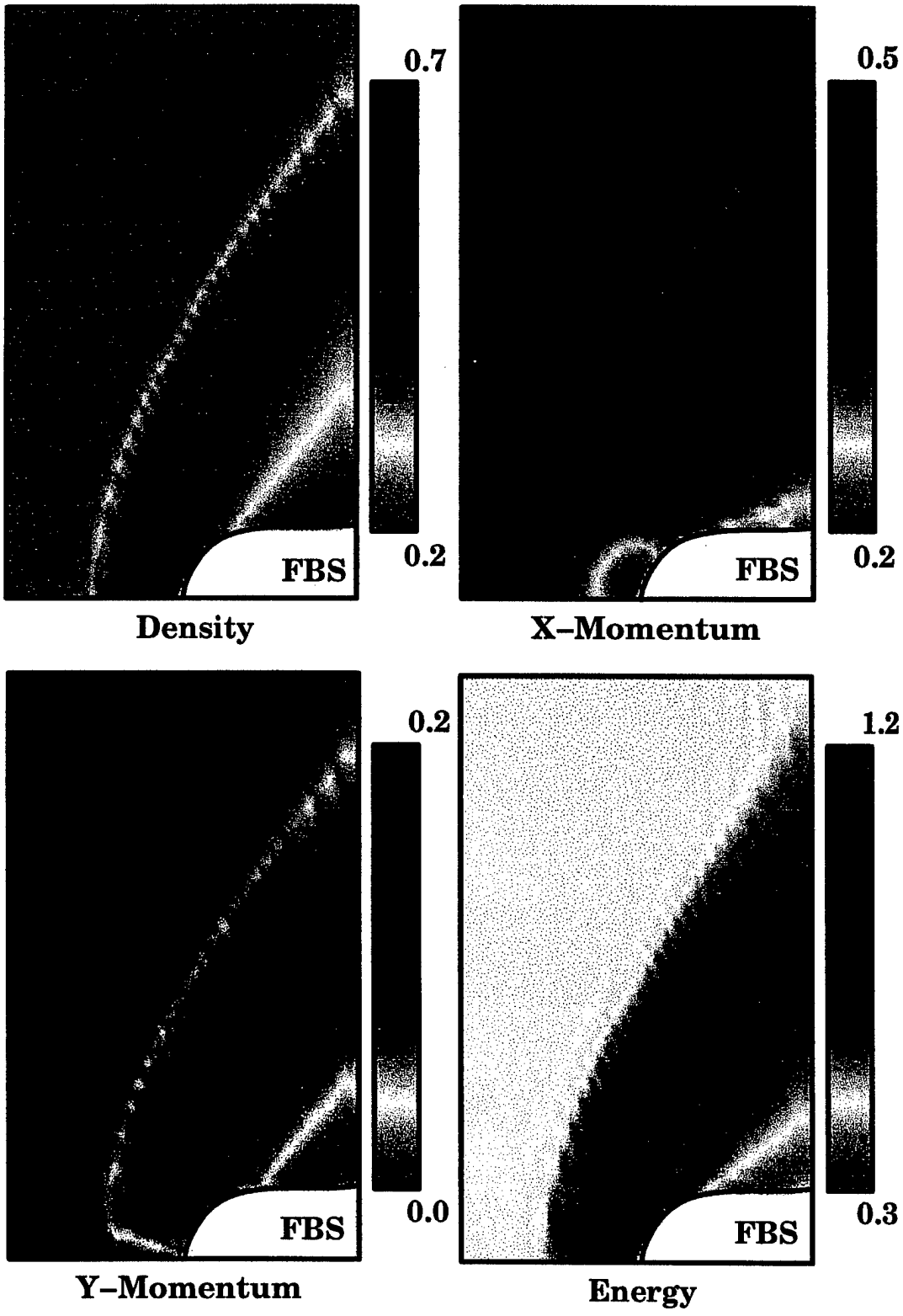


Figure 4.34: Optimal Short Forebody Design

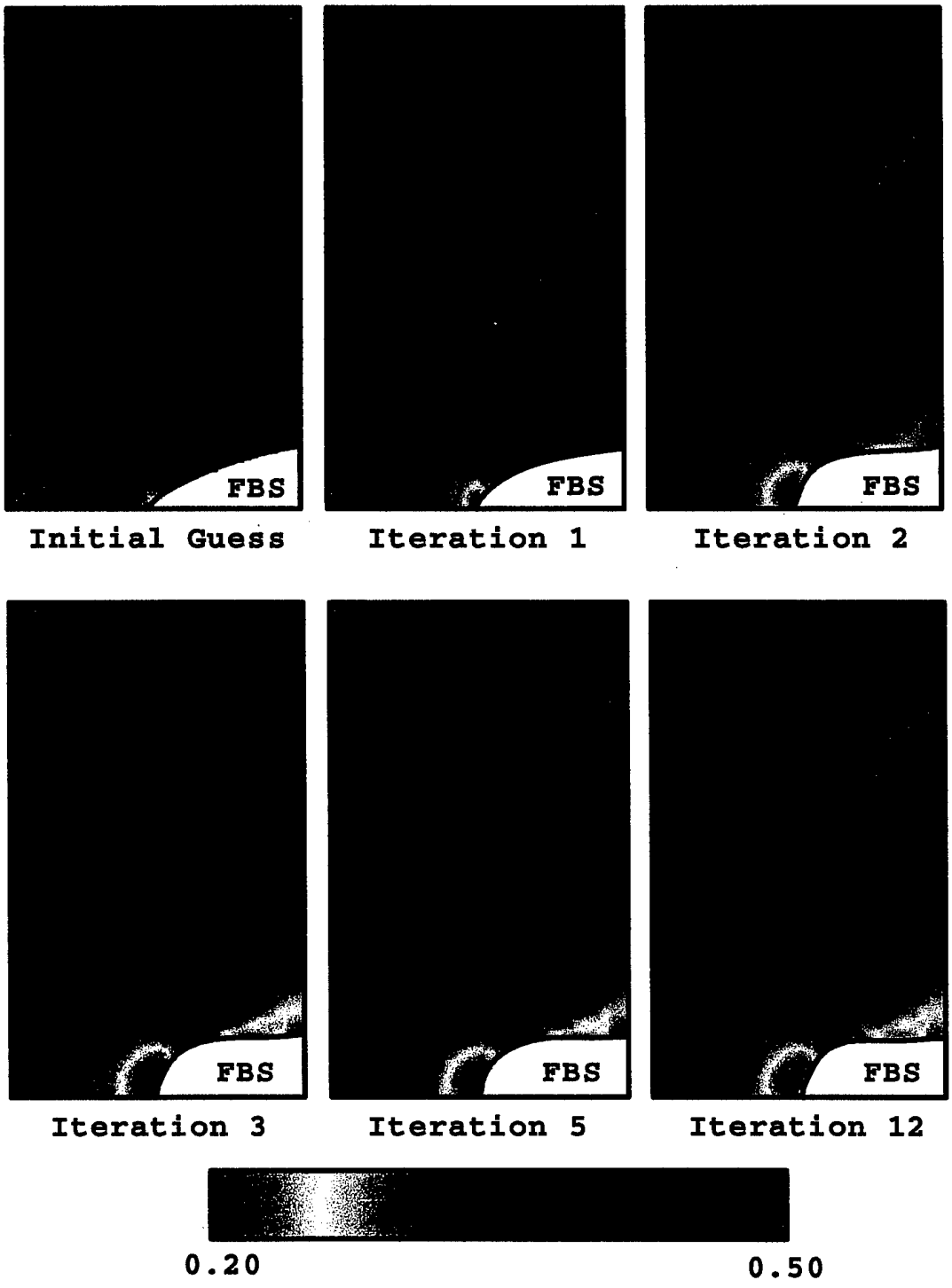


Figure 4.35: Forebody Design Iterations

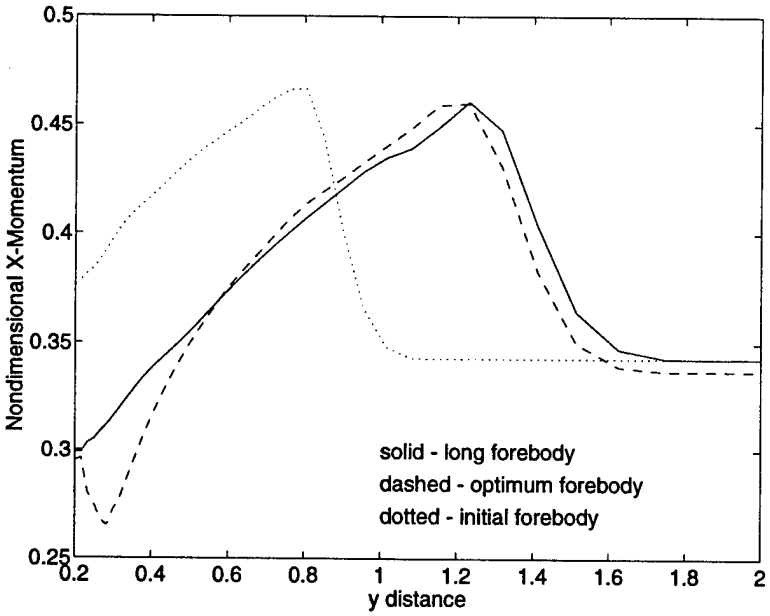


Figure 4.36: Outflow X-Component of Momentum

4.7 Asymptotically Consistent Gradients in Optimal Design

Here, the *sensitivity equation method* for approximating optimal design problems, introduced in the previous section is further analyzed. Recall that the method couples a trust-region optimization algorithm with gradients obtained by approximating the sensitivity equation. It has been shown that this method has several advantages over standard design sensitivity based methods, especially for shape optimization problems. However, derivatives obtained by approximating the sensitivity equation are not necessarily consistent and hence the optimization loop can be based on inaccurate gradient information. In the present research, we introduce the notion of *asymptotic consistency*. Convergence of the approximate design problem is considered for the case where asymptotically consistent sensitivities are used.

Optimal design problems can often be stated as finding parameters q from a prescribed set $\mathcal{Q}$ which minimize a design objective function $\mathcal{J}$ subject to a constraint on the states y of a system which is described by a differential equation. Namely, find q_* which satisfies

$$\mathcal{J}(q_*, y(q_*)) = \min_{q \in \mathcal{Q}} \mathcal{J}(q, y(q)) \tag{4.284}$$

subject to

$$\mathcal{E}(q, y(q)) = 0. \tag{4.285}$$

Many approximation methods for this problem are developed by cascading a simulation scheme for the states in an optimization algorithm. Design sensitivities can be used to compute the objective function gradients efficiently. An issue that often arises in this development is that numerical approximations of the sensitivities can produce gradients that are not the same as the gradient of the numerically approximated objective function, hence producing "non-consistent" gradients. Since the optimization algorithm is applied to the approximate model, it is usually assumed that the gradient which is provided to the optimization algorithm is the gradient of the approximate objective function. However, when the approximation of the constraint becomes parameter dependent (e.g. in shape optimization problems), the gradient needs to capture the sensitivity of the truncation error of the approximation. To do this the derivative of the discretization (mesh sensitivity) is required.

In last year's report, we introduced an efficient method for computing the gradient of the approximate objective function which does not require a mesh sensitivity. In this method, an equation for the sensitivities (the sensitivity equation) is obtained by differentiating the infinite dimensional constraint equation, which is then discretized. This method, however, does not always produce consistent gradients and the question of convergence of the optimization loop does not have a simple answer. One might expect that if the gradients were "close enough" that a convergent algorithm could be obtained (this conjecture is supported by the numerical experiments presented in last year's report). This motivates our definition of asymptotically consistent gradients.

In this work, we use a model optimization problem where the constraint is given by a one dimensional wave equation in order to illustrate these concepts. In particular, we provide a convergence result for the sensitivity equation method, which combines a trust-region optimization algorithm with asymptotically consistent gradients provided by approximating the sensitivity equation. Finally, we use numerical simulations to demonstrate that the sensitivity equation approach produces viable gradients for this problem.

4.7.1 Model Problem

To illustrate the notion of asymptotic consistency in a shape optimization setting, we introduce a model design problem with the constraint given by the one dimensional wave equation.

Problem 5 *Given* $\hat{y}(\cdot) \in L^2(0, 1.25)$ *and* $T > 0$, *find* $q_* \in [.5, 1.25]$ *which minimizes*

$$\mathcal{J}(q) = \int_0^{.5} (y(T, x; q) - \hat{y}(x))^2 \, dx, \tag{4.286}$$

where $y(t,x;q)$ is the solution of

$$\frac{\partial^2}{\partial t^2}y(t,x;q) = a^2\frac{\partial^2}{\partial x^2}y(t,x;q) \tag{4.287}$$

subject to

$$y(t,0;q) = 0, \qquad y(t,q;q) = 0, \tag{4.288}$$

$$y(0,x;q) = f(x;q) \qquad \textit{and} \qquad \frac{\partial}{\partial t}y(0,x;q) = 0 \tag{4.289}$$

on $[0,T]\times[0,q]$. Here we assume that $f(\cdot;q)\in H_0^1(0,q)$.

Note that $\mathcal{J}$ is considered as an explicit function of q through the solution of the wave equation. The above problem is a shape optimization problem since the domain depends on the design parameter q. Since we are interested in using computational techniques to solve this problem, we present the approximate problem below, beginning with an approximation of the wave equation.

Finite difference methods for approximating PDE's often use a transformation to simplify the difference algorithm in the case of irregularly shaped domains or non-uniform discretizations. While it is not necessary to introduce a transformation to approximate (4.287), we do so to illustrate the more general case. Consider a transformation

$$\mathcal{M} : \xi \to x, \tag{4.290}$$

mapping $[0,1]\to[0,q]$ and assume that $\mathcal{M}$ is an isomorphism from the "computational domain" $[0,1]$ to the "physical domain" $[0,q]$ in such a way that the discretization of the physical domain transforms to a uniform discretization of the computational domain. If we define

$$w(t,\xi;q) = y(t,\mathcal{M}(\xi;q);q), \tag{4.291}$$

then the wave equation can be transformed to the following equation

$$\begin{aligned}\frac{\partial^2}{\partial t^2}w(t,\xi;q) &= a^2\frac{\partial^2}{\partial\xi^2}w(t,\xi;q)\left[\frac{\partial}{\partial\xi}\mathcal{M}(\xi;q)\right]^{-2} \\ &\quad -a^2\frac{\partial}{\partial\xi}w(t,\xi;q)\frac{\partial^2}{\partial\xi^2}\mathcal{M}(\xi;q)\left[\frac{\partial}{\partial\xi}\mathcal{M}(\xi;q)\right]^{-3}\end{aligned} \tag{4.292}$$

with

$$w(t,0;q) = 0, \qquad w(t,1;q) = 0, \tag{4.293}$$

$$w(0,\xi;q) = f(\mathcal{M}(\xi;q);q) \qquad \text{and} \qquad \frac{\partial}{\partial t}w(0,\xi;q) = 0, \tag{4.294}$$

for $t\in[0,T]$ and $\xi\in[0,1]$. Here we have assumed $\frac{\partial}{\partial\xi}\mathcal{M}(\xi;q)\neq 0$ for $0\le\xi\le 1$.

A straight forward approximation scheme for $w(t,\xi;q)$ is given. Consider a uniform discretization of $[0,T]\times[0,1]$ with K points in the t-direction and N points in the ξ-direction, i.e.

$$\Delta t = \frac{T}{K-1} \qquad \text{and} \qquad \Delta\xi = \frac{1}{N-1}. \tag{4.295}$$

Define

$$w_{i,j} = w\left((i-1)\Delta t,(j-1)\Delta\xi;q\right), \tag{4.296}$$

$$\mathcal{M}_j = \mathcal{M}((j-1)\Delta\xi;q), \tag{4.297}$$

$$\mathcal{M}'_j = \frac{\partial}{\partial\xi}\mathcal{M}\left((j-1)\Delta\xi;q\right) \tag{4.298}$$

$$\text{and} \qquad \mathcal{M}''_j = \frac{\partial^2}{\partial\xi^2}\mathcal{M}\left((j-1)\Delta\xi;q\right), \tag{4.299}$$

for $(i,j) \in \{1,\ldots,K\} \times \{1,\ldots,N\}$ and consider the following finite difference scheme,

$$\frac{w_{i+1,j} - 2w_{i,j} + w_{i-1,j}}{\Delta t^2} = a^2 \frac{w_{i,j+1} - 2w_{i,j} + w_{i,j-1}}{\Delta\xi^2 \left[\mathcal{M}_j'\right]^2} - a^2 \frac{w_{i,j+1} - w_{i,j-1}}{2\Delta\xi \left[\mathcal{M}_j'\right]^3} \mathcal{M}_j''. \tag{4.300}$$

The boundary conditions are $w_{i,1} = w_{i,N} = 0$, for $i = 1,\ldots,K$, $w_{1,j} = f(\mathcal{M}_j; q)$ and $w_{0,j} = w_{2,j}$ for $j = 1,\ldots,N$. It can be shown that for a uniform discretization, i.e.

$$\mathcal{M} = q\xi, \tag{4.301}$$

the above scheme is convergent provided

$$\frac{a\Delta t}{q\Delta\xi} \leq 1. \tag{4.302}$$

This places a restriction on the behavior of K relative to N. Thus we will define the approximation at the Nth level by

$$w^N = \left\{w_{i,j}\right\}_{(i,j)=(1,1)}^{(K(N),N)}. \tag{4.303}$$

The approximate solution y^N to the wave equation is obtained by applying the inverse mapping $\mathcal{M}^{-1}$ to w^N. We now define the approximate optimization problem.

Problem 6 *Given data $\hat{y}$ and $T > 0$, find $q_* \in [.5, 1.25]$ which minimizes*

$$\mathcal{J}_g^N(q) = \sum_{i=1}^{g} c_i \left(y^N(T, x_i; q) - \hat{y}(x_i)\right)^2 \tag{4.304}$$

where y^N is the approximate solution to the wave equation as described above.

Note that the integral in (4.286) has been replaced by a quadrature $\{(c_i, x_i)\}_{i=1}^g$. Also note that y^N depends on both q and $\mathcal{M}(\cdot; q)$.

4.7.2 Sensitivity Equation Method

Design Sensitivities

In order to solve approximate optimization problems such as the one above, a gradient based optimization algorithm is frequently used. Thus, we need to consider methods for computing the gradient of $\mathcal{J}_g^N$. A straight forward approach is to use a finite difference approximation, i.e.

$$\frac{\partial}{\partial q}\mathcal{J}_g^N(q) \approx \frac{\mathcal{J}_g^N(q + \Delta q) - \mathcal{J}_g^N(q)}{\Delta q}. \tag{4.305}$$

Unfortunately, this approach is not practical in problems where the approximation of the PDE is computationally expensive. One way of alleviating the computational burden is to use design sensitivities, quantities which describe the influence of the design variables on the state variables. For example, we can directly compute the gradient by differentiating (4.304) as

$$\frac{\partial}{\partial q}\mathcal{J}_g^N(q) = 2\sum_{i=1}^{g} c_i \left(y^N(T, x_i; q) - \hat{y}(x_i)\right) \frac{\partial}{\partial q} y^N(T, x_i; q). \tag{4.306}$$

The quantity $\frac{\partial}{\partial q} y^N$ is the design sensitivity for the discretized state y^N.

There are several ways to compute this sensitivity. As above, one might use finite differences, yielding the approximation

$$\frac{\partial}{\partial q} y^N(T, x; q) \approx \frac{y^N(T, x; q + \Delta q) - y^N(T, x; q)}{\Delta q}. \tag{4.307}$$

It is usually more practical to compute this approximation by using

$$\begin{aligned}\frac{\partial}{\partial q}y^N(T,x;q) \approx & \frac{y^N(T,x+\frac{\partial}{\partial q}\mathcal{M}(\mathcal{M}^{-1}(x))\Delta q;q+\Delta q)-y^N(T,x;q)}{\Delta q} \\ & -\frac{\partial}{\partial x}y^N(T,x;q)\frac{\partial}{\partial q}\mathcal{M}(\mathcal{M}^{-1}(x))\end{aligned} \tag{4.308}$$

in order to avoid interpolating back to the unperturbed mesh.

This approach has the advantage that a step size Δq can be selected using error estimates for y^N. However, it is as computationally expensive as computing finite differences on $\mathcal{J}_g^N$. A more efficient approach can be obtained by differentiating the simulation scheme used to approximate the states. We demonstrate this using the scheme for the one dimensional wave equation given above (4.300).

By (implicitly) differentiating (4.300) with respect to q, and defining

$$u_{i,j} = \frac{\partial}{\partial q}w_{i,j}, \tag{4.309}$$

we achieve the following scheme for $u^N = \left\{u_{i,j}\right\}_{(i,j)=(1,1)}^{(K,N)}$,

$$\begin{aligned}\frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{\Delta t^2} = & a^2\frac{u_{i,j+1}-2u_{i,j}+u_{i,j-1}}{\Delta\xi^2\left[\mathcal{M}_j'\right]^2}-a^2\frac{u_{i,j+1}-u_{i,j-1}}{2\Delta\xi\left[\mathcal{M}_j'\right]^3}\mathcal{M}_j'' \\ & -2a^2\frac{w_{i,j+1}-2w_{i,j}+w_{i,j-1}}{\Delta\xi^2\left[\mathcal{M}_j'\right]^3}\frac{\partial}{\partial q}\mathcal{M}_j' \\ & +3a^2\frac{w_{i,j+1}-w_{i,j-1}}{2\Delta\xi\left[\mathcal{M}_j'\right]^4}\mathcal{M}_j''\frac{\partial}{\partial q}\mathcal{M}_j' \\ & -a^2\frac{w_{i,j+1}-w_{i,j-1}}{2\Delta\xi\left[\mathcal{M}_j'\right]^3}\frac{\partial}{\partial q}\mathcal{M}_j''.\end{aligned} \tag{4.310, 4.311}$$

with boundary conditions given by

$$u_{i,1} = 0, \qquad u_{i,N} = 0, \tag{4.312}$$

for $i = 1,\ldots,K$ and

$$u_{1,j} = \frac{\partial}{\partial q}f(\mathcal{M}_j;q)+\frac{\partial}{\partial x}f(\mathcal{M}_j;q)\frac{\partial}{\partial q}\mathcal{M}_j \qquad \text{and} \qquad u_{0,j} = u_{2,j}, \tag{4.313}$$

for $j = 1,\ldots,N$. The design sensitivity can then be extracted from

$$\frac{\partial}{\partial q}y_{i,j} = u_{i,j}-\frac{\partial}{\partial x}y_{i,j}\frac{\partial}{\partial q}\mathcal{M}_j, \tag{4.314}$$

by differentiating equation (4.291).

Note that derivatives of the mapping $\mathcal{M}$ with respect to the parameter q are required to compute the last three terms in expression (4.310). In more complex problems, the evaluation of $\mathcal{M}$ can involve the solution of an elliptic PDE and thus be as expensive to compute as the state. Determining the derivative of this map would then also require applying the above technique to the mesh generation scheme that produces $\mathcal{M}$, adding additional expense. Frequently, the derivatives of this mapping are computed using some type of approximation. Besides the computational savings offered over finite differences, this discrete design sensitivity approach has another advantage in that it produces consistent derivatives.

Loosely speaking, a sensitivity approach produces *consistent derivatives* (with a given approximation of the state y^N) if it produces the sensitivities of the approximate state variables with respect to the design variables, $\frac{\partial}{\partial q} y^N$. Note that this implies that these sensitivities are used to compute $\frac{\partial}{\partial q} \mathcal{J}_g^N$ in equation (4.306). Thus, when consistent derivatives are used, the gradient which is sent to the optimization algorithm is consistent with the approximate objective function $\mathcal{J}_g^N$. That is, a sensitivity approach which produces consistent derivatives, such as the discrete sensitivity approach described above, captures the sensitivity of the mesh dependent round-off errors in $\mathcal{J}_g^N$.

Another approach is based on approximating the sensitivity equation, obtained by differentiating the (continuous) state equation. For the model problem considered above, this amounts to differentiating the one dimensional wave equation and its boundary conditions. Thus if one defines

$$s(t,x;q) = \frac{\partial}{\partial q} y(t,x;q), \tag{4.315}$$

then the sensitivity equation is given by

$$\frac{\partial^2}{\partial t^2} s(t,x;q) = a^2 \frac{\partial^2}{\partial x^2} s(t,x;q) \tag{4.316}$$

subject to

$$s(t,0;q) = 0, \qquad s(t,q;q) = -\frac{\partial}{\partial x} y(t,q;q), \tag{4.317}$$

$$s(0,x;q) = \frac{\partial}{\partial q} f(x;q) \qquad \text{and} \qquad \frac{\partial}{\partial t} s(0,x;q) = 0 \tag{4.318}$$

on $[0,T] \times [0,q]$. At this point, we are free to use any approximation scheme to find s^N, however, for this example, we use the same simulation scheme which was used to determine y^N. Thus, transforming the sensitivity equation using the change of variables

$$v(t,\xi;q) = s(t,\mathcal{M}(\xi;q);q), \tag{4.319}$$

one obtains the equation

$$\begin{aligned} \frac{\partial^2}{\partial t^2} v(t,\xi;q) &= a^2 \frac{\partial^2}{\partial \xi^2} v(t,\xi;q) \left[\frac{\partial}{\partial \xi} \mathcal{M}(\xi;q)\right]^{-2} \\ &\quad - a^2 \frac{\partial}{\partial \xi} v(t,\xi;q) \frac{\partial^2}{\partial \xi^2} \mathcal{M}(\xi;q) \left[\frac{\partial}{\partial \xi} \mathcal{M}(\xi;q)\right]^{-3}, \end{aligned} \tag{4.320}$$

with

$$v(t,0;q) = 0, \qquad v(t,1;q) = -\frac{\partial}{\partial \xi} w(t,1;q) \left[\frac{\partial}{\partial \xi} \mathcal{M}(1;q)\right]^{-1}, \tag{4.321}$$

$$v(0,\xi;q) = \frac{\partial}{\partial q} f(\mathcal{M}(\xi;q);q) \qquad \text{and} \qquad \frac{\partial}{\partial t} v(0,\xi;q) = 0 \tag{4.322}$$

on $[0,T] \times [0,1]$. Defining

$$v_{i,j} = v((i-1)\Delta t, (j-1)\Delta\xi; q), \tag{4.323}$$

the corresponding finite difference scheme is given by

$$\frac{v_{i+1,j} - 2v_{i,j} + v_{i-1,j}}{\Delta t^2} = a^2 \frac{v_{i,j+1} - 2v_{i,j} + v_{i,j-1}}{\Delta\xi^2 \left[\mathcal{M}'_j\right]^2} - a^2 \frac{v_{i,j+1} - v_{i,j-1}}{2\Delta\xi \left[\mathcal{M}'_j\right]^3} \mathcal{M}''_j, \tag{4.324}$$

with boundary conditions given by $v_{i,1} = 0$ and $v_{i,N} = -\frac{\partial}{\partial \xi} w_{i,N} \left[\mathcal{M}'_N\right]^{-1}$ for $i = 1, \ldots, K$, $v_{1,j} = \frac{\partial}{\partial q} f(\mathcal{M}_j; q)$ and $v_{0,j} = v_{2,j}$ for $j = 1, \ldots, N$.

In a similar fashion as for the wave equation, the approximate sensitivities $s^{N,M}$ can be found by applying $\mathcal{M}^{-1}$ to v^N. We use the notation N, M to indicate the dependence of the approximate sensitivities on the approximation scheme for the state (N), and the approximation scheme for the sensitivity equation (M). Clearly, the above finite difference scheme requires less computational effort than (4.310), furthermore, no mesh sensitivities are required.

It is important to point out that these two sensitivity approaches do not necessarily give the same values, i.e.

$$\frac{\partial}{\partial q} y^N \neq \left(\frac{\partial}{\partial q} y\right)^{N,N}. \tag{4.325}$$

Therefore, the sensitivity equation approach does not necessarily produce consistent derivatives. Thus, one must address the problem of convergence of the optimization algorithm. Consider the derivative of the infinite dimensional objective function,

$$\frac{\partial}{\partial q}\mathcal{J}(q) = 2\int_0^{.5} (y(T,x;q) - \hat{y}(x)) \frac{\partial}{\partial q} y(T,x;q) dx. \tag{4.326}$$

Using a similar approximation of this expression to that used to approximate $\mathcal{J}(q)$, we arrive at

$$\left(\frac{\partial}{\partial q}\mathcal{J}\right)_g^{N,M}(q) = 2\sum_{i=1}^{g} c_i \left(y^N(T,x_i;q) - \hat{y}(x_i)\right) \left(\frac{\partial}{\partial q} y\right)^{N,M}(T,x_i;q). \tag{4.327}$$

Therefore, the sensitivity equation approach offers an approximation of the gradient of $\mathcal{J}$, and for a sufficiently fine discretization we might expect this gradient to be "close enough." This motivates our definition of asymptotically consistent gradients below.

We can now make precise definitions.

Definition 7 *A sensitivity approach is said to produce* **consistent derivatives** *with respect to approximations N (for the states) and M (for the sensitivities) if*

$$\frac{\partial}{\partial q}\mathcal{J}_g^N(\cdot) = \left(\frac{\partial}{\partial q}\mathcal{J}\right)_g^{N,M}(\cdot). \tag{4.328}$$

This is exactly the case for the discrete sensitivity approach.

Definition 8 *A sensitivity approach is said to produce* **asymptotically consistent derivatives** *with respect to approximations N (for the states) and M (for the sensitivities) if it satisfies*

$$\left|\frac{\partial}{\partial q}\mathcal{J}_g^N(q) - \left(\frac{\partial}{\partial q}\mathcal{J}\right)_g^{N,M}(q)\right| \to 0, \qquad \forall q \in \mathcal{Q} \tag{4.329}$$

as the approximations N and M are refined.

For the wave equation problem, if the exact solution $y^{ex}(T,x;q)$ along with its first four derivatives in x are bounded on $[0,1.25]$ and the mapping $\mathcal{M}$ satisfies

$$\frac{\partial}{\partial \xi}\mathcal{M}(\xi;q) > \alpha \qquad \text{and} \qquad \frac{\partial^2}{\partial \xi^2}\mathcal{M}(\xi;q) < \beta \tag{4.330}$$

where $\alpha, \beta > 0$ (which guarantees that the mesh on $[0,q]$ is refined with refinements of the computational space, $[0,1]$), then

$$\left|\frac{\partial}{\partial q} y^N(T,x_i;q) - \frac{\partial}{\partial q} y^{ex}(T,x_i;q)\right| \to 0 \tag{4.331}$$

and

$$\left|\left(\frac{\partial}{\partial q} y\right)^{N,N}(T,x_i;q) - \frac{\partial}{\partial q} y^{ex}(T,x_i;q)\right| \to 0 \tag{4.332}$$

which implies that the sensitivity equation approach produces asymptotically consistent derivatives for this particular approximation of the sensitivity equation.

Trust-Region Optimization Algorithm

Trust-region optimization algorithms are designed to converge quickly from initial parameters which are out of the superlinear convergence region. This is accomplished by minimizing the local quadratic model only over a region where this model is "trusted." A benefit of using these algorithms is that they provide a robust optimization algorithm even when there are model inaccuracies. In particular, Carter (ICASE Report 89-45) shows that the algorithms can converge when neither the function nor its gradient values are known exactly. Therefore, it seems natural to couple a trust-region optimization algorithm with gradients provided by approximating the sensitivity equation. We denote this strategy by the *sensitivity equation method.* A convergence result is provided below.

Hypothesis 9 *For the results below, we assume there exists a positive constant, L, (independent of N) such that*

$$\left|\frac{\partial}{\partial q}\mathcal{J}_g^N(q_1) - \frac{\partial}{\partial q}\mathcal{J}_g^N(q_2)\right| \leq L\,|q_1 - q_2| \qquad \text{for all } q_1, q_2 \in \mathcal{Q}_0, \tag{4.333}$$

where, for a given $q_0 \in \mathcal{Q}$, $\mathcal{Q}_0$ is an open convex set containing the level set

$$\mathcal{L}_0 = \left\{q \in \mathcal{Q} | \mathcal{J}_g^N(q) \leq \mathcal{J}_g^N(q_0)\right\}. \tag{4.334}$$

Theorem 10 *Apply the sensitivity equation method to an approximate optimal design problem of the form in Problem 6. Assume $\mathcal{J}_g^N$ is continuously differentiable on an open convex set $\mathcal{Q}_0$ and satisfies Hypothesis 9. Then, if the approximate gradient $\left(\frac{\partial}{\partial q}\mathcal{J}\right)_g^{N,M}$ is asymptotically consistent to the gradient $\frac{\partial}{\partial q}\mathcal{J}_g^N$, the sensitivity equation method produces iterates $\{q_k\}_{k=1}^{\infty}$ which satisfy*

$$\liminf_{k\to\infty} \frac{\partial}{\partial q}\mathcal{J}_g^N(q_k) = 0. \tag{4.335}$$

The proof of this theorem uses asymptotic consistency to obtain the error bound

$$\left|\left(\frac{\partial}{\partial q}\mathcal{J}\right)_g^{N,M}(q_k) - \frac{\partial}{\partial q}\mathcal{J}_g^N(q_k)\right| \leq \zeta < 1, \tag{4.336}$$

which can be satisfied by selecting sufficiently fine discretizations N and M. The result follows from Carter's theorem.

4.7.3 Numerical Results

In this section, we present numerical results which verify the asymptotic consistency results for the wave equation design problem in Section 4.7.2. In Problem 6, we consider $\hat{y}(x) = \sin(2\pi x)$ and $T = 1$. The quadrature is given by a 5 point trapezoidal rule: $(c_1, x_1) = (.1, .05)$, $(c_2, x_2) = (.1, .15)$, $(c_3, x_3) = (.1, .25)$, $(c_4, x_4) = (.1, .35)$ and $(c_5, x_5) = (.1, .45)$. The numerical solution y^N is an approximation of the wave equation (4.287) with $a = 1$ and $f(x; q) = \sin\left(\frac{2\pi x}{q}\right)$. For this problem, the exact solution of the wave equation is

$$y_{ex}(t, x; q) = \sin\left(\frac{2\pi x}{q}\right)\cos\left(\frac{2\pi a t}{q}\right), \tag{4.337}$$

and the exact solution of the sensitivity equation (4.316) is

$$s_{ex}(t, x; q) = -\frac{2\pi x}{q^2}\cos\left(\frac{2\pi x}{q}\right)\cos\left(\frac{2\pi a t}{q}\right) + \frac{2\pi a t}{q^2}\sin\left(\frac{2\pi x}{q}\right)\sin\left(\frac{2\pi a t}{q}\right). \tag{4.338}$$

Thus, we can compute the gradient of the objective function (using quadrature) as

$$\left(\frac{\partial}{\partial q}\mathcal{J}\right)_g(q) = 2\sum_{i=1}^{5} c_i\left[y_{ex}(T, x_i; q) - \hat{y}(x_i)\right] s_{ex}(T, x_i; q) \tag{4.339}$$

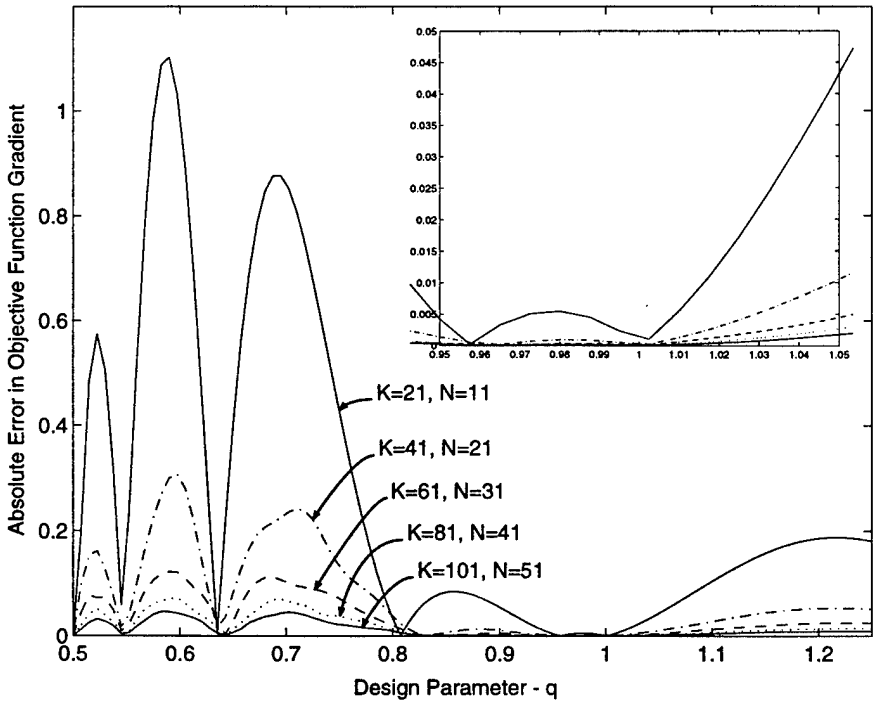


Figure 4.37: Error in Gradient Using Sensitivity Equation Approach: Uniform Mesh

using the above expressions. It is easily seen that the optimal parameter, q_*, is 1.

We now plot the absolute error in the gradient calculations using the sensitivity equation and the discrete sensitivity approaches discussed earlier. First of all, we plot the absolute error in the objective function gradient corresponding to a uniform mesh, $\mathcal{M}(\xi;q) = \xi q$. In Figure 4.37, we plot this for the sensitivity equation approach, i.e.

$$e_{se} = \left| \left(\frac{\partial}{\partial q} \mathcal{J} \right)_g (q) - \left(\frac{\partial}{\partial q} \mathcal{J} \right)_g^{N,M} (q) \right| \tag{4.340}$$

for various levels of refinement (which satisfy the stability condition (4.302)). In Figure 4.38, we present the corresponding plots for the discrete sensitivity approach, i.e.

$$e_{ds} = \left| \left(\frac{\partial}{\partial q} \mathcal{J} \right)_g (q) - \frac{\partial}{\partial q} \mathcal{J}_g^N(q) \right| \tag{4.341}$$

One observes that the error in the gradient calculations is diminishing as the mesh is refined, leading to the conclusion that the gradients computed using the sensitivity equation approach are asymptotically consistent. We point out that an example can be given where the sensitivity equation approach produces consistent derivatives, but $\frac{\partial}{\partial q}\mathcal{J}_g^N(\cdot) \not\to \frac{\partial}{\partial q}\mathcal{J}(\cdot)$.

The figures were constructed by evaluating the gradients for 101 equally spaced points in the interval $[.5, 1.25]$ (denoted by P). Asymptotic consistency is also seen by evaluating

$$D(q) = \left| \frac{\partial}{\partial q} \mathcal{J}_g^N(q) - \left(\frac{\partial}{\partial q} \mathcal{J} \right)_g^{N,M} (q) \right| \tag{4.342}$$

at these points. Table 4.10 also clearly suggests asymptotic consistency.

The same plots were made for a nonuniform mesh corresponding to $\mathcal{M}(\xi;q) = .7\xi(1-\xi)q^2 + \xi q$. This mapping satisfies conditions (4.330), therefore we know that the sensitivity equation approach

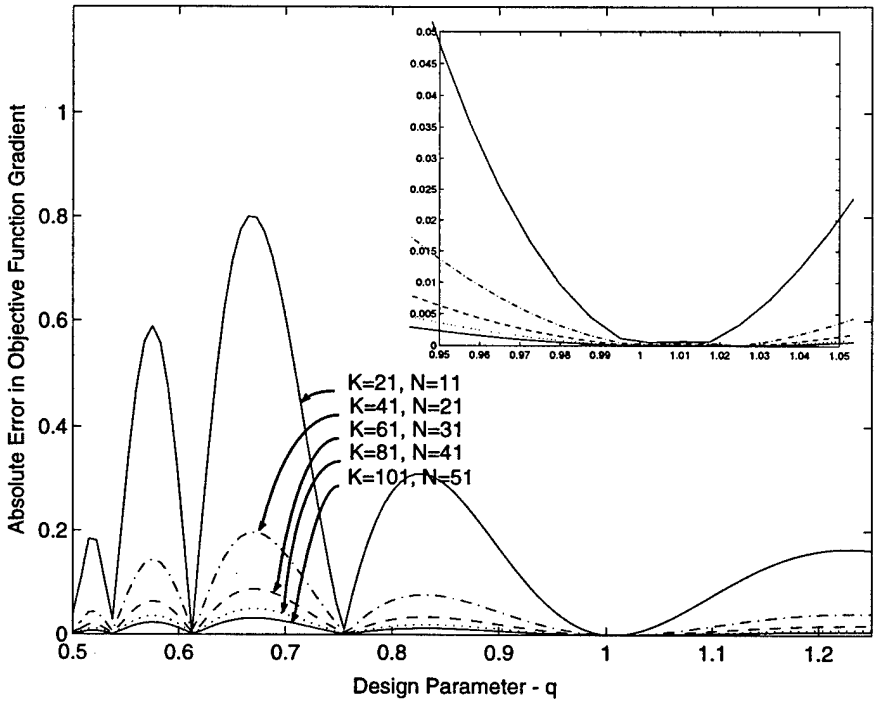


Figure 4.38: Error in Gradient Using Discrete Sensitivity Approach: Uniform Mesh

Table 4.10: Asymptotic Consistency for Sensitivity Equation Approach

N	11	21	31	41	51	61	71	81	91
K	21	41	61	81	101	121	141	161	181
$\max_P D$	.7138	.2300	.0967	.0507	.0336	.0227	.0169	.0131	.0105

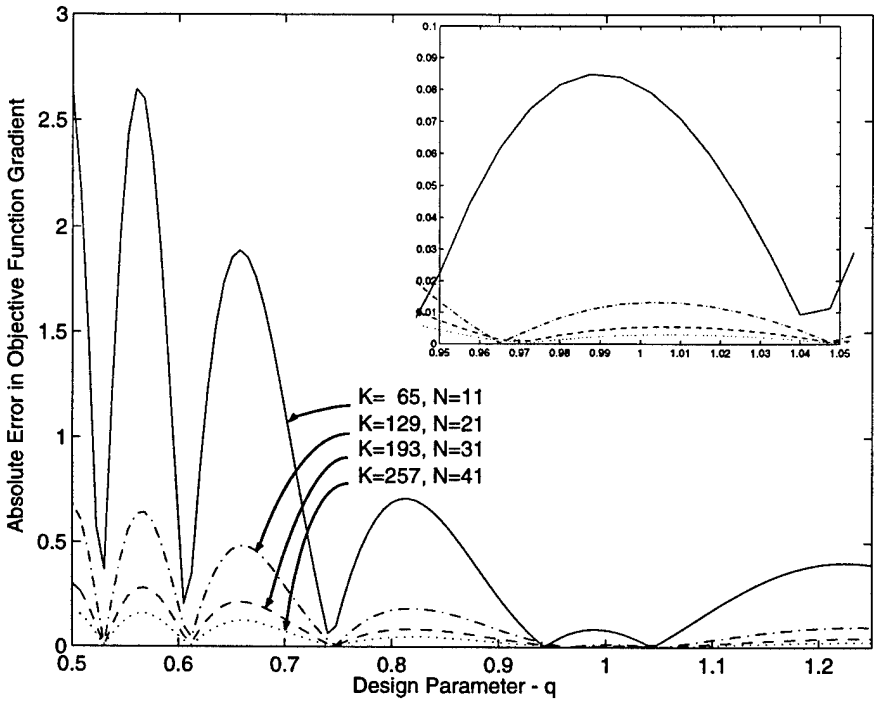


Figure 4.39: Error in Gradient Using Sensitivity Equation Approach: Non-Uniform Mesh

produces asymptotically consistent derivatives. Figures 4.39 and 4.40 show that the gradients $\left(\frac{\partial}{\partial q}\mathcal{J}\right)_g^{N,M}$ and $\frac{\partial}{\partial q}\mathcal{J}_g^N$ are both converging to $\left(\frac{\partial}{\partial q}\mathcal{J}\right)_g$ which verifies asymptotic consistency.

4.7.4 Conclusions

In this section, we presented the sensitivity equation method for optimal design. This method is efficient and does not require mesh sensitivities. Convergence of this method was studied by introducing the notion of asymptotic consistency. We demonstrated these ideas using a model design problem with a constraint given by the one dimensional wave equation.

The sensitivity equation method has been successfully used to approximate optimal design problems corresponding to shape optimization of a forebody simulator (where the constraint is given by the two dimensional Euler equations) and a channel design problem (where the constraint is given by the two dimensional incompressible Navier-Stokes equations).

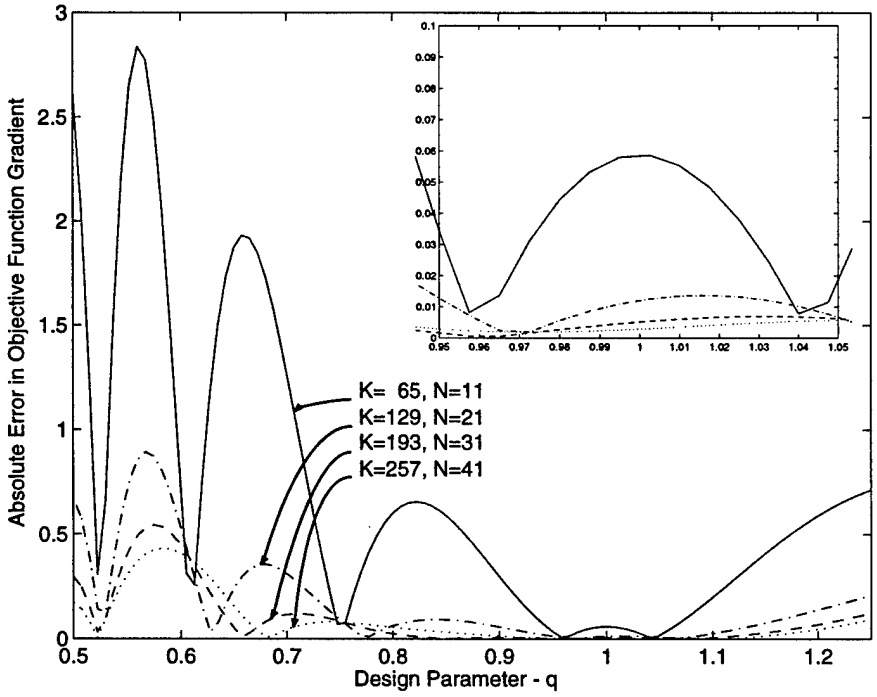


Figure 4.40: Error in Gradient Using Discrete Sensitivity Approach: Non-Uniform Mesh

4.8 Reduced Hessian Methods for Design

The purpose of this study is to implement sequential quadratic programming (SQP) algorithms to find solutions to the inverse design problem for a one-dimensional nozzle flow. The flow of the fluid within the nozzle is governed by the Euler equations, and is transonic in the sense that the flow is assumed to be supersonic at the inlet to the nozzle, and is subsonic at the outlet, with a shock existing in the flow. In a nonlinear programming (NLP) framework, the state variables are the (discrete) flow variables, which are then governed by the (discretized) Euler equations. The control variables are the design parameters that shape the area of the duct. The problem is to find an "optimal" shape for the duct, which will match a desired flow profile within the duct.

SQP methods are iterative, and treat controls and states as independent variables. The nonlinear governing equations, in this case the Euler equation, are enforced as part of the constraints and are only satisfied in the limit, as the iteration converges to a feasible solution. The difficulties in the application of SQP methods to problems, such as the one described above, are the occurence of shocks and, for two or three space dimensions, the size of the problem. We use an SQP framework based on a reduced Hessian that exploits the problem structure. These SQP methods are well suited for large scale optimal control problems. The SQP methods described there include a trust region globalization to guarantee convergence of the iteration from arbitrary starting points and to enhance its robustness. Moreover, these methods use an affine scaling interior point strategy to handle the bounds on q. The SQP methods incorporate the general structure of optimal control problems like (4.415) and allow for inexact solutions of the quadratic subproblems. In particular, structure in the linearized state equation can be easily incorporated. Moreover the implementation of these SQP algorithms allows the use of weighted scalar products.

The realm of computational fluid dynamics (CFD) offers a number of methods for solving the above state equations, which can capture the shock implicitly (e.g. Godunov scheme). The aim is to study the effect of some of these schemes on the optimization method. This design problem is difficult to solve numerically, because the duct flow has a shock. Many good numerical shock capturing schemes, such as the Godunov scheme have low continuity properties. On the other hand, efficient numerical optimization schemes require sufficiently smooth cost and constraint functions. A straightforward combination of off–the–shelf discretization schemes for the flow equations and of off–the–shelf optimization methods often leads to very unsatisfactory results. In our case, we found that the SQP method failed for converge for cases when the initial guess is far from the optimal solution. The failure of the SQP method is related to the presence of shocks in the flow. In particular, if shock capturing schemes with low continuity properties, like the Godunov scheme, are used, then the SQP method, which is designed for smooth problems, behaves poorly. This seeming incompatibility of good shock capturing schemes for the discretization of the flow equations and efficient SQP methods for the solution of the optimization problem motivated the study of a shock fitting scheme. The important extension is that we include the shock location as a state variable. The formulation gives a sharp shock and, since the shock location is an explicit variable, the map from the design parameters to the flow is differentiable. In this section, we give a rigorous analysis of the infinite dimensional design problem including existence of optimal designs, existence of Fréchet derivatives, and existence of Lagrange multipliers. In particular we will show that the co–state is discontinuous at the shock location, unless the target velocity can be matched perfectly. We discuss the discretized design problem and investigate the relation between the finite dimensional problem and the discretized one. The careful study of this relation gives valuable insight and reveals information that is shown to be important for the performance of the optimization algorithm.

4.8.1 One-Dimensional Nozzle Flow

The problem we consider is a one-dimensional flow in a duct. The flow is governed by the Euler equations

$$\mathcal{F}_x + \mathcal{G} = 0, \qquad 0 \leq x \leq 1, \tag{4.343}$$

where,

$$\mathcal{F} = \begin{pmatrix} (\rho u)A \\ (\rho u^2 + p)A \\ (\rho E + p)uA \end{pmatrix}, \quad \mathcal{G} = \begin{pmatrix} 0 \\ -pA_x \\ 0 \end{pmatrix},$$

Standard notation has been used, with ρ being the fluid density, u the velocity, $E = e + u^2/2$, where e is specific internal energy, and p is the fluid static pressure. The subscript x denotes differentiation with respect to the position along the streamwise direction, x, and $A(x)$ is a given distribution function of the cross-sectional area, which is assumed to be at least piecewise differentiable. It is assumed that the cross-sectional area $A(x)$ of the duct along the streamwise direction x is absolutely continuous and monotonically increasing:

$$A(x) > 0, \quad A_x(x) > 0, \quad x \in [0, 1]. \tag{4.344}$$

Under given conditions, equation (4.343) can be reduced to a single ordinary differential equation in u. We obtain,

$$(f(u))_x + g(u, A) = 0, \tag{4.345}$$

where

$$f(u) \equiv u + \bar{H}/u, \quad g(u, A) \equiv \frac{A_x}{A}(\bar{\gamma}u - \bar{H}/u), \tag{4.346}$$

and where

$$\bar{\gamma} = (\gamma - 1)/(\gamma + 1), \quad \bar{H} = 2H\bar{\gamma}$$

are given constants. The constant $\gamma > 1$ is the gas constant (for air, $\gamma = 1.4$), and the constant H is the total enthalpy. The flow is supersonic for $u > u_* \equiv \sqrt{\bar{H}}$ and subsonic for $u < u_*$.

In addition, we impose the following boundary conditions

$$u(0) = u_{\text{in}}, \quad u(1) = u_{\text{out}}. \tag{4.347}$$

We choose boundary data $u_{\text{in}} > u_* > u_{\text{out}}$ so that a solution u of (4.345), (4.347) has a jump from supersonic to subsonic at some point x_s. At the shock location x_s the flow is required to satisfy the Rankine–Hugoniot relation

$$f(u(x_s-)) = f(u(x_s+)) \tag{4.348}$$

or, equivalently,

$$u(x_s-) \cdot u(x_s+) = \bar{H}. \tag{4.349}$$

As usual, $u(x_s-) = \lim_{h \to 0+} u(x_s - h)$ and $u(x_s+) = \lim_{h \to 0+} u(x_s + h)$. Equation (4.345) along with the above conditions (4.347) and (4.348) defines the flow profile.

For sake of completeness, we review some existence results. While the existence result will not be needed in this form, the arguments applied for its proof give some important insight into the structure of the problem.

First we consider the initial value problems

$$(f(u))_x + g(u, A) = 0, \quad u(0) = u_{\text{in}} \tag{4.350}$$

and

$$(f(u))_x + g(u, A) = 0, \quad u(1) = u_{\text{out}}. \tag{4.351}$$

Since $f_u(u) > 0$ for $u > \sqrt{\bar{H}}$, there exists a solution of (4.350) in a neighborhood $[0, x_L)$ of $x = 0$ provided $u_{\text{in}} > \sqrt{\bar{H}}$. If $u_{\text{in}} \in (\sqrt{\bar{H}}, \sqrt{2\bar{H}})$ and $u(1) = u_{\text{out}} < \sqrt{\bar{H}}$, then (4.344) and the definitions of f, g imply that

$$u_x(x) = -g(u(x), A(x))/f_u(u(x)) \begin{cases} > 0 & x = 0, \\ < 0 & x = 1. \end{cases}$$

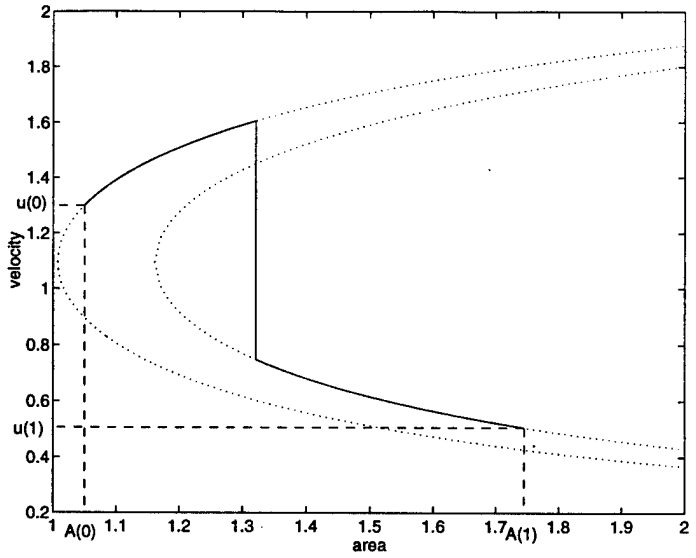


Figure 4.41: Sketch of the velocity as a function of the area.

Using the continuity of the solution and bootstrapping, we can deduce that unique solutions of (4.350) and (4.351) exist on some interval $[0, x_L)$ and $(x_R, 1]$, respectively. Moreover, the solutions are monotonically increasing and decreasing, respectively. It is easy to verify that the solutions are implicitly given by

$$A(x)u(x)\Big(2H - u^2(x)\Big)^r = \begin{cases} K_L, & x \in [0, x_L), \\ K_R, & x \in (x_R, 1], \end{cases} \tag{4.352}$$

where $r = 1/(\gamma - 1)$ and where the constants K_L, K_R are determined from $A(0)$, $u(0) = u_{\text{in}}$ and $A(1)$, $u(1) = u_{\text{out}}$. Due to the restrictions on the boundary conditions and due to the fact that $A(x) > 0$, the constants K_L, K_R are positive. Equation (4.352) defines two functions $A_L(u), A_R(u)$. It can be easily verified that these functions have a minimum at $u_* = \sqrt{\bar{H}}$ and are strictly monotone on $(0, \sqrt{\bar{H}}]$ and on $[\sqrt{\bar{H}}, \sqrt{2H})$. Thus, the initial condition $u_{\text{in}} \in (\sqrt{\bar{H}}, \sqrt{2H})$ guarantees that the solution u of (4.350) exists on $[0, x_L) = [0, \infty)$. The point x_R is the uniquely defined point satisfying $A(x_R) = A_R^*$. The situation is sketched in Figure 4.41.

Using (4.349), (4.352), and the continuity of A, the shock position x_s can be characterized by

$$w(u(x_s-)) = \frac{1}{K_L} u(x_s-)\left(2H - u^2(x_s-)\right)^r - \frac{1}{K_R}\frac{\bar{H}}{u(x_s-)}\left(2H - \left(\frac{\bar{H}}{u(x_s-)}\right)^2\right)^r = 0. \tag{4.353}$$

It is easy to see that $\lim_{u \to \sqrt{2H}-} w(u) < 0$. Hence, given A there exists a boundary conditions $u_{\text{in}} \in (\sqrt{\bar{H}}, \sqrt{2H})$, $u_{\text{out}} \in (0, \sqrt{\bar{H}})$, i.e. K_L, K_R, such that $\lim_{u \to u_{\text{in}}+} w(u) > 0$. In this case there exists $u(x_s-)$ such that $w(u(x_s-)) = 0$. Since the area A is monotonically increasing, the shock condition can then be computed from (4.352). Thus, we can conclude the following result:

Theorem 11 *Suppose the area function satisfies (4.344). Then there exist boundary conditions $u_{\text{in}} \in (\sqrt{\bar{H}}, \sqrt{2H})$ and $u_{\text{out}} \in (0, \sqrt{\bar{H}})$ such that the equations (4.345), (4.348), and (4.347) admit a unique solution u which is supersonic and monotonically increasing on $(0, x_s)$ and subsonic and monotonically decreasing on $(x_s, 1)$. Moreover, it obeys the inequalities*

$$\begin{aligned} \sqrt{\bar{H}} < u_{\text{in}} \leq \quad u(x) \quad &< \min\left\{\sqrt{2H}, \bar{H}/u_{\text{out}}\right\}, \quad x \in [0, x_s), \\ u_{\text{out}} \leq \quad u(x) \quad &< \bar{H}/u_{\text{in}} < \sqrt{\bar{H}}, \qquad\qquad x \in (x_s, 1]. \end{aligned} \tag{4.354}$$

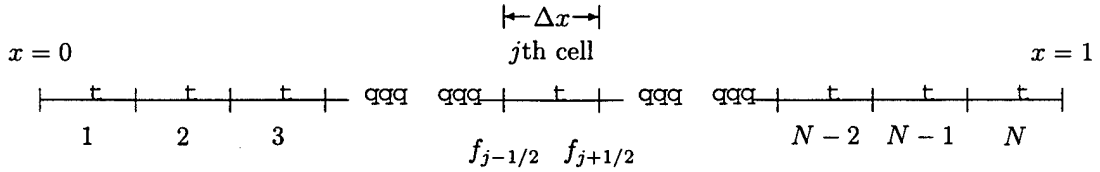


Figure 4.42: Grid Setup

The problem we are generally interested in is the design problem, wherein we are given some desired velocity profile, and the aim is to determine the area profile for the duct that will "most closely" generate such a flow. The problem can be stated thus:

$$\min_A J = \int_0^1 (u(x) - u^d(x))^2 dx \tag{4.355}$$

subject to constraints (4.345), (4.347) and (4.349).

The nature of the duct flow solution closely resembles the flow phenomena for two-dimensional inviscid flow over an airfoil. The one-dimensional transonic duct flow problem, hence, serves a useful purpose in that it provides valuable insight into the nature of the relatively complex problem of airfoil design.

4.8.2 Shock-Capturing Scheme

In the finite dimensional case, we discretize the flow, and use the discrete values of the flow at various grid points as our state. The domain $(0,1)$ is divided into N equispaced grid points, as shown in figure 4.42, yielding N subdomains, each of which must satisfy the state constraint. We have,

$$y = u = \{u_j\} \qquad j = 1, \ldots, N$$

We use a cell-centered finite volume formulation. For the jth cell (subdomain), we have,

$$f_x \approx \frac{f_{j+1/2} - f_{j-1/2}}{\Delta x}$$

This yields the following constraints

$$C(j) = \frac{f_{j+1/2} - f_{j-1/2}}{\Delta x} + g_j \qquad j = 1, \ldots, N \tag{4.356}$$

Using first order accurate interpolation, we have, for the supersonic region

$$f_{j+1/2} \approx f_j; \text{ and } f_{j-1/2} = f_{j-1}, \qquad u_j > u_s$$

and, for the subsonic region

$$f_{j+1/2} = f_{j+1}; \text{ and } f_{j-1/2} = f_j, \qquad u_j < u_s$$

where, $u_s = \sqrt{\overline{H}}$ is the sonic velocity. We use a Godunov scheme to capture the shock. This uses the following logic,

if $\quad ((u_j > u_s) \,\&\, (u_{j+1} < u_s))$ then

$\quad f_{j+1/2} = \max(f_j, f_{j+1})$

endifelseif $\quad ((u_{j-1} > u_s) \,\&\, (u_j < u_s))$ then

$\quad f_{j-1/2} = \max(f_{j-1}, f_j)$

endif

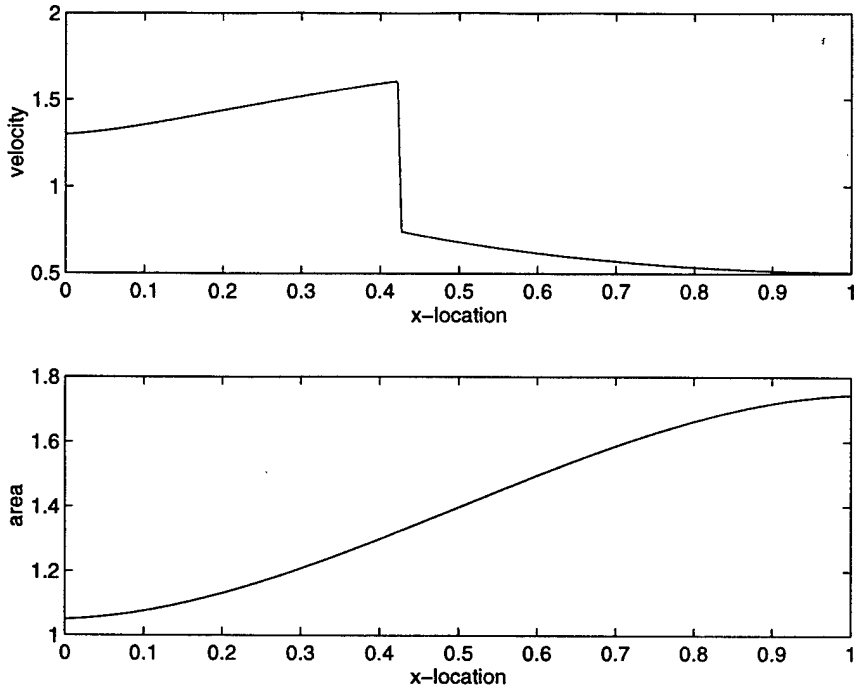


Figure 4.43: Target Data

which implicitly forces the shock condition, $f(u(x_s-)) = f(u(x_s+))$. The above system of equations yields a tridiagonal matrix for the Jacobian of the constraints. The objective function we use is,

$$J = \sum_{i=1}^{N} (u_i - u_i^d)^2 \tag{4.357}$$

Numerical Results

The area profile is parametrized in terms of $A(0), A_x(0), A(1)$ and $A_x(1)$. We use a cubic hermite polynomial, defined by the afore-mentioned parameters, to represent the area distribution. We use the following values, to compute a target solution, u^d:

$$\begin{aligned} u(0) &= 1.299 \\ A(0) &= 1.05 \\ A_x(0) &= 0.1 \\ u(1) &= 0.506 \\ A(1) &= 1.745 \\ A_x(1) &= 0.1 \end{aligned}$$

The target flow solution and the area profile are shown in figure 4.43. For the SQP runs, we use the following as our control variables:

$$q = [A_x(0) \quad A_x(1)]$$

Results, comparing the computational effort using both a *full* and *reduced* SQP algorithm, for a grid size of $N = 400$, are given in Table 4.11. An initial guess of $A_x(0) = A_x(1) = 0.09$ is

used for the parameters, and the initial flow variables are estimated to be equal to the target data. Additionally, results obtained using a *Black–box* method are also shown, for comparison. The *Black–box* method uses NLPQL, an SQP-based algorithm, with finite central differencing used to obtain gradient information. The constraints, $C = 0$, are solved using Newton's method, for a given area distribution. It was found, overall that the number of discrete evaluations of the flux, f, made the biggest contribution to the overall computational effort,and this has been used as a benchmark in quantifying the computational effort.

Table 4.11: Comparison of Computational Efforts

	Blackbox	*Full SQP*	*Reduced SQP*
# QP Iterations	5	6	7
# Flux Evaluations	7897456	270505	212588
# Constraint Evaluations	2224	29	36
# Gradient Evaluations	2192	81	62

It can be seen that the *Black–box* method is outperformed by the other methods. While the *reduced* and *full* SQP schemes consume about the same amount of computational time, the *Black–box* scheme requires about 10 times as much time to produce a converged solution. While the *reduced* SQP method outperforms the *full* SQP procedure in this case, it was found that, as the initial guess for the parameters is moved further away from the optimal value, the *reduced* SQP method requires far more iterations to converge. The *reduced* SQP method is not very robust, compared to the other two methods. Analysis showed that the difficulty arises from the lack of differentiability of the shock-capturing scheme. Using such schemes yields ill-conditioned Jacobians for the state constraints, which produce spikes in the numerical solutions. This hinders the convergence of the algorithm. An alternative, which is discussed in the following section, is to use a shock-fitting scheme.

4.8.3 A Shock Fitting Scheme

In the following we will denote the logarithmic derivative of A by q,

$$q(x) \equiv \frac{A_x(x)}{A(x)} = \frac{d}{dx}\ln(A(x)). \tag{4.358}$$

With this substitution the state equation is given by

$$(f(u))_x + g(u,q) = 0, \tag{4.359}$$

and (4.349), (4.347), where

$$g(u,q) \equiv q\,(\bar{\gamma}u - \bar{H}/u). \tag{4.360}$$

Note that instead of introducing another symbol we redefine g. Unless stated otherwise, in the following $g(u,q)$ is always given by (4.360).

Trivially, q is determined by A. On the other hand, if the area is known at some point, for example, if $A(0) > 0$ is given, then $A(x)$ can be computed from q by integrating (4.358):

$$A(x) = \exp\left(\ln(A(0)) + \int_0^x q(t)\,dt\right). \tag{4.361}$$

The function $A(x)$ defined by (4.361) is absolutely continuous and, therefore, differentiable almost everywhere. From now on, we assume that $A(0) > 0$ is given.

For a rigorous treatment of the dependence of the solution upon parameters, we have to transform the ODE, the shock condition, and the boundary conditions. We denote the velocity left of the shock by u_L, the velocity right of the shock by u_R, and, as before, the shock location by x_s. If we perform the variable transformation $x \to x_s\xi$ left of the shock and $x \to 1-(1-x_s)\xi$ right of the shock and if we set

$$u_L(\xi) = u(x_s\xi), \quad u_R(\xi) = u(1-(1-x_s)\xi), \tag{4.362}$$

then we find that

$$\frac{d}{d\xi}u_L(\xi) = x_s\frac{d}{dx}u(x_s\xi), \quad \frac{d}{d\xi}u_R(\xi) = (x_s-1)\frac{d}{dx}u(1-(1-x_s)\xi). \tag{4.363}$$

Thus, in the new spatial variables the ODE (4.359), the shock condition (4.349), and the boundary conditions (4.347) become

$$(f(u_L))_\xi + x_s g(u_L, q_L) = 0, \quad \xi \in [0,1], \tag{4.364}$$

$$(f(u_R))_\xi + (x_s-1)g(u_R, q_R) = 0, \quad \xi \in [0,1], \tag{4.365}$$

$$f(u_L(1)) = f(u_R(1)), \tag{4.366}$$

and

$$u_L(0) = u_{\text{in}}, \quad u_R(0) = u_{\text{out}}. \tag{4.367}$$

The functions q_L and q_R in (4.364), (4.365) are defined by $q_L(\xi) = q(x_s\xi)$ and $q_R(\xi) = q(1-(1-x_s)\xi)$, respectively. In the following we view q_L and q_R as independent variables. To guarantee that the corresponding area function is monotonically increasing we have to impose the conditions

$$q_L(\xi) > 0, \quad q_R(\xi) > 0, \quad \xi \in [0,1].$$

The control problem we are interested in is the design of an area function generating a flow that best approximates a desired velocity in the least squares sense. Given a desired velocity $u^d \in L^2(0,1)$ and a point x_s we introduce

$$u_L^d(x_s;\xi) = u^d(x_s\xi), \quad u_R^d(x_s;\xi) = u^d(1-(1-x_s)\xi). \tag{4.368}$$

With the transformations (4.362), (4.368) the objective function is given by

$$\begin{aligned}\int_0^1 (u(x)-u^d(x))^2\,dx &= \int_0^{x_s}(u(x)-u^d(x))^2\,dx + \int_{x_s}^1 (u(x)-u^d(x))^2\,dx \\ &= x_s\int_0^1 (u_L(\xi)-u_L^d(x_s;\xi))^2\,d\xi + (1-x_s)\int_0^1 (u_R(\xi)-u_R^d(x_s;\xi))^2\,d\xi.\end{aligned}$$

Thus, using the transformation introduced in the previous section, the control problem we have to solve is given as follows:

$$\text{Minimize } J(u_L,u_R,x_s,q_L,q_R) \equiv \frac{x_s}{2}\int_0^1 (u_L(\xi)-u_L^d(x_s;\xi))^2\,d\xi + \frac{1-x_s}{2}\int_0^1 (u_R(\xi)-u_R^d(x_s;\xi))^2\,d\xi \tag{4.369}$$

subject to the equality constraints

$$(f(u_L))_\xi + x_s g(u_L, q_L) = 0, \quad \xi \in [0,1], \tag{4.370}$$

$$(f(u_R))_\xi + (x_s-1)g(u_R, q_R) = 0, \quad \xi \in [0,1], \tag{4.371}$$

$$f(u_L(1)) = f(u_R(1)), \tag{4.372}$$

$$u_L(0) = u_{\text{in}}, \quad u_R(0) = u_{\text{out}}, \tag{4.373}$$

and to the inequality constraints

$$0 \le x_s \le 1, \tag{4.374}$$

$$0 \le q_{\text{low}} \le q_L(\xi), q_R(\xi) \le q_{\text{upp}}, \quad \xi \in [0,1]. \tag{4.375}$$

We assume that the boundary conditions obey

$$u_{\text{in}} \in (\sqrt{\bar{H}}, \sqrt{2H}), \quad u_{\text{out}} \in (\sqrt{\bar{\gamma}\bar{H}}, \sqrt{\bar{H}}). \tag{4.376}$$

The states are given by the triple (u_L, u_R, x_s) and the controls are (q_L, q_R). The equations (4.370), (4.371), (4.372), (4.373) are the state equations.

As the control space we use

$$\mathcal{Q} = L^\infty([0,1]) \times L^\infty([0,1])$$

and we denote the set of admissible controls by

$$\mathcal{Q}_{ad} = \{(q_L, q_R) \in \mathcal{Q} \mid 0 \le q_{\text{low}} \le q_L(\xi), q_R(\xi) \le q_{\text{upp}} \text{ a.e. in } [0,1]\}.$$

The set of admissible controls is closed and convex. By a solution to (4.370) (or (4.371)) we mean an absolutely continuous function which satisfies (4.370) (or (4.371)) almost everywhere on $[0,1]$.

Using the arguments applied in the previous section we can establish the following result.

Lemma 4 *Suppose that u_{in} and u_{out} obey (4.376) and that $(q_L, q_R) \in \mathcal{Q}_{ad}$. If (u_L, u_R, x_s) with $x_s \in [0,1]$ is a solution of (4.370) to (4.373), then*

$$0 < u_{\text{out}} \le u_R(\xi) \le \frac{\bar{H}}{u_{\text{in}}} < \sqrt{\bar{H}} < u_{\text{in}} \le u_L(\xi) \le \frac{\bar{H}}{u_{\text{out}}} < \sqrt{2H}, \quad \xi \in [0,1], \tag{4.377}$$

and there exists $c > 0$ which depends only on $u_{\text{in}}, u_{\text{out}}$ and $q_{\text{low}}, q_{\text{upp}}$ such that

$$|(u_L)_\xi(\xi)| \le c, \quad |(u_R)_\xi(\xi)| \le c \quad \text{a.e. on } [0,1]. \tag{4.378}$$

Proof: The estimate (4.377) follows from the monotonicity properties of the solution and the Rankine–Hugoniot relation written in the form (4.349). See also the discussion in Section 4.8.3.

From (4.370) and (4.371) we find that

$$(u_L)_\xi(\xi) = -\frac{x_s g(u_L(\xi), q_L(\xi))}{f_u(u_L(\xi))} = -\frac{x_s q_L(\xi)(\bar{\gamma}u_L(\xi) - \bar{H}/u_L(\xi))}{1 - \bar{H}/u_L^2(\xi)} \quad \text{a.e. on } [0,1],$$

$$(u_R)_\xi(\xi) = -\frac{(x_s - 1) g(u_R(\xi), q_R(\xi))}{f_u(u_R(\xi))} = -\frac{(x_s - 1) q_R(\xi)(\bar{\gamma}u_R(\xi) - \bar{H}/u_R(\xi))}{1 - \bar{H}/u_R^2(\xi)} \quad \text{a.e. on } [0,1].$$

The function $\bar{H}/u - \bar{\gamma}u = (\bar{\gamma}/u)(2H - u^2)$ is monotonically decreasing in u and positive for $u \in (0, \sqrt{2H})$. Using the estimate (4.377) it can be seen that

$$|(u_L)_\xi(\xi)| \le \frac{q_{\text{upp}}(\bar{H}/u_{\text{in}} - \bar{\gamma}u_{\text{in}})}{1 - \bar{H}/u_{\text{in}}^2}, \quad |(u_R)_\xi(\xi)| \le \frac{q_{\text{upp}}(\bar{H}/u_{\text{out}} - \bar{\gamma}u_{\text{out}})}{u_{\text{in}}^2/\bar{H} - 1} \quad \text{a.e. on } [0,1].$$

△

Thus, the state space appropriate for this design problem is given by

$$\mathcal{U} = W^{1,\infty}([0,1]) \times W^{1,\infty}([0,1]) \times I\!R$$

In the following we simply write $W^{1,\infty}, L^\infty$ instead of $W^{1,\infty}([0,1]), L^\infty([0,1])$ and we set

$$||q||_\infty = \text{ ess sup}_{[0,1]}|q(\xi)|, \quad ||u||_{1,\infty} = ||u||_\infty + ||u_\xi||_\infty.$$

We note that the shock location x_s enters the design problem in the differential equations (4.370), (4.371) and in the objective function, see also (4.368). The functions u_L, u_R, q_L, and q_R do not depend explicitly on x_s, but are implicitly coupled with the shock location through the design problem, in particular through (4.370) and (4.371).

Theorem 12 *Suppose there exist* $x_s \in (0,1)$ *and* $\bar{u} \in \left(u_{\rm in}, \bar{H}/u_{\rm out}\right)$ *such that*

$$0 \le q_{\rm low} \le \min\left\{ \frac{\left(1-\frac{\bar{H}}{u_{\rm in}^2}\right)(\bar{u}-u_{\rm in})}{x_s\left(\frac{\bar{H}}{u_{\rm in}}-\bar{\gamma}u_{\rm in}\right)}\,,\ \frac{\left(\frac{\bar{u}^2}{\bar{H}}-1\right)\left(\frac{\bar{H}}{\bar{u}}-u_{\rm out}\right)}{(1-x_s)(\frac{\bar{H}}{u_{\rm out}}-\bar{\gamma}u_{\rm out})} \right\} \tag{4.379}$$

and

$$q_{\rm upp} \ge \max\left\{ \frac{\left(1-\bar{\gamma}\right)(\bar{u}-u_{\rm in})}{x_s\left(\frac{\bar{H}}{\bar{u}}-\bar{\gamma}\bar{u}\right)}\,,\ \frac{\left(\frac{\bar{H}}{u_{\rm out}}-1\right)\left(\frac{\bar{H}}{\bar{u}}-u_{\rm out}\right)}{(1-x_s)\left(\bar{u}-\bar{\gamma}\frac{\bar{H}}{\bar{u}}\right)} \right\}. \tag{4.380}$$

Then there exists an optimal control $(q_L^*, q_R^*) \in \mathcal{Q}$ *of (4.369) - (4.375).*

Proof: First, note that the conditions (4.376) guarantee that the min in the condition (4.379) is positive.

(i) Existence of feasible points: For given $x_s \in (0,1)$ and $\bar{u} \in \left(u_{\rm in}, \bar{H}/u_{\rm out}\right)$ we set

$$u_L(\xi) = u_{\rm in} + \xi(\bar{u}-u_{\rm in}), \quad u_R(\xi) = u_{\rm out} + \xi\Big(\frac{\bar{H}}{\bar{u}} - u_{\rm out}\Big),$$

and

$$q_L(\xi) = -\frac{\left(1-\frac{\bar{H}}{u_L^2(\xi)}\right)(\bar{u}-u_{\rm in})}{x_s\left(\bar{\gamma}u_L(\xi)-\frac{\bar{H}}{u_L(\xi)}\right)}, \quad q_R(\xi) = -\frac{\left(1-\frac{\bar{H}}{u_R^2(\xi)}\right)\left(\frac{\bar{H}}{\bar{u}}-u_{\rm out}\right)}{(x_s-1)\left(\bar{\gamma}u_R(\xi)-\frac{\bar{H}}{u_R(\xi)}\right)}.$$

By construction, $(u_L, u_R, x_s, q_L, q_R)$ satisfies the constraints (4.370) to (4.374).

The function $\bar{H}/u - \bar{\gamma}u = (\bar{\gamma}/u)(2H-u^2)$ is monotonically decreasing in u and positive for $u \in (0, \sqrt{2H})$. Notice that (4.376) implies the inequalities $\bar{u} < \sqrt{2\bar{H}}$ and $\bar{H}/\bar{u} < \sqrt{2\bar{H}}$. Therefore the functions q_L, q_R obey

$$\frac{\left(1-\frac{\bar{H}}{u_{\rm in}^2}\right)(\bar{u}-u_{\rm in})}{x_s\left(\frac{\bar{H}}{u_{\rm in}}-\bar{\gamma}u_{\rm in}\right)} \le q_L(\xi) \le \frac{\left(1-\frac{\bar{H}}{2\bar{H}}\right)(\bar{u}-u_{\rm in})}{x_s\left(\frac{\bar{H}}{\bar{u}}-\bar{\gamma}\bar{u}\right)} = \frac{\left(1-\bar{\gamma}\right)(\bar{u}-\bar{u})}{x_s\left(\frac{\bar{H}}{\bar{u}}-\bar{\gamma}\bar{u}\right)}$$

and

$$\frac{\left(\frac{\bar{u}^2}{\bar{H}}-1\right)\left(\frac{\bar{H}}{\bar{u}}-u_{\rm out}\right)}{(1-x_s)(\frac{\bar{H}}{u_{\rm out}}-\bar{\gamma}u_{\rm out})} \le q_R(\xi) \le \frac{\left(\frac{\bar{H}}{u_{\rm out}}-1\right)\left(\frac{\bar{H}}{\bar{u}}-u_{\rm out}\right)}{(1-x_s)\left(\bar{u}-\bar{\gamma}\frac{\bar{H}}{\bar{u}}\right)}.$$

Thus, q_L, q_R also satisfy the bound constraints (4.375).

(ii) Existence of optimal controls: This part of the existence result uses standard techniques. Let $\{(u_L^n, u_R^n, x_s^n, q_L^n, q_R^n)\}$ be a minimizing sequence.

By Lemma 4 the states obey (4.377) for all n and the derivatives of u_L^n, u_R^n are uniformly bounded. Therefore, the sequence $\{(u_L^n, u_R^n)\}$ is equicontinuous and, by the Arzelà–Ascoli theorem, relatively compact in $C([0,1])^2$. Thus, there exists a subsequence, for simplicity also denoted $\{n\}$, with

$$\begin{aligned} (u_L^n, u_R^n) &\to (u_L^*, u_R^*) \quad \text{in } C([0,1])^2, \\ (q_L^n, q_R^n) &\rightharpoonup (q_L^*, q_R^*) \quad \text{weak-* in } (L^\infty)^2, \\ x_s^n &\to x_s^*. \end{aligned}$$

Consider the set $S_k = \{q_L^* \le q_{\rm low} - 1/k\}$. Let $m(S_k)$ be the (Lebesgue) measure of this set and let χ_{S_k} be the characteristic function. If $m(S_k) > 0$, then the definition of weak–* convergence implies that

$$m(S_k)\, q_{\rm low} \le \int_0^1 q_L^n(\xi)\, \chi_{S_k}(\xi) d\xi \to \int_0^1 q_L^*(\xi)\, \chi_{S_k}(\xi) d\xi \le m(S_k)\, (q_{\rm low} - 1/k),$$

which is a contradiction. Hence $m(S_k) = 0$ for all k. With $\{q_L^* < q_{\text{low}}\} = \cup_{k\in\mathbb{N}} S_k$ we find that $m(\{q_L^* < q_{\text{low}}\}) = 0$. Using analogous arguments we can deduce that $(q_L^*, q_R^*) \in \mathcal{Q}_{ad}$.

Since u_L^n, u_R^n satisfy (4.372), u_L^*, u_R^* satisfy (4.372). Moreover, using (4.377),

$$\frac{x_s^n(\bar{\gamma}u_L^n - \bar{H}/u_L^n)}{1 - \bar{H}/(u_L^n)^2} \to \frac{x_s^*(\bar{\gamma}u_L^* - \bar{H}/u_L^*)}{1 - \bar{H}/(u_L^*)^2} \quad \text{in } C([0,1]).$$

Hence, by taking the limit in

$$u_L^n(\xi) = u_{\text{in}} - \int_0^\xi q_L^n \, \frac{x_s^n(\bar{\gamma}u_L^n - \bar{H}/u_L^n)}{1 - \bar{H}/(u_L^n)^2} \, d\zeta$$

we find that

$$u_L^*(\xi) = u_{\text{in}} - \int_0^\xi q_L^* \, \frac{x_s^*(\bar{\gamma}u_L^* - \bar{H}/u_L^*)}{1 - \bar{H}/(u_L^*)^2} \, d\zeta,$$

i.e. u_L^* satisfies (4.370). Similarly, we can show that u_R^* satisfies (4.371).

Thus, $(u_L^*, u_R^*, x_s^*, q_L^*, q_R^*)$ is a solution to the optimal control problem. △

Fréchet Differentiability

In the following we view u_L, u_R, x_s and q_L, q_R as independent variables. Since the shock location is treated explicitly, Fréchet differentiability of the objective function and the function of constraints can be established. In this section we introduce the mathematical framework that permits us to prove Fréchet differentiability, derive the first derivatives and we prove the continuous invertibility of the partial Fréchet derivative of the constraints with respect to the state variables. The latter property is important to show that constraint qualifications hold and is essential in SQP methods, in which one has to solve linearized state equations.

We introduce the operator

$$C : \mathcal{U} \times \mathcal{Q} \to \mathcal{C}, \tag{4.381}$$

where

$$\mathcal{C} = L^\infty \times L^\infty \times \mathbb{R}^3.$$

The operator C is defined as follows: For $r = (r_L, r_R, r_s, r_{\text{in}}, r_{\text{out}}) \in \mathcal{C}$ the equality

$$C(u_L, u_R, x_s, q_L, q_R) = r$$

holds if and only if

$$(f(u_L))_\xi + x_s g(u_L, q_L) = r_L, \quad \xi \in [0,1], \tag{4.382}$$

$$(f(u_R))_\xi + (x_s - 1) g(u_R, q_R) = r_R, \quad \xi \in [0,1], \tag{4.383}$$

$$f(u_L(1)) - f(u_R(1)) = r_s, \tag{4.384}$$

and

$$u_L(0) - u_{\text{in}} = r_{\text{in}}, \quad u_R(0) - u_{\text{out}} = r_{\text{out}}. \tag{4.385}$$

The equation $C(u_L, u_R, x_s, q_L, q_R) = 0$ is equivalent to (4.364), (4.365), (4.366), (4.367).

To be able to evaluate (4.382) and (4.383) the velocities have to satisfy $u_L(x) \neq 0$, $u_R(x) \neq 0$ for all $x \in [0,1]$.

Theorem 13 *The nonlinear operator $C : \mathcal{U} \times \mathcal{Q} \to \mathcal{C}$ is Fréchet differentiable at any point $(u_L, u_R, x_s, q_L, q_R) \in \mathcal{U} \times \mathcal{Q}$ satisfying $u_L(x) \neq 0$, $u_R(x) \neq 0$ for all $x \in [0,1]$. The partial Fréchet derivatives are given by*

$$C_{(u_L,u_R,x_s)}(u_L,u_R,x_s,q_L,q_R)\,(\hat{u}_L,\hat{u}_R,\hat{x}_s) = \begin{pmatrix} (f_u(u_L)\hat{u}_L)_\xi + x_s g_u(u_L,q_L)\hat{u}_L + \hat{x}_s g(u_L,q_L) \\ (f_u(u_R)\hat{u}_R)_\xi + (x_s-1) g_u(u_R,q_R)\hat{u}_R + \hat{x}_s g(u_R,q_R) \\ f_u(u_L(1))\hat{u}_L(1) - f_u(u_R(1))\hat{u}_R(1) \\ \hat{u}_L(0) \\ \hat{u}_R(0) \end{pmatrix}$$

and

$$C_{(q_L,q_R)}(u_L,u_R,x_s,q_L,q_R)\,(\hat{q}_L,\hat{q}_R) = \begin{pmatrix} x_s g_q(u_L,q_L)\hat{q}_L \\ (x_s-1) g_q(u_R,q_R)\hat{q}_R \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Proof: To prove the differentiability of C with respect to (u_L, u_R, x_s) we have to show that

$$\Big\| C(u_L+\hat{u}_L, u_R+\hat{u}_R, x_s+\hat{x}_s, q_L, q_R) - C(u_L,u_R,x_s,q_L,q_R) \\ - C_{(u_L,u_R,x_s)}(u_L,u_R,x_s,q_L,q_R)(\hat{u}_L,\hat{u}_R,\hat{x}_s) \Big\|_{\mathcal{C}} = o\Big(\|(\hat{u}_L,\hat{u}_R,\hat{x}_s)\|_{\mathcal{U}}\Big).$$

For the first component corresponding to the equation (4.382) we obtain

$$\begin{aligned} & (f(u_L+\hat{u}_L))_\xi + (x_s+\hat{x}_s) g(u_L+\hat{u}_L, q_L) - (f(u_L))_\xi - x_s g(u_L,q_L) \\ & -(f_u(u_L)\hat{u}_L)_\xi - x_s g_u(u_L,q_L)\hat{u}_L - \hat{x}_s g(u_L,q_L) \\ = \; & (f(u_L+\hat{u}_L) - f(u_L) - f_u(u_L)\hat{u}_L)_\xi + x_s\,[g(u_L+\hat{u}_L,q_L) - g(u_L,q_L) - g_u(u_L,q_L)\hat{u}_L] \\ & + \hat{x}_s\,[g(u_L+\hat{u}_L,q_L) - g(u_L,q_L)] \\ = \; & \left(\int_0^1 (f_u(u_L+t\hat{u}_L) - f_u(u_L))\,\hat{u}_L\,dt \right)_\xi + x_s\, o\Big(\|\hat{u}_L\|_\infty\Big) + \hat{x}_s\, O\Big(\|\hat{u}_L\|_\infty\Big) \\ = \; & \int_0^1 \big(f_{uu}(u_L+t\hat{u}_L)(u_L+t\hat{u}_L)_\xi - f_{uu}(u_L)(u_L)_\xi\big)\hat{u}_L + \big(f_u(u_L+t\hat{u}_L) - f_u(u_L)\big)(\hat{u}_L)_\xi\,dt \\ & + x_s\, o\Big(\|\hat{u}_L\|_\infty\Big) + \hat{x}_s\, O\Big(\|\hat{u}_L\|_\infty\Big) \\ = \; & O\Big(\|\hat{u}_L\|_{1,\infty}^2\Big) + x_s\, o\Big(\|\hat{u}_L\|_\infty\Big) + \hat{x}_s\, O\Big(\|\hat{u}_L\|_\infty\Big). \end{aligned}$$

Similar estimates can be applied to show analogous results for the equations (4.383) and (4.384). This proves the differentiability of C with respect to (u_L, u_R, x_s).

The differentiability of C with respect to (q_L, q_R) follows easily, since the function g in (4.360) is linear in q. △

The following result concerns the invertibility of the partial Fréchet derivative $C_{(u_L,u_R,x_s)}(u_L,u_R,x_s,q_L,q_R)$.

For given $(u_L, u_R, x_s) \in \mathcal{U}$, $(q_L, q_R) \in \mathcal{Q}$, and $(r_L, r_R, r_s, r_{\text{in}}, r_{\text{out}}) \in \mathcal{C}$ we consider the system

$$(f_u(u_L)\hat{u}_L)_\xi + x_s g_u(u_L, q_L)\hat{u}_L + \hat{x}_s g(u_L, q_L) = r_L, \quad \xi \in [0,1], \tag{4.386}$$

$$(f_u(u_R)\hat{u}_R)_\xi + (x_s - 1) g_u(u_R, q_R)\hat{u}_R + \hat{x}_s g(u_R, q_R) = r_R, \quad \xi \in [0,1], \tag{4.387}$$

$$f_u(u_L(1))\hat{u}_L(1) - f_u(u_R(1))\hat{u}_R(1) = r_s, \tag{4.388}$$

and

$$\hat{u}_L(0) = r_{\text{in}}, \quad \hat{u}_R(0) = r_{\text{out}}. \tag{4.389}$$

Theorem 14 *(i) Suppose that $(u_L, u_R, x_s, q_L, q_R) \in \mathcal{U} \times \mathcal{Q}_{ad}$ is a point satisfying*

$$u_i(\xi) \neq 0, \quad u_i(\xi) \neq \sqrt{\bar{H}}, \quad \forall \xi \in [0,1], \qquad i = L, R, .$$

If

$$\begin{aligned} &\int_0^1 \exp\left(-\int_t^1 \frac{x_s g_u(u_L(s), q_L(s))}{f_u(u_L(s))}\, ds\right) g(u_L(t), q_L(t))\, dt \\ \neq\ &\int_0^1 \exp\left(-\int_t^1 \frac{(x_s - 1) g_u(u_R(s), q_R(s))}{f_u(u_R(s))}\, ds\right) g(u_R(t), q_R(t))\, dt, \end{aligned} \tag{4.390}$$

then for every $(r_L, r_R, r_s, r_{\text{in}}, r_{\text{out}}) \in \mathcal{C}$ the system (4.386), (4.387), (4.388), (4.389) admits a unique solution $(\hat{u}_L, \hat{u}_R, \hat{x}_s) \in \mathcal{U}$ which depends continuously on $(r_L, r_R, r_s, r_{\text{in}}, r_{\text{out}})$.

(ii) If, in addition, there exists constants δ, Δ with $0 < \delta < \Delta$ such that the point $(u_L, u_R, x_s, q_L, q_R)$ obeys

$$\delta \leq u_i(\xi) \leq \Delta, \quad |u_i(\xi) - \sqrt{\bar{H}}| \geq \delta, \quad \forall \xi \in [0,1], \qquad i = L, R,$$

and

$$\left| \int_0^1 \exp\left(-\int_t^1 \frac{x_s g_u(u_L, q_L)}{f_u(u_L)}\, ds\right) g(u_L, q_L) - \exp\left(-\int_t^1 \frac{(x_s - 1) g_u(u_R, q_R)}{f_u(u_R)}\, ds\right) g(u_R, q_R)\, dt \right| > \delta,$$

then there exists a constant K dependent on δ, Δ, but independent of $(u_L, u_R, x_s, q_L, q_R)$ such that

$$||(\hat{u}_L, \hat{u}_R, \hat{x}_s)||_{\mathcal{U}} \leq K ||(r_L, r_R, r_s, r_{\text{in}}, r_{\text{out}})||_{\mathcal{C}}.$$

Proof: (i) First we note that since $W^{1,\infty}(0,1) \subset C([0,1])$, there exists $\delta > 0$ such that $u_i(\xi) > \delta$, $|u_i(\xi) - \sqrt{\bar{H}}| > \delta$, for all $\xi \in [0,1]$, $i = L, R$.

The equation (4.386) is equivalent to

$$(f_u(u_L)\hat{u}_L)_\xi + \frac{x_s g_u(u_L, q_L)}{f_u(u_L)} (f_u(u_L)\hat{u}_L) = r_L - \hat{x}_s g(u_L, q_L). \tag{4.391}$$

Using the integrating factor

$$\mu_L(\xi) = \exp\left(\int_0^\xi \frac{x_s g_u(u_L(t), q_L(t))}{f_u(u_L(t))}\, dt\right),$$

the solution of (4.391) with initial condition $\hat{u}_L(0) = r_{\text{in}}$ is given by

$$\hat{u}_L(\xi) = +\frac{1}{\mu_L(\xi) f_u(u_L(\xi))} \left(r_{\text{in}} f_u(u_L(0)) + \int_0^\xi \mu_L(t)[r_L(t) - \hat{x}_s g(u_L(t), q_L(t))]\, dt\right). \tag{4.392}$$

Similarly, one can show that the solution of (4.387) with initial condition $\hat{u}_R(0) = r_{\text{out}}$ is given by

$$\hat{u}_R(\xi) = \frac{1}{\mu_R(\xi) f_u(u_R(\xi))} \left(r_{\text{out}} f_u(u_R(0)) + \int_0^\xi \mu_R(t)[r_R(t) - \hat{x}_s g(u_R(t), q_R(t))]\, dt\right), \tag{4.393}$$

where

$$\mu_R(\xi) = \exp\left(\int_0^\xi \frac{(x_s-1)g_u(u_R(t),q_R(t))}{f_u(u_R(t))}\,dt\right).$$

Inserting (4.392), (4.393) into (4.388) yields

$$\begin{aligned} & r_{\text{in}}\frac{f_u(u_L(0))}{\mu_L(1)} + \int_0^1 \exp\left(-\int_t^1 \frac{x_s g_u(u_L(s),q_L(s))}{f_u(u_L(s))}\,ds\right)[r_L(t) - \hat{x}_s g(u_L(t),q_L(t))]\,dt \\ = \; & r_{\text{out}}\frac{f_u(u_R(0))}{\mu_R(1)} + \int_0^1 \exp\left(-\int_t^1 \frac{(x_s-1)g_u(u_R(s),q_R(s))}{f_u(u_R(s))}\,ds\right)[r_R(t) - \hat{x}_s g(u_R(t),q_R(t))]\,dt. \end{aligned} \tag{4.394}$$

If the inequality (4.390) is valid, then (4.394) can be solved for $\hat{x}_s$. This proves the existence and uniqueness of the solution.

The continuous dependence of $(\hat{u}_L, \hat{u}_R, \hat{x}_s) \in \mathcal{U}$ upon $(r_L, r_R, r_s, r_{\text{in}}, r_{\text{out}}) \in \mathcal{C}$ follows from the equations (4.392), (4.393), (4.394).

(ii) The assertion follows from equations (4.392), (4.393), and (4.394). △

Corollary 1 *If $(\bar{u}_L, \bar{u}_R, \bar{x}_s, \bar{q}_L, \bar{q}_R) \in \mathcal{U} \times \mathcal{Q}$ is feasible, i.e. it satisfies the constraints (4.370) to (4.375), and if there exists $\delta > 0$ with*

$$\left|\int_0^1 \exp\left(-\int_t^1 \frac{x_s g_u(\bar{u}_L,\bar{q}_L)}{f_u(\bar{u}_L)}\,ds\right) g(\bar{u}_L,\bar{q}_L) - \exp\left(-\int_t^1 \frac{(x_s-1)g_u(\bar{u}_R,\bar{q}_R)}{f_u(\bar{u}_R)}\,ds\right) g(\bar{u}_R,\bar{q}_R)\,dt\right| > \delta,$$

then there exists $\epsilon > 0$ such that for all $(u_L, u_R, x_s) \in \mathcal{U}$, $(q_L, q_R) \in \mathcal{Q}$ with

$$||(u_L,u_R,x_s) - (\bar{u}_L,\bar{u}_R,\bar{x}_s)||_{\mathcal{U}} < \epsilon, \quad ||(q_L,q_R) - (\bar{q}_L,\bar{q}_R)||_{\mathcal{Q}} < \epsilon$$

and for all $(r_L, r_R, r_s, r_{\text{in}}, r_{\text{out}}) \in \mathcal{C}$ the system (4.386), (4.387), (4.388), (4.389) admits a unique solution $(\hat{u}_L, \hat{u}_R, \hat{x}_s) \in \mathcal{U}$. Moreover, there exists a constant K independent of $(\bar{u}_L, \bar{u}_R, \bar{x}_s, \bar{q}_L, \bar{q}_R)$ such that

$$||(\hat{u}_L,\hat{u}_R,\hat{x}_s)||_{\mathcal{U}} \le K||(r_L,r_R,r_s,r_{\text{in}},r_{\text{out}})||_{\mathcal{C}}.$$

Proof: The solutions u_L, u_R satisfy (4.377). Hence, the assertion follows from Theorem 14(ii). △

From the definitions of f and g one can see that the Fréchet derivative of C is Lipschitz–continuous for all u_L, u_R with $u_L(\xi), u_R(\xi) \ge \underline{u} > 0$ for all $\xi \in [0,1]$. Moreover, C is even twice Fréchet differentiable if $u_L, u_R > 0$.

To prove the Fréchet differentiability of the objective function we have to keep in mind that the desired velocity depends on x_s, cf. (4.368). Therefore differentiability with respect to x_s can only be guaranteed if the desired velocity u^d is sufficiently smooth, a fact that will be addressed again in the numerical examples section.

Theorem 15 *If the desired velocity u^d is differentiable with absolutely continuous derivative, then the objective function J is Fréchet differentiable. The partial Fréchet -derivatives are given by*

$$\begin{aligned} J_{u_L}(u_L,u_R,x_s,q_L,q_R)\hat{u}_L &= x_s\int_0^1 (u_L(\xi) - u_L^d(x_s;\xi))\,\hat{u}_L(\xi)\,d\xi, \\ J_{u_R}(u_L,u_R,x_s,q_L,q_R)\hat{u}_R &= (1-x_s)\int_0^1 (u_R(\xi) - u_R^d(x_s;\xi))\,\hat{u}_R(\xi)\,d\xi, \end{aligned}$$

and

$$
\begin{aligned}
&J_{x_s}(u_L, u_R, x_s, q_L, q_R) \\
&\quad = \int_0^1 \frac{1}{2}(u_L(\xi) - u_L^d(x_s;\xi))^2 - x_s\,(u_L(\xi) - u_L^d(x_s;\xi))\,(u_L^d)_x(\xi)\,\xi\,d\xi \\
&\qquad - \int_0^1 \frac{1}{2}(u_R(\xi) - u_R^d(x_s;\xi))^2 + (1-x_s)\,(u_R(\xi) - u_R^d(x_s;\xi))\,(u_R^d)_x(\xi)\,\xi\,d\xi. \qquad (4.395)
\end{aligned}
$$

The objective function J is twice Fréchet differentiable if the desired velocity u^d is twice differentiable with absolutely continuous second derivative.

Proof: The assertion follows from the definition of J using standard estimates. The proof is therefore omitted. △

We conclude this section with a brief discussion of the differentiability of the velocity function. Suppose that we have an area function $\bar{q}_L, \bar{q}_R$ and corresponding velocities $\bar{u}_L, \bar{u}_R$ and shock location $\bar{x}_s$ that satisfy the state equations (4.370), (4.371), (4.372), (4.373) and are such that (4.390) is fulfilled. Then the implicit function theorem guarantees the differentiability of the function

$$L^\infty \times L^\infty \ni (q_L, q_R) \longrightarrow (u_L, u_R, x_s) \in W^{1,\infty} \times W^{1,\infty} \times I\!R$$

that maps the area into the solution of the state equation at this point. In fact, the derivative is given by

$$(u_L(\bar{q}_L,\bar{q}_R), u_L(\bar{q}_L,\bar{q}_R), x_s(\bar{q}_L,\bar{q}_R)_{(\bar{q}_L,\bar{q}_R)} = -C_{(u_L,u_R,x_s)}(\bar{u}_L,\bar{u}_R,\bar{x}_s,\bar{q}_L,\bar{q}_R)^{-1} C_{(q_L,q_R)}(\bar{u}_L,\bar{u}_R,\bar{x}_s,\bar{q}_L,\bar{q}_R).$$

Given u_L, u_R, x_s, the velocity of the original problem can be obtained via (4.362), i.e.

$$u(x) = \begin{cases} u_L\left(\frac{x}{x_s}\right), & x \in [0, x_s), \\ u_R\left(\frac{1-x}{1-x_s}\right), & x \in [x_s, 1]. \end{cases} \qquad (4.396)$$

If one considers the map

$$W^{1,\infty} \times W^{1,\infty} \times I\!R \ni (u_L, u_R, x_s) \longrightarrow u \in L^\infty$$

that is defined by (4.396), then it is easy to see that because of the presence of a shock this map is not Fréchet differentiable. In fact it is not even continuous. This shows that differentiability is only lost when the composite function

$$(q_L, q_R) \longrightarrow (u_L, u_R, x_s) \longrightarrow u$$

is considered. If left and right velocity and shock location are treated as independent variables, then, as shown in this section, differentiability can be guaranteed under suitable assumptions.

Optimality Conditions

We define the Lagrange function

$$
\begin{aligned}
&L(u_L, u_R, x_s, q_L, q_R, \lambda_L, \lambda_R, \lambda_s) \\
&\quad = \frac{x_s}{2}\int_0^1 (u_L - u_L^d)^2\,d\xi + \frac{1-x_s}{2}\int_0^1 (u_R - u_R^d)^2\,d\xi + \int_0^1 \lambda_L[(f(u_L))_\xi + x_s g(u_L, q_L)]\,d\xi \\
&\qquad + \int_0^1 \lambda_R[(f(u_R))_\xi + (x_s - 1)g(u_R, q_R)]\,d\xi + \lambda_s[f(u_L(1)) - f(u_R(1))]. \qquad (4.397)
\end{aligned}
$$

If the shock location at the optimum obeys $x_s \in (0,1)$, then the first order necessary optimality conditions are

$$\begin{aligned} 0 &= L_{(u_L,u_R,x_s)}(u_L,u_R,x_s,q_L,q_R,\lambda_L,\lambda_R,\lambda_s)(\hat{u}_L,\hat{u}_R,\hat{x}_s), \\ 0 &\leq L_{(q_L,q_R)}(u_L,u_R,x_s,q_L,q_R,\lambda_L,\lambda_R,\lambda_s)(\hat{q}_L,\hat{q}_R), \\ 0 &= L_{(\lambda_L,\lambda_R,\lambda_s)}(u_L,u_R,x_s,q_L,q_R,\lambda_L,\lambda_R,\lambda_s)(\hat{\lambda}_L,\hat{\lambda}_R,\hat{\lambda}_s), \end{aligned} \tag{4.398}$$

for all $(\hat{u}_L,\hat{u}_R,\hat{x}_s)$ with $\hat{u}_L(0) = \hat{u}_R(0) = 0$, for all $(\hat{q}_L,\hat{q}_R)$ with $(q_L+\hat{q}_L, q_R+\hat{q}_R) \in \mathcal{Q}_{ad}$, and for all $(\hat{\lambda}_L,\hat{\lambda}_R,\hat{\lambda}_s)$.

The third equation in (4.398) yields the state equation (4.370), (4.371), (4.372), (4.373). Using integration by parts we find that the first equation in (4.398) with $\hat{u}_L(0) = \hat{u}_R(0) = 0$ yields

$$\begin{aligned} 0 &= L_{(u_L,u_R,x_s)}(u_L,u_R,x_s,q_L,q_R,\lambda_L,\lambda_R,\lambda_s)(\hat{u}_L,\hat{u}_R,\hat{x}_s) \\ &= x_s\int_0^1 (u_L-u_L^d)\hat{u}_L\,d\xi + (1-x_s)\int_0^1 (u_R-u_R^d)\hat{u}_R\,d\xi + J_{x_s}(u_L,u_R,x_s,q_L,q_R)\hat{x}_s \\ &\quad + \int_0^1 -(\lambda_L)_\xi f_u(u_L)\hat{u}_L + \lambda_L[x_s g_u(u_L,q_L)\hat{u}_L + \hat{x}_s g(u_L,q_L)]\,d\xi \\ &\quad + \int_0^1 -(\lambda_R)_\xi f_u(u_R)\hat{u}_R + \lambda_R[(x_s-1)g_u(u_R,q_R)\hat{u}_R + \hat{x}_s g(u_R,q_R)]\,d\xi \\ &\quad + (\lambda_L(1)+\lambda_s)f_u(u_L(1))\hat{u}_L(1) + (\lambda_R(1)-\lambda_s)f_u(u_R(1))\hat{u}_R(1). \end{aligned} \tag{4.399}$$

If one sets $\hat{x}_s = 0$ and varies over all $(\hat{u}_L,\hat{u}_R)$ with $\hat{u}_L(0) = \hat{u}_R(0) = \hat{u}_L(1) = \hat{u}_R(1) = 0$, then one obtains the adjoint equations

$$(\lambda_L)_\xi\, f_u(u_L) = x_s g_u(u_L,q_L)\,\lambda_L + x_s(u_L-u_L^d), \tag{4.400}$$

$$(\lambda_R)_\xi\, f_u(u_R) = (x_s-1)g_u(u_R,q_R)\,\lambda_R + (1-x_s)(u_R-u_R^d). \tag{4.401}$$

Allowing $\hat{u}_L(1), \hat{u}_R(1) \neq 0$ yields the conditions

$$\lambda_L(1) = -\lambda_s, \quad \lambda_R(1) = \lambda_s. \tag{4.402}$$

Finally, varying $\hat{x}_s$ gives

$$\int_0^1 \lambda_L g(u_L,q_L) + \lambda_R g(u_R,q_R)\,d\xi + J_{x_s}(u_L,u_R,x_s,q_L,q_R) = 0. \tag{4.403}$$

The existence of Lagrange multipliers are guaranteed if the operator of linearized constraints is onto. Thus, existence of Lagrange multipliers is expected under the assumptions of Theorem 14(i). We provide a proof of this result, since the explicit form of the Lagrange multipliers derived in the proof are of interest in connection with the discretized problem.

Theorem 16 *If the assumptions of Theorem 14(i) are valid, then the adjoint system (4.400), (4.401), (4.402), (4.403) admits a unique solution.*

Proof: Equation (4.400) is equivalent to

$$(\lambda_L)_\xi - \frac{x_s g_u(u_L,q_L)}{f_u(u_L)}\,\lambda_L = x_s\,\frac{u_L-u_L^d}{f_u(u_L)}.$$

Using the integrating factor

$$\nu_L(\xi) = \exp\left(\int_\xi^1 \frac{x_s g_u(u_L(t),q_L(t))}{f_u(u_L(t))}\,dt\right),$$

the solution of (4.400) with $\lambda_L(1) = -\lambda_s$ is given by

$$(\lambda_L)(\xi) = \frac{1}{\nu_L(\xi)} \left(-\lambda_s - \int_\xi^1 x_s \, \nu_L \, \frac{u_L - u_L^d}{f_u(u_L)} \, dt \right). \tag{4.404}$$

Similarly, the solution of (4.401) with $\lambda_R(1) = \lambda_s$ is given by

$$(\lambda_R)(\xi) = \frac{1}{\nu_R(\xi)} \left(\lambda_s - \int_\xi^1 (1 - x_s) \, \nu_R \, \frac{u_R - u_R^d}{f_u(u_R)} \, dt \right), \tag{4.405}$$

where

$$\nu_R(\xi) = \exp\left(\int_\xi^1 \frac{(x_s - 1) g_u(u_R(t), q_R(t))}{f_u(u_R(t))} \, dt \right).$$

Inserting the solutions into (4.403), we find that

$$\begin{aligned}
&\int_0^1 \exp\left(-\int_\xi^1 \frac{x_s g_u(u_L, q_L)}{f_u(u_L)} \, ds \right) g(u_L, q_L) - \exp\left(-\int_\xi^1 \frac{(x_s - 1) g_u(u_R, q_R)}{f_u(u_R)} \, ds \right) g(u_R, q_R) \, d\xi \;\; \lambda_s \\
= \;& J_{x_s}(u_L, u_R, x_s, q_L, q_R) - \int_0^1 \left\{ \int_\xi^1 (1 - x_s) \exp\left(-\int_\xi^t \frac{(x_s - 1) g_u(u_R, q_R)}{f_u(u_R)} \, ds \right) \frac{u_R - u_R^d}{f_u(u_R)} \, dt \right. \\
&\left. + \int_\xi^1 x_s \exp\left(-\int_\xi^t \frac{x_s g_u(u_L, q_L)}{f_u(u_L)} \, ds \right) \frac{u_L - u_L^d}{f_u(u_L)} \, dt \right\} d\xi.
\end{aligned} \tag{4.406}$$

Since (4.390) holds true, equation (4.406) has a unique solution λ_s. This concludes the proof of the theorem. △

The second equation in (4.398) is equivalent to

$$\lambda_L(\xi) \, x_s \left(\bar{\gamma} u_L(\xi) - \frac{\bar{H}}{u_L(\xi)} \right) \begin{cases} \geq 0 & \text{if } q_L(\xi) = q_{\text{low}}, \\ = 0 & \text{if } q_L(\xi) \in (q_{\text{low}}, q_{\text{upp}}), \\ \leq 0 & \text{if } q_L(\xi) = q_{\text{upp}}, \end{cases} \tag{4.407}$$

and

$$\lambda_R(\xi) \, (x_s - 1) \left(\bar{\gamma} u_R(\xi) - \frac{\bar{H}}{u_R(\xi)} \right) \begin{cases} \geq 0 & \text{if } q_R(\xi) = q_{\text{low}}, \\ = 0 & \text{if } q_R(\xi) \in (q_{\text{low}}, q_{\text{upp}}), \\ \leq 0 & \text{if } q_R(\xi) = q_{\text{upp}}. \end{cases} \tag{4.408}$$

The interesting result in this section is about the continuity properties of the Lagrange multiplier. The co–states λ_L and λ_R obey the shock condition (4.402). From the examination of the other equations it can be seen that $\lambda_L(1) = -\lambda_R(1) = 0$ can be guaranteed only if the desired velocities can be matched exactly, i.e. the source terms in (4.400), (4.401), and the term $\mathcal{J}_{x_s}(u_L, u_R, x_s, q_L, q_R)$ in (4.403) vanish.

The Discrete Design Problem

As in the analysis of the continuous problem we divide the interval into two subintervals $[0, x_s]$ and $[x_s, 1]$. The shock location x_s is one of the state variables. The transformation onto the fixed domain as shown at the end of Section 4.8.3, however, is not performed explicitly, but incorporated implicitly using moving grids on the left and on the right of the shock. As we will see later, this is equivalent to discretizing the fixed domain control problem (4.369) – (4.375).

For the discretization of the optimal control problem we use a cell centered grid. The subinterval $[0, x_s]$ left of the shock is subdivided into N_L equidistant subintervals of length $h_L = x_s/N_L$, the subinterval $[x_s, 1]$ right of the shock is subdivided into N_R equidistant subintervals of length $h_R = (1 - x_s)/N_R$. The point x_i denotes the midpoint of the ith cell:

$$\begin{aligned} x_i &= (i - \tfrac{1}{2})h_L, && i = 1, \ldots, N_L, \\ x_i &= x_s + (i - \tfrac{1}{2} - N_L)h_R, && i = N_L + 1, \ldots, N_L + N_R. \end{aligned} \tag{4.409}$$

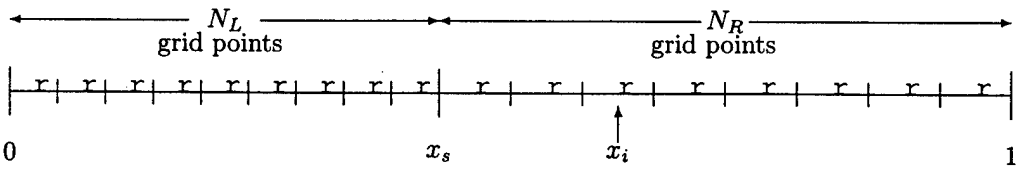


Figure 4.44: The Grid

On each cell the function q defining the area of the duct and the velocity u are approximated by constants q_i and u_i, respectively. Therefore the number of state variables is $N_L + N_R + 1$ and the number of control variables is $N_L + N_R$.

The objective function is discretized using the midpoint rule:

$$\frac{x_s}{2} \int_0^1 (u_L(\xi) - u_L^d(x_s;\xi))^2 \, d\xi + \frac{1 - x_s}{2} \int_0^1 (u_R(\xi) - u_R^d(x_s;\xi))^2 \, d\xi \;\approx\; J^h(u, x_s, q),$$

where

$$J^h(u, x_s, q) \equiv \frac{1}{2} \sum_{i=1}^{N_L} h_L \, (u_i - u^d(x_i))^2 + \frac{1}{2} \sum_{i=N_L+1}^{N_L+N_R} h_R \, (u_i - u^d(x_i))^2. \tag{4.410}$$

For the discretization of the differential equation (4.359) we use

$$\frac{f_{i+1/2} - f_{i-1/2}}{\Delta x_i} + g(u_i, q_i) = 0,$$

where Δx_i denotes the width of cell i. The fluxes at the cell boundaries are approximated as follows: In the supersonic region left of the shock we set $f_{i+1/2} = f(u_i)$ giving

$$\frac{f(u_i) - f(u_{i-1})}{h_L} + g(u_i, q_i) = 0, \quad i = 1, \ldots, N_L. \tag{4.411}$$

In the subsonic region right of the shock we set $f_{i+1/2} = f(u_{i+1})$ giving

$$\frac{f(u_{i+1}) - f(u_i)}{h_R} + g(u_i, q_i) = 0, \quad i = N_L + 1, \ldots, N_L + N_R. \tag{4.412}$$

The Rankine–Hugoniot condition is discretized as

$$f(u_{N_L}) - f(u_{N_L+1}) = 0. \tag{4.413}$$

The equations (4.411), (4.412) are also the ones used in the Godunov scheme for the supersonic and subsonic regions, respectively.

If one multiplies (4.411) by x_s and (4.412) by $(x_s - 1)$, then one can see that the resulting equations are implicit discretization schemes for (4.370) and (4.371). For (4.370) the indices $i = 0$ and $i = N_L$ correspond to the boundaries $\xi = 0$ and $\xi = 1$, respectively, whereas for (4.371) the indices $i = N_L + N_R$ and $i = N_L + 1$ correspond to the boundaries $\xi = 0$ and $\xi = 1$, respectively.

The equations (4.411) and (4.412) multiplied by x_s and $x_s - 1$, respectively, and the equation (4.413) form the $N_L + N_R + 1$ state constraints

$$C_i^h(u, x_s, q) = 0, \quad i = 1, \ldots, N_L + N_R + 1,$$

where

$$C_i^h(u, x_s, q) \equiv \begin{cases} N_L \left(f(u_i) - f(u_{i-1})\right) + x_s \, g(u_i, q_i) & i = 1, \ldots, N_L, \\ N_R \left(-f(u_{i+1}) + f(u_i)\right) + (x_s - 1) \, g(u_i, q_i) & i = N_L + 1, \ldots, N_L + N_R, \\ f(u_{N_L}) - f(u_{N_L+1}) & i = N_L + N_R + 1. \end{cases} \tag{4.414}$$

The scalars u_0 and $u_{N_L+N_R+1}$ are determined from the boundary conditions (4.373).

This leads to the finite dimensional optimal control problem

$$\begin{aligned} \min \quad & J^h(u, x_s, q) \\ \text{s.t.} \quad & C^h(u, x_s, q) = 0, \\ & 0 \le q_{\text{low}} \le q \le q_{\text{upp}}. \end{aligned} \tag{4.415}$$

The state variables in the discrete problem are $(u_1, \ldots, u_{N_L+N_R}, x_s)$ and the control variables are
$(q_1, \ldots, q_{N_L+N_R})$. One may add a state constraint $0 \le x_s \le 1$ to (4.415). However, in our numerical experiments the shock was always in the interior.

Under the assumptions of Theorem 15 the discretized objective function J^h is differentiable. In fact, due to the discretization, one can even relax the differentiability assumptions on u^d. The objective function J^h is differentiable if u^d is differentiable at x_i, $i = 1, \ldots, N_L + N_R$. In particular, it holds that

$$\begin{aligned} J^h_{x_s}(u, x_s, q) \;=\; & \sum_{i=1}^{N_L} \frac{1}{2N_L} (u_i - u^d(x_i))^2 - h_L \frac{i - \frac{1}{2}}{N_L} (u_i - u^d(x_i)) \, u_x^d(x_i) \\ & - \sum_{i=N_L+1}^{N_L+N_R} \frac{1}{2N_R} (u_i - u^d(x_i))^2 + h_R \Big(1 - \frac{i - \frac{1}{2} - N_L}{N_R}\Big) (u_i - u^d(x_i)) \, u_x^d(x_i). \end{aligned}$$

From the definition of f and g, it is easy to see that C^h is differentiable for all u, x_s, q with $u > 0$. The partial Jacobian $C^h_{(u,x_s)}$ of C^h is a bordered matrix given by

$$C^h_{(u,x_s)}(u, x_s, q) = \begin{pmatrix} B_L & 0 & e_L \\ 0 & B_R & e_R \\ d_L^T & d_R^T & 0 \end{pmatrix}, \tag{4.416}$$

where $B_L \in \mathbb{R}^{N_L \times N_L}$ is a lower bidiagonal matrix, $B_R \in \mathbb{R}^{N_R \times N_R}$ is a upper bidiagonal matrix, and $e_L, d_L \in \mathbb{R}^{N_L}$, $e_R, d_R \in \mathbb{R}^{N_R}$. The structure of the matrix reflects the left hand side of the system (4.386), (4.387), (4.388), (4.389). The partial Jacobian C_q^h of C^h is a $(N_L + N_R + 1) \times (N_L + N_R)$

'diagonal' matrix with diagonal entries given by

$$\Big(C_q^h(u,x_s,q)\Big)_{ii} = \begin{cases} x_s\, g_q(u_i,q_i) & i=1,\ldots,N_L, \\ (x_s-1)\, g_q(u_i,q_i) & i=N_L+1,\ldots,N_L+N_R. \end{cases}$$

If $C^h_{(u,x_s)}(u,x_s,q), B_L, B_R$ are nonsingular, the linear system

$$\begin{pmatrix} B_L & 0 & e_L \\ 0 & B_R & e_R \\ d_L^T & d_R^T & 0 \end{pmatrix} \begin{pmatrix} \hat{u}_L \\ \hat{u}_R \\ \hat{x}_s \end{pmatrix} = \begin{pmatrix} r_L \\ r_R \\ r_s \end{pmatrix}$$

can be solved using

$$\hat{u}_L = B_L^{-1}(r_L - e_L\,\hat{x}_s), \tag{4.417}$$

$$\hat{u}_R = B_R^{-1}(r_R - e_R\,\hat{x}_s), \tag{4.418}$$

where

$$\hat{x}_s = (d_L^T B_L^{-1} e_L + d_R^T B_R^{-1} e_R)^{-1}(d_L^T B_L^{-1} r_L + d_R^T B_R^{-1} r_R - r_s). \tag{4.419}$$

This solution procedure is the discrete version of the procedure applied in the proof of Theorem 14 to establish the existence of a unique solution of (4.386), (4.387), (4.388), (4.389). This is also the solution procedure that is used in our numerical examples.

Theorem 17 *If*

$$u_i > \sqrt{\bar{H}}, \quad i=1,\ldots,N_L, \tag{4.420}$$

then B_L is nonsingular. If

$$u_i \in (0,\sqrt{\bar{H}}), \quad i=N_L+1,\ldots,N_L+N_R, \tag{4.421}$$

then B_R is nonsingular, and if (4.420), (4.421), and

$$\begin{aligned} &\frac{1}{N_L}\sum_{j=1}^{N_L}\left(\prod_{k=j}^{N_L}\frac{N_L\, f_u(u_k)+x_s\, g_u(u_k,q_k)}{N_L\, f_u(u_k)}\right)^{-1} g(u_j,q_j) \\ \neq\ &\frac{1}{N_R}\sum_{j=N_L+1}^{N_L+N_R}\left(\prod_{k=N_L+1}^{j}\frac{N_R\, f_u(u_k)+(x_s-1)\, g_u(u_k,q_k)}{N_R\, f_u(u_k)}\right)^{-1} g(u_j,q_j), \end{aligned} \tag{4.422}$$

then the matrix $C^h_{(u,x_s)}(u,x_s,q)$ is nonsingular.

Proof: The ith equation of the system $B_L\hat{u}_L = r_L - e_L\,\hat{x}_s$ is given by

$$\Big(N_L\, f_u(u_i)+x_s\, g_u(u_i,q_i)\Big)(\hat{u}_L)_i - N_L\, f_u(u_{i-1})(\hat{u}_L)_{i-1} = -g(u_i,q_i)\hat{x}_s + (r_L)_i, \quad i=1,\ldots,N_L, \tag{4.423}$$

where $(\hat{u}_L)_0 = 0$. Condition (4.420) guarantees that $N_L\, f_u(u_i)+x_s\, g_u(u_i,q_i) > 0,\ i=1,\ldots,N_L$. With

$$z_i = \frac{N_L\, f_u(u_{i-1})}{N_L\, f_u(u_i)+x_s\, g_u(u_i,q_i)}, \quad w_i = \frac{g(u_i,q_i)}{N_L\, f_u(u_i)+x_s\, g_u(u_i,q_i)}, \quad v_i = \frac{(r_L)_i}{N_L\, f_u(u_i)+x_s\, g_u(u_i,q_i)},$$

these equation can be written as

$$(\hat{u}_L)_i - z_i(\hat{u}_L)_{i-1} = -w_i\hat{x}_s + v_i, \quad i=1,\ldots,N_L,$$

If we multiply the last difference equation by $1/Z_i$, where $Z_i = \prod_{j=1}^{i} z_j$, $i = 1, \ldots, N_L$, $Z_0 = 1$, then we obtain

$$\frac{1}{Z_i}(\hat{u}_L)_i - \frac{1}{Z_{i-1}}(\hat{u}_L)_{i-1} = -\frac{w_i}{Z_i}\hat{x}_s + \frac{v_i}{Z_i}, \quad i = 1, \ldots, N_L,$$

The solution of this difference equation is given by

$$(\hat{u}_L)_i = Z_i \left(\sum_{j=1}^{i} -\frac{w_j}{Z_j}\hat{x}_s + \frac{v_j}{Z_j} \right), \quad i = 1, \ldots, N_L.$$

In particular, it holds that

$$\begin{aligned} (\hat{u}_L)_{N_L} \;=\; & \frac{\prod_{k=1}^{N_L-1} N_L\, f_u(u_k)}{\prod_{k=1}^{N_L} N_L\, f_u(u_k) + x_s\, g_u(u_k, q_k)} \times \\ & \Big(- \sum_{j=1}^{N_L} \prod_{k=1}^{j-1} \frac{N_L\, f_u(u_k) + x_s\, g_u(u_k, q_k)}{N_L\, f_u(u_k)}\, g(u_j, q_j)\, \hat{x}_s \; + \sum_{j=1}^{N_L} \frac{v_j}{Z_j} \Big). \end{aligned} \tag{4.424}$$

The ith equation of the system $B_R \hat{u}_R = r_R - e_R \hat{x}_s$ is given by

$$\Big(N_R\, f_u(u_i) + (x_s - 1)\, g_u(u_i, q_i) \Big)(\hat{u}_R)_i - N_R\, f_u(u_{i+1})(\hat{u}_R)_{i+1} = -g(u_i, q_i)\hat{x}_s + (r_R)_i, \tag{4.425}$$

$i = N_L + 1, \ldots, N_L + N_R$, where $(\hat{u}_R)_{N_L+N_R+1} = 0$. As before we can rewrite these equations in the form

$$(\hat{u}_R)_i - z_i(\hat{u}_R)_{i+1} = -w_i\hat{x}_s + v_i, \quad i = N_L + 1, \ldots, N_L + N_R,$$

where

$$z_i = \frac{N_R\, f_u(u_{i+1})}{N_R\, f_u(u_i) + (x_s - 1)\, g_u(u_i, q_i)}, \quad w_i = \frac{g(u_i, q_i)}{N_R\, f_u(u_i) + (x_s - 1)\, g_u(u_i, q_i)},$$

$$v_i = \frac{(r_R)_i}{N_R\, f_u(u_i) + (x_s - 1)\, g_u(u_i, q_i)}.$$

Note that condition (4.421) implies $N_R\, f_u(u_i) + (x_s - 1)\, g_u(u_i, q_i) < 0$, $i = N_L + 1, \ldots, N_L + N_R$. If we multiply the last difference equation by $1/Z_i$, where $Z_i = \prod_{j=i}^{N_L+N_R} z_j$, $i = N_L + 1, \ldots, N_L + N_R$, $Z_{N_L+N_R+1} = 1$, then we obtain

$$\frac{1}{Z_i}(\hat{u}_L)_i - \frac{1}{Z_{i+1}}(\hat{u}_L)_{i-1} = -\frac{w_i}{Z_i}\hat{x}_s + \frac{v_i}{Z_i}, \quad i = N_L + 1, \ldots, N_L + N_R,$$

The solution of (4.425) is given by

$$(\hat{u}_R)_i = Z_i \left(\sum_{j=i}^{N_L+N_R} -\frac{w_j}{Z_j}\hat{x}_s + \frac{v_j}{Z_j} \right), \quad i = N_L + 1, \ldots, N_L + N_R. \tag{4.426}$$

In particular, it holds that

$$\begin{aligned} (\hat{u}_R)_{N_L+1} \;=\; & \frac{\prod_{k=N_L+2}^{N_L+N_R} N_R\, f_u(u_k)}{\prod_{k=N_L+1}^{N_L+N_R} N_R\, f_u(u_k) + (x_s - 1)\, g_u(u_k, q_k)} \times \\ & \Big(- \sum_{j=N_L+1}^{N_L+N_R} \prod_{k=j+1}^{N_L+N_R} \frac{N_R\, f_u(u_k) + (x_s - 1)\, g_u(u_k, q_k)}{N_R\, f_u(u_k)}\, g(u_j, q_j)\, \hat{x}_s \; + \sum_{j=N_L+1}^{N_L+N_R} \frac{v_j}{Z_j} \Big). \end{aligned}$$

The last equation $d_L^T \hat{u}_L + d_R^T \hat{u}_R = r_s$ of the system is equivalent to

$$f_u(u_{N_L})(\hat{u}_L)_{N_L} - f_u(u_{N_L+1})(\hat{u}_R)_{N_L+1} = r_s. \tag{4.427}$$

Using the expressions (4.424), (4.426) one can see that the equation (4.427) admits a unique solution $\hat{x}_s$ if and only if

$$\frac{1}{N_L}\prod_{k=1}^{N_L}\frac{N_L\, f_u(u_k)}{N_L\, f_u(u_k)+x_s\, g_u(u_k,q_k)}\sum_{j=1}^{N_L}\prod_{k=1}^{j-1}\frac{N_L\, f_u(u_k)+x_s\, g_u(u_k,q_k)}{N_L\, f_u(u_k)}\, g(u_j,q_j)$$

$$\neq \frac{1}{N_R}\prod_{k=N_L+1}^{N_L+N_R}\frac{N_R\, f_u(u_k)}{N_R\, f_u(u_k)+(x_s-1)\, g_u(u_k,q_k)}\sum_{j=N_L+1}^{N_L+N_R}\prod_{k=j+1}^{N_L+N_R}\frac{N_R\, f_u(u_k)+(x_s-1)\, g_u(u_k,q_k)}{N_R\, f_u(u_k)}\, g(u_j,q_j).$$

This condition is equivalent to (4.422). △

Remark 3 *Equation (4.422) is the discretized version of (4.390) with* $e^x \approx 1+x$ *and*

$$\int_{x_j-\frac{1}{2}h_L}^{1}\frac{x_s g_u(u_L(s),q_L(s))}{f_u(u_L(s))}\,ds \;\approx\; \sum_{k=j}^{N_L}\frac{x_s\, g_u(u_k,q_k)}{N_L\, f_u(u_k)},$$

$$\int_{x_j+\frac{1}{2}h_R}^{1}\frac{(x_s-1)g_u(u_R(s),q_R(s))}{f_u(u_R(s))}\,ds \;\approx\; \sum_{k=N_L+1}^{j}\frac{(x_s-1)\, g_u(u_k,q_k)}{N_R\, f_u(u_k)}.$$

The Lagrange function of the discretized design problem (4.415) is given by

$$\begin{aligned} L(u,x_s,q,\lambda) \;=\; & \frac{1}{2}\sum_{i=1}^{N_L}h_L\,(u_i-u^d(x_i))^2+\frac{1}{2}\sum_{i=N_L+1}^{N_L+N_R}h_R\,(u_i-u^d(x_i))^2 \\ & +\sum_{i=1}^{N_L}\lambda_i\Big(N_L\,(f(u_i)-f(u_{i-1}))+x_s\,g(u_i,q_i)\Big) \\ & +\sum_{i=N_L+1}^{N_L+N_R}\lambda_i\Big(N_R\,(-f(u_{i+1})+f(u_i))+(x_s-1)\,g(u_i,q_i)\Big) \\ & +\lambda_{N_L+N_R+1}\Big(f(u_{N_L})-f(u_{N_L+1})\Big). \end{aligned} \tag{4.428}$$

The equations $L_{(u,x_s)}(u,x_s,q,\lambda)=0$ are equivalent to

$$\Big(N_L\, f_u(u_i)+x_s\, g_u(u_i,q_i)\Big)\lambda_i - N_L\, f_u(u_i)\lambda_{i+1} \;=\; -h_L\big(u_i-u^d(x_i)\big), \tag{4.429}$$
$$i=1,\ldots,N_L-1,$$

$$\Big(N_L\, f_u(u_i)+x_s\, g_u(u_i,q_i)\Big)\lambda_i + f_u(u_i)\lambda_{N_L+N_R+1} \;=\; -h_L\big(u_i-u^d(x_i)\big), \quad i=N_L, \tag{4.430}$$

$$\Big(N_R\, f_u(u_i)+(x_s-1)\, g_u(u_i,q_i)\Big)\lambda_i - f_u(u_i)\lambda_{N_L+N_R+1} \;=\; -h_R\big(u_i-u^d(x_i)\big), \tag{4.431}$$

$$i=N_L+1, \tag{4.432}$$

$$\Big(N_R\, f_u(u_i)+(x_s-1)\, g_u(u_i,q_i)\Big)\lambda_i - N_R\, f_u(u_i)\lambda_{i-1} \;=\; -h_R\big(u_i-u^d(x_i)\big), \tag{4.433}$$
$$i=N_L+2,\ldots,N_L+N_R,$$

and

$$J^h_{x_s}(u,x_s,q)+\sum_{i=1}^{N_L}\lambda_i\, g(u_i,q_i)+\sum_{i=N_L+1}^{N_L+N_R}\lambda_i\, g(u_i,q_i)=0. \tag{4.434}$$

The system (4.429) to (4.434) is given by

$$\begin{pmatrix} B_L^T & 0 & d_L \\ 0 & B_R^T & d_R \\ e_L^T & e_R^T & 0 \end{pmatrix}\begin{pmatrix} \lambda_L \\ \lambda_R \\ \lambda_s \end{pmatrix}=\begin{pmatrix} r_L \\ r_R \\ r_s \end{pmatrix}, \tag{4.435}$$

where we used the notation

$$\lambda_L = (\lambda_1, \ldots, \lambda_{N_L}), \quad \lambda_R = (\lambda_{N_L+1}, \ldots, \lambda_{N_L+N_R}), \quad \lambda_s = \lambda_{N_L+N_R+1},$$

and

$$\begin{aligned} r_L &= \Big(-h_L(u_1 - u^d(x_1)), \ldots, -h_L(u_{N_L+1} - u^d(x_{N_L+1}))\Big), \\ r_R &= \Big(-h_R(u_{N_L+1} - u^d(x_{N_L+1})), \ldots, -h_L(u_{N_L+N_R} - u^d(x_{N_L+N_R}))\Big), \\ r_s &= -J^h_{x_s}(u, x_s, q). \end{aligned}$$

The system (4.429) to (4.434) are the adjoint equations of the discretized problem (4.415). However, the equations (4.429) to (4.434) are *not* consistent with the adjoint equations (4.400), (4.401), (4.403)

The inconsistency of the adjoint equations (4.429) to (4.434) of the discretized problem can be removed if we define

$$\begin{aligned} \tilde{\lambda}_i &= N_L \lambda_i, \quad i = 1, \ldots, N_L, \\ \tilde{\lambda}_i &= N_R \lambda_i, \quad i = N_L + 1, \ldots, N_L + N_R, \\ \tilde{\lambda}_i &= \lambda_i, \qquad i = N_L + N_R + 1. \end{aligned} \tag{4.436}$$

In the scaled Lagrange multipliers, the equations (4.429) to (4.433) are equivalent to

$$-\frac{\tilde{\lambda}_{i+1} - \tilde{\lambda}_i}{1/N_L} f_u(u_i) + x_s\, g_u(u_i, q_i)\tilde{\lambda}_i = -x_s\big(u_i - u^d(x_i)\big), \quad i = 1, \ldots, N_L - 1, \tag{4.437}$$

$$\Big(N_L\, f_u(u_i) + x_s\, g_u(u_i, q_i)\Big)\tilde{\lambda}_i + N_L f_u(u_i)\tilde{\lambda}_{N_L+N_R+1} = -x_s\big(u_i - u^d(x_i)\big), \quad i = N_L, \tag{4.438}$$

$$\Big(N_R\, f_u(u_i) + (x_s - 1)\, g_u(u_i, q_i)\Big)\tilde{\lambda}_i - N_R f_u(u_i)\tilde{\lambda}_{N_L+N_R+1} = -(1 - x_s)\big(u_i - u^d(x_i)\big), \quad i = N_L + 1, \tag{4.439}$$

$$-\frac{\tilde{\lambda}_{i-1} - \tilde{\lambda}_i}{1/N_R} f_u(u_i) + (x_s - 1)\, g_u(u_i, q_i)\tilde{\lambda}_i = -(1 - x_s)\big(u_i - u^d(x_i)\big), \quad i = N_L + 2, \ldots, N_L + N_R. \tag{4.440}$$

The equations (4.437) and (4.440) are consistent with the infinite dimensional adjoint equations (4.400) and (4.401). Equation (4.437) is an implicit scheme for (4.400) starting at $x = x_s$ and marching towards $x = 0$, the equation (4.440) is an implicit scheme for (4.401) starting at $x = x_s$ and marching towards $x = 1$.

The equations (4.438) and (4.439) are equivalent to

$$f_u(u_i)\tilde{\lambda}_i + f_u(u_i)\tilde{\lambda}_{N_L+N_R+1} + \frac{x_s}{N_L}\, g_u(u_i, q_i)\tilde{\lambda}_i = -h_L\big(u_i - u^d(x_i)\big), \quad i = N_L, \tag{4.441}$$

$$f_u(u_i)\tilde{\lambda}_i - f_u(u_i)\tilde{\lambda}_{N_L+N_R+1} + \frac{x_s - 1}{N_R}\, g_u(u_i, q_i)\tilde{\lambda}_i = -h_R\big(u_i - u^d(x_i)\big), \quad i = N_L + 1 \tag{4.442}$$

These equations are consistent with the initial conditions (4.402).

In the scaled Lagrange multipliers, equation (4.434) is written as

$$\sum_{i=1}^{N_L} \tilde{\lambda}_i \frac{1}{N_L}\, g(u_i, q_i) + \sum_{i=N_L+1}^{N_L+N_R} \tilde{\lambda}_i \frac{1}{N_R}\, g(u_i, q_i) = -J^h_{x_s}(u, x_s, q) \tag{4.443}$$

which correspond to the equation (4.403).

The system (4.437) to (4.440) and (4.443) is given by

$$\begin{pmatrix} B_L^T & 0 & \tilde{d}_L \\ 0 & B_R^T & \tilde{d}_R \\ \tilde{e}_L^T & \tilde{e}_R^T & 0 \end{pmatrix} \begin{pmatrix} \tilde{\lambda}_L \\ \tilde{\lambda}_R \\ \tilde{\lambda}_s \end{pmatrix} = \begin{pmatrix} \tilde{r}_L \\ \tilde{r}_R \\ \tilde{r}_s \end{pmatrix}, \tag{4.444}$$

where we used the notation

$$\tilde{\lambda}_L = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{N_L}), \quad \tilde{\lambda}_R = (\tilde{\lambda}_{N_L+1}, \ldots, \tilde{\lambda}_{N_L+N_R}), \quad \tilde{\lambda}_s = \tilde{\lambda}_{N_L+N_R+1},$$

and

$$\begin{aligned} \tilde{r}_L &= \Big(-x_s(u_1 - u^d(x_1)), \ldots, -x_s(u_{N_L} - u^d(x_{N_L})) \Big), \\ \tilde{r}_R &= \Big(-(1-x_s)(u_{N_L+1} - u^d(x_{N_L+1})), \ldots, -(1-x_s)(u_{N_L+N_R} - u^d(x_{N_L+N_R})) \Big), \\ \tilde{r}_s &= -J^h_{x_s}(u, x_s, q). \end{aligned}$$

The entries in the system matrices in (4.435) and (4.444) are related as follows:

$$\tilde{d}_L = N_L\, d_L, \quad \tilde{d}_R = N_R\, d_R, \quad \tilde{e}_L = \frac{1}{N_L}\, e_L, \quad \tilde{e}_R = \frac{1}{N_R}\, e_R.$$

Notice that

$$\begin{pmatrix} B_L & 0 & \tilde{e}_L \\ 0 & B_R & \tilde{e}_R \\ \tilde{d}_L^T & \tilde{d}_R^T & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{N_L} I & 0 & 0 \\ 0 & \frac{1}{N_R} I & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} B_L & 0 & e_L \\ 0 & B_R & e_R \\ d_L^T & d_R^T & 0 \end{pmatrix} \begin{pmatrix} N_L I & 0 & 0 \\ 0 & N_R I & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{4.445}$$

In particular, this relation shows that the system (4.444) is uniquely solvable if and only if (4.435) is uniquely solvable.

The previous considerations raise the question of the "correct" Lagrange multipliers. The Lagrange multipliers λ appear to be the correct ones if one starts with the discrete system. On the other hand, the Lagrange multipliers $\tilde{\lambda}$ appear to be the appropriate ones if one tries to establish a relation with the original, infinite dimensional problem. This discrepancy can be overcome, if one chooses the appropriate scalar product for the control space.

If we consider u, x_s as a function of the area q defined by the discrete state equation $C^h(u, x_s, q) = 0$, then we can write the discrete optimal control problem (4.415) in the reduced form

$$\begin{aligned} \min \quad & \hat{J}^h(q) \equiv J^h(u(q), x_s(q), q) \\ \text{s.t.} \quad & 0 \le q_{\text{low}} \le q \le q_{\text{upp}}. \end{aligned} \tag{4.446}$$

For the sake of presentation, we assume that for all q with $0 \le q_{\text{low}} \le q \le q_{\text{upp}}$ the equation $C^h(u, x_s, q) = 0$ has a unique solution. Using the implicit function theorem, the derivative of the reduced objective can shown to be

$$\hat{J}^h_q(q)\, \delta q = \Big(\nabla_q J^h(u(q), x_s(q), q) + C^h_q(u(q), x_s(q), q)\lambda \Big)^T\, \delta q.$$

Hence, the gradient of the reduced objective function with respect to the Euclidean scalar product is given by

$$\nabla_q \hat{J}^h(q) = \nabla_q J^h(u(q), x_s(q), q) + C^h_q(u(q), x_s(q), q)\, \lambda.$$

If we define the scalar product in the discretized control space to be the weighted Euclidean scalar product

$$\langle q_1, q_2 \rangle_{\mathcal{Q}_h} = \sum_{i=1}^{N_L} \frac{1}{N_L} (q_1)_i (q_2)_i + \sum_{i=N_L+1}^{N_L+N_R} \frac{1}{N_R} (q_1)_i (q_2)_i, \tag{4.447}$$

then we find that

$$\hat{J}_q^h(q)\, \delta q = \Big(\big(C_q^h(u, x_s, q)\big)^T \lambda \Big)^T \delta q = \langle \big(C_q^h(u, x_s, q)\big)^T \tilde{\lambda}, \delta q \rangle_{\mathcal{Q}_h}.$$

Thus, the gradient of the reduced objective function with respect to the weighted Euclidean scalar product is given by

$$\nabla_q \hat{J}^h(q) = \nabla_q J^h(u(q), x_s(q), q) + C_q^h(u(q), x_s(q), q)\, \tilde{\lambda}.$$

Moreover, with (4.445) it is easy to see that the system matrix in (4.444) is the adjoint of the Jacobian $C^h_{(u,x_s)}(u, x_s, q)$ with respect to a weighted scalar product. In fact, if we define

$$\langle \lambda_1, \lambda_2 \rangle_{\Lambda_h} = \sum_{i=1}^{N_L} \frac{1}{N_L} (\lambda_1)_i (\lambda_2)_i + \sum_{i=N_L+1}^{N_L+N_R} \frac{1}{N_R} (\lambda_1)_i (\lambda_2)_i + (\lambda_1)_{N_L+N_R+1} (\lambda_2)_{N_L+N_R+1}, \tag{4.448}$$

then

$$\begin{aligned}
&\left\langle \begin{pmatrix} \tilde{\lambda}_L \\ \tilde{\lambda}_R \\ \tilde{\lambda}_s \end{pmatrix}, \begin{pmatrix} B_L & 0 & e_L \\ 0 & B_R & e_R \\ d_L^T & d_R^T & 0 \end{pmatrix} \begin{pmatrix} \tilde{u}_L \\ \tilde{u}_R \\ \tilde{u}_s \end{pmatrix} \right\rangle_{\Lambda_h} \\
= &\begin{pmatrix} \tilde{\lambda}_L \\ \tilde{\lambda}_R \\ \tilde{\lambda}_s \end{pmatrix}^T \begin{pmatrix} \frac{1}{N_L} I & 0 & 0 \\ 0 & \frac{1}{N_R} I & 0 \\ 0^T & 0^T & 1 \end{pmatrix} \begin{pmatrix} B_L & 0 & e_L \\ 0 & B_R & e_R \\ d_L^T & d_R^T & 0 \end{pmatrix} \begin{pmatrix} N_L I & 0 & 0 \\ 0 & N_R I & 0 \\ 0^T & 0^T & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{N_L} I & 0 & 0 \\ 0 & \frac{1}{N_R} I & 0 \\ 0^T & 0^T & 1 \end{pmatrix} \begin{pmatrix} \tilde{u}_L \\ \tilde{u}_R \\ \tilde{u}_s \end{pmatrix} \\
= &\left\langle \begin{pmatrix} B_L^T & 0 & \tilde{d}_L \\ 0 & B_R^T & \tilde{d}_R \\ \tilde{e}_L^T & \tilde{e}_R^T & 0 \end{pmatrix} \begin{pmatrix} \tilde{\lambda}_L \\ \tilde{\lambda}_R \\ \tilde{\lambda}_s \end{pmatrix}, \begin{pmatrix} \tilde{u}_L \\ \tilde{u}_R \\ \tilde{u}_s \end{pmatrix} \right\rangle_{\Lambda_h}.
\end{aligned}$$

Note that $\langle q_1, q_2 \rangle_{\mathcal{Q}_h}$ and $\langle \lambda_1, \lambda_2 \rangle_{\Lambda_h}$ are the appropriate discretizations of the scalar products on $L^2(0,1) \times L^2(0,1)$ and $L^2(0,1) \times L^2(0,1) \times I\!R$, respectively.

Numerical Results

In our numerical experiments the discrete optimal control problem (4.415) is solved using a sequential quadratic programming (SQP) method. These methods solve the nonlinear constrained problem (4.369)–(4.375) or the corresponding discretized problem (4.415) by solving a sequence of quadratic programming problems. We used the reduced SQP methods, described previously, based upon a trust region globalization strategy. The algorithm uses limited memory BFGS updates for the reduced Hessian. The initial Hessian was chosen to be the identity and the number of updates stored will be denoted by L. In all computations, the trust region was active in the first few iterations. We also point out that since we do not add a regularization term like $\rho(\int_0^1 q_L^2 + \int_0^1 q_R^2)$ in the objective function, the reduced Hessian for the infinite dimensional problem can only expected to be positive semidefinite. The discretization sometimes has a regularizing effect and in this case for a fixed discretization the reduced Hessian for the discretized problem may be positive definite. However,

in this case the smallest eigenvalue converges towards zero as the discretization is refined. The lack of positive definiteness will of course effect the convergence behavior. We have made runs with a regularization term as shown above added to the objective function. As expected, the SQP algorithm required fewer iteration. However, we have omitted the regularization term here.

One issue which will be emphasized in this section is the influence of the relation between the infinite dimensional problem and its discretization onto the performance of the optimization method. Various studies have shown the appropriate implementation of the optimization algorithms for the discretized problems to be important. The underlying infinite dimensional problem dominates the discretized problems. If the discretized problems are treated as finite dimensional nonlinear programming problems, i.e., if the underlying infinite dimensional problem structure is ignored, the performance of the optimization algorithms usually deteriorates as the discretization is refined. The use of weighted scalar products that are obtained from the discretization of the proper scalar products of the infinite dimensional problem emphasize the underlying infinite dimensional character of these problems. The implementation of the SQP algorithm with weighted scalar products as discussed in the previous section is the proper application of the frameworks used in the previous references. In our numerical experiments reported below this leads to a substantial improvement in the performance of the algorithms. We also point out that differentiability of the functions and continuous invertibility of the linearized constraints are important conditions that have to hold in order to formulate the SQP method and to prove its local convergence in the neighborhood of strict local minimizers. For the infinite and finite dimensional version of the design problem, these properties have been established in this section.

In all our numerical computations we use the constants

$$\gamma = 1.4, \quad H = 3.6, \quad \text{which yield} \quad \bar{\gamma} = 1/6, \quad \bar{H} = 1.2\,.$$

The target velocity u^d is computed as follows: Given a cubic area function A uniquely determined by

$$A(0) = 1.05, \quad A_x(0) = 0.1, \quad A(1) = 1.745, \quad A_x(1) = 0.1\,,$$

we use a discretization scheme similar to the one described in this work to compute the corresponding velocity u as a solution of (4.345), (4.347), and (4.348) with inflow and outflow velocities given by

$$u_{\text{in}} = 1.299, \quad u_{\text{out}} = 0.506.$$

The discretization scheme applied to solve (4.345), (4.347), (4.348) also treats the shock location as an explicit variable and approximates the ODE using a scheme corresponding to (4.414). We use 200 subintervals left of the shock and 200 subintervals right of the shock to compute the target velocity. Note that for the construction of the target velocity the area A and not its logarithmic derivative is used. For the formulation of our optimal design problem we need continuous data. These are obtained by using spline interpolations. First we compute two cubic splines using the points $(x_1, u_1), \ldots, (x_{200}, u_{200})$ and $(x_{201}, u_{201}), \ldots, (x_{400}, u_{400})$ and then we join these two cubic splines by constructing a cubic polynomial interpolating $(x_{200}, u_{200}), (x_{201}, u_{201})$ and the derivatives of the two previously constructed splines at x_{200} and x_{201}, respectively. The so computed resulting target velocity u^d is continuously differentiable. Unless stated otherwise, we use the bound constraints $q_{\text{low}} = 0, q_{\text{upp}} = 1$.

The starting values for the SQP method are as follows: The initial logarithmic derivative q of the area is chosen to be $q = 0.5$. The initial shock location is computed from the target data and is chosen to be $x_s = \frac{1}{2}(x_{200} + x_{201})$. For the initial velocity we use a piecewise linear function. On the left of the shock the initial velocity is a linear interpolation between u_{in} at $x = 0$ and $u_{x_s} = 1.7$ at the initial estimate x_s of the shock location. On the right of the shock we use the linear interpolation $u_{x_s} = \bar{H}/1.7$ and u_{out}. This interpolation scheme guarantees that $u \in (\sqrt{\bar{H}}, \sqrt{2\bar{H}})$ left of the shock and $u < \sqrt{\bar{H}}$ right of the shock. If we would simply use $u_i = u^d(x_i)$, then these restrictions on u would not be satisfied if the initial shock location does not match the target shock location.

With these starting values and target data, and the discretization $N_L = N_R = 100$ the initial function value is $J^h(u, x_s, q) \approx 0.9 * 10^{-3}$ and the norm of the residual is $||C^h(u, x_s, q)||_{\Lambda_h} \approx 0.13$.

Here, the residual is computed using the weighted norm induced by (4.448). The bound constraints were never active. The necessary optimality conditions show that in this case the Lagrange multipliers $\lambda_1, \ldots, \lambda_{N_L+N_R}$ at the optimum are zero. See also (4.407), (4.408). If the truncation criteria $||C_q^T \lambda||_{I\!R^{N_L+N_R}} < \epsilon$ is used, then the Lagrange multipliers $\lambda_1, \ldots, \lambda_{N_L+N_R}$ are of the order ϵ. Analogous statements hold for the Lagrange multipliers $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{N_L+N_R}$.

In the first set of computations we study the importance of the scalar products for the numerical computations. In these computations the bounds on q were inactive. We use the weighted scalar products introduced in Section 4.8.3 and their corresponding norms. The scalar product $\langle \cdot, \cdot \rangle_{\mathcal{Q}_h}$ is used in the truncation criteria $||C_q^* \tilde{\lambda}||_{\mathcal{Q}_h} < 10^{-5}$ and, more importantly, for the computations of the BFGS updates. The scalar product $\langle \cdot, \cdot \rangle_{\Lambda_h}$ is used to compute quantities like $\langle \tilde{\lambda}, C \rangle_{\Lambda_h}$. These computations are compared with the ones in which the discretized problem (4.415) is solved as a nonlinear programming problem in $I\!R^{N_L+N_R} \times I\!R^{N_L+N_R+1}$. The truncation criteria is $||C_q^T \lambda||_{I\!R^{N_L+N_R}} < 10^{-5}$.

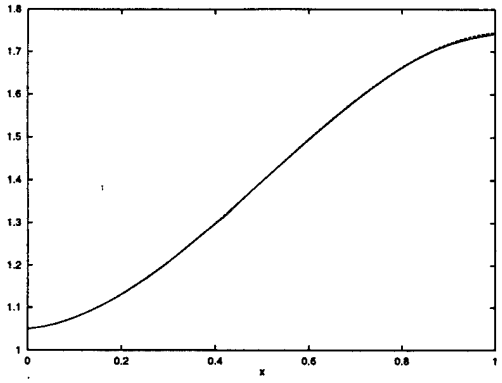
First, we observe that the SQP method with weighted scalar products requires significantly fewer iterations to converge. The results are summarized in Table 4.12 in which we compare the two SQP versions for various choices of numbers of updates stored.

Table 4.12: Number of SQP iterations versus number L of updates stored ($N_L = 100, N_R = 100$).

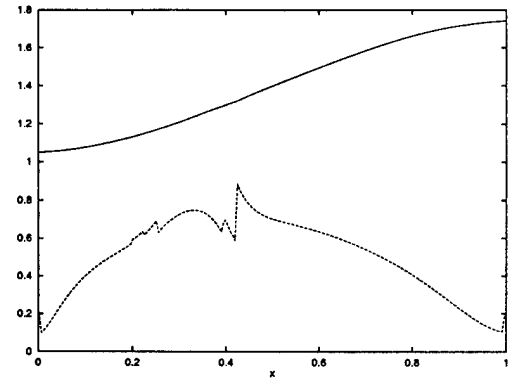
Using weighted scalar products		Using Euclidean scalar products	
L	Iterations	L	Iterations
10	68	10	88
20	54	20	75
30	45	30	52
40	45	40	52

The superiority of the SQP method with weighted scalar products over the one with Euclidean scalar products, not only shows in the number of iterations, but also in the quality of the computed solution. Typical results are shown in Figures 4.45 and 4.46. The differences between computed velocity and target velocity and between computed area and the cubic area function are significantly larger for the computations using Euclidean scalar products. Moreover, the results computed using weighted scalar products and limited memory BFGS updates with $L = 30$ or $L = 40$ are virtually identical, whereas, the logarithmic derivatives of the area functions computed using Euclidean scalar products and $L = 30$ or $L = 40$ were significantly different. This shows that the relation between the infinite dimensional problem and its discretization is not only of theoretical interest, but also promises significant advantages from a computational point of view. As we have noted before, the reason for this behavior is that the underlying infinite dimensional problem dominates the discretized problems. If the discretized problems are treated as finite dimensional nonlinear programming problems, i.e., if the underlying infinite dimensional problem structure is ignored, then the problems often become artificially ill–conditioned and the performance of the optimization algorithms usually deteriorates as the discretization is refined. In our examples, an ill–conditioning is indicated by the oscillating parameter functions q_L, q_R shown in e.g. Figure 4.46. The use of weighted scalar products that are obtained from the discretization of the proper scalar products of the infinite dimensional problem take the underlying infinite dimensional problem structure into account. The resulting implementation of the SQP method is consistent with the formulation of the SQP method in the infinite dimensional framework.

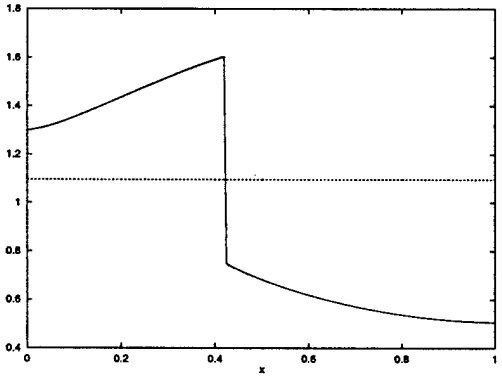
The next results concern target data with errors. Although the target data u^d was constructed using a different discretization for the area than the one used in the optimal design problem, the



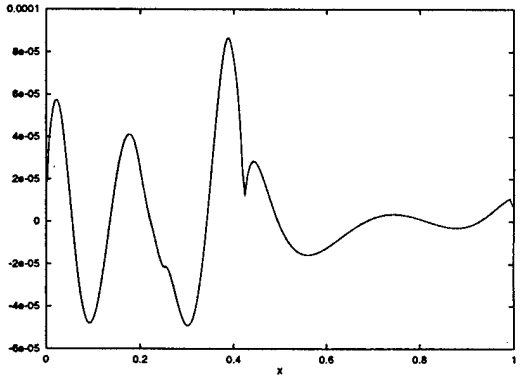
Computed area A.



Computed area A and its logarithmic derivative q.



Computed velocity, and sonic speed $\sqrt{\bar{H}}$.



Adjoint

Figure 4.45: Computed area, velocity and adjoint using weighted scalar products and $L = 40, N_L = 100, N_R = 100$.

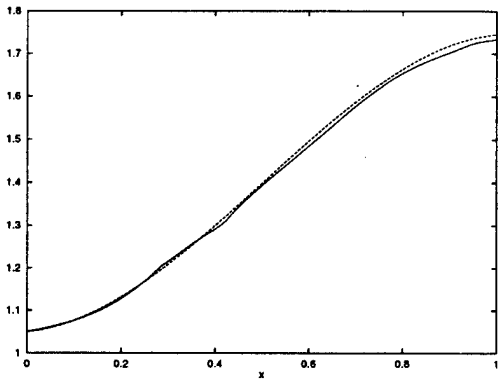
target data is almost feasible in the sense that we can find an area and a velocity profile such that $u \approx u^d$. In the following we will use nonfeasible target data. This is done by modifying the procedure for the computation of the target data used previously. Now we compute two cubic splines using the points $(x_1, u_1), \ldots, (x_{200-s}, u_{200-s})$ and $(x_{201+s}, u_{201+s}), \ldots, (x_{400}, u_{400})$ and then we join these two cubic splines by constructing a cubic polynomial interpolating $(x_{200-s}, u_{200-s}), (x_{201+s}, u_{201+s})$ and the derivatives of the two previously constructed splines at x_{200+s} and x_{201+s}, respectively. This gives target data that are "smoother" around the shock.

Computations corresponding to the following figures were done using weighted scalar products and the parameters $L = 40$ and $N_L = 100, N_R = 100$. The target data were computed using $s = 10$. The bounds $q_{\text{low}} = 0$ and $q_{\text{upp}} = 1$ were both active for some indices, as can be seen in Figure 4.47. The SQP method converged after 74 iterations. The function value at truncation was $J^h = 0.79 * 10^{-3}$, the norm of the constraints was $||C^h||_{\Lambda_h} = 0.75 * 10^{-7}$.

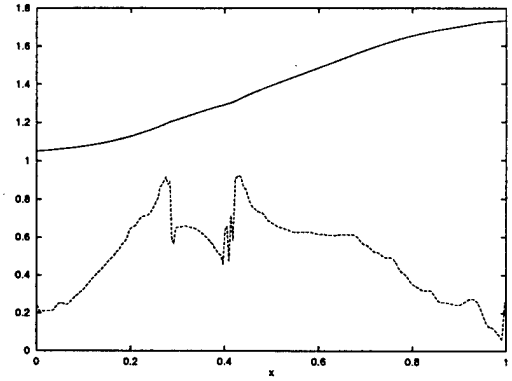
The fact that the logarithmic derivative is zero is due to the fact that the target velocity is not monotonically increasing left of the estimated shock location. In fact, if we consider the infinite dimensional problem, then the state equation (4.364) implies that

$$q_L = \frac{(1 - \bar{H}/u^2)u_x}{x_s(\bar{\gamma}u - \bar{H}/u)}.$$

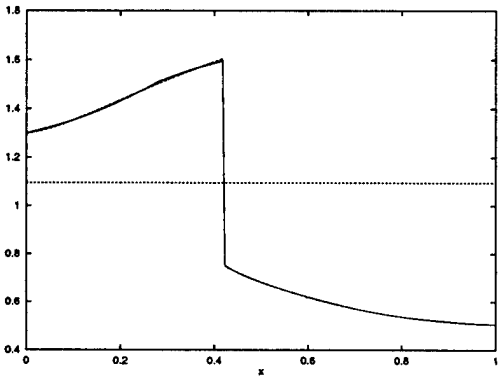
If the target velocity u^d is decreasing left of the computed shock, then, in order to be close to u^d, the computed velocity tries to imitate the nature of the target flow and, hence, the logarithmic derivative q_L of the area tries to become negative. See the previous equation for q_L. Of course, the



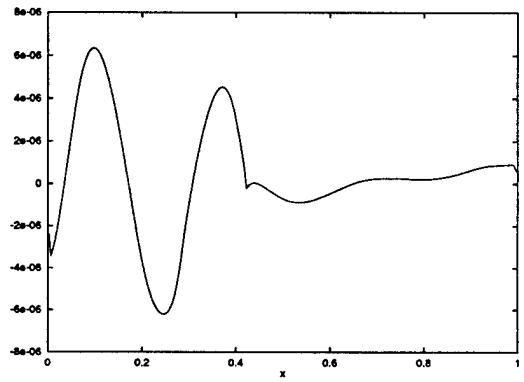
Computed area A.



Computed area A and its logarithmic derivative q.



Computed velocity, and sonic speed $\sqrt{\tilde{H}}$.



Adjoint

Figure 4.46: Computed area, velocity and adjoint using Euclidean scalar products and $L = 40, N_L = 100, N_R = 100$.

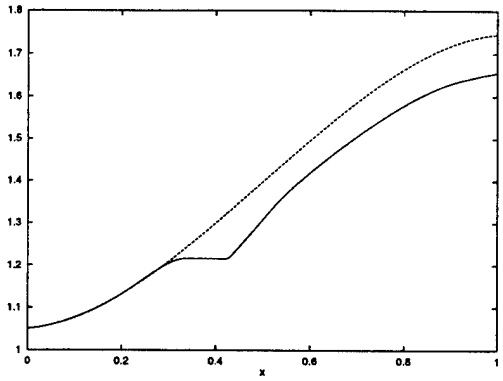
constraints prevent q_L from becoming negative. Similar reasoning can be used to explain why the logarithmic derivative of the area to the right of the computed shock, q_R is at its upper bound, q_{upp}.

These results should not be surprising, as the lagrangian (4.428) is linear in the design variable, (q_L, q_R), and optimal control theory tells us that this is a candidate for "bang-bang" control. In these cases, the bounds play an important role in the solution of the problem, as in most cases, a solution would not exist, without bounds. The region where the bound constraints on the design variables are inactive appears to correspond to a case of singular control, and we find the flows are perfectly matched in these regions. In this case the Lagrange multipliers $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{N_L+N_R}$ will generally not be zero in regions where the bounds are active, c.f. (4.407), (4.408). This behavior can be observed in Figure 4.47.

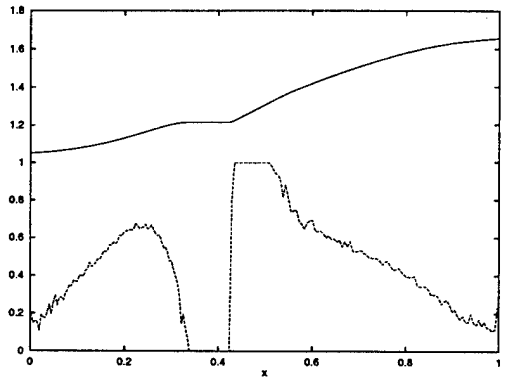
The presence of the lower bound at $q_{\text{low}} = 0$ is important in this case, for the results to make physical sense. The magnitude of the upper bound also seems to be important, as might be expected. We tried to run the same problem with $q_{\text{low}} = 0$ and $q_{\text{upp}} = 10$. However, the SQP algorithm stopped because the maximum number of iterations 100 was exceeded. The reason is that a spike evolves in the function q right of the estimated shock.

A similar situation prevails for the case where the discretization for the computed solution, N_L, N_R, exceeds the number of discrete grid points used to represent the target data. For the numerical example discussed above, when N_L or N_R exceeds 200, we find that some subintervals exist in the region around the shock where the left and right target velocities are connected by a cubic spline. Since this cubic spline causes a smoothing, effects similar to those observed with the perturbed, smooth target data discussed previously were observed, for some cases. Thus, if a target

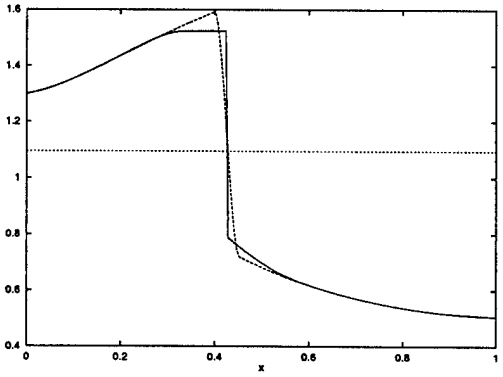
velocity similar to the one used here has to be identified, then it seems to be important that the discretization of the problem is sufficiently coarse relative to the target data.



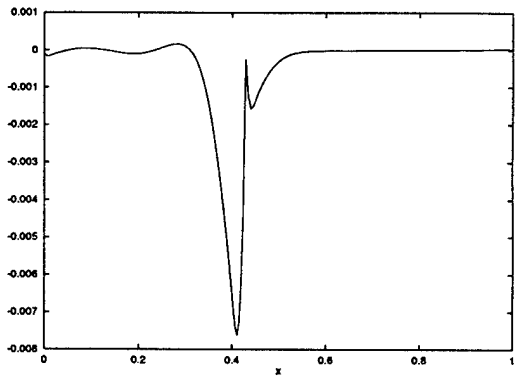
Computed versus 'exact' area A.



Computed area A and its logarithmic derivative q.



Computed velocity, and sonic speed $\sqrt{\tilde{H}}$.



Adjoint

Figure 4.47: Computed area, velocity, and adjoint using smoothed data.

4.8.4 Conclusions

In this study we have studied a design optimization problem involving a compressible flow with a shock. The differentiability of the constraint functions and the formulation of the optimality conditions in the presence of shocks is a difficult issue. Our study indicates that shock-capturing schemes with poor continuity properties result in poor convergence behavior for optimization algorithms. As an alternative, we used a formulation that treats the shock location as an explicit state variable. This allowed us to perform a rigorous mathematical analysis of the problem. We were able to sharply resolve the discontinuity while preserving differentiability of the map from design parameters to flow solution. Moreover, under suitable conditions we have established that the linearization of this map is invertible, that this inverse is uniformly bounded in a neighborhood of feasible points and that the usual first order necessary optimality conditions are valid. An important finding of our study is that the co–state is discontinuous at the shock location, unless the target velocity can be matched perfectly.

The structure of the infinite dimensional problem is inherited by its discretization. However, important observations can be made concerning the numerical solution of the discretized optimal control problem. One can view the discretized optimal control problem as a nonlinear programming problem in $\mathbb{R}^{N_L+N_R} \times \mathbb{R}^{N_L+N_R+1}$. On the other hand, one can establish the relation between the original, infinite dimensional problem and its discretization. This leads to slight reformulations

of the optimality conditions and the introduction of weighted constraints that correspond to the infinite dimensional formulation of the problem. This has been proven valuable for the numerical performance of the SQP algorithm used for the solution of the optimization problem. The use of weighted scalar products, i.e. the use of the infinite dimensional nature of the problem, reduced the number of iterations significantly and improved the quality of the computed solution.

Our results show that while a straightforward off–the–shelf application of SQP methods will likely fail, a careful analysis of the problem and an incorporation of the problem structure allows the successful application of these powerful methods. The extension of the result presented in this section to the full Euler equations is part of our ongoing research.

4.9 Airfoil Design by an All–At–Once Method

In this section we further investigate the applicability of the all–at–once formulation of the optimization problem now in the context of an airfoil design problem. In many treatments of the airfoil design problem, the flow variables q are viewed as functions of the design parameters w. This function $q(w)$ is implicitly defined by the governing equations $R(q, w) = 0$, in our case the steady state 2-D Euler equations. The optimization formulation describing the airfoil design problem is then posed in the design variables w. This is called the black-box approach. The Euler equations are not visible to the optimizer, but hidden by eliminating the flow variables, *i.e.*by expressing the flow variables q as functions of the designs variables w. An alternative to this approach is the all–at–once formulation in which one views flow variables q and design variables w as independent variables in the optimization problem. The Euler equations coupling these two are included into the optimization formulation as a constraint along with other constraints such as geometric constraints, drag constraints, etc. The optimizer is now responsible for computing a point which is feasible and optimal at the same time, *i.e.*move towards feasibility and optimality at once, rather than moving along the manifold of feasible points towards optimality. Comparisons between these two approaches on other problems have shown that the all–at–once approach can be substantially faster. The reason is that viewing q and w as independent variables, allows the optimizer to violate the Euler equations during the iterations. These are only required to be satisfied at the solution. This makes the optimization problem less nonlinear and often results in fewer iterations. Our experience indicates that an all-at-once optimization approach requires only three to four times as many iterations to solve the design problem as compared to the effort required for the solution of a single analysis problem. It is also important to note that an optimizer implementing the all–at–once approach requires roughly the same problem information as an optimizer applied to the black-box approach, except that the all–at–once approach does not require solutions to the nonlinear flow equations. We give a more detailed presentation of the relations in the next section.

Rather than formulating the airfoil design problem, its discretization, and a solution algorithm and then implement all components from scratch, we decided to build upon existing codes. In our implementation of the all–at–once method for our airfoil design problem we combine the optimizer, TRICE, with the flow code, ErICA. This imposes certain limits on the choice of problem formulation and discretization, but we believe this to be a realistic approach. As we have indicated above, various issues have to be addressed when solving the airfoil design problem. We focus on the optimization formulation. Our airfoil parameterization is obtained by choosing a set of basis airfoils and computing an optimized airfoil as a linear combination of those. Moreover, grid generation and discretization of the Euler equation was done to limit difficulties arising from nonsmoothness and inconsistencies. Further gains in efficiency and accuracy can be achieved by using more refined discretization techniques and improving the coupling of the flow solver with the optimizer. This was beyond the scope of this study and is planned for future investigations.

This paper is organized as follows: In Section 4.9.1 we discuss optimization formulations and their relations. This section also provides further motivation for the all–at–once approach and reviews some existing optimization approaches. The governing equations and the flow code ErICA are discussed in Section 4.9.1. The design problem is formulated in Section 4.9.2 and Section 4.9.3 contains a description of the optimizer TRICE. Section 4.9.4 contains some implementation issues that have to be resolved when combining an optimizer with a flow code for our situation. Section 4.9.5 presents some numerical results and contains a discussion of our numerical experiments and open issues.

4.9.1 Optimization Problem

There are several ways to cast the design problem outlined in the introduction into an optimization problem. Two formulations will be discussed in this section. The main purpose of this section is to provide a background for the discussion of our approach to the airfoil design problem and for a comparison with other approaches in the literature. In this section we proceed as follows: First,

we present the two formulations and their relation in an abstract framework. Then we discuss the applicability to the airfoil design problem.

The first formulation of the airfoil design problem is given by

$$\begin{aligned} \min \quad & J(q,w), && (4.449)\\ \text{s.t.} \quad & R(q,w) = 0, && (4.450)\\ & G(q,w) \leq 0. && (4.451) \end{aligned}$$

Here q represent the flow variables and w are the design parameters. The constraint function R represents the Euler equations. The inequality constraints (4.451) represent geometric constraints, drag constraints and the like. The special case in which G does not depend on the flow parameters q deserves attention. In this section we assume that the functions $J : \mathbb{R}^{n_q} \times \mathbb{R}^{n_w} \to \mathbb{R}$, $R : \mathbb{R}^{n_q} \times \mathbb{R}^{n_w} \to \mathbb{R}^{n_q}$, and $G : \mathbb{R}^{n_q} \times \mathbb{R}^{n_w} \to \mathbb{R}^{n_g}$ are twice continuously differentiable at the points under consideration. However, we note that for the formulation and execution of the optimization algorithm applied to our airfoil design problem, we only need first derivatives. For R, G we denote differentiation with respect to a variable by using the variable as a subscript, *e.g.*$R_q(q,w)$ denotes the partial Jacobian of R with respect to q. In addition to the differentiability assumption, we make the assumption that $R_q(q,w)$ is invertible at all points (q,w) under consideration.

Under the assumption of the implicit function theorem, the constraint (4.450) locally defines a function $q : \mathbb{R}^{n_w} \to \mathbb{R}^{n_q}$ as the solution of

$$R(q(w), w) = 0. \tag{4.452}$$

If the equation (4.450) has a unique solution $q(w)$ for all $w \in \mathbb{R}^{n_w}$ under consideration (typically, (4.451) represents an explicit restriction of the design space and therefore not the whole $\mathbb{R}^{n_w}$ is relevant), then we can eliminate the flow variables q and formulate (4.449)–(4.451) in the following reduced form:

$$\begin{aligned} \min \quad & \widehat{J}(w) = J(q(w), w), && (4.453)\\ \text{s.t.} \quad & \widehat{G}(w) = G(q(w), w) \leq 0. && (4.454) \end{aligned}$$

The optimization formulation (4.449)–(4.451) corresponds to the all–at–once (AAO) approach (also called the simultaneous analysis and design (SAND) approach). The optimization formulation (4.453), (4.454) corresponds to the black–box approach (also called the nested analysis and design (NAND)) and it corresponds to the multidiscipline feasible and individual discipline feasible approach.

In the following, we present optimality conditions for (4.449)–(4.451) and we discuss the relation between these two problems. These results are known and can be found in a similar form. Let

$$L(q, w, \lambda, \mu) = J(q,w) + \lambda^T R(q,w) + \mu^T G(q,w) \tag{4.455}$$

be the Lagrangian corresponding to (4.449)–(4.451). If a constraint qualification is met, then for an optimal point (q,w) of (4.449)–(4.451) there exist λ, μ such that

$$\begin{aligned} \nabla_q J(q,w) + R_q(q,w)^T\lambda + G_q(q,w)^T\mu &= 0,\\ \nabla_w J(q,w) + R_w(q,w)^T\lambda + G_w(q,w)^T\mu &= 0,\\ R(q,w) &= 0,\\ G(q,w) &\leq 0,\\ \mu &\geq 0,\\ G(q,w)^T\mu &= 0. \end{aligned} \tag{4.456}$$

If G does not depend on q, then the first equation in (4.456) reduces to

$$\nabla_q J(q,w) + R_q(q,w)^T \lambda = 0. \tag{4.457}$$

Equation (4.457) is called the *adjoint equation* and, if G does not depend on q, defines the Lagrange multiplier (or the co–state) λ. With λ given by (4.457), the term $\nabla_w J(q,w) + R_w(q,w)^T\lambda$ of the second equation in (4.456) is called the *reduced gradient.*

A commonly used constraint qualification is the linear independent constraint qualification (LICQ): Let $G^I(q,w)$ denote the vector of functions of $G(q,w)$ which are active at (q,w). Then LICQ is satisfied if the gradient of the component functions in $R(q,w)$ and $G^I(q,w)$ are linearly independent. If G does not depend on q and if the rows of $G^I(w)$ are linear independent, which is, *e.g.*the case if $G(w) = \pm w$, then our assumption that $R_q(q,w)$ is invertible implies that LICQ is satisfied.

The second order necessary [sufficient] optimality conditions are given by (4.456) and

$$\begin{pmatrix} s_q \\ s_w \end{pmatrix}^T H(q,w,\lambda,\mu) \begin{pmatrix} s_q \\ s_w \end{pmatrix} \underset{[>]}{\geq} 0 \tag{4.458}$$

for all s_q, s_w satisfying

$$R_q(q,w)s_q + R_w(q,w)s_w = 0, \tag{4.459}$$
$$G_q(q,w)s_q + G_w(q,w)s_w = 0. \tag{4.460}$$

Unless noted otherwise, $G(q,w)$ would usually refer to the set of active constraints, $G^I(q,w)$. Here $H(q,w,\lambda,\mu)$ denotes the Hessian of the Lagrangian

$$H(q,w,\lambda,\mu) = \nabla^2_{(q,w)} L(q,w,\lambda,\mu).$$

Points satisfying the homogeneous state equation (4.459) can be characterized by $\begin{pmatrix} s_q \\ s_w \end{pmatrix} = T(q,w)s_w$, where

$$T(q,w) = \begin{pmatrix} -R_q(q,w)^{-1}R_w(q,w) \\ I \end{pmatrix}. \tag{4.461}$$

With this, (4.458), (4.459), (4.460) can be rewritten as

$$s_w T(q,w)^T H(q,w,\lambda,\mu) T(q,w) s_w \geq [>]\, 0 \tag{4.462}$$

for all s_w satisfying

$$[-G_q(q,w)R_q(q,w)^{-1}R_w(q,w) + G_w(q,w)]s_w = 0. \tag{4.463}$$

The matrix $T(q,w)^T H(q,w,\lambda,\mu)T(q,w)$ is called the *reduced Hessian.* The term on the left hand side of (4.463) can either be computed by calculating the sensitivities $R_q(q,w)^{-1}R_w(q,w)$ or using an adjoint approach. If we define

$$\Lambda = -R_q(q(w),w)^{-T} G_q(q(w),w)^T, \tag{4.464}$$

then the left hand side of (4.463) can be written in the form $\Lambda^T R_w(q(w),w) + G_w(q(w),w)$. In particular if n_g is smaller than n_w the adjoint equation based approach seems more attractive than the sensitivity equation approach.

It is known that derivatives for the reduced problem (4.453), (4.454) are related to the reduced quantities of the problem (4.449)–(4.451). For example, the gradient $\nabla\widehat{J}(w)$ of the reduced problem is equal to the reduced gradient $\nabla_w J(q,w) + R_w(q,w)^T\lambda$, with λ given by (4.457), at $q = q(w)$. Moreover, the Hessian $\widehat{H}(w,\mu) = \nabla^2_w \widehat{L}(w,\mu)$ of the Lagrangian $\widehat{L}(w,\mu) = \widehat{J}(w) + \widehat{G}(w)^T\mu$ of the reduced problem is equal to the reduced Hessian $T(q,w)^T H(q,w,\lambda,\mu)T(q,w)$. Finally, the Jacobian $\widehat{G}_w(w)$ is equal to the matrix on the left hand side of (4.463) at $q = q(w)$.

The all–at–once approach decouples state and design variables. An optimizer for (4.449)–(4.451) can use this decoupling and is allowed to violate constraints during the iteration. This can result in substantial gains in performance. However, the optimizer must achieve feasibility and optimality at the same time. This requires carefully designed optimization codes to maintain robustness. In computational experiments (Frank and Shubin, 1992), the black–box formulation always performed more robustly than the implementation of the all–at–once approach for a one–dimensional duct design problem. Concerning the applicability of the all–at–once approach and the black–box approach, it should be mentioned that most of the quantities needed to implement the all–at–once approach also have to be provided for an implementation of the black–box approach. This is indicated by the relations between derivatives for the reduced problem (4.453), (4.454) and the reduced quantities of the problem (4.449)–(4.451) summarized above.

In the context of airfoil design problems both formulations (4.449)–(4.451) and (4.453), (4.454) have been used, however, currently the black–box formulation (4.453), (4.454) seems to be dominant. In both cases only the equality constrained problem (4.449), (4.450) is considered. The optimization methods are derived from the optimality system for (4.449), (4.450).

We use the all–at–once formulation (4.449)–(4.451). For our particular design problem the constraints G are simple constraints on the design variables. The exact problem formulation will be introduced in the subsequent sections. We use an SQP method for the solution of the all–at–once formulation. This SQP method uses an interior point strategy to handle the inequality constraints and employs a trust–region strategy for globalization of convergence and to enhance robustness. See also Section 4.9.3. If only equality constraints are present, then the Newton based methods are related to the SQP methods. Besides the capability of handling inequalities on the designs, other main differences are that the SQP methods use a trust–region globalization and, in addition to exact second derivatives, provide quasi–Newton approximations to the full and reduced Hessian of the Lagrangian. First and second order convergence results have been proved and the influence of inexact derivatives have been analyzed.

For the formulation of the airfoil design problem as an optimization problem, several other issues are of great importance. These are the issues of discretization, differentiability, and unique solvability of state equations and linearized state equations. We give a more detailed description below. For general airfoil design problems comprehensive, rigorous treatments of these issues are still missing. In the case of a one-dimensional duct design problem, which is related to the airfoil design problem, such a comprehensive, rigorous treatment can be found in Section 4.8. It is shown that an understanding of these issues can be used to improve robustness and efficiency of the optimization code. These improvements are based on the understanding of the problem, of its discretization, and of the optimization method. They are achieved with very little programming effort and almost no additional computing effort per iteration.

The airfoil design problem originally is an infinite dimensional problem. Therefore, the optimization formulation and optimization algorithm have to be combined with a discretization scheme. Various approaches are possible. Two of those are the optimize–then–discretize approach in which the optimization algorithm is formulated in the infinite dimensional setting and then discretization are applied to the individual steps, and the discretize–then–optimize approach in which one first discretizes the problem and then applies an optimization algorithm to the discretized problem. The processes of discretization and optimization are usually not interchangeable and therefore these two approaches are different. The numerical solution of an infinite dimensional problem requires a careful study of the problem at hand. Several issues have to be kept in mind. In the optimize–then–discretize approach the derivatives after discretization are usually not the derivatives of the discretized functions. Therefore optimization algorithms have to cope with inexact derivative information. See Section 4.7. The discretize–then–optimize approach often neglects the fact that the infinite dimensional problem structure still influences the finite dimensional problem. If this influence is not incorporated properly, then the optimization problem typically becomes artificially ill–conditioned and one observes a severe degradation in performance and robustness of the optimizer.

Derivatives of constraint functions and solutions q to the state equations are used in the formulation of optimality conditions and in efficient optimizers. See, for example, gradient computations using sensitivities or adjoint equations. For problems governed by the Euler equations, differentiability in the infinite dimensional context is problematic, due to the presence of shocks. This might be different for the discretized Euler equations. If smoothing procedures (*e.g.*introduction of artificial viscosity) are applied in discretization schemes for the Euler equations, the resulting finite dimensional system may be differentiable. However, since the discretization schemes used in CFD codes are very complex, 'derivatives' and 'adjoint equations' should be treated with care and usually must be understood formally.

It is also important to keep in mind that the formulations (4.449)–(4.451) and (4.453), (4.454) are only equivalent if (4.450) has a unique solution $q(w)$ for all $w \in I\!R^{n_w}$ under consideration. If $R(q,w) = 0$ represents the (discretized) Euler equations, this assumption seems to be rather strong in view of non-uniqueness results for discretized Euler equations. The existence and uniqueness of the solution q of $R(q,w) = 0$ for given w is often also related to the existence and uniqueness of the solution s_q of the linearized state equations $R_q(q,w)s_q + R_w(q,w)s_w + R(q,w) = 0$ for given $(q,w), s_w$.

As we have noted before, for a one-dimensional duct design problem, which is related to the airfoil design problem, the above issues have been rigorously discussed in Sectionrhess. For general airfoil design problems these issues are subject of current research. In our approach to the airfoil design problem we parameterize the airfoil using linear combinations of existing airfoils. This can be viewed as a reduced basis approach leading to a low dimensional ($n_w = 4$) design space. Our grid generation scheme leads to grids which depend smoothly on the design parameters. Our application programs are based on the package ErICA for the simulation of flows over airfoils governed by the Euler equations. Among the discretization schemes available in that package, we use the schemes with better smoothness properties. We use the discretize-then-optimize approach. Since we have a low dimensional design space and a rather simple grid generation scheme, we believe this is sensible. However, given our experiences, we believe this approach has to be rethought if more complex discretization schemes are used. More details on the discretization schemes are provided in Sections 4.9.1 and 4.9.4.

Analysis Problem, Discretization, and Flow Code

In this section we discuss the analysis problem underlying our design problem and its discretization. The analysis problem is the flow q around the airfoil governed by the steady state Euler equations for a perfect gas. We also outline the flow code ErICA used for the solution of the analysis problem. Although our optimization formulation is based on the all–at–once approach and our optimizer never needs to solve the Euler flow equations, we will extract several subtasks from the flow code ErICA. The presentation of the ErICA code will help to describe these tasks.

The unsteady Euler equations for a perfect gas, written in integral conservation law form is given by

$$\frac{\partial}{\partial t}\int_\Omega Q\,dS + \int_{\delta\Omega} \hat{F}\cdot\hat{n}\,ds = 0 \tag{4.465}$$

where, in Cartesian coordinates,

$$\hat{F} = \mathcal{F}\hat{j} + \mathcal{G}\hat{k}$$

and

$$Q = \begin{Bmatrix} \rho \\ \rho u \\ \rho v \\ \rho e_o \end{Bmatrix},\ \mathcal{F} = \begin{Bmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ (\rho h_o)u \end{Bmatrix},\ \mathcal{G} = \begin{Bmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ (\rho h_o)v \end{Bmatrix}$$

with velocity components u, v, density ρ, total energy per unit mass $e_o = e+(u^2+v^2)/2$, with e being the internal energy per unit mass and pressure p, which for a perfect gas may be expressed by the relation, $p = (\gamma-1)\rho e$. The total enthalpy per unit mass is given by $h_o = a^2/(\gamma-1)+u^2/2+v^2/2 = e_o + p/\rho$, where $a = \sqrt{\gamma p/\rho}$ is the sonic velocity. Here, Q represents the conserved variables with $q = [\rho\ u\ v\ p]^T$ denoting the primitive variables, and, $\mathcal{F}$ and $\mathcal{G}$ represent the inviscid fluxes. The problem domain is denoted by Ω and $\delta\Omega$ represents the boundary of the domain.

Discretization of the Euler Equations

For a given airfoil configuration, the shape of which is represented as a function of the design variables w, the analysis problem corresponds to the solution of the Euler equations of flow. The flow is simulated numerically using the solver ErICA (EuleR Inviscid Code for Aerodynamics) which was developed by Narducci.

Computational simulations were performed on a C-type grid, which is wrapped around the airfoil. We only sketch the grid generation to fix some notation. The grid is generated algebraically, by the following procedure: We first distribute points on bottom boundary, corresponding to the airfoil surface and the trailing edge wake, and on the top boundary of the computational grid, corresponding to the far-field boundary. Once the boundaries are defined, we connect corresponding pairs of points on the top and bottom boundaries using straight lines. Grid cells and nodes are numbered by (j,k), where j refers to the horizontal position and k refers to the vertical position in the grid. Indices with $k = 1$ refer to nodes or cells on or at the airfoil, respectively. A typical 201×53 grid is shown in Figures 4.48 and 4.49 with 121 points on the airfoil surface. In practical CFD codes more sophisticated grid generation schemes are used. Eventually, such grid generations have to be incorporated. However, in a first attempt to apply the all–at–once methodology to airfoil design, we preferred this simple grid because of the relative ease of generating the grid and because of its guaranteed smooth dependence on the design parameters (airfoil).

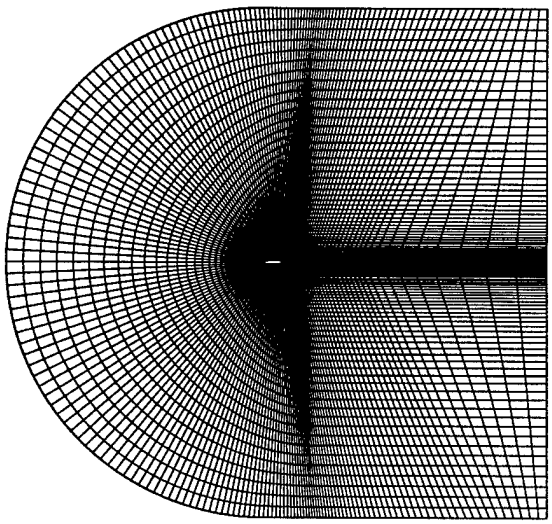


Figure 4.48: The 201×53 grid.

Given the grid, the ErICA code is used for the solution of the steady state Euler equations. In our version of ErICA a finite volume discretization using an upwind scheme with Van Leer Flux

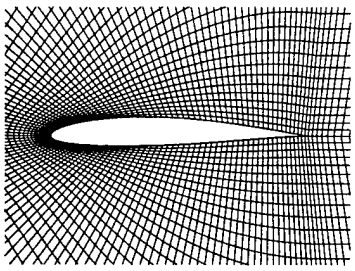


Figure 4.49: Close-up View of the 201 × 53 grid.

Vector Splitting is applied to compute the residuals. A MUSCL (Monotone Upstream-centered Scheme for Conservation Laws) differencing approach is used to interpolate the values of the state variables q from the cell centers to the cell faces. In order to fully capture the shock, we use third order interpolation of the fluxes. We use Van Albada's limiter to suppress oscillations in the flow solution. While other discretization schemes are available in ErICA, we selected these because of their better smoothness properties. See the discussion at the end of Section 4.9.1. A pseudo–time marching scheme is used to compute solution to the steady state Euler equations.

As mentioned above, we use a cell-centered, finite volume formulation to rewrite the governing equations (4.465). For the (j,k)th grid-cell the (semi-discretized) residual in terms of the primitive variables is given by

$$S_{jk} M \frac{\partial q}{\partial t} + R_{jk}(q,w) = 0, \tag{4.466}$$

where S_{jk} is the area of the (j,k)th subdomain, $M = \frac{\partial}{\partial q} Q$ is the Jacobian of the mapping between the conserved and the primitive variables, and $R_{jk} = \sum_{sides} (\hat{F} \cdot \hat{n}) \Delta s$ is the residual, where $\hat{F}$ is the inviscid flux and Δs is the length of the side. Summation is done over all sides of cell (j,k). Requiring (4.466) for all cells yields

$$SM \frac{\partial q}{\partial t} + R(q,w) = 0. \tag{4.467}$$

The residual is computed as

$$R_{jk}(q,w) = \left[(\hat{F} \cdot \hat{n}) \Delta s\right]_{j-1/2} + \left[(\hat{F} \cdot \hat{n}) \Delta s\right]_{j+1/2} + \left[(\hat{F} \cdot \hat{n}) \Delta s\right]_{k-1/2} + \left[(\hat{F} \cdot \hat{n}) \Delta s\right]_{k+1/2}, \tag{4.468}$$

where $[\hat{F} \cdot \hat{n}]_{j \pm 1/2}$ correspond to the inviscid flux, $\hat{F} \cdot \hat{n}$, across the vertical cell faces, and $[\hat{F} \cdot \hat{n}]_{k \pm 1/2}$ corresponds to the flux across the horizontal cell faces, respectively. For a given cell face, we have,

$$\hat{F} \cdot \hat{n} = \left\{ \begin{array}{c} \rho U \\ \rho u U + \hat{n}_x p \\ \rho U v + \hat{n}_y p \\ (\rho h_0) U \end{array} \right\},$$

where $U (= \hat{n}_x u + \hat{n}_y v)$ is the velocity normal to the cell face, and $\hat{n}_x$ and $\hat{n}_y$ are the Cartesian components of the normal to the cell face. As noted above, the residual is computed using an upwind scheme, with Van Leer Flux Vector Splitting, with third order interpolation via MUSCL differencing to interpolate the values of the state variables, q, from the cell centers to the cell faces, and with Van Albada's limiter to suppress oscillatory behavior in the flow solution.

The steady state solution corresponding to the semi-discrete equations (4.467) is computed using a pseudo–time marching scheme applied to (4.467). We consider the implicit scheme

$$\left[\frac{S}{\Delta t}M + \left(\widetilde{\frac{\partial R}{\partial q}}\right)^n\right]\Delta q = -R^n, \tag{4.469}$$

where $\Delta q = q^{n+1} - q^n$; $q^n = q(n\Delta t)$. Here $\widetilde{\frac{\partial R}{\partial q}}$ denotes an approximation of the Jacobian $\frac{\partial}{\partial q}R$. See below. The scheme (4.469) may be regarded as a simplified implicit Euler scheme since,

$$SM\frac{\Delta q}{\Delta t} = -R^{n+1} \approx -R^n - \left(\frac{\partial R}{\partial q}\right)^n \Delta q \approx -R^n - \left(\widetilde{\frac{\partial R}{\partial q}}\right)^n \Delta q.$$

The equation (4.469) is 'solved' by applying one step on an Alternating Direction Implicit (ADI) scheme. The resulting time–marching scheme is not accurate in time, but this is not required since we are only interested in steady state solutions.

For a moment, suppose that $(\widetilde{\frac{\partial}{\partial q}R})^n = (\frac{\partial}{\partial q}R)^n$ and that $\hat{A} = \frac{\partial}{\partial q}\hat{F}$. Then the equation of (4.469) corresponding to cell (j,k) is given by

$$\begin{aligned}\left\{\frac{S}{\Delta t}M + \left(\left[(\hat{A}\cdot\hat{n})\Delta s\right]_{j-1/2} + \left[(\hat{A}\cdot\hat{n})\Delta s\right]_{j+1/2}\right.\right.\\ \left.\left. + \left[(\hat{A}\cdot\hat{n})\Delta s\right]_{k-1/2} + \left[(\hat{A}\cdot\hat{n})\Delta s\right]_{k+1/2}\right)^n\right\}\Delta q_{jk} \;&=\; -R(q^n_{jk}).\end{aligned} \tag{4.470}$$

See (4.466), (4.468). Instead of using $\hat{A} = \frac{\partial}{\partial q}\hat{F}$, we make two simplifications to derive $\hat{A}$. These increase the efficiency with which one step of the pseudo–time marching scheme can be performed. The first simplification is as follows. In the residual computation, flux terms like $[(\hat{F}\cdot\hat{n})\Delta s]_{j+1/2}$ are calculated using Van Leer Flux Vector Splitting and MUSCL differencing with cubic interpolation of the values of the state variables q from the cell centers to the cell faces. ErICA analytically computes the Jacobians of the flux terms obtained using linear instead of cubic interpolation. The second simplification in computing $\hat{A}$ is made by partly suppressing the influence of the ghost-cells. Emphasizing the influence of the boundary conditions, the residual can be written as $R(q,w) = \check{R}(q,q_g(q),w)$, where q_g are the values of the flow variables on the ghost cells. The boundary conditions are used to express these as functions of the flow variables q in the interior and of w: $q_g = q_g(q,w)$. Thus, the derivative of the residual is of the form

$$\frac{\partial R(q,w)}{\partial q} = \frac{\partial \check{R}(q,q_g,w)}{\partial q} + \frac{\partial \check{R}(q,q_g,w)}{\partial q_g}\frac{\partial q_g(q,w)}{\partial q}. \tag{4.471}$$

In ErICA the approximation

$$\frac{\partial R(q,w)}{\partial q} \approx \frac{\partial \check{R}(q,q_g,w)}{\partial q} \tag{4.472}$$

is used to obtain $\hat{A}$.

Let $\hat{A} \approx \frac{\partial}{\partial q}\hat{F}$ denote the approximate flux-Jacobians derived using the two simplifications outlined above. If we set $\widetilde{A} = \widetilde{\frac{\partial R}{\partial q}}$, then $\widetilde{A}$ is a block pentadiagonal matrix. We split $\widetilde{A} = \widetilde{A}_k + \widetilde{A}_j$, where $\widetilde{A}_j$ corresponds to the terms $\left[(\hat{A}\cdot\hat{n})\Delta s\right]_{k-1/2} + \left[(\hat{A}\cdot\hat{n})\Delta s\right]_{k+1/2}$ in (4.470) and $\widetilde{A}_k$ corresponds to the terms $\left[(\hat{A}\cdot\hat{n})\Delta s\right]_{j-1/2} + \left[(\hat{A}\cdot\hat{n})\Delta s\right]_{j+1/2}$ in (4.470). The subscript j in $\widetilde{A}_j$ indicates that the matrix includes information of $\widetilde{A}$ along constant j–lines (vertical grid–lines). Similarly, $\widetilde{A}_k$ includes information of $\widetilde{A}$ along constant k–lines (horizontal grid–lines). We also define $T = \frac{1}{\Delta t}SM$. Now, (4.469) can be written as

$$\left[T + \widetilde{A}^n_k + \widetilde{A}^n_j\right]\Delta q = -R^n. \tag{4.473}$$

We do not solve (4.473), but approximately factor

$$T + \widetilde{A}_k^n + \widetilde{A}_j^n \approx \left[T + \widetilde{A}_k\right]^n T^{-1} \left[T + \widetilde{A}_j\right]^n$$

and solve

$$\left[T + \widetilde{A}_k\right]^n T^{-1} \left[T + \widetilde{A}_j\right]^n \Delta q = -R^n \tag{4.474}$$

The step Δq is computed by solving

$$\begin{aligned} \left[T + \widetilde{A}_k\right]^n \Delta q_{1/2} &= -R^n, \\ \left[T + \widetilde{A}_j\right]^n \Delta q &= T\,\Delta q_{1/2}. \end{aligned} \tag{4.475}$$

The new flow iterate is

$$q^{n+1} = q^n + \Delta q. \tag{4.476}$$

The two simplifications in the flux-Jacobians $\frac{\partial}{\partial q}\hat{F}$ leading to $\hat{A} \approx \frac{\partial}{\partial q}\hat{F}$ guarantee that (after symmetric permutation) the matrices on the left hand sides of (4.475) are block tridiagonal. Thus, each subproblem in (4.475) requires a block tridiagonal matrix inversion, which involves a block LU factorization and a block matrix solve; the latter consists of forward and backward substitutions. An outline of the ErICA algorithm for the solution of the governing Euler flow equations, $R(q, w) = 0$ for given w, is given in Algorithm 18.

Algorithm 18 (ErICA)

1 Given ϖ_i.

- 1.1 Generate C–grid, including ghost cells.
- 1.2 Compute direction cosines and lengths for each cell face and the areas of each cell.

2 Given q^n. Compute residual:

- 2.1 Impose Boundary Conditions.
- 2.2 Compute $R_{jk} = \sum_{\text{sides}} (\hat{F} \cdot \hat{n}) \Delta s$.
- 2.3 Compute $\|R\|$.
- 2.4 If $\|R\| < \text{tol}$, then output the result and stop; otherwise goto 3.

3 Euler Implicit Time Integration.

- 3.1 k = constant lines:
  Solve $\left[T + \widetilde{A}_k\right]^n \Delta q_{1/2} = -R^n$.
- 3.2 j = constant lines:
  Solve $\left[T + \widetilde{A}_j\right]^n \Delta q = T \Delta q_{1/2}$.
- 3.2 Update State: $q^{n+1} = q^n + \Delta q^n$. Set $n = n + 1$ and goto 2.

Airfoil Shape Parameterization

The airfoil geometry is represented as the weighted combination of six shape functions

$$y(x/c) = \sum_{i=1}^{6} \varpi_i y_i(x/c). \tag{4.477}$$

Four of the shape functions are pre-existing airfoils, namely, NACA 2412, NACA 64_1-412, NACA 65_2-415 and NACA 64_2-$A215$. The shape functions y_1–y_4 are standard. The two additional shape functions y_5, y_6 are used to impose certain geometric closure conditions at the trailing edge of the airfoil. These shapes are given by

$$y_5 = \begin{cases} x/c, & \text{on upper surface,} \\ 0, & \text{on lower surface,} \end{cases}$$

$$y_6 = \begin{cases} 0, & \text{on upper surface,} \\ -x/c, & \text{on lower surface.} \end{cases}$$

Since these functions are used to close the airfoil at the trailing edge, the weights ϖ_5 and ϖ_6 are fixed in terms of ϖ_1–ϖ_4. We require that

$$y_{us}(1) = y_{ls}(1) = 0,$$

which yield the following relations

$$\begin{aligned} \varpi_5 &= -[y_{1_{us}}(1)\varpi_1 + y_{2_{us}}(1)\varpi_2 + y_{3_{us}}(1)\varpi_3 + y_{4_{us}}(1)\varpi_4], \\ \varpi_6 &= [y_{1_{ls}}(1)\varpi_1 + y_{2_{ls}}(1)\varpi_2 + y_{3_{ls}}(1)\varpi_3 + y_{4_{ls}}(1)\varpi_4], \end{aligned}$$

where subscripts us and ls refer to the upper and lower surfaces, respectively. An efficient approach to implement the above is to close each airfoil y_1–y_4 individually to obtain $\hat{y}_1$–$\hat{y}_4$, and use these as our design bases. We have

$$y(x/c) = \sum_{i=1}^{4} \varpi_i \hat{y}_i(x/c). \tag{4.478}$$

4.9.2 The Design Problem

Given $\{\varpi_i\}$, the flow q around the airfoil is governed by the Euler equations for a perfect gas. The design problem is formulated as follows:

$$\max_{\varpi} C_L(q, \varpi) \tag{4.479}$$

such that

$$R(q, \varpi) = 0, \tag{4.480}$$

$$C_D(q, \varpi) \leq C_{D_{\max}}, \tag{4.481}$$

$$S_{\min} \leq S(\varpi) \leq S_{\max}, \tag{4.482}$$

$$\delta_{TE}(\varpi) \geq \delta_{\min}. \tag{4.483}$$

where $\varpi = \{\varpi_i\}$, and $q = [\rho\ u\ v\ p]^T$ denote the primitive variables of flow, with the usual notation. Equation (4.480) refers to the discretized steady state Euler equations of flow. The drag, C_D, in this case, is the wave drag. The lower limit on the area, S, is imposed so that the airfoil does not become too thin, a requirement for structural integrity. The upper limit is imposed to avoid thick, unrealistic airfoils. Equation (4.483) represents a bound on the trailing edge angle imposed to avoid situations in which the upper surface can go below the lower surface. See below. The free-stream conditions are based on Mach number $M = 0.75$ flow at angle of attack $\alpha = 0$.

The state equation (4.480) was discussed in Section 4.9.1. We briefly describe the computation of the aerodynamic forces used to compute $C_L(q, w)$ and $C_D(q, w)$, and the trailing edge condition (4.483). These are fairly standard, but are included for completeness.

The aerodynamic forces are computed by numerically integrating the pressure over the surface of the airfoil. The normalized forces normal and tangential to the airfoil chord line are respectively given by,

$$\begin{aligned} C_N &= -\frac{2}{\rho_\infty V_\infty} \sum_{j=j_o}^{j_f} p_{jk}(x_{j+1,1} - x_{j,1}), \\ C_T &= \frac{2}{\rho_\infty V_\infty} \sum_{j=j_o}^{j_f} p_{jk}(y_{j+1,1} - y_{j,1}), \end{aligned}$$

where (j,k), $k=1$, $j=j_o,\ldots,j_f$, denote the grid points on the airfoil surface. We compute the lift and drag forces respectively as,

$$\begin{aligned} C_L &= C_N \cos\alpha - C_T \sin\alpha, \\ C_D &= C_N \sin\alpha + C_T \cos\alpha. \end{aligned} \tag{4.484}$$

The dependence of the lift and drag coefficients on the state q and the design ϖ can be determined from (4.484).

The area of the airfoil (for unit chordlength) is given by

$$S = \int_0^1 y_{\rm us} dx - \int_0^1 y_{\rm ls} dx.$$

Here $y_{\rm us}$, $y_{\rm ls}$ represent the upper and the lower surface, respectively. Representing the airfoil in terms of the basic (closed) airfoils (4.478), we have

$$\begin{aligned} S &= \int_0^1 \sum_{i=1}^4 \varpi_i \hat{y}_{i_{\rm us}} dx - \int_0^1 \sum_{i=1}^4 \varpi_i \hat{y}_{i_{\rm ls}} dx \\ &= \sum_{i=1}^4 \varpi_i \left(\int_0^1 \hat{y}_{i_{\rm us}} dx - \int_0^1 \hat{y}_{i_{\rm ls}} dx \right) = \sum_{i=1}^4 \varpi_i S_i. \end{aligned}$$

where S_i correspond to the areas of the individual airfoils. The areas of the individual airfoils can be computed at the beginning of the design cycle. For given ϖ the area is then simply computed as the weighted sum of the areas of the given (closed) airfoils.

Our parametric representation of the airfoil (4.478) allows for situations where the upper surface can go below the lower surface of the airfoil. Such situations were actually encountered in our preliminary attempts at optimization. Hence, we need to impose an additional constraint to prevent such physically incompatible configurations to arise. This is done by constraining the trailing edge angle of the airfoil. The trailing edge angle is given by $\tan^{-1}(y'_{\rm ls}(1)) - \tan^{-1}(y'_{\rm us}(1))$, which is approximated by

$$\delta_{TE} = y'_{\rm ls}(1) - y'_{\rm us}(1).$$

Using the (approximate) trailing edge angles δ_{TE_i} of the four basic airfoils, this can be written as

$$\delta_{TE} = \sum_{i=1}^4 \varpi_i \delta_{TE_i}. \tag{4.485}$$

It was found sufficient to impose the requirement (4.483) to ensure that the upper surface does not go below the lower surface. Note that while the above approximation of the trailing edge angle is fairly crude, it does yield a constraint that is easy to compute and achieves the desired effect.

4.9.3 Optimization Algorithm

The optimization algorithm used for our computation is a version of the trust-region interior-point SQP methods called TRICE for solving

$$\begin{aligned} \min \quad & J(q,w), && (4.486)\\ \text{s.t.} \quad & R(q,w)=0, && (4.487)\\ & w_{\min} \le w \le w_{\max}. && (4.488) \end{aligned}$$

Clearly, (4.486)–(4.488) is a particular case of (4.449)–(4.451). In this section we give a brief description of the algorithm. We leave out many technical details and focus on how the algorithm interfaces with the flow solver. An important aspect of the TRICE implementation is that application specific subtasks in the optimization are separated from the optimizer TRICE and can be provided by the user. In our case this allows us to provide approximate solutions to linearized state equations and adjoint equations computed by a modification of the flow code ErICA. See also Section 4.9.4. Since (4.488) corresponds to (4.451) with a G independent of q, the Lagrange multiplier λ is determined by (4.457). We use the equation (4.457) to define $\lambda = \lambda(q,w)$.

If we define a diagonal scaling matrix $D(q,w) \in \mathbb{R}^{n_w \times n_w}$ with diagonal elements

$$\Big(D(q,w)\Big)_{ii} = \begin{cases} (w_{\max}-w)_i^{\frac{1}{2}} & \text{if} \quad \big(T(q,w)^T \nabla J(q,w)\big)_i < 0, \\ (w-w_{\min})_i^{\frac{1}{2}} & \text{if} \quad \big(T(q,w)^T \nabla J(q,w)\big)_i \ge 0, \end{cases} \tag{4.489}$$

then the first order optimality conditions (4.456) can be equivalently written as

$$\begin{aligned} & R(q,w)=0, \\ & D(q,w)^2 T(q,w)^T \nabla J(q,w) = 0, \end{aligned} \tag{4.490}$$

and $w_{\min} \le w \le w_{\max}$.

This class of algorithms generate a sequence of iterates (q_k, w_k), where w_k is strictly feasible with respect to the bounds, *i.e.*$w_{\min} < w_k < w_{\max}$ (hence the term interior–point method). The algorithms can be motivated by applying Newton's method to the system of nonlinear equations (4.490) where the w component is kept strictly feasible with respect to the bounds, *i.e.*$w_{\min} < w < w_{\max}$. The step $s = (s_q, s_w)$ is decomposed in to a quasi-normal step s^{n} and a tangential step s^{t}. The role of the quasi-normal step s^{n} is to move towards feasibility. It is of the form $s^{\mathrm{n}} = (s_q^{\mathrm{n}}, 0)$. The q-component s_q^{n} is related to the Newton step applied to solve $R(q,w_k)=0$, for given w_k. The role of the tangential step is to move towards optimality. It is of the form $s^{\mathrm{t}} = T(q_k,w_k)s_w = (-R_q(q_k,w_k)^{-1}R_w(q_k,w_k)s_w, s_w)$, where $T(q_k,w_k)$ is the representation of the null-space of the linearized state equation defined in (4.461). The w–component s_w of s^{t} is related to a quasi–Newton step for the reduced problem (4.453), (4.454).

The matrix $D(q,w)$ is in general not differentiable, but this nondifferentiability is benign and does not interfere with the fast convergence of Newton's method. A linearization of (4.490) around q_k, w_k gives

$$(R_q)_k s_q + (R_w)_k s_w = -R_k, \tag{4.491}$$

$$\Big(D_k^2 T_k^T \nabla^2_{(q,w)} \ell_k + \big[0 \mid E_k\big]\Big) \begin{pmatrix} s_q \\ s_w \end{pmatrix} = -D_k^2 T_k^T \nabla J_k. \tag{4.492}$$

Here we have used the subscript k to denote evaluation of functions at q_k, w_k. In (4.492), 0 denotes the $n_w \times n_q$ matrix with zero entries, $\ell(q,w,\lambda) = J(q,w) + \lambda^T c(q,w)$, $\nabla^2_{(q,w)}\ell(q,w,\lambda) = \frac{d}{d(q,w)}[J(q,w)+\lambda^T c(q,w)]$, and $E(q,w)$ is the $\mathbb{R}^{n_w \times n_w}$ diagonal matrix

$$\Big(E(q,w)\Big)_{ii} = \big|\big(T(q,w)^T \nabla J(q,w)\big)_i\big|$$

replacing the in general not existing term $[\frac{d}{d(q,w)}D^2(q,w)]T(q,w)^T\nabla J(q,w)$.

Since the solution of the linearized state equation (4.491) can be written as

$$s = s^{\mathrm{n}} + T_k s_w, \tag{4.493}$$

where $s^{\mathrm{n}} = (-(R_q)_k^{-1}R_k, 0)^T$ and T_k is given by (4.461).

By using (4.493) we can rewrite the linear system (4.491)–(4.492) as

$$s = s^{\mathrm{n}} + T_k s_w, \tag{4.494}$$

$$\left(D_k T_k^T \nabla^2_{xx}\ell_k T_k D_k + E_k\right) D_k^{-1} s_w = -D_k T_k^T\left(\nabla^2_{(q,w)}\ell_k s^{\mathrm{n}} + \nabla J_k\right), \tag{4.495}$$

The Newton–like step now is the solution of (4.494), (4.495) with D_k replaced by $\bar{D}_k$, where $\bar{D}_k$ is defined by (4.489) with $T_k^T\nabla J_k$ replaced by $T_k^T[\nabla^2_{(q,w)}\ell_k s^{\mathrm{n}} + \nabla J_k]$. This change of the diagonal scaling matrix is based on the form of the right hand side of (4.495).

One can that if (q_k, w_k) is close to a nondegenerate minimizer (q_*, w_*) which satisfies the second order sufficient optimality conditions, the matrix on the left hand side of (4.495) is positive definite. Therefore, (4.495) can also be interpreted as the optimality condition of a quadratic program in s_w. To globalize the convergence and to enhance robustness of the algorithm, a trust–region globalization is added. Let Δ_k be the trust radius at iteration k. The q–component of s^{n} is computed by approximately solving

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}\|(R_q)_k(s^{\mathrm{n}})_q + R_k\|^2 \\ &\text{subject to} && \|(s^{\mathrm{n}})_q\| \le \Delta_k. \end{aligned} \tag{4.496}$$

Given s^{n}, the step in w is computed by approximately solving

$$\begin{aligned} &\text{minimize} && \left(T_k^T\left(H_k s_k^{\mathrm{n}} + \nabla J_k\right)\right)^T s_w + \tfrac{1}{2}s_w^T\left(T_k^T H_k T_k + E_k\bar{D}_k^{-2}\right)s_w \\ &\text{subject to} && \|\bar{D}_k^{-1}s_w\| \le \delta_k. \end{aligned} \tag{4.497}$$

Of course, we also have to require that the new iterate is in the interior of the box constraints. To ensure that $w_k + s_w$ is strictly feasible with respect to the box constraints we choose $\sigma_k \in [\sigma, 1)$, $\sigma \in (0,1)$, and compute s_w with $\sigma_k(w_{\min} - w_k) \le s_w \le \sigma_k(w_{\min} - w_k)$. The quadratic minimization problems (4.496) and (4.497) only needs to be solved approximately. For example, an approximate solution of (4.496) is given by

$$(s^{\mathrm{n}})_q = -t[(R_q)_k]^{-1}R_k, \tag{4.498}$$

where $t = 1$ if $\|[(R_q)_k]^{-1}R_k\| \le \Delta_k$ and $t = \Delta_k/\|[(R_q)_k]^{-1}R_k\| \le \Delta_k$ otherwise.

An approximate solution of (4.497) can be computed using a modified conjugate gradient method. If the reduced Hessian $T_k^T H_k T_k$ is approximated by a quasi-Newton update, then the solution of (4.497) is relatively inexpensive. The cost of computing the 'reduced' gradient $T_k^T(H_k s_k^{\mathrm{n}} + \nabla J_k)$ dominates the cost of solving (4.497).

The main steps of the trust–region interior–point SQP scheme are outlined in algorithm 19.

We briefly sketch the information that the SQP algorithm 19 requires from the application programs. If (4.498) is used, then step 2.1 requires the solution of a linearized state equation. The computation in step 2.2 involves the solution of an adjoint equation, see the definition (4.461) of $T(q,w)$. If a quasi–Newton approximation is used to replace the the reduced Hessian $T_k^T H_k T_k$, then step 2.3 can be implemented very efficiently using a modified conjugate gradient method. The application of T_k in step 2.4 requires the solution of another linearized state equation. See (4.461). Computations involving the solution of one linearized state equation and one adjoint equation may be needed in step 2.8. This depends on the update used. The influence of inexact derivatives is important in our application since we use a pseudo–time marching (iterative) scheme to compute approximate solutions to linearized state equations and to the adjoint equations. Moreover, additional approximations, outlined in Section 4.9.1, are applied in these computations as well.

Algorithm 19 (TRICE)

1 Given H_0 and Δ_0.

2 For $k = 0, 1, 2, \ldots$ do

2.1 Compute s_k^{n} by approximately solving (4.496).

2.2 Compute $T_k^T(H_k s_k^{\mathrm{n}} + \nabla J_k)$.

2.3 Compute s_w with $\sigma_k(w_{\min} - w_k) \leq s_w \leq \sigma_k(w_{\max} - w_k)$ by approximately solving (4.497).

2.4 Compute $s = s^{\mathrm{n}} + s^{\mathrm{t}} = s^{\mathrm{n}} + T_k s_w$.

2.5 Compute λ by solving (4.457) with $q = q_k + s_q$, $w = w_k + s_w$.

2.6 Update the trust region radius Δ_k and decide if $q_k + s_q$, $w_k + s_w$ can be accepted as the new iterate.

2.7 If s is rejected set $q_{k+1} = q_k$, $w_{k+1} = w_k$, and $\lambda_{k+1} = \lambda_k$.
Otherwise, s is accepted; set $q_{k+1} = q_k + s_q$, $w_{k+1} = w_k + s_w$ and $\lambda_{k+1} = \lambda$.

2.8 If exact second order information is not used, update the (reduced) Hessian approximation.

A final remark on the handling of inequality constraints is in order. In our case the design space is small $n_w = 4$ and other approaches such as projection methods or active set methods can likely be used with similar performance to handle the inequality constraints (4.488). However, wing designs in industrial settings may involve up to 500 design parameters. In this case an interior point approach promises to be superior.

4.9.4 Numerical Implementation

Handling the Inequality Constraints and Reformulation of the Optimization Problem

The current version of TRICE only solves problems of the form (4.486)–(4.488). Hence we need to recast the design problem (4.479)–(4.482) into the form (4.486)–(4.488). This is done by handling (4.481) as a "soft" constraint using a penalty term and by transforming the design parameters.

Instead of including the drag constraint (4.481) we add a penalty term $P(\varrho(C_D/C_{D_{\max}} - 1))$ to the objective, where

$$P(z) = \begin{cases} 0 & z \leq 0, \\ z^2 & z > 0. \end{cases} \tag{4.499}$$

Here ϱ is a (scalar) penalty constant which can be used to increase emphasis on the drag violation. The addition of $P(z)$ to the objective function J has the desired effect of penalizing the objective when the drag constraint is violated. Note that this is a "soft" constraint, in the sense that the optimizer will allow the drag constraint (4.481) to be violated as long as the penalty term added is not too large. This can be addressed to some extent by controlling the penalty constant ϱ.

The remaining constraints can be addressed using a mapping between the control variable, w, and the design weights, ϖ. Rather than use the design weights as our control variables, we use the area of the airfoil and its trailing edge angle as control variables, as shown below. This enables us to address the issue of ensuring that the lower bound on the area and the trailing edge angle remain strictly enforced, by making use of the fact that we can place bounds on the control variables. We

use, as our control variables,

$$w = W \begin{Bmatrix} 2 \cdot (\varpi_1 - \varpi_3) \\ 2 \cdot (\varpi_2 - \varpi_4) \\ 2 \cdot \bar{S} \\ 0.5 \cdot \bar{\delta}_{TE} \end{Bmatrix} \tag{4.500}$$

where

$$\bar{S} = \frac{S}{S_{min}}, \qquad \bar{\delta}_{TE} = \frac{\delta_{TE}}{\delta_{min}},$$

and the scalar factor W has been added in order to be able to experiment with the scaling. Note that this results in a control space that is simply a result of a linear combination of the design weights, as should be evident from equations (4.485) and (4.485). Now, enforcing bounds

$$2W \leq w_3 \leq 2W \cdot \frac{S_{max}}{S_{min}}$$

strictly enforces the bounds on the area of the airfoil (4.482). Similarly, the bound,

$$0.5W \leq w_4$$

imposes the constraint (4.483). The mapping (4.500) can be used to translate between w and ϖ. Note that the above mapping was chosen so as to yield a low condition number for the transformation matrix providing the mapping between the weights and the controls. This was done with the philosophy that a change in the controls should produce roughly the same amount of change in the weights. This choice did yield improved performance in the optimization algorithm.

The original design problem (4.479)–(4.482) is now recast as

$$\begin{aligned} \min_{q,w} \quad & J(q,w) = -C_L(q,w) + P(\varrho(C_D(q,w)/C_{D_{max}} - 1)) \\ \text{s.t.} \quad & R(q,w) = 0, \\ & w_{\min} \leq w \leq w_{\max}, \end{aligned}$$

which is of the form (4.486)–(4.488).

Solution of the Linearized State and Adjoint Equations

To solve the design problem using Algorithm 19 described above, we need to be able to do the following:

- Provide an update s_q in the state variable, given an update in the control variable s_w by solving the linearized state constraint.
- Solve the adjoint equation at a given point.

These tasks can be performed using the same problem-solving structure applied in the flow solver ErICA. These tasks can be easily extracted from the flow code and only relatively few changes are needed.

The modification of the ErICA code to compute approximate solutions to the linearized state equation is shown in Algorithm 20. The scheme outlined in Algorithm 20 solves an approximation

$$\left(\widetilde{A}_j + \widetilde{A}_k + \widetilde{A}_g\right) s_q + \frac{\partial R}{\partial w}(q,w) s_w + R(q,w) \equiv \bar{R}(s_q, s_w, q, w) = 0 \tag{4.501}$$

of the linearized Euler equations using a pseudo-time marching scheme analogous to the one applied in ErICA. Here q, w and s_w are given and an approximate solution s_q has to be computed. Note that

$\frac{\partial}{\partial q}R(q,w)$ is replaced by $\widetilde{A}_j + \widetilde{A}_k + \widetilde{A}_g$. While in the residual computation, flux terms are calculated using Van Leer Flux Vector Splitting and MUSCL differencing with cubic interpolation of the values of the state variables q from the cell centers to the cell faces, only linear interpolation is used to calculate approximate Jacobians, see Section 4.9.1. This leads to the matrices $\widetilde{A}_j, \widetilde{A}_k$. However, boundary conditions are included in the residual computations, *cf.* (4.471). This is reflected above by the matrix $\widetilde{A}_g$. The equation (4.501) is solved by driving an *unsteady* form of the linearized Euler equations

$$SM\frac{\partial s_q}{\partial t} = -(\widetilde{A}_j + \widetilde{A}_k + \widetilde{A}_g)s_q - \frac{\partial R}{\partial w}(q,w)s_w - R(q,w) \equiv -\bar{R}(s_q, s_w, q, w) = 0 \tag{4.502}$$

towards steady state. The factor SM is added to the transient term in order to make the above equation consistent with the discretized Euler equations (*cf.* (4.467)). The pseudo-time marching scheme used is identical to the one used in marching the nonlinear Euler equations, described above in equations (4.474)–(4.476), with the nonlinear residual, R replaced by the linearized residual, $\bar{R}$. We can view this algorithm as simply an iterative method for solving the linearized state equation. Note that a relaxation factor β is used to update the solution in the iterative process. Our numerical experiments showed that using $\beta = 1.25$ yielded improved convergence rates. Also, note that there is an external loop in the iterative process monitoring the residual. This is in order to ensure that the residual does not diverge. If the norm of the residual is greater than some predetermined value, $\bar{R}_{\max}$, then the iterative process is restarted with a reduced time step Δt. This is necessitated by the fact that the Jacobians can be ill-conditioned if the solution is far from feasible, resulting in a divergent iteration. Reducing the time step had the effect of alleviating the ill-conditioning. Using $\bar{R}_{\max} = 12\|\bar{R}^0\|$ seemed adequate for our purposes. A factor of 12 is used because in some instances, the iterative process initially increased the residual but managed to recover. Clearly, these rules are somewhat ad-hoc and more sophisticated techniques could have been applied to increase efficiency. Since we are concerned with more fundamental issues arising in the all-at–once approach, optimizing performance is beyond the scope of this study. Note that Algorithm 20 only involves one Jacobian evaluation and one nonlinear residual evaluation. The Jacobian undergoes one block LU factorization and the iterative loop only involves block matrix solves, and evaluation of the linearized residual, which simply requires relatively cheap block matrix multiplications and additions.

Similarly, for the solution of the approximate adjoint equation

$$(\widetilde{A}_j + \widetilde{A}_k + \widetilde{A}_g)\lambda + + \nabla_q J(q,w) \equiv \Lambda(q,w) = 0 \tag{4.503}$$

consider a "pseudo" time dependent adjoint equation,

$$SM^T\frac{\partial \lambda}{\partial t} = -\left(\left(\widetilde{A}_j + \widetilde{A}_k + \widetilde{A}_g\right)^T \lambda + \nabla_q J(z_k)\right) \equiv -\Lambda,$$

where J is the objective. As in the solution of the linearized state equation, we replace $\frac{\partial}{\partial q}R$ by $\widetilde{A}_j + \widetilde{A}_k + \widetilde{A}_g$. We use the approximate factorization algorithm used in ErICA to iterate this equation in time, until the residual of the adjoint equation is sufficiently small, ideally $\Lambda = 0$. The procedure is as follows. We have,

$$\left[T + \widetilde{A}_j + \widetilde{A}_k\right]^T \Delta\lambda = -\Lambda^n$$

where $\widetilde{A}_j$ and $\widetilde{A}_k$ are Jacobian terms. The matrix of the left is factored approximately according to spatial directions

$$\left[T + \widetilde{A}_j\right]^T T^{-T}\left[T + \widetilde{A}_k\right]^T \Delta\lambda = -\Lambda^n$$

This system is solved using the sequence

$$\begin{aligned}\left[T + \widetilde{A}_j\right]^T \Delta\lambda_{1/2} &= -\Lambda^n \\ \left[T + \widetilde{A}_k\right]^T \Delta\lambda &= T^T\Delta\lambda_{1/2}\end{aligned} \tag{4.504}$$

Algorithm 20 (Linearized State Equation Solver)

1 Given q, w, s_w, tol.

1.1 Generate grid.

1.2 Compute: $\widetilde{A}_j(q,w)$, $\widetilde{A}_k(q,w)$, $\widetilde{A}_g(q,w)$, $R_w(q,w)$.

1.3 Compute: $\bar{R}^0 = R(q,w) + R_w(q,w)s_w$.

1.4 Set: $n = 0$, ${s_q}^n = 0$.

2 LU Decomposition.

2.1 Compute: $L_k U_k = \left[T + \widetilde{A}_k\right]$.

2.2 Compute: $L_j U_j = \left[T + \widetilde{A}_j\right]$.

3 Euler Implicit Time Integration.

3.1 k = constant lines:

Solve $L_k U_k \Delta s_{q_{1/2}} = -\bar{R}^n$.

3.2 j = constant lines:

Solve $L_j U_j \Delta s_q = T \cdot \Delta s_{q_{1/2}}$.

3.2 Update s_q: $s_q^{n+1} = s_q^n + \Delta s_q$. Set $n = n + 1$ and goto 4.

4 Compute Linearized Residual

4.1 Compute: $\bar{R}^n = \left[\widetilde{A}_j + \widetilde{A}_k + \widetilde{A}_g\right] {s_q}^n + \bar{R}^0$

4.2 Compute: $\|\bar{R}^n\|$.

If $\|\bar{R}^n\| <$ tol, set $s_q = s_q^n$. Return.

Else, if $\|\bar{R}^n\| \leq \bar{R}_{\max}$, goto 3.

Else, restart. Set: $\Delta t = 0.5\Delta t$, $n = 0$, ${s_q}^n = 0$, and goto 2.

$$\lambda^{n+1} = \lambda^n + \beta\Delta\lambda$$

The above iteration is performed until the residual of the adjoint equation, Λ, is reduced to zero. The adjoint computation is outlined in Algorithm 21. Note, that while computing the residual of the adjoint equation, and the linearized state equation, we include the terms $\widetilde{A}_g$ corresponding to the boundary conditions.

Since we are interested in the solution of the steady state adjoint equation (4.503), we could have just as well reversed the sequence above, *i.e.*solve

$$\begin{aligned} \left[T + \widetilde{A}_k\right]^T \Delta\lambda_{1/2} &= -\Lambda^n \\ \left[T + \widetilde{A}_j\right]^T \Delta\lambda &= T^T \Delta\lambda_{1/2} \end{aligned} \tag{4.505}$$

and set $\lambda^{n+1} = \lambda^n + \beta\Delta\lambda$. This would give us an algorithm which is exactly the same as that used in the linearized state algorithm, except the matrices would be transposed. However even though the pseudo-time marching is just an iterative scheme for solving (4.503), we preferred to use the transpose of the pseudo-time process (4.502) to calculate the adjoints. Numerical experiments showed that (4.504) had a slightly superior convergence behavior than (4.505).

Once again an outer loop monitors divergence of the residual. We use $\Lambda_{\max} = 12\|\Lambda^0\|$. Numerical experiments showed that the computation of the adjoint was more susceptible to producing divergent results, and hence care has to be taken in choosing the value of the relaxation factor β. We choose $\beta = \min(1.25, 1-0.12\log(10\|\Lambda^i\|))$, which has the desired effect of underrelaxing when the solution is crude, in order to reduce the possibility of divergence, and overrelaxing when the solution is refined in order to increase speed of convergence. Also, note that we do not start with $\lambda = 0$. Rather, we start from the previously computed estimate of the adjoint variable. Our experiments showed that this yielded significant savings in terms of the number of iterations. However, if the iteration proves to be divergent, then we reset λ to zero. As we have noted already for the linearized state solver, these rules are somewhat ad-hoc and more sophisticated techniques could have been applied to increase efficiency. This will be done in future studies. As with the procedure for the linearized state equation, this iteration only involves a single Jacobian evaluation.

Algorithm 21 (Adjoint Equation Solver)

1 Given q, w, tol.

1.1 Generate grid.

1.2 Compute: $\widetilde{A}_j(q,w)$, $\widetilde{A}_k(q,w)$, $\widetilde{A}_g(q,w)$, $R_w(q,w)$.

1.3 Compute: $\Lambda^0 = \nabla_q J(q,w)$.

1.4 Set: $n = 0$, $\lambda^n = \lambda_{\text{prev}}$.
(λ_{prev} is the Lagrange multiplier estimate at the previous iteration)

2 LU Decomposition.

2.1 Compute: $L_k U_k = \left[T + \widetilde{A}_k\right]$.

2.2 Compute: $L_j U_j = \left[T + \widetilde{A}_j\right]$.

3 Compute Adjoint Residual

3.1 Compute: $\Lambda^n = \left[\widetilde{A}_j + \widetilde{A}_k + \widetilde{A}_g\right]^T \lambda^n + \Lambda^0$

3.2 Compute: $\|\Lambda^n\|$.
If $\|\Lambda^n\| <$ tol, set $\lambda = \lambda^n$. Return.
Else, if $\|\Lambda^n\| > \Lambda_{\max}$, restart. Set: $\Delta t = 0.5\Delta t$, $n = 0$, $\lambda^n = 0$, goto 2.

4 Euler Implicit Time Integration.

4.1 j = constant lines:
Solve $L_j U_j \Delta\lambda_{1/2} = -\Lambda^n$.

4.2 k = constant lines:
Solve $L_k U_k \Delta\lambda = T \cdot \Delta\lambda_{1/2}$.

4.3 Update Adjoint: $\lambda^{n+1} = \lambda^n + \Delta\lambda$. Set $n = n+1$ and goto 3.

4.9.5 Numerical Results and Discussion

This section reports on some of numerical experiments conducted using the TRICE interior-point trust-region SQP optimization algorithm (see Section 4.9.3) coupled with the modification of the ErICA flow code (see Sections 4.9.1 and 4.9.3) to solve the airfoil design problem stated in Sections 4.9.2 and 4.9.4. The presentation of results is followed by a discussion of observed difficulties, possible

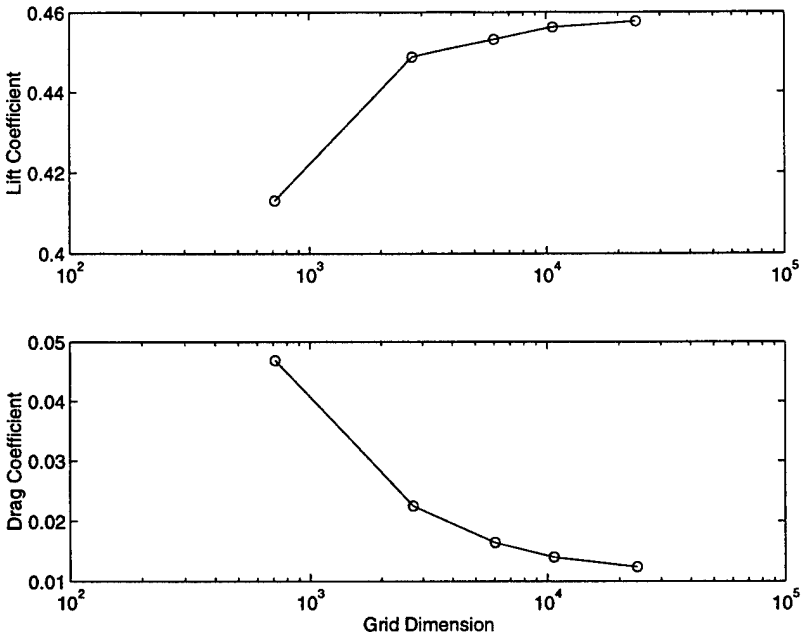
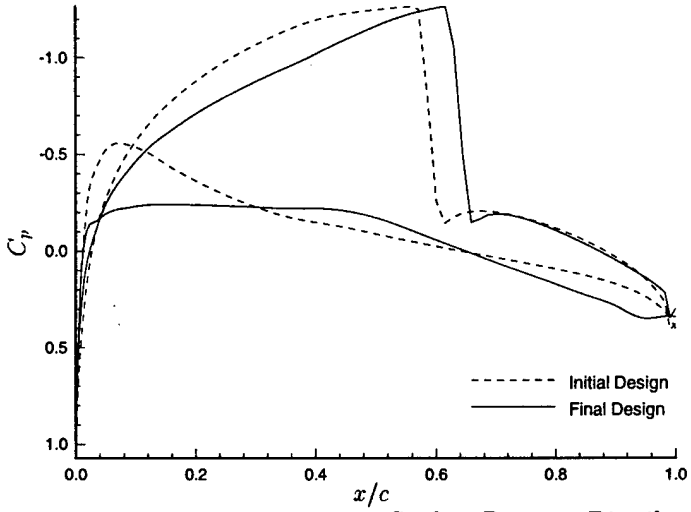


Figure 4.50: Grid Convergence Study for NACA 2412

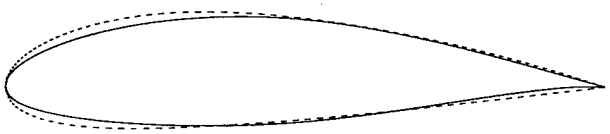
Table 4.13: Numerical Results for Airfoil Design Problem

Grid Size	$C_{D_{max}}$	Data	w_1	w_2	w_3	w_4	C_L	C_D	S
51 × 14	0.04	Initial	1.0000	0.0000	0.0000	0.0000	0.4132	0.0469	0.0823
		Final	0.2538	0.1793	-0.1369	0.5793	0.5263	0.0403	0.0770
101 × 27	0.02	Initial	0.2538	0.1793	-0.1369	0.5793	0.5722	0.0251	0.0770
		Final	0.2812	0.1049	-0.0760	0.5314	0.5227	0.0201	0.0750
151 × 40	0.014	Initial	0.2812	0.1049	-0.0760	0.5314	0.5266	0.0159	0.0750
		Final	0.3059	0.0593	-0.0313	0.5006	0.5037	0.0142	0.0750
201 × 53	0.012	Initial	0.3059	0.0593	-0.0313	0.5006	0.5044	0.0126	0.0750
		Final	0.3153	0.0429	-0.0154	0.4893	0.4955	0.0120	0.0750
301 × 79	0.010		0.3153	0.0429	-0.0154	0.4893	0.4959	0.0103	0.0750

is relatively efficient in finding the given solutions. Consider, for example, the computational effort required for the 51 × 14 grid. We require 3786 residual evaluations, 5004 Jacobian evaluations, 10008 block LU factorizations and 454948 block matrix solves. Compare this to the effort required to obtain a single analysis solution: We require approximately 1000 pseudo–time integration steps to produce a converged solution which requires 1000 residual evaluations, 1000 Jacobian evaluations, 2000 block LU factorizations and 2000 block matrix solves. Discounting the discrepancy in the number of solves, the computational effort required by TRICE is roughly equal to the effort required to perform 5–6 flow analyses, which is very cheap. It should be noted that though we



a. Surface Pressure Distribution



b. Initial airfoil (dashed) and optimal airfoil (solid)

Figure 4.51: Results obtained using TRICE for Airfoil Design Problem

require a large number of block matrix solves, this just involves forward and backward substitutions which is fairly cheap. As we have indicated earlier, this paper concerned with the feasibility of the all–at–once approach for airfoil design. Computational efficiency was not a prime concern for the interface of the ErICA flow subroutines with the TRICE optimizer. Several improvements can be made. For example, we recompute the Jacobian information every time that we require to solve the linearized state equation or the adjoint equation. Since the Jacobian will only change if the iterate (q, w) changes, a more efficient implementation would require only one Jacobian evaluation for each iteration of Algorithm 19. Moreover, our pseudo-time-stepping scheme for solving the linearized state and the adjoint equation can be improved which would lead to fewer pseudo-time-steps and fewer LU solves. Similar instances of implementing the code in a more efficient manner should yield significant savings from the elimination of redundant computations.

Earlier, we have pointed out that the optimizer struggled to achieve the desired tolerances. We now discuss possible reasons for this and also possible remedies. We distinguish among three groups. The first group is related to the information that is provided to the optimizer, the second group is related to the optimization formulation, and the third group is related to the airfoil design problem and its discretization.

The TRICE optimizer requires from the user the solution of linearized state equation and adjoint equation. We extract this information from the ErICA code. However, we use simplifications in the computation of the constrain residuals. While the residual $R(q, w)$ is evaluated using Van Leer Flux Vector Splitting and MUSCL differencing with cubic interpolation of the values of the state variables q from the cell centers to the cell faces, the 'Jacobians' of the flux terms are obtained using linear instead of cubic interpolation. Thus we only use approximate Jacobians. These approximations become better as the grid is refined, but on a given grid a certain error level cannot be removed.

Table 4.14: Computational History for Airfoil Design Problem using TRICE

Grid Size	Total Iterations	Successful Iterations	Number of Restarts	# Residual Evaluations	# Jacobian Evaluations	# LU Factorizations	# Solves
51 x 14	388*	211	12	3786	5004	10008	454948
101 x 27	44	26	1	313	437	874	73588
151 x 40	27	11	1	506	518	1036	81953
201 x 53	19	5	1	386	425	850	137768

*Iteration was stopped with a norm of the reduced gradient of $1.5 \cdot 10^{-2}$ due to lack of progress.

This explains our earlier observation that the optimizer had more problems finding a solution relative to the given tolerances on the coarse grid than on the fine grid. The Jacobians used are only asymptotically correct and the discrepancies between true Jacobians and Jacobians used become smaller as the mesh is refined. On coarse grids, we try to oversolve the problem. We expect a significant improvement in performance if true Jacobians are used. However, in that case some complications may arise from the differentiation of the Van Albada flux limiter. This matter will be investigated in depth in future research. We point out that this behavior does not contradict the ability of the optimizer to handle inexact information. The optimizer can only perform successfully for arbitrary stopping tolerances if the degree of inexactness can be adjusted by the optimizer to the progress it makes towards computing the solution and thereby to the required tolerance. One needs to adjust the stopping tolerances to the accuracy in function values. In fact, we could have relaxed the tolerance on the coarse grid and thereby reduced the number of coarse grid iterations, while maintaining the performance on the finer grids. However, since an exact error bound for the quality of the Jacobians used is not available, we believe that one tends to try to oversolve the problem. Hence, the performance displayed in Table 4.14 is what one should expect in the experimentation phase of the algorithm. These experiments can be used to define grid-dependent tolerances which will lead to a better performance than that shown in Table 4.14.

We used the optimizer TRICE because of its capability to accept solutions of linearized state equations and adjoint equations computed using application specific solvers. A reformulation of the problem presented in Section 4.9.4 was necessary, since the current version of TRICE only solves problems of the form (4.486)–(4.488). Some inefficiencies and difficulties might be attributed to this. It is expected that these will be resolved with future version of the optimizer for solving the more general problems (4.479)-(4.483). A high percentage of the number of block matrix solves required to solve the design problem (refer Table 4.14) can be attributed to the large number of iterations required to obtain a converged solution for the adjoint equation. This is especially true for the finer grids, which may be caused by ill-conditioning in the grid. We return to this issue below. The high number of solves can also be partly attributed to the penalty function approach we use to address the drag constraint (4.499). When the drag constraint is violated the gradient of the objective function with respect to the state variables becomes very large, which in turn means that the residual of the adjoint equation is very large and requires a large number of iterations to converge. The penalty term (4.499) also causes objective function to change abruptly when the drag constraint is violated. This effect is currently inadequately reflected in the model (4.497) used to compute the step and led to a large number of unsuccessful iterations.

The third group of reasons for difficulties in the solution process is somewhat related to the first one and concerns the airfoil design problem and its solution. There are known cases where nonunique solutions of the discretized Euler equations can be obtained for certain airfoils. While the all–at–once approach never requires the solution of the Euler equation, our implementation

uses the solution to the linearized equations and adjoint equations. Existence of these solutions and their dependence upon right hand side data need to be investigated. Another possible reason for the difficulties in convergence behavior is the fact that the simple algebraic grid that we use may be ill-conditioned. Since the primary purpose of the present study was to demonstrate the concept of using the all–at–once approach for solving the design problem, we have not investigated the effect of the grid on the solution process. As a result, no attempt has been made to ascertain the quality of the grid. It was, in fact, observed that the convergence behavior exhibited by the flow solver deteriorates as the grid is refined, which could be an indication of ill-conditioning. Other grid generation techniques and airfoil surface discretizations should be investigated in this context. An inclusion of such techniques in the all-at–once approach requires a careful analysis of grid sensitivities which are needed in the computation of $R_w(q,w)$ and $J_w(q,w)$. Finally, as we have pointed out in Section 4.9.1, the airfoil design problem is an infinite dimensional problem. The infinite dimensional problem, its discretization, and the optimization approach have to be analyzed jointly to derive robust and efficient solution methods. In a simplified model problem the benefits of such an analysis were demonstrated previously by the authors and were shown to lead to 10–15% reductions in optimization iterations. Provisions for the inclusion of infinite dimensional problem structure into the optimizer have been made.

4.9.6 Conclusion

We have implemented the all–at–once approach to solve an optimum airfoil design problem. The airfoil design problem was formulated as a constrained optimization problem in which flow variables and design variables are viewed as independent variables and in which the coupling steady state 2-D Euler equation is included as a constraint. To implement this approach, we have combined an existing optimization algorithm, TRICE, with an existing flow code, ErICA. Details of the implementation were given and difficulties arising in the implementation were discussed. Our numerical results indicate that the cost of solving the design problem is approximately six times the cost of solving a single analysis problem. This is consistent with the expectation that the decoupling of flow variables and design variables in the all–at–once approach makes the problem less nonlinear and can increase the efficiency with which the design problem is solved. Difficulties observed in the solution process were discussed and some future research issues were addressed.

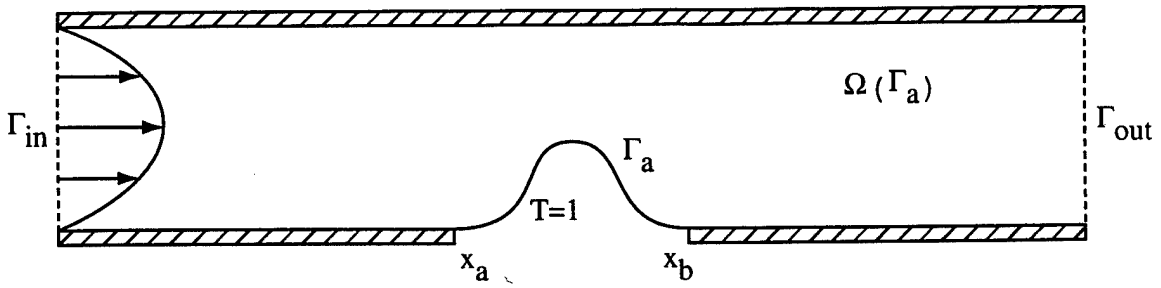


Figure 4.52: Optimal Shape Design Problem

4.10 Optimal Shape Design in Forced Convection Using Adaptive Finite Elements

In this section, we study an optimal design problem involving the shape of a channel wall. We seek the shape of the wall which maximizes the ratio of heat transferred off of it to the pressure head required to drive the flow. The problem is solved with the sensitivity equation method, coupling a trust-region optimization algorithm with gradient information supplied through the (continuous) sensitivity equation. To ensure sufficient accuracy at all of the intermediate designs, both the flow and the sensitivity equations are approximated with an adaptive finite element method, where the adaptation is performed based on error estimates using a local projection of the flow and sensitivity quantities.

4.10.1 Problem Description

Shape Design Problem

We introduce a model problem which will be used to illustrate the optimal design method discussed in the next section. This shape optimization problem, illustrated in Figure 4.52, is to find the shape of an obstruction, Γ_a, in the channel wall which maximizes the ratio of the heat transferred off of it to the pressure head required to drive the flow. For this problem, we model the forced convection using the 2D steady Navier-Stokes and energy equations,

$$\mathbf{u} \cdot \nabla \mathbf{u} = -\nabla p + \frac{1}{Re} \Delta \mathbf{u} \tag{4.506}$$

$$\nabla \cdot \mathbf{u} = 0 \tag{4.507}$$

$$\mathbf{u} \cdot \nabla T = \frac{1}{RePr} \Delta T, \tag{4.508}$$

where the non-dimensional flow quantities are the velocity vector $\mathbf{u} = (u, v)$, the pressure p and the temperature T. The two fluid parameters Re and Pr represent the Reynolds number and Prandtl number, respectively. These equations are solved on the parameter dependent domain $\Omega(\Gamma_a)$ subject to the following boundary conditions:

$$\begin{aligned} u = 4y(1-y) \quad & \text{on } \Gamma_{\text{in}}, & T = 0 \quad & \text{on } \Gamma_{\text{in}}, \\ \tau(\mathbf{u}) \cdot \hat{\mathbf{n}} - p\hat{\mathbf{n}} = 0 \quad & \text{on } \Gamma_{\text{out}}, & T = 1 \quad & \text{on } \Gamma_a, \\ u = 0 \quad & \text{on } \Gamma / (\Gamma_{\text{in}} \cup \Gamma_{\text{out}}), & k\nabla T \cdot \hat{\mathbf{n}} = 0 \quad & \text{on } \Gamma / (\Gamma_{\text{in}} \cup \Gamma_a), \end{aligned}$$

and $v = 0$ on Γ, where $\hat{\mathbf{n}}$ is the unit outward normal, k is the thermal conductivity and $\tau(\mathbf{u})$ is the fluid stress given by

$$\tau(\mathbf{u}) = \mu \left(\nabla \mathbf{u} + (\nabla \mathbf{u})^T \right).$$

We consider a parameterization of the obstruction shape,

$$\Gamma_a(a) = \left\{ (x,y) \middle| y = \frac{a}{2}\left(1 - \cos\left(2\pi \frac{x - x_a}{x_b - x_a}\right)\right),\ x \in (x_a, x_b) \right\},$$

which results in a smooth bottom surface. For brevity, further dependencies on Γ_a will be referred to using a. The solution of the above equations (4.507)-(4.508) on the parameter dependent domain implicitly defines the flow quantities as a function of the obstruction shape. This dependence will be denoted by $u(x,y;a)$, $v(x,y;a)$, etc. The set of all admissible parameters is given by

$$\mathcal{A} = \{\Gamma_a(a) |\ a \in (-1,1)\},$$

the restriction of the parameter a is used to provide a reasonable geometry.

The heat transferred off of Γ_a can be measured as

$$Q(a) = \int_{\Gamma_a} k \nabla T(\cdot;a) \cdot \hat{\mathbf{n}}(\cdot;a)\ \mathrm{d}\Gamma_a.$$

Using the integral form of the energy equation and applying the temperature boundary conditions, this can be written equivalently as

$$Q(a) = \int_{\Gamma_{\text{out}}} \rho c_p u(\cdot;a) T(\cdot;a)\ \mathrm{d}\Gamma_{\text{out}} - \int_{\Gamma_{\text{in}}} \rho c_p u(\cdot;a) T(\cdot;a)\ \mathrm{d}\Gamma_{\text{in}} = \int_{\Gamma_{\text{out}}} \rho c_p u(\cdot;a) T(\cdot;a)\ \mathrm{d}\Gamma_{\text{out}},$$

the second term vanishing due to the prescription of $T = 0$ at Γ_{in}. The pressure head required to drive the flow can be measured as

$$P(a) = \int_{\Gamma_{\text{in}}} p(\cdot;a) u(\cdot;a)\ \mathrm{d}\Gamma_{\text{in}} - \int_{\Gamma_{\text{out}}} p(\cdot;a) u(\cdot;a)\ \mathrm{d}\Gamma_{\text{out}} = \int_{\Gamma_{\text{in}}} p(\cdot;a) u(\cdot;a)\ \mathrm{d}\Gamma_{\text{in}},$$

the second term vanishing due to the imposition of $p = 0$ at Γ_{out}. We have also used the assumption that the velocity is fully developed at the inflow and outflow. These expressions allow us to define our design objective function as

$$\mathcal{J}(a) = \frac{Q(a)}{P(a)} = \frac{\int_{\Gamma_{\text{out}}} \rho c_p u(\cdot;a) T(\cdot;a)\ \mathrm{d}\Gamma_{\text{out}}}{\int_{\Gamma_{\text{in}}} p(\cdot;a) u(\cdot;a)\ \mathrm{d}\Gamma_{\text{in}}}. \tag{4.509}$$

The optimal design problem we consider is

Problem : (Optimal Shape Design)

Find the shape of the obstruction, Γ_a^* (or a^*) such that

$$\mathcal{J}(a^*) \geq \mathcal{J}(a)$$

for all $\Gamma_a \in \mathcal{A}$.

At this time, we introduce another design objective function

$$\mathcal{J}_1(a) = c_1 Q(a) - c_2 P(a) \tag{4.510}$$

for positive constants c_1 and c_2, and we also consider its maximization. The creation of an appropriate objective function is an art, one that design engineers need to consider carefully. In both $\mathcal{J}$ and $\mathcal{J}_1$, we have attempted to balance our real design objective, the maximization of $Q(a)$, with a physical constraint on the pumping work $P(a)$. Without this "constraint," the solution would be an obstruction leaving an infinitesimally small channel area bringing all of the passing fluid as close to the "hot" obstruction as possible. To produce a realistic solution, the objective has been modified to include the physical constraint. However, as we shall see, how this constraint is implemented has a dramatic impact on the optimal design.

Approximate Shape Design Problem

Since the flow and energy equations cannot be solved in closed form for this problem, we describe an approximate design problem. In this work, we considered an adaptive finite element strategy to approximate the flow and energy equations. A brief summary of this strategy is provided below for completeness and to introduce notation for later discussions. The equations are written in weak form,

$$\begin{aligned} a(\mathbf{u},\mathbf{v}) - \langle p, \nabla\cdot\mathbf{v}\rangle + \langle \mathbf{u}\cdot\nabla\mathbf{u},\mathbf{v}\rangle &= 0 && (4.511)\\ \langle \nabla\cdot\mathbf{u}, w\rangle &= 0 && (4.512)\\ d(T,\chi) + \langle \mathbf{u}\cdot\nabla T,\chi\rangle &= 0, && (4.513) \end{aligned}$$

for test functions $\mathbf{v}$, w and χ, where the bilinear forms are

$$a(\mathbf{u},\mathbf{v}) = \frac{1}{Re}\int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v}\; d\Omega$$

and

$$d(T,\chi) = \frac{1}{RePr}\int_\Omega \nabla T : \nabla\chi\; d\Omega,$$

and $\langle\cdot,\cdot\rangle$ is the standard L^2 inner product on Ω. The equations are solved using a mixed finite element method with Crouzeix-Raviart triangular elements (which have a piecewise smooth enriched quadratic basis for the velocity and temperature and a piecewise continuous bilinear basis for the pressure). The incompressibility constraint is treated with an augmented Lagrangian technique.

The result of this approximation is a set of nonlinear algebraic equations which are solved using Newton's method. Thus, the nonlinear term in (4.511) is replaced by the term

$$\langle \mathbf{u}^c\cdot\nabla\mathbf{u},\mathbf{v}\rangle + \langle \mathbf{u}\cdot\nabla\mathbf{u}^c,\mathbf{v}\rangle$$

in computing the Jacobian at the velocity $\mathbf{u}^c$ and similarly for the nonlinear term in the energy equation, which is replaced by

$$\langle \mathbf{u}^c\cdot\nabla T,\chi\rangle + \langle \mathbf{u}\cdot\nabla T^c,\chi\rangle\,.$$

Note that the same Jacobian is obtained for Newton's method regardless of when the linearization is performed (before or after the approximation).

For a given mesh (triangularization of the domain $\Omega(a)$), the resulting nonlinear system is solved for the quantities $\mathbf{u}^n$, p^n and T^n. The initial mesh which is generated for this approximation is unlikely to provide an accurate solution to the flow equations since the areas where there are sharp gradients of the flow variables are not known a priori. This may be particularly true if an automatic shape design algorithm is being constructed where minimal human intervention is used at the intermediate design calculations. Thus an adaptive mesh refinement strategy is coupled with the finite element algorithm. The strategy uses an estimate for the error on the current mesh to design a new mesh which has a prescribed error on each element, producing an "optimal mesh" which provides the best possible solution for the given computational resources. Of course, computing on this new mesh gives a better estimate for the error, so this strategy is used in an iterative process creating successively better meshes and interpolating the solutions from the previous mesh.

To completely define the adaptive strategy, we briefly discuss the error estimator and the mesh size selection criteria we use. For computational efficiency, the Zienkiewicz-Zhu local projection technique is used to project the discontinuous flow and temperature gradients and the pressure onto a continuous finite element space. The rationale is that these quantities are continuous in the exact solutions, therefore the difference between these computed quantities and their continuous projections (denoted by an overbar),

$$\left\| \left(\tau(\mathbf{u}^N) - \overline{\tau(\mathbf{u}^N)}\right) : \left(\tau(\mathbf{u}^N) - \overline{\tau(\mathbf{u}^N)}\right) \right\|_2, \quad \left\| \left(\nabla T^N - \overline{\nabla T^N}\right)\cdot\left(\nabla T^N - \overline{\nabla T^N}\right) \right\|_2$$

and

$$\left\| p^N - \overline{p^N} \right\|_2,$$

provide reasonable predictions of the finite element error. While there are many ways these error estimates can be used to define a new mesh, we use a conservative approach where a new mesh size is predicted based on each of these error estimates and the minimum mesh size is selected.

With a suitable approximation of the physical quantities available, denoted by $\mathbf{u}^N$, p^N and T^N, we now define the approximate shape optimization problem. Define

$$Q^N(a) = \int_{\Gamma_{\text{out}}} \rho c_p u^N(\cdot;a) T^N(\cdot;a)\ \mathrm{d}\Gamma_{\text{out}}$$

$$P^N(a) = \int_{\Gamma_{\text{in}}} p^N(\cdot;a) u^N(\cdot;a)\ \mathrm{d}\Gamma_{\text{in}}.$$

Then the approximate design objective function is

$$\mathcal{J}^N(a) = \frac{Q^N(a)}{P^N(a)} \qquad \text{or} \qquad \mathcal{J}_1^N(a) = c_1 Q^N(a) - c_2 P^N(a).$$

The approximate optimal design problem is then
Problem : (Approximate Optimal Shape Design)
Find the shape of the obstruction, Γ_a^* (or a^*) such that

$$\mathcal{J}^N(a^*) \geq \mathcal{J}^N(a)$$

for all $\Gamma_a \in \mathcal{A}$.

In the next section, we consider an optimal design method for solving this approximate optimal shape design problem.

4.10.2 The Sensitivity Equation Method

The sensitivity equation method is used to find the optimal parameter a^* in the approximate optimal shape design problem above. This method couples a trust-region optimization algorithm with gradient information provided by approximating the sensitivity equation, a partial differential equation describing the influence of the parameter a on the flow variables. This optimization algorithm was selected for its robustness properties, particularly for its convergence in the presence of errors in the gradient information. This robustness has been used to show convergence of the overall sensitivity equation method for some problems.

We begin by implicitly differentiating the flow equations with respect to the design parameter a. Defining $\mathbf{s}_u = \left(\frac{\partial u}{\partial a}, \frac{\partial v}{\partial a}\right)$, $s_p = \frac{\partial p}{\partial a}$ and $s_T = \frac{\partial T}{\partial a}$, we arrive at

$$\mathbf{u} \cdot \nabla \mathbf{s}_u + \mathbf{s}_u \cdot \nabla \mathbf{u} = -\nabla s_p + \frac{1}{Re} \Delta \mathbf{s}_u \tag{4.514}$$

$$\nabla \cdot \mathbf{s}_u = 0 \tag{4.515}$$

$$\mathbf{s}_u \cdot \nabla T + \mathbf{u} \cdot \nabla s_T = \frac{1}{RePr} \Delta s_T. \tag{4.516}$$

The associated boundary conditions are

$$\begin{aligned}
\mathbf{s}_u &= 0 && \text{on } \Gamma / (\Gamma_a \cup \Gamma_{\text{out}}), \\
\tau(\mathbf{s}_u) \cdot \hat{\mathbf{n}} - s_p \hat{\mathbf{n}} &= 0 && \text{on } \Gamma_{\text{out}}, \\
\mathbf{s}_u &= -\nabla \mathbf{u} \cdot (\phi_x, \phi_y)^T && \text{on } \Gamma_a, \\
s_T &= 0 && \text{on } \Gamma_{\text{in}}, \\
k \nabla s_T \cdot \hat{\mathbf{n}} &= 0 && \text{on } \Gamma / (\Gamma_{\text{in}} \cup \Gamma_a), \\
s_T &= -\nabla T \cdot (\phi_x, \phi_y)^T && \text{on } \Gamma_a.
\end{aligned}$$

The form of all of the boundary conditions are immediately obvious except for those on Γ_a. We look at the conditions for s_T, the form for $\mathbf{s}_u$ is obtained in a similar fashion. On the boundary Γ_a, $T = 1$. The value of T on this surface is fixed, regardless of the position of Γ_a. Thus, the material derivative of T on this surface with respect to a is zero, implying

$$s_T + \nabla T \cdot \left(0, \frac{1}{2} - \frac{1}{2}\cos\left(2\pi\frac{x - x_a}{x_b - x_a}\right)\right)^T \equiv s_T + \nabla T \cdot (\phi_x, \phi_y)^T = 0$$

where the vector (ϕ_x, ϕ_y) describes how the boundary coordinates change with the parameter a. This expression provides the boundary condition for s_T.

The *linear* sensitivity equations above have the same form as the linearization of the flow and energy equations leading to their efficient solution. This is because the linearization of the term $\mathbf{u} \cdot \nabla T$ is implemented in the same way regardless of the order of approximation and linearization. Furthermore, the boundary condition types (Dirichlet, Neumann) are the same for the state and sensitivity equations. Thus, using the finite element scheme to approximate these sensitivity equations leads to an efficient solver for the sensitivity information.

In this work, we extend the adaptation strategy to include the sensitivity equations. There is a trade-off here, since the sensitivity equations need to be solved on every intermediate mesh (requiring the work of one Newton iteration on that mesh). This is balanced by the fact that more accurate sensitivity information is computed.

Analogous to the flow and energy equations, the local projection technique is used to project the velocity sensitivity gradients, the pressure sensitivity and the thermal flux sensitivity onto the continuous finite element basis. Then, the following error estimates

$$\left\| \left(\tau(\mathbf{s}_u^N) - \overline{\tau(\mathbf{s}_u^N)}\right) : \left(\tau(\mathbf{s}_u^N) - \overline{\tau(\mathbf{s}_u^N)}\right) \right\|_2, \qquad \left\| \left(\nabla s_T^N - \overline{\nabla s_T^N}\right) \cdot \left(\nabla s_T^N - \overline{\nabla s_T^N}\right) \right\|_2$$

and

$$\left\| s_p^N - \overline{s_p^N} \right\|_2.$$

are used to predict new mesh sizes. Now, the minimum mesh size predicted by the three error norms above along with the three error norms for the flow and energy equations is used in constructing the next mesh.

Gradient Calculation

The gradient of the design objective function is given by

$$\frac{\partial}{\partial a}\mathcal{J}(a) = \frac{Q'(a)P(a) - Q(a)P'(a)}{P^2(a)}$$

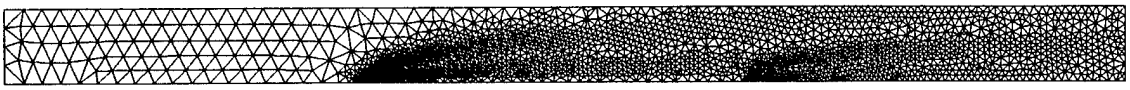
where

$$Q'(a) = \int_{\Gamma_{\text{out}}} \rho c_p \left(\frac{\partial u}{\partial a}(\cdot; a)T(\cdot; a) + u(\cdot; a)s_T(\cdot; a)\right) \, \mathrm{d}\Gamma_{\text{out}}$$

and

$$P'(a) = \int_{\Gamma_{\text{in}}} \left(s_p(\cdot; a)u(\cdot; a) + p(\cdot; a)\frac{\partial u}{\partial a}(\cdot; a)\right) \, \mathrm{d}\Gamma_{\text{in}}.$$

This can be approximated using the finite element solutions for the state and sensitivity equations described above. The resulting approximation, denoted by $\left(\frac{\partial}{\partial a}\mathcal{J}\right)^N$, is used by the trust-region optimization algorithm to produce the next iterate on the parameter a. The above expressions for $Q'(a)$ and $P'(a)$ can obviously also be used to evaluate the gradient of $\mathcal{J}_1$.

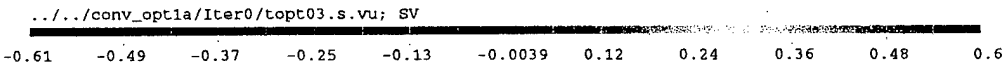


a.) Adapting for flow variables only

b.) Adapting for flow and sensitivity variables

Figure 4.53: Mesh Comparison

a.) u-velocity sensitivity



b.) v-velocity sensitivity

Figure 4.54: Velocity Sensitivities

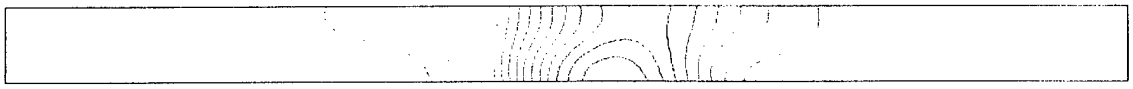
4.10.3 Numerical Results

First Design Objective

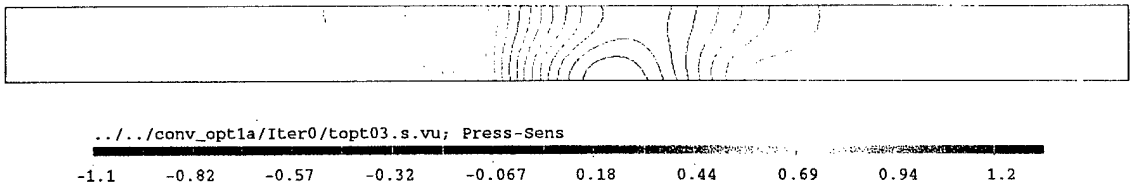
We solve the problem of maximizing $\mathcal{J}^N(a)$ described in Section 4.10.1 using the sensitivity equation method. The channel has unit width and the obstruction is located at $x_a = 5$, $x_b = 10$. Using five channel widths past the obstruction assures that we satisfy our assumption of a fully developed velocity profile at the outflow. The non-dimensional parameters were selected as $Re = 100$ and $Pr = 0.71$.

In Figure 4.53, we show the meshes which are produced with and without adaptation on the sensitivity variables. Note that for this geometry, the velocity and pressure are solved adequately with an arbitrarily coarse mesh (being quadratic and linear, respectively). Thus the only adaptation that occurs is for the temperature gradient. However, the sensitivity variables require more refinement above the curve Γ_a due to the boundary conditions, see Figures 4.54, 4.55 and 4.56. Note that the last two figures show essentially the same contours whether adaptation is performed on the sensitivity quantities or not. There is some difference above Γ_a, but this difference is lost downstream. Furthermore, as seen in Table 4.15, the sensitivity information is computed much more accurately according to our error estimators. However, this comes at a cost of not calculating the temperature as accurately.

Table 4.16 shows the convergence history of the optimization algorithm. The adaptation cycle

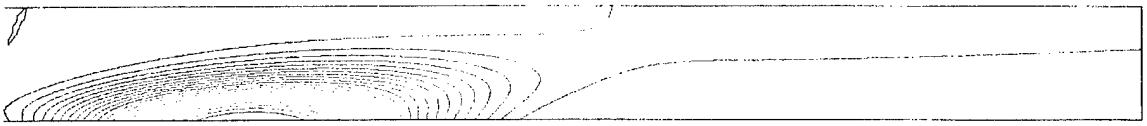


a.) Adapting for flow variables only

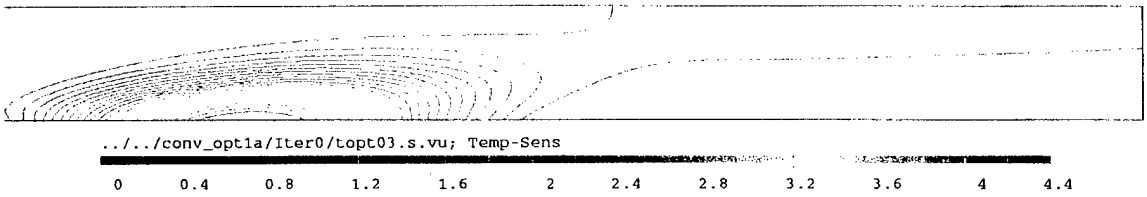


b.) Adapting for flow and sensitivity variables

Figure 4.55: Pressure Sensitivity Comparison



a.) Adapting for flow variables only



b.) Adapting for flow and sensitivity variables

Figure 4.56: Temperature Sensitivity Comparison

Table 4.15: Relative Error Estimates for Iteration 0

Adapting only on Flow Quantities							
Cycle	Nodes	$\tau(\mathbf{u})$	p	∇T	$\tau(\mathbf{s}_u)$	s_p	∇s_T
0	1105	2.897E-09	6.199E-10	1.423E-01	5.432E-02	7.920E-03	5.665E-02
1	1432	4.041E-09	6.494E-10	6.311E-02	4.352E-02	5.238E-03	4.709E-02
2	1617	4.451E-09	6.944E-10	3.527E-02	4.650E-02	6.062E-03	4.179E-02
3	2062	5.246E-09	7.357E-10	1.962E-02	3.501E-02	4.368E-03	3.142E-02
4	2893	6.777E-09	8.888E-10	1.240E-02	2.532E-02	2.887E-03	1.733E-02
5	4567	9.437E-09	1.130E-09	7.345E-03	1.561E-02	1.836E-03	1.217E-02
6	7051	1.227E-08	1.390E-09	5.281E-03	1.078E-02	1.266E-03	6.343E-03
7	11961	1.691E-08	1.825E-09	4.432E-03	6.738E-03	8.315E-04	4.334E-03

Adapting on Flow and Sensitivity Quantities							
Cycle	Nodes	$\tau(\mathbf{u})$	p	∇T	$\tau(\mathbf{s}_u)$	s_p	∇s_T
0	1105	2.897E-09	6.199E-10	1.423E-01	5.432E-02	7.920E-03	5.665E-02
1	2315	6.624E-09	8.614E-10	6.132E-02	1.574E-02	1.612E-03	1.309E-02
2	5026	1.102E-08	1.272E-09	2.775E-02	5.384E-03	6.200E-04	5.548E-03
3	11321	1.709E-08	1.927E-09	1.655E-02	1.998E-03	2.321E-04	2.423E-03

Table 4.16: Optimization Iteration History for $\mathcal{J}$

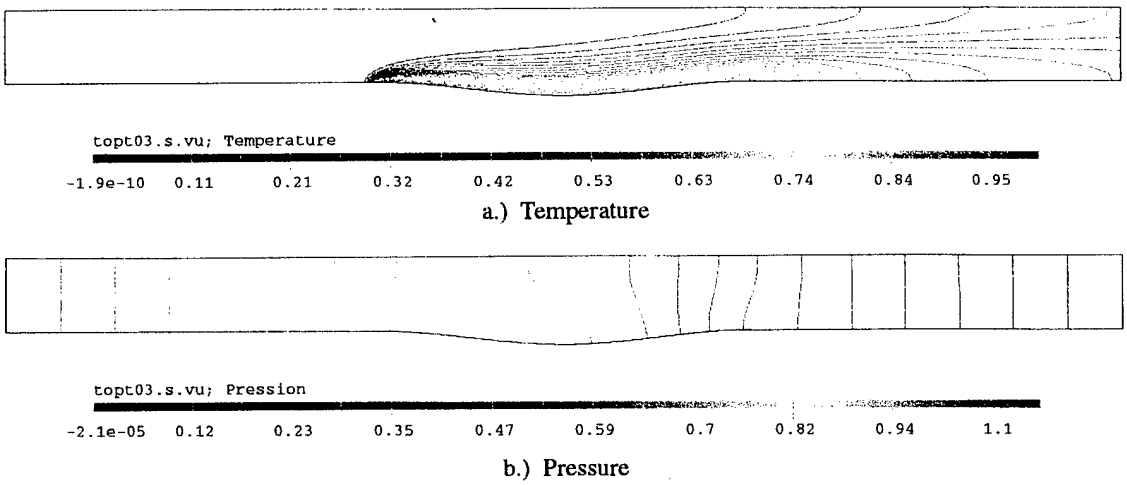
Iteration	a	$\mathcal{J}^N$	$\left(\frac{\partial}{\partial a}\mathcal{J}\right)^N$	Nodes
0	0.0000000	0.255931086	-0.0804729674	1105
		0.254790356	-0.0945435238	2315
		0.254388631	-0.0954557818	5026
		0.254263625	-0.0956619185	11321
1	-0.1430605	0.260084443	-0.0015143637	1157
		0.258715346	-0.0015915621	2451
		0.258307956	-0.0018519338	5296
		0.258189818	-0.0019365850	12518
2	-0.1793729	0.260349693	0.0113225917	1197
		0.258938717	-0.0002685903	2481
		0.258532311	-0.0036352185	5498
		0.258416015	-0.0045778374	12902
3	-0.1906139	0.260388891	0.0127604991	1205
		0.258958392	0.0024282222	2565
		0.258546684	0.0005718905	5484
		0.258431073	-0.0002657475	13056
4	-0.1913066	0.260385697	0.0130685564	1205
		0.258953119	0.0045400995	2545
		0.258542547	0.0010437568	5653
		0.258426702	0.0000278092	13352

was used to obtain about four significant digits in the objective function evaluation. The gradient was computed accurately to about 0.0001. We see that we run into the limits of our accuracy at the third iteration. The objective function does not change in the fourth significant digit and the gradient which is computed is about the same order as our error tolerance.

We plot the temperature and pressure contours for the third iteration in Figure 4.57. Note that although we have increased our design objective function, we did so by affecting the "pressure constraint" rather than by increasing the heat transferred off of the "obstruction." In fact Q decreased from 0.203 to 0.192, but P decreased further from 0.800 to 0.744. This motivated us to consider a different design objective.

Second Design Objective

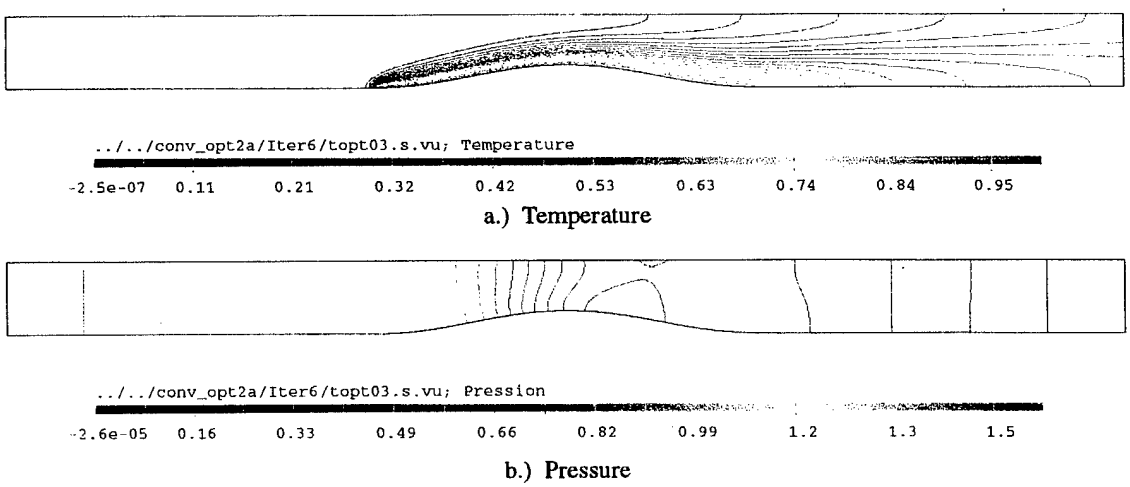
In this section, we consider the maximization of $\mathcal{J}_1(a)$ using the values $c_1 = 15$ and $c_2 = 1$. Applying the sensitivity equation method to this case produced the iteration history given in Table 4.17. Note that as above, there are about four significant digits in the objective function calculation, while the gradient is accurate to about 0.01. Thus, while the third to the seventh iterations produce little change in the objective function, the optimization algorithm proceeds until failure after the sixth



a.) Temperature

b.) Pressure

Figure 4.57: Temperature and Pressure at Optimal Shape



a.) Temperature

b.) Pressure

Figure 4.58: Temperature and Pressure at Optimal Shape

iteration, where the inaccuracy in the gradient produces an invalid search direction.

The plots of temperature and pressure at the sixth iteration are given in Figure 4.58. Note that in this case we get a real obstruction which does increase the heat that is being convected off of it (from $Q = 0.203$ to $Q = 0.228$). This however comes at a cost in P (from 0.800 to 1.044). Obviously a family of shapes is determined by changing the ratio of c_1 to c_2.

Table 4.17: Optimization Iteration History for $\mathcal{J}_1$

Iteration	a	$\mathcal{J}_1^N$	$\left(\frac{\partial}{\partial a}\mathcal{J}_1\right)^N$	Nodes
0	0.0000000	2.271173030	1.3064437434	1105
		2.257484269	1.1235263540	2315
		2.252663573	1.1081328015	5026
		2.251163495	1.1041584585	11321
1	0.1182964	2.332193132	0.8366165676	1105
		2.318839773	0.8678131107	2443
		2.313896697	0.9097510136	5394
		2.312315720	0.9306183575	12717
2	0.2365927	2.375239985	0.3257817000	1113
		2.363818277	0.3428584385	2384
		2.359223826	0.4864326970	5351
		2.357770739	0.5206597715	12881
3	0.3021628	2.387989142	-0.4680732831	1089
		2.375589441	-0.1528290993	2163
		2.370743613	0.0193999924	5124
		2.369002747	0.0819619660	12435
4	0.3144133	2.387976109	-0.6790688256	1113
		2.375312658	-0.3908063610	2268
		2.370630444	-0.1000848872	5164
		2.369235810	-0.0262166400	12560
5	0.3114445	2.387795285	-0.5065145289	1093
		2.375820227	-0.2045254290	2251
		2.370857409	-0.0818018460	5193
		2.369237831	-0.0070671398	12293
6	0.3103488	2.389464148	-0.7051882603	1093
		2.376070276	-0.2654940962	2210
		2.370971444	-0.0658611218	5302
		2.369269552	0.0116997992	12474
7	0.3110319	2.389018166	-0.8444888935	1105
		2.375518023	-0.2885576888	2314
		2.370771675	-0.0753567197	5319
		2.369228213	0.0002796022	12922

Chapter 5

Personnel Supported

The following people were supported in part under AFOSR Grant F49620-93-1-0280:

5.1 Senior Investigators

H.T. Banks - Professor, Department of Mathematics, North Carolina State University
John A. Burns - Professor, Department of Mathematics, Virginia Tech
Eugene M. Cliff - Professor, Aerospace and Ocean Engineering Department, Virginia Tech
Harley H. Cudney - Assistant Professor, Mechanical Engineering Department, Virginia Tech
Max Gunzburger - Professor, Department of Mathematics, Virginia Tech
Matthias Heinkenschloss - Assistant Professor, Department of Mathematics, Virginia Tech
Terry Herdman - Professor, Department of Mathematics, Virginia Tech
Daniel Inman - Professor, Engineering Science and Mechanics Department, Virginia Tech
Janet Peterson - Associate Professor, Department of Mathematics, Virginia Tech

5.2 Associate Investigators

Chris Beattie - Department of Mathematics, Virginia Tech
Mary Bradley - Department of Mathematics, Brown University
Ferenc Hartung - Instructor, Department of Mathematics, Virginia Tech
Belinda King - Assistant Professor, Department of Mathematics, Oregon State University
Robert Miller - Assistant Professor, Department of Mathematical Sciences, University of Arkansas
Robert Rogers - Associate Professor, Department of Mathematics, Virginia Tech
David Stewart - Assistant Professor, Department of Mathematics, Virginia Tech

5.3 Post Doctoral Fellows

Jeff Borggaard - Research Assistant Professor, Department of Mathematics
Paul Gilmore - Research Assistant Professor, Department of Mathematics
Yuh Roung Ou - Research Assistant Professor, Aerospace and Ocean Engineering Department
A. Diana Rubio - Research Assistant Professor, Department of Mathematics
Xiaonan Wu - Research Assistant Professor, Department of Mathematics

5.4 Graduate Students

Justin Appel - Department of Mathematics
Thomas Bail - Department of Mathematics
Pavel Bochev - Department of Mathematics
Jeff Borggaard - Department of Mathematics
Mark Borisuk - Department of Mathematics
John Burkardt - Department of Mathematics
Yanzhao Cao - Department of Mathematics
Graciela Cerezo - Department of Mathematics
Shihchung Chiang - Department of Mathematics
Joe Colvin - Mechanical Engineering
Jennifer Deang - Department of Mathematics
Gregory W. Diehl - Mechanical Engineering
Wei Huang - Department of Mathematics
Kyehong Kang - Department of Mathematics
Rossitza Karamikhowa - Department of Mathematics
Martin Khumbah - Department of Mathematics
Hong-Chul Kim - Department of Mathematics
Karen Klaimon - Department of Mathematics
Hyesuk Kwon - Department of Mathematics
Hyung-Chun Lee - Department of Mathematics
Eric Lengyel - Department of Mathematics
Donald Leo - Engineering Science and Mechanics
George Moss - Department of Mathematics
Shana Olds - Department of Mathematics
Steven Pugh - Department of Mathematics
Shaohong Qu - Department of Mathematics
Edgardo Ramirez-Gomez - Department of Mathematics
Diana Rubio - Department of Mathematics
Michael Sarver - Department of Mathematics
Lena Sadtchikova - Aerospace and Ocean Engineering Department
Ajit Shenoy - Aerospace and Ocean Engineering Department
Lisa Stanley - Department of Mathematics
Hartono Sumali - Mechanical Engineering Department
Rosalyn Swiggett - Department of Mathematics
Lan Zhang - Department of Mathematics

5.5 Undergraduate Students

Thomas Bail - Department of Mathematics
Sarah Burrowbridge - Department of Mathematics
Carrie Queen - Department of Mathematics
Lyle Smith - Department of Mathematics

5.6 Support Personnel

Melissa Chase - Program Support Technician
Kristine Gross - Office Services Assistant

Denise Wirt - Program Support Technician

5.7 Degrees Awarded

Justin Appel, Ph.D., Department of Mathematics
Sensitivity Calculations for Conservation Laws with Application to Discontinuous Fluid Flows, 1997.

Thomas Bail, M.S., Department of Mathematics
A Disturbance-Rejection Problem Involving a 2-D Airfoil, 1997.

Jeffrey Borggaard, Ph.D., Department of Mathematics
The Sensitivity Equation Method for Optimal Design, 1994.

Pavel Bochev, Ph.D. Department of Mathematics
Least Squares Finite Element Methods for the Stokes and Navier-Stokes Equations

Mark Borisuk, Ph.D., Department of Mathematics
Bifurcation Analysis of a Model of Frog Egg Cell Cycle, 1997.

John Burkardt, Ph.D., Department of Mathematics
Sensitivity Analyses and Computational Shape Optimization for Incompressible Flows, 1995.

Yanzhao Cao, Ph.D., Department of Mathematics
Analysis and Numerical Approximations of Exact Controllability Problems for Systems Governed by Partial Differential Equations

Graciela Cerezo, Ph.D., Department of Mathematics
Solution Representation and Identification for Singular Neutral Functional Differential Equations, 1996.

Shihchung Chiang, Ph.D., Department of Mathematics
Numerical Solutions for a Class of Singular Integro-Differential Equations, 1996.

Jennifer Deang, Ph.D., Department of Mathematics
A Study of Inhomogeneities and Anisotropies in Superconductors, 1997.

Sophie Dufresne, M.S., Department of Aerospace and Ocean Engineering
Optimization of an Airfoil's Performance through Moving Boundary Control, 1993.

Michael Feldman, M.S., Aerospace and Ocean Engineering Department
Efficient Low-Speed Flight in a Wind Field

Wei Huang, Ph.D., Department of Mathematics
Compensator Design for a System of Two Connected Beams, 1994.

Rossitza Karamikhowa, Ph.D., Department of Mathematics
A Finite Element Analysis of a High-Kappa, High-Field, Ginzburg-Landau Type Model of Superconductors

Hongchul Kim, Ph.D., Department of Mathematics
Analysis and Finite Element Approximation of an Optimal Shape Control Problem for the Steady State Navier-Stokes Equations, 1993.

Hyung-Chun Lee, Ph.D., Department of Mathematics
Analysis, Finite Element Approximation, and Computation of Optimal and Feedback Flow Control Problems, 1994.

Hamadi Marrekchi, Ph.D. Department of Mathematics
Dynamics Compensators for a Nonlinear Conservation Law, 1993.

Steven Pugh, M.S., Department of Mathematics
Finite Element Approximations of Burgers' Equation, 1995

A. Diana Rubio, Ph.D., Department of Mathematics
Distributed Parameter Control of Thermal Fluids, 1997.

Lena Sadtchikova, M.S., Aerospace and Ocean Engineering Department
Optimal Shape Design with Domain Decomposition

Michael Sarver, M.S., Department of Mathematics

William Waldron, Ph.D., Aerospace and Ocean Engineering Department
Optimal Vertical-Plane Booster Guidance Including Pitch Dynamics

Lan Zhang, Ph.D., Department of Mathematics
Parameter Identification in Linear and Nonlinear Parabolic Partial Differential Equations, 1995.

Chapter 6

Publications

1. Appel, J., Sensitivity Calculations for Conservation Laws with Application to Discontinuous Fluid Flows, Ph.D. Thesis, Department of Mathematics, Virginia Polytechnic Insititute and State University, Blacksburg, VA, 1997.

2. Appel, J., Godfrey, A., Cliff, E., and Gunzburger, M., Optimization-Based Design in High Speed Flows, *CFD For Design and Optimization: Proceedings 1995 ASME International Mechanical Engineering Congress And Exposition*, Ed. O.Baysal, FED-Vol. 232, pp. 61-68 (1995).

3. Appel, J. and Gunzburger, M., Sensitivity Calculation in Flows with Discontinuities, *Proceedings of the 14th AIAA Applied Aerodynamics Conference* New Orleans, LA, June 1996.

4. Appel, J. and Gunzburger, M., Difficulties in Sensitivity Calculations for Flows with Discontinuities, *AIAA J.* 35, 1997, pp. 842-848.

5. Bail, T., A Disturbance-Rejection Problem Involving a 2-D Airfoil, M.S. Thesis, Department of Mathematics, Virginia Polytechnic Insititute and State University, Blacksburg, VA, 1997.

6. Balogh, A., Burns, J., Gilliam, D. and Shubov, V., A Note on Numerical Stationary Solutions for the Viscous Burgers' Equation, *Journal of Mathematical Systems, Estimation and Control*, to appear.

7. Banks, H.T., Gilliam, D.S., and Shubov, V.I., Global Solvability for Damped Abstract Nonlinear Hyperbolic Systems, CRSC-TR95-25, August, 1995, also in *Differential and Integral Equations*, 10 (1997), pp. 309-332.

8. Banks, H.T., Inman, D.J., Leo, D.J. and Wang, Y., An Experimentally Validated Damage Detection Theory in Smart Structures, *Journal of Sound and Vibration*, to appear.

9. Banks, H.T. and Ito, K., Approximation in LQR Problems for Infinite Dimensional Systems with Unbounded Input Operators, CRSC-TR94-22, November, 1994 *J. Math. Systems, Estimation and Control*, 7 (1997), pp. 119-122.

10. Banks, H.T. and Kurdila, A.J., Hysteretic Control Influence Operators Representing Smart Material Actuators: Identification and Approximation, CRSC-TR96-23, August, 1996 *Proc. 35th IEEE Conf. on Decision and Control*, December,1996, pp. 3711-3716.

11. Banks, H.T., Kurdila, A.J., and Webb, G., Identification of Hysteretic Control Influence Operators Representing Smart Actuators, Part I: Formulation, CRSC-TR96-14, April, 1996, *Math Problems in Engineering*, to appear.

12. Banks, H.T. and Wang, Y., Damage Detection and Characterization in Smart Material Structures, CRSC-TR93-17, November 1993; also in *Control and Estimation of Distributed Parameter Systems; Nonlinear Phenomena*, Birkhäuser ISNM, to appear.

13. Banks, H.T., Wang, Y., Inman, D.J. and Slater, J.D., Approximation and Parameter Identification for Damped Second Order Systems with Unbounded Input Operators, CRSC-TR93-9, May 1993; *Control: Theory and Adv. Tech.*, to appear.

14. Banks, H.T. and Zhang, Y., Computational Methods for a Curved Beam with Piezoceramic Patches, CRSC-TR96-22, July, 1996 *J. Intelligent Material systems and Structures*, to appear.

15. Battermann, A., and Heinkenschloss, M., Preconditioners for Karush-Kuhn-Tucker Matrices Arising in the Optimal Control of Distributed Systems, Technical Report, 18 pages, October 1996, also to appear W. Desch, F. Kappel, K. Kunisch, eds., *Optimal Control of Partial Differential Equations, Vorau 1996*, Birkhäuser Verlag, Basel, Boston, Berlin.

16. Berdaguer, E., Cerezo, G., Herdman, T. and Turi, J., Parameter Identification Techniques for Singular Neutral Equations, *Applied Mechanics in the Americas, Vol. II, Dynamics and Vibrations/Optimization and Control*, 1995, pp. 322-327.

17. Bikdash, M., Cliff, E.M. and Speyer, J.L., An Algorithm for Space-Based Tracking with Bearings-Only Measurements, *Proceedings of the AIAA Guidance, Navigation and Control Conference*, Baltimore, MD, August 1995, pp. 656-669.

18. Bochev, P., Least Squares Finite Element Methods for the Stokes and Navier-Stokes Equations, Ph.D. Thesis, Department of Mathematics, Virginia Polytechnic Insititute and State University, Blacksburg, VA, 1997.

19. Bochev, P. and Gunzburger, M., Least-squares Methods for the Velocity-Pressure-Stress Formulation of the Stokes Equations; *Comput. Methods Appl. Mech. Engng.* 126 1995, pp. 267-287.

20. Bocvarov, S., Cliff, E.M. and Lutze, F.H., Hierarchal Modeling Approach in Aircraft Trajectory Optimization, *Proc. 13th IFAC Symposium on Automatic Control in Aerospace*, September, 1994, pp. 15-20.

21. Bocvarov, S., Cliff, E.M. and Lutze, F.H., Aircraft Time-Optimal Heading Reversal Maneuvers, *Proc. AIAA Guidance, Navigation and Control Conf.*, Scottsdale, AZ, August 1994, pp. 146-153.

22. Bocvarov, S., Cliff, E.M. and Lutze, F.H., Significance of the Dihedral Effect in Rapid Fuselage - Reorientation Maneuvers, *J. Aircraft*, Vol. 31, No. 3, May - June 1994, pp. 548-555.

23. Bocvarov, S., Cliff, E.M. and Lutze, F.H., A Hierarchal-Modeling Approach in Trajectory Optimization, in Proceedings of the IFAC Symposium on Automatic Control in Aerospace, Palo Alto, CA, September, 1994, pp. 15-20.

24. Bocvarov, S., Cliff, E.M. and Lutze, F.H., Aircraft Time-Optimal Heading-Reversal Maneuvers, in *Proceedings of the AIAA Guidance, Navigation and Control Conference*, Scottsdale, AZ, August 1994, pp. 146-153.

25. Borggaard, J.T., On the Presence of Shocks on Domain Optimization of Euler Flows, in *Flow Control*, M. Gunzburger, Ed., Springer-Verlag, 1995, pp. 35-48.

26. Borggaard, J., The Sensitivity Equation Method for Optimal Design, Ph.D. Thesis, Department of Mathematics, Virginia Polytechnic Institute and State University, 1994.

27. Borggaard, J., Burkardt, J., Burns, J., Cliff, E., Gunzburger, M., Kim, H., Lee, H., Peterson, J., Shenoy, A. and Wu, X., Algorithms for Flow Control and Optimization, *Optimal Design and Control*, J. Borggaard, J. Burkardt, M. Gunzburger and J. Peterson Eds., Birkhauser, 1995, pp. 97-116.

28. Borggaard, J. and Burns, J., A Sensitivity Equation Approach to Shape Optimization in Fluid Flows, in *Flow Control*, M. Gunzburger, Ed., Springer-Verlag, 1995, pp. 49-78.

29. Borggaard, J. and Burns, J., A Sensitivity Equation Approach for the Optimal Design of Nozzles, in *Proceedings of the AIAA 5th Symposium on Multidisciplinary Analysis and Optimization*, September, 1994, pp. 232-241.

30. Borggaard, J. and Burns, J., Asymptotically Consistent Gradients in Optimal Design, *Proceedings of the Workshop on Multidisciplinary Design Optimization*, N. Alexandrov Ed., to appear.

31. Borggaard, J. and Burns, J., A PDE Sensitivity Equation Method for Optimal Aerodynamic Design, *Journal of Computational Physics*, to appear.

32. Borggaard, J., Burns, J., Cliff, E. and Gunzburger, M., Sensitivity Calculations for a 2D, Inviscid, Supersonic Forebody Problem, in *Identification and Control of Distributed Parameter Systems*, H.T. Banks, R. Fabiano and K. Ito, Eds., SIAM Publications, Philadelphia, PA, 1993, 14-24.

33. Borggaard, J., Herdman, T. and Turi, J., On Control Design for a Fluid-Structure Interaction Problem, in *Proceedings of the 1993 IEEE Conference on Aerospace Control Systems*, May, 1993, pp. 236-242.

34. Borggaard, J., Herdman, T. and Turi, J., On an Application of the Boundary Element Method to Study Flow Induced Vibrations, *Applied Mechanics in the Americas, Vol. II, Dynamics and Vibration/ Optimization and Control*, 1995, pp. 317-321.

35. Borggaard, J., Herdman, T. and Turi, J., On Active Control of Flow Induced Vibrations, *Proceedings of the 34th IEEE Conference on Decision and Control*, December 1995, pp. 3725–3729.

36. Borggaard, J. and Pelletier, D., On Optimal Design Using an Adaptive Finite Element Method, *Proceedings of the First International Conference on Nonlinear Problems in Aviation and Aerospace*, to appear.

37. Borggaard, J. and Pelletier, D., Computing Design Sensitivities Using an Adaptive Finite Element Method, *Proceedings of the 27th AIAA Computational Fluid Dynamics Conference*, AIAA Paper 96-1938, June 1996.

38. Borisuk, M., Bifurcation Analysis of a Model of Frog Egg Cell Cycle, Ph.D. Thesis, Department of Mathematics, Virginia Polytechnic Insititute and State University, Blacksburg, VA, 1997.

39. Burazanis, M., Cliff, E.M. and Well, K.-H., A Frenet-based Agility Model, *Proceedings of the AIAA Atmospheric Flight Mechanics Conference*, Baltimore, MD, August 1995, pp. 479-485.

40. Burkardt, J., Sensitivity Analyses and Computational Shape Optimization for Incompressible Flows, Ph.D. Thesis, Department of Mathematics, Virginia Polytechnic Institute and State University, 1995.

41. Burkardt, J., and Gunzburger, M., Sensitivity Discrepancy for Geometric Parameters; *CFD For Design and Optimization: Proceedings 1995 ASME International Mechanical Engineering Congress And Exposition*, Ed. O.Baysal, FED-Vol. 232, pp. 9-15 (1995).

42. Burkardt, J., Gunzburger, M. and Peterson, J., Discretization of Cost and Sensitivities in Shape Optimization, *Computation and Control IV, Proceedings of the Fourth Bozeman Conference on Computation and Control*, Bowers and Lund, eds., Birkhauser, 1995, pp. 43-56.

43. Burns, J. and Ito, K., On Well-Posedness of Integro-Differential Equations in Weighted L^2 Spaces, *Journal of Differential and Integral Equations*, Vol. 8, No. 3, 1995, pp. 627–646.

44. Burns, J., Kang, S., Kachroo, P., and Ozbay, K., System Dynamics and Feedback Control Formulations for Real Time Dynamic Traffic Routing with an Application Example, *Journal of Mathematical and Computer Modeling*, to appear.

45. Burns, J. and King, B.B., A Note on the Regularity of Solutions of Infinite Dimensional Riccati Equations, *Applied Math Letters*, Vol. 7, No. 6, 1994, pp. 13-17.

46. Burns, J.A. and King, B.B., Optimal Sensor Location for Robust Control of Distributed Parameter Systems, *33rd IEEE Conference on Decision and Control*, December 1994, pp. 3967-3972.

47. Burns, J. and King, B., A Note on the Mathematical Modeling of Damped Second Order Systems, *Journal of Mathematical Systems, Estimation and Control*, to appear.

48. Burns, J. and King, B. B., A Reduced Basis Approach to the Design of Low Order Feedback Controllers for Nonlinear Continuous Systems, *Journal of Vibration and Control*, to appear.

49. Burns, J. and King, B. B., A Comparison of Minmax and LQG Control for a Hybrid Nonlinear Continuous Systems, *Applied Mechanics in the Americas*, Vol. II, L. A. Godoy, S. R. Idelshon, P. A. Laura and D. Mook, Eds., AMM and AMCA, Santa Fe, Argentina, 1995, pp. 503–506.

50. Burns, J. and King, B. B., Representation of Feedback Operators for Hyperbolic Systems , *Computation and Control IV*, K.L. Bowers and J. Lund, Eds., Birkhauser, 1995, pp. 57–73.

51. Burns, J.A., King, B.B. and Ou, Y.R., A Computational Approach to Sensor/Actuator Location for Feedback Control of Fluid Flow Systems, *SPIE Conference on Sensing, Actuation and Control in Aeropropulsion*, Orlando, FL, April 1995, pp. 60-70.

52. Burns, J., King, B, and Rubio, D., Regularity of Feedback Operators for Boundary Control of Thermal Processes, *First International Conference on Nonlinear Problems in Aeronautics and Aerospace*, S. Sivasundaram, ed., Embry-Riddle Aeronautical Press, May, 1996, pp. 67–73.

53. Burns, J.A. and Ou, Y.R., Effect of Rotation Rate on the Forces of a Rotating Cylinder: Simulation and Control, ICASE Report No. 93-11, 1993.

54. Burns, J.A. and Ou, Y.R., Active Control of Vortex Shedding, in *Proceedings of the 32nd AIAA Aerospace Sciences Meeting*, Reno, Nevada, Paper AIAA 94-0182, January, 1994, 1-10.

55. Burns, J.A. and Ou, Y.R., Feedback Control of the Driven Cavity Problem Using LQG Designs, *33rd IEEE Conference on Decision and Control*, December 1994, pp. 289-294.

56. Burns, J. and Rubio, D., Control of the Boussinesq Equations in a Thermal Loop, *Applied Mechanics in the Americas*, Vol. VI, M. Rysz, L. Godoy and L. Suarez, eds. University of Iowa Press, January, 1997, pp. 142–145.

57. Burns, J. A. and Rubio, D., Control of the Boussinesq Equations in a Thermal Loop, *Fifth Pan American Conference on Applied Mechanics*, to appear.

58. Burns, J.A. and Spies, R., A Numerical Study of Parameter Sensitivities in Landau-Ginzburg Models of Phase Transitions in Shape Memory Alloys, *Journal of Intelligent Material Systems and Structures*, Vol. 5, 1994, pp. 321-332.

59. Cao, Y., Analysis and Numerical Approximations of Exact Controllability Problems for Systems Governed by Partial Differential Equations, Ph.D. Thesis, Department of Mathematics, Virginia Polytechnic Institute and State University, 1996.

60. Cashin, T.P., A Study of Aircraft Agility, Master's Thesis, Department of Aerospace and Ocean Engineering, Virginia Polytechnic Institute and State University, 1994.

61. Cerezo, G., Solution Representation and Identification for Singular Neutral Functional Differential Equations, Ph.D. Thesis, Department of Mathematics, Virginia Polytechnic Insititute and State University, Blacksburg, VA, 1996.

62. Chiang, S., Numerical Solutions for a Class of Singular Integro-Differential Equations, Ph.D. Thesis, Department of Mathematics, Virginia Polytechnic Institute and State University, 1996.

63. Cliff, E. M., Heinkenschloss, M., and Shenoy, A., An Optimal Control Problem for Flows with Discontinuities, Technical Report, 27 pages, September 1995, also in *Journal of Optimization Theory and Applications* August 1997, to appear.

64. Cliff, E. M., Heinkenschloss, M. and Shenoy, A., On the Optimality System for a 1D Euler Flow Problem, *Proceedings of the AIAA 6th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, AIAA Paper 96–3993, September 1996.

65. Deang, J., A Study of Inhomogeneities and Anisotropies in Superconductors, Ph.D. Thesis, Department of Mathematics, Virginia Polytechnic Insititute and State University, Blacksburg, VA, 1997.

66. Dennis, J.E. and Heinkenschloss, M., and Vicente, L.N., Trust–Region Interior–Point SQP Algorithms for a Class of Nonlinear Programming Problems, Interdisciplinary Center for Applied Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, ICAM Report No. 94–12–01, 1994, Appeared also as TR94–45, Department of Computational and Applied Mathematics, Rice University, and also submitted for publication to *SIAM J. on Control and Optimization.*

67. Diehl, G. and Cudney, H.H., The Effects of Shaped Piezoceramic Actuators on the Excitations of Beams, in *Proceedings of the AIAA/ASME Adaptive Structures Forum and 35th AIAA / ASME / ASCE / AHS/ ASC Structures, Structural Dynamics, and Materials Conference,* pp. 270-278, Hilton Head, SC, April, 1994.

68. Du, Q., Gunzburger, M. and Hou, L., Analysis and Finite Element Approximation of Optimal Control Problems for a Ladyzhenskaya Model for Stationary, Incompressible, Viscous Flows, *J. Comp. Appl. Math.* 61 1995, pp. 323-343.

69. Dufresne, S., M.S. Thesis, Aerospace and Ocean Engineering Department, Virginia Polytechnic Insititute and State University, Blacksburg, VA, 1993.

70. Dufresne, S., Cliff, E.M.. and Pelletier, D., Optimization of an Airfoils Performance Through Moving Boundary Control, AIAA 96-2399, *14th AIAA Applied Aerodynamics Conference*, New Orleans, LA, June 1996.

71. Fan, Y., Cliff, E.M., Lutze, F.H. and Anderson, M.R., Mixed H_2/H_∞ Optimal Control for an Elastic Aircraft, *Journal of Guidance, Control and Dynamics*, Vol. 19, No. 3, May-June 1996, pp. 650-655, also *Proceedings of the AIAA Guidance, Navigation and Control Conference*, Baltimore, MD, August 1995, pp. 580-589.

72. Feldman, M., Efficient Low-Speed Flight in a Wind Field, M.S. Thesis, Aerospace and Ocean Engineering Department, Virginia Polytechnic Institute and State University, 1996.

73. Feldman, M. and Cliff, E. M. , Energy-Modeled Flight in a Wind-Field, *Proc. of the First International Conf. on Nonlinear Problems in Aviation and Aerospace*, Embry-Riddle University, Daytona Beach FL, May 1996, pp. 187–193.

74. Fursikov, A., Gunzburger, M. and Hou, L., Boundary Value Problems and Optimal Boundary Control for the Navier-Stokes System: the Two-dimensional Case, *SIAM J. Cont. Optim.*, to appear.

75. Gunzburger, M., Flow Control and Optimization; in *Computational Fluid Dynamics Review*, Wiley, West Sussex, 1995, pp. 548-566.

76. Gunzburger, M., Navier-Stokes Equations for Incompressible Laminar Flow: Finite Element Methods, to appear in *Handbook on Computational Fluid Mechanics*, Academic, Boston, 1996, pp. 99-157.

77. Gunzburger, M. and Bochev, P., Analysis of Wrighted Least-Squares Finite Element Methods for the Navier-Stokes Equations, to appear in *Proc. 14th IMACS World Congress on Computational and Applied Mathematics*, Georgia Tech, Atlanta, 1994, pp. 584-587.

78. Gunzburger, M., Erlebacher, G., Hussaini, M., Joslin, R. and Nicolaides, R., Active Control of Instabilities in Laminar Boundary-Layer Flow: Use of Sensors and Spectral Controller, *AIAA J.* 33 1995, pp. 1521-1523.

79. Gunzburger, M. and Hou, L., Finite Dimensional Approximation of a Class of Constrained Nonlinear Optimal Control Problems, *SIAM J. Cont. Opt.* 34, 1996, pp. 1001-1043.

80. Gunzburger, M., Hou, L. and Ravindran, S., Analysis and Approximation of Optimal Control Problems for a Simplified Ginzburg-Landau Model of Superconductivity; to appear.

81. Gunzburger, M. and Kim H., Existence of An Optimal Solution of a Shape Control Problem for the Stationary Navier-Stokes Equations, to appear.

82. Gunzburger, M. and Kim, H., Sensitivity Analysis of a Shape Control Problem for the Stationary Navier-Stokes equations, to appear.

83. Gunzburger, M. and Lee, H., Analysis, Approximation, and Computation of a Coupled Solid/Fluid Temperature Control Problem, to appear in *Comput. Methods Appl. Mech. Eng.*

84. Gunzburger, M. and Lee, H., Feedback Control of Karman Vortex Shedding, *J. Appl. Mech* 63, 1996, pp. 828-835.

85. Gunzburger, M. and Lee, H.C., Analysis of Some Boundary Value Problems Associated with Feedback Control Problems, to appear.

86. Gunzburger, M. and Lee, H.C., Feedback Control of Fluid Flows, *Proc. 14th IMACS World Congress on Computational and Applied Mathematics*, Georgia Tech, Atlanta, 1994, pp. 716-719.

87. Hartung, F., Herdman, T. and Turi, J., Identification of Parameters in Hereditary Systems: A Numerical Study, *Proceedings of the 3rd IEEE Mediterranean Symposium on New Directions in Control and Automation*, Vol. 1, 1995, pp. 291–298.

88. Hartung, F., Herdman, T. and Turi, J., Parameter Identification in Classes of Hereditary Systems of Neutral Type, *Journal of Applied Mathematics and Computation*, to appear.

89. Hartung, F., Herdman, T. and Turi, J., Identification of Parameters in Hereditary Systems, *Proceedings of the ASME Fifteenth Biennial Conference on Mechanical Vibration and Noise*, 1995, pp. 1061–1066.

90. Heinkenschloss, M. The Numerical Solution of a Control Problem Governed by a Phase Field Model. *Optimization Methods and Software*, Vol. 7, 1997, pp. 211–263.

91. Heinkenschloss, M., Projected Sequential Quadratic Programming Methods, *SIAM J. on Optimization*, Vol. 6, 1996, pp. 373–417.

92. Heinkenschloss, M., A Trust Region Method for Norm Constrained Problems, Technical Report, 27 pages, August 1994, *SIAM J. on Numer. Analysis* to appear.

93. Heinkenschloss, M., Formulation and Analysis of a Sequential Quadratic Programming Method for the Optimal Dirichlet Boundary Control of Navier–Stokes Flow, Technical Report, 18 pages, May 1997. Submitted for publication to *Optimal Control: Theory, Algorithms, and Applications*, W. W. Hager and P. M. Pardalos, Editors.

94. Heinkenschloss, M. and Tröltzsch, F., Analysis of the Lagrange–SQP–Newton Method for the Control of a Phase Field Equation, Interdisciplinary Center for Applied Mathematics, Virginia Polytechnic Institute and State University, 24061, ICAM Report No., 95–03–01, 1995, also submitted for publication to *Numerical Functional Analysis and Optimization.*

95. Heinkenschloss, M. and Vicente, L. N., Analysis of Inexact Trust–Region Interior–Point SQP Algorithms, Technical Report, 34 pages, May 1995. Submitted for publication to *SIAM J. on Control and Optimization.*

96. Herdman, T. L. and Kang, K., A Structured Sequential Quadratic Programming and its Application to a Flow Matching Problem, *Proceedings of the First International Conference on Nonlinear Problems in Aviation and Aerospace*, pp. 255-261.

97. Herdman, T.L., Cerezo, G., Berdaguer, E. and Turi, J., Parameter Identification Techniques for Singular Neutral Equations, *Applied Mechanics in the Americas, Vol. II, Dynamics and Vibration/Optimization and Control*, 1995, pp. 322-327.

98. Herdman, T.L., Cerezo, G., Berdaguer, E. and Turi, J., Collocation Techniques for the Approximation of Singular Neutral Functions, *Proc. 33rd IEEE Conference on Decision and Control*, 1994, pp. 2534-2536.

99. Herdman, T. L. and Kang, K., An Optimization Based Approach to Flow Matching for Burger's Equation with a Forcing Term, *Proceedings of the ASME Fifteenth Biennial Conference on Mechanical Vibration and Noise*, 1995, pp. 1083–1086.

100. Herdman, T.L. and Turi, J., A "Natural" State-Space for an Aeroelastic Control System, *Journal of Integral Equations and Applications*, Vol. 7, No. 4, Fall 1995, pp. 413–424.

101. Hou, L. and Peterson, J., Matching Electric Current in Electrically Conducting Fluid Flows by Using an Optimal Control Formulation, accepted.

102. Huang, W., Compensator Design for a System of Two Connected Beams, Ph.D. Thesis, Department of Mathematics, Virginia Polytechnic Insititute and State University, Blacksburg, VA, 1994.

103. Inman, D.J. and Kress, A., Eigenstructure Assignment Using Inverse Eigenvalue Methods, *AIAA Journal of Guidance, Control and Dynamics*, Vol. 18, No. 3, 1995, pp. 625-627.

104. Inman, D.J., Structural Control by Eigenstructure Assignment, in *Proceedings of the 10th Symposium on Structural Dynamics*, VPI&SU, May, 1995.

105. Karamikhowa, R., A Finite Element Analysis of a High-Kappa, High-Field, Ginzburg-Landau Type Model of Superconductors, Ph.D. Thesis, Department of Mathematics, Virginia Polytechnic Institute and State University, 1995.

106. Kim, H., Analysis and Finite Element Approximation of an Optimal Shape Control Problem for the Steady State Navier-Stokes Equations, Ph.D. Thesis, Department of Mathematics, Virginia Polytechnic Insititute and State University, Blacksburg, VA, 1993.

107. King, B. B., Regularity of Feedback Operators for Parabolic Control Problems, *Proceedings of the AMS*, to appear.

108. King, B.B. and Y.R. Ou, Nonlinear Dynamic Compensator Design for Flow Control in a Driven Cavity, *Proceedings of the 34th IEEE Conference on Decision and Control*, December 1995, pp. 3741–3746.

109. Kumar, R., Seywald, H. and Cliff, E.M., Near-Optimal, Three-Dimensional Air-to-Air Missile Guidance, *Journal of Guidance, Control and Dynamics*, Vol. 18, No. 3, May-June 1995, pp. 449-456.

110. Kumar, R., Seywald, H., Cliff, E.M., and Kelley, H.J., Three-Dimensional Air-to-Air Missile Trajectory Shaping, *Journal of Guidance, Control and Dynamics*, Vol. 18, No. 3, May-June 1995, pp. 457-464.

111. Lallement, G. and Inman, D.J., A Tutorial on Complex Eigenvalues, *Proceedings of the 13th IMAC*, February, 1995, pp. 490-495.

112. Layton, W., Lenferink, K. and Peterson, J., A two-level Newton finite element method for approximating electrically conducting incompressible fluid flows, to appear.

113. Lee, H-C., Analysis, Finite Element Approximation, and Computation of Optimal and Feedback Flow Control Problems, Ph.D. Thesis, Department of Mathematics, Virginia Polytechnic Insititute and State University, Blacksburg, VA, 1994.

114. Leo, D.J. and Inman, D.J., Linear Controller Design for Structures with Uncertain Transient Disturbances, in *Proceedings AIAA SDM Conference*, April 1994.

115. Leo, D.J. and Inman, D.J., Optimal Collocated Control of a Smart Antenna, *Proceedings 33rd IEEE Conference on Decision and Control*, December, 1994, pp. 109-114.

116. Lomenzo, R.A., Sumali, H. and Cudney, H.H., Maximizing Mechanical Power Transfer From Piezoelectric Stacked Actuators to Structures, in *Adaptive Structures and Material Systems*, edited by G. P. Carman and E. Garcia, Publication AD-Vol. 35, pp. 229236, ASME, NY, NY, 1993.

117. Marrekchi, H., Dynamic Compensators for a Nonlinear Conservation Law, Ph.D. Thesis, Department of Mathematics, Virginia Polytechnic Institute and State University, 1993.

118. Ou, Y.R. and Sritharan, S.S., On the Robustness of the Navier-Stokes Global Attractor, in *14th IMACS World Congress*, pp. 868-871, Atlanta, GA, July 1994.

119. Ou, Y.R. and Sritharan, S.S., Upper Semicontinuous Global Attractors for Viscous Flow, *Dynamic Systems and Applications*, 1995, accepted.

120. Pugh, S., Finite Element Approximations of Burgers' Equation, M.S. Thesis, Department of Mathematics, Virginia Polytechnic Institute and State University, 1995.

121. Rubio, A. D., Boundary Control for the Chaotic Flow of a Thermal Convection Loop, *Proceedings of the 34th IEEE Conference on Decision and Control*, New Orleans, December 1996, pp. 3734–3737.

122. Rubio, A. D., Distributed Parameter Control of Thermal Fluids, Ph.D. Thesis, Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA, 1997.

123. Sadtchikova, E., Optimal Shape Design with Domain Decomposition, M.S. Thesis, Aerospace and Ocean Engineering Department, Virginia Polytechnic Institute and State University, 1996.

124. Sadtchikova, E., Shenoy, A. and Cliff, E., Computational Issues in Optimization-Based Design, *Proceedings of the 34th IEEE Conference on Decision and Control*, New Orleans, December 1996, pp. 445–450.

125. Shenoy, A., Optimization Techniques Exploiting Problem Structure: Applications to Aerodynamic Design, Ph.D. Thesis, Aerospace and Ocean Engineering Department Virginia Polytechnic Institute and State University, Blacksburg, VA, 1997.

126. Shenoy, A.R. and Cliff, E.M., An Optimal Control Formulation for a Flow Matching Problem, ICAM Report 93-07-01, Virginia Tech, July, 1993, also *Proceedings of the AIAA 5th Symposium on Multidisciplinary Analysis and Optimization*, Panama City, FL, AIAA-94-4306, September 1994

127. Shenoy, A., Cliff, E. M. and Heinkenschloss, M., Thermal-Fluid Control via Finite-Dimensional Approximation, *Proceedings of the 31st AIAA Thermophysics Conference*, AIAA Paper 96–1910, June 1996.

128. Shenoy, A., Cliff, E. M., and Heinkenschloss, M., Thermal-Fluid Control via Finite-Dimensional Approximation, AIAA Paper 96–1910, *31st AIAA Thermophysics Conference*, June 17–20, 1996, New Orleans, LA.

129. Sumali, H. and Cudney, H.H., Segmented Two-Dimensional Modal-Filtering Sensors, in *Vibration and Control of Mechanical Systems*, edited by C. A. Tan and L. A. Bergman, Publication DE-Vol. 61, pp. 5966, published by ASME, New York, N.Y., 1993.

130. Sumali, H. and Cudney, H.H., Maximizing Mechanical Power Transfer From Piezoelectric Stacked Actuators to Structures, in *Adaptive Structures and Material Systems*, edited by E. Garcia and H. H. Cudney, ASME, NY, NY, 1994.

131. Tadi, M. and Burns, J., Feedback Controller for a Flexible Structure Using Piezoceramic Actuators, *Journal of Dynamics and Control*, Vol. 5, No. 4, 1995, pp. 401-419.

132. Ulbrich, M., Ulbrich, S., and Heinkenschloss, M., Global convergence of affine-scaling interior-point Newton methods for infinite-dimensional nonlinear problems with pointwise bounds, Technical Report TR97–04, Department of Computational and Applied Mathematics, Rice University, Houston, Texas, 30 pages, 1997. Submitted for publication to *SIAM J. on Control and Optimization.*

133. Waldron, W., Optimal Vertical-Plane Booster Guidance Including Pitch Dynamics, Ph.D. Thesis, Aerospace and Ocean Engineering Department, Virginia Polytechnic Institute and State University, 1995.

134. Wu, X., Cliff, E. M. and Gunzburger, M. D., An Optimal Design Problem for Two-Dimensional Flow in a Duct, *Optimal Control, Applications and Methods*, Vol 17, 1996, pp. 329-339.

135. Zhang, L., Parameter Identification in Linear and Nonlinear Parabolic Partial Differential Equations, Ph.D. Thesis, Department of Mathematics, Virginia Polytechnic Insititute and State University, Blacksburg, VA, 1995.

Chapter 7

Interactions and Transitions

One of the major components of **CODAC** is the active corporation with Air Force laboratories, facilities and industry. During the past ten years most of the Center's core participants have developed strong scientific associations with engineers and scientists at various Air Force facilities and industries. These interactions have been mutually beneficial and we feel that such collaborations between Air Force scientists, industry and academic researchers is the most direct mechanism for transitioning basic research. We briefly summarize some of the current joint efforts in laboratory/industry/academic interactions.

7.1 Participation and Presentations at Meetings

John Burns

1. *Computational Issues in Optimal Design*, Center for Research on Computation and its Applications, Montreal, Quebec, September, 1994.
2. *A Practical Approach to Feedback Control of Distributed Parameter Systems*, Oregon State University, Corvallis, OR, October, 1994.
3. *A Report on New Results in Control and Design*, Air Force Conference on Dynamics and Control, Dayton, OH, June, 1994.
4. *Control of Fluid Flows*, National SIAM Meeting, San Diego, CA, July, 1994.
5. *Computational Methods for Optimal Sensor Location*, Fourth International Conference on Computation and Control, Bozeman, MT, August, 1994.
6. *Sensor Location Problems for Hyperbolic Systems*, IEEE Conference on Decision and Control, Orlando, FL, December, 1994.
7. *Physics Based Models as a Framework for Optimal Design and Control*, IMA Workshop on Control of Materials Processing, Minneapolis, MN, January, 1995.
8. *A Control Theory Approach to Shape Optimization*, ICASE/LaRC Multidisciplinary Design Optimization Workshop, Hampton, VA, March, 1995.
9. *A New Approach to Control Design for Fluid Flows*, SPIES Conference on Sensing and Control of Aerosystems, Orlando, FL, April, 1995.
10. *Low Order Observers for Nonlinear PDE Systems*, SIAM Conference on Control, St. Louis, MO, April, 1995.

11. *Reduced Basis Approach for Optimal Feedback Control*, Conference on Dynamics and Control, Minneapolis, MN, June, 1995.

12. *Sensitivity Equations for Optimal Design and Control*, IMA Workshop in Optimal Design, Minneapolis, MN, July, 1995.

13. *Optimization Problems in the Design of Local Feedback Controllers*, SIAM National Meeting, Charlotte, NC, October, 1995.

14. Texas Tech University, Lubbock, TX, November, 1995.

15. *The Sensitivity Equation Method: A Short Course*, Southeastern SIAM Conference, Clemson, SC, March, 1996.

16. *Applied Mathematics in the Defense of the Nation*, Basic Research in the National Defense, Washington, DC, May, 1996.

17. *Optimal Control Approaches to MDO*, SIAM Conference on Optimization, Victoria, BC, May, 1996.

18. *Computational Approaches to the Problem of Controller Reduction for DPS*, MNTS 1996 Conference on Control, St. Louis, MO, June, 1996.

19. *Numerical Methods for Gradient Computations via Sensitivity Equations*, Second World Congress on Nonlinear Analysis, Athens, Greece, July, 1996.

20. *Accurate Numerical Methods for Sensitivity Equations in Optimal Design*, International Conference on Distributed Parameter Control, Vorau, Austria, July, 1996.

21. *A Note on the Mathematical Modeling of Internal Damping*, Fifth International Conference on Computation and Control, Bozeman, MT, August, 1996.

22. *Models of Elastic Structures*, Universitat Trier, Trier, Germany, September, 1996.

23. *Optimal Design for Flows*, Nestle Research and Development Headquarters, Stuttgart, Germany, September, 1996.

24. *The Sensitivity Equation Method in Industrial Applications*, Deutche Mathematiker-Vereinigung Annual Meeting, Jena, Germany, September, 1996.

25. *A Short Course on Optimal Design*, Lousianna State University, Batton Rouge, LA, October, 1996.

26. *Set Valued Integration*, St. Mary's College, St. Mary's, MD, October, 1996.

27. *Control of Fluid/Structure Interaction*, University of Arkansas, Fayetteville, AK, November, 1996.

28. *Semigroups Generated by Second Order Hyperbolic Systems*, Showme Lectures, Rolla, MO, November, 1996.

29. *Projection Schemes for Optimal Design and Analysis*, IFIP Conference on Optimization, Gainsville, FL, February, 1997.

Gene Cliff

1. *Aircraft Time-Optimal Heading Reversal Maneuvers*, AIAA Guidance, Navigation and Control Conf., Scottsdale, AZ, August, 1994.

2. *Heirarchal Modeling Approach in Aircraft Trajectory Optimization*, 13th IFAC Symposium on Automatic Control in Aerospace, September 12-16, 1994.

3. *Optimization in Flight Performance and Beyond*, Control and System Science Seminar, University of Minnesota, April 28, 1995.

4. *Optimization - Some Aerospace Applications*, Honeywell Research Center, June 6, 1995.

5. *Optimal Aerodynamic Design for Flows with Shocks*, Meeting on Optimal Control and Variational Methods, Mathematisches Forschungsinstitut, Oberwolfach, Germany, 21-27 January, 1996.

6. *Energy-Modeled Flight in a Wind Field*, First International Meeting on Nonlinear Problems in Aviation and Aerospace, Daytona Beach, FL, May 1996.

7. *Optimization of an Airfoil's Performance Through Moving Boundary Control*, 14th AIAA Applied Aerodynamics Conf., New Orleans, LA, June 1996.

8. *Some Aerospace Uses of Optimization*, Mathematics Department Colloquium, Lousianna State University, Baton Rouge, LA, June 1996.

9. *On An Optimality System for 1-D Euler Flows*, Second World Congress on Nonlinear Analysis, Athens, Greece, July 1996.

10. *Energy-Modeled Flight in a Wind Field*, Workshop on Trajectory Optimization, AIAA Atmospheric Flight Mechanics Conference, San Diego, CA, July 1996.

Max Gunzburger

1. *Control of the Time-Dependent Navier-Stokes Equations*, Canadian Applied Mathematics Society Annual Meeting, Montreal June, 1994.

2. *Feedback Control of Karman Vortex Shedding*, IMACS World Congress, Atlanta July, 1994.

3. *Modeling and Analysis of Type-II Superconductivity*, Westinghouse Research Laboratories February, 1995.

4. *Flow Control and Optimization*, Iowa State University March, 1995.

5. *Analysis and Computations for Flow Control and Optimization*, Centre de Recherche en Calcul Applique, Montreal April, 1995.

6. *Modeling and Analysis of Type-II Superconductivity*, USAF Rome Laboratory, Hanscom AFB April, 1995.

7. *Finite Dimensional Approximation of Optimal Control Problems*, Third SIAM Conference on Control and its Applications, St. Louis April, 1995.

8. *Shape Control Problems for the Navier-Stokes Equations*, Third SIAM Conference on Control and its Applications, St. Louis April, 1995.

Matthias Heinkenschloss

1. *Optimization Methods for Constrained Control Problems.* SIAM Conference on Control and its Applications. April, 1995, St. Louis

2. *Optimization Methods for Large Scale Constrained Control Problems.* SCICade 95 – International Conference on Scientific Computing and Differential Equations. March, 1995, Stanford University.

3. *Optimization in Function Spaces.* Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA, September, 1994

4. *SQP Methods for Constrained Control Problems.* 15th International Symposium on Mathematical Programming. August 15–19, 1994, University of Michigan, Ann Arbor, Michigan.

5. *Projected Sequential Quadratic Programming Methods.* Seventh German–French Conference on Optimization, June, 1994, Dijon, France.

6. *Projizierte SQP-Verfahren für Optimierungsprobleme mit Ungleichungen.* Technical University of Dresden, Germany, June 1994

7. *Reduzierte SQP-Verfahren für Steuerungsprobleme mit Ungleichungen.* University of Trier, Germany, June 1994

8. *Trust–Region Innere–Punkt Verfahren zur Lösung von großdimensionalen Steuerungsproblemen*, Colloquium Talk, Technische Universität Chemnitz, Germany, June, 1995.

9. *Die Numerische Lösung Nichtlinearer Kontrollprobleme bei Partiellen Differentialgleichungen*, Colloquium Talk, Universität Trier, Germany, June, 1995.

10. *Die Numerische Lösung Nichtlinearer Kontrollprobleme bei Partiellen Differentialgleichungen*, Colloquium Talk, IWR Interdisziplinäres Zentrum für Wissenschaftliches Rechnen (Interdisciplinary Center for Scientific Computing), Universität Heidelberg, Germany June, 1995.

11. *Preconditioners for KKT Systems Arising in Optimal Control*, International Conference on Control and Estimation of Distributed Parameter Systems, Vorau, Austria, July, 1996.

12. SIAM Conference on Optimization, Victoria, Canada, May 20–22, 1996.

13. *SQP Methods for the Control of Fluids*, SIAM Annual Meeting, Charlotte, NC, October, 1995.

14. *The Design of SQP Algorithms for the Solution of Optimal Control Problems*, Workshop on Iterative Methods for Large Scale Nonlinear Problems, Utah State University, September, 1995.

15. *Trust–Region Interior–Point SQP Methods for Optimal Control and Parameter Identification Problems*, IMA Workshop on Large Scale Optimization IMA, University of Minnesota, July, 1995.

16. *Optimal Design for Flows with Discontinuities*, ICIAM 95, Hamburg, Germany, July, 1995.

17. *Optimization Methods for Constrained Control Problems*, SIAM Conference on Control and its Applications, St. Louis, MO, April, 1995.

18. *Optimization Methods for Large Scale Constrained Control Problems*, SCICade 95 – International Conference on Scientific Computing and Differential Equations, Stanford University, March, 1995.

19. *SQP Interior–Point Methods for Optimal Control Problems*, Numerical Analysis Conference – M. J. D. Powell Fest, University of Cambridge, England, July, 1996.

20. *SQP Methods for the Control of Fluids*, 8th French–German Conference on Optimization, University of Trier, Germany, July, 1996.

Terry Herdman

1. Organized special *Fluid-Structure Interactions*, Pan American Congress on Applied Mechanics, Buenos Aires, Argnetina, January, 1995.

2. *On an Application of the Boundary Element Method to Study Flow Induced Vibrations*, Fluid Structure Interaction Session, Pan American Congress on Applied Mechanics, Buenos Aires, Argentina, January, 1995.

3. *A Comparison of MinMax and LQR Control for a Hybrid Nonlinear Continuous System*, Control Session, Pan American Congress on Applied Mechanics, Buenos Aires, Argnetina, January, 1995.

4. *Parameter Identification in Classes of Hereditary Systems of Neutral Type*, Conference on Differential Equations and Computational Simulations, Mississippi State, April 1995.

5. SIAM Control and Its Applications, St. Louis, MO, April, 1995.

6. *Parameter Identification for Neutral Equations*, Second International Conference on Dynamical Systems and Applications, Atlanta, GA, May, 1995.

7. *Identification of Parameters in Hereditary Systems*, The 3rd IEEE Mediterranean Symposium on New Directions in Control and Automation, Special Session on Infinite Dimensional Systems, Limassol, Cyprus, July, 1995.

8. *Identification of Parameters in Hereditary Systems*, ASME Conference on Mechanical Vibrations and Noise, Special Session on Parameter Identification, September, 1995.

9. *An Optimization Based Approach to Flow Matching for Burger's Equation with Forcing Term*, ASME Conference on Mechanical Vibrations and Noise, Special Session on Parameter Identification, September, 1995.

10. *Applications of Neutral Functional Differential Equations*, International Conference on Mathematics Applied to Industry and Medicine, Buenos Aires, Argentina, November, 1995.

11. *Singular Neutral Integral Equations* Institut fur Praktische Mathematik, University of Karlsruhe, Colloquium Lecture, Karlsruhe, Germany, July, 1995.

12. *A Quasi-linearization Approach to Parameter Identification in a Hybrid System*, The Second World Congress of Nonlinear Analysts, Athens, Greece, July, 1996.

13. *Solution Representations and High Order Schemes for Singular Neutral Equations*, Volterra Centennial International Conference on the Numerical Solutions of Volterra and Delay Equations.

Robert Miller

1. *Optimization of Sensor Locations for Elastic Structures*, SIAM National Meeting, Kansas City, MO, July, 1996

Belinda King

1. *Robust Feedback Control of Parabolic Systems*, National SIAM Meeting, San Diego, CA, July, 1994.

2. *Representation Theorems for Feedback Operators*, Fourth International Conference on Computation and Control, Bozeman, MT, August, 1994.

3. *Low Order Controllers for Continuous Systems*, IEEE Conference on Decision and Control, Orlando, FL, December, 1994.

4. *Feedback Control of Nonlinear Systems*, SIAM Conference on Control, St. Louis, MO, April, 1995.

5. *Reduced Basis Approach for Feedback Control of Hyperbolic Systems*, SIAM National Meeting, Charlotte, NC, October 1995

6. *Nonlinear Dynamic Compensator Design for Flow in A Driven Cavity*, IEEE Conference on Decision and Control, New Orleans, LA, December 1996

John Burkardt

1. *Discretization of Cost and Sensitivities in Shape Optimization*, Fourth International Conference on Computation and Control, Bozeman, MT, August, 1994.

Jeff Borggaard

1. *A Sensitivity Equation Approach for the Optimal Design of Nozzles*, AIAA 5th Symposium on Multicisciplinary Analysis and Optimization, Panama City, FL, September, 1994.

2. *The Sensitivity Equation Method for Optimal Design*, Oregon State University, Mathematics Department Colloquium, Corvallis, OR, February, 1995.

3. *The Sensitivity Equation Method for Optimal Design*, SIAM Annual Meeting, Charlotte, NC, October, 1995.

4. *Optimal Design Using the Boundary Element Method*, AMS Fall Southern Sectional Meeting, Greensboro, NC, November, 1995.

5. *On Active Control of Flow Induced Vibrations*, 34th IEEE Conference on Decision and Control, New Orleans, LA, December, 1995.

6. *A PDE Sensitivity Equation Approach to Optimal Design*, Mathematics Department Colloquium, Virginia Tech, Blacksburg, VA, May, 1996.

7. *On Optimal Design Using an Adaptive Finite Element Method*, First International Conference on Nonlinear Problems in Aviation and Aerospace, Daytona, FL, May, 1996.

8. *Computing Design Sensitivities Using an Adaptive Finite Element Method*, 27th AIAA Computational Fluid Dynamics Conference, New Orleans, LA, June, 1996.

9. *Optimal Design Using Adaptive Finite Elements*, Industrial Materials Institute, Montreal, Canada, February, 1997.

Yuh-Roung Ou

1. *On the Robustness of the Navier-Stokes Global Attractors*, The 14th IMACS World Congress, Atlanta, GA, July 1994.

2. *Feedback Control of the Driven Cavity Problem using LQR Designs*, the 33rd IEEE Conference on Decision and Control, Orlando, FL December 14-16, 1994.

3. Workshop of Dynamics and Control of Turbulence and Combustion: Basic Research and Industrial Applications, sponsored by Rocketdyne Division/Rockwell International, Canoga Park, CA, April 5, 1995.

4. *High-Performance Computing Experience at NASA and Virginia Tech.*, Technical Lecture at the Naval Command, Control and Ocean Surveillance Center, San Diego, CA, April 7, 1995.

Justin Appel

1. *Optimization-Based Design in High Speed Flows*, 1995 ASME International Mechanical Engineering Congress and Exposition, San Francisco, CA, November, 1996.

2. *Sensitivity Calculation in Flows with Discontinuities*, 14th AIAA Applied Aerodynamics Conference, New Orleans, LA, June, 1996.

3. *Calculating Nearby Flows using Sensitivities*, SIAM 1996 Annual Meeting, Kansas City, MO, July, 1996.

Lena Sadtchikova

1. *Computational Issues in Optimization-Based Design*, 34th IEEE Conference on Decision and Control, New Orleans, December 1995.

Ajit Shenoy

1. *An Optimal Control Formulation of a Flow Matching Problem*, AIAA 5th Symposium on Multicisciplinary Analysis and Optimization, Panama City, FL, September, 1994.

2. *Thermal-Fluid Control via Finite-Dimensional Approximation*, 31st AIAA Thermophysics Conference, New Orleans, June 1996.

3. *On the Optimality System for a 1-D Euler Flow Problem*, 6th AIAA/NASA/USAF Multidisciplinary Analysis & Optimization Symposium, Bellevue, September 1996.

Diana Rubio

1. *Regularity of Feedback Operators for Boundary Control of Thermal Processes*, First International Conference on Nonlinear Problems in Aviation and Aerospace, Daytona, FL, May, 1996.

7.2 Air Force Interactions

Arnold Engineering Development Center

This interaction was a continuation of the work on the design and improvement of wind tunnel testing facilities and procedures. The technical work has been described as part of the broader applications being pursued under industrial liason with **Sverdrup**. **CODAC** began working with AEDC in the development of a sensitivity module for the PARC code used in optimal forebody design problems. This effort was extended to other groups and CFD codes at AEDC. In particular, we obtained XAIR from Sverdrup and we have been working to include sensitivity modules in this CFD code. XAIR is the primary CFD tool used by the Sverdrup group at AEDC. In addition, we are continuing our work with AEDC in the development of an unstructured grid CFD code for 3D problems.

As a side benefit to the **CODAC** - AEDC effort, Dr. Steve Keeling (Sverdrup) has started several joint projects with our partners at North Carolina State (Banks and Fitzpatrick). This interaction came about through an initial meeting in Blacksburg. During this meeting Burns, Fitzpatrick and Keeling focused on the use of surrogates as one approach to optimal design. This discussion also included the application of convexity based methods used by Keeling. Keeling and Fitzpatrick have used these ideas in several design projects at AEDC. The result of this effort is that VA Tech will focus on sensitivity and adjoint methods while Fitzpartick and Keeling concentrate on surrogates and convexity approaches. This plan addresses the need to develop methods for the quick-and-dirty designs as well as more complex methods needed if one is to make radical steps in optimization of these systems. **CODAC**, North Carolina State and Sverdrup continue to interact on these problems.

Phillips Laboratory

During the past four years, **CODAC** personnel have worked jointly on several efforts with researchers at the Space Experiments Directorate of Phillips Laboratory (PL/SX). One aspect was an

independent validation and verification (IVV) effort in support of the Miniature Sensor Technology Integration Program (MSTI). While it is not funded under the URI program, it does involve personnel from CODAC, as well as Dr. Jason L. Speyer - a participating researcher in the AFOSR's URI effort at UCLA.

The MSTI program is a technology demonstration effort that highlights the use of small, relatively inexpensive space-assets for tracking theater ballistic missiles. While MSTI will employ various infrared passive detectors, the essential feature is that the sensor is capable of only angular measurements of the relative target position (no direct range measurement). The challenge is to create a filter algorithm/structure that can develop accurate reliable estimates of the target's position and velocity based on angles–only measurements.

For the MSTI–2 satellite, which was launched on 8 May 1994, this problem is exacerbated by the fact that sensor limitations will permit tracking only during the boost–phase of the flight. That is, tracking of the cold - nonthrusting - missile is not possible. Our work has included the implementation of advanced filter structures based on a global linearization of the state-to-measurement map and on a disturbance attenuation formulation. Theoretical aspects of this were developed by Professor Speyer under earlier AFOSR support. Beyond this we have also developed a Matlab–based software suite and implemented several new models for the target motion. We expect to test our algorithms with data from upcoming flight experiments.

A second project, related to PL/SX's larger ballistic missile defense mission, was a sensitivity analysis to characterize the accuracy of a new approach to target tracking. The idea is to make better use of early launch detection measurements and so to improve track-estimation. This can permit early launch of ground-based interceptors.

Phillips Laboratory

This is a joint research effort that is based on work in optimal design and control of distributed parameter systems currently in progress at **CODAC**. The Structures and Controls Division of Phillips Laboratory is responsible for the design and fabrication of an Advanced Controls Technology Experiment (ACTEX). A problem of controlling the payload fairing noise during launch has been identified by Phillips Laboratory as an important issue in launch vehicles. scientists at Phillips Laboratory. We are investigating the use of distributed parameter control and various active control actuators as possible solutions to this problem. Mr. Christopher Niezrecki (an ASSERT student) from Virginia Tech spent the summer at Phillips Laboratory working on this project. The work at **CODAC**will focus on the use of distributed-parameter control theory to optimally design and locate sensors and actuators and on design of embedded actuators and sensors.

Wright Aeronautical Laboratories

CODAC personnel have provided support to Dr. Banda's group at the Flight Dynamics Laboratory in two areas. In February, Drs. Burns and Cliff were invited by Dr. Banda to participate in technical interactions at Cal Tech. The purpose of these meetings was to better understand issues of low-speed, compressor-control research; the group at Wright Labs will be pursuing control studies for high-speed compressors. In 1996, Dr. Cliff participated in an invited workshop at WPAFB on guidance and control issues for uninhabited autonomous flight vehicles.

7.3 Transitions and Industrial Interactions

Aerosoft, Inc.

AeroSoft develops and licenses user-friendly software for CFD applications. In addition, the company provides consulting, applications-oriented analysis and customer training to a range of government and industrial clients. The General Aerodynamics Simulation Program (**GASP**) is a principal product. By user-option, **GASP** incorporates models with a variety of flow-physics, including finite-rate chemical reactions and turbulence models. **CODAC** has a continuing relationship

with AeroSoft, focusing on the development of a software product for determining flow-field sensitivity by post-processing CFD solutions. The underlying formulation relies on the SEM approach to accept flow-solution and grid information which defines a linear boundary value problem (the Sensitivity Equation) for the sensitivity. AeroSoft is part of the new **PRET** center.

Analytical Mechanics Associates

AMA develops software and supplies technical consulting services for the aerospace industry. CODAC worked under sub-contract with AMA to improve stability and efficiency of algorithms for the numerical solution of aerospace trajectory optimization problems. The research centers around alternative techniques for approximating infinite-dimensional control problems. Current methods are based on a standard differential-equation formulation of the dynamics and introduce approximations for the control variables and (possibly) for the state variables. We are investigating a formulation based on differential-inclusions, wherein it is only necessary to approximate the state variables. Evidence indicates that such formulations exhibit a wider domain of convergence and that they can faithfully capture certain 'singular-control' aspects that commonly appear in aerospace trajectory-shaping.

Aurora Flight Sciences Inc

Aurora builds remotely piloted vehicles for atmospheric research applications. One current effort is directed toward a vehicle that will fly over the South Pole from a bases in New Zealand. We assisted in the development of software for near real-time trajectory optimization. The main issue is that the flight speed of the vehicle is about the same magnitude as the prevailing winds. In these cases it requires careful flight planning to permit a successful scientific mission and safe return of the vehicle.

Our approach was to transcribe the infinite-dimensional optimal control problem to an approximating finite-dimensional one. Our parameterization of the path includes the notion of way-points, loosely points in space at which the speed and heading of the vehicle is specified. This will interface nicely with the flight control strategy used in the guidance block. Our work focused on the development and validation of mathematical models for this problem, as well as on the necessary software.

It should be noted that such autonomous vehicles are also of interest in gathering military intelligence over hostile areas.

BEAM Engineering and Applied Research

During the past four years scientists at **CODAC** and BEAM have worked on several projects involving optimal design and control. Two directions that make up the present thrust are in the development of mathematical tools for performing engineering analyses and in the construction of low order local observers for feedback control of fluids. BEAM is developing software for a variety of industrial applications, mostly in the area of fluid mechanics. However, we are also working on software for material processing with the goal of attacking metal forging problems. This joint research will make use of Sensitivity Equation procedures for design optimization. We now have a formal agreement with BEAM and have expanded our interactions through the new **PRET** center.

Sverdrup

The Sverdrup operations at Tullahoma, Tennessee include a responsibility for designing advanced ground test facilities for use by the Air Force as well as civilian aerospace companies. Additional responsibilities include the development of software capable of attacking complicated flow problems such as those present during store separation. As part of this effort, Sverdrup scientists and engineers are actively involved in the development of advanced CFD methodology. We are continuing our joint research and software development program with Sverdrup which focuses on the following problems.

Currently, **CODAC** has a copy of XAIR (MicroCraft's primary CFD code) and we worked with Sverdrup to modify this code so that it can be used to compute sensitivity derivatives with respect to shape and flow parameters. These derivatives can be used in optimization and design studies,

and also by aerodynamicists in studying flow properties. This effort is the first step towards the development of optimal design methodologies that are to be applied to the design of wind tunnel components. The goal is to enhanced test section flow quality. This effort complimented the effort by Keeling and Fitzpatrick on convexity based methods and the long term goal is to develop algorithms for radical advances in design.

Tektronix Graphics, Printing and Imaging Division

Tektronix is a major manufacturer of electronic instrumentation; its GPI Division develops and manufactures the Phaser family of printers. The Phaser 340 is a radical new ink-jet print engine that is based on off-set printing ideas. However, there remains several control and design problems that, if resolved, could greatly enhance speed and quality. The device is intended to produce droplets of ink in response to commands. The performance issues center the around the rate at which droplets are produced and the quality of the droplets (size, shape, velocity). The physical phenomena are largely fluid mechanics, including surface tension and free-surface effects in the drop formation process. The activation mechanism, whereby a pressure pulse is imparted to the fluid, is also of interest.

We are continuing our joint research effort to improve the analytical design tools for this class of problems. We have formulated related optimization problems (e.g. to produce a drop of given mass and momentum in minimum time) and constructed a model to be used for feedback control design. Our earlier work on fluid control, including the experience with shape optimization at AEDC, has played a key role in the initial effort. Tektronix is providing partial support for Dr. Paul Gilmore to work jointly with Tektronix and **CODAC**. Dr. Gilmore is currently located at Tektronix in Wilsonville and has made considerable progress on the development of new control and optimization algorithms for the printhead design and control.

Although the primary motivation for this work is to improve ink jet performance, much of the basic research is also applicable to fuel injection devices and many combustion problems. It is important to note that this same technology is applicable to the design of certain propulsion systems and fuel injection devices common in many Air Force systems. This project provides an example of dual usage where basic research on Air Force problems also has payoffs in industrial development.

7.4 Coupling Activities

Captain Dawn Stewart of the Air Force Academy joined **CODAC** in the fall of 1995 and is working on her Ph.D. in applied mathematics. Her research will involve the development of computational methods for shape optimization in metal forging. Captain Greg Agnes of Wright-Patterson Air Force Base (WL-FIBG) and Dino Schiulli of the AF Phillips Lab joined our Ph.D. program in Engineering Science and Mechanics in the fall of 1994, as part of their Air Force duties.

Chapter 8

Inventions and Patents

No patents nor trademarks were applied for during this period.

Chapter 9

Honors and Awards

- Professor John Burns was appointed the first Hatcher Endowed Professor of Mathematics.
- Dr. John Burkardt received the Mathematics Department's Outstanding Teacher Award for 1994-95.

Chapter 10

Visitors

The Center offers an unparalleled potential for strengthening the educational and scientific infrastructure by training students and post-doctoral researchers in an interdisciplinary team approach to scientific and engineering research. The Center provides unique opportunities for theoretical, computational, and experimental research. Through the interactions with Air Force laboratories, industrial partners and a strong visitors program, students are exposed to real applications. The combined theoretical, computational, and experimental approach provides a meaningful interdisciplinary research experience. The visitors program is the centerpiece for the educational and outreach programs.

This component of **CODAC** includes a program of visits by Air Force laboratory scientists, industrial partners and scholars from all over the world. During the period from 1 May 1993 through 30 April 1997, **CODAC** hosted more than 150 visitors from more than 13 countries.

10.1 1993-1994 Visitors

May, 1993

Kirsten Morris
Department of Applied Mathematics
University of Waterloo

Ralph Showalter
Department of Mathematics
University of Texas

June, 1993

Qiang Du
Department of Mathematics
Michigan State University

Kazufumi Ito
Center for Research in Scientific Computation
North Carolina State University

September, 1993

David Dew-Hughes
Department of Engineering Science
Oxford University

A.J. Meir
Department of Mathematics
Auburn University

October, 1993

Elijah Polak
Department of Electrical Engineering and Computer Science
University of California-Berkeley

Ekkehard Sachs
Department of Mathematics
University of Trier

Srdjan Stojanovic
Department of Mathematics
University of Cincinnati

November, 1993

Jack Benek
Calspan Corporation
Arnold Engineering Development Center

Mark Briski
U.S. Air Force
Arnold Engineering Development Center

Ben Fitzpatrick
Department of Mathematics
NC State University

Peter Hoffman
Calspan Corporation
Arnold Engineering Development Center

Steve Keeling
Calspan Corporation
Arnold Engineering Development Center

Zuhair Nashed
Department of Mathematical Sciences
University of Delaware

Elijah Polak
Department of Electrical Engineering and Computer Science
University of California-Berkeley

Ekkehard Sachs
Department of Mathematics
University of Trier

Srdjan Stojanovic
Department of Mathematics
University of Cincinnati

Mohsen Tadi
Department of Chemistry
Princeton University

December, 1993

Gal Berkooz
Department of Mathematics
Cornell University

Srdjan Stojanovic
Department of Mathematics
University of Cincinnati

January, 1994

Srdjan Stojanovic
Department of Mathematics
University of Cincinnati

February, 1994

Elena Fernandez
CAC-Comision Nacional
de Energia Atomica

Andrei Fursikov
Department of Mechanics and Mathematics
Moscow State University

Steve Hou
Department of Mathematics
York University

Manfred Laumen
Department of Mathematics
University of Trier

David Ross
Eastman Kodak Company

Srdjan Stojanovic
Department of Mathematics
University of Cincinnati

Tom Svobodny
Department of Mathematics
Wright State University

Janos Turi
Programs in Mathematical Sciences
University of Texas-Dallas

James Turner
Department of Mathematics
Ohio State University

March, 1994

Elena Fernandez
CAC-Comision Nacional
de Energia Atomica

Andrei Fursikov
Department of Mechanics and Mathematics
Moscow State University

Belinda King
Department of Mathematics
Oregon State University

Manfred Laumen
Department of Mathematics
University of Trier

Jason Speyer
Department of Mechanical Aerospace and Nuclear Engineering
University of California-Los Angeles

Srdjan Stojanovic
Department of Mathematics
University of Cincinnati

April, 1994

Eyal Arian
NASA Langley Research Center

Gal Berkooz
Department of Mathematics
Cornell University

Dennis Brewer
Department of Mathematical Sciences
University of Arkansas

Daniela Calvetti
Department of Mathematics
Stevens Institute of Technology

Jonathan Chapman
Mathematical Institute
Oxford University

Michel Delfour
Cent. Rech. Math.
Universite de Montreal, Canada

John Dennis
Computational and Applied Mathematics
Rice University

Qiang Du
Department of Mathematics
Michigan State University

Richard Fabiano
Department of Mathematics
Texas A & M University

Omar Ghattas
Department of Civil Engineering
Carnegie Mellon University

Andreas Griewank
Department of Mathematics
Dresden University, Germany

Jaroslav Haslinger
Department of Mathematics
Charles University
Czech Republic

David Hudak
The Analytical Science Corporation
Washington, DC

Tim Kelley
Department of Mathematics
North Carolina State University

Manfred Laumen
Department of Mathematics
University of Trier

Perry Newman
NASA Langley Research Center

Elijah Polak
Department of Electrical Engineering and Computer Science
University of California-Berkeley

Jason Speyer
Dept. of Mechanical Aerospace and Nuclear Engineering
University of California-Los Angeles

Janos Turi
Programs in Mathematical Sciences
University of Texas-Dallas

10.2 1994-1995 Visitors

May, 1994

John Ockendon
Mathematical Institute
Oxford University

June, 1994

Dieter Dinkler Institut für Statik und Dynamik
Universität Stuttgart

Belinda King
Department of Mathematics
Oregon State University

Arik Melikyan
Institute for Problems in Mechanics
IPM RAS

Helena Wisniewski
Advanced Program Development
Titan Corporation

July, 1994

Dieter Dinkler
Institut für Statik und Dynamik
Universität Stuttgart

Gyou-Bong Lee
Department of Mathematics
Keonyan University

Hans Josef Pesch
Department of Mathematics
University of Munich

Fredi Troltzsch
Department of Mathematics
Technical University of Chemnitz-Zwickau

August, 1994

Dieter Dinkler
Institut für Statik und Dynamik
Universität Stuttgart

Qiang Du
Department of Mathematics
Michigan State University

Steve Hou
Department of Mathematics
York University

Zhuangyi Liu
Department of Mathematics and Statistics
University of Minnesota

A.J. Meir
Department of Mathematics
Auburn University

James Turner
Department of Mathematics
Ohio State University

September, 1994

Dieter Dinkler
Institut für Statik und Dynamik
Universität Stuttgart

Qiang Du
Department of Mathematics
Michigan State University

Messoud Efendiev
Department of Mathematics
University of Berlin

Steve Hou
Department of Mathematics
York University

William Layton
Department of Mathematics
University of Pittsburgh

A.J. Meir
Department of Mathematics
Auburn University

Rudolf Scherer
Institute of Practical Mathematics
University of Karlsruhe

Victor Shubov
Department of Mathematics
Texas Tech University

Klaus Well
Institute for Flight Mechanics and Control
Universität Stuttgart

October, 1994

Dieter Dinkler
Institut für Statik und Dynamik
Universität Stuttgart

John Rice
Department of Computer Science
Purdue University

Janos Turi
Programs in Mathematical Sciences
University of Texas - Dallas

Luther White
Department of Mathematics
University of Oklahoma

November, 1994

Thomas Bewley
Department of Mathematics
Stanford University

Ben Fitzpatrick
Department of Mathematics
North Carolina State University

Kazufumi Ito
Department of Mathematics
North Carolina State University

Stephen Keeling
CFD Department
Micro Craft Technology
Arnold Air Force Base

Belinda King
Department of Mathematics
Oregon State University

Carlos Neto
Department of Electrical Engineering
University of California - Berkeley

December, 1994

Vassilios Dougalis
Department of Mathematics
National Technical University
Athens, Greece

Belinda King
Department of Mathematics
Oregon State University

January, 1995

Andrei Fursikov
Department of Mechanics and Mathematics
Moscow State University

Steve Hou
Department of Mathematics and Statistics
York University

James Turner
Department of Mathematics
Ohio State University

February, 1995

Qiang Du
Department of Mathematics
Michigan State University

Andrei Fursikov
Department of Mechanics and Mathematics
Moscow State University

Janos Turi
Programs in Mathematical Sciences
University of Texas-Dallas

March, 1995

Ferenc Hartung
Programs in Mathematical Sciences
University of Texas-Dallas

Belinda King
Department of Mathematics
Oregon State University

Danny Sorensen
Department of Computational and Applied Mathematics
Rice University

Janos Turi
Programs in Mathematical Sciences
University of Texas-Dallas

Luis Vicente
Department of Computational and Applied Mathematics
Rice University

April, 1995

Dennis Brewer
Department of Mathematics
University of Arkansas

10.3 1995-1996 Visitors

May, 1995

Dominique Pelletier
Département de génie Mécanique
Ecole Polytechnique de Montréal

June, 1995

Rossitza Karamikhova
Bell Helicopter
Arlington, TX

Robert E. Miller
Department of Mathematical Sciences
University of Arkansas

July, 1995

Sungkwon Kang
Department of Mathematics
Chosun University

August, 1995

Sungkwon Kang
Department of Mathematics
Chosun University

September, 1995

Jonathan Chapman
Mathematical Institute
Oxford University

Rich Fabiano
Department of Mathematics
University of St. Thomas

Belinda King
Department of Mathematics
Oregon State University

John Ockendon
Mathematical Institute
Oxford University

Peter Görtz
Universität Karlsruhe
Institute für Praktische Mathematik

October, 1995

Martin Berggren
FFA the Aeronautical Research Institute of Sweden
Computational Aerodynamics

Chris Byrnes
School of Engineering and Applied Sciences
Washington University

Janos Turi
Programs in Mathematical Sciences
University of Texas at Dallas

November, 1995

Qiang Du
Department of Mathematics
Michigan State University

Zhuangyi Liu
Department of Mathematics
University of Minnesota/Duluth

Rick Newsome
Beam Technologies
Ithaca, NY

Jiongmin Yong
Department of Mathematics
University of Tennessee

December, 1995

Paul Gilmore
Department of Mathematics
Florida State University

February, 1996

Pablo Jacovkis
Departamento de Computacion Facultdad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Rosita Wachenchauzer
Departamento de Computacion
Facultdad de Ciencias Exactus y Naturales
Universidad de Buenos Aires

March, 1996

Arik Melikyan
Institute for Problems in Mechanics
Russian Academy of Sciences

Janos Turi
Programs in Mathematical Sciences
University of Texas at Dallas

April, 1996

Dennis Brewer
Department of Mathematical Sciences
University of Arkansas

Hermann Brunner
Department of Math and Statistics
Memorial University of Newfoundland

10.4 1996–1997 Visitors

May, 1996

Martin Berggren
FFA, Computational Aerodynamics Department
Bromma, Sweden

Fariba Fahroo
The Naval Postgraduate School

September, 1996

Mary E. Bradley
Brown University

Elena M. Fernandez-Berdaguer
Comision de Energia Atomica
Buenos Ares, Argentina

Ruben D. Spies
INTEC-PEMA
Santa Fe, Argentina

October, 1996

Beth E. Bradley
Brown University

James C. Ellenbogen
The Mitre Corporation
McLean, Virginia

Richard Fabiano
Department of Mathematical Sciences
University of North Carolina-Greensboro

Dominique Pelletier
Department of Mechanical Engineering
Ecole Polytechnique de Montreal

Ruben D. Spies
INTEC-PEMA
Santa Fe, Argentina

November, 1996

Jacalyn Huband
University of Charleston

February, 1997

Ben Fitzpatrick
Department of Mathematics
North Carolina State University

Patrick Justen
Department of Mathematics
Universitat Trier
Trier, Germany

Pedro Morin
INTEC-PEMA
Santa Fe, Argentina

Suely Oliveira
Department of Computer Science
Texas A & M

Gunther Peichl
Department of Mathematics
Karl-Franzens-University of Graz
Graz, Austria

Janos Turi
Programs in Mathematical Sciences
University of Texas-Dallas

March, 1997

Jeanne Atwell
Department of Mathematics
Oregon State University

Carmen Chicon
Department of Mathematics
University of Missouri

Ruth Curtain
Department of Mathematics
University of Groningen
The Netherlands

Mary Gallo
Department of Mathematics
Oregon State University

Belinda King
Department of Mathematics
Oregon State University

Pedro Morin
INTEC-PEMA
Santa Fe, Argentina

Job Oostveen
Department of Mathematics
University of Groningen
The Netherlands

Tim Randolph
Mathematics Department
University of Missouri

Ekkehard Sachs
Department of Mathematics
Universitat Trier

April, 1997

H.T. Banks
Center for Research in Scientific Computation
North Carolina State University

Dennis Brewer
Department of Mathematical Sciences
University of Arkansas

Hermann Brunner
Department of Mathematics and Statistics
Memorial University of Newfoundland

Richard Fabiano
Department of Mathematics
University of North Carolina-Greensboro

Jerome Goldstein
Department of Mathematical Sciences
University of Memphis

Gisele Goldstein
Center for Earthquake Research and Information
University of Memphis

David Gilliam
Department of Mathematics
Texas Tech University

Belinda King
Department of Mathematics
Oregon State University

Zhuangyi Liu
Department of Mathematics
University of Minnesota-Duluth

Robert Miller
Department of Mathematical Sciences
University of Arkansas

Dominique Pelletier
Department of Mechanical Engineering
Ecole Polytechnique de Montreal

Jason Speyer
Mechanical, Aerospace and Nuclear Engineering
University of California-Los Angeles

Mohsen Tadi
Department of Mathematics, Statistics and Computer Science
University of Illinois at Chicago