RATES OF CONVERGENCE FOR DISCRETE APPROXIMATIONS TO PROBLEMS IN CONTROL THEORY

by

William Ward Hager

B.S. Harvey Mudd College
1970

M.S. Massachusetts Institute of Technology
1971

SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE

DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1974

Signature of Author..... William W. Hager
Department of Mathematics, May 3, 1974

Certified by..... Gilbert Strang
Thesis Supervisor

Accepted by.....
Chairman, Departmental Committee on Graduate Students

RATES OF CONVERGENCE FOR DISCRETE APPROXIMATIONS TO PROBLEMS IN CONTROL THEORY

by

William Ward Hager

Submitted to the Department of Mathematics on May 3, 1974 in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Abstract

Rates of convergence of the solution to discrete approximations of the primal and dual control problem are studied. The convergence theory is preceded by an analysis of the Lagrangian dual control problem where it is shown that in "convex problems," the Slater condition implies the existence of a dual solution in problems with inequality constraints on the state and the control. Furthermore when an optimal primal solution exists, a complementary slackness condition holds and a minimum principal relating the dual variables to the state and control minimizing the dual function is proved. Convergence rates for the discretization of the dual problem by the finite element method are derived and it is shown that in piecewise polynomial spaces with fixed grid points, the solution to the discrete problem converges to the continuous solution at rate bounded by $ch^{3/2}$ in constrained problems where h is the maximum grid interval in the space. However, if the grid points are free parameters in the discrete set, then the full convergence rate expected for the piecewise polynomial spaces can be achieved. In a numerical example, it is observed and then proved rigorously that the error in the Ritz-Trefftz approximate exhibits a boundary layer phenomenon with the error $O(h)$ near points where the constraints change from binding to non-binding and $O(h^2)$ elsewhere. Finally discrete approximations to the primal unconstrained control problem are studied in which the differential equation is replaced by onestep or multistep integration schemes. It is shown that the convergence rate for the onestep

method depends on the behavior of the scheme on the right end of each grid interval while the convergence of the multistep method depends on the order of accuracy of a "built-in" initial condition at the right endpoint for the discrete costate variable.

Thesis Supervisor: Sanjoy K. Mitter
Title: Professor of Electrical Engineering

Acknowledgments

I wish to express my sincere thanks to Professor Sanjoy Mitter for his time, patience, encouragement, and wisdom in guiding my thesis work and to Professor Gilbert Strang for his encouragement and for bringing to my attention recent results concerning variational inequalities and the stability of the solution set for systems of equalities and inequalities. Also Tom Magnanti and Richard Vinter are appreciated for many helpful conversations. I am also indebted to the fellowship at Park Street Church Boston whose love and warmth was an integral part of this paper and special thanks belong to my parents for their constant love, faith, and understanding.

CONTENTS

Abstact 2

Acknowledgments 4

Overview 7

Chapter 1: Lagrange Duality Theory for Convex Control Problems 12

I. Introduction 13

II. Duality Theory 18

III. Minimum Principals 29

IV. Appendix - Regularity of Dual Solution and Existence Theorems for the Primal Problem 38

Bibliography 55

Chapter 2: The Ritz-Trefftz Method for State and Control Constrained Optimal Control Problems 56

I. Introduction 57

II. Formulation of the Problem 63

III. Existence of Solution 68

IV. Convergence of State and Control 72

V. Convergence of Dual Variables 94

VI. Error in Free Boundary 99

VII. Summary and Conclusions 108

VIII. Appendices 110

Appendix 1: Interpolation from Below 110

Appendix 2: Stability of Solution to Primal and Dual Quadratic Programing Problem 117

Bibliography 125

Chapter 3: Numerical Examples for the Ritz-Trefftz Method with Fixed Grid Points 127

I. Introduction 128

II. Problem Description 130

III. Convergence Results 134

Chapter 4: Rates of Convergence for Discrete Approximations to Unconstrained Control Problems 150

I. Introduction 151

II. Relationship Between Continuous and Discrete Necessary Conditions 154

III. Upper Bounds on Convergence Rates 166

IV. Numerical Examples 170

V. Improved One Step Schemes? 180

VI. Cost Convergence 181

Bibliography 182

OVERVIEW

The numerical solution to the following state and control constrained optimization problem is studied:

$$(P) \qquad \inf \int_0^1 L(x(t),u(t))\,dt$$

$$\text{s.t.} \quad \dot{x}(t) = f(x(t),u(t))$$

$$x(0) = x_o$$

$$K_c(u(t),t) \leq 0$$

$$K_s(x(t),t) \leq 0$$

If (P) is to be solved numerically, then it must be discretized and a knowledge of the rate of convergence of the solution to a discrete approximation to (P) to the continuous solution is essential to evaluate the efficiency of numerical algorithms.

Papers by Budak [1] and Cullum [2] proved convergence (without rate estimates) for the solution to the approximation of (P) corresponding to the Euler integration scheme. This author has analyzed rates of convergence both for the solution to discrete approximations of unconstrained control problems using standard onestep and multistep integration procedures and for the solution of the dual problem associated with (P) using the finite

element method.

The solution of the dual problem first required a development of Lagrange duality theory. In chapter 1, it is proved that in "convex problems," a solution to the dual problem exists if an interior point assumption is satisfied; and if a solution to the primal problem exists, then a complementary slackness condition holds. Also a minimum principal and an adjoint condition is developed which characterizes the state and control variables corresponding to given dual variables. The appendix also proves the existence of a bounded measureable solution to a class of control and state constrained problems.

Using the foundation built in chapter 1, the following chapter then proves convergence estimates in problems with quadratic cost and linear dynamics and constraints for the solution of the dual problem using the finite element method. Although the dual function is only semi-definite, a solution to the Ritz-Trefftz problem is proved to exist. Furthermore, for control constrained problems, the dual function is shown to be positive definite in an appropriate subspace. First order convergence of the Ritz-Trefftz method in piecewise polynomial subspaces is easily proved; however, higher order convergence requires not only that higher order spaces be used, but

also that the grid points in a neighborhood of changes in binding constraints be left as free parameters in the maximization procedure. The convergence proofs are based on a theorem on the regularity of the solution to the control problem and results on the approximation of either non-negative or monotone functions by non-negative or monotone polynomials. The regularity theorem, in turn, is based on a lemma concerning the stability of the solution to the primal and dual quadratic programing problem to changes in the data. The chapter concludes with a bound on the error in the free boundary (or time when a constraint becomes binding) for the Ritz-Trefftz problem.

Chapter 3 then studies the convergence properties of the Ritz-Trefftz method in two numerical examples- a state and a control constrained optimal control problem. In both examples the L^2 error of the Ritz-Trefftz solution is bounded by $ch^{3/2}$ for higher order piecewise polynomial subspaces where h is the grid interval associated with the space; similarly the sup norm of the Ritz-Trefftz error is bounded by ch. Away from the region where the constraint changes from binding to non-binding, however, the Ritz-Trefftz solution to the control constrained problem converged at rate bounded by ch^2 while in the state constrained problem the error was only bounded by $ch^{3/2}$.

The different convergence behavior for the two problems is accounted for by the differences in the constraints for the dual problem - the control constraints lead to necessary conditions for the finite dimensional problem that uncouple while the necessary conditions in the state constrained problem do not uncouple.

Although the dual problem has simpler constraints than the primal problem and does not require an integration of the system dynamics, the dual problem has the disadvantage of not necessarily solving the primal problem unless the constrints and cost functional are convex. Chapter 4 begins an analysis of discrete approximations to the unconstrained primal problem by considering the effect of replacing the differential equation by onestep integration schemes based on quadrature, multistep procedures, and Taylor series approximations. It is demonstrated that the convergence rate for onestep schemes depends on the behavior of the scheme at the end of each grid interval while the convergence rate for multistep methods depends on the behavior of the approximation near the right endpoint. The solution to the Taylor series approximation, on the other hand, diverged from the continuous solution. Finally the optimality of three Range-Kutta schemes is proved and two numerical examples

confirming the theory are analyzed.

References

1. B. M. Budak, E. M. Berkovich, and E. N. Solov'eva, "The Convergence of Difference Approximations for Optimal control Problems", USSR Computational Mathematics and Mathematical Physics (9), No. 3, p. 30-65

2. J. Cullum, "Finite Dimensional Approximations of State Constrained Continuous Optimal Control Problems," SIAM J. Control (10), p. 649-670

CHAPTER 1

Lagrange Duality Theory for Convex Control Problems

I. INTRODUCTION

The Lagrange dual of the following control problem is studied:

$$\begin{aligned} &\inf\ f(x,u) \\ &\text{s.t. } \dot{x}(t) = A(t)x(t) + B(t)u(t) \\ &\qquad x(0) = x_0 \\ &\qquad K_c(u(t),t) \leq 0 \\ &\qquad K_s(x(t),t) \leq 0 \end{aligned}$$

where $f(\cdot,\cdot)$, $K_c(\cdot,t)$, and $K_s(\cdot,t)$ are all convex. Rockafellar [7] has derived duality results for convex state constrained control problems using conjugate functions. The theory in this paper goes beyond Rockafellar's results since the constraints are given explicitly by inequalities above and, hence, the multipliers associated with the constraints can be characterized. Also a slightly different form of the dual problem, the Lagrange dual, is studied herein. The theory in this paper provides the foundation for an analysis of the numerical solution of the dual problem by the Ritz method in [1].

In Section 2, the principal result, based on the Hahn-Banach Theorem, proves that the dual problem has an optimal solution if there exists an interior point for the constraint

set; and if the primal problem has an optimal solution, then a complementary slackness condition holds. The optimal multipliers p and v corresponding to the system dynamics and state constraints are shown to have bounded variation while the multiplier w corresponding to the control constraints lies in L^1.

Section 3 then proves that a minimum principal holds and while (p,v) are only of bounded variation, the combination $q(t) = K_s(x^*(t),t)_x^T v(t) - p(t)$ is absolutely continuous where x^* solves the primal problem; furthermore q satisfies the conventional adjoint equation for state constrained control problems. This result has important consequences for the solution of the dual problem using the Ritz method in [1] since the convergence rate of the discrete approximation depends upon the smoothness of the dual variables; hence if the dual problem is reformulated in terms of q rather than p, then a superior convergence rate is achieved.

The Appendix contains several technical lemmas concerning the regularity of the dual variables and the existence of bounded, measureable solutions to a class of control problems.

Notation

The following notation is used for spaces of real valued functions on $[0,1]$:

$\mathcal{A}$	absolutely continuous functions
BV	functions of bounded variation
NBV	functions of normalized bounded variation
C	continuous functions
C^∞	infinitely differentiable functions
L^p	functions with $\int_0^1 \lvert f(t)\rvert^p \, dt < \infty$
B	functions bounded and measureable.

If W is any of the spaces above and x is a vector-valued function, then the notation $x \in W$ means that the components of x lie in W. Similarly, the notation $x \in W(R^n)$ means that x is a vector-valued function with n components and each component lies in W.

If $y \in R^m$, then define the E^1 norm by

$$|y| = \sum_{k=1}^{m} |y_k|$$

where y_k or $(y)_k$ denotes the k-th component of a vector. The following function norms are used:

$$||f||_{L^p} = \{ \int_0^1 |f(t)|^p \, dt \}^{1/p}$$

$$||f|| = \sup_{t\in[0,1]} |f(t)|.$$

If $x,y \in R^m$, the inner product $(\cdot,\cdot)$ is defined by

$$(x,y) = \sum_{k=1}^{m} x_k y_k.$$

If f,g are vector-valued measureable functions on $[0,1]$, $v \in BV$, and $h \in C$, then define:

$$((f,g)) = \int_0^1 (f(t),g(t)) \, dt$$

$$[v,h] = \int_0^1 h(t) \, dv(t)$$

The distance between two sets A and B in some norm, $||\cdot||_W$ is defined by:

$$dist\{A,B\}_W = \inf_{\substack{a\in A \\ b\in B}} ||a-b||_W.$$

The total variation of a function f on the interval $[0,t]$ is denoted $TV(t,f)$ and $TV(f) \equiv TV(1,f)$.

The complement, closure, and boundary of a set are denoted $A^c, \overline{A}$ and ∂A, respectively. Another notation for closure is $cl\{A\}$.

The notation $A \prec f \prec B$ means that f is a real valued function with support contained in B, $f(a) = 1$ for $a \in A$, and $||f|| \leq 1$.

II. DUALITY THEORY

The following control problem is considered:

$$
\begin{aligned}
(P) \qquad & \inf\ c(x,u) \\
& \text{s.t. } c(x,u) = \int_0^1 h(x(t),u(t),t)\,dt \\
& \dot{x}(t) = A(t)x(t) + B(t)u(t) \\
& x(0) = x_0 \\
& K_c(u(t),t) \le 0 \\
& K_s(x(t),t) \le 0 \\
& x \in \mathcal{A}(R^n), \quad u \in B(R^m)
\end{aligned}
$$

where h, K_c, and K_s have range in R, R^{m_c}, and R^{m_s}, respectively, and the matrices A and B are of the appropriate dimensions. Note that there exists a control $u \in B$ that solves (P) if the feasible controls lie in a compact set in R^m or h satisfies a growth assumption (see Appendix and [7]). The dual function L is given by

$$
\begin{aligned}
(1) \qquad L(p,w,v) = \inf\ & c(x,u) + ((p,\dot{x}-Ax-Bu)) + ((w,K_c(u))) \\
& + [v,K_s(x)] \\
& \text{s.t. } x(0) = x_0 \\
& x \in \mathcal{A}, \quad u \in B.
\end{aligned}
$$

The dual problem corresponding to (P) is then

$$(D) \qquad \sup L(p,w,v)$$
$$\text{s.t. } p,v \in BV, \quad w \in L^1$$
$$w \geq 0, v \text{ non-decreasing}$$

In order that all the terms in (P) and (1) above make sense, assumptions must be made concerning the functions appearing in these problems. Theorem 1 will require the following:

(A1) $h(\cdot,\cdot,t)$, $K_s(\cdot,t)$, and $K_c(\cdot,t)$ are convex for $t \in [0,1]$, $A(\cdot)$ and $B(\cdot)$ have components in L^1, and $h(\cdot,\cdot,\cdot)$, $K_s(\cdot,\cdot)$, and $K_c(\cdot,\cdot)$ are all continuous.

(A2) There exists a control $\bar{u} \in C$ and a corresponding trajectory $\bar{x}$ such that $(K_c(\bar{u}(t),t))_j < a < 0$ and $(K_s(\bar{x}(t),t))_j < a < 0$ for some "a", for all $t \in [0,1]$, and for all components of K_c and K_s.

Proposition 1 below, the "weak duality theorem," is easily verified. This is followed by the principal theorem, or "strong duality" result.

Proposition 1

$c(x,u) \geq L(p,w,v)$ whenever (x,u) are feasible in (P) and (p,w,v) are feasible in (D).

Theorem 1

Suppose (A1) and (A2) hold and the optimal value $\hat{c}$ of (P) is finite. Then there exists (p,w,v) that are optimal in (D) with $L(p,w,v) = \hat{c}$. If (x,u) are optimal in (P), then complementary slackness holds:

$$w(t)_j = 0 \quad \text{whenever} \quad K_c(u(t),t)_j < 0$$

v_j is constant on every interval where

$$K_s(x(t),t)_j < 0.$$

Hence, (x,u) achieve the minimum in (1) for (p,w,v).

Rather than prove Theorem 1 directly, we first consider a slightly more general and notationally tractable problem:

$$(\overline{P}) \qquad \inf f(x,u)$$

$$\text{s.t. } \dot{x}(t) = A(t)x(t) + B(t)u(t)$$

$$x(0) \in X_0$$

$$K_s(x(t),t) \leq 0$$

$$u(t) \in U(t) \quad \text{for all} \quad t \in [0,1]$$

where f is a functional defined on $\mathcal{A} \times B$. The corresponding dual function is

$$(2) \qquad L(p,v) = \inf f(x,u) + ((p,\dot{x}-Ax-Bu)) + [v,K_s(x)]$$
$$\text{s.t. } x \in \mathcal{A}, \quad u \in B$$
$$x(0) \in X_0, \quad u(t) \in U(t) \text{ for all } t \in [0,1].$$

The dual problem is then:

$$(\bar{D}) \qquad \sup \bar{L}(p,v)$$
$$\text{s.t. } p,v \in BV$$
$$v \text{ non-decreasing.}$$

Define $X = \{x \in \mathcal{A}: K_s(x(t),t) \leq 0 \text{ for all } t \in [0,1]\}$ and $U = \{u \in B: u(t) \in U(t) \text{ for all } t \in [0,1]\}$ and make the following assumption:

(A3) $f(\cdot,\cdot)$, $K_s(\cdot,t)$, $U(t)$, and X_0 are convex for all $t \in [0,1]$ and there exists a control $\bar{u} \in C$, a corresponding trajectory $\bar{x}$, and constants $M, \rho, \alpha > 0$ such that $\bar{u} \in U$, $\bar{x} \in X$, $\bar{x}(0) \in X_0$, $\text{dist}\{\bar{x}, \partial X\}_C > \alpha > 0$, $K_s(\bar{x}(t),t)_j < -\alpha < 0$, and $f(x,\bar{u}) < M$ whenever $||x-\bar{x}|| \leq \rho$. Also $K_s(\cdot,\cdot)$ is

continuous and $A(\cdot), B(\cdot)$ have components in L^1.

Lemma 1

Suppose (A3) holds and $\hat{c}$, the optimal value of $(\overline{P})$ is finite. Then there exists (p,v) that are optimal in $(\overline{D})$ and $\overline{L}(p,v) = \hat{c}$. If (x,u) is optimal in $(\overline{P})$, then v_j is constant on every interval where $K_s(x(t),t)_j < 0$; hence (x,u) achieves the minimum in (2) for (p,v).

Proof: Lemma 1 follows from an application of the Hahn-Banach Theorem.

Step 1: Construction of convex sets.

Define the sets:

$$Y = \{(a,b,c): \ a\in R^1, b\in L^1(R^n), c\in C(R^{m_s}), a\leq \hat{c}, b=0, c\leq 0\}$$

$$Z = \{(a,b,c): \ a\in R^1, b\in L^1(R^n), c\in C(R^{m_s}) \text{ and there exists } x\in\mathcal{A}, u\in B \text{ with } x(0)\in X_0,$$

$$a\geq f(x,u), \quad u(t)\in U(t)$$

$$b(t)=\dot{x}(t)-A(t)x(t)-B(t)u(t)$$

$$c(t)\geq K_s(x(t),t) \text{ for all } t\in[0,1]\}.$$

The sets Y, Z are contained in the space $E^1 \times L^1 \times C$ with norm N defined by $N(a,b,c) = |a| + ||b||_{L^1} + ||c||$. By the convexity of X_0, K_s, U, and f, it follows that Z is convex while Y on the other hand, is trivially convex.

Step 2: $(M+1, 0, 1)$ is an interior point of Z where "0" and "1" denote vectors of n zeroes and m_s ones, respectively.

For $d \in L^1(R^n)$, define x_d to be the solution to $\dot{y}(t) = A(t)y(t) + B(t)\bar{u}(t) + d(t)$, $y(0) = \bar{x}(0)$ where $(\bar{x}, \bar{u})$ was given in (A3). Now $x_d - \bar{x}$ satisfies the equation $\dot{y}(t) = A(t)y(t) + d(t)$, $y(0) = 0$ and since A has components in L^1, the solution to this equation can be bounded: $||x_d - \bar{x}|| \leq \gamma ||d||_{L^1}$ for some constant γ. Choose $0 < e \leq \min(.5, \rho/\gamma, \alpha/\gamma)$. Then it is proved that all points inside the ball of radius e centered at $(M+1, 0, 1)$ lie in Z. Suppose $N(a-M-1, b, c-1) \leq e$; then $x_b, \bar{u}$ satisfy all the conditions above for (a,b,c) to be contained in Z; i.e.,

$$\dot{x}_b = Ax_b + B\bar{u} + b$$

$$||x_b - \bar{x}|| \leq \gamma ||b||_{L^1} \leq \gamma e \leq \rho, \alpha$$

$$K_s(x_b(t), t)_j \leq 0 \leq 1 - e \leq c_j(t) \quad \forall\, t \in [0,1]$$

$$a \geq M + 1 - e \geq M + \tfrac{1}{2} \geq f(x_b, \bar{u}) \quad \text{since} \quad ||x_b - \bar{x}|| \leq \rho.$$

Step 3: No point of Y lies in Z°, the interior of Z.

Suppose $(a,b,c) \in Y$ and $(a,b,c) \in Z^\circ$. Thus, $a \leq \hat{c}$, $b = 0$, and $c \leq 0$ and for e small enough $(a-e,b,c) \in Z$. Hence, there exists (x,u) with $f(x,u) \leq a - e \leq \hat{c} - e$ and (x,u) satisfy all the conditions given in the definition of Z. These conditions imply that (x,u) are feasible in $(\overline{P})$ and hence this contradicts the optimality of $\hat{c}$.

Now by the Hahn-Banach Theorem [4], there exists a hyperplane separating Z and Y; i.e., there exist $r \in R^1$, $p \in L^\infty$, and $v \in NBV$ satisfying:

$$(3) \qquad (r,a_1) + ((p,b_1)) + [v,c_1] \geq (r,a_2) + ((p,b_2)) + [v,c_2]$$

for all $(a_1,b_1,c_1) \in cl\,(Z)$, $(a_2,b_2,c_2) \in cl(Y)$

Step 4: Properties of the hyperplane.

By choosing particular points in Y and Z, properties of the separating hyperplane will be exhibited:

(a) $r \geq 0$.

Substitute $a_2 = \hat{c} - 1$, $a_1 = f(\overline{x},\overline{u})$, $b_1 = b_2 = c_1 = c_2 = 0$ in (3) where $(\overline{x},\overline{u})$ were given in (A3).

(b) v is monotone non-decreasing.

For notational convenience, v is assumed scalar valued although for vector-valued functions the proof is identical.

Since $v \in BV$, the set of continuous point has full measure. Suppose t, s are points of continuity of v and let c_d denote the continuous function that is -1 on $[t+d, s-d]$, 0 outside $[t,s]$, and linear on $[t,t+d]$ and $[s-d,s]$. Now

$$[v, c_d] = v(t+d)-v(s-d)+z_d$$

where

$$|z_d| \leq |TV(s,v)-TV(s-d,v)| + |TV(t,v)-TV(t+d,v)|$$

Since t and s are also points of continuity of $TV(t,v)$ (see [6]), $\lim_{d\to 0} |z_d| = 0$ and $\lim_{d\to 0} [v, c_d] = v(t)-v(s)$. Substituting into (3), $(\hat{c}, 0, 0) \in cl(Z)$ and $(\hat{c}, 0, c_d) \in Y$ and letting $d \to 0$ yield $v(t) \leq v(s)$.

Given any point t of discontinuity in v, since left and right limits exist, $v(t)$ can be set to its left limit without changing the value of the functional. Hence v is monotone non-decreasing everywhere.

(c) If (x,u) is optimal in $(\overline{P})$, then v_j is constant on every interval where $K_s(x(t),t)_j < 0$.

Substitute $a_1 = a_2 = \hat{c}$, $b_1 = b_2 = c_2 = 0$, and $c_1(t) = K_s(x(t),t)$ in (3). Then $[v, K_s(x)] \geq 0$ and (c) follows from (b).

(d) $r > 0$.

Suppose $r = 0$. Substituting $b_1 = b_2 = c_2 = 0$ and $c_1(t) = K_s(\overline{x}(t),t)$ in (3) yields $[v, K_s(\overline{x})] \geq 0$. Since $K_s(\overline{x}(t),t)_j < -\alpha < 0$, (b) implies that $v = 0$. Substituting $b_1 = -p$ and $b_2 = 0$ in (3) yields $-((p,p)) \geq 0$. Hence, $p = 0$ a.e.. This is impossible since r,p,v cannot all vanish so that $r > 0$ and (3) can be normalized with $r = 1$.

(e) $\overline{L}(p,v) = \hat{c}$

Substitute $a_1 = c(x,u)$, $b_1 = \dot{x}-Ax-Bu$, $c_1 = K_s(x)$, $a_2 = \hat{c}$, $b_2 = c_2 = 0$ in (3) and recall that $r = 1$ from (d). This yields $\overline{L}(p,v) \geq \hat{c}$. By weak duality, $\overline{L}(p,v) \leq \hat{c}$ so that $\overline{L}(p,v) = \hat{c}$.

(f) $p = q$ a.e. where $q \in BV$.

This proof is more technical than (a) to (e) and appears in Lemma 1A of the Appendix so the proof of Lemma 1 is complete.

■

Proof of Theorem 1: In the problem (P) with explicit control constraints, proceed exactly as in the proof of Lemma 1. A fourth component $d \in C$ is added to the sets Y and Z where $d \leq 0$ in Y and $d(t) \geq K_c(u(t),t)$ in Z. (Note that $d \in C$ and not $d \in B$--if d were chosen in B, then the Hahn-Banach Theorem would produce a multiplier in the

dual of B which is a miserable space. By choosing $d \in C$, the dual multiplier lies in BV.)

Continuing as in Lemma 1, the Hahn-Banach Theorem yields:

$$(4) \qquad c(x,u) + ((p,\dot{x}-Ax-Bu)) + [v,K_s(x)] + [w,K_c(u)] \geq \hat{c}$$

for all $x \in \mathcal{A}$ with $x(0) = x_0$ and $u \in C$ where $v,w \in BV$ and are both non-decreasing. Note that to prove p,v,w are optimal in (D), it must be shown that (4) holds for all $u \in B$ not just $u \in C$ and also that $w \in \mathcal{A}$ so that $[w,K_c(u)] = ((\dot{w},K_c(u)))$. Then exactly as in (e) above, weak duality implies that $L(p,\dot{w},v) = \hat{c}$.

First, it is proved that the infimum of the left side of (4) over $x \in \mathcal{A}$ and $u \in C$ equals $\hat{c}$. Let u^k be a minimizing sequence for (P) and let x^k be the corresponding trajectories. In Lemma 2A of the Appendix it is shown that given $e > 0$, there exists $y \in C$ with $K_c(y(t),t) \leq 0$, $|y(t)-u^k(t)| \leq e$ except on a set of measure less than e, and $||y|| \leq ||\bar{u}|| + ||u^k||$. Thus as $e \to 0$, $c(x^k,y) \to c(x^k,u^k)$, $((p,\dot{x}^k-Ax^k-By)) \to 0$, $[v,K_s(x^k)] \leq 0$, and $[w,K_c(y)] \leq 0$. Since $c(x^k,u^k) \to \hat{c}$, then the claim is verified.

In Lemma 3A of the Appendix it is shown that if the infimum of the left side of (4) over $u \in C$ equals $\hat{c}$, then

$w \in \mathcal{A}$ so that $[w, K_c(u)]$ equals $((w, K_c(u)))$. By Lusin's Theorem [8,p.53], given $u \in B$ and $e > 0$, there exists $y \in C$ with $y = u$ except for a set of measure less than e and $||y|| \leq ||u||$. Hence (4) holds for $u \in B$ and (p, w, v) are optimal in (D) as noted above. The complementary slackness conditions follow as in the proof of Lemma 1.

■

III. MINIMUM PRINCIPALS

In order to solve the dual problem numerically, the x and u that achieve the infimum in (1) must be characterized. This leads to a minimum principal and an adjoint condition. Theorem 2 below proves that the minimization over u in (1) can be taken under the integral sign.

Theorem 2

Suppose (A1) and (A2) hold, (p,w,v) are feasible in (D) with $L(p,w,v) > -\infty$, and $x^* \in \mathcal{A}$ and $u^* \in B$ achieve the minimum in (1) corresponding to (p,v,w). Then the control minimum principal holds:

$$(5) \qquad h(x^*(t),u^*(t),t)-(p(t),B(t)u^*(t))+(w(t),K_c(u^*(t),t)) \leq h(x^*(t),z,t)-(p(t),B(t)z)+(w(t),K_c(z,t))$$

for all $z \in R^m$ and a.e. t.

Similarly if $\overline{L}(p,v) > -\infty$, $f(\cdot,\cdot) = c(\cdot,\cdot)$, $U(t) = \{b \in R^m: K_c(b,t) \leq 0\}$, and $x^* \in \mathcal{A}$ and $u^* \in B$ achieve the minimum in (2) corresponding to (p,v), then

$$(6) \qquad h(x^*(t),u^*(t),t)-(p(t),B(t)u^*(t)) \leq h(x^*(t),z,t)-(p(t),B(t)z)$$

for all $z \in U(t)$ and a.e. t.

Proof: Only (5) will be proved since (6) is similar. Let $\bar{c} = L(p,w,v)$ where by definition

$$(7) \qquad L(p,w,v) = \inf \int_0^1 \{ h(x(t),u(t),t) + (p(t),\dot{x}(t)-A(t)x(t)-B(t)\dot{u}(t)) + (w(t),K_c(u(t),t)) \} dt + [v,K_s(x)]$$
$$\text{s.t. } x \in \mathcal{A},\ u \in B,\ x(0) = x_0$$

Let E denote the intersection of the Lebesgue points of each term in the integrand of (7) evaluated at (x^*,u^*). Suppose (5) is violated at $s \in E$ by $z \in R^m$. Let G denote a ball of diameter g centered at s, I(G,u) the integral in (7) evaluated at $x=x^*$ over the the ball G, and $J(u(\cdot))$ the integrand in (7) evaluated at $x=x^*$. Since s is a Lebesgue point of $J(u^*(\cdot))$, then $I(G,u^*)= J(u^*(s))g + o(g)$. Define v to be a control that agrees with u^* outside G and equals z inside G. Then s is easily seen to be a Lebesgue point of $J(v(\cdot))$ so that $I(G,v)=J(v(s))g+o(g)$ and $I(G,v) < I(G,u^*)$ for g sufficiently small. This violates the optimality of (x^*,u^*) in (7) so that the minimum principal holds on E. Since E has full measure, then the proof is complete.

□

Note that Theorem 2 holds for all (p,w,v) that are feasible in the dual problem while the standard necessary conditions only hold for some (p,w,v). Also observe that it is not possible to carry out the minimization over x under the integral sign in (1) due to the presence of the $\dot{x}$ term. The following lemma will be needed before the adjoint conditions can be derived:

Lemma 2

Suppose (A1) and (A2) hold, (p,w,v) are feasible in (D) with $L(p,w,v) > -\infty$, (x^*,u^*) achieve the minimum in (1) for (p,w,v), $K_s(\cdot,\cdot)$ is twice continuously differentiable, and $K_s(\cdot,\cdot)_x$ denotes the gradient of K_s with respect to x. Then $p(\cdot) - K_s(x^*(\cdot),\cdot)_x^T v(\cdot) \in \mathcal{A}$ and $p(1^-)-K_s(x^*(1),1)_x^T v(1^-) = 0$. If K_s is affine, then the existence of (x^*,u^*) is not required.

Proof: By the definition of L,

$$(8) \qquad L(p,w,v) \leq c(x,u)+((p,\dot{x}-Ax-Bu))+((w,K_c(u)))+[v,K_s(x)]$$

for all $x \in \mathcal{A}$ with $x(0) = x_0$ and $u \in B$. Each term on the right side of (8) is convex and furthermore the $[v,K_s(x)]$

term is differentiable in x. Standard necessary conditions (see [3]) then imply:

$$(9)\quad c(x^*,u^*)+((p,\dot{x}^*-Ax^*-Bu^*))+((w,K_c(u^*))) \le$$
$$c(x,u)+((p,\dot{x}-Ax-Bu))+[v,K_s(x^*)_x(x-x^*)]+((w,K_c(u)))$$

for all $x \in \mathcal{A}$ with $x(0) = x_0$ and $u \in B$. Note that equality holds in (9) for $x = x^*$ and $u = u^*$.

By Fubini's theorem,

$$(10)\quad \int_0^1 (p(t),\dot{x}(t))dt = (p(1^-),x(1))-(p(0^+),x_0)-\int_{0^+}^{1^-} x(t)^T dp$$

$$(11)\quad \int_{0^+}^{1^-} x(t)^T K_s(x^*(t),t)_x^T\, dv = \int_{0^+}^{1^-} x(t)^T d(K_s(x^*(t),t)_x^T v)$$
$$- \int_{0^+}^{1^-} v(t)^T K_s(x^*(t),t)_x x(t)dt.$$

Also if v is normalized so that $v(1) = 0$, then

$$(12)\quad \int_0^1 (x(t)-x^*(t))^T K_s(x^*(t),t)_x^T\, dv =$$
$$\int_{0^+}^{1^-} (x(t)-x^*(t))^T K_s(x^*(t),t)_x^T\, dv$$
$$-(x(1)-x^*(1))^T K_s(x^*(1),1)_x^T v(1^-).$$

Combining (9), (10), (11), and (12) yields:

$$(13) \qquad c(x,u)-((p,Ax+Bu))-((v,\dot{K}_c(x^*)_x x)) +$$

$$\int_{0^+}^{1^-} x(t)^T d(K_s(x^*)_x^T v-p) + ((w,K_c(u))) + (p(1^-) -$$

$$K_s(x(1),1))_x^T\, v(1^-),x(1)) \geq \bar{c} > -\infty$$

where $\bar{c} > -\infty$ is a constant depending only on $x^*, u^*, p, w,$ and v. Again equality holds in (13) for $x = x^*$ and $u = u^*$. If K_s is affine, then (13) holds without even assuming the existence of (x^*,u^*) and $\bar{c}$ only depends on $L(p,w,v)$.

Now, it is shown that $p(1^-) = K_s(x(1),1)_x^T\, v(1^-)$: Suppose inequality holds and g is the function shown in Figure 1. Inserting into (13) $x=\bar{x}+g(K_s(x(1),1)_x^T v(1^-)-p(1^-))$ where $\bar{x}(0) = x_0$ and $\bar{x} \in \mathcal{A}$ then yields a contradiction as $b,d \to 0$ due to the presence of the boundary term.

Finally, Lemma 4A in the Appendix can be applied to (13) to show that $K_s(x^*(\cdot),\cdot)_x^T v(\cdot)-p(\cdot) \in \mathcal{A}$. ∎

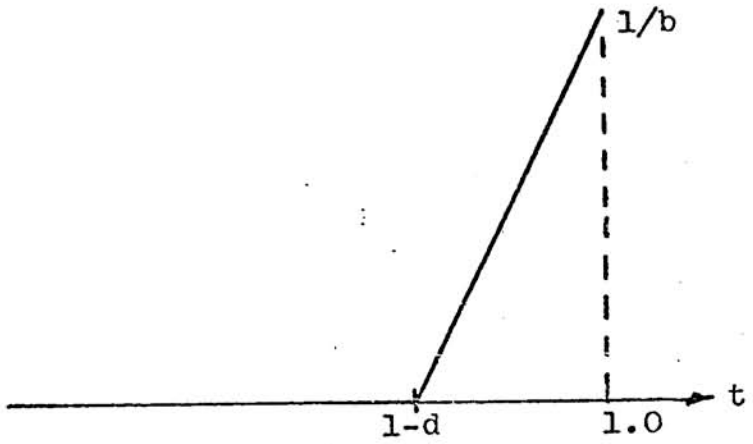

FIGURE 1: The Function g Which Depends on the Parameters b and d

Theorem 3

Suppose (A1) and (A2) hold, (p,w,v) are feasible in (D) with $L(p,w,v) > -\infty$, $x^* \in \mathcal{A}$ and $u^* \in B$ achieve the minimum in (1) corresponding to (p,w,v), and $K_s(\cdot,\cdot)$ is twice continuously differentiable. Then the state minimum principal holds:

$$(14)\quad h(x^*(t),u^*(t),t) \leq (\dot{q}(t)+A^T(t)q(t) - ((\dot{K}_s(x^*(t),t)_x^T + A^T(t)K_s(x^*(t),t)_x^T)v(t), z-x^*(t)) + h(z,u^*(t),t) \quad \text{a.e. } t \text{ and}$$

for all $z \in R^n$ where $q(t) = K_s(x^*(t),t)_x^T v(t) - p(t)$. If $h(\cdot,u,t)$ is differentiable, then the adjoint equation holds:

$$(15)\quad \dot{q}(t) = -A^T(t)q(t) - h(x^*(t),u^*(t),t)_x + (\dot{K}_s(x^*(t),t)_x^T + A^*(t)^T K_s(x^*(t),t)_x^T)v(t), \quad \text{a.e. } t$$

$$q(1) = 0.$$

Proof: In Lemma 2 it was observed that $q \in \mathcal{A}$ so that $[q,x] = ((\dot{q},x))$. From (13),

$$(16) \qquad \int_0^1 h(x(t),u^*(t),t)-(p(t),A(t)x(t)) - (v(t),\dot{K}_s(x^*(t),t)_x x(t)) + (\dot{q}(t),x(t))\, dt \geq c$$

for all $x \in \mathcal{A}$ with $x(0) = x_0$ where $c > -\infty$ is a constant depending only on $x^*,u^*,p,w,$ and v. As noted after (13), equality holds in (16) for $x = x^*$. Exactly as in Theorem 2, $x^*(t)$ must yield the pointwise minimum for the integral. There is one technical point though since in Theorem 2, u was contained in B while in (16) x lies in $\mathcal{A}$. However, if $z \in R^n$ yields a better minimum for the integrand of (16) at $t = s$, then Urysohn's Lemma can be used to construct a function in $\mathcal{A}$ that agrees with z near s and x^* away from s so that the optimality of x^* is again violated. The adjoint equation is obtained simply by setting the derivative of the right side of (14) to zero at $x = x^*$. ∎

The condition (15) above is the familar adjoint equation for state constrained problems given in [5] and [2]. These standard necessary condition conditions only assert that (15) holds for some $(p,w,v,)$ where (x^*,u^*) are optimal in (P)

while Theorem 3 holds for all (p,w,v) feasible in (D). The numerical solution of the dual problem using the Ritz method is analyzed in [1].

IV. APPENDIX

Lemma 1A

Suppose (A3) is satisfied, the optimal value of $(\overline{P})$, $\hat{c}$, is finite, and the functions $p \in L^{\infty}$ and $v \in BV$ satisfy the conditions: v is monotone non-decreasing and $\overline{L}(p,v) = \hat{c}$. Then $p = q$ a.e. where $q \in BV$.

Proof: For notational convenience p will be assumed scalar valued although the proof for vector valued functions is identical. Let R denote the set of Lebesque points of p and suppose that p has infinite variation on this set. Thus, given a constant b, there exists $t_0 < t_1 \ldots < t_N$ such that

$$(17) \qquad \sum_{\substack{N \geq j \geq 1 \\ j \text{ odd}}} |p(t_{j-1}) - p(t_j)| > b$$

and $p(t_{j+1}) < p(t_j) > p(t_{j-1})$ for j even. Let α, ρ, M be as given in (A3), let $c = \min(\rho, \alpha)$, and x_e be as pictured in Figure 2. As $e \to 0$, $\dot{x}_e \to -\sum_{j=0}^{N} (-1)^j d(t-t_j)$ where $d(\cdot)$ is the delta function. Since t_j are Lebesgue points of p and $p(t_{j+1}) < p(t_j) > p(t_{j-1})$ for j even, then

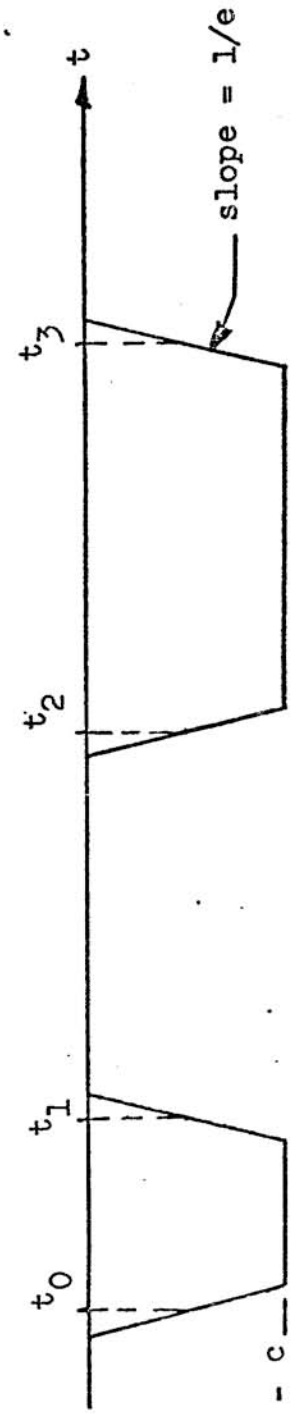

FIGURE 2: The Function x_e

$$\lim_{e\to 0} (\dot{(}p,\dot{x}_e)) = \sum_{\substack{N\geq j\geq 1 \\ j \text{ odd}}} p(t_j)-p(t_{j-1}) < -b.$$

From the definition of $\bar{L}$,

$$f(\bar{x}+x_e,\bar{u})+((p,\dot{x}_e+\dot{\bar{x}}-Ax_e-A\bar{x}-B\bar{u}))+[v,K_s(\bar{x}+x_e)] \geq \hat{c}.$$

Now $||x_e|| \leq \min(\alpha,\rho)$ so that (A3) implies $K_s(\bar{x}(t)+x_e(t),t) \leq 0$ and $f(\bar{x}+x_e,\bar{u}) \leq M$. Thus,

$$(18) \qquad -b > \lim_{e\to 0} ((p,\dot{x}_e)) \geq \hat{c} - M + ((p,\dot{\bar{x}}-A\bar{x}-Ax_e-B\bar{u})).$$

As $b \to \infty$ this yields a contradiction since the right side of (18) is finite so the variation of p on R is finite. Since R has full measure (see [8, p. 158]), then $t \in R^c$ implies the existence of a sequence $\{t_j\} \subset R$ such that $t_j \to t^+$. p has finite variation on R so that $\lim_{j\to\infty} p(t_j)$ exists. Define a new function q by

$$q(t) = \begin{cases} p(t) & \text{if } t \in R \\ \lim_{i\to\infty} p(t_i) & \text{if } t \notin R \text{ where } \{t_i\} \subset R \text{ and } t_i \to t^+ \end{cases}$$

Now $q(t) = p(t)$ a.e. and q has the same variation on $[0,1]$ as p has on R. ∎

Lemma 2A

Suppose K: $R^m \times [0,1] \to R^n$ and is continuous on $R^m \times [0,1]$, $K(\cdot,t)$ is convex for $t \in [0,1]$, $u \in B(R^m)$ and $K(u(t),t) \leq 0$ for all $t \in [0,1]$, and there exists $\bar{u} \in C(R^m)$ such that $K(\bar{u}(t),t)_j < a < 0$ for some "a" and for all $t \in [0,1]$ and $j = 1,\ldots,n$. Then given $e > 0$ there exists $v \in C(R^m)$ with $|u(t)-v(t)| < e$ except on a set of measure less than e, $K(v(t),t) \leq 0$ for all $t \in [0,1]$, and $||v|| \leq ||\bar{u}|| + ||u||$.

Proof: Let $w = b\bar{u} + (1-b)u$ where $b > 0$ is small enough that $||u-w||_C \leq e$. By the convexity of $K(\cdot,t)$, $K(w(t),t)_j \leq ba < 0$ and by Lusin's Theorem [8, p. 53] there exists $y \in C$ with $y = w$ except on an open set E^c of measure less than e and $||y|| \leq ||w||$. Since $K(y(\cdot),\cdot)$ is uniformly continuous on $[0,1]$, there exists a constant $d > 0$ such that if $|t-s| < d$, then $|K(y(t),t)-K(y(s),s)| < ba$. Outer regularity of the Lebesgue measure implies the existence of an open set D containing E with measure of $D-E$ less than d. Since $K(y(t),t)_j = K(w(t),t)_j \leq ba < 0$ on E and

any point in D is at most d away from a point in E, then $K(y(t),t) \leq 0$ for $t \in D$. From Urysohn's Lemma, there exists $g \in C(R)$ with $E \prec g \prec D$. Define $v = gy + (1-g)\bar{u}$. For $t \in D$, v is a convex combination of two functions that satisfy $K(z,t) \leq 0$ so the convexity of K implies that $K(v(t),t) \leq 0$ on D. On the other hand, $v = \bar{u}$ on D^c so $K(v(t),t) \leq 0$ for $t \in [0,1]$. By construction $v(t) = y(t) = w(t)$ for $t \in E$ so that $|u(t)-v(t)| < e$ except on a set of measure less than e. Similarly the bound on $||v||$ in the lemma statement is easily verified. ∎

Let $g(x,u)$ denote the first three terms in (4) and let $K(\cdot,\cdot) = K_c(\cdot,\cdot)$.

Lemma 3A

Suppose the optimal value $\hat{c}$ to the problem

$$\text{(19)} \qquad \inf\ g(x,u) + [w,K(u)]$$
$$\text{s.t. } x \in \mathcal{A},\ u \in C$$
$$x(0) = x_0$$

is finite and there exists (x^k,u^k) feasible in (19) such that $g(x^k,u^k) \to \hat{c}$ and $K_c(u^k(t),t) \leq 0$ for all $t \in [0,1]$. Assume

(A2) holds and $w \in BV$ is non-decreasing; then $w \in \mathcal{A}$.

Proof: To keep the notation simple, K is assumed to have range in R^1. The main idea of the proof is the following: In [8, p. 166] it is proved that $w = r + s$ where $r \in \mathcal{A}$, $s \in NBV$, $\dot{s} = 0$ a.e., and s is non-decreasing. If $s \neq 0$, then $TV(s) = s(1) \neq 0$. Choose k sufficiently large that $g(x^k,u^k) - \hat{c} < |a|s(1)/2$ where a was given in (A2) and define $(x,u) = (x^k,u^k)$. Since $\dot{s} = 0$ a.e., a sequence of closed sets F^k with measure converging to zero can be chosen with almost all of the variation of s occuring on F^k. Then a function $v^k \in C$ is constructed that agrees with $\bar{u}$, the interior control, on F^k and u just outside F^k. Thus, $\int_{F^k} K(v^k(t),t)ds \to < a\, s(1)$, $\int_{F^k} K(v^k(t),t)\, dr \to 0$ since the measure of F^k converges to zero while v^k is uniformly bounded, $\int_{(F^k)^c} K(v^k(t),t)\, dw \leq 0$ since $K(v^k(t),t) \leq 0$ and w is non-decreasing, and $g(x,v^k) \to g(x,u) < \hat{c} + |a|s(1)/2$. Combining these results, $g(x,v^k) + [w,K(v^k)] < \hat{c} + as(1)/2$ for k sufficiently large which contradicts the optimality of $\hat{c}$ since $a < 0$. Thus $s = 0$ and $w = r \in \mathcal{A}$.

Begin the construction by choosing a closed set $E \subset [0,1]$ with $\dot{s} = 0$ on E and the measure of E^c less than e. Around each $q \in E$ construct an open ball B of radius $b > 0$ small enough that

$$(20) \qquad |s(t)-s(T)| \leq e|t-T|$$

for t,T in a ball of radius $2b$ about q. Since $\dot{s}$ vanishes on E, this construction is possible, and since E is compact, a finite subcover of balls B_j of radius b_j can be selected. Let D_j be an open ball containing B_j and of radius $b_j + d$ where $d < b_j$. By Urysohn's Lemma, there exists $z \in C$ with $D^c \prec z \prec \overline{B}^c$ where $D = \bigcup_j D_j$ and $B = \bigcup_j B_j$. Define $v = (i-z)u + z\bar{u}$. By (20), the variation of s on D is bounded by e since s is monotone. Hence, the variation of s on D^c is at least $s(1) - e$. Also the measure of D^c and B^c is less than e and $v = u$ on B. By choosing a sequence of e's converging to zero, then a sequence of v's is obtained satisfying all the properties stated above. ∎

Let $g(x)$ denote the first three terms in (13) evaluated at $u = u^*$.

Lemma 4A

Suppose $q \in NBV$ and

$$(21) \qquad \inf g(x) + [q,x] > -\infty$$
$$\text{s.t. } x \in \mathcal{A}, \quad x(0) = x_0.$$

Then $q \in \mathcal{A}$.

Proof: Again to keep notation simple, assume q is scalar valued. The proof uses the same construction employed in Lemma 3A. Assume q is continuous from the left and express $q = r + s$ where $r \in \mathcal{A}$, $s \in NBV$, and $\dot{s} = 0$ a.e. Suppose $s(t) > 0$ for some $t \in [0,1]$. Using the construction of Lemma 3A on the interval $[0,t]$ generates sets D and B such that the variation of s on D is at most e. Again construct z satisfying $(DU(t,1])^c \prec z \prec (\overline{B}U[t,1])^c$ and define $x_N = (1-z)\overline{x} - Nz$ where $\overline{x}$ is feasible in (21) and $N \in R$. As $e \to 0$, then $[q,x_N] \to -Ns(t)$ and $g(x_N) \to g(x)$. Now let $N \to \infty$ and (21) is contradicted. If $s(t) < 0$, then let $N \to -\infty$ and again (21) is contradicted. Thus $s=0$ and $q = r \in \mathcal{A}$.

The following two theorems prove the existence of an optimal control for problem (P). First the control problem is stated in a slightly more general form:

$$(22) \qquad \inf \int_0^1 L(x(t),u(t),t)\,dt + h(x(0),x(1))$$

$$\text{s.t. } \dot{x}(t) = A(t)x(t)+B(t)u(t) \quad \text{for all} \quad t \in [0,1]$$

$$(u(t),x(t)) \in K(t) \quad \text{a.e.} \quad t \in [0,1]$$

$$x(t) \in G(t) \quad \text{for all} \quad t \in [0,1]$$

$$u \in L^1(R^m), \quad x \in \mathcal{A}(R^n)$$

where $L: R^n \times R^m \times [0,1] \to R$, $h: R^{2n} \to R$, and $(A(t),B(t))$ are matrices of the appropriate sizes. Rockafellar [7] proves an existence theorem for this problem when certain recession functions satisfy a boundedness condition, so while the results that follow are not new, the method of proof is of interest. The proof is direct in that only results of basic real analysis are utilized and machinery from convex analysis is not required; the approach is similar to that in Lee and Markus [9, p. 259-307], however, they develop two distinct theories for problems with compact constraints and problems satisfying a growth assumption on the cost functional with no constraints. The approach below is more unified and also treats problems with non-compact constraints. Theorem 1A shows that if a bounded minimizing sequence exists, then an optimal solution exists while Theorem 2A provides a number of conditions guaranteeing the existence of a bounded minimizing sequence.

Theorem 1A

Assume the following:

(A4) $L(\cdot,\cdot,t)$ and $h(\cdot,\cdot)$ are convex and both $h(\cdot,\cdot)$ and $L(\cdot,\cdot,\cdot)$ are continuous. The components of $A(\cdot)$ lie in L^1 and the components of $B(\cdot)$ lie in L^∞. $G(t)$ and $K(t)$ are convex and closed for all $t \in [0,1]$.

If there exists a feasible minimizing sequence (x^k,u^k) such that $(x(0)^k,u^k)$ is bounded in $E^1 \times L^p$ for some $p > 1$, then there exists an optimal control u^* and corresponding trajectory x^*.

Proof: Let $x[u,y]$ denote the solution to

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \qquad x(0) = y. \tag{23}$$

Define

$$S = \{(u,y): u \in L^p, y \in R^n \text{ and } (x[u,y],u) \text{ satisfy the constraints}\}. \tag{24}$$

Step 1: S is closed, convex, and hence weakly closed.

Since the differential equation is linear, $x[u,y]$ is a linear function of (u,y). By the convexity of $K(t)$ and $G(t)$, S is easily proved convex.

Now consider the closedness property. First, it is proved that the transition matrix, $F(t,s)$, corresponding to the system dynamics is uniformly bounded for $t,s \in [0,1]$. If x satisfies the equation $\dot{x} = A(t)x(t)$, then

$$(25) \qquad |x(t)| = \left|\int_0^t A(s)x(s)\,ds + x(0)\right|$$

$$\leq \int_0^t \left(\sum_{i,j} |A_{ij}(s)|\right)|x(s)|\,ds + |x(0)|$$

$$\leq e^a|x(0)|$$

where $a = \int_0^1 \left(\sum_{i,j} |A_{ij}(s)|\right) ds$. Since $A_{ij}(\cdot)$ lie in L^1, "a" is finite and there exists $c < \infty$ such that $|F(t,s)|<c$ for all $(t,s) \in [0,1]$.

Suppose $u^k \to u$ in L^p and $y^k \to y$ in E^1. Define $x^k = x[u^k,y^k]$ and $x = x[u,y]$. Then

(26) $|x(t)-x^k(t)| =$

$$|F(t,0)(y-y^k) + \int_0^t F(t,s)B(s)(u(s)-u^k(s))ds|$$

$$\leq c|y-y^k| + c\Big(\sum_{i,j} ||B_{ij}||_{L^\infty}\Big)||u-u^k||_{L^1}.$$

Thus, $\lim_{k\to 0} ||x-x^k|| = 0$. Combining this with the fact that $x^k(t) \in G(t)$ for all $t \in [0,1]$ and $G(t)$ is closed yields $x(t) \in G(t)$ for all $t \in [0,1]$. Similarly $||u-u^k||_{L^1} \to 0$ implies that $u^k(t) \to u(t)$ a.e. $t \in [0,1]$ and since $(x^k(t),u^k(t)) \in K(t)$ a.e. t, the closedness of $K(t)$ implies that $(x(t),u(t)) \in K(t)$ a.e. $t \in [0,1]$. Thus $[x,u]$ satisfy the constraints and hence S is closed.

Step 2: Existence

Let $[u^k,y^k]$ denote the bounded minimizing sequence in $L^p \times E^1$. A weakly convergent subsequence also subscripted by k can be extracted converging to $[u,y]$ and since S is weakly closed, then $(u,y) \in S$. Let $x^k = x[u^k,y^k]$ and $x = x[u,y]$. It remains to be shown that $[x,u]$ yields optimal cost.

Lee and Markus [9, p. 259] prove that if $L(\cdot,\cdot,t)$ is convex and $L(\cdot,\cdot,\cdot)$ is continuous, then $\int_0^1 L(x(t),u(t),t)dt$ is weakly lower-semi-continuous in L^1. Since $h(\cdot,\cdot)$ is

continuous, the proof will be completed if it can be shown that $x^k \to x$ weakly in L^1 and $x^k(0) \to x(0)$, $x^k(1) \to x(1)$. In fact, it is proved that $x^k(t) \to x(t)$ for all $t \in [0,1]$ and $||x-x^k||_{L^1} \to 0$.

Define $G: R \times R \to R$ by:

$$G(t,s) = \begin{cases} 1 & \text{for } s \leq t \\ 0 & \text{for } s > t \end{cases}$$

Then if F is the transition matrix again,

$$x^k(t)-x(t) = F(t,0)(y^k-y) + \int_0^1 G(t,s)F(t,s)B(s)$$
$$(u^k(s)-u(s))\, ds.$$

Since $[G(t,\cdot)\, F(t,\cdot)B(\cdot)]_{ij} \in L^\infty$, and L^∞ is contained in the dual of L^p, the right hand side above approaches zero. Thus $\lim_{k\to 0} |x^k(t)-x(t)| = 0$ for all $t \in [0,1]$. Since u^k is uniformly bounded in L^p and y^k is uniformly bounded in E^1, then the same inequality used in (26) implies that $||x_k||$ and $||x||$ are uniformly bounded say by the constant c. From Egoroff's Theorem, given $e > 0$, there is a measureable set $E \subset [0,1]$ such that the measure of E^c is less

than e and x^k converge uniformly to x on E. Thus

$$\lim_{k\to\infty} \int_0^1 |x^k(t)-x(t)|dt = \lim_{k\to\infty} \int_E |x^k(t)-x(t)|dt + \int_{E^c} |x^k(t)-x(t)|dt$$

$$\leq 2ce$$

Now let $e \to 0$ and the claim that $||x-x^k||_{L^1} \to 0$ has been proved. ∎

Define $U(t) = \{u: (x,u)\in K(t) \text{ for some } x\in G(t)\}$ and $X(t) = \{x: x\in G(t), (x,u)\in K(t) \text{ for some } u\}$.

Theorem 2A

Suppose that there exists a feasible control $\bar{u}$ and corresponding trajectory $\bar{x}$ such that

$$\int_0^1 L(\bar{x}(t),\bar{u}(t),t)dt + h(\bar{x}(0),\bar{x}(1)) = \bar{L} < \infty.$$

If (A4) and any of the following is satisfied, then any minimizing sequence that satisfies the constraint $(x(t),u(t)) \in K(t)$ for all t is bounded:

(A5) There exists bounded sets C_1 and C_2 such that $U(t) \subset C_1$ for <u>all</u> $t \in [0,1]$ and for <u>some</u> $t \in [0,1]$ $X(t) \subset C_2$.

(A6) $L(x,u,t) \geq c_1|u|^p + g(x)$ for some $p>1$ and $c_1 > 0$ where $g(x) \to \infty$ as $|x| \to \infty$ and both $g(\cdot)$ and $h(\cdot,\cdot)$ are bounded from below.

(A7) There exists a bounded set C such that $X(t) \subset C$ for <u>some</u> $t \in [0,1]$ and $L(x,u,t) \geq c_1|u|^p + c_2$ for all (x,u,t) and for some $p > 1$ and $c_1 > 0$. Also $h(\cdot,\cdot)$ is bounded from below.

Proof: Let S be defined as in the proof of Theorem 1A and let (u_k, y_k) be a minimizing sequence. Suppose (A5) holds; then $||u^k||_{L^p}$ is bounded immediately. Let $z^k = x[u_k, y_k](r)$ where at time r the state lies inside of C_2 and let c be a uniform bound on $|F(t,s)|$ where $F(t,s)$ is the transition matrix corresponding to the system dynamics; then

$$(27) \quad |y^k| = |x[u^k,y^k](0)| = |F(0,r)z^k + \int_r^0 F(0,s)B(s)u^k(s)ds|$$

$$\leq c|z^k| + c\Big(\sum_{i,j} ||B_{ij}||\Big)||u^k||_{L^1}.$$

Since $|z^k|$ and $||u^k||_{L^1}$ are uniformly bounded, then y^k is uniformly bounded.

Suppose (A6) holds. Given $e > 0$, then for k sufficiently large

$$e + \bar{L} \geq \int_0^1 L(x^k(t),u^k(t),t)dt + h(x^k(0),x^k(1)) \tag{28}$$

$$\geq c_1||u^k||^p_{L^p} + c + \int_0^1 g(x^k(t))dt \tag{29}$$

$$\geq c_1||u^k||^p_{L^p} + c$$

where $x^k = x[u^k,y^k]$ and c includes the lower bounds for g and h. Thus $||u^k||_{L^p}$ is uniformly bounded. Also from (29) there exists a time t_k such that

$$g(x^k(t_k)) \leq e + \bar{L} - c_1||u^k||^p_{L^p} - c.$$

Since $g(x) \to \infty$ as $|x| \to \infty$, then $|x^k(t_k)|$ is uniformly bounded in k. Then exactly as in Equation (27), y^k is uniformly bounded.

Suppose (A7) holds. Then as in (28), $||u^k||_{L^p}$ is uniformly bounded and as in (27), $|y^k|$ is uniformly bounded.

Notice that Theorem 1A combined with Theorem 2A only proves the existence of an optimal solution to the control problem in an L^p space; on the other hand, the complementary slackness conditions in Theorem 1 requires that the optimal control be bounded and measureable. If (A5) holds, then the optimal control is trivially bounded. When (A6) or (A7) hold, then the optimal control can also be proved bounded by the minimum principal as follows: Note that Lemma 1 can be proved with the controls in L^p without making any changes in the proof and furthermore, the minimum principal (6) also holds. Since $p \in BV$, then a bound on the components of B in (6) implies that u* is bounded when (A6) or (A7) holds.

REFERENCES

1. W. W. Hager, "The Ritz-Trefftz Method for Optimal Control Problems with State and Control Constraints," to appear.

2. M. R. Hestenes, Calculus of Variations and Optimal Control Theory, John Wiley and Sons, New York, 1966.

3. J. L. Lions, Contrôle Optimal de Systèms gouvernés par des équations aux dérivées partielles, Dunod, Paris, 1968.

4. D. G. Luenberger, Optimization by Vector Space Methods, John Wiley and Sons, New York, 1969.

5. L. W. Neustadt, "A General Theory of Extremals," J. of Computer and Systems Sciences, 3(1969), p. 57-59.

6. I. P. Natanson, The Theory of Functions of a Real Variable, Frederick Ungar, New York, 1964.

7. R. T. Rockafellar, "State Constraints in Convex Control Problems of Bolza," SIAM J. Control, 10(1972), p. 691-715.

8. W. Rudin, Real and Complex Analysis, McGraw Hill, New York, 1966.

9. E. B. Lee and L. Markus, Foundations of Optimal Control Theory, John Wiley and Sons, New York, 1967.

CHAPTER 2

The Ritz-Trefftz Method for State and Control Constrained Optimal Control Problems

I. Introduction

The convergence properties of the Ritz-Trefftz method are examined in state and control constrained optimal control problems. Recall that the Lagrange dual to the problem

$$\begin{aligned} &\inf\ f(x) \\ &\text{s.t. } g(x) \leq 0 \\ &\qquad h(x) = 0 \end{aligned}$$

is given by

$$\begin{aligned} &\sup\ \mathcal{L}(p,\lambda) \\ &\text{s.t. } \lambda \geq 0 \\ &\qquad \mathcal{L}(p,\lambda) = \inf_x\ f(x)+\langle\lambda,g(x)\rangle+\langle p,h(x)\rangle \end{aligned}$$

where λ and p are linear functionals in the appropriate dual spaces. In the Ritz-Trefftz method, the dual problem is discretized by requiring the dual variables to lie in a finite dimensional subspace of the entire space.

Bosarge, et.al., [2] analyzed unconstrained control problems and derived rates of convergence for the solution of the finite dimensional problem to the solution of the

continuous problem. This paper analyzes the Ritz-Trefftz method from a fundamentally different approach which permits state and control constraints. The theory is illustrated using the linear quadratic control problem with linear inequality control and state constraints.

Section II formulates the dual problem. An important distinction between the primal and dual problems is that although the primal cost functional may be strictly convex, the dual cost functional is in general only convex. Hence, the question of existence of a solution to the finite dimensional dual problem is not obvious as in the primal problem. Section III proves that if the Slater condition holds, then the Ritz-Trefftz problem will have a solution.

In Sections IV and V, the rates of convergence of the solution to the finite dimensional problem to the solution of the continuous problem are derived. The finite dimensional dual variables are proved to converge to the continuous dual solution in a seminorm defined by the quadratic part of the dual functional. This result then leads to an estimate for the rate of convergence of the finite dimensional primal solution to the continuous solution. For non-linear problems, the dual variables converge in a intrinsic seminorm defined by the second derivative of the dual function evaluated at the optimal solution.

It is also noted in Section V that for problems without state constraints, the dual functional is positive definite

in a subspace of the total space.

Finally, Section VI estimates the error in the times when constraints become binding. Appendix 1 contains a result on interpolation in polynomial spaces of a function $f \geq 0$ by a function f^h satisfying $f \geq f^h \geq 0$ and the interpolation of a monotone function with a monotone polynomial from above. This result is required in the convergence estimates. Appendix 2 proves an estimate on the sensitivity of the solution to the dual and primal quadratic programing problem to changes in the data. This lemma is used to prove a theorem on the regularity of the solution to control constrained problems.

Nomenclature

Define the following norms, inner products, and linear functionals:

$$(u,v) = \sum_{k=1}^{\ell} u_k v_k \quad \text{where} \quad u = (u_1,\ldots,u_\ell), \quad v = (v_1,\ldots,v_\ell)$$

$$|v|^2 = (v,v)$$

$$||f||_{L^p} = \{ \int_0^1 |f(t)|^p \, dt \}^{1/p}$$

$$||f||_{H^p} = \{ \sum_{k=0}^{p} || \frac{d^k f}{dt^k} ||^2_{L^2} \}^{\frac{1}{2}}$$

$$||f||_C = \sup_{0 \le t \le 1} |f(t)|$$

$$||f||_{C^p} = ||f||_{C^{p-1}} + ||f^{(p)}||_C$$

$||f||_{BV}$ = total variation of f on $[0,1]$

$$((f,g)) = \int_0^1 (f(t),g(t))dt$$

$$\langle \nu, f \rangle_C = \int_0^1 f(t)d\nu(t).$$

Let L^p, H^p, and BV denote the spaces of real valued functions on $[0,1]$ for which $||\cdot||_{L^p}$, $||\cdot||_{H^p}$, and $||\cdot||_{BV}$ are finite. Let $\mathcal{A}$ denote the space of absolutely continuous functions and C^p be the set of functions with p continuous derivatives on $[0,1]$. Finally, PC^p represents the space of functions which have p continuous derivatives everywhere except for a finite number of points where the p-th derivative has a simple jump discontinuity.

If v is a vector, then the k-th component of v is denoted v_k or $(v)_k$. If f is a vector valued function and $\mathbb{b}$ is any of the spaces above, then the notation $f \in \mathbb{b}$ means that each component of f lies in $\mathbb{b}$.

A sequence of spaces S^h parameterized by h is of degree p and continuity q if for all $v \in C^p$, there

exists $v^h \in S^h$ such that

$$(1) \qquad ||v_k - v_k^h||_{C^r} \leq ch^{p-r}||v||_{C^p}$$

for all r satisfying $0 \leq r \leq q$ and for all components of v where c is a numeric constant not a function of h. The parameter h is usually the grid interval and v^h is the interpolate of v. Many examples of spaces S^h and estimates of the form (1) above can be found in [1] and [8].

For some sets S^h in which the grid points are free parameters, the estimate (1) holds for a larger space, PC^p. For example, suppose S^h consists of polynomials of degree $p-1$ on grid intervals of width h with derivatives of order $q-1$ continuous across the grid points. For $q-1 < p/2$, S^h is of degree p and continuity q. However, if the grid points in S^h are free parameters, then given any $v \in PC^p$, there exists $v^h \in S^h$ such that (1) holds. Simply place grid points at the points of discontinuity in the p-th derivative of v and use the standard interpolate. Furthermore, if $v \in PC^q$ and $v \in PC^p$ between the points of discontinuity in the q-th derivative, then v^h constructed as above satisfies (1) with $||v||_{C^p}$

replaced by the C^p norm of v over the <u>interior</u> of the grid intervals.

The constant c is used throughout to designate a generic constant.

II. FORMULATION OF THE PROBLEMS

The following problem is analyzed in this paper:

$$\min\ c(x,u)$$

$$\text{s.t. } c(x,u) = \int_0^1 \tfrac{1}{2}\, x(t)^T Q x(t) + \tfrac{1}{2}\, u(t)^T R u(t)\, dt$$

$$\dot{x}(t) = Ax(t) + Bu(t)$$

$$x(0) = x_o$$

$$K_s x(t) + b_s \leq 0$$

$$K_c u(t) + b_c \leq 0$$

$$u \in L^2, \quad x \in \mathcal{A}$$

where $x\colon [0,1] \to R^n$, $u\colon [0,1] \to R^m$ and the matrices are of the appropriate dimensions. The analysis is not significantly changed if the matrices are time varying and piecewise continuous and $c(x,u)$ is any non-linear, twice differentiable, strictly convex functional. Terminal constraints and terminal costs do not alter the analysis and were omitted to keep the notation less cumbersome. The matrices Q and R are assumed positive definite and the solution set for the inequality constraints is non-empty.

Define the dual function

$$(2) \qquad \mathcal{L}(p,\lambda,\nu) = \inf\ c(x,u) + ((p,\dot{x}-Ax-Bu)) + ((\lambda,K_c u+b_c)) + \langle \nu, K_s x+b_s \rangle_C$$

$$\text{s.t. } u \text{ bounded, measureable}$$

$$x \in \mathcal{A},\ x(0) = x_o.$$

The dual problem corresponding to (P) is:

$$(D) \qquad \sup\ \mathcal{L}(p,\lambda,\nu)$$

$$\text{s.t. } \nu(1) = 0,\ \nu \in BV,\ \nu \text{ non-decreasing}$$

$$\lambda \geq 0,\ \lambda \in L^1$$

$$p \in BV.$$

Make the following assumption:

(A1) There exists a control $\bar{u} \in C^0$ and a corresponding trajectory $\bar{x} \in C^1$ such that $(K_s\bar{x}(t) + b_s)_j < a < 0$ and $(K_c\bar{u}(t) + b_c)_j < a < 0$ for all constraints.

Now define $q(\cdot) = K_s^T \nu(\cdot) - p(\cdot)$. Then in [4] it is shown that (A1) implies the existence of (x^*,u^*) solving (P) and (p^*,λ^*,ν^*) solving (D) and the following hold:

(a.) (x^*,u^*) achieve the infimum in (2) for (p^*,λ^*,ν^*);

(b.) $\mathcal{L}(p,\lambda,\nu) \leq c(x,u)$ whenever (x,u) are feasible in (P) and (p,λ,ν) are feasible in (D) and $\mathcal{L}(p^*,\lambda^*,\nu^*)=c(x^*,u^*)$;

(c.) $(\lambda^*(t), K_c u^*(t)+b_c) = 0$ for all $t \in [0,1]$;

(d.) $\nu_j(t)$ is constant on intervals where $(K_s x^*(t)-b_s)_j<0$

(e.) unless $q \in \mathcal{A}$ and $q(1) = 0$, then $\mathcal{L}(p,\lambda,\nu)=-\infty$; furthermore if $\mathcal{L}(p,\lambda,\nu) > -\infty$ and (x,u) achieve the infimum in (2), then

$$(3) \quad \dot{q}(t) = -A^T q(t) - Qx(t) + A^T K_s^T \nu(t)$$
$$q(1) = 0$$

$$(4) \quad Ru(t) + B^T q(t) - B^T K_s^T \nu(t) + K_c^T \lambda(t) = 0.$$

In light of the relations above, it will be more convenient to express the dual function in terms of (q,λ,ν) rather than (p,λ,ν). Let $\mathcal{L}(q,\lambda,\nu)$ denote the new dual function. Hence the dual problem reduces to:

$$(D) \qquad \sup \ \mathcal{L}(q,\lambda,\nu)$$

$$\text{s. t. } q(1) = 0,\ q \in \mathcal{A}$$

$$\nu(1) = 0,\ \nu \in BV,\ \nu \text{ non-decreasing}$$

$$\lambda \in L^1,\ \lambda \geq 0$$

After an integration by parts, $\mathcal{L}$ becomes:

$$(5) \quad \mathcal{L}(q,\lambda,\nu) = \inf_{x,u} \ c(x,u) + ((\dot{q},x)) + ((q-K_s^T\nu, Ax+Bu)) + ((\lambda, K_c u+b_c)) + (q(0)-K_s^T\nu(0), x_o) - (\nu(0), b_s).$$

Define $\ell_1(\cdot)$ and $\ell_2(\cdot)$ by :

$$(6) \qquad \ell_1(q,\lambda,\nu) = A^T K_s^T \nu - A^T q - \dot{q}$$

$$\ell_2(q,\lambda,\nu) = B^T K_s^T \nu - B^T q - K_c^T \lambda.$$

From (3) and (4), the x and u achieving the minimum in (2) satisfy:

$$(8) \qquad x = Q^{-1}\ell_1(q,\lambda,\nu)$$

$$(9) \qquad u = R^{-1}\ell_2(q,\lambda,\nu)$$

Hence,

$$(10)\qquad \mathcal{L}(q,\lambda,\nu) = -\tfrac{1}{2}((\ell_1(q,\lambda,\nu), Q^{-1}\ell_1(q,\lambda,\nu))) - \tfrac{1}{2}((\ell_2(q,\lambda,\nu), R^{-1}\ell_2(q,\lambda,\nu))) + ((\lambda, b_c)) + (q(0)-K_s^{T}\nu(0), x_o) - (\nu(0), b_s).$$

III. EXISTENCE OF SOLUTION

In the Ritz-Trefftz method, the constraint $(q, \lambda, \nu) \in S^h$ is added to the above constraints on the dual problem, (D). Note that for the primal problem, the cost functional is strictly convex and hence the existence of a solution to the finite dimensional problem is trivial. However, for the dual problem, the quadratic part of the cost functional is only semi-definite. This section examines the existence of a solution to the finite dimensional problem:

$$
\begin{aligned}
(\hat{D}) \qquad & \sup \mathcal{L}(q, \lambda, \nu) \\
& \text{s.t. } q(1) = 0,\ q \in \mathcal{A} \\
& \qquad \nu(1) = 0,\ \nu \in BV,\ \nu \text{ non-decreasing} \\
& \qquad \lambda \in L^1,\ \lambda \geq 0 \\
& \qquad (q, \lambda, \nu) \in S^h .
\end{aligned}
$$

Throughout this section h is assumed fixed so that the superscript in S^h will be omitted.

Suppose that $S = S^q \times S^\lambda \times S^\nu$ is spanned by the functions $\{\phi^k\}_{k=1,\ldots N}$ where $[\phi^k]^T = [\phi^{k,q}{}^T, \phi^{k,\lambda}{}^T, \phi^{k,\nu}{}^T]$. Define the following partitioning of the basis functions:

$$\phi = [\phi^1|\phi^2|\ldots|\phi^N]$$

$$\begin{bmatrix} \phi^q \\ \phi^\lambda \\ \phi^\nu \end{bmatrix} = \begin{bmatrix} \phi \\ \cdot \end{bmatrix}$$

Finally, define $\hat{\mathcal{L}}(\delta) = \mathcal{L}(\phi^q\delta, \phi^\lambda\delta, \phi^\nu\delta)$ for $\delta \in R^N$, F to be the set of feasible dual variables in (D), and $\hat{F} = \{\delta \in R^N: \phi\delta \in F\}$.

The existence theorem will require the following assumption:

(A2) S^q, S^λ, S^ν lie in PC^1, PC^0, and PC^0 respectively, $S \cap F$ is non-empty, and if δ satisfies $\phi^\nu(1)\delta = 0$, then

$$(11) \qquad ||\phi^\nu\delta||_{BV} + ||\phi^\lambda\delta||_{L^1} + ||\phi^q\delta||_{L^2} \to \infty$$

as $|\delta| \to \infty$.

The relation (11) above just requires that the basis functions are linearly independent.

Theorem 1: If (A1) and (A2) hold, then there exists an optimal solution to $(\hat{D})$.

Proof: For notational convenience, x and u will be assumed scalar valued functions although the method of proof is identical for vector valued functions. Let δ^k be a maximizing sequence for $\hat{\mathcal{L}}$ and choose $\bar{c}$ such that $\hat{\mathcal{L}}(\delta^k) \geq \bar{c}$ for $k \geq \bar{k}$. Let $\lambda^k = \phi^\lambda \delta^k$, $\nu^k = \phi^\nu \delta^k$, $q^k = \phi^q \delta^k$, and $p^k = K_s^T \nu^k - q^k$. Since $\hat{\mathcal{L}}(\delta)$ is the infimum of the Lagrangian over (x,u), then for $k \geq \bar{k}$

$$(12)\quad \bar{c} \leq \hat{\mathcal{L}}(\delta^k) \leq c(\bar{x},\bar{u}) + ((\lambda^k, K_c\bar{u}+b_c)) + \langle \nu^k, K_s\bar{x}+b_s\rangle_C$$
$$\leq c(\bar{x},\bar{u}) + a||\lambda^k||_{L^1} + a||\nu^k||_{BV}$$

where $(\bar{x},\bar{u})$ and $a < 0$ were given in (A1). This gives a bound for $||\lambda^k||_{L^1}$ and $||\nu^k||_{BV}$.

Let x^k satisfy the equation

$$(13)\quad \dot{x}(t) - Ax(t) - B\bar{u}(t) = \epsilon\, p^k(t)/||p^k||_{L^2}, \qquad x(0) = x_o$$

where $\epsilon < 0$ is small enough that $K_s x^k(t) + b_s \leq 0$ and $c(x^k,\bar{u}) < c(\bar{x},\bar{u}) + 1$. Then

$$(14)\quad \bar{c} \leq \hat{\mathcal{L}}(\delta^k) \leq c(x^k,\bar{u}) + ((\lambda^k, K_c\bar{u}+b_c)) + \langle \nu^k, K_s x^k + b_s\rangle_C$$
$$+ ((\dot{x}^k - Ax^k - B\bar{u}, p^k))$$
$$< c(\bar{x},\bar{u}) + 1 + \epsilon||p^k||_{L^2}.$$

Thus, $||p^k||_{L^2}$ is bounded and hence $||q^k||_{L^2} = ||K_s^T \nu^k - p^k||_{L^2}$ is bounded. By (A2) it follows that δ^k is uniformly bounded and hence there exists a subsequence converging to δ^*.

$\hat{F}$ is easily seen to be closed so that $\delta^* \in \hat{F}$. By the continuity requirement in (A2), $||\dot{\phi}^{q,k}||_C$, $||\phi^{\lambda,k}||_C$, and $||\phi^{\nu,k}||_C$ are bounded and hence $\phi\delta^k \to \phi\delta^*$ in $H^1 \times H^0 \times H^0$. Finally, the continuity of $\mathcal{L}(q,\lambda,\nu)$ in $H^1 \times H^0 \times H^0$ implies that $\mathcal{L}(\delta^*) = \lim_{k\to\infty} \hat{\mathcal{L}}(\delta^k)$. Thus, δ^* has optimal cost and is feasible. ∎

If the grid points in S^h are free parameters although constrained from becoming too close together (e.g. if h is the grid interval for a uniform mesh, then a constraint that the grid points are separated by at least h/2 is acceptable), and (11) holds uniformly in the mesh, then a similar proof demonstrates the existence of a solution to the finite dimensional problem in this larger set.

IV. CONVERGENCE OF STATE AND CONTROL

Define $w = (q, \lambda, \nu)$ and let $w^* = (q^*, \lambda^*, \nu^*)$ denote an optimal dual solution. Expanding $\mathcal{L}(w)$ about w^* yields:

$$(15) \qquad \mathcal{L}(w) = \mathcal{L}(w^*) - \ell(w-w^*) - a(w-w^*, w-w^*)$$

where $-\ell(\cdot)$ and $-2a(\cdot,\cdot)$ are the first and second derivatives of the operator $\mathcal{L}$ evaluated at w^*. Inspection of $\mathcal{L}$ in (10) reveals that $\ell(\cdot)$ and $a(\cdot,\cdot)$ are given by

$$(16) \quad \ell(w) = ((Q^{-1}\ell_1(w^*), \ell_1(w))) + ((R^{-1}\ell_2(w^*), \ell_2(w))) - ((b_c, \lambda)) + (q(0) - K_s^T \nu(0), x_o) - (\nu(0), b_s)$$

$$(17) \quad a(w,w) = \tfrac{1}{2}((\ell_1(w), Q^{-1}\ell_1(w))) + \tfrac{1}{2}((\ell_2(w), R^{-1}\ell_2(w))).$$

Note that Equations 8 and 9 imply that ℓ can be rewritten in the form:

$$(18) \quad \ell(w) = ((x^*, \ell_1(w))) + ((u^*, \ell_2(w))) - ((b_c, \lambda)) + (q(0) - K_s^T \nu(0), x_o) - (\nu(0), b_s).$$

Rearranging terms in (18) and integrating the $\dot{q}$ term by parts yields:

$$(19)\quad \ell(w) = ((K_s Ax^* + K_s Bu^*, \nu)) - ((K_c u^* + b_c, \lambda)) - (K_s x_o + b_s, \nu(0))$$

$$= ((K_s \dot{x}^*, \nu)) - ((K_c u^* + b_c, \lambda)) - (K_s x_o + b_s, \nu(0))$$

$$(20)\quad = -\langle \nu, K_s x^* + b_s \rangle_C - ((K_c u^* + b_c, \lambda)).$$

Note that the term involving q, $((\dot{x}^* - Ax^* - Bu^*, q))$, vanished since (x^*, u^*) satisfy the differential equation.

Since w^* is optimal in the dual problem, the necessary condition $\ell(w - w^*) \geq 0$ holds for all $w \in F$, the feasible region. Let w^h minimize $\mathcal{L}$ over $F \cap S^h$; then for all $v^h \in F \cap S^h$,

$$(21)\quad \mathcal{L}(w^*) - \mathcal{L}(v^h) \geq \mathcal{L}(w^*) - \mathcal{L}(w^h)$$

$$= a(w^h - w^*, w^h - w^*) + \ell(w^h - w^*)$$

$$(22)\quad \geq a(w^h - w^*, w^h - w^*)$$

Let e_u and $e_\ell > 0$ denote upper and lower bounds for the eigenvalues of both Q^{-1} and R^{-1}. Then from the definition of $a(\cdot,\cdot)$ in (17),

$$(23)\quad a(w^h-w^*, w^h-w^*) \geq \tfrac{1}{2}e_\ell ||\ell_1(w^h-w^*)||^2_{L^2} + \tfrac{1}{2}e_\ell ||\ell_2(w^h-w^*)||^2_{L^2}$$

If x^h and u^h are the state and control corresponding to w^h in (8) and (9), then

$$(24)\quad ||x^h-x^*|| = ||Q^{-1}\ell_1(w^h-w^*)|| \leq e_u||\ell_1(w^h-w^*)||$$

$$(25)\quad ||u^h-u^*|| = ||R^{-1}\ell_2(w^h-w^*)|| \leq e_u||\ell_2(w^h-w^*)||$$

Combining these inequalities yields the following <u>fundamental error estimate for the Ritz-Trefftz method</u>:

$$(26)\quad ||x^h-x^*||^2_{L^2} + ||u^h-u^*||^2_{L^2} \leq \frac{2e_u^2}{e_\ell}\,[\mathcal{L}(w^*)-\mathcal{L}(v^h)]$$

for all $v^h \in F \cap S^h$. An upper bound is determined for the right side above using approximation; note that the right side of (26) can also be expanded about w^*:

$$(27)\quad \mathcal{L}(w^*) - \mathcal{L}(v^h) = \ell(v^h-w^*) + a(v^h-w^*, v^h-w^*)$$

The order to which w^* can be interpolated depends on the properties of the set S^h and the smoothness of w^*. The smoothness properties are examined first by studying a few illustrative problems. Consider the state constrained problem:

$$\min \int_0^1 u^2(t)\, dt \tag{28}$$

$$\text{s.t. } \dot{x}(t) = u(t)$$

$$x(0) = .25$$

$$x(t) \geq \alpha(t)$$

$$\alpha(t) = \begin{cases} t \text{ for } 0 \leq t \leq .5 \\ 1-t \text{ for } .5 \leq t \leq 1.0 \end{cases}$$

The optimal solutions x^* and u^* are shown in Figure 1. Note that $\dot{x}^*$ and u^* both have jumps and hence $x^* \in H^1$ and $u^* \in H^0$. It will be seen shortly that in general the smoothness of u^* and x^* depends on the smoothness of the derivative of K_s and b_s. If

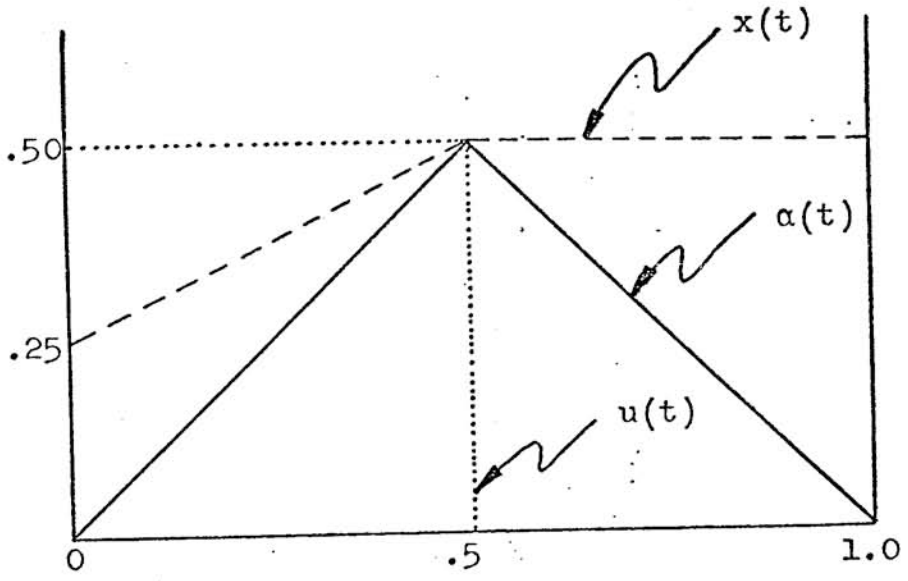

FIGURE 1: Optimal Solution to (28)

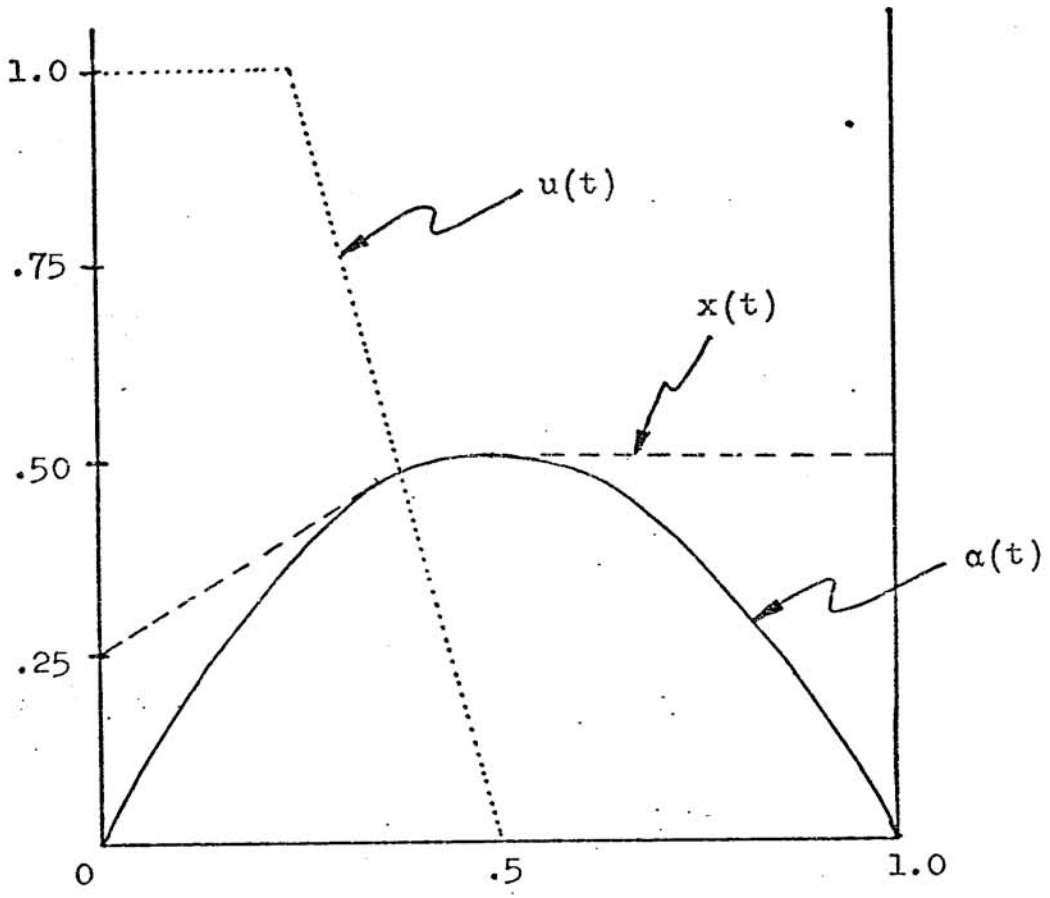

Figure 2: Optimal Solution to (28) for

$\alpha(t) = 2t(1-t)$

$\alpha(t) = 2t(1-t)$, then $u^* \in H^1$ and $x^* \in H^2$ as shown in Figure 2.

This particular problem is a one-dimensional form of the "obstacle problem" studied by Lewy and Stampacchia [5]. In a much more general framework, they show that if $\alpha \in H^2$, then $u^* \in H^1$ and $x^* \in H^2$. The most general conditions guaranteeing $u^* \in H^1$ and $x^* \in H^2$ for a state constrained control problem are still not known to the author; however, if no control constraints are present and the system is completely controllable, then certainly $u^* \in H^1$ and $x^* \in H^2$ when condition (A3) below is satisfied.

Now consider the control constrained problem:

$$\min \int_0^1 u^2(t)\, dt \tag{29}$$

$$\text{s.t. } \dot{x}(t) = u(t)$$

$$x(0) = 0,\ x(1) = 1$$

$$\beta(t) \geq u(t) \geq \alpha(t).$$

Since q^* satisfies $\dot{q} = 0$, then q^* is a constant. Also $u^*(t)$ satisfies:

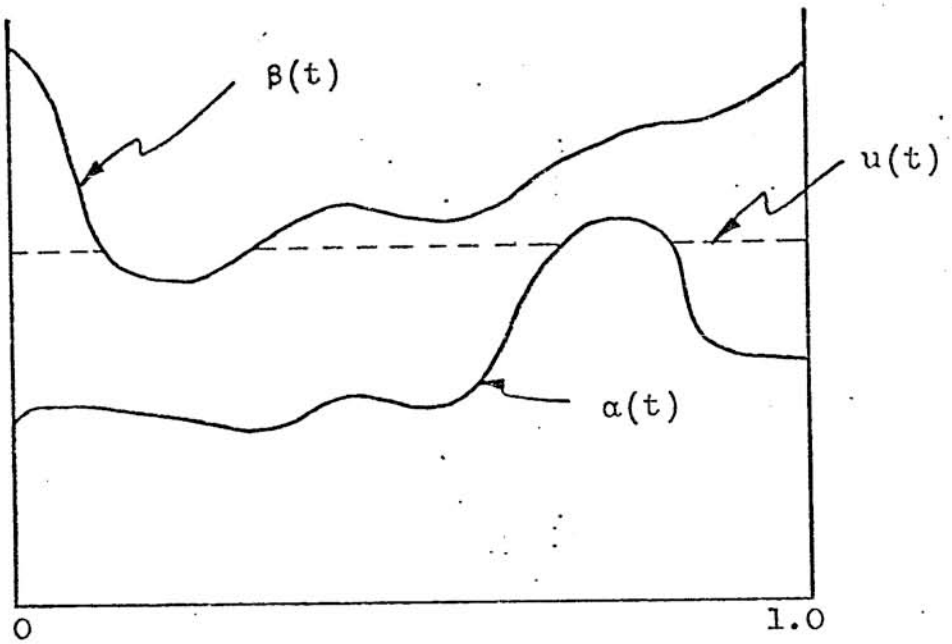

FIGURE 3: Typical Solution to (29).

(30) $\quad q^*u^*(t)+u^*(t)^2 \leq q^*v+v^2$ for all v such that

$$\alpha(t) \leq v \leq \beta(t)$$

Thus, $u^*(t)$ is a constant until a constraint forces the control away from this constant value. A typical solution to this problem is as shown in Figure 3. Note that $\beta,\alpha \in H^1$ implies $u^* \in H^1$ and $x^* \in H^2$. Below, it is proved that this result holds in general.

The proof of these regularity properties is contained in the necessary conditions. At some time t, $(x^*,u^*,q^*,\lambda^*,\nu^*)$ satisfy

(31)
$$\dot{x}(t) = Ax(t)+Bu(t)$$
$$\dot{q}(t) = -A^T q(t)-Qx(t)+A^T K_s^T \nu(t)$$
$$u(t)+R^{-1}B^T q(t)-R^{-1}B^T K_s^T \nu(t)+R^{-1}K_c^T \lambda(t) = 0$$
$$\overline{K_c u(t)+b_c} = 0$$
$$\hat{\lambda} = 0$$

(32)
$$\overline{K_s x(t)+b_s} = 0$$
$$\hat{\nu} = \text{constant}$$

where a bar, "$\overline{\quad}$", over a vector means that only components corresponding to binding constraints are included while a hat, "^", means that only non-binding components

are considered. Henceforth all variables are optimal so that the star, "*", will be omitted.

The condition (32) on the binding state constraints implies $(\overline{K_s \dot{x}(t)}) = 0$ or a state constraint would be violated. Replacing (32) by this condition and rearranging terms results in:

$$(33)\quad \begin{bmatrix} I & 0 & -B & 0 & 0 \\ 0 & I & 0 & 0 & -A^T\overline{K}_s^T \\ 0 & 0 & I & R^{-1}\overline{K}_c^T & -R^{-1}B^T K_s^T \\ 0 & 0 & -\overline{K}_c & 0 & 0 \\ \overline{K}_s & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x} \\ \dot{q} \\ u \\ \overline{\lambda} \\ \overline{\nu} \end{bmatrix} = \begin{bmatrix} Ax \\ -A^T q - Qx + A^T \widehat{K_s^T \nu} \\ -R^{-1}B^T(q - \widehat{K_s^T \nu}) \\ \overline{b}_c \\ 0 \end{bmatrix}$$

After eliminating the K_c and K_s blocks in the first and third columns, the following system is obtained:

$$(34)\quad \begin{bmatrix} I & 0 & -B & 0 & 0 \\ 0 & I & 0 & 0 & -A^T\overline{K_s}^T \\ 0 & 0 & I & R^{-1}\overline{K_c}^T & -R^{-1}B^T\overline{K_s}^T \\ 0 & 0 & 0 & \overline{K_c}R^{-1}\overline{K_c}^T & -K_cR^{-1}B^T\overline{K_s}^T \\ 0 & 0 & 0 & -\overline{K_s}BR^{-1}\overline{K_c}^T & \overline{K_s}BR^{-1}B^T\overline{K_s}^T \end{bmatrix} \begin{bmatrix} \dot{x} \\ \dot{q} \\ u \\ \overline{\lambda} \\ \overline{\nu} \end{bmatrix} =$$

$$\begin{bmatrix} Ax \\ -A^T\dot{q}-Qx+A^T\overline{K_s}^T\overline{\nu} \\ -R^{-1}B^T(q-\overline{K_s}^T\overline{\nu}) \\ -\overline{K_c}R^{-1}B^T(q-\overline{K_s}^T\overline{\nu})+\overline{b_c} \\ -\overline{K_s}Ax+B\overline{K_s}R^{-1}B^T(q-\overline{K_s}^T\overline{\nu}) \end{bmatrix}.$$

Note that this system is non-singular provided the columns of the matrix $[\overline{K_c}^T : -B^T\overline{K_s}^T]$ are linearly independent; i.e., let $V = \overline{K_c}$ and $W = \overline{K_s}B$ and the lower right 2 x 2 submatrix in (34) reduces to:

$$\begin{bmatrix} VR^{-1}V^T & -VR^{-1}W^T \\ -WR^{-1}V^T & WR^{-1}W^T \end{bmatrix}.$$

Operating on this matrix from the left and right by the vector $\begin{bmatrix} y \\ z \end{bmatrix}$ yields:

$$(y^T V - z^T W) R^{-1} (V^T y - W^T z).$$

This expression vanishes for $y, z \neq 0$ if and only if the columns of $[V^T : W^T]$ are linearly independent as stated above.

Observe that if no state constraints are present, then this independence is equivalent to requiring that the gradients of the binding control constraints are linearly independent while if no control constraints are present and $B = I$, then this requires that the gradients of the binding state constraints are linearly independent. Formally, assume:

(A3) If x and u lie on the boundary of the sets $\{x: K_s x + b_s \leq 0\}$ and $\{u: K_c u + b_c \leq 0\}$, respectively, and if $\overline{K_c}, \overline{K_s}$ are the rows of K_c and K_s corresponding to the binding constraints, then the columns of $[\overline{K_c}^T : B\,\overline{K_s}^T]$ are linearly independent.

Now, it is easily seen that (x, u, q, λ, ν) are analytic on any interval where the binding constraints are the same. Since the matrix on the left side of (34) is non-singular, then $\dot{x}$ and $\dot{q}$ can be solved for in terms of x and q.

Thus, x and q are described by a linear first-order constant coefficient system whose solution is analytic of course. Hence (ν, λ, u) are also analytic. Note that if the data is time varying, then $(\dot{x}, \dot{q}, u, \lambda, \nu)$ will be only as smooth as $R, Q, \dot{K}_s, \dot{b}_s, K_c, b_c, A, B$.

For future reference, introduce the following assumption:

(A4) There are only a finite number of times where binding constraints change.

For practical problems (A4) always holds. It is easy, however, to concoct a problem where (A4) is violated.

Theorem 2: Suppose (A3) holds. If (x, q, u, λ) are optimal in the control constrained problem, then $q \in H^3$, $x \in H^2$, and $(u, \lambda) \in H^1$. If state constraints are also present, then $(\dot{x}, \dot{q}, u, \lambda, \nu) \in BV$. Furthermore, (x, q, u, λ, ν) are analytic on every interval where the binding constraints do not change. If (A4) holds, then in the state constrained problem $(x, q) \in PC^1$ and $(u, \lambda, \nu) \in PC^0$. If no state constraints are present, then $q \in PC^3$, $x \in PC^2$, and $(u, \lambda) \in PC^1$.

Proof: For the general state and control constrained problem, the necessary conditions (31) above imply that at $t \in [0,1]$, $u(t)$ solves the following problem and $\lambda(t)$ is the Kuhn-Tucker multiplier corresponding to the constraints:

$$(35) \qquad \min \tfrac{1}{2} u^T R u + u^T B^T (q(t) - K_s^T \nu(t))$$

$$\text{s.t. } K_c u + b_c \leq 0.$$

In Appendix 2, it is proved that (A3) implies that there exists a constant c such that

$$(36) \quad |u(t_1)-u(t_2)| + |\lambda(t_1)-\lambda(t_2)| \leq c|q(t_1)-q(t_2)| + c|\nu(t_1)-\nu(t_2)|.$$

Since $\nu, q \in BV$, then (36) implies that $u, \lambda \in BV$. Furthermore, if no state constraints are present, then the ν term does not appear in (36) and since $q \in \mathcal{A}$, then $u, \lambda \in \mathcal{A}$. Both x and q satisfy $\dot{x} = Ax + Bu$ and $\dot{q} = -A^T q - Qx$ so it follows that $\dot{x}, \dot{q} \in \mathcal{A}$. Hence, $\dot{q}$ is uniformly bounded on $[0,1]$ and (36) implies that $\dot{u}$ and $\dot{\lambda}$ are uniformly bounded on $[0,1]$ or $u, \lambda \in H^1$. Again, by the equation for $\dot{x}$, $x \in H^2$ and by the equation for $\dot{q}$, it follows that $q \in H^3$.

Finally, the analyticity result was proved earlier and the PC properties follow from the analyticity property and the results above. □

Next the approximation problem posed above will be analyzed. The approximating sets to be studied will consist

of polynomials of degree p-1 on each grid interval and derivatives of order q continuous across each grid point. As shown in [8], for $q< p/2$, this space is of degree p and continuity q+1. Assume that $q^*, x^* \in H^2$ and $u^*, \lambda^*, \nu^* \in H^1$; as shown above, this is an appropriate assumption for many problems.

Case 1: $S^{q,h}$ consists of continuous piecewise linear functions and $S^{\lambda,h}$ and $S^{\nu,h}$ consists of piecewise constant functions where h is the largest grid interval.

Let $v^h = (q^h, \lambda^h, \nu^h)$ be the interpolate of (q^*, λ^*, ν^*). Note that v^h lies in the feasible region for the dual problem and by spline approximation theory, (q^h, λ^h, ν^h) approximates (q^*, λ^*, ν^*) to order 1. Thus $a(v^h - w^*, v^h - w^*) = O(h^2)$. Now consider the linear term in the expansion (27). From the complementary slackness condition, $(\lambda^*(t), K_c u^*(t) + b_c) = 0$ for all $t \in [0,1]$. If the binding constraints do not change on the grid interval $[t_k, t_{k+1}]$, then either $\lambda_j(t) = \lambda^*_j(t) = 0$ or $(K_c u^*(t) + b_c)_j = 0$ on $[t_k, t_{k+1}]$ and hence $(\lambda^h(t) - \lambda^*(t), K_c u^*(t) + b_c)$ vanishes on the interval. If the j-th constraint changes from binding to non-

binding at $s \in [t_k, t_{k+1}]$, then by the fundamental theorem of calculus,

$$(K_c u^*(t)+b_c)_j = \int_s^t (K_c \dot{u}^*(t))_j dt$$

$$\leq h^{\frac{1}{2}} \{ \int_{t_k}^{t_{k+1}} (K_c \dot{u}^*(t))_j^2 \, dt \}^{\frac{1}{2}}$$

$$\lambda^h{}_j(t) - \lambda^*{}_j(t) \leq h^{\frac{1}{2}} \{ \int_{t_k}^{t_{k+1}} \dot{\lambda}^*{}_j(t)^2 \, dt \}^{\frac{1}{2}}$$

$$(37) \quad \int_{t_k}^{t_{k+1}} (\lambda^h(t)-\lambda^*(t), K_c u^*(t)+b_c)dt \leq$$

$$h^2 \{ \int_{t_k}^{t_{k+1}} |\dot{\lambda}^*(t)|^2 dt \int_{t_k}^{t_{k+1}} c|\dot{u}^*(t)|^2 dt \}^{\frac{1}{2}}$$

where c depends on the norm of K_c. Summing (37) over k and using the inequality $2ab \leq a^2 + b^2$, yields:

$$(38) \quad ((\lambda^h-\lambda^*, K_c u^*+b_c)) \leq \tfrac{1}{2}h^2[c||u^*||^2_{H^1} + ||\lambda^*||^2_{H^1}]$$

Now consider the $\langle \nu^h-\nu^*, K_s x^*+b_s \rangle_C$ term. Again the complementary slackness condition implies that

$$\int_{t_k}^{t_{k+1}} (K_s x^*(t)+b_s)d(\nu^h-\nu^*) = 0$$

on every interval where there is no change in the binding constraints. Now suppose that at $r \in [t_k, t_{k+1}]$, the j-th constraint changes from binding to non-binding. Since $(K_s x^*(r)+b_s)_j = 0$ and $x^* \in \mathcal{O}$, the state constraint would be violated unless $(K_s \dot{x}^*(r))_j = 0$ and another application of the fundamental theorem of calculus yields:

$$|v^h{}_j(t) - v^*{}_j(t)| \leq h^{\frac{1}{2}} \{ \int_{t_k}^{t_{k+1}} (\dot{v}^*{}_j(t))^2 \, dt \}^{\frac{1}{2}}$$

$$(K_s x^*(t)+b_s)_j \leq \tfrac{1}{2} h^{3/2} \{ \int_{t_k}^{t_{k+1}} (K_s \ddot{x}^*(t))_j^2 \, dt \}^{\frac{1}{2}}$$

Exactly as in (38) above, the following estimate holds:

$$\langle v^h - v^*, K_s x^* + b_s \rangle_C \leq \frac{h^2}{4} (c||x^*||^2_{H^2} + ||v^*||^2_{H^1})$$

Thus $\ell(v^h - w^*) = 0(h^2)$ and both the linear and the quadratic term in the expansion of $\mathcal{L}(w^*) - \mathcal{L}(v^h)$ is of order 2. Hence by (26) it follows that x^h and u^h approximate x^* and u^* to order 1.

Case 2: $S^{q,h}$ consists of continuous piecewise quadratic functions, $S^{\lambda,h}$ and $S^{\nu,h}$ consist of continuous piecewise linear functions, and (A4) holds.

Again if v^h is the interpolate of w^*, then v^h is feasible in the dual problem. On intervals between points where the constraints change from binding to non-binding, w^* is analytic by Theorem 2 and by spline approximation theory, $(\dot{q}^h, \lambda^h, \nu^h)$ approximates $(\dot{q}^*, \lambda^*, \nu^*)$ to order 2. On mesh intervals containing points where constraints change from binding to non-binding, $(\ddot{q}^*, \dot{\lambda}^*, \dot{\nu}^*)$ is possibly discontinuous and hence $(\dot{q}^h, \lambda^h, \nu^h)$ only approximates $(\dot{q}^*, \lambda^*, \nu^*)$ to order 1. Since there are only a finite number of these change points, then $a(v^h-w^*, v^h-w^*) = O(h^3)$. An analysis of the linear term, $\ell(v^h-w^*)$, as in case 1 reveals that $(K_c u^*(t)+b_c, \lambda^h(t)-\lambda^*(t)) = 0$ and $(K_s x^*(t)+b_s)^T d(\nu^h(t)-\nu^*(t)) = 0$ on intervals where there is no change in the binding constraints. On intervals where the j-th binding constraint changes to non-binding, $\lambda^h{}_j(t)$ only approximates $\lambda^*{}_j(t)$ to order 1 since $\dot{\lambda}^*{}_j(t)$ is possibly discontinuous on the grid interval while $(K_c u^*(t)+b_c)_j$ is of order 1 since the constraint is binding somewhere on the interval. Hence

$$\int_{t_k}^{t_{k+1}} (\lambda^h(t)-\lambda^*(t), K_c u^*(t)+b_c)\, dt = O(h^3)$$

Since there are only a finite number of points where there is a change in binding constraints, then $((\lambda^h-\lambda^*, K_c u^*+b_c)) = O(h^3)$. A similar statement holds for the $\langle \nu^h-\nu^*, K_s x^*+b_s\rangle_c$ term so that $\mathcal{L}(v^h-w^*) = O(h^3)$. Thus both the linear term and the quadratic term in the expansion of $\mathcal{L}(w^*)-\mathcal{L}(v^h)$ is of order 3 and hence (26) implies that x^h and u^h approximate x^* and u^* to order 3/2.

Case 3: $S^{q,h}$ consists of continuous piecewise polynomials of degree r, $S^{\lambda,h}$ and $S^{\nu,h}$ consist of piecewise polynomials of degree r-1, the grid points are free parameters, and (A4) holds.

Distribute the grid points on [0,1] so that there is a mesh point at any time where a change occurs in the binding constraints and let $v^h = (q^h, \lambda^h, \nu^h)$ where q^h is an interpolate of q^* and λ^h and ν^h are the interpolates of λ^* and ν^* given in Corollary 1A and Corollary 4A in the appendix; i.e. $\lambda^* \geq \lambda^h \geq 0$,

$\dot{\nu}^* \geq \dot{\nu}^h \geq 0$, $\nu^h(1) = 0$, and λ^h and ν^h approximates λ^* and ν^* to order r. Thus v^h is feasible in the dual problem and $a(v^h-w^*, v^h-w^*) = 0(h^{2r})$. Since there are no changes in the binding constraints on the interior of grid intervals and since λ^h and $\dot{\nu}^h$ lie below λ^* and $\dot{\nu}^*$, then λ^h_j vanishes whenever the j-th control constraint is non-binding and $\dot{\nu}^h_j$ is zero whenever the j-th state constraint is non-binding. Thus $\ell(v^h-w^*) = 0$. Hence $\mathcal{L}(w^*)-\mathcal{L}(v^h) = a(v^h-w^*, v^h-w^*)$ is of order 2r and x^h and u^h approximates x^* and u^* to order r. These results are summarized in:

Theorem 3: If $x^*, q^* \in H^2$, $u^*, \lambda^*, \nu^* \in H^1$, and $S^{q,h}$ consists of piecewise linear functions continuous across the grid points and $S^{\lambda,h}$ and $S^{\nu,h}$ consist of piecewise constant functions, then

$$||x^*-x^h||_{L^2} \leq ch \tag{39}$$

$$||u^*-u^h||_{L^2} \leq ch$$

$$\mathcal{L}(w^*)-\mathcal{L}(w^h) \leq ch^2 \tag{40}$$

If $S^{q,h}$ consists of continuous piecewise quadratic functions, $S^{\nu,h}$ and $S^{\lambda,h}$ consist of piecewise linear continuous functions, and (A4) holds, then

(41) $$||x^*-x^h|| \leq ch^{3/2}$$

$$||u^*-u^h|| \leq ch^{3/2}$$

(42) $$\mathcal{L}(w^*) - \mathcal{L}(w^h) \leq ch^3$$

If (A4) holds and $S^{q,h}$ consists of continuous piecewise polynomials of degree r while $S^{\lambda,h}$ and $S^{\nu,h}$ consist of piecewise polynomials of degree r-1 with no continuity requirements, then

(43) $$||x^*-x^h|| \leq ch^r$$

$$||u^*-u^h|| \leq ch^r$$

$$\mathcal{L}(w^*)-\mathcal{L}(w^h) \leq ch^{2r}$$

■

A result in section VI will prove that if $\dot{\lambda}^*$ or $\dot{\nu}^*$ has a discontinuity and the closest grid point to the discontinuity lies at least a distance sh away for $s \in (0,1]$, then the L^2 norm of the error of the best polynomial approximation to λ^* or ν^* is <u>at least</u> $ch^{3/2}$ where c depends on the degree of the polynomials in the space. By theorem 3, the space of piecewise linear functions for $S^{\lambda,h}$ and $S^{\nu,h}$ and piecewise quadratic functions for $S^{q,h}$ actually achieves this rate. Thus the bound is tight

and higher order spaces would appear to not be of advantage if the grid points are fixed unless the constant c in the convergence rates is substantially smaller in the higher order spaces.

V. CONVERGENCE OF DUAL VARIABLES

The question of convergence of the dual variables will now be considered. The principal result in this section is the following: in control constrained problems, if λ is constrained to the space spanned by the columns of K_c and q satisfies $q(1) = 0$, then the dual function $\mathcal{L}$ is positive definite.

The first lemma below proves that in cases with no state or control constraints, $\mathcal{L}$ is positive definite in the space with $q(1) = 0$:

Lemma 1: There exists c_1, c_2 such that

$$(44) \qquad c_2 ||p||_{H^1} \geq ||\dot{p} + A^T p||_{H^0} \geq c_1 ||p||_{H^1}$$

for all $p \in H^1$ with $p(1) = 0$.

Proof: The upper bound is immediate so now consider the lower bound. Define w by $w = \dot{p} + A^T p$. Then since $p(1) = 0$,

$$p(t) = \int_1^t e^{A^T(\sigma-t)}\, w(\sigma)\, d\sigma.$$

Since $|e^{A^T(\sigma-t)}|$ is uniformly bounded for $t, \sigma \in [0,1]$, then

$$|p(t)| \le c \int_0^1 |w(\sigma)|\, d\sigma \le c\, ||w||_{H^0}. \tag{45}$$

Thus, squaring and integrating (45),

$$||p||_{H^0}^2 \le c^2\, ||w||_{H^0}^2. \tag{46}$$

From the definition of w,

$$\begin{aligned} \dot{p}(t) &= w(t) - A^T p(t) \\ |\dot{p}(t)| &\le |w(t)| + c\,|p(t)| \\ |\dot{p}(t)|^2 &\le 2\,|w(t)|^2 + 2\,c\,|p(t)|^2 \\ ||\dot{p}||_{H^0}^2 &\le 2\,||w||_{H^0}^2 + 2\,c\,||p||_{H^0}^2. \end{aligned} \tag{47}$$

Combining (46) and (47),

$$(48) \qquad ||\dot{p}||_{H^0}^2 \leq (2+2c^3)\ ||w||_{H^0}^2.$$

Adding (46) and (48) yields the lemma. ∎

The next lemma proves that $\mathcal{L}$ is positive definite in control constrained problems in the subspace where $q(1) = 0$ and λ lies in the space spanned by the columns of K_c.

Lemma 2: There exists $c_1, c_2 > 0$ such that

$$(49) \quad c_2||p||_{H^1}^2 + c_2||\lambda||_{H^0}^2 \geq ||\dot{p}+A^T p||_{H^0}^2 + ||B^T p + K_c^T \lambda||_{H^0}^2 \geq c_1||p||_{H^1}^2 + c_1||\lambda||_{H^0}^2$$

for all $p \in H^1$ and $\lambda \in H^0$ satisfying the following condition:

(50) $p(1) = 0$ and $\lambda(t)$ is perpendicular to the null space of K_c^T for all $t \in [0,1]$.

Proof: Once again the upper bound is immediate so now consider the lower bound. Let (p^k, λ^k) be a minimizing sequence for the following problem:

$$\text{(51)} \qquad \inf\ c||p||_{H^1}^2 + ||B^T p + K_c^T \lambda||_{H^0}^2$$

$$\text{s.t. } ||p||_{H^1}^2 + ||\lambda||_{H^0}^2 = 1$$

and condition (50) above

where $c > 0$ satisfies $||\dot{p}+A^T p||_{H^0}^2 \geq c||p||_{H^1}^2$ as in Lemma 1. Suppose that the extremand converges to zero as $k \to \infty$. Then $c||p^k||_{H^1}^2 \to 0$ and hence $||p^k||_{H^0} \to 0$ and $||\lambda^k||_{H^0} \to 1$. Now,

$$\text{(52)} \qquad c||p^k||_{H^1}^2 + ||B^T p^k + K_c^T \lambda^k||_{H^0}^2 \geq ||K_c^T \lambda^k||_{H^0}^2$$

$$+ c||p^k||_{H^1}^2 + ||B^T p^k||_{H^0}^2 - 2||B^T p^k||_{H^0}||K_c^T \lambda^k||_{H^0}.$$

Since $||\lambda^k||_{H^0} \to 1$ and λ^k is perpendicular to the null space of K_c^T, then $||K_c^T \lambda^k||_{H^0}$ is bounded away from zero. This yields a contradiction in (52), however, since the left side converges to zero by assumption while all the

terms on the right involving p^k converge to zero except for $||K_c^T \lambda^k||_{H^0}$.

■

Lemma 2 is now used to prove a dual variable convergence theorem. Let $\lambda^\perp$ denote the component of λ perpendicular to the null space of K_c^T and let (q^h, λ^h) denote the solution to the Ritz-Trefftz problem.

Theorem 4: Theorem 3 is valid with $||q^*-q^h||_{H^1}$ and $||\lambda^{*\perp}-\lambda^{h\perp}||_{H^0}$ replacing $||x^*-x^h||_{L^2}$ and $||u^*-u^h||_{L^2}$.

Proof: From the definition of ℓ_1 and ℓ_2 and Lemma 2, there exists $c > 0$ such that

$$(53) \quad ||\ell_1(w^*-w^h)||_{H^0}^2 + ||\ell_2(w^*-w^h)||_{H^0}^2 \geq c||q^*-q^h||_{H^1}^2 + c||\lambda^{*\perp}-\lambda^{h\perp}||_{H^0}^2.$$

The analogue of the fundamental error estimate given in (26) is the following:

$$||q^*-q^h||_{H^1}^2 + ||\lambda^{*\perp}-\lambda^{h\perp}||_{H^0}^2 \leq \frac{2}{e_\ell} [\mathcal{L}(w^*)-\mathcal{L}(v^h)]$$

for all $v^h \in F \cap S^h$. The interpolation results of the previous section yield upper bounds for the right side above exactly as in Theorem 3.

■

VI. ERROR IN FREE BOUNDARY

One final numerical question examined is the following: If the grid points are free parameters in the set S^h, then at what rate do the grid points in the Ritz-Trefftz solution converge to a point where there is a change in the binding constraints.

For problems with just one constraint, the boundary between the region where the constraint is binding and where the constraint is non-binding is referred to as the free boundary. For our purposes, the free boundary will refer to any point where there is a change in the binding constraints. As noted in the examples of Section IV, a change in the binding constraint is characterized by a change in a derivative of x and u. This section then analyzes the rate at which the grid points converge to times where the m-th derivative of u has a jump.

After a somewhat lengthy computation, it is shown that for polynomial spaces of degree $r+1, r, r$ in $S^{q,h}, S^{\lambda,h}, S^{\nu,h}$, respectively, the free boundary is determined to order $2r/(2m+1)$. Thus, an increase in the order of the discontinuity results in an increase in the difficulty in determining the boundary as would be expected.

Recall that in Theorem 3 it was proved that if $S^{q,h}$, $S^{\lambda,h}$, and $S^{\nu,h}$ are spaces of polynomials with free grid

points and the spaces have degree $r+1, r, r$, respectively, then (A4) implies:

$$||x^* - x^h|| \leq ch^r$$

(54)

$$||u^* - u^h|| \leq ch^r.$$

Examining the equations defining x^h and u^h, it is seen that x^h and u^h are polynomials of degree $r+1$ and r, respectively. Now suppose that u^* has a discontinuity in its m-th derivative at $\bar{t}$. Then the analysis that follows proceeds to determine how far the two grid points bordering the grid interval containing $\bar{t}$ can be moved away form $\bar{t}$ and still maintain the error bound ch^r above.

To begin, let $s \in [0,1]$ and consider the problem of minimizing the L^2 distance from the space S^h of polynomials of degree ℓ on $[-hs, h-hs]$ to the following function:

$$\phi_{m,s}(t) = \begin{cases} 0 & \text{for } -sh \leq t \leq 0 \\ t^m & \text{for } 0 \leq t \leq h-hs. \end{cases}$$

That is, consider the minimization problem:

$$(55) \qquad \min_{\alpha} \int_{-sh}^{h-sh} (\phi_{m,s}(t) - \sum_{k=0}^{\ell} \alpha_k t^k)^2 \, dt.$$

Let $y = t/h$ and $\beta_k = \alpha_k h^k$. Then the extremand in (55) reduces to

$$(56) \qquad h \int_{-s}^{1-s} (\phi_{m,s}(hy) - \sum_{k=0}^{\ell} \beta_k y^k)^2 \, dy.$$

Carrying out the integration results in a quadratic form

$$(57) \qquad h[\beta^T H \beta - 2h^m b^T \beta + \frac{h^{2m}}{2m+1} (1-s)^{2m+1}]$$

where

$$H_{j,k} = \frac{1}{k+j+1} [(1-s)^{k+j+1} - (-s)^{k+j+1}]$$

$$b_k = \frac{(1-s)^{k+m+1}}{k+m+1}.$$

Now it will be assumed that $\ell \geq m$. This assumption is justified intuitively since to estimate an m-th order discontinuity, it would seem necessary that the approximating polynomials be at least of degree m. The mathematics will also point this out later in the development. Thus, if $\ell \geq m$, b can be expressed as $b = H\delta^m + Hq^m$ where

$$(\delta^m)_j = \begin{cases} 0 & \text{for } j \neq m \\ 1 & \text{for } j = m \end{cases}$$

$$(Hq^m)_j = \frac{(-s)^{j+m+1}}{j+m+1} .$$

Hence the following equalities hold:

$$(58) \qquad b^T H^{-1} b = (H\delta^m + Hq^m)^T H^{-1} (H\delta^m + Hq^m)$$

$$= \frac{(1-s)^{2m+1} - (-s)^{2m+1} + 2(-s)^{2m+1}}{2m+1} + q^{m^T} Hq^m .$$

We will only be interested in the value of (57) for small s. As $s \to 0$, H approaches a constant nonsingular matrix that is not a function of s. Hence, $q^m Hq^m = q^m H\, H^{-1}\, Hq^m = O(s^{2m+2})$ since the largest term in Hq^m is $O(s^{m+1})$. Equation 58 then reduces to

$$(59) \qquad \frac{(1-s)^{2m+1} + (-s)^{2m+1}}{2m+1} + O(s^{2m+2}).$$

The quadratic (57) above is minimized for $\beta = h^m H^{-1} b$ so that the optimal value for the extremand (57) is

$$(60) \qquad h^{2m+1}\left[\frac{(1-s)^{2m+1}}{2m+1} - b^T H^{-1} b\right].$$

Substituting (59) for the last term in (60) proves

Lemma 3:

$$\int_{-hs}^{h-hs} |\phi_{m,s}(t)-\phi^h(t)|^2 dt \geq h^{2m+1}\left[\frac{s^{2m+1}}{2m+1} + O(s^{2m+2})\right]$$

for all $\phi^h \in S^h$. ∎

Now let $f_{r,s}$ denote a function on $[-sh, h-sh]$ with a discontinuity in the r-th derivative at the origin and derivatives of order $(\ell+1)$ away from the origin that are bounded by M. If $z \in S^h$, then the distance from $f_{m,s} - z$ to S^h is the same as the distance from $f_{m,s}$ to S^h. Construct z so that its value and first ℓ derivatives agree with the same derivatives of $f_{m,s}$ from the left at $t = 0$. Define $\hat{f}_{m,s} = f_{m,s} - z$ and let $m!\gamma_k$ be the k-th derivative of $\hat{f}_{m,s}$ from the right. Then

$$(61) \qquad \hat{f}_{m,s}(t) = \sum_{k=r}^{\ell} \gamma_k \phi_{k,s}(t) + R(t)$$

where R(t) is a remainder term. By the Taylor series remainder formula, $|R(t)| \leq ch^{\ell+1}$ where c depends on M.

For any $\phi^h \in S^h$,

$$(62) \qquad ||\hat{f}_{m,s} - \phi^h|| \geq ||\sum_{k=r}^{\ell} \gamma_k \phi_{k,s} - \phi^h|| - ||R||$$

where $||\cdot||$ is the L^2 norm on $[-hs, h-hs]$. Thus,

$$(63) \qquad \inf_{\phi^h \in S^h} ||f_{m,s} - \phi^h|| \geq \inf_{\phi^h \in S^h} ||\sum_{k=r}^{\ell} \gamma_k \phi_{k,s} - \phi^h|| - ||R||$$

Lemma 4:

$$||f_{m,s} - \phi^h|| \geq O(h^{\ell + \frac{3}{2}}) + \{\gamma_m^2 h^{2m+1} [\frac{s^{2m+1}}{2m+1} + O(s^{2m+2})]\}^{\frac{1}{2}}.$$

Proof: The $O(h^{\ell + \frac{3}{2}})$ term arises from the $||R||$ in (63). Proceeding exactly as in Lemma 3, the optimal value for the minimization problem

$$\min_{\phi^h \in S^h} ||\sum_{k=r}^{\ell} \gamma_k \phi_{k,s} - \phi^h||^2$$

is

$$(64) \qquad ||\sum_{k=r}^{\ell} \gamma_k \phi_{k,s}||^2 - hb^T H^{-1} b$$

where

$$b = H[\sum_{k=r}^{\ell} \gamma_k h^k \delta^k + \sum_{k=r}^{\ell} \gamma_k h^k q^k].$$

Let $\beta^j = \gamma_j h^j [\delta^j + q^j]$. Then for $j,k \geq r$

$$(65) \qquad h\beta^j H \beta^k = \frac{\gamma_j \gamma_k h^{j+k+1}}{j+k+1} [(1-s)^{j+k+1} + (-s)^{j+k+1} + O(s^{j+k+2})].$$

The (j,k)-th term in $||\sum_{k=r}^{\ell} \gamma_k \phi_{k,s}||^2$ for $j,k \geq r$ is:

$$(66) \qquad \frac{\gamma_j \gamma_k h^{j+k+1}}{j+k+1} (1-s)^{j+k+1} .$$

The lemma then follows immediately by subtracting (65) from (66).

■

Theorem 5: Suppose that the constraint set $S^h = S^{q,h} \times S^{\lambda,h} \times S^{\nu,h}$ where $S^{q,h}$, $S^{\lambda,h}$, and $S^{\nu,h}$ lie in PC^1, PC^0, and PC^0 and consist of polynomials of degree $\ell+1$, ℓ, and ℓ, respectively, on the grid intervals. Also assume that (A4) is satisfied, the grid points are free parameters, and the distance between grid points is constrained to lie between h and $2h$. Then the distance between an m-th order discontinuity in u^*, the optimal

control, and the closest grid point associated with the optimal solution, u^h, to the Ritz-Trefftz problem is at most ch^p where $p = 2(\ell+1)/(2m+1)$ provided $m \leq \ell$.

Proof: Suppose that there is a discontinuity in u^* of order m at $t' \in [t_k, t_{k+1}]$ and let $||\cdot||$ denote the L^2 norm on $[t_k, t_{k+1}]$. Translate the system so that $t' = 0$ and s is as defined above. If v^h minimizes $||u^*-w^h||$ over all polynomials w^h of degree ℓ on $[t_k, t_{k+1}]$, then by (54)

$$||u^*-v^h|| \leq ||u^*-u^h|| \leq ch^{\ell+1}. \tag{67}$$

Using the lower bound for $||u^*-v^h||$ from Lemma 4,

$$O(h^{\ell+\frac{3}{2}}) + \{h^{2m+1}(s^{2m+1}+O(s^{2m+2}))\}^{\frac{1}{2}} \leq ch^{\ell+1}. \tag{68}$$

Hence

$$s^{2m+1} + O(s^{2m+2}) \leq ch^{2\ell+2-2m-1}. \tag{69}$$

Since $m \leq \ell$, then the right side of (69) converges to zero as $h \to 0$ and hence the left side of (69) converges to zero. Note that $||u^*-v^h|| = 0$ only is possible when $s = 0$ since a function with a discontinuity can never be

approximated exactly by a polynomial. Using this then it is easy to see that $s \to 0$ as $h \to 0$ and the $O(s^{2m+2})$ term is small relative to s^{2m+1}. Since the nearest grid point is a distance hs away from the origin, the estimate in the theorem follows directly. ∎

Note that if S^h is some subset of the polynomial space described in the theorem, then the result also holds since the inequality (67) is satisfied. For example, S^h could be a polynomial space with higher continuity requirements.

VII. SUMMARY AND CONCLUSIONS

In implementing the Ritz-Trefftz method numerically, the following observations were made:

1. For state constrained problems, the variable q is far superior to p since q possesses one degree more smoothness.

2. For control constrained problems, the dual function $\mathcal{L}$ is positive definite in the subspace where $q(1) = 0$ and λ lies in the space spanned by the columns of K_c.

3. For higher order Ritz-Trefftz spaces, in order to attain the full convergence rate possible in the space, the grid points in a neighborhood of a discontinuity of the dual variables should be left as free parameters.

Some advantages of this dual method over the primal method where u is restricted to lie in a subspace are

1. If the basis functions in the Ritz-Trefftz subspace are "patch bases" (i.e., zero everywhere except for a few grid intervals), then the quadratic part of $\mathcal{L}$ will possess a band structure while the quadratic part of the primal cost functional usually does not possess this structure.

2. The constraints in the dual problem, $\lambda \geq 0$ and $\overset{\circ}{\nu} \geq 0$, are in general simpler than the constraints in the primal problem. This is especially evident in non-linear problems.

Part II of this paper will analyze the numerical implementation of the Ritz-Trefftz method in some specific problems.

VIII. APPENDICES

APPENDIX 1: Interpolation From Below

Strang [9] proved a result for approximating a function of two variables, $f \geq 0$, by a piecewise linear function f^h satisfying $f \geq f^h \geq 0$. The following lemma applies to more general spaces, however, only functions of one variable are considered.

Lemma 1A: Suppose S^h contains any polynomial of degree less than p on $[0,h]$, $f: [0,h] \to [0,\infty)$, $f \in C^p$. Then there exists $f^h \in S^h$ with $f \geq f^h \geq 0$ and

$$(70) \qquad \sup_{t\in[0,h]} |f(t)-f^h(t)| \leq \sup_{s\in[0,h]} |f^p(s)| \frac{h^p}{p!} .$$

Proof: Define the set

$$F = \{g: f \geq g \geq 0 \text{ and } g \text{ is a polynomial of degree less than } p\}.$$

Define the "number of times g touches f" to be the sum of all zeroes of $f-g$ (a zero of multiplicity m is counted as m zeroes).

Let $g \in F$ touch f at ℓ times. If $\ell = p$, then the following standard result in approximation theory [11] proves the lemma:

(71) Suppose g is a polynomial of degree less than p and $f^{(j)}(t_k) = g^{(j)}(t_k)$ for $0 \leq j \leq q_k - 1$ and $k = 1,2,\ldots,m$ where $\sum_{k=1}^{m} q_k = p$. Then

$$f(t) = g(t) + \frac{f^{(p)}(\xi(t))}{p!} \prod_{k=1}^{m} (t-t_k)^{q_k}$$

where $\xi(t) \in [0,h]$.

The proof proceeds by proving the following claim:

(72) If $\ell < p$, then there exists a polynomial e of degree less than p with $e \geq 0$, $f \geq g+e$, and $g+e$ touches f at least $\ell+1$ times.

Thus by induction there exists $g \in F$ that touches f at p times so that (71) can be applied.

Proof of (72): If $\ell = 0$, then the claim is obvious. Now suppose g touches f at $t_1,\ldots,t_m$ and the multiplicity of the zero of $f-g$ at t_k is q_k; i.e., $g^{(j)}(t_k) = f^{(j)}(t_k)$ for $0 \leq j \leq q_k - 1$ and $1 \leq k \leq m$ and $\sum_{k=1}^{m} q_k = \ell$. First, it is proved that if $0 < t_k < h$, then

q_k is even.

Suppose q_k were odd. Then, expanding f in a Taylor series about t_k,

$$(73)\quad f(t) - g(t) = (f^{(q_k)}-g^{(q_k)})(t_k)\,\frac{(t-t_k)^{q_k}}{q_k!} + o(|t-t_k|^{q_k}).$$

Since q_k is odd, then the constraint $f-g \geq 0$ is violated in a neighborhood of t_k unless $f^{(q_k)}(t_k) = g^{(q_k)}(t_k)$ since $(t-t_k)^{q_k}$ changes sign at t_k. This is a contradiction.

Define

$$(74)\qquad h_\epsilon(t) = \epsilon \prod_{j=1}^{m} (t-t_j)^{q_j}$$

where $\epsilon \geq 0$. If $t_m = h$, then replace the factor $(t-h)$ by $(h-t)$ in (74). Note that $h_\epsilon \in S^h$ since it is a polynomial of degree ℓ less than p and since q_k is even for $0 < t_k < h$, then all the factors in the product in (74) are non-negative.

By a Taylor series expansion about t_k as in (73), it is seen that the constraint $f-g \geq 0$ implies that $[f^{(q_k)}(t_k)-g^{(q_k)}(t_k)](t-t_k)^{q_k} \geq 0$ for $t \in [0,h]$. Furthermore, since t_k has multiplicity q_k, then strict

inequality holds for $t \neq t_k$. Thus, for $\epsilon > 0$ small enough,

$$(75) \qquad [f^{(q_k)}(t_k)-g^{(q_k)}(t_k)-\epsilon](t-t_k)^{q_k} \geq 0$$

for $t \in [0,1]$. Hence, there exists $\bar{\epsilon}_j$ and an open ball B_j centered at t_j such that $f(t)-g(t)-h_\epsilon(t) \geq 0$ for $t \in B_j$ and $\epsilon \leq \bar{\epsilon}_j$. Now since f and g are continuous, there exists t^* solving

$$(76) \qquad \min_{\substack{t\in[0,h] \\ t\notin B_j, j=1,\ldots,m}} |f(t)-g(t)|.$$

Furthermore, since f and g only touch at $t_1,\ldots,t_m$, then the minimizing value, d, is positive.

Since $\sup_{t\in[0,h]} |h_\epsilon(t)| \leq \epsilon h^\ell$, then if we define $\bar{\epsilon}_o = d/h^\ell$, it follows that $f(t)-g(t)-h_\epsilon(t) \geq 0$ for $t \in [0,h]$, $t \notin B_j$, $j = 1,\ldots,m$, and $\epsilon \leq \bar{\epsilon}_o$.

Now let $\delta = \min\{\bar{\epsilon}_o, \bar{\epsilon}_1, \ldots, \bar{\epsilon}_m\}$. Then $f(t)-g(t)-h_\epsilon(t) \geq 0$ for $\epsilon \leq \delta$ and $t \in [0,h]$. Since $h_\epsilon \geq 0$, then $g+h_\epsilon \in F$ for $\epsilon \leq \delta$. As ϵ increases, there eventually exists $\bar{\epsilon}$ such that for $\epsilon > \bar{\epsilon}$, $g+h_\epsilon \notin F$. Hence, $g+h_{\bar{\epsilon}}$ must touch f at least $\ell+1$ times or the same procedure

as above could be applied to $g+h_{\varepsilon}$. This completes the proof. ∎

Now if f is defined on $[0,1]$, then an application of Lemma 1A to each grid interval yields:

Corollary 1A: Suppose S^h contains any polynomial of degree less than p on grid intervals of width h, $f: [0,1] \to [0,\infty)$, $f \in C^p$, then there exists $f^h \in S^h$ with $f \geq f^h \geq 0$ and

$$(77) \qquad ||f-f^h||_C \leq ||f||_{C^p} \frac{h^p}{p!} .$$

∎

Also note that since the method of proof of Lemma 1A pushes f^h up at all points until it interpolates f in the sense of (71), then the same procedure solves the general restricted range problem:

Corollary 2A: Suppose S^h contains any polynomial of degree less than p on grid intervals of width h, $f: [0,1] \to R$, $f \in C^p$, $f(t) \geq \alpha(t)$ for $t \in [0,1]$, and there exists $g^h \in S^h$ with $f \geq g^h \geq \alpha$. Then there exists $f^h \in S^h$ such that $f \geq f^h \geq \alpha$ and (77) holds. ∎

If f is monotone non-decreasing, then Lemma 1A can be used to interpolate the derivative of f and hence prove:

Corollary 3A: Suppose S^h contains any polynomial of degree less than p on $[0,h]$ where $p>1$ and $f:[0,h]\to R$ is monotone non-decreasing and lies in C^p; then there exists $u^h, \ell^h \in S^h$ with (u^h, ℓ^h) monotone non-decreasing, $\ell^h \le f \le u^h$, $\ell^h(0) = f(0)$, $u^h(h) = f(h)$, and the following estimate holds for $f^h = u^h$ or ℓ^h:

$$\sup_{t\in[0,h]} |f(t)-f^h(t)| \le \sup_{t\in[0,h]} |f^p(t)|\frac{h^p}{(p-1)!}$$

Proof: Let g^h be the interpolate of $\dot{f}$ in the space of polynomials of degree p-2 that is given in Lemma 1A. Then

$$\sup_{s\in[0,h]} |g^h(s)-\dot{f}(s)| \le \sup_{s\in[0,h]} |f^{(p)}(s)|\frac{h^{p-1}}{(p-1)!}.$$

Let $\ell^h(t) = f(0) + \int_0^t g^h(t)dt$. Then integrating $g^h-\dot{f}$ yields:

$$|\ell^h(s)-f(s)| = \left| \int_0^s g^h(t)-\dot{f}(t)\, dt \right|$$
$$\le \sup_{s\in[0,h]} |f^{(p)}(s)|\frac{h^p}{(p-1)!}$$

A similar estimate holds for u^h.

■

Again if f is defined on [0,1], then Corollary 3A applied to each grid interval yields a global approximation:

Corollary 4A: Suppose S^h contains any polynomial of degree less than p on grid intervals on width h, $p>1$, and $f:[0,1] \to R$ is monotone non-decreasing and lies in C^p; then there exists $u^h, \ell^h \in S^h$ with $u^h \geq f \geq \ell^h$, (u^h, ℓ^h) monotone non-decreasing, u^h and ℓ^h agreeing with f at the grid points on (0,1] and [0,1), respectively, and the following estimate holds for $f^h = u^h$ or ℓ^h:

$$||f-f^h||_C \leq ||f||_{C^p} \frac{h^p}{(p-1)!}$$

■

Appendix 2: Stability of Solution to Primal and Dual Quadratic Programing Problem

J. W. Daniel [3] bounds the change in the solution of a definite quadratic program in terms of the change in the data in the problem. In the development of regularity properties for the solution of the continuous optimization problem in Section 4, bounds are also needed on the change in the dual multipliers. This appendix proves that the change in both the primal solution and the dual multipliers can be bounded by the change in the data, and, unlike Daniel's results, the proof follows directly from the Kuhn-Tucker conditions.

The following quadratic programing problem is considered:

$$(P) \qquad \min\ v^T R v + r^T v$$
$$\text{s.t. } Av + a \leq 0$$
$$Bv + b = 0.$$

The matrices and vectors are all assumed of compatible size and R is assumed positive definite. Let $Q = \{R, r, A, a, B, b\}$, the data set, and let u, u^*, and u' denote the solution to the problem (P) with data Q, Q^*, and Q', respectively. If S and T denote the sets $\{s_1, \ldots, s_n\}$ and $\{t_1, \ldots, t_n\}$, respectively, then define $S \sim T = \{s_1 - t_1, \ldots, s_n - t_n\}$,

$cS = \{cs_1,\ldots,cs_n\}$, and $|S| = \max\{|s_1|,\ldots,|s_n|\}$. In our development, the elements of S will be vectors and matrices and $|\cdot|$ will be the "sup" norm. Also define $F = \{v: A^*v + a^* \leq 0,\ B^*v + b^* = 0\}$.

The theorem is based on the following mild restriction: There exists $v \in F$ and a submatrix $\bar{A}$ consisting of the rows of A^* corresponding to the binding constraints for the inequality $A^*v + a^* \leq 0$ such that

(A) if $x \geq 0$ and $x^T\bar{A} + y^TB^* = 0$, then $x = y = 0$.

Robinson [6] introduces this assumption in the study of the effects of perturbations on the solution set for systems of equations and inequality restrictions. He and Meyer [7] prove that if (A) holds for <u>some</u> $v \in F$, then (A) holds for <u>all</u> $v \in F$ and furthermore [6] there exists $w \in F$ such that $A^*w + a < 0$. Thus if (A) holds for the data (A^*,a^*,B^*), then it also holds for data in a neighborhood since $Aw + a < 0$ and the rows of B are linearly independent for $|A-A^*|$, $|B-B^*|$, and $|a-a^*|$ sufficiently small.

<u>Theorem 1A</u>

Suppose (A) holds. Then there exists a constant D such that whenever $|Q\sim Q^*| \leq D$ and $|Q'\sim Q^*| \leq D$, the solution

(u,u') and Kuhn-Tucker vector (p,p') corresponding to the data (Q,Q') satisfies the estimate:

(1) $$|u-u'| \leq c|Q-Q'| \quad \text{and} \quad |p-p'| \leq c|Q-Q'|$$

where D depends on $Q^* - \{r^*\}$ and c depends on Q^*.

As Daniel and others note, it is easy to prove that the solution to (P) depends continuously on the data for D sufficiently small so we will assume continuous dependence and then prove the Lipschitz estimate (1). The method of proof is identical with or without the presence of equality constraints so to keep the notation less cumbersome, it will be assumed that the equality constraints are vacuous.

If J is a collection of row numbers, then let $A(J)$ denote the submatrix of A consisting of those rows corresponding to elements of J. Also define

$$c(A,a) = \{J: \ J \text{ is a possible collection of binding constraints for the region defined by } Av + a \leq 0\}.$$

Recalling the comments preceding the statement of the theorem, it is possible to choose D sufficiently small that

$|Q \sim Q^*| \leq D$ implies Q satisfies (A), R is positive definite, and the solution of (P) corresponding to the data Q is a continuous function of the data in the ball of radius D about Q^*. The Theorem is based on the following simple result:

Lemma 2A

The optimal value, m, of the following optimization problem is positive:

$$(2) \qquad \min \left| \begin{pmatrix} R & A(J)^T \\ A(J) & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} \right|$$

$$\text{s.t.} \quad |(y^T, z^T)| = 1$$

$$z \geq 0$$

$$J \in C(A, a)$$

$$|Q \sim Q^*| \leq D.$$

Proof of Lemma

This property is proved by contradiction. Suppose that there exists a feasible sequence z^k, y^k, R^k, A^k, a^k, J^k, such that the objective function in (2) converges to zero. Since J^k is a subset of a finite set, then there exists a

subsequence (also subscripted by k) such that $J^k = J$, a fixed set. Since the sequence lies in a compact set, then there exists a subsequence (also subscripted by k for convenience) converging to $S = (z,y,R,A,a)$ and by the continuity of the extremand, the objective function vanishes for S. Thus

$$(3) \qquad Ry + A(J)^T z = 0$$

$$(4) \qquad A(J)y = 0.$$

Since the constraints in (2) are continuous, then S satisfies these conditions and by the construction of D, R is positive definite. Solving (3) for y and inserting the result into (4) implies that $z^T A(J) R^{-1} A(J)^T z = 0$ and since $R > 0$, then $A(J)^T z = 0$. Also by the construction of D, (A,a) are close enough to (A^*,a^*) that condition (A) holds and hence $z = 0$. Finally by (3), $z = 0$ implies $y = 0$ which violates the constraint that $|(y^T,z^T)| = 1$.

∎

Proof of Theorem

Consider the quadratic program with data $Q(s) = sQ + (1-s)Q'$ where $|Q-Q^*| \leq D$, $|Q'-Q^*| \leq D$, and $s \in [0,1]$ and

let $u(s)$, $p(s)$, and $J(s)$ denote the solution, Kuhn-Tucker vector, and binding constraint set for this problem. The necessary conditions are then:

$$(5) \qquad \begin{bmatrix} R(s) & A(s)(J(s))^T \\ A(s)(J(s)) & \end{bmatrix} \begin{bmatrix} \dot{u}(s) \\ p(s)(J(s)) \end{bmatrix} = \begin{bmatrix} -r(s) \\ -a(s)(J(s)) \end{bmatrix}.$$

The equation (5) is of the form $M(s)y(s) = f(s)$ where the following conditions hold whenever $s,t \in [0,1]$ and both $M(s)$ and $M(t)$ are of compatible size: $|M(s)y| \geq m|y|$, $|M(s)-M(t)| \leq d|s-t|$, and $|f(s)-f(t)| \leq d|s-t|$ where m is the optimal value for (2) and $d = 2|Q \sim Q'|$. Defining $M_k = M(s_k)$, $y_k = y(s_k)$, and $f_k = f(s_k)$, then the following bound holds:

$$(6) \qquad |y_1-y_2| \leq cd|s_1-s_2|$$

when c depends on Q^*. This is proved as follows:

$$A_1y_1 - A_2y_2 = f_1 - f_2$$

$$(A_1-A_2)y_1 + A_2(y_1-y_2) = f_1 - f_2$$

$$m|y_1-y_2| \le |f_1-f_2| + |A_1-A_2|\,|y_1|$$

$$\le d|s_1-s_2| + |A_1-A_2|\,|f_1|/m$$

$$\le d|s_1-s_2| + (|f(0)|+D)d|s_1-s_2|/m$$

$$\le d|s_1-s_2|(1 + (|a^*|+|r^*|+2D)/m).$$

The estimate (1) is now at hand. Define $s_0 = 0$ and s_k by:

$$s_k = \sup\ s$$

$$\text{s.t. } s_{k-1} \le s \le 1$$

$$J(s) = J(s_k^+).$$

Eventually there exists s_N with $s_N = 1$. By assumption $u(s)$ is a continuous function of s for $s \in [0,1]$ so if $p(\cdot)$ is also continuous then (1) follows immediately from (6) since

$$(u^T, p^T(J(0))) = y(0)^T,\ (u'^T, p'^T(J(1))) = y(1)^T, \text{ and}$$

$$|y(0)-y(1)| \le \sum_{j=1}^{N} |y_{j-1}-y_j| \le cd \sum_{j=1}^{N} |s_{j-1}-s_j| = cd.$$

The continuity of $p(\cdot)$ again follows from the lemma above: Suppose that there exists a sequence s_k converging to s from the left with J^- being the binding constraint set for all k. Since $u(\cdot)$ is continuous, then $J^- \subset J \equiv J(s)$ and $(A(s)u(s^-) + a(s))_j = 0$ for all $j \in J$. Also by the complementary slackness condition that $p(\cdot)$ satisfies, $p(s_k)_j = 0$ for $j \notin J^-$. Suppose that $p(s_k)$ does not converge to $p(s)$. Since $M(s_k)y(s_k) = f(s_k)$, then $|y(s_k)| \leq |f(s_k)|/m$ by the lemma and since $f(s_k)$ is uniformly bounded, then there exists a subsequence of the $y(s_k)$ converging to $y^- \neq y(s)$. Note that since the binding constraint sets are identical for all k, then $M(s_k)$ and $f(s_k)$ converge to limits M^- and f^- and $M^-y^- = f^-$. Now M^- and $M(s)$ are identical except that $M(s)$ may have a few extra rows and columns. If y^- is augmented with zeroes corresponding to those extra columns, then the augmented vector, z, satisfies $M(s)z = f(s)$ since for those extra rows in $M(s)$ corresponding to $j \in J\text{-}J^-$, the equality $(A(s)u(s^-)+a(s))_j = 0$ holds. The matrix $M(s)$ is positive definite by the lemma and so the system $M(s)y = f(s)$ has a unique solution. Thus $z = y(s)$ which is a contradiction.

■

BIBLIOGRAPHY

1. J. P. Aubin, Approximation of Elliptic Boundary Value Problems, Interscience, John Wiley and Son, New York, 1972

2. W. E. Bosarge Jr., et.al., "The Ritz-Galerkin Procedure for Non-Linear Control Problems", SIAM J. Numer. Anal. 10(1973), pp 94-111

3. J. W. Daniel, "Stability of the Solution of Definite Quadratic Programs," Math. Programing, 5(1973), pp41-53

4. W. W. Hager, "Duality Theory for Convex Control Problems", to appear

5. H. Lewy and S. Stampacchia, "On the Regularity of the Solution of a Variational Inequality," Comm. Pure and Appl. Math., 22(1969), pp 153-188

6. S. M. Robinson, "Perturbations in Finite-Dimensional Systems of Linear Inequalities and Equations," Mathematics Research Center, University of Wisconsin-Madison, Report 1357, June, 1973.

7. S. M. Robinson, and R. R. Meyer, "Lower Semicontinuity of Multivalued Linearization Mappings," SIAM J. Control., 11(1973), pp 525-533.

8. G. Strang and G. Fix, An Analysis of the Finite Element Method, Prentice-Hall, Englewood Cliffs, New Jersey, 1973.

9. G. Strang, "One Sided Approximation and Plate Bending", The International Symposium on Computing Methods, IRIA, Rocquencourt, France, 1973.

10. U. Mosco and G. Strang, "One Sided Approximation and Variational Inequalities," Bul. AMS, 80(1974), p.308-312.

11. J. F. Traub, Iterative Methods for Solutions of Equations, Prentice-Hall, Englewood Cliffs, New Jersey, 1964, p244.

CHAPTER 3

Numerical Examples for the Ritz-Trefftz Method with Fixed Grid Points

I. INTRODUCTION

In part I of this paper it was observed that for a class of optimal control problems with inequality state and control constraints, the dual multitpliers q corresponding to state dynamics, λ corresponding to the inequality control constraints, and ν corresponding to inequality state constraints were contained in the spaces H^2, H^1, and H^1 respectively. If the dual problem was solved using the Ritz method, the dual multipliers were restricted to a piecewise polynomial space of degree r+1, r, and r in q, λ, and ν respectively, and the grid points in the space were fixed, then it was observed that the Ritz-Trefftz approximation to the optimal dual solution was in general accurate to order at most 3/2. The error bounds were given in the H^1 norm for the q variable and the H^0 norm for the λ and ν variable.

Let the term "static region" refer to the time intervals away from the region where the constraints change from binding to non-binding. Analogously the "dynamic region" denotes the time intervals in a neighborhood of the points where constraints change from binding to non-binding. Two simple problems, a state and a control constrained problem, were studied to determine the answers to the following questions:

(Q1) Does the solution to the Ritz-Trefftz problem converge at a rate greater than 3/2 in the static region for higher order spaces?

(Q2) At what rate does the Ritz-Trefftz solution converge to the continuous solution in the sup norm both in the static and the dynamic region.

Computationally it was found that in the control constrained problem, the Ritz-Trefftz method converged at order bounded by 2 in the static region instead of 3/2 . The error in the finite dimensional solution decreased from $O(h)$ in the dynamic region to $O(h^2)$ within a few grid intervals. For state constrained problems, on the other hand, the L^2 error was $O(h^{3/2})$ even in the static region. This difference in the convergence behavior reflects the difference in the type of dual constraints for the two problems: $\lambda_k \geq 0$ for the control constrained problem and $\nu_{k+1}-\nu_k \geq 0$ for the state constrained problem. The latter constraint tends to spread out errors while the former tends to localize errors.

In the sup norm, the Ritz-Trefftz solution converged at order 1 in the dynamic region and in the static region converged at order bounded by 3/2 in the state constrained problem and 2 in the control constrained problem.

II. PROBLEM DESCRIPTION.

The following two problems were studied:

$$(P1) \qquad \min \ \tfrac{1}{2} \int_0^1 u^2(t)dt$$

$$\text{s.t. } \dot{x}(t) = u(t)$$

$$x(0) = 0$$

$$x(t) \geq -1/4 + \sin(\pi t)$$

$$(P2) \qquad \min \ \tfrac{1}{2} \int_0^1 u^2(t)dt$$

$$\text{s.t. } \dot{x}(t) = u(t)$$

$$x(0) = 0$$

$$x(1) = 1/6 + \sqrt{3}/\pi$$

$$u(t) \geq \sin(\pi t)$$

where $u:[0,1] \to R$ and $x:[0,1] \to R$. The constant $1/6 + \sqrt{3}/\pi$ will be denoted by c^* for convenience. The corresponding dual problems are:

$$(D1) \qquad \max \ -\tfrac{1}{2} \int_0^1 \nu^2(t)dt + \int_0^1 (\sin(\pi t) - 1/4)\, d\nu$$

$$\text{s.t. } \nu(1) = 0, \ \nu \text{ non decreasing}$$

$$\nu \text{ of bounded variation}$$

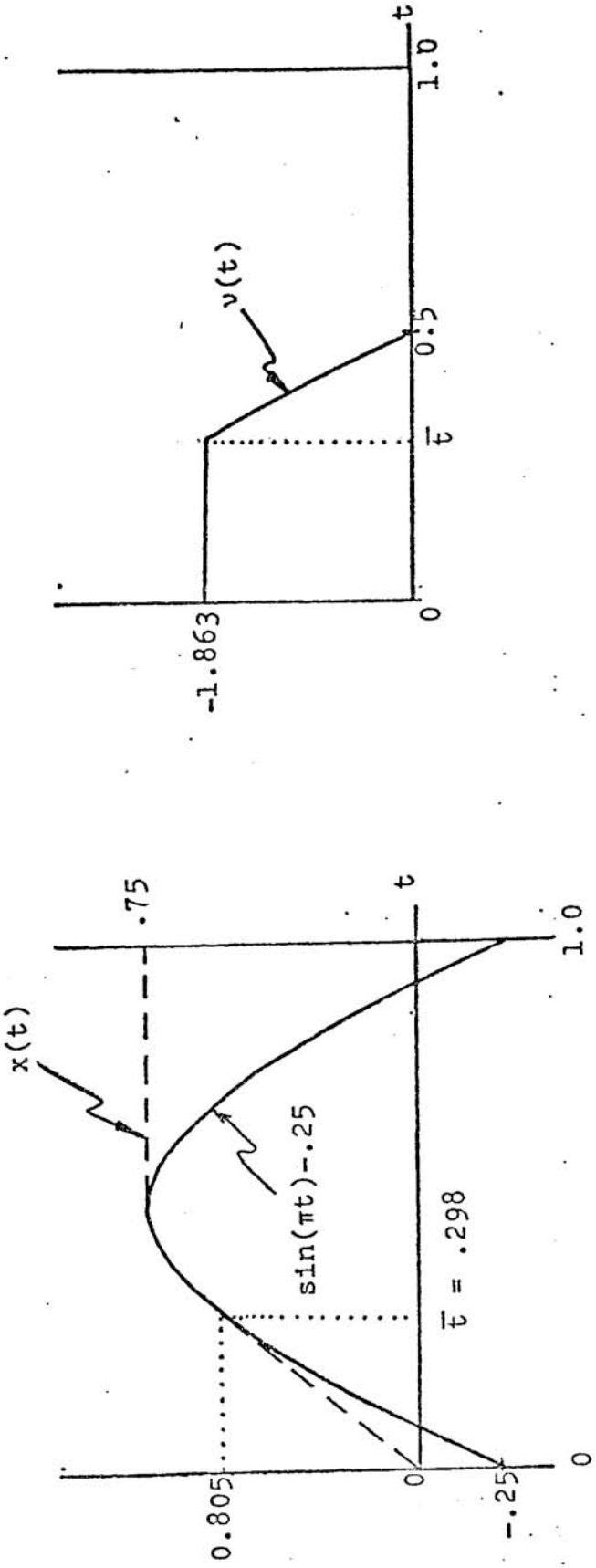

Figure 1: The Exact Solution to Problem P1.

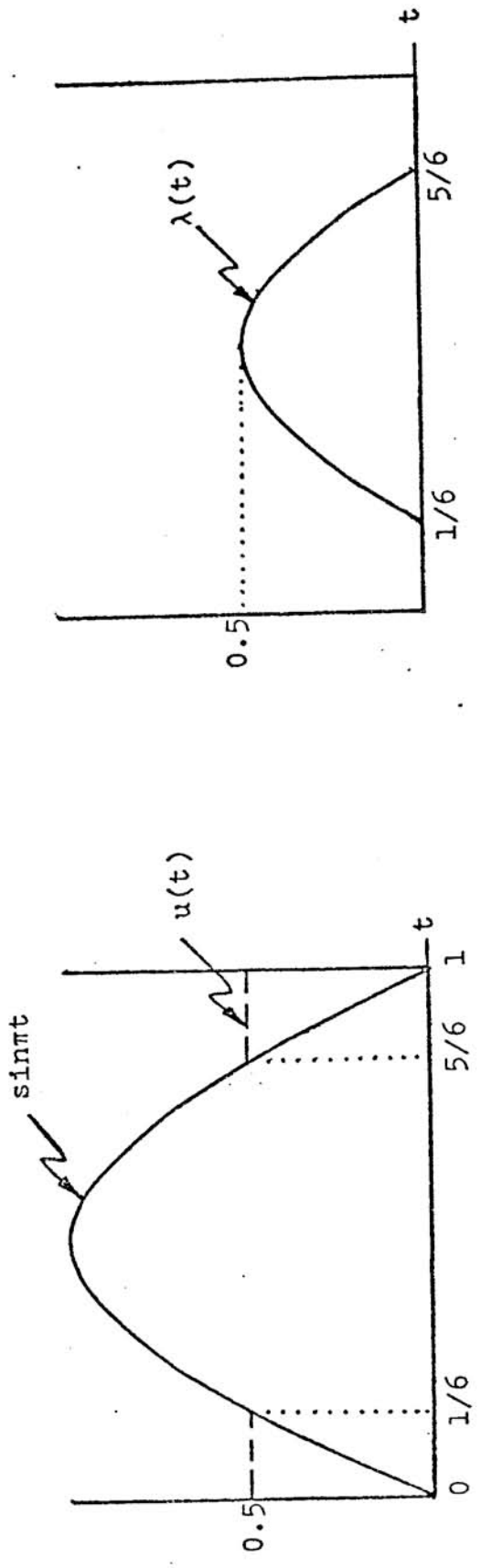

Figure 2: The Exact Solution to Problem P2.

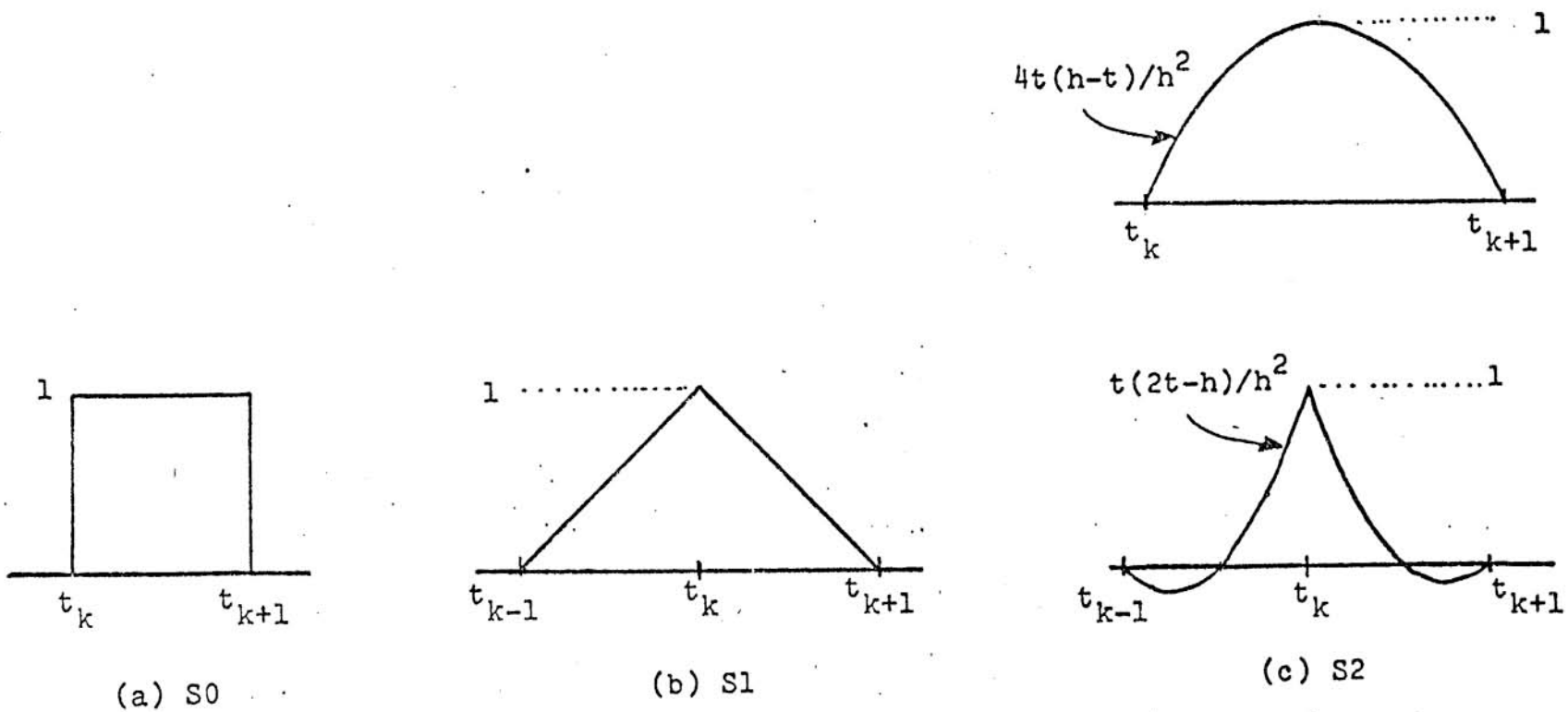

Figure 3: The Basis Functions for the Spaces S0,S1, and S2.

$$(D2) \qquad \max \int_0^1 -\tfrac{1}{2}(\lambda(t)-q)^2 + \lambda(t)\sin(\pi t)dt - c*q$$

$$\text{s.t.} \quad \lambda \geq 0$$

$$\lambda \in L^1$$

where $q \in R$, $\lambda:[0,1] \to R$, and $\nu:[0,1] \to R$. The exact solutions to these problems are shown in Figures 1 and 2.

The spaces Sk for k=0,1, and 2 consisting of polynomials of degree k on each grid interval that are continuous at the grid points were utilized in the analysis. The basis functions for S0, S1, and S2 are shown in Figure 3. Note that S2 consists of two characteristic basis functions.

III. CONVERGENCE RESULTS

As noted in part I of this paper, the rate of convergence of the Ritz-Trefftz solution to the continuous solution depends very strongly on the distance between the grid points and the points where binding constraints change to non-binding constraints in the optimal dual solution. These latter points will be referred to as the break points. In the computational experiments, the grid points were always

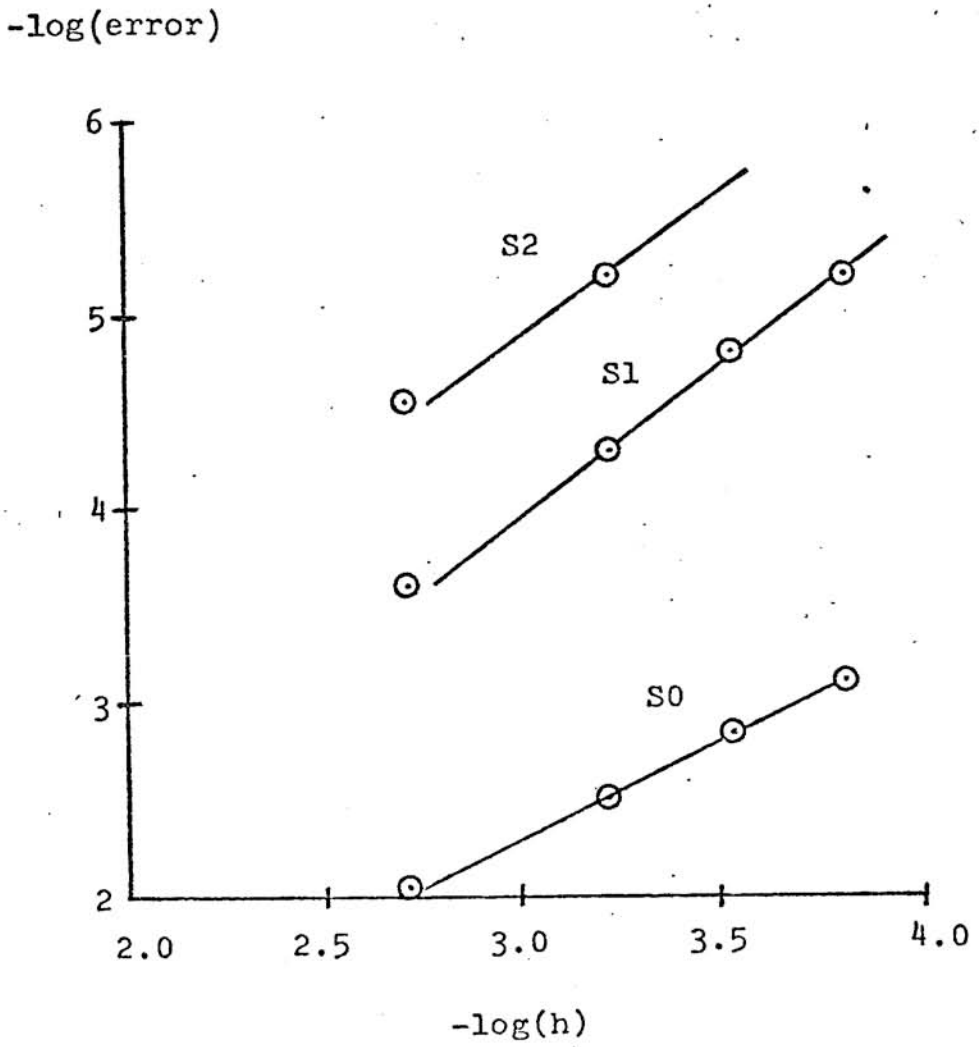

Figure 4: The Log of the L^2 Error of v^h as a Function of Log(h).

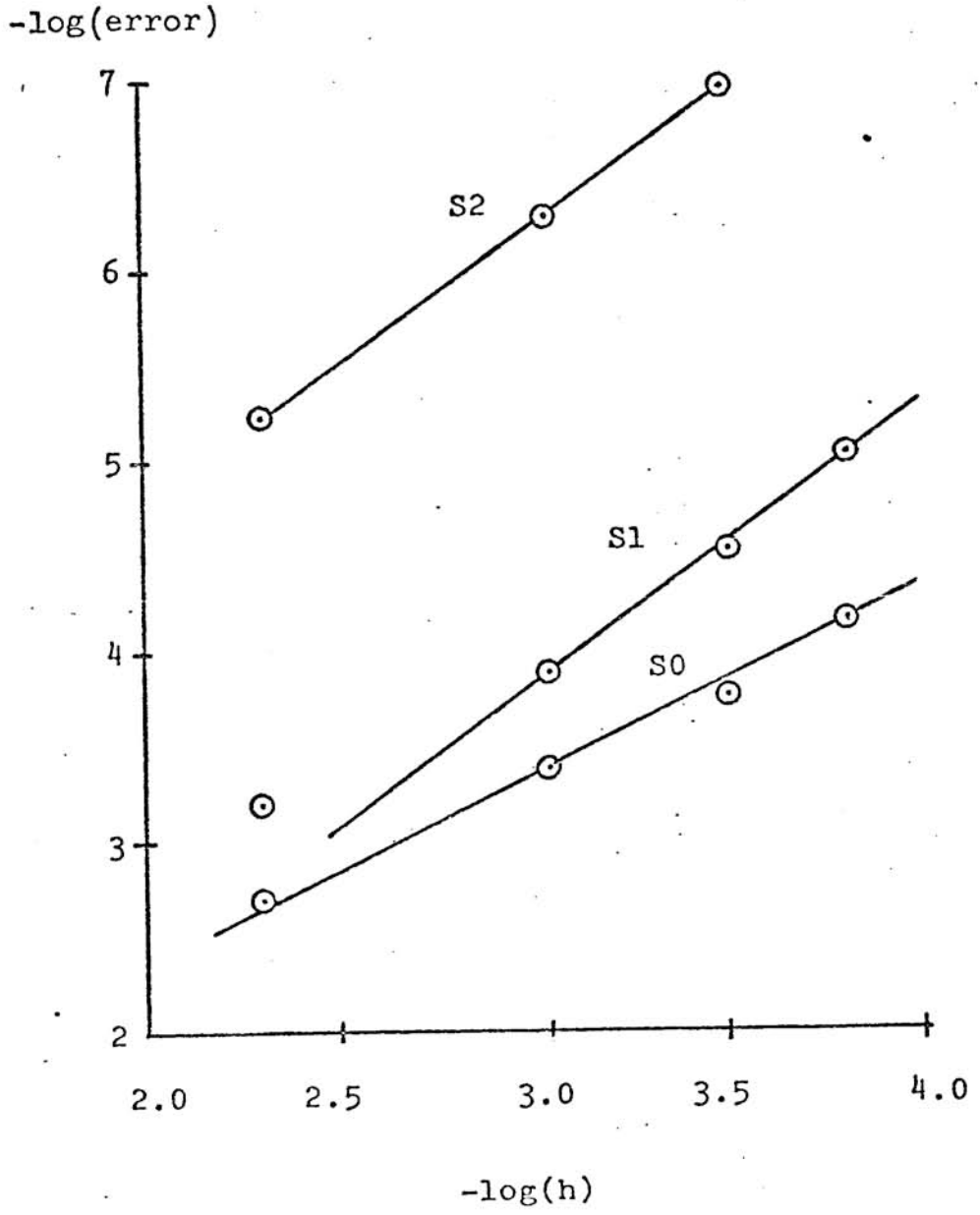

Figure 5: The Log of the L^2 Error of λ^h as a Function of Log(h).

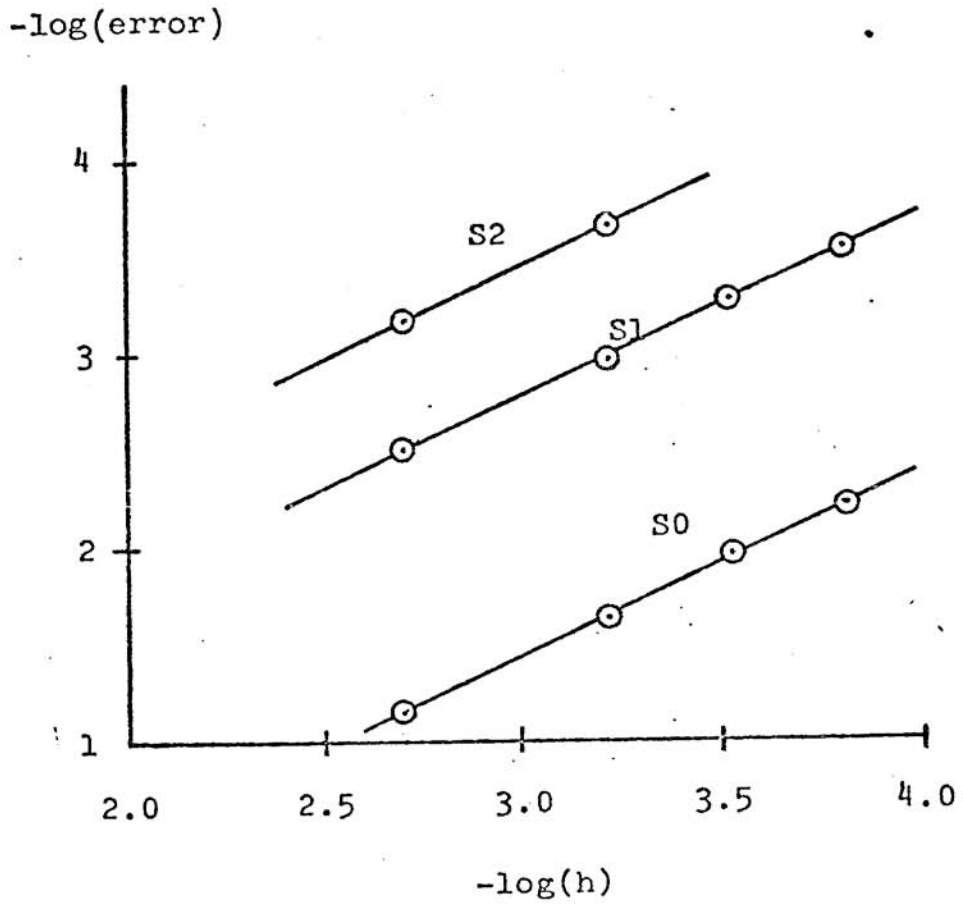

Figure 6: The Log of the Sup Error of v^h as a Function of Log(h).

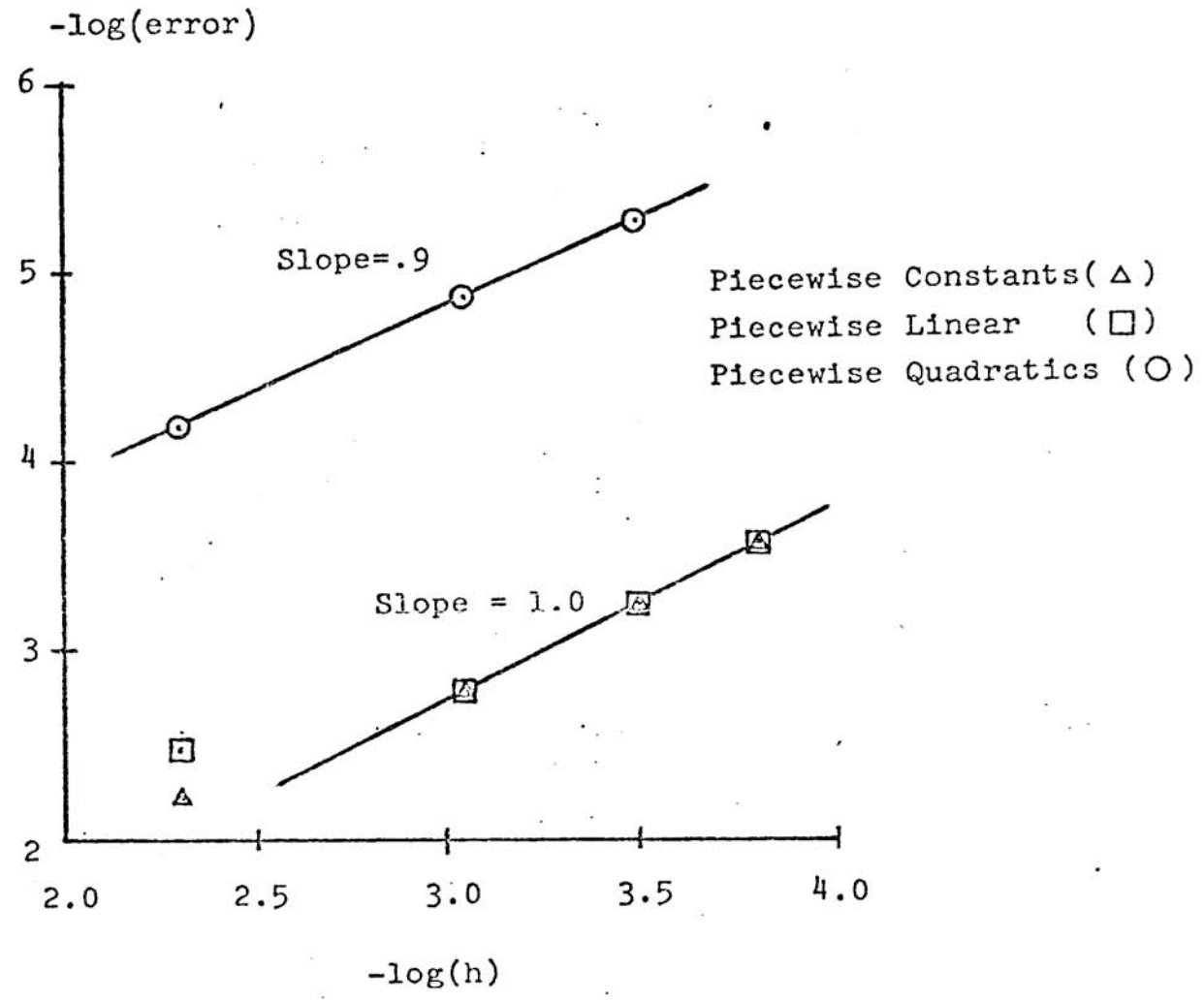

Figure 7: The Log of the Sup Error of λ^h as a Function of Log(h).

	S0	S1	S2
P1	1.0	1.5	1.5
P2	1.0	1.5	1.4

TABLE I: Convergence Rates for λ^h and ν^h in the L^2 Norm

chosen so that the break points occurred almost exactly in the middle of a grid interval.

Let ν^h and (λ^h, q^h) denote the optimal Ritz-Trefftz solution corresponding to problems (D1) and (D2) respectively where h is the grid interval in the finite dimensional subspace. Note that the Ritz-Trefftz approximation u^h to the control is given by $u^h = -\nu^h$ and $u^h = \lambda^h - q^h$ in problems (P1) and (P2) respectively. The rate of convergence of λ^h to λ^* and u^h to u* where u* and λ^* are the continuous solutions to (P2) and (D2) respectively were observed to be identical. In the graphs that follow, the convergence rates were presented in terms of the variables ν^h and λ^h.

In Figures 1 and 2 it is seen that the exact solutions to (D1) and (D2) have the property that the dual constraints are non-binding in the middle of the interval and binding on the ends of the interval. The Ritz-Trefftz solution exhibits the same type of behavior for both problems and the region of binding dual constraints in the finite dimensional problem agreed with the region of binding constraints in the continuous problem to within one grid interval.

The log of the L^2 error of the Ritz-Trefftz solution is plotted as a function of the log of the grid interval in Figures 4 and 5. Note that nearly linear graphs are

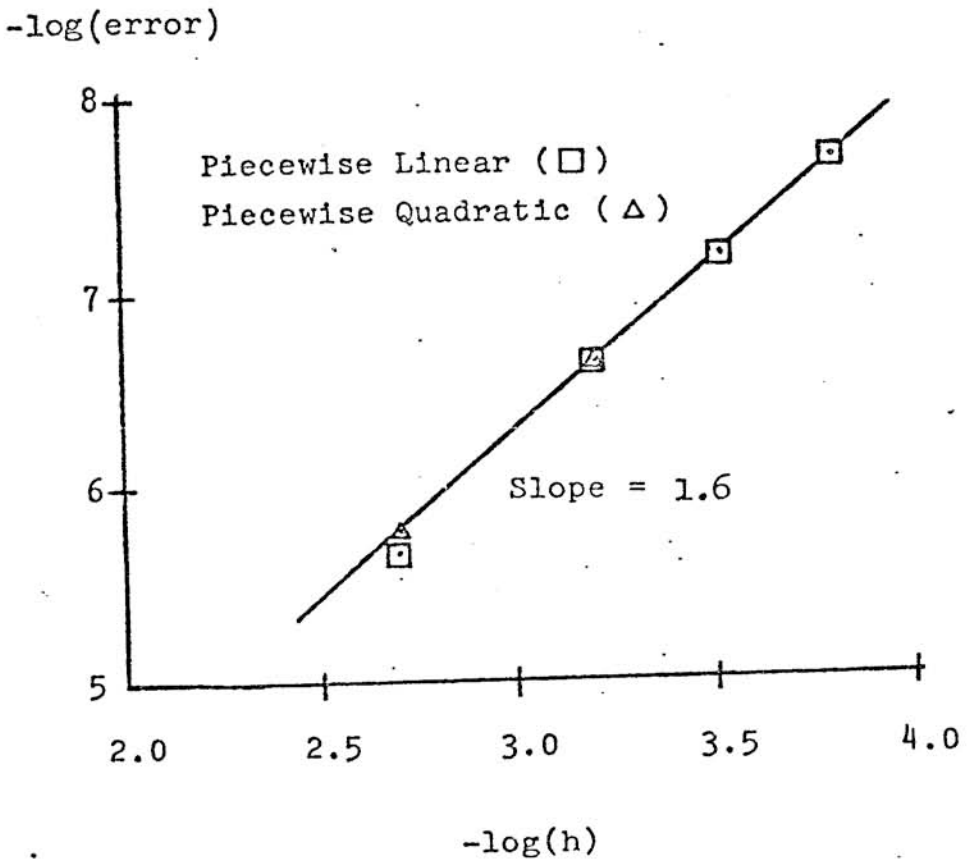

Figure 8: The Log of the Error in v^h at $t=0$ as a Function of Log(h).

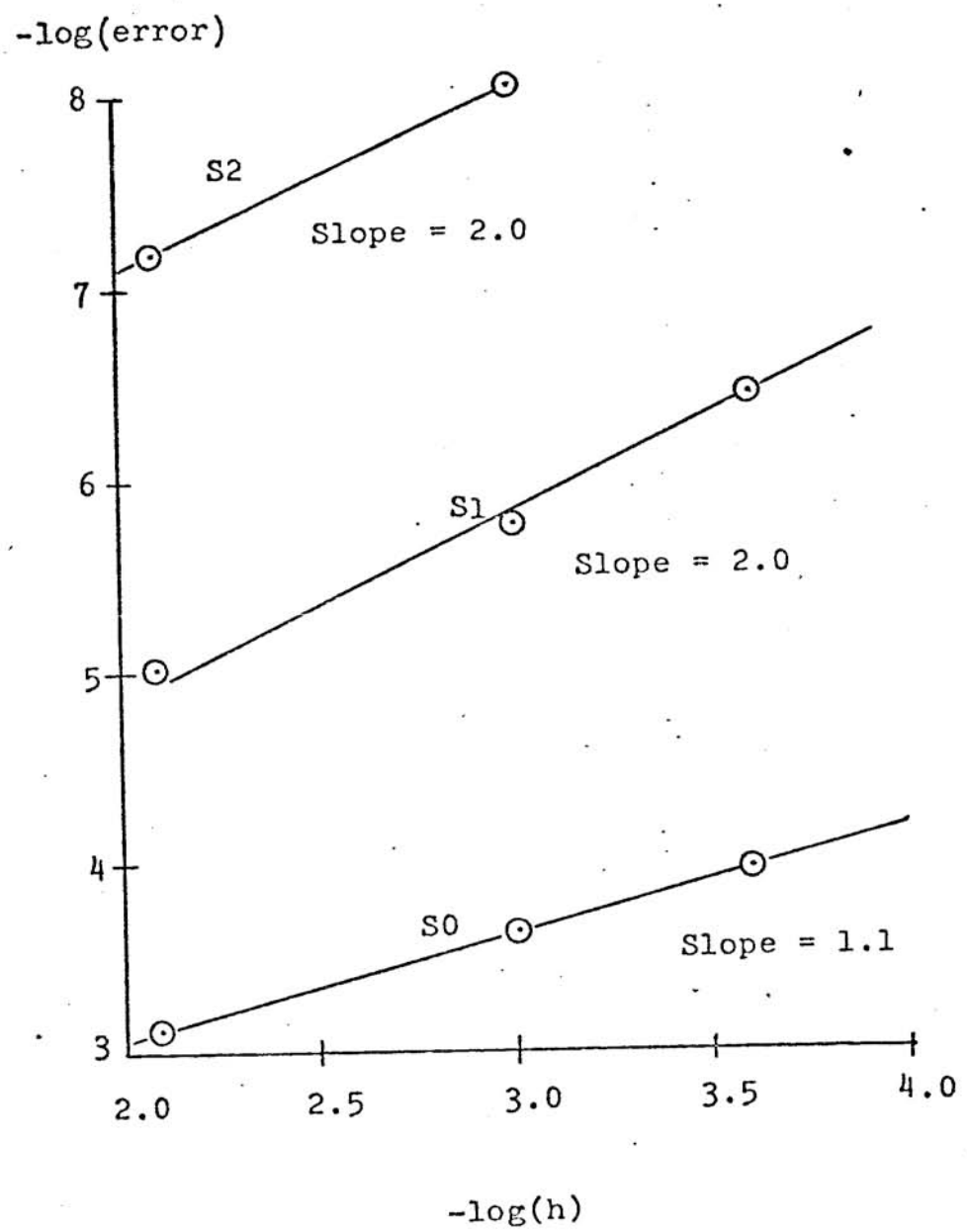

Figure 9: The Sup Error in λ^h a Distance $\frac{1}{6}$ From the Break Points.

obtained which indicates that the error bounds are tight. The asymptotic convergence rate for the data in these figures is then listed in Table I. The error bound of $h^{3/2}$ is exactly as predicted. The optimal cost for the finite dimensional problem also converged to the optimal cost for the continuous problem at rate 2 for the space S0 and 3 for the spaces S1 and S2 which again was expected.

The errors in the Ritz-Trefftz solution in the sup norm are given in Figures 6 and 7. For the spaces S1 and S2 , the error in the sup norm must converge at rate at least 1.0 or the $h^{3/2}$ error in the L^2 norm would be violated. Figures 6 and 7 indicate the 1.0 is indeed the convergence rate obtained.

The error in ν^h at $t=0$ is plotted in Figure 8. Since ν^h and ν^* are constant on the interval [0,.25], then the error at $t=0$ must converge at rate at least 3/2 for the spaces S1 and S2 or the global convergence rate of 3/2 in the L^2 norm would be violated. Figure 8 shows that this is indeed the convergence rate observed.

The error in λ^h in the static region is plotted in Figure 9 and the convergence rate, indicated in the figure, is 2 for both the space S1 and S2 . The reason for the difference in the behavior in (P1) and (P2) lies in the form of the necessary conditions for the two problems. These conditions are now analyzed in detail.

Consider the space S1 and expand λ^h in terms of the

basis functions g_k where g_k is as shown in Figure 3b; i.e.

$$\lambda^h = \sum_{k=0}^{N} \lambda_k g_k$$

where N is the number of grid intervals.

Let a_k denote the Kuhn-Tucker multiplier vector corresponding to the dual constraint $\lambda_k^h \geq 0$. The necessary conditions for the Ritz-Trefftz problem are then

(1) $$a + H\lambda^h + eq^h + b = 0$$

(2) $$he^T\lambda^h - q^h = c^*$$

$$\lambda^h \geq 0 \ , \ a \geq 0 \ a^T\lambda^h = 0$$

where

$$H_{jj} = -\frac{2}{3} \quad \text{for} \quad j \neq 0,N$$

$$H_{00} = H_{NN} = -\frac{1}{3}$$

$$H_{j+1,j} = H_{j,j+1} = -\frac{1}{6}$$

$$b_j = [2\sin(\pi t_k) - \sin(\pi t_{k-1}) - \sin(\pi t_{k+1})]/h^2\pi^2 \quad \text{for} \quad j \neq 0,N$$

$$b_0 = b_N = [h - \sin(\pi h)/\pi]/h^2\pi$$

$$e_j = 1 \quad \text{for} \quad j \neq 0,N$$

$$e_0 = e_N = .5$$

Suppose $\bar{\lambda}^h = [\lambda_k^h, \ldots, \lambda_m^h]$ are the components of λ^h corresponding to the non-binding constraints. Then $a_k = \ldots = a_m = 0$ and hence

$$\bar{H}\,\bar{\lambda}^h + \bar{e}q^h + \bar{b} = 0 \tag{3}$$

where $\bar{H}$, $\bar{e}$, and $\bar{b}$ are the appropriate blocks from H, e, and b. Now the optimal dual solution λ^*, q^* satisfies (1) and (2) to second order except for the k^{th} and the m^{th} equation which are only satisfied to first order since $\dot{\lambda}^*$ is discontinuous. If we define $\bar{\lambda}^* = [\lambda^*(t_k), \ldots \lambda^*(t_m)]$, then

$$\bar{H}\,\bar{\lambda}^* + \bar{e}q^* + \bar{b} = 0(h^2) + 0(h)d \quad \text{where} \tag{4}$$

$$d=[1,0,\ldots,0,1]$$

It can be shown that the sup norm of $\bar{H}^{-1}$ (maximum row sum) is bounded by 3.0 (independent of h). Subtracting (4) from (3) and using this bound yields:

$$\bar{\lambda}^h = \bar{\lambda}^* - \bar{H}^{-1}\,\bar{e}\,(q^h - q^*) + 0(h^2) + 0(h)\bar{H}^{-1}d \tag{5}$$

Substituting (5) in (2) produces:

$$\bar{e}^T\bar{\lambda}^* - h\bar{e}^T\bar{H}^{-1}\bar{e}(q^h - q^*) + 0(h^2) - q^h + 0(h^2)\bar{e}^T\bar{H}^{-1}d = c^* \tag{6}$$

By the continuous necessary conditions:

$$q^* = \int_0^1 \lambda^*(t)dt - c^* \quad . \tag{7}$$

By approximation theory

$$(8) \qquad \int_0^1 \lambda^*(t)dt = h\bar{e}^T\bar{\lambda}^* + O(h^2)$$

Combining (6), (7), and (8):

$$(9) \qquad (h\bar{e}^T\bar{H}^{-1}\bar{e} + 1)(q^h - q^*) = O(h^2) + O(h^2)\bar{e}^T\bar{H}^{-1}d$$

Now it can be shown that $\bar{H}^{-1}\bar{e}$ is a vector with entries bounded uniformly in h so that

$$O(h^2)\bar{e}^T\bar{H}^{-1}d = O(h^2) .$$

Also the computational results on the support of λ^h mentioned at the very beginning of this section reveal that $m - k \sim (2/3)N$. It can then be shown that $h\bar{e}^T\bar{H}^{-1}\bar{e}$ is bounded uniformly in h from -1 and hence (9) implies that $|q^h - q^*| = O(h^2)$.

Since $\bar{H}^{-1}\bar{e}(q^h - q^*) = O(h^2)$, the predominant error term in (5) is the $h\bar{H}^{-1}d$ term. Define $f = (1,0,\ldots,0)$. If $\bar{H}Y = f$ is solved by Gaussian elimination, then after eliminating the subdiagonal terms, the k^{th} row of the system of equations converges to:

$$.621Y_k + .167Y_{k+1} = (.536)^k .$$

Thus $Y_k \sim (.536)^k$ for k large. The contribution to the error in $\bar{\lambda}^h$ from the $hH^{-1}d$ term is then of order h^2 when $(H^{-1}d)_k \sim (.536)^k \sim h$. For $h = 1/33$, then $k \sim 5$.

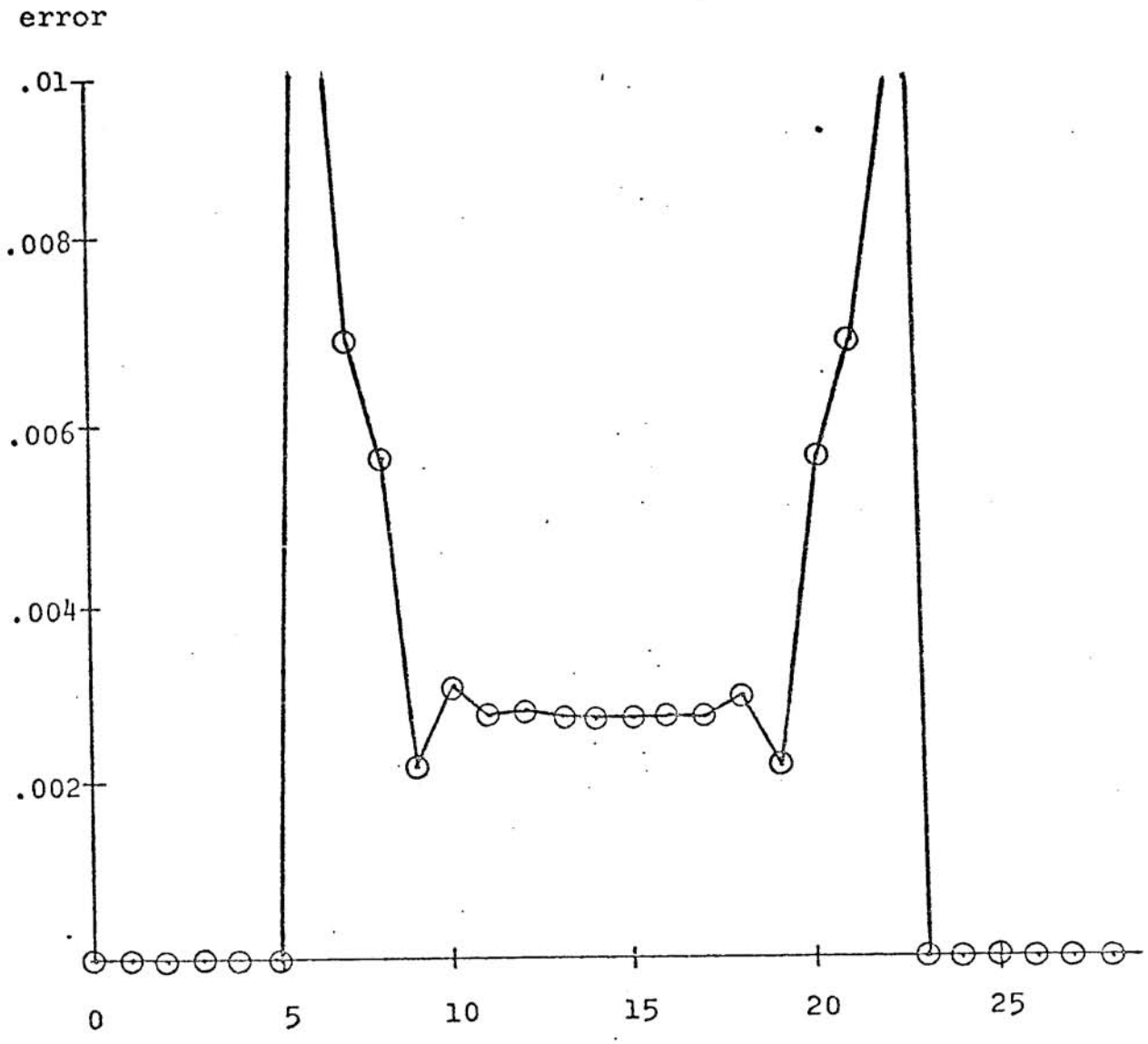

Figure 10: Error in λ^h as a Function of Time for $h=1/33$.

Figure 10 shows that indeed the error in λ^h decreases to $O(h^2)$ in about 5 grid intervals.

One immediately asks, "why does not the Ritz-Trefftz problem corresponding to S2, the space of piecewise quadratics, converge at order 3 in the static region instead of 2?" The first part of the argument above works equally well for the piecewise quadratic space since λ^* satisfies the discrete necessary conditions to order 3 except for the intervals containing the break points where an error term of order 1 arises. Also the latter part of the argument works since the "max row sum" norm of $\bar{H}^{-1}$ can be bounded independent of h and the $\bar{H}^{-1}\bar{d}$ term in (5) results in a vector whose entries decay quickly to zero. Unfortunately, however, q^h only approximates q^* to order 2 and hence the $H^{-1}\bar{e}(q^h-q^*)$ error term in (5) is dominating away from the break points.

If we tried to prove that $|q^h-q^*| = O(h^3)$ using a proof as above, the difficulty would arise at (8) since if λ^I is the interpolate of λ^* in S2, then

$$\int_0^1 (\lambda^*(t) - \lambda^I(t))dt = O(h^2)$$

instead of the $O(h^3)$ which would be required for cubic convergence of q^h.

The critical factor that permitted the higher convergence rate away from the break points in the problem (D2) was that the necessary conditions uncoupled so that the equations k,...m could be studied independently. In the state constrained problem, the constraints $\nu_{k+1} - \nu_k \geq 0$ lead to a system that doesn't uncouple and qualitatively the errors appear to be smeared out rather than localized.

CHAPTER 4

Rates of Convergence for Discrete Approximations to Unconstrained Control Problems

I. INTRODUCTION

Discrete approximations of the following unconstrained control problem (P) are studied:

$$(1) \qquad \min\ (x)_0(1)$$
$$\text{s.t. } \dot{x}(t) = f(x(t),u(t))$$
$$x(0) = x_0$$

where $x\colon R \to R^n$, $u\colon R \to R^m$, $f\colon R^n \times R^m \to R^n$, and $(x)_0(\cdot)$ is the 0-th component of the vector $x(\cdot)$.

If the problem (P) has a solution (x^*,u^*), then there exists a function p^* generated by the adjoint equation and the following necessary conditions (N) hold:

$$(2) \qquad \dot{x}^*(t) = f(x^*(t),u^*(t))$$
$$(3) \qquad \dot{p}^*(t) = -f_x(x^*(t),u^*(t))^T p^*(t)$$
$$x(0) = x_0$$
$$p(1) = (1,0,0,\ldots,0)$$
$$(4) \qquad f_u(x^*(t),p^*(t))^T p^*(t) = 0.$$

Solving the conditions (N) is equivalent to solving a two-point boundary value problem since (4) specifies $u^*(t)$ as a function of $x^*(t)$ and $p^*(t)$ while the differential equation has n conditions specified at one end and n at the

other end.

The solution of two-point boundary value problems has been studied very thoroughly in the literature (see [5] and [7]); however, the author believes it to be far superior numerically to discretize the original problem (P) directly rather than discretize the necessary conditions. The problem (P) has the advantage that gradients of the cost function are easy to compute and hence algorithms such as steepest descent and conjugate gradients are easily applied. The principal disadvantage to the solution of the two point boundary value problem is that many algorithms require the inversion of a transition matrix for the linearized system at $t = 1$ and the condition number of this matrix becomes exponentially small as a function of time.

This paper derives upper bounds on the convergence rates for the solution of a discrete approximation of (P) to the continuous solution. Onestep, multistep, and Taylor series schemes are analyzed. It was discovered that the convergence rate of the onestep procedures depended on the behavior of the scheme at the end of each grid interval while the convergence rate of the multistep schemes was determined by its behavior at the right end of the interval $[0,1]$. The Taylor series approximate, on the other hand, diverged from the continuous solution.

The upper bounds are derived by considering the relationship between the discrete necessary conditions and the continuous necessary conditions; and in fact, these upper bounds are the actual convergence rates observed in the numerical examples studied at the end of this paper. The convergence theory is then used to prove the optimality of three Range-Kutta schemes and explain the bad convergence property of a fourth Range-Kutta scheme and two Milne schemes. The optimality results show that if the problem (P) is to be solved by the finite element method over the space of controls that are piecewise quadratic polynomials and the differential equation is to be integrated using a onestep scheme based on quadrature, then the integration procedure must be at least fourth order or the third order convergence associated with piecewise quadratic polynomial spaces will not be achieved.

II. RELATIONSHIP BETWEEN CONTINUOUS AND DISCRETE NECESSARY CONDITIONS

Assuming that the requirements needed for the continuous [6] and the discrete [1] minimum principal are satisfied, the relationship between these two sets of necessary conditions will now be examined.

A. One Step Integration Schemes Based on Quadrature

Assume the integration scheme is given by:

$$x(j,k) = x(0,k) + h \sum_{m=0}^{j-1} a(j,m)\, f(m,k) \tag{5}$$

$$\text{for} \quad j = 1,2,..,r \quad \text{and} \quad k = 0,1,...,N-1$$

where (j,k) denotes the j-th variable on the k-th grid interval and $f(m,k) = f(x(m,k),u(m,k))$, $x(r,k) = x(0,k+1)$, and $f(r,k) = f(0,k+1)$. Integration schemes of this form, which includes the Range-Kutta procedures, are discussed in [3].

If $q(j,k)$ is the dual multiplier associated with (5), $G(j,k) \equiv f_x(x(j,k),u(j,k))^T$, and f_x denotes the gradient of f with respect to x, then the discrete minimum principal or Kuhn Tucker conditions reduce to the following system of

equations:

$$(6) \qquad \sum_{j=1}^{r} [I+ha(j,0)G(r,k)]q(j,k+1)-q(r,k) = 0$$

$$(7) \qquad h\,G(j,k) \sum_{m=j+1}^{r} a(m,j)q(m,k) - q(j,k) = 0$$

for $1 \leq j < r$

$$(8) \qquad f_u(j,k)^T \sum_{m=j+1}^{r} a(m,j)q(m,k) = 0$$

for $0 \leq j < r$.

Define the variables:

$$(9) \qquad p(j,k) = \sum_{m=j+1}^{r} \frac{a(m,j)}{a(r,j)}\, q(m,k), \quad \text{for} \quad 0 \leq j < r$$

$$p(r,k) = p(0,k+1)$$

$$a(r,r) = a(r,0).$$

Note that (8) reduces to

$$(10) \qquad f_u(j,k)^T\, p(j,k) = 0.$$

Now multiply (7) by $a(j,m)/a(r,m)$, sum from $j = m+1$ to $r-1$, and add the result to (6) to determine:

$$(11) \qquad p(m,k) = h \sum_{j=m+1}^{r} a(j,m) \frac{a(r,j)}{a(r,m)} G(j,k)\, p(j,k) + \sum_{j=1}^{r} q(j,k+1).$$

Changing from the q to the p variable in (6) yields:

$$(12) \qquad p(r-1,k) = h\, a(r,0)\, G(r,k)\, p(r,k) + \sum_{j=1}^{r} q(j,k+1).$$

Subtracting (12) from (11), produces:

$$(13) \qquad p(m,k) = p(r-1,k) + h \sum_{j=m+1}^{r-1} a(j,m) \frac{a(r,j)}{a(r,m)} G(j,k)p(j,k)$$

$$\text{for} \quad m = 0,1,\ldots,r-2.$$

Summing (7) from $j = 1$ to $r-1$ results in:

$$(14) \qquad \sum_{j=1}^{r-1} q(j,k) = h \sum_{j=1}^{r-1} a(r,j)\, G(j,k)\, p(j,k).$$

Substituting (14) in (12) yields:

$$(15) \qquad p(r-1,k-1) = p(r-1,k) + h \sum_{j=0}^{r-1} a(r,j)\, G(j,k)\, p(j,k).$$

The terminal condition is

$$(16) \qquad p(r-1,N) = (1,0,0,\ldots,\ ).$$

Note that the difference relations (13), (15), and (16) run backward in time and furthermore all the standard one-step schemes such as those in Table I satisfy the identity:

$$(17) \qquad a(j,m)\,\frac{a(r,j)}{a(r,m)} = a(r-1-m,r-1-j)$$

$$a(r,j) = a(r,r-1-j)$$

so that (13) and (15) are the same scheme as (5) except that it moves backward in time:

$$(18) \qquad p(m,k) = p(r-1,k) + h \sum_{j=m+1}^{r-1} a(r-1-m,r-1-j)G(j,k)p(j,k)$$

$$p(-1,k) = p(r-1,k-1).$$

Also note that the scheme (18) is structured so that $p(r-1,k-1)$ not $p(0,k)$ approximates $p(t_k)$, the solution to (3) evaluated at t_k, the k-th grid point; and since $G(j,k)$ is a function of $x(j,k)$ where $x(j,k)$ was generated by the forward Equation 5, then equations (5) and (18) represent an implicit scheme for the system of differential equations (2) and (3). i. e. if the initial condition $p(-1,0)$ is given, then (5) and (18) can both be solved in the forward direction for x and p although the procedure is implicit. Lemma 1 will now prove that this implicit scheme approximates the system (2) and (3) to the same order as (5) approximates (2). Recall that a difference approximation is said to be of order b if the exact solution to the differential equation satisfies the difference equation to within ch^{b+1} where h is the grid interval and c is a constant depending on the derivatives of x^* and u^*.

The equation (18) is of the form

$$(19) \quad p(r-1,k-1) = F(G(r-1,k),\ldots,G(0,k))\ p(r-1,k)$$

where F consists of sums and products of the matrices that are its arguments. A feature common to Range-Kutta schemes such as those in Table I is the following

symmetry property:

$$(20) \quad F(A_1, A_2, \ldots A_r) = F(A_r^T, A_{r-1}^T, \ldots A_1^T)^T$$

Lemma 1:

If the onestep scheme (5) is of order b and (20) holds, then for given fixed $u(\cdot)$, the schemes (5) and (18) are b-th order approximations to the system (2) and (3).

Proof: Since (5) is a b-th order scheme for (2), then (5) combined with the scheme

$$q(k+1) = F(G(0,k)^T, \ldots, G(r-1,k)^T)\, q(k) \tag{21}$$

is an explicit b-th order "forward marching" scheme for the system

$$\dot{x}(t) = f(x(t),u(t)) \tag{22}$$

$$\dot{p}(t) = f_x(x(t),u(t))\, p(t).$$

If $H(A,t,s)$ is the transition matrix for the system $\dot{y}(t) = A(t)\, y(t)$ where $A = -f_x(x,u)^T$, then it follows that

$$F(G(0,k)^T, \ldots, G(r-1,k)^T) = H(-A^T, t_k+h, t_k) + O(h^{b+1}). \tag{23}$$

where t_k is the k-th grid point and h is the grid interval. We wish to prove that

$$(24) \qquad F(G(r-1,k),\ldots,G(0,k)) = H(A,t_k,t_k+h) + O(h^{b+1}).$$

Since (20) holds, then (23) implies that

$$(25) \qquad F(G(r-1,k),\ldots,G(0,k)) = F(G(0,k)^T,\ldots,G(r-1,k)^T)^T$$
$$= H(-A^T,t_k+h,t_k)^T + O(h^{b+1}).$$

If $H(-A^T,t_k+h,t_k)^T = H(A,t_k,t_k+h)$, then (24) follows immediately. This last identity is proved by showing that both sides of the equality satisfy the same differential equation and initial conditions; let $s = t_k$ and observe that:

$$\frac{d}{dh} H(A,s,s+h) = - H(A,s,s+h)A$$

$$\frac{d}{dh} H(-A^T,s+h,s)^T = [-A^T H(-A^T,s+h,s)]^T.$$

∎

B. Multistep Schemes

Assume the multistep scheme is of the form:

$$(26) \qquad \sum_{j=0}^{r} a(j)x(k+j) = h \sum_{j=0}^{r} b(j)\, f(k+j)$$

for $k = 0,1,\ldots,N-r$

where $f(k) = f(x(k),u(k))$. The integration of equations by (26) is studied in [2]. Multistep schemes require starting procedures such as the one step schemesdiscussed earlier to generate the initial conditions $x(0),\ldots,x(r-1)$, however, to simplify the analysis, it will be assumed that these starting conditions are known exactly.

If $q(k)$ is the multiplier corresponding to (26) and $G(k) = f_x(x(k),u(k))^T$, then the discrete necessary conditions are:

$$(27) \qquad \sum_{j=r}^{0} a(j)q(k-j) = e(k) + h \sum_{j=r}^{0} b(j)G(k)q(k-j)$$

$$(28) \qquad \sum_{j=r}^{0} f_u(k)^T b(j)q(k-j) = 0$$

for $k = 0,1,\ldots N$ where $q(j) \equiv 0$ for $j > N-r$, $e(j) = 0$ for $j \neq N$, and $e(N) = (1,0,0,\ldots,0)$. Define

$$(29) \qquad p(k) = \sum_{j=r}^{0} b(j)\, q(k-j).$$

Then (28) reduces to

$$(30) \qquad f_u^{T}(k)\, p(k) = 0.$$

Now multiply (27) by $b(m)$, replace k with $k-m$ and sum from $m = 0$ to r to obtain:

$$(31) \qquad \sum_{m=0}^{r} \sum_{j=r}^{0} a(j)b(m)q(k-m-j) = \sum_{m=0}^{r} b(m)e(k-m) + h \sum_{m=0}^{r} \sum_{j=r}^{0} b(j)b(m)G(k-m)q(k-j-m).$$

Interchanging the order of summation on the left side of (31) and using (29) results in:

$$(32) \qquad \sum_{m=0}^{r} a(m)p(k-m) = \sum_{m=0}^{r} h\, b(m)G(k-m)p(k-m) + b(m)e(k-m).$$

The $e(k-m)$ term vanishes for $k < N$ so that (32) is the same multistep scheme as (26) applied to the Equation 3 except that it moves in the backward direction.

Note that since $q(j) = 0$ for $j > N-r$, then by (29), $p(j) = 0$ for $j > N$. Hence (32) has a built-in terminal condition so that once $x(j)$ has been determined, then (32) can be solved starting at the right endpoint. As Henrici [2] proves, the accuracy of convergent multistep schemes is the minimum of the order of accuracy of the initial conditions and the order of the scheme. The order of accuracy of the built-in procedure near the right endpoint is listed in Table II for several multistep schemes.

C. Taylor Series Schemes

Only the following Taylor series integration scheme is considered:

$$(33) \qquad x(k+1) = x(k) + h\, f(k) + .5\, h^2 (f_x(k)f(k) + f_u(k)\dot{u}(k)).$$

The corresponding discrete necessary conditions are:

$$(34) \quad p(k-1) = p(k)+[hf_x(k)+.5h^2(f_x(k)f(k))_x+.5h^2(f_u(k)\dot{u}(k))_x]^T p(k)$$

$$(35) \qquad [(f_x(k)f(k))_u + (f_u(k)\dot{u}(k))_u]^T p(k) = 0$$

(36) $$f_u(k)^T\, p(k) = 0.$$

These relations will be discussed in more detail in the next section.

III. UPPER BOUNDS ON CONVERGENCE RATES

To derive upper bounds on the rate of convergence of the solution of the discrete system to the solution of the continuous system, we simply assume that the discrete state and control approximate the continuous state and control to some order and then examine the discrete necessary conditions to determine whether there is a contradiction. The proofs in this section are not completely rigorous, although the validity of the statements should be clear. The following implicit function theorem is used [4]:

(I) Suppose $g: R^n \times R^m \to R^n$, $g(a^*,b^*) = 0$, $\frac{\partial g(a^*,b^*)}{\partial a}$ is non-singular and g is continuously differentiable in a neighborhood of (a^*,b^*). Then $a(b)$, the solution to $g(a,b) = 0$ exists for b in some neighborhood B of b^* and $|a(b_1) - a(b_2)| \leq c|b_1 - b_2|$ for some c and for all $b_1, b_2 \in B$.

Theorem 1

Suppose that (x^*,u^*) solves (P) and p^* is generated in (3) by (x^*,u^*), $f(x,u)^T p$ is twice continuously differentiable, and $[f(x^*(t),u^*(t))^T p^*(t)]_{uu}$ is non-singular for all $t \in [0,1]$. Also assume that the onestep scheme (5)

is accurate to order b and the difference $|x(r,k)-x(r-1,k)|$ is of order s. Then the solution to the discrete minimization problem is in general accurate to order at best $m = \min(s,b)$.

Proof: Since the onestep scheme (18) for the discrete costate is of order b as proved in Lemma 1, then in general $|p(r-1,k)-p^*(t_{k+1})|$ is at best $O(h^b)$; (recall that $p(r-1,k)$ not $p(0,k+1)$ approximates $p^*(t_{k+1})$). Now $u(0,k)$ satisfies

$$f_u(x(0,k),u(0,k))^T p(0,k) = 0 \tag{37}$$

while x^*,u^*,p^* satisfies:

$$f_u(x^*(t_k),u^*(t_k))^T p^*(t_k) = 0. \tag{38}$$

By the property of the integration scheme assumed above $|p(r-1,k-1)-p(0,k)|$ is $O(h^s)$ and hence $|p(0,k)-p^*(t_k)| = O(h^m)$. Since $|x(0,k)-x^*(t_k)|$ is at best $O(h^b)$, then from (I), the solution $u(0,k)$ to (37) agrees with $u^*(t_k)$, the solution to (38), to order at best m.

Now consider $x(r\text{-}1,k), u(r\text{-}1,k), p(r\text{-}1,k)$ which satisfies:

$$f_u(x(r\text{-}1,k),u(r\text{-}1,k))^T p(r\text{-}1,k) = 0$$

where $|p(r\text{-}1,k)-p^*(t_{k+1})|$ is at best $O(h^b)$ while $|x(r\text{-}1,k)-x^*(t_{k+1})|$ is at best $O(h^m)$. Hence $|u(r\text{-}1,k)-u^*(t_{k+1})|$ is at best $O(h^m)$. ∎

Note, however, that if f_u is not a function of x, then the fact that $|x(r\text{-}1,k)-x^*(t_k)| = O(h^m)$ does not enter into the analysis in the last paragraph above and hence $u(r\text{-}1,k)$ can be accurate to order b. In a numerical example in Section 4, this higher convergence rate is indeed observed in a problem where the state and control uncouple so that f_u does not depend on x.

Theorem 2

Suppose that (x^*,u^*) solves (P), p^* is generated in (3) by (x^*,u^*), $f(x,u)^T p$ is twice continuously differentiable, and $[f(x^*(t),u^*(t))^T p^*(t)]_{uu}$ is non-singular for all $t \in [0,1]$. Also assume that the multistep scheme (26) is of order b and the "built-in" initial conditions at the right

endpoint for the discrete costate equation (27) are accurate to order s. Then the solution to the discrete minimization problem is in general accurate to order at best $m = \min(b,s)$.

Proof: As noted in Section 2, $p(k)$ is in general accurate to order at best m. Since $f(x(k),u(k))^T p(k) = 0$ and $f_u(x^*(t_k),u^*(t_k))^T p^*(t_k) = 0$, then (I) again implies that $|u(k)-u(t_k)|$ is at best $O(h^m)$. □

Now the Taylor series integration scheme (33) is examined. If x^*,u^*,p^* are optimal in (P), then by differentiating (4), the following condition holds:

$$(38) \qquad [(f_x(x^*,u^*)f(x^*,u^*))_u + (f_u(x^*,u^*)\dot{u}^*)_u$$

$$-f_x(x^*,u^*)f_u(x^*,u^*)]^T p^* = 0.$$

If $(u(k),\dot{u}(k))$ is near $(u^*(t_k),\dot{u}^*(t_k))$; then $(x(k),p(k))$ will be near $(x^*(t_k),p^*(t_k))$ since the integration scheme (33) and (34) are at least first order. There is now a contradiction since $(x(k),u(k),\dot{u}(k),p(k))$ satisfies (35) while $(x^*(t_k),u^*(t_k),\dot{u}^*(t_k),p^*(t_k))$ satisfies a different equation, (38), above. In fact, it is observed numerically that the discrete solution diverges from the continuous solution.

IV. NUMERICAL EXAMPLES

The convergence of the schemes listed in Table I and Table II were studied numerically in two simple control problems with linear dynamics and quadratic cost:

$$(P1) \qquad \min \int_0^1 .5\, u(t)^2 + x(t)^2 \, dt$$

$$\text{s.t. } \dot{x}(t) = .5x(t) + u(t)$$

$$x(0) = 1$$

$$(P2) \qquad \min \int_0^1 5x(t)^2/8 + u(t)x(t)/2 + u(t)^2/2 \, dt$$

$$\text{s.t. } \dot{x}(t) = .5x(t) + u(t)$$

$$x(0) = 1.$$

The solution of (P1) and (P2) are respectively:

$$(S1) \qquad u(t) = \frac{e^{t/2}(2+ce^{3t})[(1-ce^{3t})(e^{-3t}-c)^2]^{1/3}}{(1-ce^{3t})(c-1)}$$

$$\text{where } c = -2/e^3$$

$$(S2) \qquad u(t) = \tanh(1-t)(\tanh(1)\sinh(t) - \cosh(t))$$

The solutions are shown in Figures 1 and 2 and plots of the convergence for the various discretizations are given in Figures 3, 4, and 5. The error that is plotted is the L^2 error of the most accurate discrete control parameters. For example, in multistep schemes, the controls near the right end are of a lower order than the controls several grid intervals away from t=1 as should be expected. The error near t=1 is excluded in the data presented.

The convergence rates agreed roughly with the predicted upper bounds in section3: In problem (P1), the state and control terms in the cost function uncoupled so that the maximum convergence rate equalled the order of the integration scheme as anticipated at the end of Theorem 1. On the other hand, in (P2) the cost does not uncouple so that the convergence rate is bounded by the order of the difference $|x(r-1,k)-x(r,k)|$, as Theorem 1 predicted.

For the multistep schemes in Table II, the Milne's three and five point schemes were only accurate to $O(1)$ at the right endpoint and numerically it was observed that these schemes did not converge to the correct solution. The Improved Adams and Modified Euler schemes, however, converged to the correct solutions at order 4 and 2 respectively as expected.

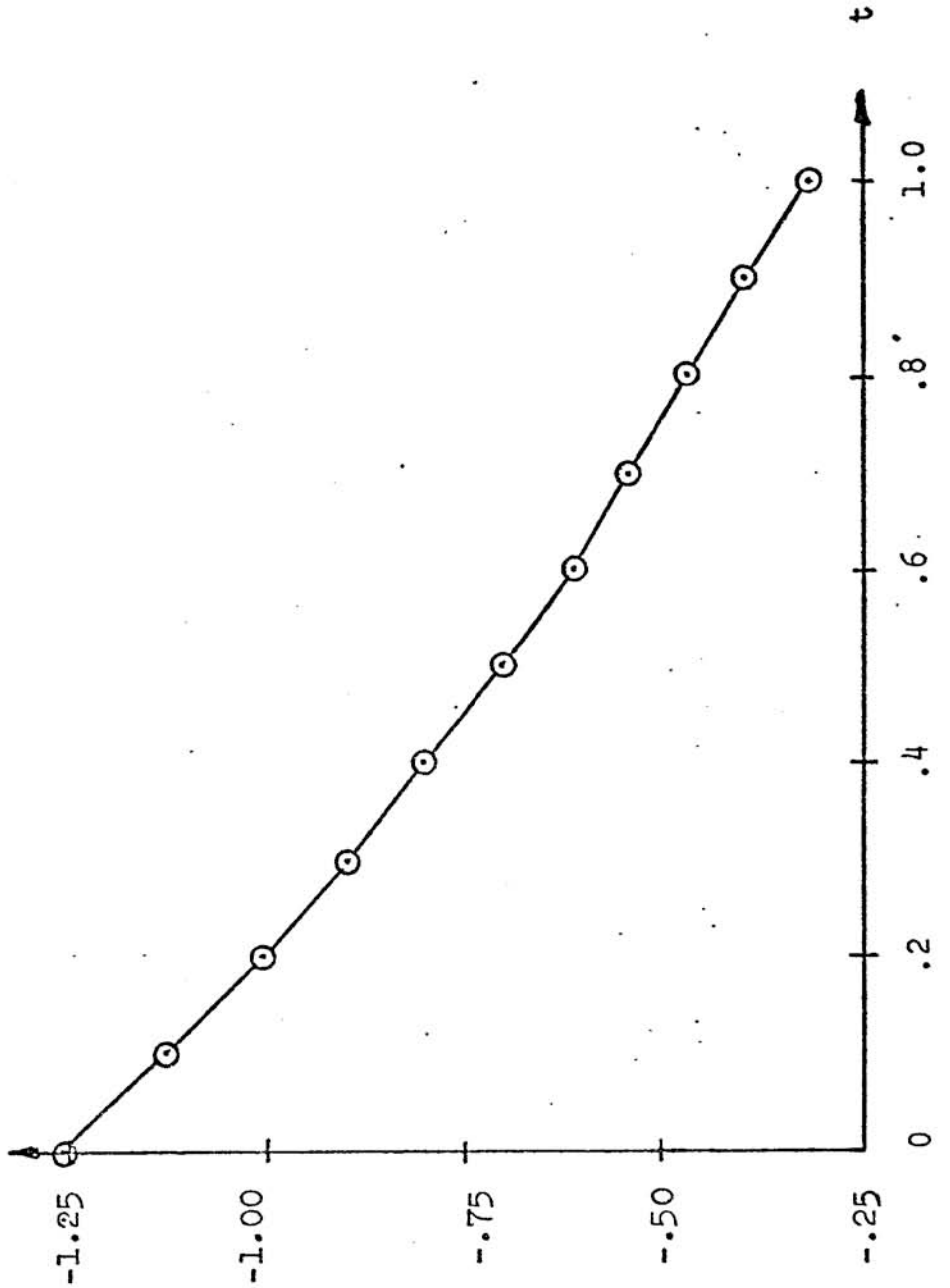

Figure 1: The Optimal Control for Problem (P2)

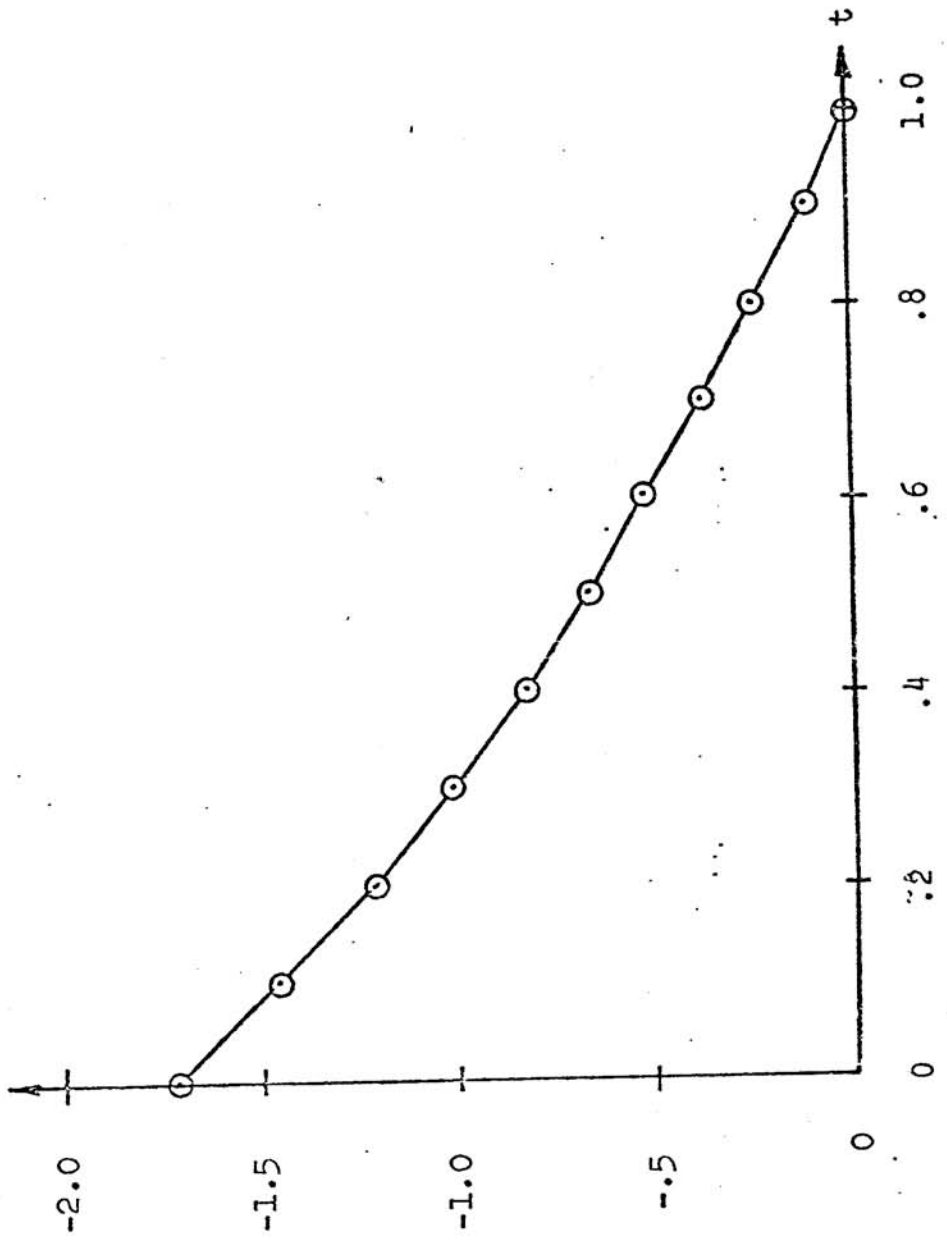

Figure 2: The Optimal Control for Problem (P1)

NAME OF SCHEME	THE MATRIX a(i,j)	ORDER OF SCHEME	ORDER OF DIFFERENCE x(r-1,k)-x(r,k)	OBSERVED CONVERGENCE RATE IN (P1)	OBSERVED CONVERGENCE RATE IN (P2)
MODIFIED EULER	1 0 1/2 1/2	2	2	~1.8	~1.9
KUTTA	1/2 0 0 -1 2 0 1/6 2/3 1/6	3	2	~2.7	~2.0
RANGE-KUTTA 1	1/2 0 0 0 0 1/2 0 0 0 0 1 0 1/6 1/3 1/3 1/6	4	3	~3.5	~2.9
RANGE-KUTTA 2	1/3 0 0 0 -1/3 1 0 0 1 -1 1 0 1/8 3/8 3/8 1/8	4	2	~3.5	~1.9

TABLE I: One Step Schemes Based on Quadrature

NAME OF SCHEME	THE VECTOR a(i)	THE VECTOR b(i)	ORDER OF SCHEME	ACCURACY OF TERMINAL CONDITION	OBSERVED CONVERGENCE RATE IN (P1)
MODIFIED EULER	(1,-1)	(1/2,1/2)	2	2	~2.0
MILNE'S THREE POINT METHOD	(1,0,-1)	(1/3,4/3,1/3)	3	0	0
IMPROVED ADAMS	(1,-1,0,0)	(9/24,19/24,-5/24, 1/24)	4	4	~3.9
MILNE'S FIVE POINT METHOD	(1,0,0,0,-1)	(14/45,64/45,24/45, 64/45,14/45)	5	0	0

TABLE II: Multi-step Schemes

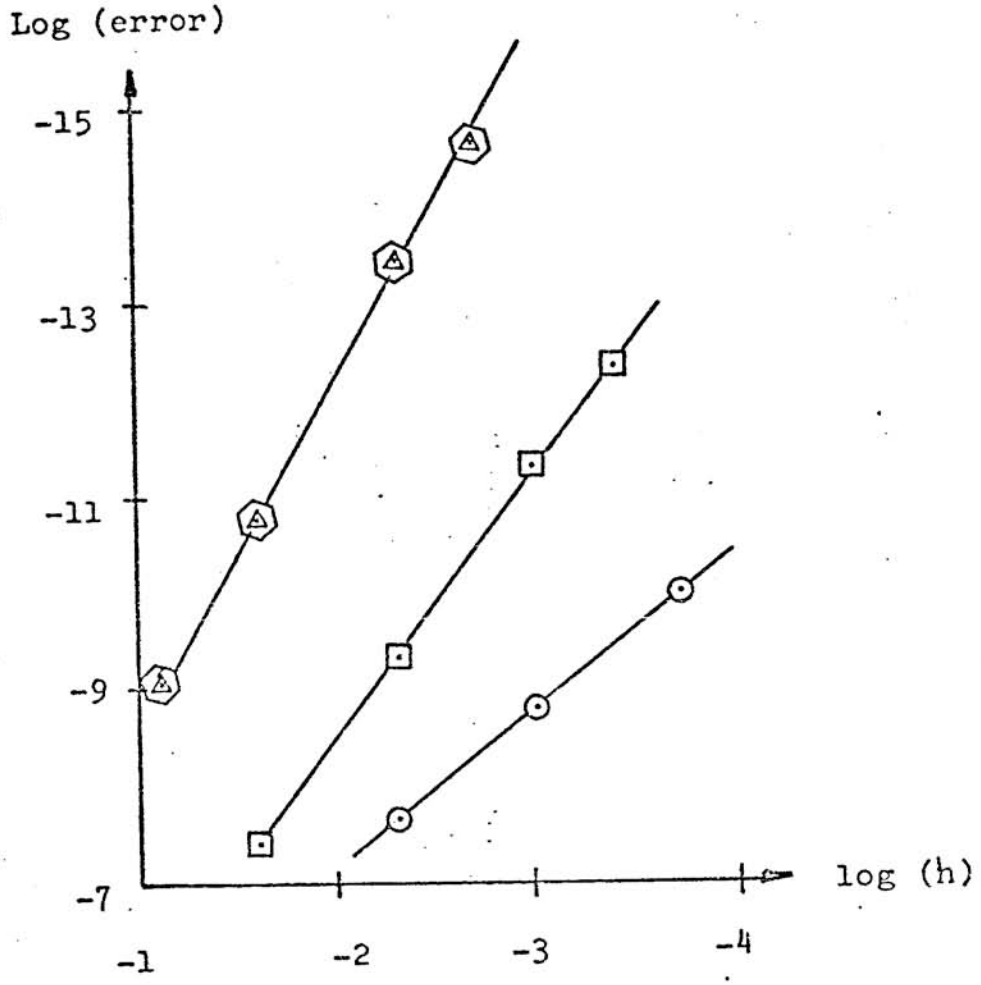

Figure 3: Convergence of Discrete Control for One Step Schemes in (P1)

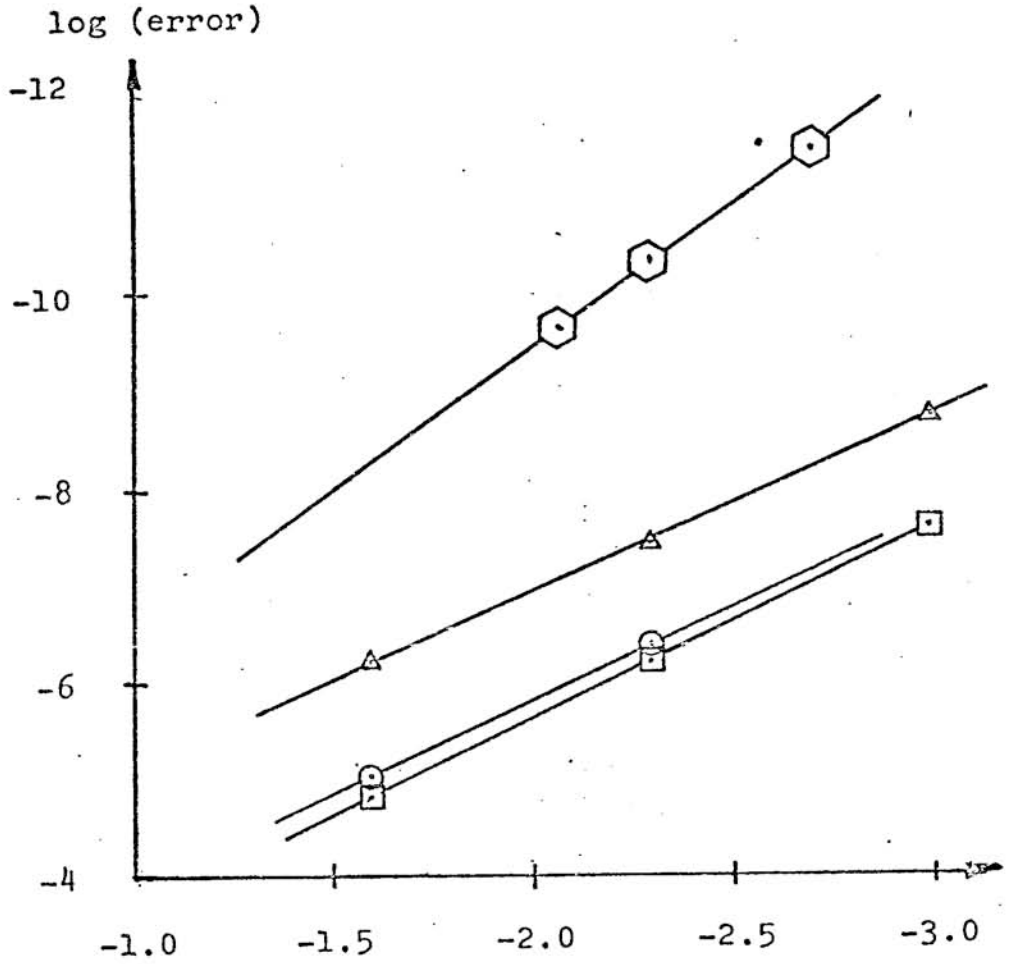

Figure 4: Convergence of Discrete Control for One Step Schemes in (P2)

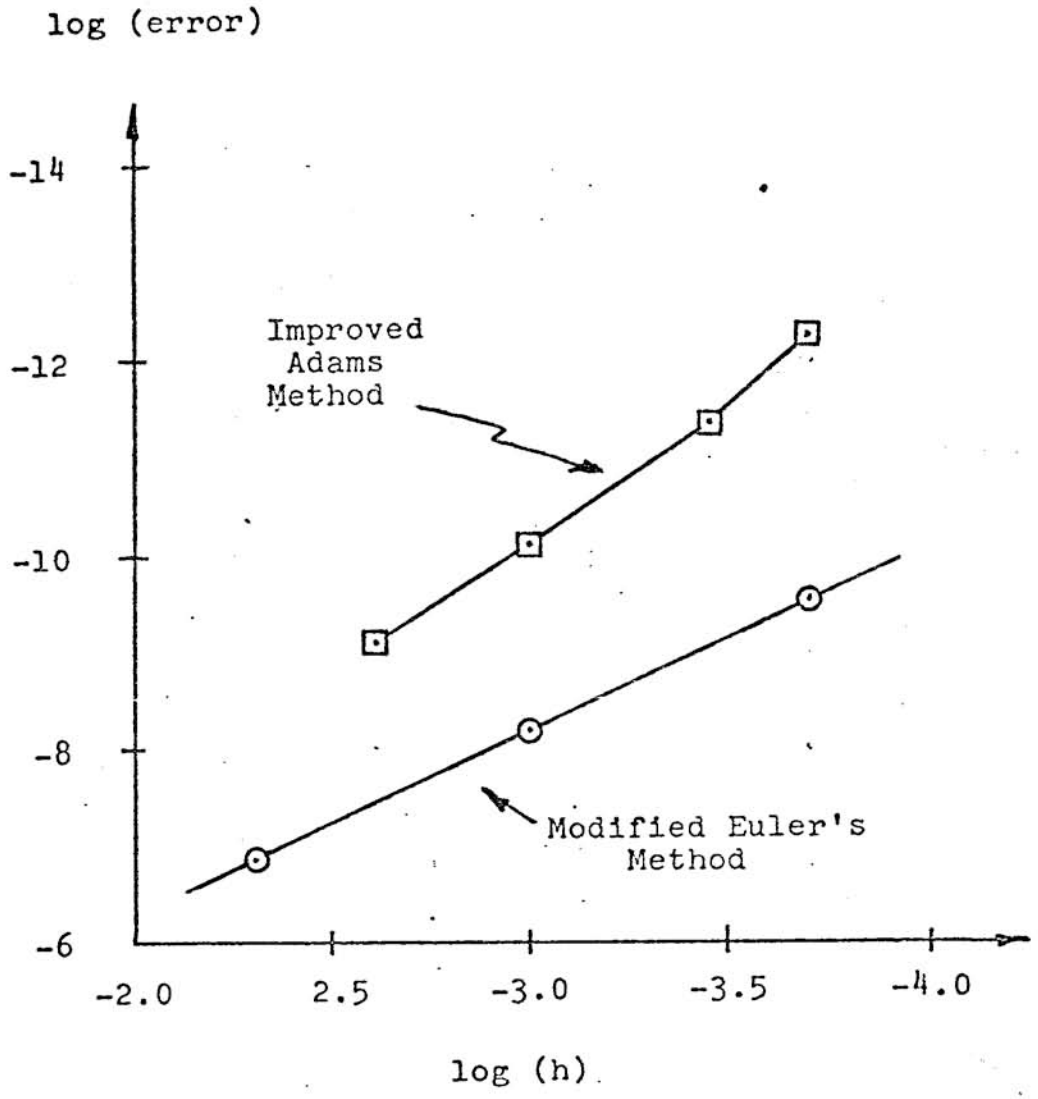

Figure 5: Convergence of Discrete Control for Multistep Methods in (P1)

Note that for a given order of convergence, the discrete problem corresponding to the multistep schemes involves fewer variables since intermediate values for the control between the grid points are not needed. The one step methods are advantageous, of course, for problems with discontinuities at known points; the error in one step procedures depends on the derivatives of u* between the grid points while the error in the multistep scheme depends on the global derivatives of u* (unless the scheme is restarted at the point of discontinuity which increases programing difficulty).

V. IMPROVED ONE STEP SCHEMES?

Since the convergence rate for one step schemes depends on the minimum of the order of the scheme and the order of the difference $|x(r-1,k)-x(r,k)|$, then one immediately asks if it would be possible to construct a third order three point scheme like the Kutta scheme in Table I but with $|x(r-1,k)-x(r,k)|$ of order 3. Unfortunately this is impossible since the equations resulting from imposing the two order of accuracy conditions are inconsistant. (Of course Modified Euler's Method has the same optimal order, but uses one less intermediate point in the integration procedure.) Likewise with 4 point schemes, if one tries to construct a fourth order discretization, then the number of equations resulting from the accuracy requirements exceeds the number of unknowns, the $a(i,j)$. Hence Modified Euler's Method, Kutta, and Range-Kutta 1 are optimal 2, 3, and 4 point schemes respectively of the form given by (5).

VI. COST CONVERGENCE

In the solution of unconstrained control problems using the finite element method, if the discrete controls converge at order b, the the cost usually converges at order 2b; however, if the integration scheme is only accurate to order s, then the discrete cost can only converge at order s. This was also observed numerically; for example, the optimal cost for the discrete problem corresponding to Improved Adam's Method only converged at order 4 and not 8.

References

1. M.D. Canon, C.D. Cullum, and E. Polak, Theory of Optimal Control and Mathematical Programming, McGraw-Hill, New York, 1970

2. P. Henrici, Error Propogation for Difference Methods, John Wiley, New York, 1963

3. E.Isaacson and H.B. Keller, Analysis of Numerical Methods, John Wiley and Sons, New York, 1966

4. R.C. James, University Mathematics, Wadsworth Publishing Co., Belmont, Ca., 1966

5. H.B. Keller, Numerical Methods for Two-Point Boundary-Value Problems, Blaisdell Publishing Co., Waltham, Ma., 1968

6. L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze, and E.F. Mishchenko, The Mathematical Theory of Optimal Processes, John Wiley and Sons, New York, 1965

7. S.M. Roberts and J.S. Shipman, Two-Point Boundary-Value Problems : Shooting Methods, American Elsevier, New York, 1972

"God gives wise men their wisdom and scholars their intelligence. He reveals profound mysteries beyond man's understanding. He knows all hidden things, for he is light, and darkness is no obstacle to him."

Daniel 2:21-22