AD-A063 962 AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OHIO SCH--ETC F/G 20/13
AN INVESTIGATION OF THE METHOD OF FINITE ELEMENTS WITH ACCURACY--ETC(U)
DEC 78 J C GENECZKO

UNCLASSIFIED AFIT/GNE/PH/78D-15 NL

AFIT/GNE/PH/78D-15

(1)

LEVEL II

AD A063962

DDC FILE COPY

AN INVESTIGATION OF THE METHOD OF FINITE ELEMENTS WITH ACCURACY COMPARISONS TO THE METHOD OF FINITE DIFFERENCES FOR SOLUTION OF THE TRANSIENT HEAT CONDUCTION EQUATION USING OPTIMUM IMPLICIT FORMULATIONS

THESIS

AFIT/GNE/PH/78D-15

Joseph C. Geneczko
Capt USAF

Approved for public release; distribution unlimited.

(14) AFIT/GNE/PH/78D-15

(6) AN INVESTIGATION OF THE METHOD OF FINITE ELEMENTS WITH ACCURACY COMPARISONS TO THE METHOD OF FINITE DIFFERENCES FOR SOLUTION OF THE TRANSIENT HEAT CONDUCTION EQUATION USING OPTIMUM IMPLICIT FORMULATIONS.

(9) THESIS Master's thesis,

(12) 179 p.

Presented to the Faculty of the School of Engineering

of the Air Force Institute of Technology

Air University

in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

by

(10) Joseph C. Geneczko, B.S.

Capt USAF

Graduate Nuclear Engineering

(11) December 1978

Approved for public release; distribution unlimited.

Preface

This report is the result of my investigation of the finite-element method, with a quadratic interpolation function, for solution of the one-dimensional transient heat conduction equation. Although order of accuracy improvements over the linear interpolation formulation did not materialize, the results achieved were significant in that they were previously postulated, easily accounted for by elementary mathematical analysis, and verified the accuracy of solution by finite-elements. Failure to achieve greater accuracy was a function of the solution method; solution improvement by quadratic interpolation requires special treatment of the internal nodes and time domain.

I would like to express my appreciation and gratitude to Dr. Bernard Kaplan of the Air Force Institute of Technology for his guidance in my performance of this thesis, and to Dr. W. Kessler of the Air Force Materials Laboratory for sponsoring this research project. Also, I am deeply grateful to Drs. John Jones and David Hardin, also of the Air Force Institute of Technology, for their technical advice on many special occasions, and to Sharon Gabriel for her precise typing achievement.

Finally, I wish to express my gratitude to my wife, Linh, for her invaluable moral support and insistence to maintain a proper perspective in the accomplishment of this continuous work.

Contents

		Page
Preface		ii
List of Tables		v
Abstract		vi
I.	Introduction	1
	Background	1
	Problem	2
	Scope	3
	Assumptions	3
	Approach	4
II.	Theory	5
	The Physical Problem	5
	Finite Element Background Theory	10
	Finite Element Problem Approach	13
	Quadratic Finite Element Application	17
	Quadratic Solution Interpretation	45
	Error Analysis	46
	Stability Analysis	49
III.	Procedure	59
	General Approach	59
	Computer Application	62
IV.	Results	65
	Stability Analysis	65
	Error Analysis	65
	Plots	69
V.	Conclusions and Recommendations	74
	Conclusions	74
	Recommendations	78
Bibliography		80
Appendix A:	The Analytical Solution of the Primary Problem	81
Appendix B:	Elementary Variational Calculus Review	85
Appendix C:	Derivation of Quadratic Constants	90

Page

Appendix D: Assembly Theory for Matrices..... 92

Appendix E: Derivation of the Truncation Error for the Finite-Element Formulation..... 97

Appendix F: Treatment of Internal Nodes for Static Problems..... 102

Appendix G: Comparison of Linear and Quadratic Factors and Equivalence of Linear and Quadratic Interpolation Function Analysis..... 104

Appendix H: Computer-Generated Plots of Results..... 110

Appendix I: Alternative Formulation of the Time Response..... 164

Vita..... 169

List of Tables

Table		Page
I	Oscillation and Instability Limits By Eigenvalue Definition	51
II	Oscillation and Instability Limits for the Fourier Modulus in the Finite Element Method	54
III	Oscillation and Instability Limits for the Fourier Modulus in the Finite Difference Formulation	56
IV	Oscillation and Instability Limits for the Fourier Modulus in the Finite Element Method and Quadratic System	58
V	Error Comparisons for the Various Methods for $\theta = .08$ and $p = 1.0$	66
VI	Error Comparisons for the Various Methods for $\theta = .04$ and $p = .5$	67
VII	Error Comparisons for the Various Methods for $\theta = .16$ and $p = 2.0$	67
VIII	Discretization Error Ratio Comparison for the Optimum Implicit Scheme	70
D-I	System Numbering - The Correspondence Between Local and Global Numbering Schemes	93
D-II	Relationship Between Local and Global Elements of Matrix $\underline{A}$.	95
G-I	Equivalency of Optimum Alpha Values For Quadratic and Linear Interpolation Functions	109

Abstract

The one-dimensional transient heat conduction equation, with Dirichlet boundary conditions, is solved by the method of finite-elements, employing a quadratic interpolation function. The numerical solutions are investigated with respect to accuracy and stability, and compared to like results attained by the method of finite-differences, and the finite-element method with linear interpolation. The version of the finite-element method used was based on a variational principle which is stationary in time; the temporal behavior of the differential equation is treated with a finite-difference approximation. This method is equivalent to the method of Galerkin, called the Method of Weighted Residuals. The inherent discontinuity between the initial condition and boundary conditions was accounted for by substituting the exact analytical solution at the first time step and numerically computating from there. An equivalency relationship between the two finite-element methods is shown to exist. The finite-difference version of the Crank-Nicolson method is found to be more accurate than the finite-element version; for the fully implicit method, the opposite is found to be true. In the optimum implicit method, both finite-element solutions are shown equivalent to the finite-difference solution for a Fourier modulus of one. For other values of this parameter, the finite-element solution is more accurate.

AN INVESTIGATION OF THE METHOD OF FINITE ELEMENTS WITH ACCURACY COMPARISONS TO THE METHOD OF FINITE DIFFERENCES FOR SOLUTION OF THE TRANSIENT HEAT CONDUCTION EQUATION USING OPTIMUM IMPLICIT FORMULATIONS

I. Introduction

Background

Most engineering problems reduce to finding solutions of mathematical problems. Specifically, one translates a physical phenomenon into a differential equation, the solution of which yields the unknown value. Although analytical solutions are exact and desired, factors such as mixed geometry and computer limitations often prevent the application of analytical techniques. If one is willing to accept certain inaccuracies to be explained later, numerical techniques can be reasonably employed to obtain the desired solutions.

An accurate numerical technique is the method of finite-elements, in which the problem is recast as an integral to be minimized. Exactly, the finite-element method converts the original partial differential equation into a variational integral which must be minimized. The solution of the original partial differential equation is employed in this minimization process. A resultant set of algebraic equations is then solved by digital computer. Anyone familiar with the method of finite-differences should already note certain operational analogies, the main difference in the two

techniques being that, in finite-differences the solutions are evaluated at the nodes, while in finite-elements, the solutions are taken along the nodal intervals as well.

One problem suited to the application of finite-element procedures is that of transient heat conduction. The irregular geometry involved in the study of temperature variation and control in such pieces of hardware as jet engine burner baskets and rocket nozzles necessitates the use of numerical procedures to attain data such as required by the Air Force Materials Laboratory.

There exist several schemes to the finite-element solution of the transient heat conduction problem. These approaches include the Crank-Nicolson method, the Euler method, and the fully implicit method. Recently, Martin (Ref 7:52) developed an "optimum implicit method" which was shown to be the most accurate approach for his problem. Basically, the Martin method is a finite-element procedure in analogy to the Crandall method (Ref 3:318-320) of finite-differences.

Problem

The primary objective of this project was to solve the one-dimentional transient heat conduction problem, with a known analytical solution, using modifications of the Martin solution by the finite-element method. A quadratic interpolation was used and accuracy and stability comparisons made to Martin's linear interpolation solutions. Comparisons of accuracy and stability to the Crank-Nicolson finite-difference solution were also made, where instability

is defined as the tendency for oscillation errors of the numerical technique to grow unbounded, thus destroying that solution.

Scope

The problem analysis included a comparison of the Crank-Nicolson finite-difference and finite-element methods, using a quadratic interpolation function in the latter; a comparison of the Crandall optimum implicit finite-element method (Martin, linear) to the optimum implicit method using a quadratic interpolation function; and, a comparison of the Crank-Nicolson and optimum implicit methods where both employed quadratic interploation functions.

Assumptions

Three assumptions of note are: (1) the physical properties of the material of interest do not change in time or space; (2) no heat generation occurs within the material; and (3) the application of a constant dimensional mesh spacing to the numerical calculation is satisfactory for heat conduction problems.

Assumption (1) is justified in that unchanging material properties is a usual design feature. Assumption (2) is valid because there would be no difficulty should a heat generation factor exist. Such a term could be added to the given equation as long as it was constant with respect to time and space. Assumption (3) is the greatest limitation on applicability because not all problems have the same geometry and thus the same mesh spacing. For this one-dimensional problem, the assumption is valid if no inhomogeneities

exist in the material.

Approach

Basically, greater accuracy for the finite-element solution of the transient heat conduction problem was attempted by using a quadratic interpolation function and employing the various schemes noted earlier. The major obstacle was to apply the finite-element theory to such a function and to derive the basis for the finite-element numerical formulation. The second major problem was to derive the optimum implicit theory for its application. Finally, computer programs were written to perform the comparisons mentioned.

II. Theory

The Physical Problem

The transient heat conduction equation is a specific form of the general linear second order partial differential equation

$$Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = G \qquad (1)$$

where the discriminant, $B - 4AC$, equals zero. This also establishes the equation of interest as parabolic (Ref 2:97). The terms A through G are constants for functions of x and y only.

The transient heat conduction equation states that the overall change in the internal energy of a system is equal to the heat gain, plus internally generated heat, minus heat loss. As time derivatives, the equation states

$$\dot{U}_{stored} = \dot{U}_{in} + \dot{U}_{gen} - \dot{U}_{out} \qquad (2)$$

These terms can be replaced by rate equations. Of interest here is the conduction rate term

$$\dot{q} = -kA\frac{\partial T}{\partial x} \qquad (3)$$

where

$\dot{q}$ = rate of heat flow in the x direction

k = coefficient of thermal conductivity

A = area normal to the x direction through which heat flows

T = temperature

x = space variable

and, the heat storage term

$$\dot{U}_{store} = \rho V c \frac{\partial T}{\partial t} \qquad (4)$$

where

$\dot{U}_{store}$ = rate of heat storage

ρ = density

V = volume

c = specific heat

t = time

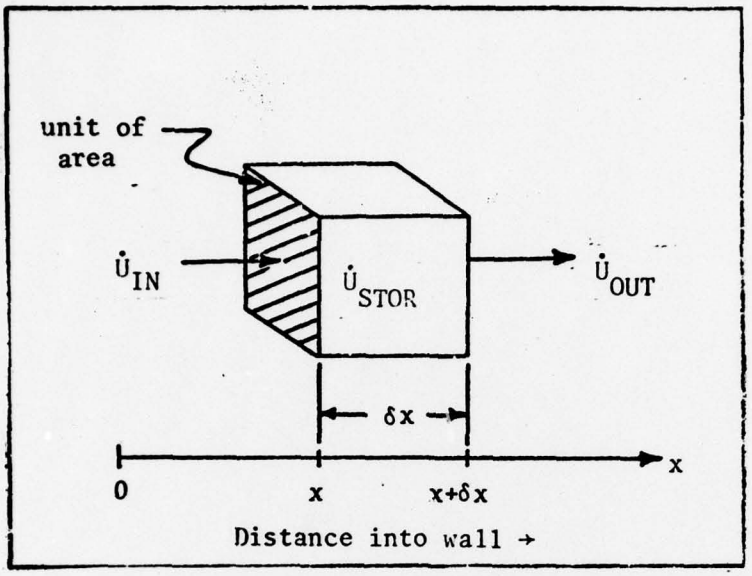

Figure 1. A Unit of Volume From a Wall with Large Dimensions in the y and z Directions.

Figure 1 (Ref 7:7) indicates that

$$\rho cA\frac{\partial T}{\partial t} = \frac{[kA\partial T/\partial x]_{x+\Delta x} - [kA\partial T/\partial x]_{x}}{\Delta x} \tag{5}$$

for no internal heat generation. As x goes to zero, the standard parabolic heat equation is attained as

$$\rho cA\frac{\partial T}{\partial t} = \frac{\partial}{\partial x}\left[kA\frac{\partial T}{\partial x}\right] \tag{6}$$

where, if ρ , k , and c are spatially constant with constant cross-sectional area, the attained result is

$$\frac{\partial T}{\partial t} = \frac{k}{\rho c} \frac{\partial^2 T}{\partial x^2} \tag{7}$$

The physical problem of note is completed by applying initial and boundary conditions. For purposes of this study, Dirichlet boundary conditions are used, where the function itself is specified at the boundaries. The exact problem considered is that of a parallel sided plane wall, infinite in all directions normal to the direction of heat flow. The wall is heated until a steady state temperature of 1000° F is attained throughout the continuum, and then cooled to a continuous temperature of 0° F. The boundary conditions for a wall of length x , 0 to L are

$$T(0,t) = T(L,t) = T_B \quad , \quad T > 0 \tag{8}$$

The initial condition is

$$T(x,t) = T_i \quad , \quad t = 0 \tag{9}$$

T_B and T_i are the specified boundary and initial conditions of 0° F and 1000° F, respectively.

Without loss of generality, the problem can be normalized to

$$\frac{\partial u}{\partial \theta} = \frac{\partial^2 u}{\partial \bar{x}^2} \tag{10}$$

where the corresponding boundary and initial conditions are

$$u(o,\theta) = u(1,\theta) = 0 \quad , \quad \theta > 0 \tag{11}$$

and

$$u(\bar{x},\theta) = 1 \quad , \quad \theta = 0 \tag{12}$$

and where $\bar{x}$ is normalized position, θ is normalized time, and u is normalized temperature.

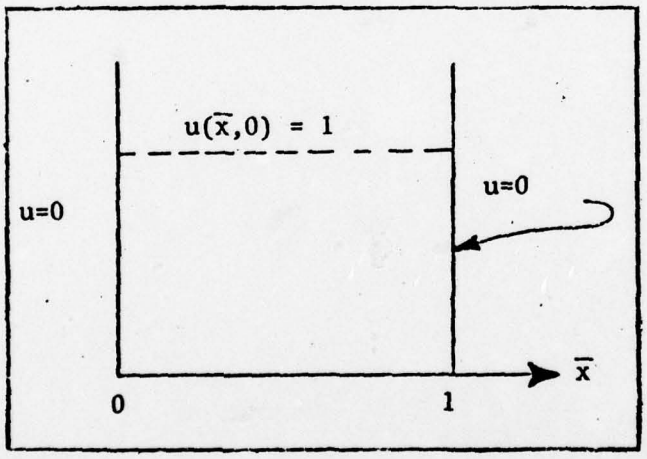

Figure 2. A Schematic Diagram of the Problem.

Figure 2 shows the normalized problem. Of immediate concern is the obvious discontinuity between the initial condition and the boundary conditions. Discussion of this dilemma will be postponed until later.

Because the numerical solutions will be compared to the exact analytical solution, this exact solution must be found. Separation of variables yields

$$u(\bar{x},\theta) = \sum_{m=1}^{\infty} \frac{4}{(2n-1)\pi} \sin [(2n-1)\pi\bar{x}] e^{[2n-1)\pi]^2\theta} \qquad (13)$$

The complete derivation is in Appendix A. Martin verified this problem (Ref 7:13) and discussed the truncation error of its computer solution (Ref 11:660). Figure 3 depicts the exact analytical solution.

Finite Element Background Theory

Unlike the finite-difference method, which envisions the solution region as an array of grid points, the finite-element method envisions the solution region as built up of many small, interconnected elements. A finite-element model of a problem gives a piecewise approximation to the governing equations. The basic premise of the finite-element method is that a solution region can be analytically modeled or approximated by replacing it with an assemblage of discrete elements. That is, the finite-element discretization procedures reduce the problem to one of a finite number of unknowns by dividing the solution region into elements and by expressing the unknown field variable in terms of assumed approximating functions within each element (Ref 5:5-6).

The approximating interpolation functions are defined in terms of the field variable values at the nodal points. The nodal

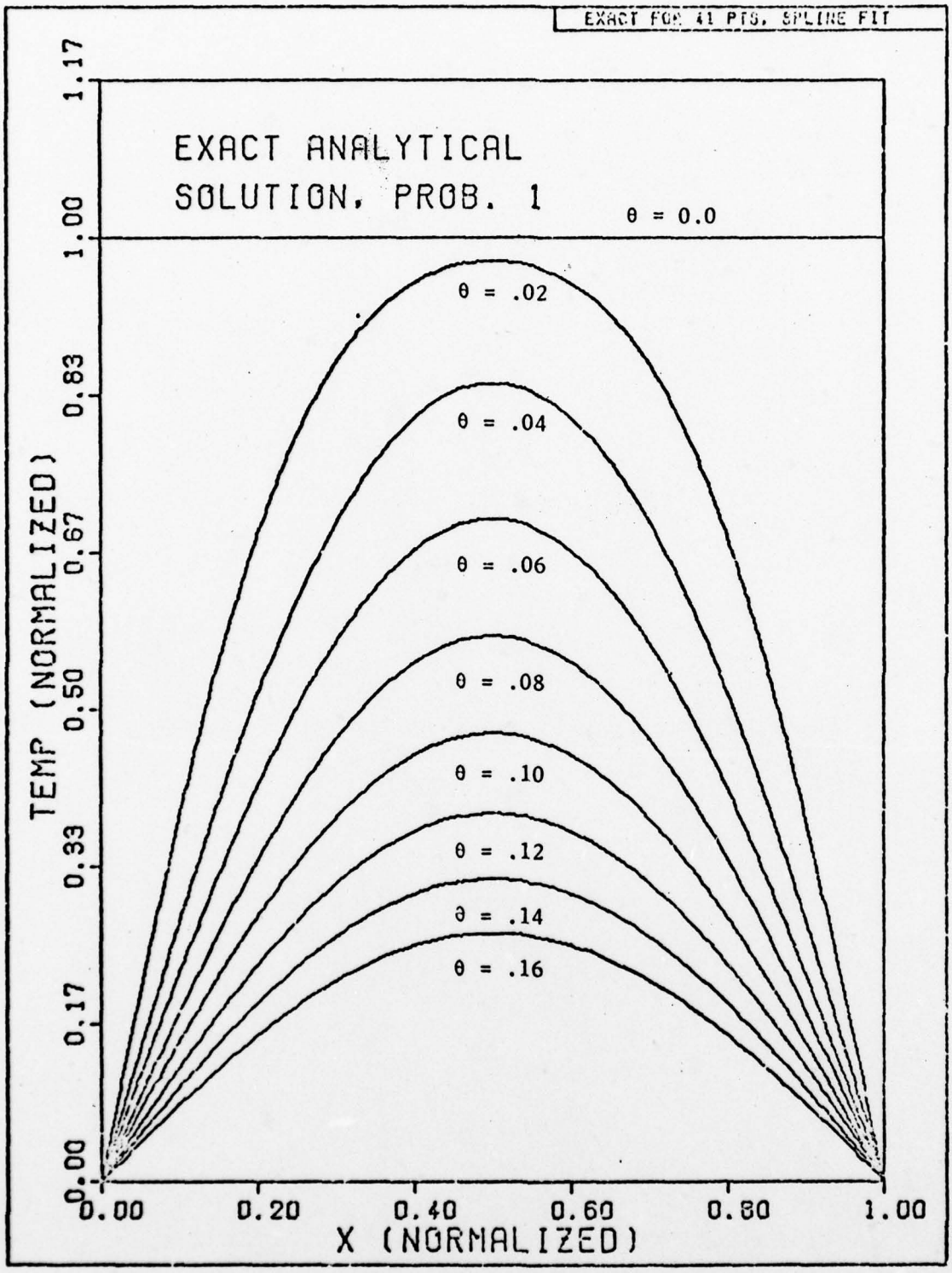

Figure 3. Analytic Solution of the Problem.

values of the field variable and the interpolation functions for the elements completely define the bahavior of the field variable within the elements.

The finite-element approach can be formulated in several ways, three of which are mentioned here. The most elementary approach is the direct approach. This procedure requires little mathematical manipulation and is used extensively in structural mechanics. Its main contribution to the heat problem lies in its reliance on matrix algebra and the formation of stiffness matrices. These matrices are employed in this thesis and discussed later.

The second appraoch, also employed in this thesis, is the mathematical or variational method. The variational basis dictates the criteria to be satisfied by the element interpolation functions. This method is one of several used to solve continuum problems. In the classical variational formulation, the problem is to find the unknown function or functions which extremize or make stationary a functional or system of functionals subject to the same given boundary conditions. This procedure is equivalent to solution of a system of differential equations because the functions that satisfy the differential equations and boundary conditions also extremize the functionals (Ref 5:67). Of course, the problem must be posed in variational form. Creation of the variational statement will be discussed in the following section.

The third approach is a particular form of the Method of Weighted Residuals, called Galerkin's Method. It is a general method used to formulate the finite-element equations without any

reliance on classical variational principles. In fact, this is its main advantage. Generally, the method requires an assumption about the general behavior of the dependent field variable. This approximation is substituted into the original differential equation, and any residual error made to vanish over the average. The resultant equations are now solved to yield the approximate solution. Martin (Ref 7:123-126) showed this method to be equivalent to the method of variations.

Whatever method or combination of methods is selected, Huebner reduces the finite-element procedure to the following steps, defined in the text (Ref 5:7-9):

(1) Discretize the continuum.

(2) Select interpolation functions.

(3) Find the element properties.

(4) Assemble the element properties to obtain the system equations.

(5) Solve the system equations.

(6) Make additional computations if desired.

Finite Element Problem Approach

General Approach. Myers (Ref 8:321-322) notes that, while in finite-difference theory the main concern is the approximation of derivatives by differences, the main concern of finite-elements involves the three concepts of minimization of functions, variational calculus, and, if necessary, the approximation of integrals.

Minimization of functions involves the elementary process of taking a derivative and setting the result equal to zero. Also

quite elementary and well-known is the approximation of integrals by such procedures as the trapezoid rule or Simpson's rule. In the finite-element method, the problem to be solved is cast as an integral to be minimized. A numerical approximation of the integral may be used to obtain the solution. The principles of variational calculus are briefly reviewed in Appendix B, in that this method is of primary use in the solution of the given problem.

Background. The method of approach is based on the variational principle as mentioned earlier and as used by Myers (Ref 8). The finite-element procedure is illustrated by solving the problem of concern directly. In review, note that the physical problem was stated in normalized form as

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial \theta} \tag{10}$$

$$u(0,\theta) = u(1,\theta) = 0 \quad , \quad \theta > 0 \tag{11}$$

$$u(x,\theta) = 1 \quad , \quad \theta = 0 \tag{12}$$

where $x = \bar{x}$. The finite-element method begins with a variational statement of the problem rather than the differential equation. Therefore, the variational statement corresponding to Equations (10) through (12) must first be found.

To find this variational statement, it is noted that the functional to be minimized is of the form (Ref 7:118-122):

$$I(\tilde{u}) = \int_0^1 [F(x,\tilde{u},\frac{d\tilde{u}}{dx})]\, dx \tag{14}$$

where $\tilde{u}$ represents a set of possible functions which satisfy Equation (14), as explained in Appendix B. For some fixed point in time

$$\tilde{u}(x) = u(x) + \eta v(x) \tag{15}$$

Chain rule differentiation of Equation (14) yields

$$\frac{\partial I}{\partial \eta} = \int_0^1 \left[\frac{\partial F}{\partial \tilde{u}} \frac{\partial \tilde{u}}{\partial \eta} + \frac{\partial F}{\partial \tilde{u}_x} \frac{\partial u_x}{\partial \eta}\right] dx \tag{16}$$

Differentiation of Equation (15) into

$$\frac{\partial \tilde{u}}{\partial \eta} = v(x), \quad \text{and,} \quad \frac{\partial \tilde{u}_x}{\partial \eta} = \frac{\partial v}{\partial x} \tag{17}$$

when substituted into (16) and then integrated by parts yields

$$\frac{\partial I}{\partial \eta} = \int_0^1 \left[\frac{\partial F}{\partial \tilde{u}} v(x) - v(x)\frac{\partial}{\partial x}\left(\frac{\partial F}{\partial \tilde{u}_x}\right)\right] dx \tag{18}$$

At the minimum point, $\tilde{u} = u$ and $\eta = 0$ and

$$\frac{\partial I}{\partial \eta} = 0 \tag{19}$$

For this last equation to be valid, the bracketed expression of Equation (18) must hold for any arbitrary $v(x)$ which satisfies the boundary conditions. This results in the Euler-Lagrange equation

$$\frac{\partial F}{\partial u} - \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) = 0 \tag{20}$$

By comparing this equation to Equation (10) rewritten as

$$\left(\frac{\partial u}{\partial \theta}\right) - \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right) = 0 \tag{21}$$

it is noted that

$$\frac{\partial F}{\partial u} = \frac{\partial}{\partial \theta}(u) \tag{22}$$

and

$$\frac{\partial F}{\partial u_x} = \frac{\partial u}{\partial x} \tag{23}$$

Integration of both Equations (22) and (23) yields, respectively

$$F = \frac{\partial}{\partial \theta}\left(\frac{u^2}{2}\right) + f(u_x) \tag{24}$$

and

$$F = \frac{(u_x)^2}{2} + g(u) \tag{25}$$

The functions f and g are found by comparing these last two equations. The functional, F, is

$$F = \frac{1}{2}\left[\frac{\partial}{\partial \theta}(u^2) + \left(\frac{\partial u}{\partial x}\right)^2\right] \tag{26}$$

The desired variational statement to the differential equation is then

$$I = \frac{1}{2}\int_0^1 \left[\frac{\partial}{\partial \theta}(u^2) + \left(\frac{\partial u}{\partial x}\right)^2\right] dx \tag{27}$$

Quadratic Finite Element Application

Finite Element Formulation. With the variational statement established, the finite-element formulation can be started to obtain an approximate solution for the temperature as a function of x. Figure 4 shows the physical problem and displays an appropriate finite-element arrangement for solution of Equation (10). In the figure, the interval is divided into E elements $(E = 6)$, with N nodes $(N = 7)$. The exact solution is best considered as a continuous line running from the origin to the last node. For example, see Figure 3.

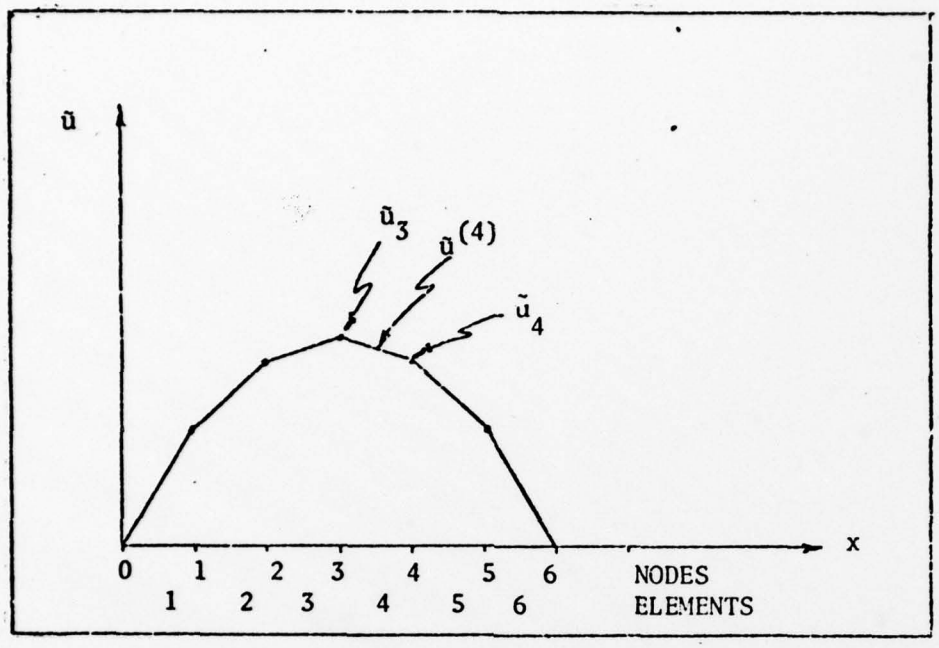

Figure 4. Finite Element Arrangement for Solution of Eq. (10).

The integral of Equation (27) is evaluated by breaking it up into E subintegrals over each of the E elements. For example,

$$I = I^{(1)} + I^{(2)} + \cdots + I^{(e)} + \cdots + I^{(E)} = \sum_{e=1}^{E} I^{(e)} \qquad (28)$$

where the integral $I^{(e)}$ over a typical finite-element (e) is given by

$$I^{(e)} = \frac{1}{2} \int_{i}^{i+1} \left[\frac{\partial}{\partial \theta} \left(u^{(e)} \right)^2 + \left(\frac{\partial u^{(e)}}{\partial x} \right)^2 \right] dx \qquad (29)$$

Equation (27) may be written, therefore, as

$$I(u) = \frac{1}{2}\int_0^1 \left[\frac{\partial (u^{(E)})^2}{\partial \theta} + \left(\frac{\partial u^{(E)}}{\partial x}\right)^2\right] dx \tag{30}$$

For simplicity, Equation (30) may be divided as

$$I_1 = \frac{1}{2}\int_0^1 \left(\frac{\partial u^{(E)}}{\partial x}\right)^2 dx \tag{31}$$

and

$$I_2 = \frac{1}{2}\int_0^1 \frac{\partial}{\partial \theta}\left(u^{(E)}\right)^2 dx \tag{32}$$

An extensive algebraic procedure is performed in Myers (Ref 8:334-339). The following formulation is developed using the matrix procedures of that same source. Note that the elements are represented by e, $e = 1$ to E, and written as superscripts $^{(e)}$. The nodes are represented by i, $i = 1$ to N, and written as subscripts $_i$.

As observed, the integral to be minimized is a function of the nodal temperatures, that is

$$I = I(u_1, u_2 \cdots u_i, u_j, \cdots u_N) \tag{33}$$

To find the minimum, I is differentiated with respect to the nodal temperatures and set equal to zero. If

$$\underline{u}^{(E)} = \begin{Bmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ u_{N-1} \\ u_N \end{Bmatrix} \tag{34}$$

then, for minimization

$$\frac{dI}{d\underline{u}^{(E)}} = \frac{dI_1}{d\underline{u}^{(E)}} + \frac{dI_2}{d\underline{u}^{(E)}} \tag{35}$$

The problem is depicted in Figure 5, in which the interval is divided into E elements, each considered separately, as for example, the temperature distribution of the element between nodes i and i + 1 . It should also be noted that it is here that the quadratic interpolation is introduced and depicted by the curve of alternating dots and dashes. If the node i + 1 is defined as node j , then the imaginary node of the quadratic function may be defined as k and inserted as depicted.

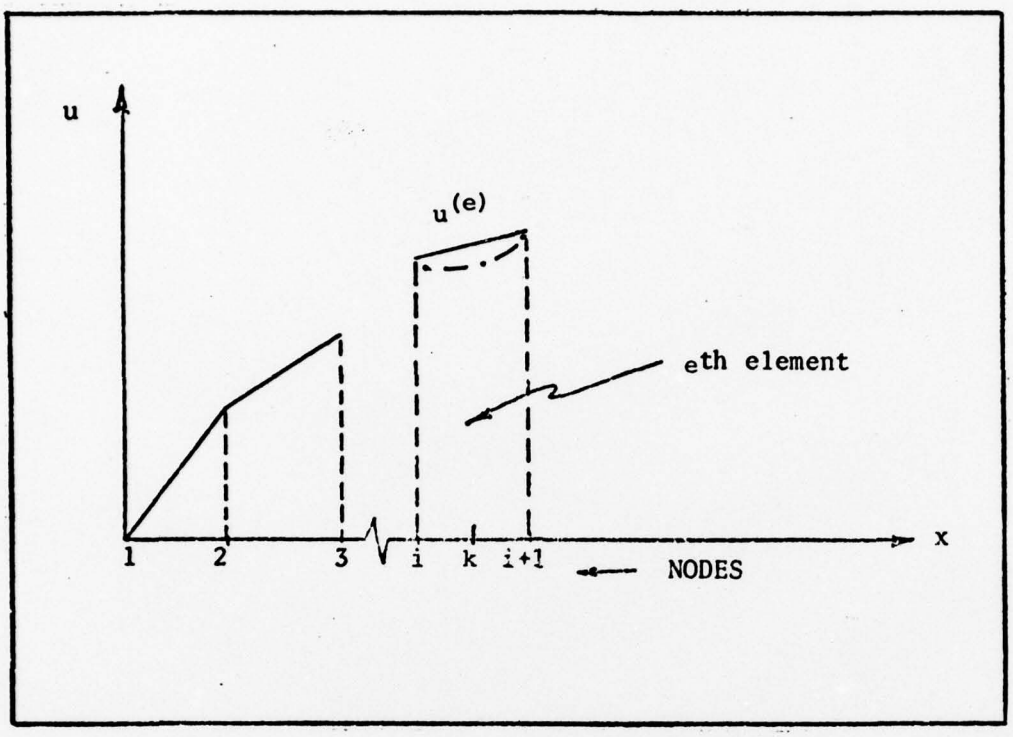

Figure 5. Arrangement of the Elements

Quadratic Interpolation Function. Although linear interpolation functions are easiest to mathematically formulate, greater accuracy can be expected by employing higher order element interpolations. The first higher order element is formulated by placing an interior node between the exterior nodes and employing a quadratic interpolation function as shown in Figure 5. Polynomials are most widely used as the interpolation functions because of their mathematical simplicity.

For example, in the one-dimensional problem of this thesis, a general nth-order polynomial may be written as

$$P_n(x) = \sum_{i=0}^{T_n^{(1)}} \alpha_i x^i \tag{36}$$

where the number of terms in the polynomial is $T_n^{(1)} = n + 1$ (Ref 5:131). Whereas for the linear case, the polynomial is written as

$$P_1(x) = \alpha_0 + \alpha_1 x \tag{37}$$

the quadratic polynomial is

$$P_2(x) = \alpha_o + \alpha_1 x + \alpha_2 x^2 \tag{38}$$

Huebner (Ref 5:79-81) lists requirements for the application of interpolation functions. These requirements, called compatibility and completeness, stem from the need to ensure that Equation (28) holds and that the approximate solution converges to the correct solution when an increasing number of smaller elements are used.

Quadratic Matrix Formulation. With the internal node required of the first higher order quadratic interpolation function, Equation (34) may be written as

$$\underline{u}^{(E)} = \begin{Bmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ u_i \\ u_k \\ u_j \\ \cdot \\ \cdot \\ \cdot \\ u_{N-1} \\ u_N \end{Bmatrix} \tag{39}$$

The derivative of $I^{(e)}$ with respect to $\underline{u}$ is a column matrix that is mostly zero because $I^{(e)}$ depends only on the particular u_i , u_k , and u_j . If the horizontal position of u_k is assumed midway between u_i and u_j , then its x location is defined as

$$x_k = \frac{x_i + x_j}{2} \tag{40}$$

Instead of differentiating the elemental integrals with respect to each component of $\underline{u}^{(E)}$, a matrix $\underline{D}^{(E)}$ is defined by

$$\underline{D}^{(e)} = \begin{bmatrix} 0 & 0 & 0 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 0 & 0 & 0 \end{bmatrix} \begin{matrix} \\ \\ \\ \text{ith row} \\ \text{kth row} \\ \text{jth row} \\ \\ \\ \\ \end{matrix} \tag{41}$$

and used as follows:

$$\frac{dI_1}{d\underline{u}^{(E)}} = \sum_{e=1}^{E} \underline{D}^{(e)} \frac{dI_1^{(e)}}{d\underline{u}^{(e)}} \tag{42}$$

and

$$\frac{dI_2}{d\underline{u}^{(E)}} = \sum_{e=1}^{E} \underline{D}^{(e)} \frac{dI_2^{(e)}}{d\underline{u}^{(e)}} \tag{43}$$

where $I^{(e)}$ is the portion of I defined on the eth interval, (i,k) and (k,j). Also, it is important to note that for the quadratic function

$$\underline{u}^{(e)} = \begin{Bmatrix} u_i \\ u_k \\ u_j \end{Bmatrix} \tag{44}$$

That is, for a particular element, e , the quadratic temperature distribution equation is given as

$$u^{(e)} = c_1^{(e)} + c_2^{(e)} x + c_3^{(e)} x^2 \tag{45}$$

where the temperature at each exterior node is defined by

$$u_i^{(e)} = c_1^{(e)} + c_2^{(e)} x_i + c_3^{(e)} x_i^2 \tag{46}$$

and

$$u_j^{(e)} = c_1^{(e)} + c_2^{(e)} x_j + c_3^{(e)} x_j^2 \tag{47}$$

By Equation (40), the temperature at the imaginary node location can be written as

$$u_k^{(e)} = c_1^{(e)} + c_2^{(e)} \left(\frac{x_i + x_j}{2} \right) + c_3^{(e)} \left(\frac{x_i + x_j}{2} \right)^2 \tag{48}$$

In the continuing matrix formulation, written in moderate detail because of its absence from other literature, $u^{(e)}$ may be written

$$u^{(e)} = \underline{p}^T c^{(e)} \tag{49}$$

where

$$\underline{p}^T = \lfloor 1 \; x \; x^2 \rfloor \tag{50}$$

and

$$\underline{c}^{(e)} = \begin{Bmatrix} c_1^{(e)} \\ c_2^{(e)} \\ c_3^{(e)} \end{Bmatrix} \tag{51}$$

These coefficients, Equation (51), are found by solving Equations (46), (47), and (48) simultaneously to yield (see Appendix C)

$$c_1^{(e)} = \frac{1}{\Delta x^2} [(x_i x_j + x_j^2) u_i - 4(x_i x_j) u_k + (x_i x_j + x_i^2) u_j] \tag{52}$$

$$c_2^{(e)} = \frac{1}{\Delta x^2} [-(x_i + 3x_j) u_i + 4(x_i + x_j) u_k - (3x_i + x_j) u_j] \tag{53}$$

$$c_3^{(e)} = \frac{1}{\Delta x^2} [2u_i - 4u_k + 2u_j] \tag{54}$$

where, for all intervals assumed equal

$$(\Delta x)^2 = (x_i - x_j)^2 \tag{55}$$

The coefficients are eliminated by substituting Equation (45) into Equation (44). The result is

$$u^{(e)} = \underline{p}^T \underline{R}^{(e)} \underline{u}^{(e)} \tag{56}$$

where

$$\underline{R}^{(e)} \underline{\Delta} \frac{1}{\Delta x^2} \begin{bmatrix} (x_i x_j + x_j^2) & -4(x_i x_j) & (x_i x_j + x_i^2) \\ -(x_i + 3x_j) & 4(x_i + x_j) & -(3x_i + x_j) \\ 2 & -4 & 2 \end{bmatrix} \tag{57}$$

Also, it is noted that

$$\underline{R}^{(e)} = \underline{p}^{(e)-1} \tag{58}$$

where

$$\underline{p}^{(e)} = \frac{\underline{u}^{(e)}}{\underline{c}^{(e)}} \tag{59}$$

By next taking the following derivative

$$\frac{\partial u^{(e)}}{\partial x} = \underline{p_x}^T \underline{R}^{(e)} \underline{u}^{(e)} \tag{60}$$

where

$$\underline{p_x}^T = \frac{\partial}{\partial x}(\underline{p}^T) = [\,0\ 1\ 2x\,] \tag{61}$$

and substituting it into $I_1^{(e)}$, the result is

$$I_1^{(e)} = \frac{1}{2}\int_{x_i}^{x_j}\left(\underline{p_x}^T\,\underline{R}^{(e)}\,\underline{u}^{(e)}\right)^2 dx \tag{62}$$

I_1 is then differentiated with respect to $\underline{u}^{(e)}$ to yield (Ref 7:41)

$$\frac{dI_1^{(e)}}{d\underline{u}^{(e)}} = \int_{x_i}^{x_j}\left(\underline{p_x}^T\,\underline{R}^{(e)}\,\underline{u}^{(e)}\right)\frac{d}{d\underline{u}^{(e)}}\left(\underline{p_x}^T\,\underline{R}^{(e)}\,\underline{u}^{(e)}\right)\,dx \tag{63}$$

or

$$\frac{dI_1^{(e)}}{d\underline{u}^{(e)}} = \int_{x_i}^{x_j}\left(\underline{p_x}^T\,\underline{R}^{(e)}\right)^T\left(\underline{p_x}^T\,\underline{R}^{(e)}\,\underline{u}^{(e)}\right)dx \tag{64}$$

where the order of the scalar terms $(\underline{p_x}\,\underline{R}^{(e)}\,\underline{u}^{(e)})$ has been rearranged. Because

$$(\underline{A}\,\underline{B})^T = \underline{B}^T\,\underline{A}^T \tag{65}$$

that is, the transpose of the product of two matrices is the product of the transposes in the reverse order, and because

$\underline{R}^{(e)}$ and $\underline{u}^{(e)}$ are independent of x and can be removed from the integral, Equation (64) is equivalent to

$$\frac{d\underline{l}_1^{(e)}}{d\underline{u}^{(e)}} = \underline{R}^{(e)T} \int_{x_i}^{x_j} \underline{p_x}\, \underline{p_x}^T dx \; \underline{R}^{(e)} \; \underline{u}^{(e)} \tag{66}$$

It is now necessary to perform the operations indicated by the last equation. Taking the bracketed product yields

$$\int_{x_i}^{x_j} \begin{Bmatrix} 0 \\ 1 \\ 2x \end{Bmatrix} \begin{bmatrix} 0 & 1 & 2x \end{bmatrix} dx = \int_{x_i}^{x_j} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 2x \\ 0 & 2x & 4x^2 \end{bmatrix} dx \tag{67}$$

Performing the integration over each term and factoring out $x = x_j - x_i = x_{ij}$ yields

$$\int_{x_i}^{x_j} \underline{p_x}\, \underline{p_x}^T dx = x_{ij} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & (x_j+x_i) \\ 0 & (x_j+x_i) & \frac{4}{3}(x_j^2+x_i x_j+x_i^2) \end{bmatrix} \tag{68}$$

Next, the pre-multiplication by $\underline{R}^{(e)T}$ is performed to yield

$$\frac{1}{\Delta x}\begin{bmatrix} 0 & -\Delta x & (5/3x_i^2-4/3x_ix_j-1/3x_j^2) \\ 0 & 0 & (-4/3x_i^2+8/3x_ix_j-4/3x_j^2) \\ 0 & \Delta x & (-1/3x_i^2-4/3x_ix_j+5/3x_j^2) \end{bmatrix} \tag{69}$$

which, by extracting $(x_i - x_j)$ from all terms of the third column is simplified to

$$\begin{bmatrix} 0 & -1 & -1/3(5x_i+x_j) \\ 0 & 0 & -4/3\Delta x \\ 0 & 1 & \frac{1}{3}(x_i+5x_j) \end{bmatrix} = \underline{R}^{(e)T}\int_{x_i}^{x_j} \underline{p_x}\,\underline{p_x}^T\,dx \tag{70}$$

Finally, the post-multiplication of this result by $\underline{R}^{(e)}$ is performed to yield

$$\underline{R}^{(e)T}\int_{x_i}^{x_j} \underline{p_x}\,\underline{p_x}^T\,dx\underline{R}^{(e)} = \frac{1}{3\Delta x}\begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix} \tag{71}$$

or

$$\frac{dI_1^{(e)}}{d\underline{u}^{(e)}} = \frac{1}{3\Delta x} \begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix} \underline{u}^{(e)} \tag{72}$$

By defining the element stiffness matrix (also called element conduction matrix (Ref 8:352) as

$$\underline{K}^{(e)} \triangleq \frac{1}{3\Delta x} \begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix} \tag{73}$$

Equation (42) may be written

$$\frac{dI_1}{d\underline{u}^{(E)}} = \sum_{e=1}^{E} \underline{D}^{(e)} \underline{K}^{(e)} \underline{u}^{(e)} \tag{74}$$

or

$$\frac{dI_1}{d\underline{u}^{(E)}} = \sum_{e=1}^{E} \underline{D}^{(e)} \underline{K}^{(e)T} \underline{u}^{(E)} \tag{75}$$

Strang and Fix (Ref 10:55) verify this result without justification. Further defining $\underline{K}$ as a global stiffness matrix

$$\underline{K} \triangleq \sum_{e=1}^{N} \underline{D}^{(e)} \underline{K}^{(e)} \underline{D}^{(e)T} \tag{76}$$

and substituting it into Equation (75) yields

$$\frac{dI_1}{d\underline{u}^{(E)}} = \underline{K}\,\underline{u}^{(E)} \tag{77}$$

Equation (43) is now approached in a similar manner. If the elemental representation of Equation (32) is differentiated by Leibnitz' rule for differentiation of integrals, the result is

$$I_2^{(e)} = \frac{1}{2}\frac{d}{d\theta}\int_{x_i}^{x_j} (u^{(e)})^2\, dx \tag{78}$$

or

$$I_2^{(e)} = \frac{1}{2}\frac{d}{d\theta}\int_{x_i}^{x_j} (\underline{p}^T \underline{R}^{(e)} \underline{u}^{(e)})^2\, dx \tag{79}$$

The derivative of this last equation is taken as follows to yield (Ref 8:355)

$$\frac{dI_2^{(e)}}{d\underline{u}^{(e)}} = \frac{d}{d\theta} \int_{x_i}^{x_j} (\underline{p}^T \underline{R}^{(e)} \underline{u}^{(e)}) (\underline{p}^T \underline{R}^{(e)})^T dx \tag{80}$$

or

$$\frac{dI_2^{(e)}}{d\underline{u}^{(e)}} = \frac{d}{d\theta} \underline{R}^{(e)T} \int_{x_i}^{x_j} \underline{p_x}\, \underline{p_x}^T dx\, \underline{R}^{(e)}\, \underline{u}^{(e)} \tag{81}$$

Again, it is necessary to perform the operations indicated by the last equation. Taking the bracketed product yields

$$\int_{x_i}^{x_j} \begin{Bmatrix} 1 \\ x \\ x^2 \end{Bmatrix} \begin{bmatrix} 1 & x & x^2 \end{bmatrix} dx = \int_{x_i}^{x_j} \begin{bmatrix} 1 & x & x^2 \\ x & x^2 & x^3 \\ x^2 & x^3 & x^4 \end{bmatrix} dx \tag{82}$$

Performing the integration over each term yields

$$\int_{x_i}^{x_j} \underline{p_x}\, \underline{p_x}^T dx = \begin{bmatrix} (x_j - x_i) & \frac{1}{2}(x_j^2 - x_i^2) & \frac{1}{3}(x_j^3 - x_i^3) \\ \frac{1}{2}(x_j^2 - x_i^2) & \frac{1}{3}(x_j^3 - x_i^3) & \frac{1}{4}(x_j^4 - x_i^4) \\ \frac{1}{3}(x_j^3 - x_i^3) & \frac{1}{4}(x_j^4 - x_i^4) & \frac{1}{5}(x_j^5 - x_i^5) \end{bmatrix} \tag{83}$$

If each term is factored with an (x_j-x_i) term and this term further factored from the entire matrix as Δx , the result is Equation (84)

$$\Delta x \begin{bmatrix} 1 & \frac{1}{2}(x_j+x_i) & \frac{1}{3}(x_j^2+x_jx_i+x_i^2) \\ \frac{1}{3}(x_j+x_i) & \frac{1}{3}(x_j^2+x_jx_i+x_i^2) & \frac{1}{4}(x_j^3+x_jx_i^2+x_j^2x_i+x_i^3) \\ \frac{1}{3}(x_j^2+x_jx_i+x_i^2) & \frac{1}{4}(x_j^3+x_jx_i^2+x_j^2x_i+x_i^3) & \frac{1}{5}(x_j^4+x_j^3x_i+x_j^2x_i^2+x_jx_i^3+x_i^4) \end{bmatrix}$$

Next, the pre-multiplication by $\underline{R}^{(e)T}$ is performed to yield Equation (85)

$$\frac{1}{\Delta x} \begin{bmatrix} \frac{x_i^2-2x_ix_j+x_j^2}{6} & \frac{x_i^3-2x_i^2x_j+x_ix_j^2}{6} & \frac{9x_i^4-16x_i^3x_j+4x_i^2x_j^2+4x_ix_j^3-x_j^4}{60} \\ \frac{2x_i^2-4x_ix_j+2x_j^2}{3} & \frac{x_i^3-x_ix_j^2-x_i^2x_j+x_j^3}{3} & \frac{3x_i^4-2x_i^3x_j-2x_i^2x_j^2-2x_ix_j^3+3x_j^4}{15} \\ \frac{x_i^2-2x_ix_j+x_j^2}{6} & \frac{x_i^2x_j-2x_ix_j^2+x_j^3}{6} & \frac{-x_i^4+4x_i^3x_j+4x_i^2x_j^2-16x_ix_j^3+9x_j^4}{60} \end{bmatrix}$$

which, when each term is factored with an $(x_i-x_j)^2/3$ and this factor removed from the matrix as $(\Delta x)^2/3$, the result is that

$$\underline{R}^{(e)T}\int_{x_i}^{x_j} \underline{p_x}\,\underline{p_x}^T\,dx = \frac{\Delta x}{3}\begin{bmatrix} \frac{1}{2} & \frac{x_i}{2} & \frac{(9x_i^2+2x_ix_j-x_j^2)}{20} \\ 2 & (x_i+x_j) & \frac{(3x_i^2+4x_ix_j+3x_j^2)}{5} \\ \frac{1}{2} & \frac{x_j}{2} & \frac{9x_j^2+2x_ix_j-x_i^2)}{20} \end{bmatrix} \quad (86)$$

Finally, the post-multiplication of this result by $\underline{R}^{(e)}$ is performed to yield

$$\underline{R}^{(e)T}\int_{x_i}^{x_j} \underline{p_x}\,\underline{p_x}^T\,dx\underline{R}^{(e)} = \frac{\Delta x}{15}\begin{bmatrix} 2 & 1 & -\frac{1}{2} \\ 1 & 8 & 1 \\ -\frac{1}{2} & 1 & 2 \end{bmatrix} \quad (87)$$

or

$$\frac{dI_2^{(e)}}{d\underline{u}^{(e)}} = \frac{d}{d\theta}\,\frac{\Delta x}{15}\begin{bmatrix} 2 & 1 & -\frac{1}{2} \\ 1 & 8 & 1 \\ -\frac{1}{2} & 1 & 2 \end{bmatrix} \quad (88)$$

By defining the element mass matrix as

$$\underline{M}^{(e)} \triangleq \frac{\Delta x}{15} \begin{bmatrix} 2 & 1 & -\frac{1}{2} \\ 1 & 8 & 1 \\ -\frac{1}{2} & 1 & 2 \end{bmatrix} \tag{89}$$

and employing Equation (88), Equation (43) may be written

$$\frac{dI_2}{d\underline{u}^{(E)}} = \sum_{e=1}^{E} \underline{D}^{(e)} \frac{d}{d\theta} (\underline{M}^{(e)} \underline{D}^{(e)T} \underline{u}^{(E)}) \tag{90}$$

in analogy to Equation (75). Also, since $\underline{M}^{(e)}$ and $\underline{D}^{(e)T}$ are independent of θ

$$\frac{dI_2}{d\underline{u}^{(E)}} = \sum_{e=1}^{E} \underline{D}^{(e)} \underline{M}^{(e)} \underline{D}^{(e)T} \frac{d}{d\theta} (\underline{u}^{(E)}) \tag{91}$$

Further, in analogy to Equation (76), a global mass matrix may be defined as

$$\underline{M} \triangleq \sum_{e=1}^{E} \underline{D}^{(e)} \underline{M}^{(e)} \underline{D}^{(e)T} \tag{92}$$

which, if substituted into Equation (91), yields

$$\frac{dI_2}{d\underline{u}^{(E)}} = \underline{M} \frac{d}{d\theta} (\underline{u}^{(E)}) \tag{93}$$

The process of minimization is now employed by setting the derivative, $dI/d\underline{u}^{(E)}$, equal to zero:

$$\frac{dI}{d\underline{u}^{(E)}} = \frac{dI_1}{d\underline{u}^{(E)}} + \frac{dI_2}{d\underline{u}^{(E)}} = 0 \tag{94}$$

or

$$\underline{K}\,\underline{u}^{(E)} + \underline{M} \frac{d}{d\theta} (\underline{u}^{(E)}) = 0 \tag{95}$$

This equation, also derived in Strang and Fix (Ref 10:243), may be written as

$$\underline{M} \frac{d}{d\theta} (\underline{u}^{(E)}) = - \underline{K}\,\underline{u}^{(E)} \tag{96}$$

and represents a system of ordinary differential equations for the nodal temperatures as functions of time. At time zero, the initial temperature distribution is given which is substituted into the right side. The system of equations is then solved directly for the initial time derivatives necessary to minimize I at that instant. These derivatives are then used to move ahead in time (Ref 8:404).

If the number of elements, E , is large, the system may be solved using finite-differences to approximate the time derivative. Myers states three schemes, the Euler explicit scheme, the Crank-Nicolson scheme, and the fully implicit scheme, which may be used in this finite difference application. Martin applies these methods directly to the system of equations (96) and notes that there exists a general scheme to accommodate the methods (Ref 7:47-49). This scheme is given by

$$(\underline{M} + \underline{K}\alpha\Delta\theta)(\underline{u}^{(E)})^{k+1} = (\underline{M} - \underline{K}(1-\alpha)\Delta\theta)(\underline{u}^{(E)})^{k} \tag{97}$$

where

$\Delta\theta$ = change in time

$(\underline{u}^{(E)})^{k+1}$ = temperature at time step, k+1

$(\underline{u}^{(E)})^{k}$ = temperature at time step, k

α = method designation parameter; that is:

$\alpha = 0$ for Euler, $\alpha = .5$ for Crank-Nicolson,

$\alpha = 1$ for fully implicit.

If the matrices $\underline{A}$ and $\underline{B}$ are defined by

$$\underline{A} = \underline{M} + \underline{K}\alpha\Delta\theta \tag{98}$$

and

$$\underline{B} = \underline{M} - \underline{K}(1-\alpha)\Delta\theta \tag{99}$$

Equation (97) may be written

$$\underline{A}\ (\underline{u}^{(E)})^{k+1} = \underline{B}\ (\underline{u}^{(E)})^{k} \tag{100}$$

The matrices $\underline{A}$ and $\underline{B}$ may now be generated as

$$\underline{A} = \frac{\Delta x}{15}\begin{bmatrix} 2 & 1 & -\frac{1}{2} \\ 1 & 8 & 1 \\ -\frac{1}{2} & 1 & 2 \end{bmatrix} + \frac{\alpha\Delta\theta}{3\Delta x}\begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix} \tag{101}$$

and

$$\underline{B} = \frac{\Delta x}{15}\begin{bmatrix} 2 & 1 & -\frac{1}{2} \\ 1 & 8 & 1 \\ -\frac{1}{2} & 1 & 2 \end{bmatrix} - \frac{(1-\alpha)\Delta\theta}{3\Delta x}\begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix} \tag{102}$$

which, when the Fourier modulus

$$p = \frac{\Delta\theta}{(\Delta x)^2} \tag{103}$$

is defined and applied, become

$$\underline{A} = \frac{\Delta x}{15} \begin{bmatrix} 2+35\alpha p & 1-40\alpha p & -\frac{1}{2}+5\alpha p \\ 1-40\alpha p & 8+80\alpha p & 1-40\alpha p \\ -\frac{1}{2}+5\alpha p & 1-40\alpha p & 2+35\alpha p \end{bmatrix} \tag{104}$$

and

$$\underline{B} = \frac{\Delta x}{15} \begin{bmatrix} 2-35(1-\alpha)p & 1+40(1-\alpha)p & -\frac{1}{2}-5(1-\alpha)p \\ 1+40(1-\alpha)p & 8-80(1-\alpha)p & 1+40(1-\alpha)p \\ -\frac{1}{2}-5(1-\alpha)p & 1+40(1-\alpha)p & 2-35(1-\alpha)p \end{bmatrix} \tag{105}$$

The matrices (104) and (105) are similar to each other and similar to their linear counterparts depicted in Appendix G for the case where the interval has been divided into three elements and four external nodes. Note that p is as given by Equation (103), and not by Equation (50).

Assembling the matrices for the quadratic case, three elements and four external nodes, yields

$$
\begin{bmatrix}
2+35\alpha p & 1-40\alpha p & -\tfrac{1}{2}+5\alpha p & 0 & 0 & 0 & 0 \\
1-40\alpha p & 8+80\alpha p & 1-40\alpha p & 0 & 0 & 0 & 0 \\
-\tfrac{1}{2}+5\alpha p & 1-40\alpha p & 4+70\alpha p & 1-40\alpha p & -\tfrac{1}{2}+5\alpha p & 0 & 0 \\
0 & 0 & 1-40\alpha p & 8+80\alpha p & 1-40\alpha p & 0 & 0 \\
0 & 0 & -\tfrac{1}{2}+5\alpha p & 1-40\alpha p & 4+70\alpha p & 1-40\alpha p & -\tfrac{1}{2}+5\alpha p \\
0 & 0 & 0 & 0 & 1-40\alpha p & 4+80\alpha p & 1-40\alpha p \\
0 & 0 & 0 & 0 & -\tfrac{1}{2}+5\alpha p & 1-40\alpha p & 2+35\alpha p
\end{bmatrix}
\begin{Bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \end{Bmatrix}^{k+1}
$$

$$
=
\begin{bmatrix}
2-35p(1-\alpha) & 1+40p(1-\alpha) & -\tfrac{1}{2}-5p(1-\alpha) & 0 & 0 & 0 & 0 \\
1+40p(1-\alpha) & 8-80p(1-\alpha) & 1+40p(1-\alpha) & 0 & 0 & 0 & 0 \\
-\tfrac{1}{2}-5p(1-\alpha) & 1+40p(1-\alpha) & 4-70p(1-\alpha) & 1+40p(1-\alpha) & -\tfrac{1}{2}-5p(1-\alpha) & 0 & 0 \\
0 & 0 & 1+40p(1-\alpha) & 8-80p(1-\alpha) & 1+40p(1-\alpha) & 0 & 0 \\
0 & 0 & -\tfrac{1}{2}-5p(1-\alpha) & 1+40p(1-\alpha) & 4-70p(1-\alpha) & 1+40p(1-\alpha) & -\tfrac{1}{2}-5p(1-\alpha) \\
0 & 0 & 0 & 0 & 1+40p(1-\alpha) & 4-80p(1-\alpha) & 1+40p(1-\alpha) \\
0 & 0 & 0 & 0 & -\tfrac{1}{2}-5p(1-\alpha) & 1+40p(1-\alpha) & 2-35p(1-\alpha)
\end{bmatrix}
\begin{Bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \end{Bmatrix}^{k}
\qquad (106)
$$

where u_2 , u_4 , u_6 represent the solutions at the internal nodes. Because of the constant temperature condition on the boundaries, and since, except for the first instant in time, the imposed boundary temperature is zero, the matrices become

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & a_{22} & a_{23} & 0 & 0 & 0 & 0 \\ 0 & a_{32} & a_{33} & a_{34} & a_{35} & 0 & 0 \\ 0 & 0 & a_{43} & a_{44} & a_{45} & 0 & 0 \\ 0 & 0 & a_{53} & a_{54} & a_{55} & a_{56} & 0 \\ 0 & 0 & 0 & 0 & a_{65} & a_{66} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{Bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \end{Bmatrix}^{k+1}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & b_{22} & b_{23} & 0 & 0 & 0 & 0 \\ 0 & b_{32} & b_{33} & b_{34} & b_{35} & 0 & 0 \\ 0 & 0 & b_{43} & b_{44} & b_{45} & 0 & 0 \\ 0 & 0 & b_{53} & b_{54} & b_{55} & b_{56} & 0 \\ 0 & 0 & 0 & 0 & b_{65} & b_{66} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{Bmatrix} u_B \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_B \end{Bmatrix}^{k} \tag{107}$$

where the lower case letters represent the value at the indicated position in $\underline{A}$ or $\underline{B}$. Matrix assembly theory is given in Appendix D.

Optimum Implicit Conditions. For this system of equations, an optimum α parameter representation can be found in analogy to the derivation of the linear optimum α equation attained by Martin and

shown in Appendix G. This quadratic derivation, depicted in Appendix E, is also determined by finding the truncation error across two connecting elements using a Taylor series expansion. It should be noted that the expansion is now a ten point representation, based on the five point rows of the matrices $\underline{A}$ and $\underline{B}$, each point being considered an integral node, and separated by Δx. If α is chosen according to the formula so derived, that is

$$\alpha = \frac{1}{3}\left(2 - \frac{1}{4p}\right) \tag{108}$$

then the resulting expression in Equation (107) is fourth order accurate at the nodes. The truncation error at these locations is proportional to $(\Delta x)^4$. The Euler, Crank-Nicolson, and fully implicit schemes are only second order accurate.

Appendix E also considers a truncation error expansion about the internal node of an element as based on the three point rows of the matrices $\underline{A}$ and $\underline{B}$. The resulting expression indicates that for this quadratic approach, only second order accuracy is attainable within an element. For second order accuracy, Appendix E states that

$$\frac{(160)}{p} = 0 \tag{109}$$

or, if α is chosen according to this equation, the resulting expression in Equation (107) is no more accurate than the other schemes just mentioned.

The dilemma posed by Equation (109), that is, of lesser accuracy for solution by quadratic interpolation function than by linear interpolation function, can be eliminated by treating the internal node problem. One possible course of action would be to reduce the system matrices of this time dependent problem to matrices representing only external nodes and corresponding temperatures, in analogy to the method of "static condensation" summarized in Appendix F.

A more direct method of elimination would be to treat all nodes as external nodes. The derivation leading to Equation (108) indicates that the order of accuracy attained by employing the quadratic interpolation function is equal to the order of accuracy attained by using the linear interpolation function. This fact tends to indicate that, as such, the quadratic approach is equivalent to the linear approach. Appendix G verifies this hypothesis by displaying the equivalence of quadratic and linear alphas. Appendix G also contains the linear optimum alpha equation and the linear matrices $\underline{A}$ and $\underline{B}$ corresponding to Equation (106). The following sections are based on this equivalance, which states that for this problem approach, the quadratic interpolation function solution may be found as equal to the corresponding linear solution across an interval $\Delta x/2$. That is, the linear solutions of double the number of nodes are equivalent to the quadratic solutions of the original mesh spacings.

Quadratic Solution Interpolation

The failure of the quadratic approach to achieve greater than fourth order accuracy was predicted by Martin as due to the nature of the solution approach. That is, by the Method of Weighted Residuals, even though the spatial variation has been handled by quadratic interpolated finite-elements, finite-differences were employed for the time variation as indicated in Equation (96). This factor and others are discussed in Appendix I. The following paragraphs discuss a variation of the quadratic derivations leading to greater accuracy of the equivalent linear solutions without applying smaller mesh spacings.

If the internal nodes were not considered in the derivations leading to the quadratic optimum alpha, the result would be

$$\alpha = \frac{1}{3}\left(\frac{1}{2} - \frac{1}{4p}\right) \tag{110}$$

where the Fourier modulus is now defined as

$$p = \frac{\Delta\theta}{\left(\frac{\Delta x}{2}\right)^2} = 4\left[\frac{\Delta\theta}{(\Delta x)^2}\right] \tag{111}$$

That is, the quadratic modulus, now interpreted as four times the linear modulus, when substituted into Equation (110), yields optimum alpha values of one-fourth the value found by the corresponding application of Equation (108). This is logical, because at any node, the particular point of solution can be found to be a function of αp or $(\alpha)(\Delta\theta)$. If alpha is decreased to one-fourth its

original value, the modulus, or here for fixed position, the time increment, must be cubed.

The implementation of these factors into the linear interpolation approach, that is, the employment of Equation (111) and the equivalent alphas of Equation (110), yields more accurate results for the selected original cases of $p = .5$ and $p = 2.0$. For the case $p = 1.0$, for which the alphas are, in fact, equal, the results are of equal accuracy. The results section and Appendix H display these findings.

Error Analysis

Background. One basic objective of this project was to compare the finite-element solutions using a quadratic interpolation function versus the solutions attained by using a linear interpolation function. To make this comparison and the other comparisons mentioned in the introduction, the quadratic function solutions must be attained and compared to the previously attained linear function solutions, and exact analytical solutions. Relative error magnitudes between these methods, and within a method for differing alpha parameters, were used to complete the accuracy study of this thesis. The stability study is discussed in a later section.

General. In the finite-element method, there are two ways to improve the accuracy of the approximate solution. The first way is to decrease the nodal interval, Δx, analogous to the ordinary finite-difference approach. The second way (this thesis study) is to apply a higher order polynomial approximating function. By its

nature, any error in the finite-element method must be measured over the entire interval and not just at the nodes.

Three error norms were chosen, as for the finite-difference linear approximation study (Ref 7:54-59), by which to complete the accuracy portion of this project.

Finite Element Error Analysis. The first measure of error is the pointwise or discretization error. It is defined as the difference between the exact analytical solution and discrete approximate solution at the node points. Alien to the concept of finite-elements, which attempts to minimize the error everywhere in an element, this error can be estimated in the pointwise sense by treating the individual equations in the system (107) as if they were simple difference equations.

Also, it is noted that the pointwise error is composed of a round-off error and truncation error. Within the limits of stability, the latter is the much larger of the two and the pointwise error is considered to be the truncation error. This truncation error, defined as that which results from elimination of the higher order derivative when Taylor's series is used to approximate a differential equation, is derived in Appendix E for the ith equation in the system. It applies to all nodes of a Dirichlet problem and is found to be

$$e_t(i\Delta x, k\Delta\theta) = \left(10p - 15\alpha p - \frac{15}{12}\right)(\Delta x)^2 \left.\frac{\partial^2 u}{\partial x^2}\right|_{\substack{x=i\Delta x \\ \theta=k\Delta\theta}} + O(\Delta x)^4 \qquad (112)$$

The second error measure is the discrete Tchebycheff norm, or the maximum error between the exact solution and the finite-element solution at any node. It is estimated by Equation (107).

The generalized mean error [also estimated by Equation (107)] is the last error measure employed. It indicates whether or not a higher or lower order of convergence would be noted for the first interior node, compared to the convergence at all of the nodes. The generalized mean error consists of the sum of the absolute values of the discretization errors at each non-zero node. It compares the pointwise error at the first interior node, $x = .1$, to the error at all the nodes.

Equation (112) can be used to ascertain the order of accuracy of the Crank-Nicolson, optimum implicit (Crandall), fully implicit, and explicit finite-element schemes. For the quadratic Crank-Nicolson equivalent, $\alpha = .6666$, and

$$e_t\,(i\Delta x, k\Delta\theta) = \left(-\frac{15}{12}\right)(\Delta x)^2 \left.\frac{\partial^2 u}{\partial x^2}\right|_{\substack{x=i\Delta x \\ \theta=k\Delta\theta}} + O(\Delta x)^4 + \cdots \tag{113}$$

or second order accurate. For the optimum implicit scheme, $\alpha = (2-1/4p)/3$, and

$$e_t\,(i\Delta x, k\Delta\theta) = 0\,(\Delta x)^4 + \cdots \tag{114}$$

or fourth order accurate. For the explicit scheme, $\alpha = 0$, and

$$e_t\,(i\Delta x, k\Delta\theta) = \left(10p - \frac{15}{12}\right)(\Delta x)^2 \left.\frac{\partial^2 u}{\partial x^2}\right|_{\substack{x=i\Delta x \\ \theta=k\Delta\theta}} + O(\Delta x)^4 + \cdots \qquad (115)$$

or second order accurate. Finally, for the pure implicit formulation, $\alpha = 1$, and

$$e_t\,(i\Delta x, k\Delta\theta) = \left(-5p - \frac{15}{12}\right)(\Delta x)^2 \left.\frac{\partial^2 u}{\partial x^2}\right|_{\substack{x=i\Delta x \\ \theta=k\Delta\theta}} + O(\Delta x)^4 + \cdots \qquad (116)$$

or, also second order accurate. Only the optimum implicit scheme is fourth order accurate with respect to truncation error.

Stability Analysis

General. The stability analysis is derived from a consideration of the round-off error, that error inherent in computer operations due to the finite number of significant figures it can manage. The error in the solution is the thing of interest. If the magnitude of the difference between the exact numerical solution and the truncated numerical solution grows exponentially as the calculation proceeds, then the numerical scheme is termed unstable.

The basic approach is to write Equation (107) in terms of error vectors (Ref 7:62-65) where the vector $\underline{e}_o$ represents the round-off error, and $\underline{e}_1$ represnets the new error after solution of the set of equations. With this,

$$\underline{A}\,\underline{e}_1 = \underline{B}\,\underline{e}_0 \tag{117}$$

or

$$\underline{e}_1 = \underline{C}\,\underline{e}_0 \tag{118}$$

where $\underline{C} = (\underline{A}^{-1}\underline{B})$. By expanding $\underline{e}_0$ in terms of the eigenvecotrs, $\underline{\phi}_i$, of $\underline{C}$, then

$$\underline{e}_1 = \underline{C} \sum_{i=1}^{n} c_i \phi_i \tag{119}$$

where c_i is a constant and $\underline{\phi}_i$ is the ith eigenvector of $\underline{C}$.
By the definition of an eigenvalue, Equation (117) is written

$$\underline{e}_1 = \sum_{i=1}^{n} c_i \lambda_i \underline{\phi}_i \tag{120}$$

where λ_i is the ith eigenvalue of the matrix $\underline{C}$. Likewise,

$$\underline{e}_2 = \underline{C}\,\underline{e}_1 = \sum_{i=1}^{n} c_i \lambda_i^2 \underline{\phi}_i \tag{121}$$

and, after k computations

$$\underline{e}_k = \sum_{i=1}^{n} c_i \lambda_i^k \underline{\phi}_i \tag{122}$$

This demonstrates that the eigenvalues of the iteration matrix, $\underline{C}$, determine the growth or decay of round-off errors. This procedure is summarized in the following Table I.

TABLE I

Oscillation and Instability Limits
By Eigenvalue Definition

Eigenvalue	Condition of Eigenfunction
$\lambda > 1$	Steady, unbounded growth, of same sign.
$0 > \lambda > 1$	Steady decay, of same sign.
$-1 < \lambda < 0$	Steady decay, alternating signs (stable oscillations).
$\lambda < -1$	Steady growth, alternating signs (unstable oscillations).

Finite-Element Analysis. Employing the given boundary conditions, Equation (107) is written

$$\underline{A}\,\underline{u}^{k+1} = \underline{B}\,\underline{u}^{k} \tag{123}$$

where the first and last equations have been dropped and the number of unknowns reduced by two. Because of the requirements of the equivalency argument, $\underline{A}$ and $\underline{B}$ are taken as in Equation (G-3), tridiagonal in form, and as shown in Equation (124).

$$\begin{bmatrix} 0 & 0 & & & \\ 0 & b & a & & \\ 0 & a & b & a & \\ & & a & b & 0 \\ & & & a & 0 \end{bmatrix} \tag{124}$$

The eigenvalues of such a matrix are given by

$$\lambda_n = b + 2a \cos\left(\frac{n\pi}{N+1}\right), \quad n = 1,2,\cdots N \tag{125}$$

where N is the matrix order (Ref 9:65). Also, the eigenvalues of the iteration matrix, $\underline{C} = \underline{A}^{-1}\underline{B}$, in Equation (123) are given by

$$(\lambda_C)_n = \frac{(\lambda_B)_n}{(\lambda_A)_n} \tag{126}$$

If Equation (125) is substituted into Equation (126) for the limiting case of $N \to \infty$, for which

$$\lim_{n=N\to\infty} \cos\left(\frac{N\pi}{N+1}\right) = -1 \tag{127}$$

the result, called the critical or minimum eigenvalue, is found to be

$$(\lambda c)_{\infty} = \frac{1+12\alpha p}{1-12(1-\alpha)p} \tag{128}$$

This is the equation derived by Martin in his study of finite-element solutions by linear interpolation. Therefore, the same oscillation and stability limits of that approach apply here and are shown in Table II and Figure 6. For quick reference, Table III and Figure 7 show these limits for the finite-difference formulation.

It would be interesting to note the stability and oscillation limits if the direct quadratic approach could be used. Ignoring the error of that approach, the quadratic system matrices were written in a modified condensed form, analogous to that of Appendix F. This modified condensation was nothing more than condensing both sides of Equation (106), holding the system constant in time. Assuming such a time constant system could represent this transient problem at successive isolated moments, the critical eigenvalue was found to be

$$(\lambda c)_{\infty} = \frac{2-15(1-\alpha)p}{2+15\alpha p} \tag{129}$$

The stability and oscillation limits are presented in Table IV. In general, these data show that solutions attained by this quadratic approach would be more restrictive in the oscillation limit for the optimum implicit scheme.

TABLE II

Oscillation and Instability Limits for the Fourier Modulus in the Finite-Element Method.

Limits	Euler	Crank-Nicolson	Optimum Implicit	Pure Implicit
Oscillation Limit, $p < x$ for no oscillation	.08333	.16667	.33333	Never Oscillates
Stability Limit, $p < x$ for a stable scheme	.16667	Always Stable	Always Stable	Always Stable

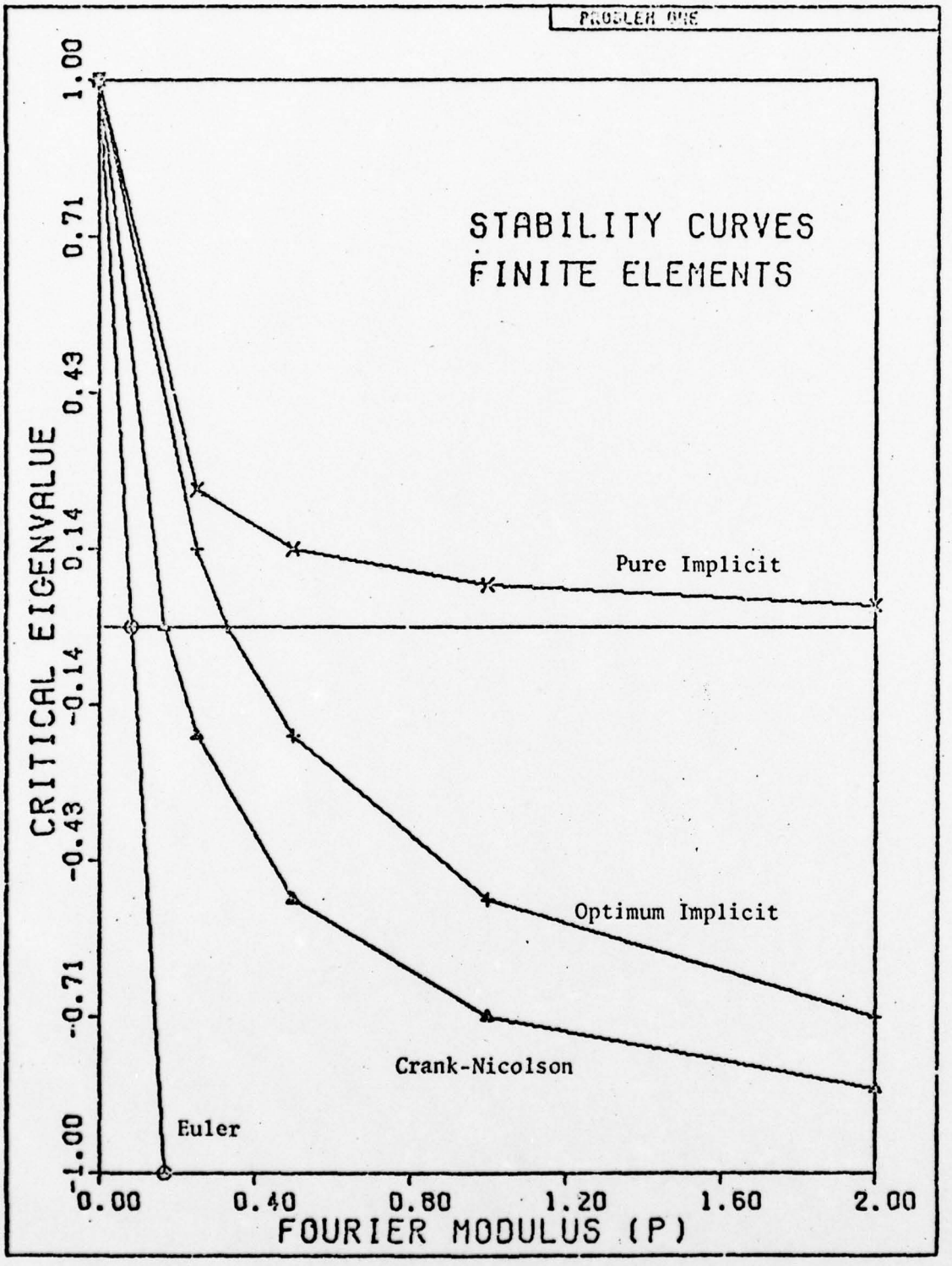

Figure 6. Stability Curves for Finite Elements.

TABLE III

Oscillation and Instability Limits for
the Fourier Modulus in
the Finite-Difference Formulation.

Limits	Explicit	Crandall	Crank-Nicolson	Pure-Implicit
Oscillation Limit, $p < x$ for no oscillation	0.25	.3333	.5	No Oscillations
Stability Limit, $p < x$ for stable scheme	0.5	Always Stable	Always Stable	Always Stable

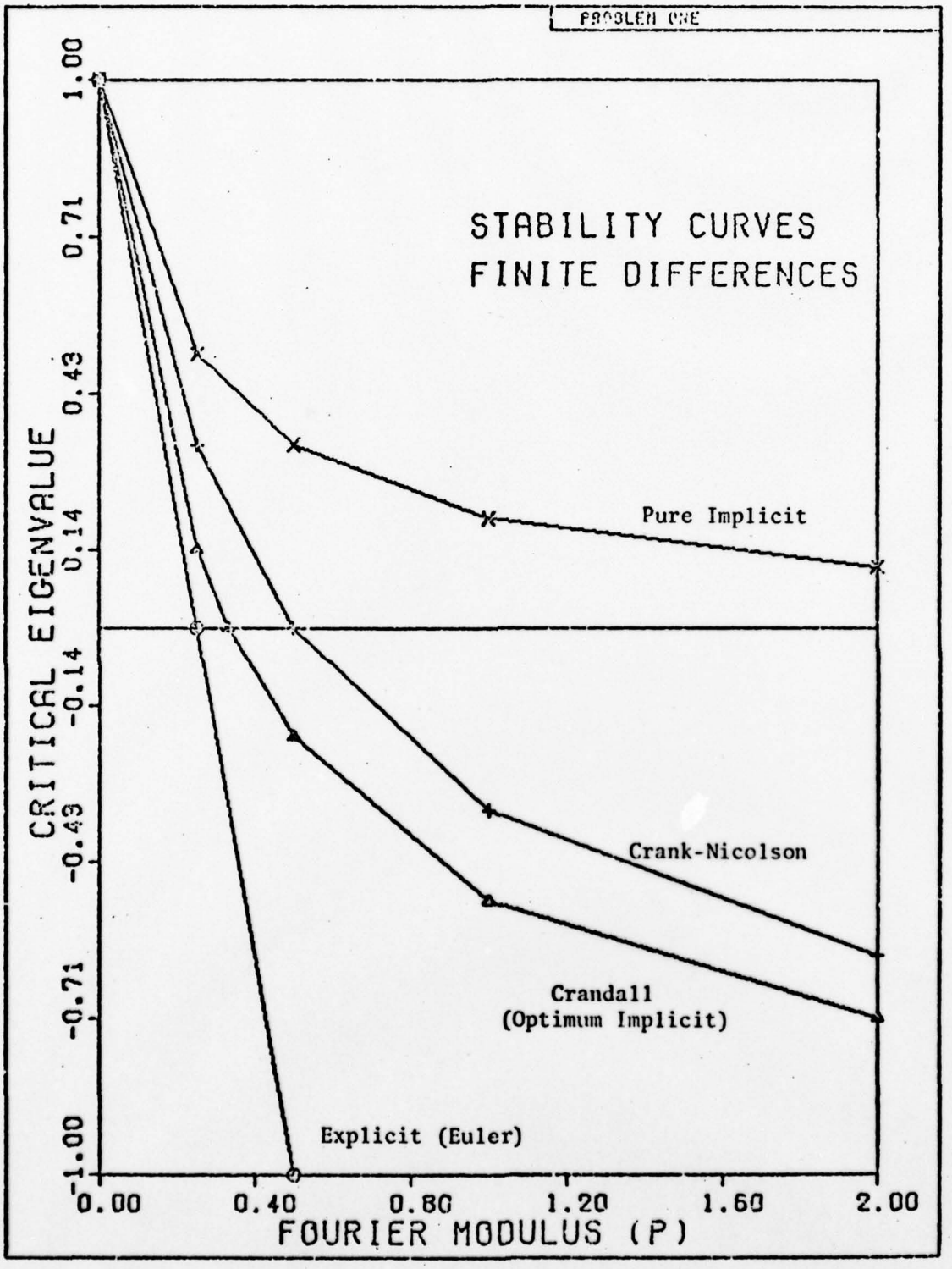

Figure 7. Stability Curves for Finite Differences.

TABLE IV

Oscillation and Instability Limits for the Fourier Modulus in the Finite-Element Method and Quadratic System

Limits	Euler	Crank-Nicolson	Optimum Implicit	Pure Implicit
Oscillation Limit, $p < x$ for no oscillation	.13333	.39999	.31999	Never Oscillates
Stability Limit, $p < x$ for a stable scheme	.26666	Always Stable	Always Stable	Always Stable

III. Procedure

General Approach

Initial Phase. Phase one of this thesis project began with a thorough study of finite-element theory, especially the calculus of variations. Also studied were the theory of heat conduction and the applicability of the finite-element procedure to this problem. Finite-difference theory was reviewed. A literature search was initiated and discussions conducted with several faculty memebers to ascertain the best approach for attacking this relatively unfamiliar problem. Meyers (Ref 8) was the primary text employed during the study, and although of an introductory level, proved the best source for procedural comparison. Huebner's text (Ref 5) was valuable in later portions of the project.

Second Phase. Phase two was started by examining the thesis proposal and determining how best to complete the proposal objectives. It was noted that this project was basically one of comparison; that is, to compare the finite-element results attained here using a quadratic interpolation function, to those attained previously (by Martin) using a linear interpolation function. The primary problem then was twofold; a quadratic analysis procedure had to be derived, something not visible in detail in the available literature, and this analysis applied to the same heat conduction problem in a manner that accurate comparisons could be made. The comparisons had to be as analogous as possible since errors in the results were assumed to be functions of truncation and round-off.

The greatest results, at least in theory, to allow an analogous approach, were the derivations of the system matrices, $\underline{A}$ and $\underline{B}$, and the truncation error e_t. Because of this truncation error analysis, the finite-element equations could be treated as difference equations to yield order of accuracy and optimum α values. Just as for the linear case, accuracy of $O(\Delta x)^4$ could be attained, and a theoretical expression for α, dependent only on the Fourier modulus, p, was achieved.

Also initiated during this phase was the reprogramming of Martin's finite-element method to handle any variations created by the quadratic interpolation functions.

Third Phase. The third phase was the actual study of the error and stability in the quadratic finite-element formulation. Theory and procedures for this study are noted in the appropriate locations of Section II and Section IV.

One factor that hindered a good error analysis, most obvious by noting Figure 3, was the discontinuity between the initial condition and boundary conditions. Several methods of handling this problem are referred to by Martin (Ref 7:72). Because the elimination of this discontinuity was not the primary goal of this thesis, the problem was by-passed by substituting the exact analytical solution at the first time step, a procedure similar to that suggested by Smith (Ref 9:48-49). In effect, this transformed the original problem into a new problem in which no discontinuity existed between the initial condition and the boundary conditions.

Of course, the dominant factor in viewing the error and stability analysis was the noted equivalance of the chosen quadratic approach to the linear approach. This factor affected all the procedural phases.

Fourth Phase. Phase four was the actual comparison of results as mentioned previously. Basically, greater accuracy was originally expected by using a quadratic interpolation function. Results using this function and the linear function were compared for selected values of the Fourier modulus, p, since both the linear and quadratic optimum α values were functions of this parameter only. The same appropriate mesh spacings were also used.

Error and stability definitions are those mentioned in Section II. Also employed for graphical depiction and comparison of error was the discretization error ratio (DER), defined as the ratio of the discretization error incurred when one subdivision of the space domain is used, to the error at the same point when the number of nodes has been doubled (Ref 7:84-85). Discretization error is basically composed of truncation error and round-off error, the latter factor approaching dominance as Δx is made smaller.

Discretization error ratio can best be understood in the following sense. If the error for some norm is given by

$$e = \xi(\Delta x)^2 \tag{130}$$

then the result of decreasing the interval size by a factor of 1/2 will be

$$\frac{e_1}{e_{\frac{1}{2}}} = \frac{\xi_1(\Delta x)^2}{\xi_{\frac{1}{2}}\left(\frac{\Delta x}{2}\right)^2} \tilde{=} 4 \qquad (131)$$

Similarly, the effect of halving the interval size when the error is given by

$$e = \xi(\Delta x)^4 \qquad (132)$$

is

$$\frac{e_1}{e_{\frac{1}{2}}} = \frac{\xi_1(\Delta x)^4}{\xi_{\frac{1}{2}}\left(\frac{\Delta x}{2}\right)^4} \tilde{=} 16 \qquad (133)$$

Thus, a discretization error ratio of 4 indicates $0(\Delta x)^2$ accuracy; a DER of 16 indicates $0(\Delta x)^4$ accuracy.

Selected for comparison were the error and stability results attained in the computations leading to the presentations of DER versus TIME for Fourier modulus values of .5 , 1.0 , and 2.0 , and α values of the Crank-Nicolson, optimum implicit, and fully implicit schemes.

Computer Application

Computer. The computer system used for this project was designed by Control Data Corporation, CDC. It consists of input and output devices, peripheral processors, and two central processors which operate in parallel, the CDC 6613 and CDC CYBER 74. Each has

131,000 60-bit words of central memory. Magnetic disc and drum storage were used as temporary storage devices.

Computer Programs. The major computer program employed was a modification of Martin's program for solution by finite-elements. Actually, two major modifications were required. The first was to allow solution of the linear interpolation problem by a general library subroutine matrix solver. This problem, which involved symmetric tridiagonal system matrices, was originally solved by the Thomas method. Since the quadratic interpolation problem was not tridiagonal, this modification was required to maintain continuity in the error and stability analyses. The library subroutine, LEQT1B, factors the system matrices ($\underline{A}$) into the $\underline{L}$-$\underline{U}$ decomposition of a rowwise permutation of $\underline{A}$, and then solves the system.

The second modification was to rewrite the program to formulate the system matrices representative of the quadratic interpolation function. A new subroutine was generated to create these pentagonal matrices. The same matrix solver mentioned above was also used here.

Because the point of interest of this investigation was the behavior and accuracy of the finite-element solution as a function of the parameter α, no effort was made to compare costs and run times for the two interpolation approaches or the various schemes within each. Figure 8 is a flow chart of the finite-element approach used. Based on previous knowledge and as implied above, only option (OPT) one was used. Several other programs were also used for plotting purposes and to assist in the stability analysis.

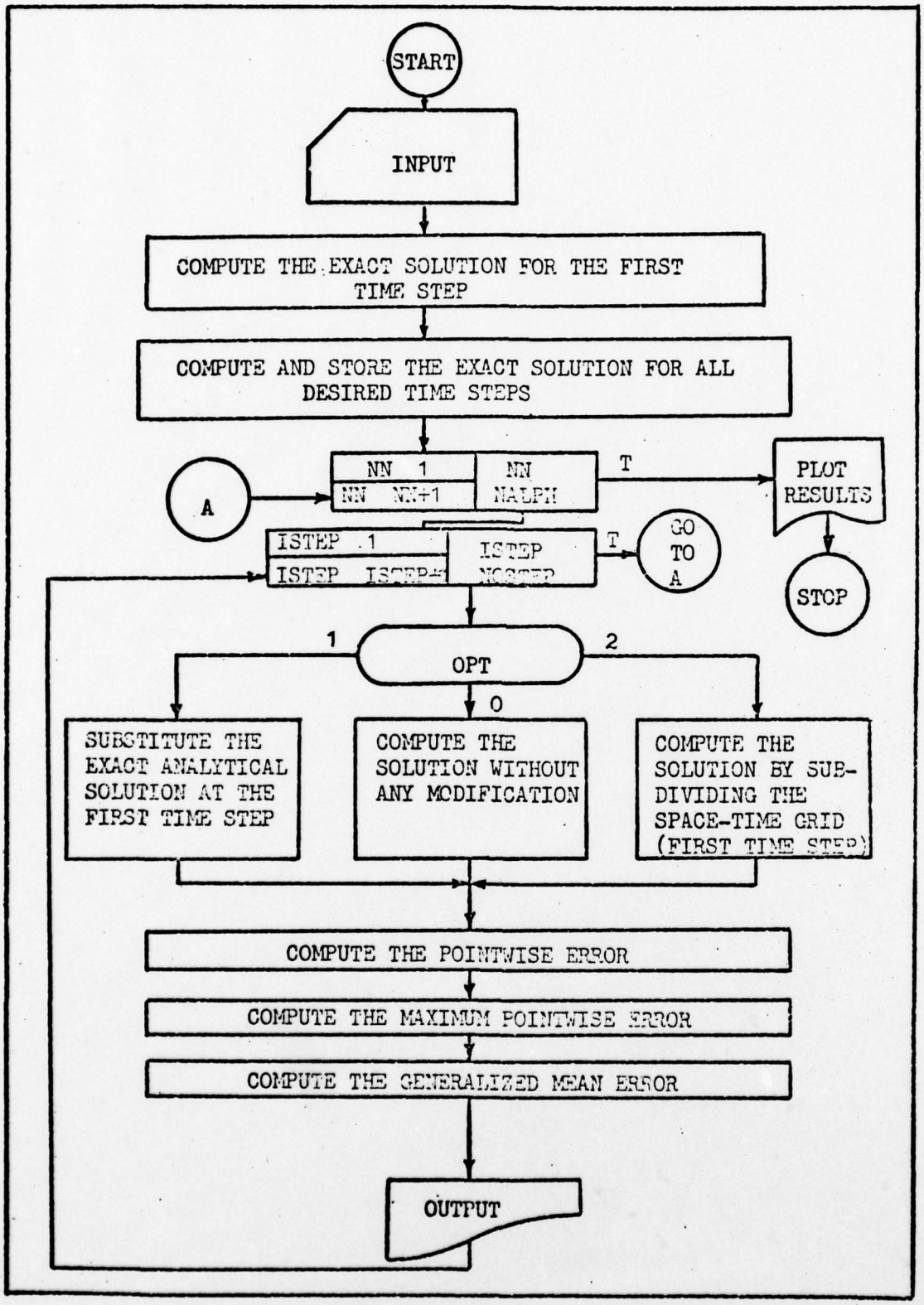

Figure 8. Finite Element Computer Program Flow Chart

IV. Results

Stability Analysis

The results of the stability analysis are shown in Tables II and III and Figures 6 and 7. These are the same results found by Martin, since the failure of direct quadratic application to yield at least fourth order accuracy required the use of a linear interpolation based on equivalence of alpha values. Basically, it should be noted that the stability curves of the finite-difference and finite-element optimum implicit schemes coincide; and, that while for finite-differences the optimum implicit scheme is less stable than the Crank-Nicolson scheme, in finite-elements this situation is reversed. As before, the finite-element method is more restrictive with respect to stability in the general sense.

Error Analysis

The theory leading to this error analysis is presented in that chapter. Basically, this analysis is just a comparison of accuracy as specified in the thesis objectives. Plots verifying these results, here presented in Tables V through VII, as well as plots displaying the accuracy found by direct application of the quadratic formulation, are found in Appendix H. It should be noted that the tabular values specified as optimum quadratic and identified by a "Q" are those found by applying the equivalent linear optimum alpha values of quadratic interpolation as found with Equation (110). It should also be noted that Table V is based on a 20 interval space domain, while Tables VI and VII are based on 40 interval space domains.

For all tables, the following abbreviations apply:

CN = Crank-Nicolson Method

OI = Optimum Implicit Method

FI = Fully Implicit Method

FD = Finite Differences

FE = Finite Elements

L = Linear

Q = Quadratic

TABLE V

Error Comparisons for the Various Methods for $\theta = .08$ and $p = 1.0$

Method	Pointwise Error $x = .1$	Maximum Error at Any Node	Generalized Mean Error
CN-FD	-3.0351×10^{-4}	8.5043×10^{-4}	5.6708×10^{-3}
CN-FE-L	3.2920×10^{-4}	8.8078×10^{-4}	5.9831×10^{-3}
CN-FE-Q	3.2920×10^{-4}	8.8078×10^{-4}	5.9831×10^{-3}
OI-FD	1.4103×10^{-5}	2.1640×10^{-5}	1.5280×10^{-4}
OI-FE-L	1.4103×10^{-5}	2.1640×10^{-5}	1.5280×10^{-4}
OI-FE-Q	1.4103×10^{-5}	2.1640×10^{-5}	1.5280×10^{-4}
FI-FD	-2.2702×10^{-3}	5.8835×10^{-3}	4.0493×10^{-2}
FI-FE-L	-1.6021×10^{-3}	4.2338×10^{-3}	2.8906×10^{-2}
FI-FE-Q	-1.6021×10^{-3}	4.2338×10^{-3}	2.8906×10^{-2}

TABLE VI

Error Comparisons for the Various

Methods for $\theta = .04$ and $p = .5$

Method	Pointwise Error $x = .1$	Maximum Error at Any Node	Generalized Mean Error
CN-FE-L	2.1667×10^{-3}	2.9165×10^{-4}	1.5362×10^{-3}
CN-FE-Q	2.1667×10^{-3}	2.9165×10^{-4}	1.5362×10^{-3}
OI-FE-L	1.89648×10^{-5}	2.32381×10^{-6}	1.39456×10^{-5}
OI-FE-Q	1.89605×10^{-5}	2.32323×10^{-6}	1.39429×10^{-5}
FI-FE-L	-4.3141×10^{-4}	5.8037×10^{-4}	3.0579×10^{-3}
FI-FE-Q	-4.3141×10^{-4}	5.8037×10^{-4}	3.0579×10^{-3}

TABLE VII

Error Comparisons for the Various

Methods for $\theta = .16$ and $p = 2.0$

Method	Pointwise Error $x = .1$	Maximum Error at Any Node	Generalized Mean Error
CN-FE-L	6.6945×10^{-4}	2.1656×10^{-4}	1.3675×10^{-3}
CN-FE-Q	6.6945×10^{-4}	2.1656×10^{-4}	1.3675×10^{-3}
OI-FE-L	1.59502×10^{-5}	5.14685×10^{-6}	3.25294×10^{-5}
OI-FE-Q	1.58979×10^{-5}	5.12994×10^{-6}	3.24226×10^{-5}
FI-FE-L	-7.1636×10^{-3}	2.3172×10^{-3}	1.4632×10^{-2}
FI-FE-Q	-7.1636×10^{-3}	2.3172×10^{-3}	1.4632×10^{-2}

Table V shows that for the optimum implicit scheme the error values are identical. The fact that the linear finite-element errors equal the quadratic finite-element errors is understandable since both approaches involved the same alpha values for the Fourier modulus of one. Tables VI and VII indicate a small increase in the accuracy of the optimum implicit schemes when the equivalent alpha to the quadratic approach is employed, as previously discussed. The Crank-Nicolson and fully implicit schemes show no changes when this technique is used. The overall result is as predicted by Equation (112); that is, that fourth order accuracy is only possible for this quadratic approach.

With the slightly greater accuracy just mentioned, the results may be stated in the following manner. The optimum implicit methods for both finite-differences and finite-elements, for $p = 1.0$, yield the same accuracy as found by Martin. Assuming this argument to be true for all values of the Fourier modulus in the linear domain, then the quadratic finite-element solution employing the equivalent linear alpha value (OI-FE-Q) is more accurate than the finite-difference solution for values of p other than 1.0, or, rather, for solutions for which the quadratic optimum alpha does not equal the linear optimum alpha. The finite-difference Crank-Nicolson scheme is more accurate than its finite-element version, but for the fully implicit scheme, the opposite is true. The optimum implicit finite-element scheme is always more accurate than the Crank-Nicolson finite-element scheme.

Plots

General. The plots are presented in Appendix H in the order of sections discussed here, and display the discretization error ratio as a function of time. The first section presents those figures (H-1 through H-12) associated with the finite-difference method solution of the problem. Section two contains the figures (H-13 through H-24) associated with the linear finite-element solution. It will be noted that, if the quadratic solution is considered to be that of double the number of nodes, then each even numbered figure of this section can be considered to be the more accurate solution representation to its linear counterpart as given by the odd numbered figures.

Section three presents those figures (H-25 through H-36) which represent the finite-element solutions attained when $\Delta x/2$ is incorporated into the quadratic formulation after the system matrices are formulated. It will be remembered that this process was accomplished so that a quadratically determined equivalent alpha could be applied to the system without considering the internal node. The accuracy of these plots, entitled "Quadli," is noted to be slightly greater for appropriate values of the Fourier modulus.

For these first three sections, the figures for $p = .5$ are not shown for the sake of brevity and to avoid redundancy. However, actual DER values are displayed in Table VIII for selected times and serve to verify the theory leading to Tables V through VII and Figures (H-13) through (H-36).

TABLE VIII

Discretization Error Ratio Comparison

For the Optimum Implicit Scheme

Method	Modulus/Time	DER Maximum Error at Any Node	DER Generalized Mean Error
FE-L	1.0/.04	15.1029	15.2466
FE-Q	1.0/.04	15.1029	15.2466
FE-L	1.0/.08	15.4987	15.6123
FE-Q	1.0/.08	15.4987	15.6123
FE-L	.5/.02	14.5045	15.2072
FE-Q	.5/.02	14.5061	15.2088
FE-L	.5/.04	15.4714	15.6005
FE-Q	.5/.04	15.4743	15.6028
FE-L	2.0/.08	15.0418	15.1991
FE-Q	2.0/.08	15.0662	15.2327
FE-L	2.0/.16	15.6262	15.6250
FE-Q	2.0/.16	15.6648	15.6636

The fourth section presents those figures (H-37 through H-48) which display the less than second order accuracy attained for direct quadratic solution when Equation (109) is not accounted for, and the second order accuracy attained when it is taken into consideration. A Fourier modulus of one was chosen for this display.

Section I Results. This section gives the results of the problem as solved by the finite-difference method. Figures (H-1) through (H-12), as annotated for the cases of $p = 1.0$ and $p = 2.0$, show that for each error norm, the fully implicit and the Crank-Nicolson methods approach second order accuracy, while for the optimum implicit method, with the value of alpha as determined by Crandall (Ref 3:319), fourth order accuracy is approached. For $p = 1.0$, $\alpha_{opt} = .41667$; for $p = 2.0$, $\alpha_{opt} = .45833$.

Section II Results. This section gives the results of the problem as solved by the finite-element method, linear interpolation. Figures (H-13) through (H-24), annotated as above, show that for each error norm, the fully implicit and the Crank-Nicolson methods approach second order accuracy, while for the optimum implicit method, with the value of alpha as determined by Martin [Equation (G-5)], fourth order accuracy is approached. For $p = 1.0$, $\alpha_{opt} = .58333$; for $p = 2.0$, $\alpha_{opt} = .54167$.

Section III Results. This section gives the results of the problem as solved by the finite-element method, with optimum alphas determined by quadratic interpolation, Equation (110). Equivalent linear optimum alphas were used and applied to the theory leading to the Section II results to produce Figures (H-25) through (H-36), again annotated as above. Results are similar to those of Section II, but, except for $p = 1.0$, a noted increase in accuracy is observed, especially if viewed in relation to Table VIII. By Equation (111), for $p = 1.0$, α_{opt} is annotated .58333. Similarly for $p = 2.0$, α_{opt} is annotated .6250.

Section IV Results. The results of this section are divided into two parts and are presented to verify the information presented by Equations (108) and (109). Figures (H-37) through (H-42) show a low order of accuracy for direct application of the quadratic formulation without considering the requirements of Equation (109). For this case of $p = 1.C$, $\alpha_{opt} = .58333$ as derived by Equation (108). It should be noted as well that the solutions by the linear Crank-Nicolson alpha value and fully implicit alpha value yield approximately the same accuracy. The fact that the accuracy observed for $\alpha = .58333$ is not the greatest of the three schemes is insignificant since no optimum scheme is really being applied.

Figures (H-43) through (H-48) show the second order accuracy attained if Equation (109) is accounted for by inserting $\alpha = .66666$ at the internal nodes. The accuracy for the optimum scheme of $\alpha_{opt} = .58333$ is noted to be greater than the other schemes, but again, only second order accurate.

Summary. In all the plots, and data from which they were constructed, it should be remembered that the exact analytical solution was used at the first time step to eliminate the problem of discontinuity between the initial and boundary conditions. This was determined by Martin to be the best procedure for handling this situation. Also, in all the plots, except Figures (H-37) through (H-42), the greatest accuracy was observed for the derived optimum alpha values. The results of Figures (H-43) through (H-48), supplemented by the findings of Appendix I, indicate that the internal nodes must be specially handled if fourth order accuracy is to be achieved.

This point should be emphasized in light of the fact that, even though fourth order accuracy is noteworthy, it is no better than that attained with the direct linear interpolation approach, which was also fourth order accurate. It would be logical and less expensive, therefore, to apply the linear finite-element solution technique.

Briefly, it was found in theory (Appendix E), and as computationally displayed by the above Section IV referenced figures, that despite the more rigorous solution development inherent in the quadratic interpolation approach, problems introduced by the existence of the internal nodes prevented direct application of the system matrices. Indeed, Equation (109) indicated an accuracy less than fourth order at the internal nodes, a consideration of which allowed for general second order accuracy.

In this limit of second order accuracy (Table G-I), a direct relationship was found between the quadratic and linear optimum alpha values yielding for a particular selection of the parameter p, a value of alpha which gave the same degree of implicitness for the quadratic approach as for the linear approach. The application of this fact to the quadratic formulation, therefore, allowed for the same or slightly greater accuracy over the linear interpolation formulation at the external nodes, or at every other node, as displayed by the Section III referenced figures.

V. Conclusions and Recommendations

Conclusions

Stability Analysis. The equivalent orders of accuracy for the linear and quadratic finite-element formulations, coupled with the necessity to handle the quadratic interpolation with an equivalent linear system to achieve such accuracy, indicates that the quadratic system may be handled as a linear system of twice the number of nodes. As such, the general stability results are as before. The finite-difference method is more stable than the finite-element method.

Error Analysis. This error analysis is based on results achieved when the discontinuity between the initial and boundary conditions has been eliminated by substitution of the exact analytical solution at the first time step. The overall result of this solution procedure is that for the optimum implicit formulation, fourth order accuracy is attainable, while for the other schemes studied, only second order accuracy is possible. For the case where the discontinuity has not been accounted for, the results are as stated by Martin (Ref 7:92-95).

Specifically, for the optimum implicit scheme, the linear finite-element errors equal the finite-difference errors for all values of the Fourier modulus. The quadratic finite-element errors equal these values for $p = 1.0$, but for the other values of the modulus, an increase in accuracy is noted. The fact that the linear finite-element errors equal the quadratic finite-element errors for

$p = 1.0$ is understandable since, for this case, both approaches involved the same optimum alpha value. The errors for solution by the Crank-Nicolson scheme, linear finite-elements, equal the errors for such solution with quadratic finite-elements. The same is true for the fully implicit scheme.

Also, for the fully implicit scheme and available data, solution by the finite-element method is more accurate than solution by the finite-difference method. However, for the Crank-Nicolson scheme, the opposite is found to be true. The optimum implicit finite-element scheme is always more accurate than the Crank-Nicolson finite-element scheme.

In considering the accuracy comparison just presented, it should be kept in mind that the overall order of accuracy for both linear and quadratic finite-element interpolation procedures was found to be equal. This indicated a possible equivalence of both methods, or a limitation on the achievement of greater accuracy when employing this general solution approach. Indeed, the truncation error analysis predicted the minimum acceptable fourth order accuracy only if the internal nodes were accounted for without direct application as such in the quadratic system. This treatment of the internal nodes established the basis for the equivalence of the linear and quadratic systems.

One possible method of better handling the internal nodes and eliminating the three point rows of the system matrices $\underline{A}$ and $\underline{B}$ would be to use the method of splines where every node, or in spline theory, knot, would be connected to another knot by an

appropriate function, usually as a minimum, a bi-cubic spline, and no distinction made between internal and external points. Such a procedure would establish new system matrices, less sparse, and containing rows of an equal number of elements.

However, in general, a truncation error analysis using the elements of these new matrices might not show order of accuracy improvement, because even though spline theory was used to eliminate the difficulties in the space dimension introduced by employment of the quadratic interpolation function, it still stands that the time derivatives are approximated by a finite-difference expression for which no parallel increase in accuracy can be attained by going to higher order polynomial approximating temperature distributions. This fact, along with the beliefs of Strang and Fix (Ref 10:244-245) that the handling of time by methods other than the Galerkin or variational approach may couple all time levels and destroy the property of propagation forward in time, tends to indicate that greater accuracy may not be attainable for this problem approach. Certainly the equivalency of the linear and quadratic finite-element optimum alpha formulations, in light of Martin's observation that for this scheme, the linear finite-element method and the finite-difference method are, in fact, the same (Ref 7:92), adds verification to these statements.

Two approaches worth considering as feasible alternatives for handling of the temporal response would be to use spline theory in the time domain, and the use of a continuous time model. In the first approach, the theoretical procedures leading to the expression

of the unknown solution are the same. For space and time, Strang and Fix (Ref 10:243) give this expression as

$$u(x,t) = \sum_{1}^{N} Q_j(t)\phi_j(x) \tag{134}$$

which should be compared with Equation (I-1). If $Q_j(t)$ were now represented by a bi-cubic spline, and $\phi_j(x)$ represented by either such a spline or finite-elements, the formulation, although somewhat unconventional, is free of the difficulties before mentioned, and promises to yield accuracy of a higher order.

The second approach is the continuous time model. In this model, Equation (96) is written as

$$\frac{d}{d\theta}\left(\underline{u}^{(E)}\right) = -\underline{M}^{-1}\underline{K}\underline{u}^{(E)} \tag{135}$$

and solved directly. The space dimension is still discretized and is represented by the elements of the column matrix, $\underline{u}^{(E)}$. $\underline{M}$ and $\underline{K}$ are as before. It should be immediately obvious that any problems with the discontinuity between the initial and boundary conditions are no longer present. The initial or normalized initial condition need only be placed in the right side of Equation (135) and the calculation allowed to proceed. In fact, a hand calculation for the minimum case of the interval kept as one element with two external nodes, yields at the centered internal node and a time .01 seconds later, a temperature value of .9991112 . Comparing this value to

the exact analytical solution for the case of 10 elements, the error is found to be 7.48×10^{-5}, which is more accurate by an order of magnitude than the maximum error at any node, any time, any scheme.

In summary, the results achieved and presented in graphical and tabular form are consistent with the theory derived and the overall structure of finite-elements.

Recommendations

The limitation on accuracy presented in this thesis was postulated as inherent in the problem approach, or more specifically, in the handling of the time domain by finite-differences after establishment of the recurrence relation. Any notable increase in accuracy requires elimination of the difficulty. The use of a continuous time model, independent of the requirements to employ the alpha parameter, appears to be the easiest immediate solution. Spline theory could be used to eliminate the problems associated with working with internal nodes. Of course, splines could also be used in time as previously mentioned, an approach which would be more lengthy and difficult since new and less sparse matrices would be created, but nonetheless, an approach that promises to yield high order of accuracy without having to eliminate the concept of time steps inherent in finite-element theory.

The employment of a quadratic finite-element interpolation function, with its additional internal nodal variables and correspondingly more complex system matrices, yields accuracy equal to that of the less exhaustive linear finite-element formulation

for this general solution approach to the transient heat conduction equation. Its use, therefore, is not recommended. Also, with such suggestions considered, the use of finite-element interpolation functions of an order greater than linear is not recommended without first appropriately treating the time response.

Bibliography

1. Churchill, Ruel V. Fourier Series and Boundary Value Problems. New York: McGraw-Hill Book Co., 1969.

2. Clark, Melville, Jr. and Kent F. Hansen. Numerical Methods of Reactor Analysis. New York: Academic Press, 1964.

3. Crandall, Stephen H. "An Optimum Implicit Recurrence Formula for the Heat Conduction Equation." Quarterly of Applied Mathematics, 13: 318-320 (1955).

4. Felippa, Carlos A. and Ray W. Clough. "The Finite Element Method In Solid Mechanics." Numerical Solution of Field Problems In Continuum Physics. SIAM-AMS Proceedings, II: 210-252 (1970).

5. Huebner, Kenneth H. The Finite Element Method for Engineers. New York: John Wiley & Sons, 1975.

6. Kohler, W. and J. Pittr. "Calculation of Transient Temperature Fields With Finite Elements In Space and Time Dimensions." International Journal For Numerical Methods In Engineering, 8: 625-631 (1974).

7. Martin, Charles R. An Investigation of the Numerical Methods of Finite Differences and Finite Elements For Digital Computer Solution of the Transient Heat Conduction (Diffusion) Equation Using Optimum Implicit Formulations. Unpublished thesis, Wright-Patterson Air Force Base, Ohio: Air Force Institute of Technology, March 1978.

8. Myers, Glen E. Analytical Methods in Conduction Heat Transfer. New York: McGraw-Hill Book Co., 1971.

9. Smith, G. D. Numerical Solution of Partial Differential Equations. London: Oxford University Press, 1965.

10. Strang, Gilbert and George J. Fix. An Analysis of the Finite Element Method. Englewood Cliffs, NJ, 1973.

11. Thomas, George B. Calculus and Analytic Geometry. Reading, Massachusetts: Addison Wesley Publishing Co., Inc., 1962.

12. Varga, Richard S. Matrix Iterative Analysis. Englewood Cliffs, NJ: Prentice Hall, 1962.

13. Zienkiewicz, O. C. The Finite Element Method in Engineering Science. London: McGraw-Hill Book Co., 1971.

APPENDIX A

The Analytical Solution of the Primary Problem

The given physical problem in normalized form and appropriate conditions is

$$\frac{\partial u}{\partial \theta} = \frac{\partial^2 u}{\partial x^2} \tag{A-1}$$

$$u(x,\theta) = 1, \quad \theta = 0 \tag{A-2}$$

$$u(0,\theta) = u(1,\theta) = 0, \quad \theta > 0 \tag{A-3}$$

where $x = \overline{x}$.

By separation of variables (Ref 1:34), it is assumed that $u(x,\theta) = X(x)\,\textcircled{H}\,(\theta)$. Taking the appropriate partial derivatives and substituting into the equations above yields

$$\frac{X''}{X} = \frac{\textcircled{H}'}{\textcircled{H}} = \gamma \tag{A-4}$$

and

$$X(0)\,\textcircled{H}\,(\theta) = 0, \quad \theta > 0 \tag{A-5}$$

and

$$X(1)\,\textcircled{H}\,(\theta) = 0, \quad \theta > 0 \tag{A-6}$$

where γ is a separation constant.

For any $\theta > 0$,

$$X(0) - X(1) = 0 \tag{A-7}$$

is the boundary condition and the Sturm-Liouville problem becomes

$$X'' - \gamma X = 0 \tag{A-8}$$

Only for the case $\gamma < 0$ does a solution exist. If γ is assumed equal to $-\alpha^2$ where $\alpha \neq 0$, the solution of (A-8) is

$$X(x) = A \cos(\alpha x) + B \sin(\alpha x) \tag{A-9}$$

Application of (A-7) at 0 yields

$$A = 0 \tag{A-10}$$

Application of (A-7) at 1 yields

$$\sin(\alpha) = 0 \tag{A-11}$$

That is, for a non-trivial solution, B cannot equal zero.
Equation (A-11) is satisfied by an infinity of α values; however, by the previous restrictions

$$\alpha_n = n\pi \quad , \quad n = 1,2,3,... \tag{A-12}$$

Therefore, for (A-8), the eigenvalues are

$$\gamma_n = -(\alpha_n)^2 = -(n\pi)^2 \quad , \quad n = 1,2,3,... \tag{A-13}$$

and the solutions are

$$X_n(x) = B_n \sin(n\pi x) \quad , \quad n = 1,2,3,... \tag{A-14}$$

The remaining problem

$$\textcircled{H}' - \gamma \textcircled{H} = 0 \tag{A-15}$$

is solved by integration to yield

$$\textcircled{H}(\theta) = C_n \exp[-(n\pi)^2\theta] \tag{A-16}$$

By superposition of Equations (A-14) and (A-16), the general solution is

$$u(x,\theta) = \sum_{n=1}^{\infty} (B_n \cdot C_n) \sin(n\pi x) \exp[-(n\pi)^2\theta] \tag{A-17}$$

The initial condition is now applied in an infinite Fourier series of Equation (A-14) to yield

$$u(x,0) = \sum_{n=1}^{\infty} A_n \sin(n\pi x) \cdot 1 = 1 \qquad \text{(A-18)}$$

Here, theta equals zero and the constants are combined. Equation (A-18) is a Fourier sine series where it can be shown that (Ref 7:103)

$$A_n = \frac{2}{n\pi} [1 - (-1)^n] \qquad \text{(A-19)}$$

For $n = 2m - 1$, the solution, (A-17), becomes

$$u(x,\theta) = \sum_{m=1}^{\infty} \frac{4}{(2m-1)\pi} \sin [(2m-1)\pi x] \exp \{-[(2m-1)\pi]^2 \theta\} \qquad \text{(A-20)}$$

APPENDIX B

Elementary Variational Calculus Review

A simple problem of variational calculus is to find a function $u(x)$ that minimizes the integral

$$I = \int_{x=0}^{L} F(x,u(x),u'(x))\,dx \tag{B-1}$$

with the boundary conditions

$$u(0) = u_o \tag{B-2}$$

and

$$u(L) = u_L \tag{B-3}$$

and where u' denotes differentiation with respect to x (Ref 8:322-325).

$u(x)$ is found by considering every possible continuous function that satisfies the boundary conditions and selecting the one that minimizes I. If $\varepsilon\eta(x)$ is called the variation (see Figure B-1), being zero for the case when I is minimum, then the set of possible functions is represented by $\tilde{u}(x,\varepsilon)$ where

$$\tilde{u}(x,\varepsilon) = u(x) + \varepsilon\eta(x) \tag{B-4}$$

The function $\eta(x)$ is restricted to be exact at the boundaries; that is,

$$\eta(o) = \eta(L) = 0 \tag{B-5}$$

This restriction insures that

$$\tilde{u}(o,\varepsilon) = u(o) \quad \text{and} \quad \tilde{u}(L,\varepsilon) = u(L) \tag{B-6}$$

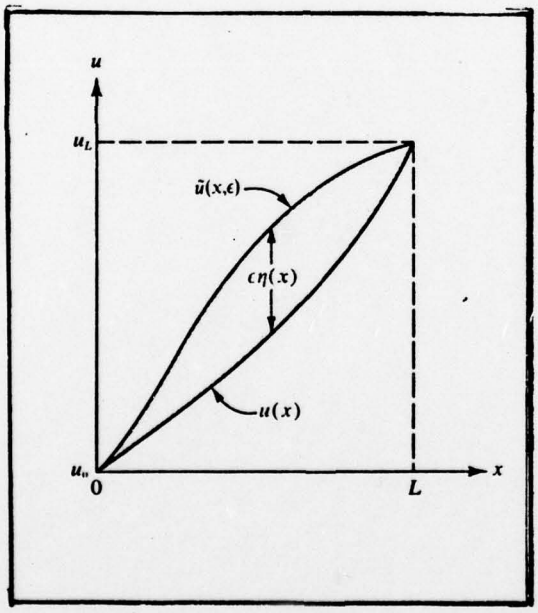

Figure B-1. Desired Variational Solution u(x) and Trial Function $\tilde{u}(x,\varepsilon)$.

As mentioned, the function sought is attained from the function set (Equation (B-4)) for the case of zero variation; or, that is, $\varepsilon = 0$. The desired value of $u(x)$ is such that

$$\tilde{u}(x,0) = u(x) \tag{B-7}$$

Also, the corresponding derivative term of Equation (B-1) is noted as

$$\tilde{u}'(x,\varepsilon) = u'(x) + \varepsilon\eta'(x) \tag{B-8}$$

Next, Equations (B-4) and (B-8) are substituted into Equation (B-1) to yield

$$I(\varepsilon) = \int_{x=0}^{L} F(x,\tilde{u}(x,\varepsilon),\tilde{u}'(x,\varepsilon))\, dx \tag{B-9}$$

where I is a function of ε because it is still a parameter after the x integration. By Equation (B-7), Equation (B-9) reduces to Equation (B-1) when ε equals zero. That is, $I(\varepsilon)$ is a minimum when ε equals zero.

The process of minimization is pursued by employing Leibnitz' rule and the chain rule in differentiating $I(\varepsilon)$ with respect to ε to yield

$$\frac{dI(\varepsilon)}{d\varepsilon} = \int_{x=0}^{L} \frac{\partial}{\partial\varepsilon} F(x,\tilde{u}(x,\varepsilon),\tilde{u}'(x,\varepsilon))\, dx \tag{B-10}$$

and

$$\frac{dI(\varepsilon)}{d\varepsilon} = \int_{x=0}^{L} \left[\frac{\partial F}{\partial \tilde{u}} \frac{\partial \tilde{u}}{\partial \varepsilon} + \frac{\partial F}{\partial \tilde{u}'} \frac{\partial \tilde{u}'}{\partial \varepsilon} \right] dx \tag{B-11}$$

By Equations (B-4) and (B-8), it is noted that

$$\frac{\partial \tilde{u}}{\partial \varepsilon} = \eta(x) \qquad \text{and} \qquad \frac{\partial \tilde{u}'}{\partial \varepsilon} = \eta'(x) = \frac{\partial \eta(x)}{\partial x} \tag{B-12}$$

which, when substituted into Equation (B-11), yields

$$\frac{dI(\varepsilon)}{d\varepsilon} = \int_{x=0}^{L} \left[\frac{\partial F}{\partial \tilde{u}} \eta(x) + \frac{\partial F}{\partial \tilde{u}'} \frac{d\eta(x)}{dx} \right] dx \tag{B-13}$$

Integration by parts of the second term yields

$$\frac{dI(\varepsilon)}{d\varepsilon} = \int_{x=0}^{L} \frac{\partial F}{\partial \tilde{u}} \eta(x) dx + \left[\frac{\partial F}{\partial \tilde{u}'} \eta(x) \right]_{x=0}^{L} - \int_{x=0}^{L} \eta(x) \frac{d}{dx} \left[\frac{\partial F}{\partial \tilde{u}'} \right] dx \tag{B-14}$$

The integrated term vanishes by Equation (B-5). Thus, upon recombining the two integrals, Equation (B-14) becomes

$$\frac{dI(\varepsilon)}{d\varepsilon} = \int_{x=0}^{L} \eta(x) \left[\frac{\partial F}{\partial \tilde{u}} - \frac{d}{dx} \left(\frac{\partial F}{\partial \tilde{u}'} \right) \right] dx \tag{B-15}$$

Setting the derivative to zero for $\varepsilon = 0$ yields

$$\tilde{u} = u \tag{B-16}$$

AD-A063 962 AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OHIO SCH--ETC F/G 20/13
AN INVESTIGATION OF THE METHOD OF FINITE ELEMENTS WITH ACCURACY--ETC(U)
DEC 78 J C GENECZKO
UNCLASSIFIED AFIT/GNE/PH/78D-15 NL

2 OF 2
AD
A063982

END
DATE
FILMED
4 -79
DDC

and therefore

$$\left.\frac{dI}{d\varepsilon}\right|_{\varepsilon=0} = \int_{x=0}^{L} \eta(x) \left[\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u'}\right)\right] dx = 0 \qquad \text{(B-17)}$$

Because $\eta(x)$ is arbitrary, the term in brackets must be zero to ensure that the integral is zero. Therefore, for I to be a minimum

$$\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u'}\right) = 0 \qquad \text{(B-18)}$$

The solution $u(x)$ to this Euler-Lagrange differential equation is the function that minimizes the original integral.

APPENDIX C

Derivation of Quadratic Constants

The first step in the matrix formulation of the finite-element procedure is to find the constants, Equations (52), (53), and (54). The process is tedious and complicated. Several steps of the process are here presented to assist in understanding the mathematical sequence employed.

After establishing Equations (46), (47), and (48) as representing the nodal temperatures of an element, they are solved simultaneously to find $c_1^{(e)}$, $c_2^{(e)}$, and $c_3^{(e)}$. $c_3^{(e)}$ is first found by writing Equation (48) as

$$c_1^{(e)} + c_2^{(e)} \left(\frac{x_i+x_j}{2}\right) = u_k - c_3^{(e)} \left(\frac{x_i+x_j}{2}\right)^2 \qquad \text{(C-1)}$$

and adding Equations (46) and (47) to yield

$$\frac{u_i+u_j}{2} = c_1^{(e)} + c_2^{(e)} \left(\frac{x_i+x_j}{2}\right) + c_3^{(e)} \frac{(x_i^2+x_j^2)}{2} \qquad \text{(C-2)}$$

Substituting Equation (C-2) into (C-1) for $c_1^{(e)} + c_2^{(e)}\left(\frac{x_i+x_j}{2}\right)$ and expanding yields

$$\frac{u_i+u_j-2u_k}{2} = \frac{c_3^{(e)}}{2}\left[(x_i^2+x_j^2) - \frac{(x_i^2+2x_ix_j+x_j^2)}{2}\right] \qquad \text{(C-3)}$$

which is condensed and solved for $c_3^{(e)}$ as Equation (54).

Next, $c_2^{(e)}$ is found by subtracting Equation (46) minus Equation (47) to yield

$$u_i - u_j = c_2^{(e)} (x_i - x_j) + c_3^{(e)} (x_i^2 - x_j^2) \qquad \text{(C-4)}$$

Substituting Equation (54) into (C-4) for $c_3^{(e)}$ gives

$$u_i - u_j = c_2^{(e)} (x_i - x_j) + 2\frac{(u_i - 2u_k + u_j)}{(x_i - x_j)^2} (x_i^2 - x_j^2) \qquad \text{(C-5)}$$

which, when expanded and solved for $c_2^{(e)}$, yields Equation (53).

Finally, with $c_2^{(e)}$ and $c_3^{(e)}$ known, $c_1^{(e)}$ is found by writing Equation (46) for $c_1^{(e)}$ as

$$c_1^{(e)} = u_i - c_2^{(e)} x_i - c_3^{(e)} x_i^2 \qquad \text{(C-6)}$$

and substituting Equations (53) and (54) for $c_2^{(e)}$ and $c_3^{(e)}$, respectively, to yield Equation (52).

APPENDIX D

Assembly Theory for Matrices

Equation (106) was assembled as representative of the problem for the case of three elements and seven total nodes, four external. Certain specific theory applies to the assembly of these matrices $\underline{A}$ and $\underline{B}$, presented here in summary and treated exactly by Huebner (Ref 5:43-50). Figure D-1 depicts the situation of this case, with

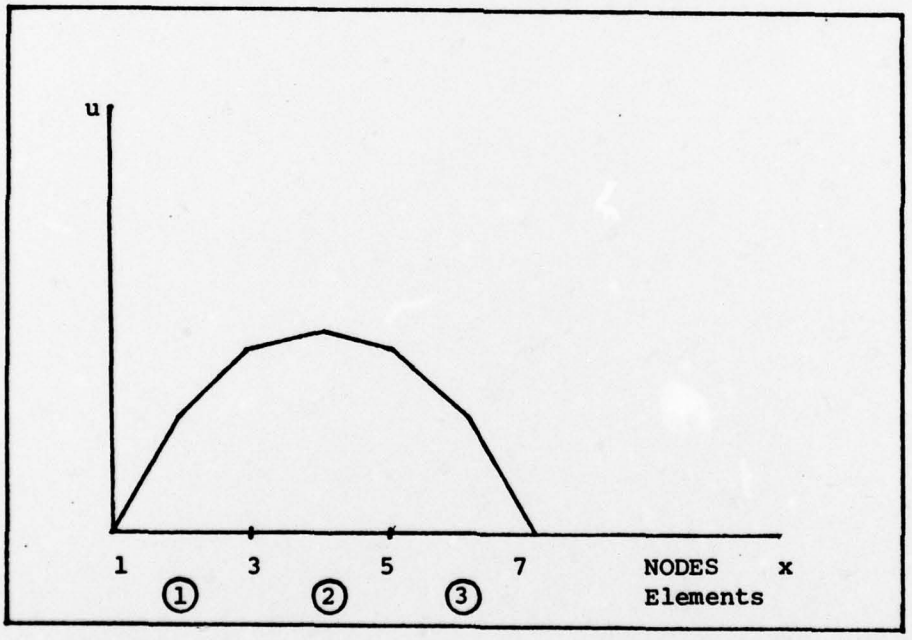

Figure D-1. Example of One Dimensional Problem of Three Elements and Four External Nodes.

elements ①, ②, and ③, and external nodes 1, 3, 5, and 7. It should be noted that the following procedures, although based on this simple case, are in principle the general procedures that apply to all

finite-element systems.

The first step is to specify a numbering scheme for the system shown in the figure. This scheme which relates the local position of the nodes to the system or global position is illustrated in Table D-I.

TABLE D-I

System Numbering - The Correspondence Between Local and Global Numbering Schemes

Element	Scheme	
	Local	Global
1	1	1
	2	2
	3	3
2	1	3
	2	4
	3	5
3	1	5
	2	6
	3	7

The next step is to place the element submatrix for element one into its assembled position. For this element, the local and global numbering schemes are, by coincidence, the same. That is, for element one (matrix $\underline{A}$)

$$[\underline{A}]^{(1)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & 0 & 0 & 0 & 0 \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(1)} & 0 & 0 & 0 & 0 \\ a_{31}^{(1)} & a_{32}^{(1)} & a_{33}^{(1)} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad \text{(D-1)}$$

where the superscripts indicate the element number. The second step is then continued for elements two and three by relating the local and global elements as depicted in Table D-II and writing the respective matrices, Equations (D-2) and (D-3) as

$$[\underline{A}]^{(2)} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & & \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & a_{35}^{(2)} & 0 & 0 \\ 0 & 0 & a_{43}^{(2)} & a_{44}^{(2)} & a_{45}^{(2)} & 0 & 0 \\ 0 & 0 & a_{53}^{(2)} & a_{54}^{(2)} & a_{55}^{(2)} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad \text{(D-2)}$$

$$[\underline{A}]^{(3)} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & a_{55}^{(3)} & a_{56}^{(3)} & a_{57}^{(3)} \\ 0 & 0 & 0 & 0 & a_{65}^{(3)} & a_{66}^{(3)} & a_{67}^{(3)} \\ 0 & 0 & 0 & 0 & a_{75}^{(3)} & a_{76}^{(3)} & a_{77}^{(3)} \end{bmatrix} \qquad \text{(D-3)}$$

TABLE D-II

Relationship Between Local and Global Elements of Matrix $\underline{A}$

Local Position	Global Position	
	Element Two	Element Three
a_{11}	a_{33}	a_{55}
a_{12}	a_{34}	a_{56}
a_{13}	a_{35}	a_{57}
a_{21}	a_{43}	a_{65}
a_{22}	a_{44}	a_{66}
a_{23}	a_{45}	a_{67}
a_{31}	a_{53}	a_{75}
a_{32}	a_{54}	a_{76}
a_{33}	a_{55}	a_{77}

Finally, $\underline{A}$ is obtained (as for Equation (106)) by adding Equations (D-1) through (D-3), representing contributions from each element. The mathematical statement of this is

$$[\underline{A}] = \sum_{e=1}^{M} [\underline{A}]^{(e)} = [\underline{A}]^{(1)} + [\underline{A}]^{(2)} + [\underline{A}]^{(3)} + \ldots \qquad \text{(D-4)}$$

where M is the total number of elements. The assembled $[\underline{A}]$ is then

$$[\underline{A}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 0 & 0 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 & 0 & 0 & 0 \\ a_{31} & a_{32} & (a_{33}+a_{11}) & a_{12} & a_{13} & 0 & 0 \\ 0 & 0 & a_{21} & a_{22} & a_{23} & 0 & 0 \\ 0 & 0 & a_{31} & a_{32} & (a_{33}+a_{11}) & a_{12} & a_{13} \\ 0 & 0 & 0 & 0 & a_{21} & a_{22} & a_{23} \\ 0 & 0 & 0 & 0 & a_{31} & a_{32} & a_{33} \end{bmatrix} \qquad \text{(D-5)}$$

The matrix $[\underline{B}]$ is similarly attained.

It should be noted that the internal nodes are not added in the assembly process. This is inherent in the procedure since the relationship between external nodes is approximated. Only the external nodes need be added to create the assembled matrix.

APPENDIX E

Derivation of the Truncation Error for the Finite Element Formulation

To derive the error, the system of equations (107), written for the minimum case of two elements, is treated as a set of difference equations (Ref 7:109-117). The general five point expression is given by

$$A_1 u_{i-1,k+1} + A_2 \left(\frac{u_{i-1,k+1} + u_{i,k+1}}{2} \right) + A_3 u_{i,k+1} + A_2 \left(\frac{u_{i+1,k+1} + u_{i,k+1}}{2} \right) + A_1 u_{i+1,k+1}$$

$$= B_1 u_{i-1,k} + B_2 \left(\frac{u_{i-1,k} + u_{i,k}}{2} \right) + B_3 u_{i,k} + B_2 \left(\frac{u_{i+1,k} + u_{i,k}}{2} \right) + B_1 u_{i+1,k} \qquad \text{(E-1)}$$

where, from Equation (106)

$$A_1 = -\tfrac{1}{2} + 5\alpha p$$

$$A_2 = 1 - 40\alpha p$$

$$A_3 = 4 + 70\alpha p$$

$$B_1 = -\tfrac{1}{2} - 5(1-\alpha)p$$

$$B_2 = 1 + 40(1-\alpha)p$$

and

$$B_3 = 4 - 70(1-\alpha)p$$

The subscripts i and k are used to position the nodal variables at the ten points represented by the terms in Equation (E-1). The procedure of this appendix is to write the equations of these nodal variables in a Taylor series centered about $u_{i,k}$, and then to find the truncation error as the difference between the exact partial differential equation and the difference equation (E-1).

This procedure, outlined by Crandall (Ref 3:319), may be stated as

$$e_t = \left\{ \frac{1}{k} \left[A_1 u_{i-1,k+1} + A_2 \left(\frac{u_{i-1,k+1} + u_{i,k+1}}{2} \right) + A_3 u_{i,k+1} \right.\right.$$

$$\left. + A_2 \left(\frac{u_{i+1,k+1} + u_{i,k+1}}{2} \right) + A_1 u_{i+1,k+1} \right]$$

$$- \frac{1}{k} \left[B_1 u_{i-1,k} + B_2 \left(\frac{u_{i-1,k} + u_{i,k}}{2} \right) + B_3 u_{i,k} \right.$$

$$\left. + B_2 \left(\frac{u_{i+1,k} + u_{i,k}}{2} \right) + B_1 u_{i+1,k} \right]$$

$$\left. - \frac{1}{k} \left(\frac{\partial u}{\partial \theta} - \frac{\partial^2 u}{\partial x^2} \right) \right\} \qquad \text{(E-2)}$$

where

$$u_{i,k+1} = u_{i,k} + k u_\theta + \frac{k^2}{2} u_{2\theta} + \frac{k^3}{3!} u_{3\theta} + \cdots \qquad \text{(E-3)}$$

$$u_{i\pm1,k} = u_{i,k} \pm hu_x + \frac{h^2}{2} u_{2x} \pm \frac{h}{3!} u_{3x} + \ldots \qquad \text{(E-4)}$$

and

$$u_{i\pm1,k+1} = u_{i,k} \pm hu_x + h^2(p+\tfrac{1}{2})\, u_{2x} + h^3(p+\tfrac{1}{6})\, u_{3x}$$

$$+ h^4\left(\frac{p^2}{2} + \frac{p}{2} + \frac{1}{24}\right) u_{4x} \pm h^5\left(\frac{p^2}{2} + \frac{p}{6} + \frac{1}{120}\right) u_{5x} + \ldots \qquad \text{(E-5)}$$

The subscripts here indicate differentiation with respect to the specified independent variable for the indicated number of times at the point $(i\Delta x, k\Delta\theta)$. The terms k and h represent $\Delta\theta$ and Δx , respectively. It is also noted that the last term of Equation (E-2) equals zero and can be dropped.

By substituting Equations (E-3) through (E-5) into Equation (E-2) and employing the relations

$$k = p(h)^2 \quad \text{or} \quad \Delta\theta = p(\Delta x)^2 \qquad \text{(E-6)}$$

$$u_\theta = u_{2x} \qquad \text{(E-7)}$$

$$u_{2\theta} = u_{\theta 2x} = u_{4x} \qquad \text{(E-8)}$$

and

$$u_{3\theta} = u_{2\theta 2x} = u_{\theta 4x} = u_{6x} \qquad \text{(E-9)}$$

each power of h may be collected together in the truncation error. The coefficients of all the odd powers of h cancel to zero. The coefficients of the even powers of h cancel to zero up to and including order two. The resulting expression is

$$e_t = h^2 \left(10p - 15\alpha p - \frac{15}{12}\right) u_{4x} + 0(h^4) \qquad \text{(E-10)}$$

or, the truncation error is of order h^2 unless

$$\left(10p - 15\alpha p - \frac{15}{12}\right) = 0 \qquad \text{(E-11)}$$

or

$$\alpha = \frac{1}{3}\left(2 - \frac{1}{4p}\right) \qquad \text{(E-12)}$$

This last expression is the quadratic analog to Martin's optimum alpha equation. It also indicates fourth order accuracy.

The system of equations (107) also possesses three point rows based on single elements. The effect of these terms on the system accuracy can be determined by performing a truncation error analysis across the elements, centered on the internal node. The procedure is the same, but now

$$A_1 = 1-40\ p$$

$$A_2 = 8+80\ p$$

$$B_1 = 1+40(1-\alpha)p$$

and

$$B_2 = 8-80(1-\alpha)p$$

After substitution, the result is

$$e_t = h\left(\frac{160}{p}\right)u_{2x} + 0(h^2) \qquad \text{(E-13)}$$

or, the truncation error is of order h unless

$$\left(\frac{160}{p}\right) = 0 \qquad \text{(E-14)}$$

This last expression states that only second order accuracy can be attained within an element if p is made large. Appendix G correlates this with the linear optimum alpha equation. It should be noted here that if p is made large, Equation (E-12) approaches a limiting alpha value as does the linear optimum alpha equation.

APPENDIX F

Treatment of Internal Nodes For Static Problems

Higher order polynomial interpolation functions with their associated internal nodes can be used to improve the overall field variable representation within an element. Once employed to attain the equations or matrices for application of the finite-element method, it is noted that these internal nodes do not connect with the nodes of other elements during the assembly process. Consequently, the degrees of freedom (here, for example, normalized temperature) associated with internal nodes do not affect interelement continuity (Ref 5:155-156).

Because of this, these internal nodal degrees of freedom may be eliminated at the element level before assembly to reduce the overall size of the assembled system matrices. The decision to do this depends on the nature of the problem and especially on the shape of the element. This elimination process, called "Static Condensation," (Ref 4:220-221) is presented here to emphasize the fact that internal nodes need not be retained after information they supply is ascertained.

For one quadratic element

$$\begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ a_7 & a_8 & a_9 \end{bmatrix} \begin{Bmatrix} u_1 \\ u_2 \\ u_3 \end{Bmatrix} = \begin{Bmatrix} R_1 \\ R_2 \\ R_3 \end{Bmatrix} \qquad \text{(F-1)}$$

where u_2 is the internal nodal degree of freedom and the R's represent the corresponding resultant actions. $\underline{A}$ can be rearranged and partitioned to segregate this internal value as

$$\left[\begin{array}{c|c} [a_{11}] & [a_{12}] \\ \hline [a_{21}] & [a_{22}] \end{array}\right] \left\{\begin{array}{c} \{x_1\} \\ \hline [x_2] \end{array}\right\} = \left\{\begin{array}{c} \{R_1\} \\ \hline [R_2] \end{array}\right\} \qquad \text{(F-2)}$$

where $\{x_1\}$ is a column vector of the external nodal values, and $\{R_1\}$ is the corresponding vector of resultant nodal actions.

Equation (F-2) may be expanded into

$$[a_{11}]\ \{x_1\} + [a_{12}]\ [x_2] = \{R_1\} \qquad \text{(F-3)}$$

and

$$[a_{21}]\ \{x_1\} + [a_{22}]\ [x_2] = [R_2] \qquad \text{(F-4)}$$

If Equation (F-4) is solved for $[R_2]$ and this result substituted into Equation (F-3), the result is

$$\left[[a_{11}] - [a_{12}]\ [a_{22}]^{-1}\ [a_{21}]\right] \{x_1\} = \{R_1\} - [a_{12}]\ [a_{22}]^{-1}\ [R_2] \qquad \text{(F-5)}$$

which is the condensed system

$$[A_c]\ \{x_1\} = [R_c] \qquad \text{(F-6)}$$

APPENDIX G

Comparison of Linear and Quadratic Factors and Equivalence of Linear and Quadratic Interpolation Function Analysis

To insure accurate and consistent error comparisons between the linear and quadratic finite-element formualtions, where possible, procedures and derivations of the latter were performed analogous to those of the former. The following paragraphs list several of these factors, culminating in a direct relationship and equivalance between the linear and quadratic interpolation function alpha values and solutions.

The quadratic element stiffness matrix and element mass matrix are given respectively by Equations (73) and (89). The corresponding linear matrices are

$$\underline{K}^{(e)} \triangleq \frac{1}{\Delta x} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{G-1}$$

and

$$\underline{M}^{(e)} \triangleq \frac{x}{6} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \tag{G-2}$$

The assembled system matrices, $\underline{A}$ and $\underline{B}$, for the quadratic case of three elements and four external nodes, are given by Equation (106). The corresponding linear system is

$$\begin{bmatrix} 2+6\alpha p & 1-6\alpha p & 0 & 0 \\ 1-6\alpha p & 4+12\alpha p & 1-6\alpha p & 0 \\ 0 & 1-6\alpha p & 4+12\alpha p & 1-6\alpha p \\ 0 & 0 & 1-6\alpha p & 2+6\alpha p \end{bmatrix} \begin{Bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{Bmatrix}^{k+1}$$

$$= \begin{bmatrix} 2-6(1-\alpha)p & 1+6(1-\alpha)p & 0 & 0 \\ 1+6(1-\alpha)p & 4-12(1-\alpha)p & 1+6(1-\alpha)p & 0 \\ 0 & 1+6(1-\alpha)p & 4-12(1-\alpha)p & 1+6(1-\alpha)p \\ 0 & 0 & 1+6(1-\alpha)p & 2-6(1-\alpha)p \end{bmatrix} \begin{Bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{Bmatrix}^{k} \qquad \text{(G-3)}$$

For the quadratic case, a truncation error analysis across two adjoining elements employs the coefficients of Equation (106) and yields

$$\alpha = \frac{1}{3}\left(2 - \frac{1}{4p}\right) \qquad \text{(G-4)}$$

for fourth order accuracy. The analogous linear truncation error analysis employs the coefficients of Equation (G-3) and yields

$$\alpha = \frac{1}{2}\left(1 + \frac{1}{6p}\right) \qquad \text{(G-5)}$$

for fourth order accuracy.

Without analogy, the quadratic formulation establishes that within an element or at the internal node

$$\frac{160}{p} = 0 \qquad \text{(G-6)}$$

for second order accuracy.

For both formulations as the value of the Fourier modulus, p, is increased, the accuracy of the finite-element solution decreases. This is true by definition, since for a selected value of Δx, p is increased by increasing Δt, or enlarging the element. In the linear case, as p becomes large, the optimum alpha value approaches .5 which is the second order accuracy Crank-Nicolson value. In the quadratic case, as p becomes large, the optimum alpha value approaches .6666 which is shown in the results section and Appendix H to also yield only second order accuracy. This is in agreement with Equation (G-6) which states that second order accuracy is attainable if p is made large. Appendix H also shows that, if a quadratic solution employing an optimum alpha of Equation (G-4), or an alpha of a different scheme, is attempted without considering the requirements of Equation (G-6) for internal nodes, the result is less than second order accurate.

Fourth order accuracy can only be expected if the internal nodes are accounted for without direct application in the solution process. Appendix F considers one procedure called "static condensation." For this transient heat conduction problem, the internal nodes can be treated in a similar manner as external nodes. The following

paragraphs show that an equivalency relationship exists between the quadratic alphas of Equation (G-4) and the linear alphas of Equation (G-5). This equivalency indicates that solution by quadratic interpolation is equal to solution by linear interpolation across an interval of $\Delta x/2$, and that solution by quadratic interpolation can be interpreted as more accurate than the corresponding linear solution across an interval Δx only by virtue of smaller mesh spacing. That is, at the external nodes, the quadratic solution is equivalent to the linear solution of twice the number of nodes.

The equivalency between quadratic and linear solutions is easily found by considering the relationship between the respective optimum alphas as p is made large. For the quadratic case

$$\lim_{p \to \text{large}} \alpha_{opt,Q} = .6666 \qquad \text{(G-7)}$$

Similarly, for the linear case

$$\lim_{p \to \text{large}} \alpha_{opt,L} = .5 \qquad \text{(G-8)}$$

Then

$$\lim_{p \to \text{large}} \left| \alpha_{opt,Q} - \alpha_{opt,L} \right| = .1666 \qquad \text{(G-9)}$$

If n is introduced as the parameter defining the p step increments as the limit is approached, then

$$\left|\alpha_{opt,Q} - \quad _{opt,L}\right| = nv \qquad \text{(G-10)}$$

where v may be called the variation and is that value required to satisfy Equation (G-10) as n is increased. As is noted

$$v = \frac{.16666}{4} = .04167 \qquad \text{(G-11)}$$

The quadratic limit is approached from below in the same steps as the linear limit is approached from above. Table G-I displays this and shows that for each quadratic alpha of Equation (G-4), the corresponding linear alpha is as derived by Equation (G-5). For example, for step one, $n = 1$, and $nv = .04167$. The quadratic case then yields

$$.6666 - (1)(.04167) = .625 \qquad \text{(G-12)}$$

while, for the linear case

$$.5000 + (1)(.04167) = .54167 \qquad \text{(G-13)}$$

The result of Equation (G-12) is equal to the value derived from Equation (G-4) for $p = 2.0$. The result of Equation (G-13) is equal to the value derived from Equation (G-5) for $p = 2.0$.

TABLE G-I

Equivalency of Optimum Alpha Values
For Quadratic and Linear Interpolation Functions

n	Quadratic (.6666) $\alpha_{opt,Q}$	Linear (.5000) $\alpha_{opt,L}$	p
1	.6250	.54167	2
2	.5833	.5833	1
3	.54167	.625	.6666
4	.5000	.6666	.5

It is noted that for the Fourier modulus of one, the optimum alpha values are, in fact, equal.

The result of Table G-I is that for each attained quadratic optimum alpha at each node, there is an equivalent linear optimum alpha which agrees with the value derived by linear interpolation. Therefore, whether the solution approach be quadratic or linear, the degree of implicitness or explicitness as defined by alpha is the same.

APPENDIX H

Computer-Generated Plots of Results

This appendix contains the graphical results of this project, presented as plots of Discretization Error Ratio versus Time, for selected alpha values of the Crank-Nicolson, optimum implicit, and fully implicit schemes. Each of the three error norms, before mentioned, are so displayed.

As in Martin's study, the pointwise error is measured at $x = 0.1$. The generalized mean is the sum of the absolute values of the pointwise errors at nine evenly spaced nodes. It shows the effect of the parameter α on the pointwise error over the whole interval. The maximum error at any node, or discrete Tchebycheff norm, shows the effect of this same parameter on the maximum deviation at any node between the time solution and the numerical solution.

Discretization error ratio was previously defined. For its calculation, solutions were attained and compared for space domain intervals of 10 and 20, and, 20 and 40. On each plot, $P = p = \frac{\Delta\theta}{(\Delta x)^2}$ is the Fourier modulus and DX represents the space interval Δx . Of course, alpha is the parameter of note, sometimes referred to as the "degree of implicitness," because it is the measure of the weight placed on the temperatures in the new time step of the numerical scheme.

Each section of graphs is introduced with a short descriptive note. In all graphs, the exact analytical solution has been substituted at the first time step to eliminate the discontinuity between the initial and boundary conditions. Graphs annotated CDF and CDH are for the

finite-difference cases of $p = 1.0$ and $p = 2.0$, respectively. Graphs annotated CET and CEY are for the finite-element cases of $p = 1.0$ and $p = 2.0$, respectively.

Section I

Results for the Problem Using Finite Differences

This section shows the graphical results for the solution of the problem by finite-differences. Run identifiers are CDF and CDH .

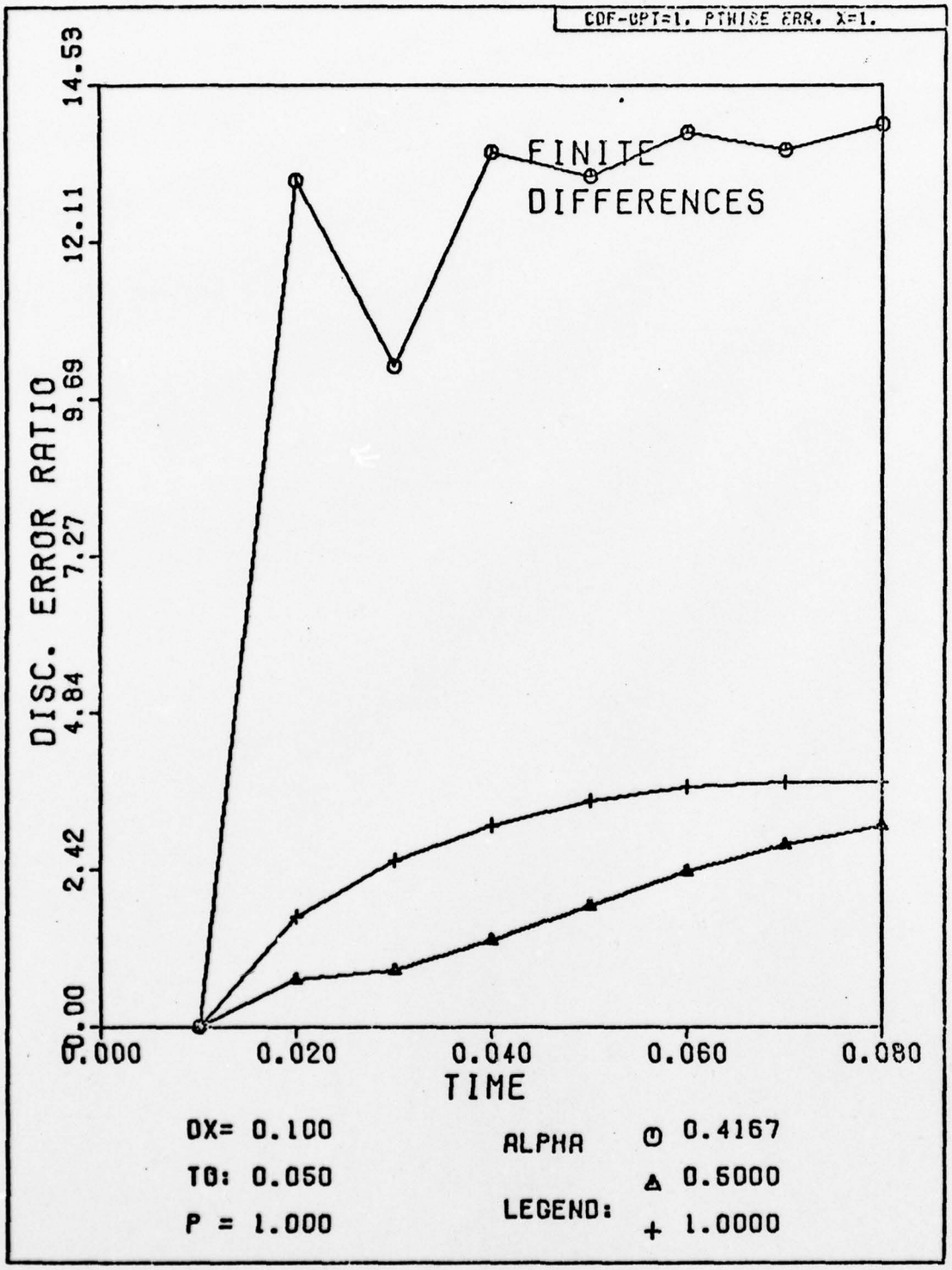

Fig. H-1. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

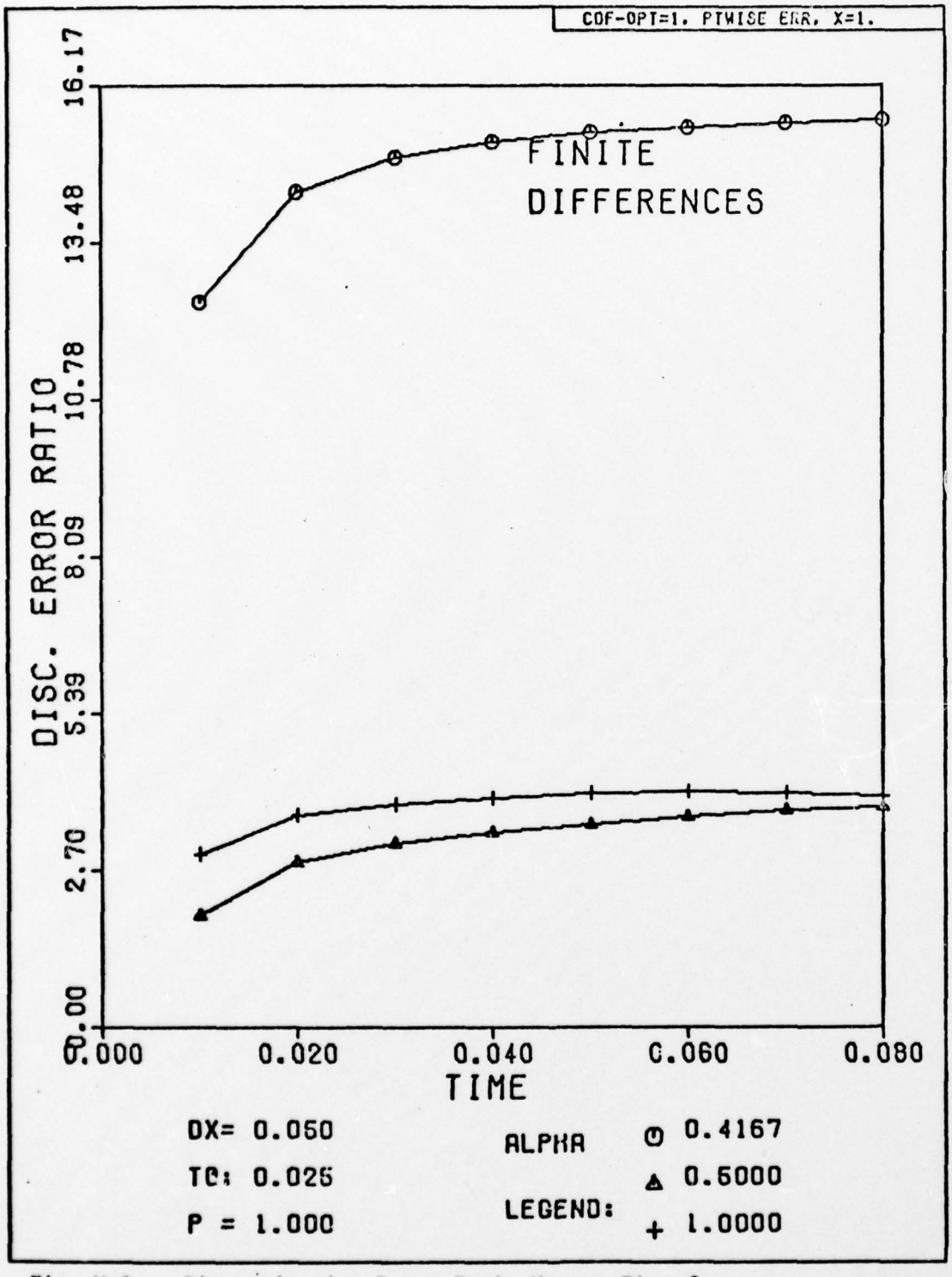

Fig. H-2. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

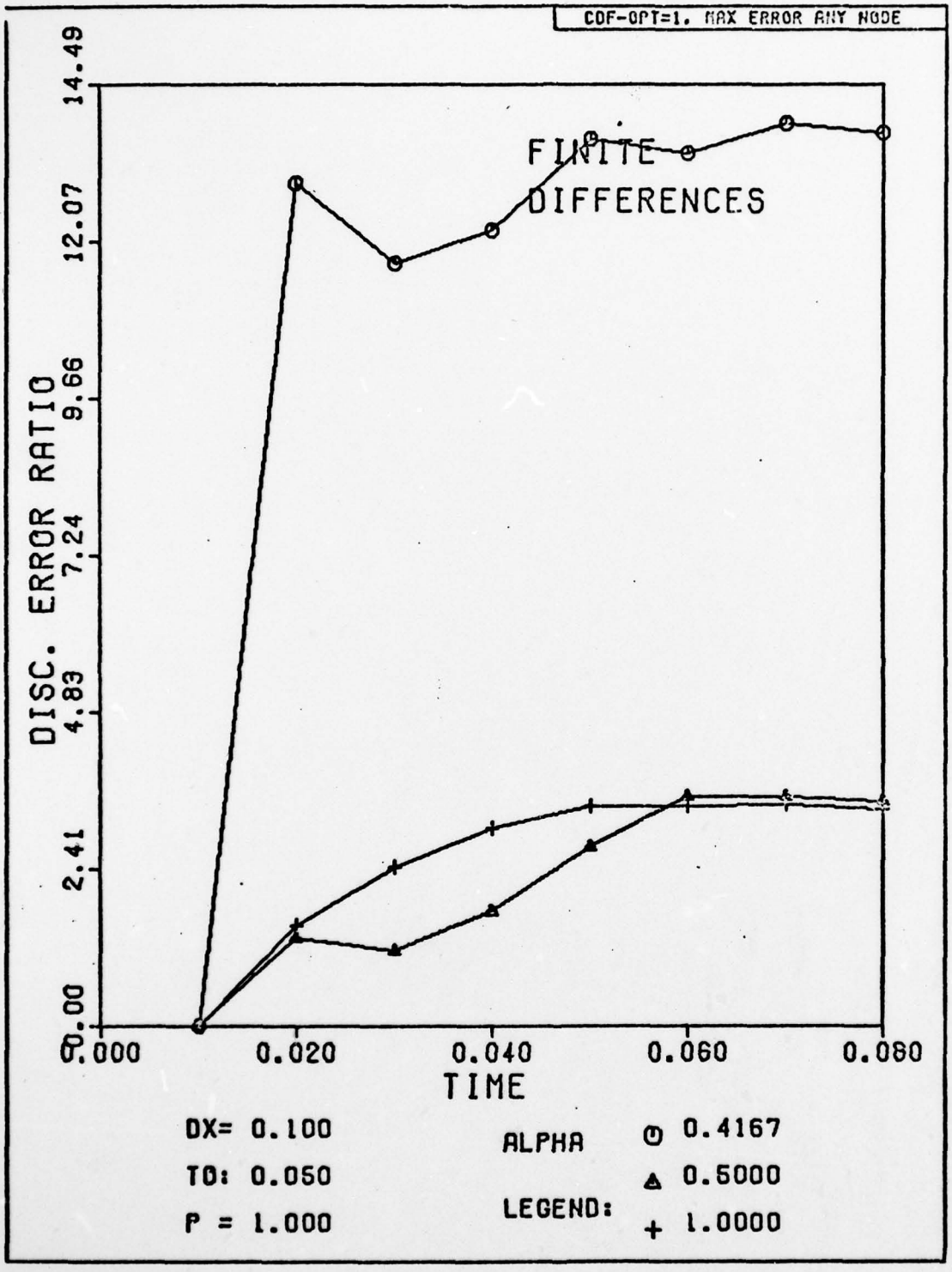

Fig. H-3. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

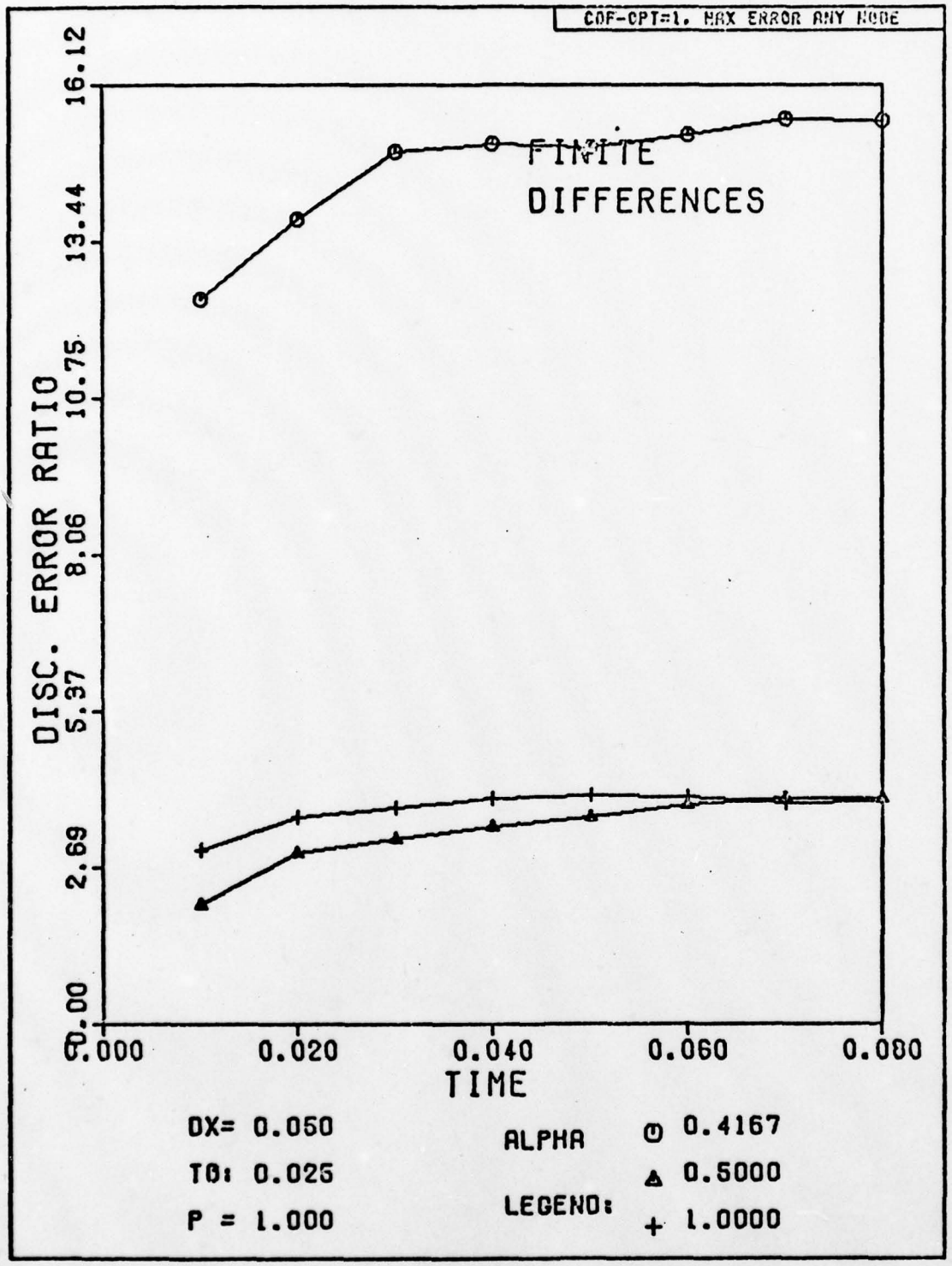

Fig. H-4. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

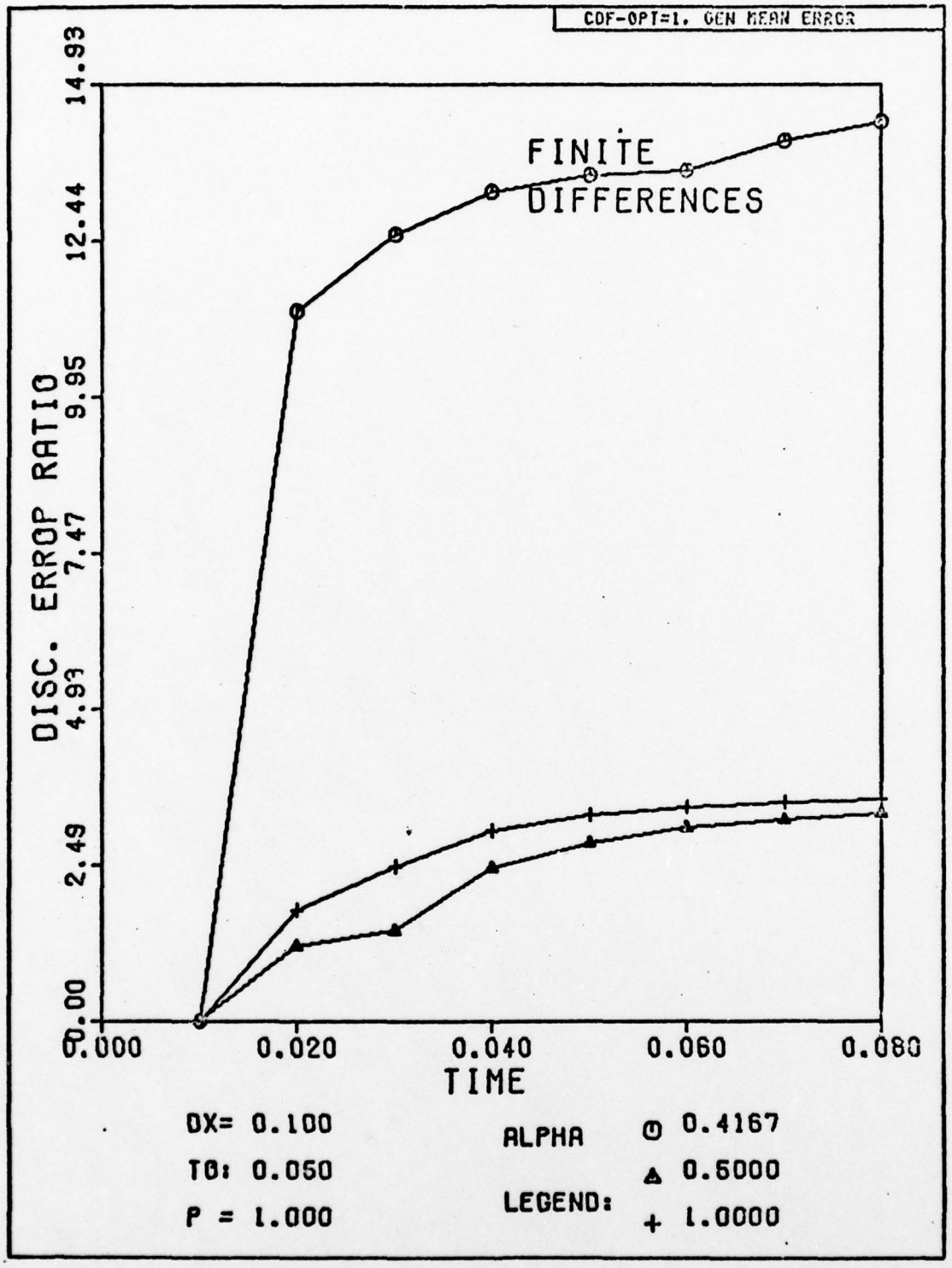

Fig. H-5. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

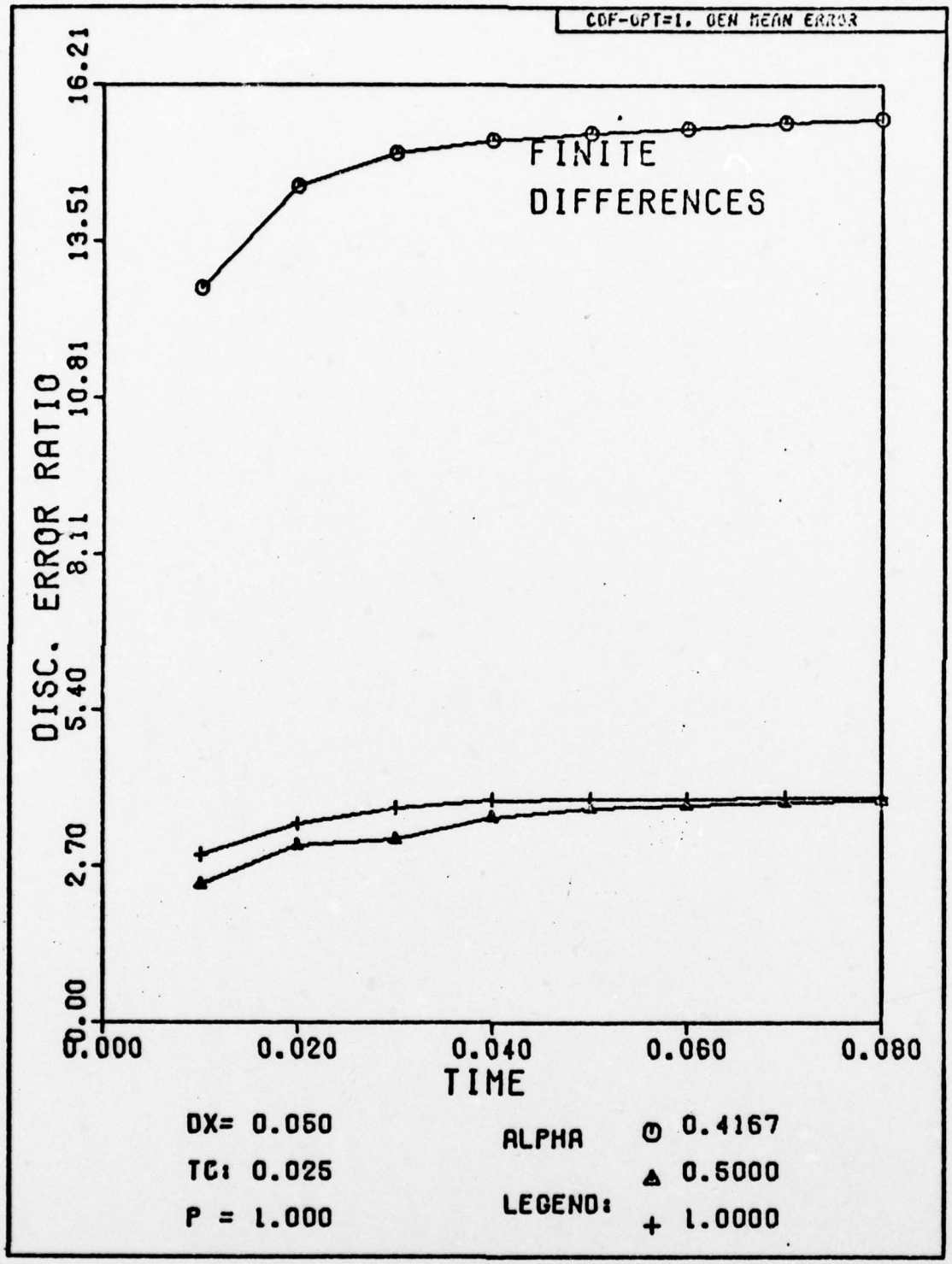

Fig. H-6. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

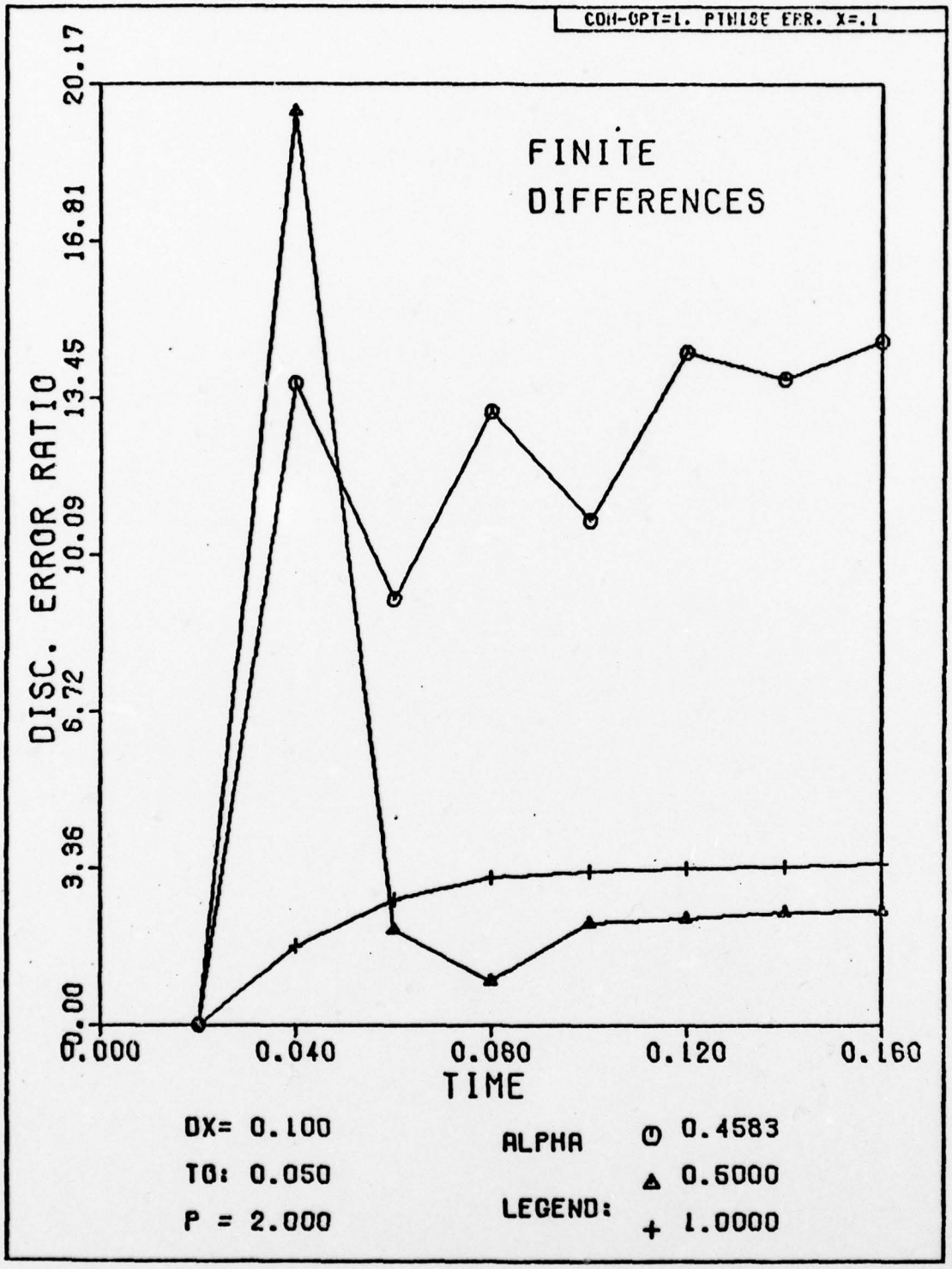

Fig. H-7. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

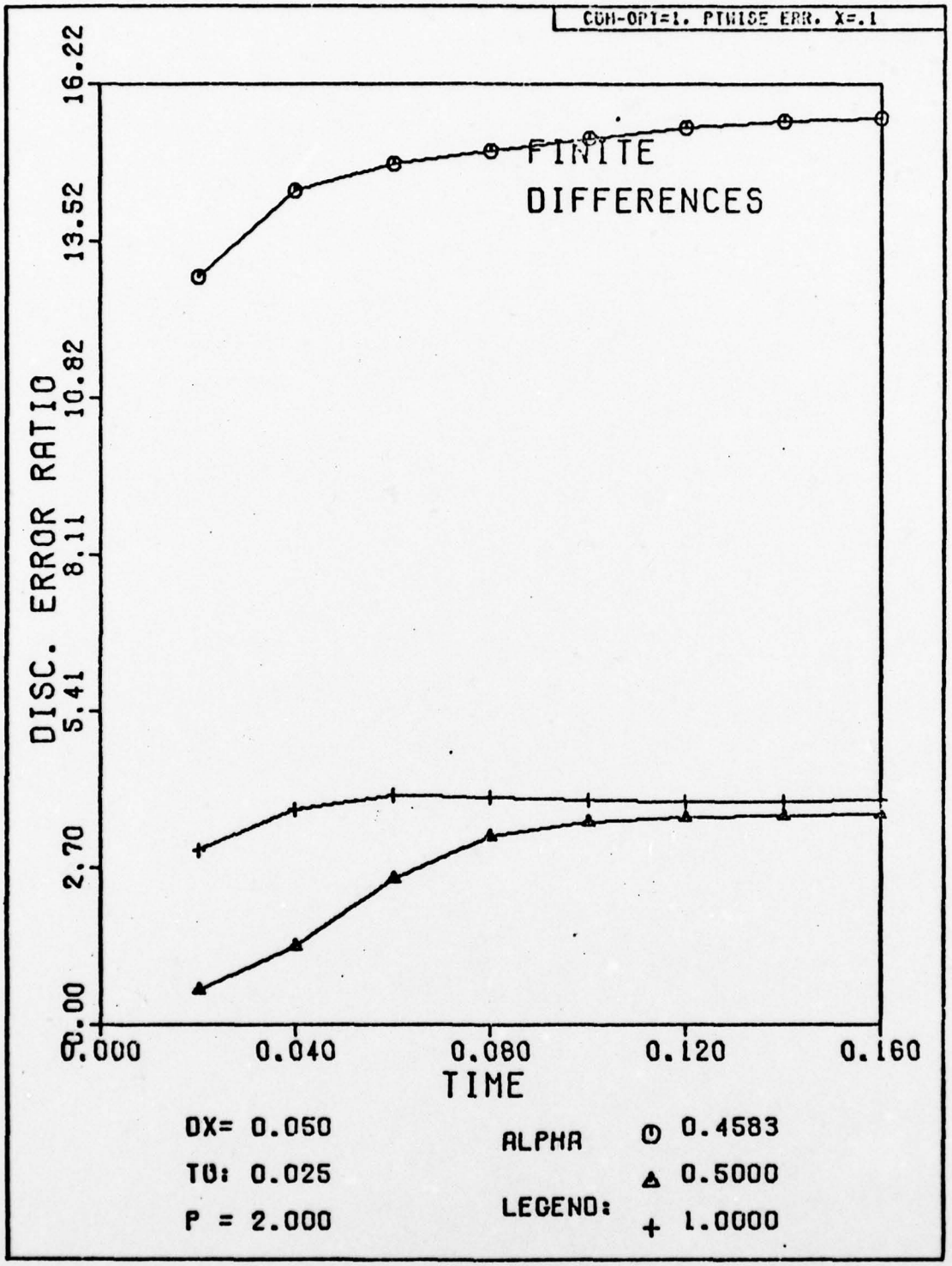

Fig. H-8. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

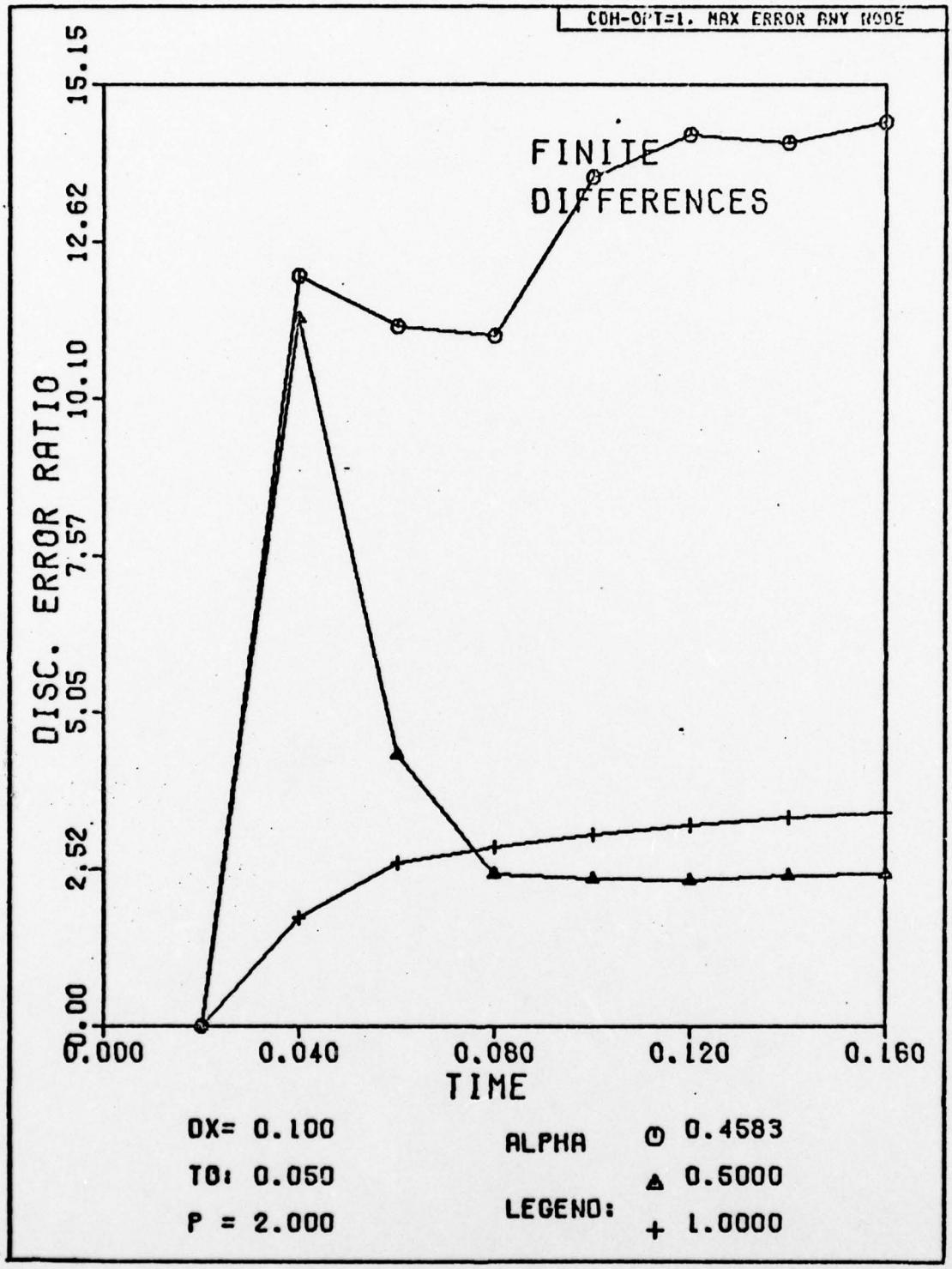

Fig. H-9. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

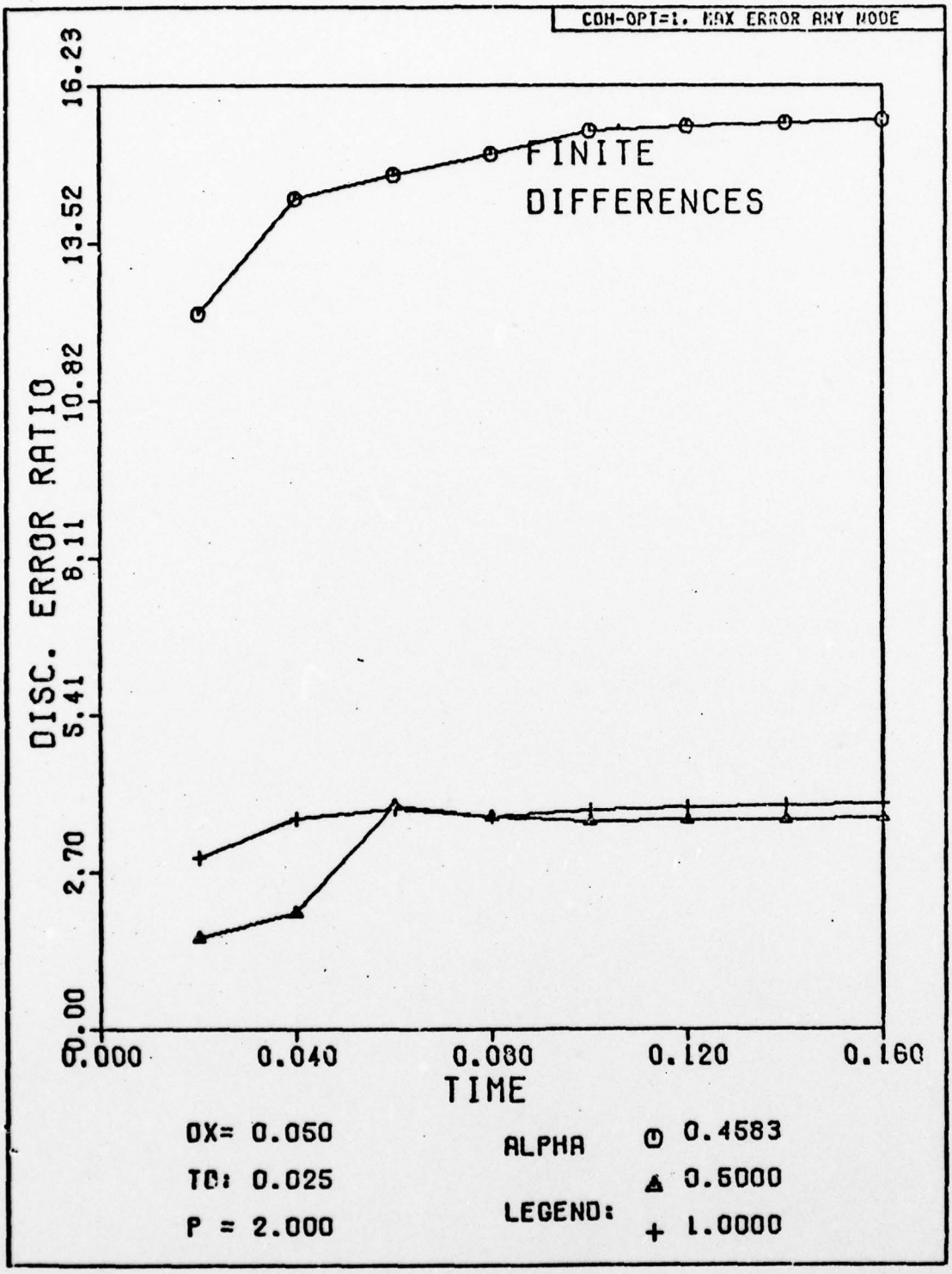

Fig. H-10. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

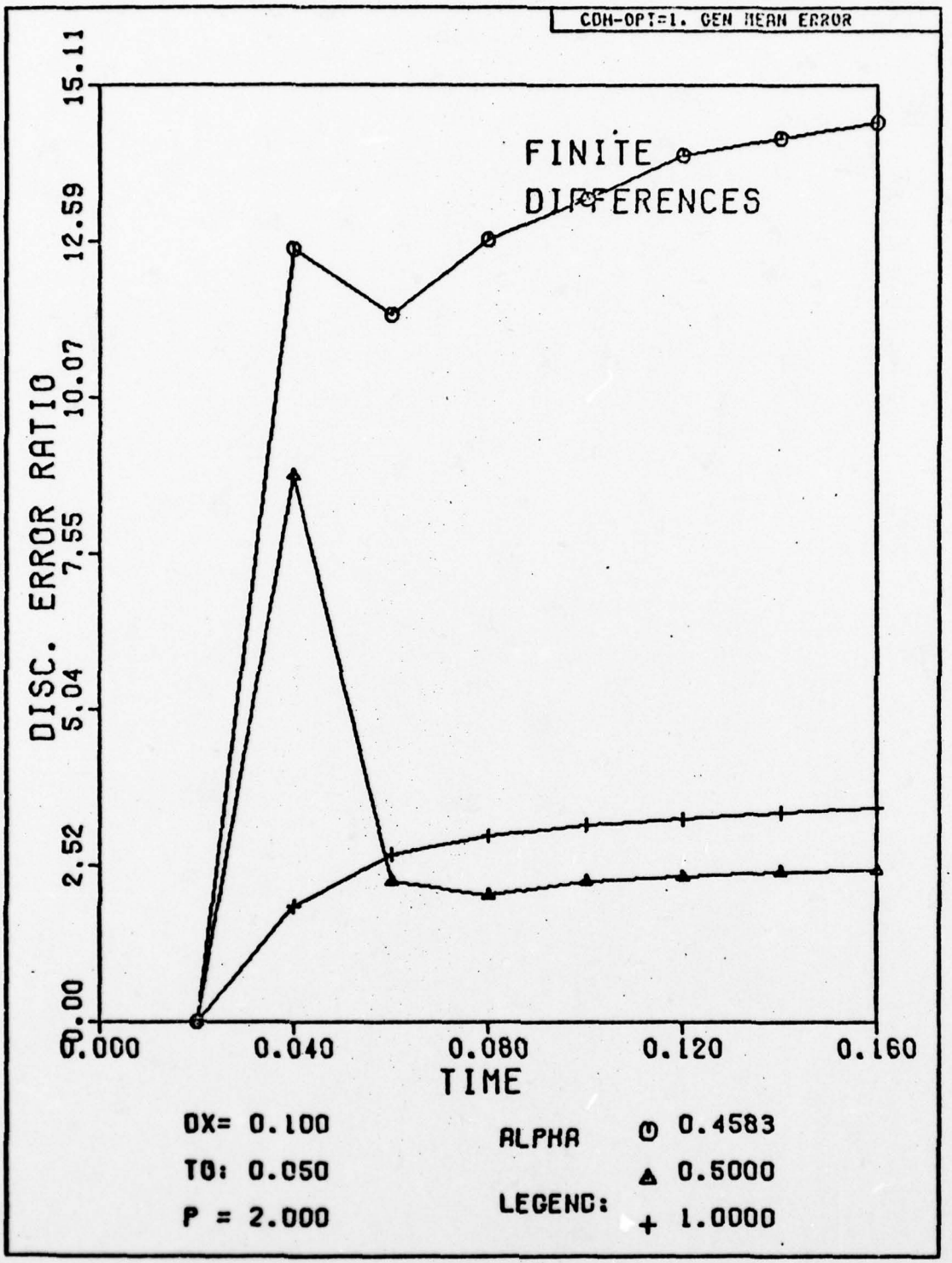

Fig. H-11. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

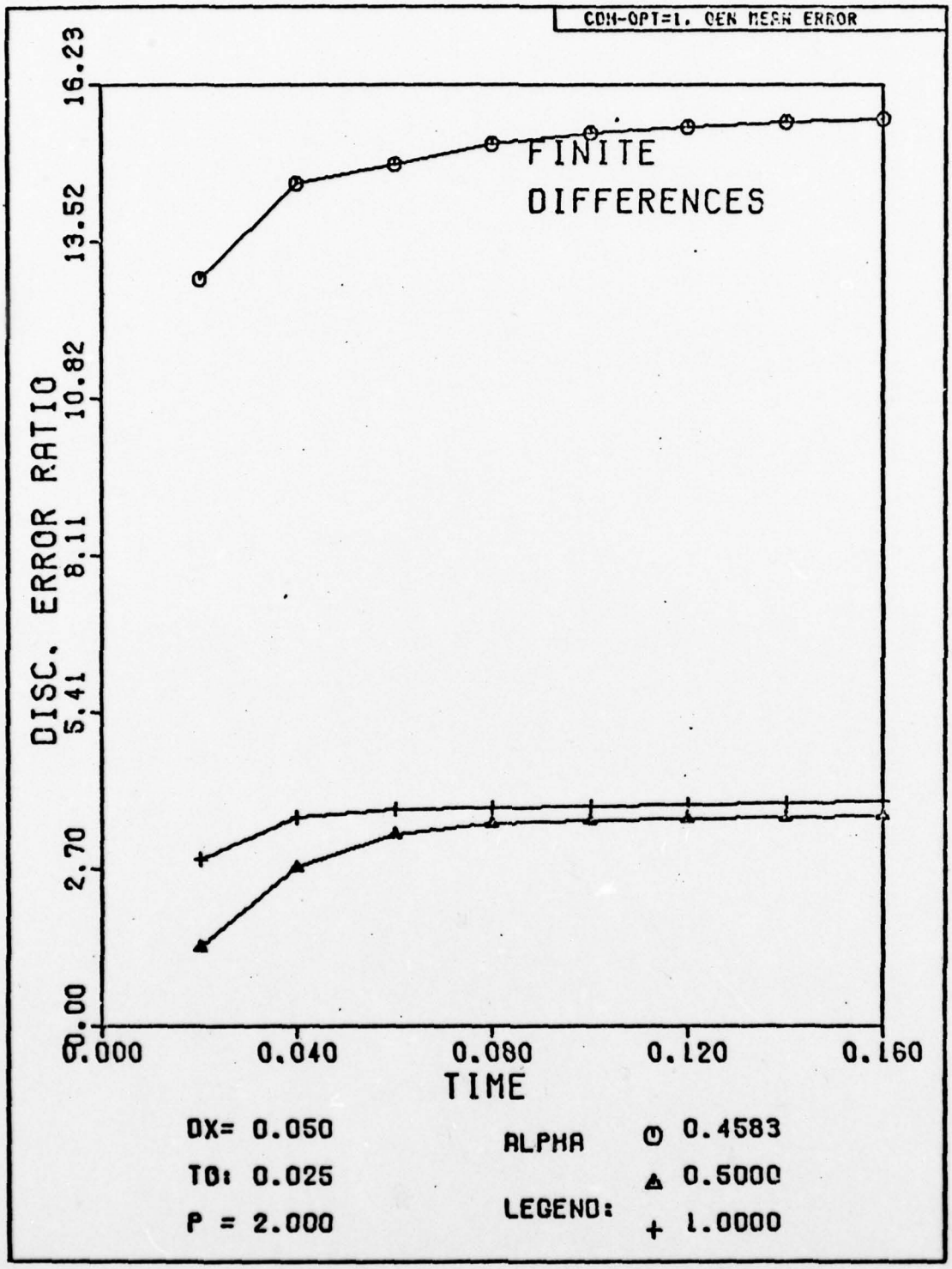

Fig. H-12. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

Section II

Results for the Problem Using Finite Elements and a Linear Interploation Function

This section shows the graphical results for the solution of the problem by finite-elements, linear interpolation. Run identifiers are CET and CEY .

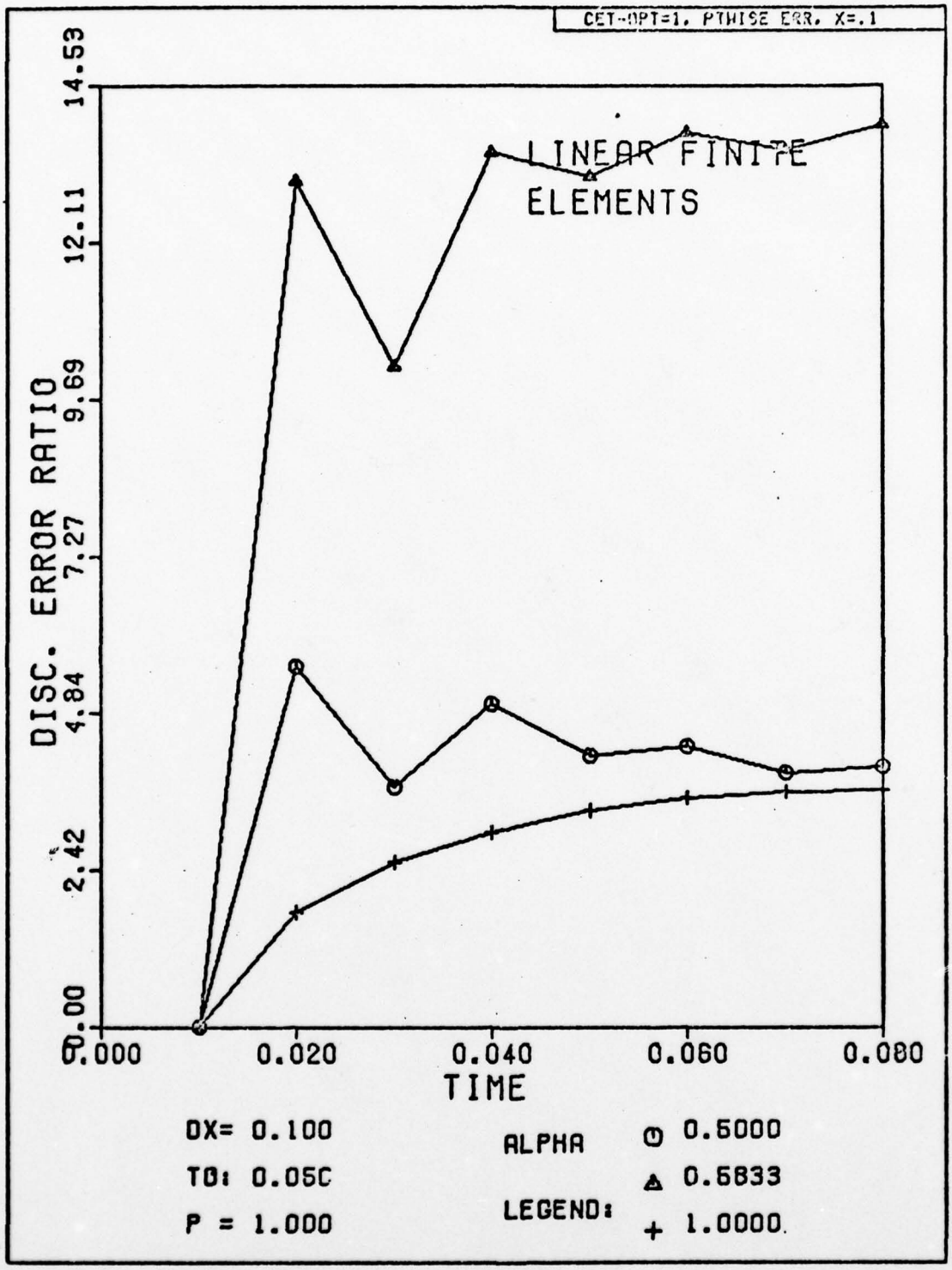

Fig. H-13. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

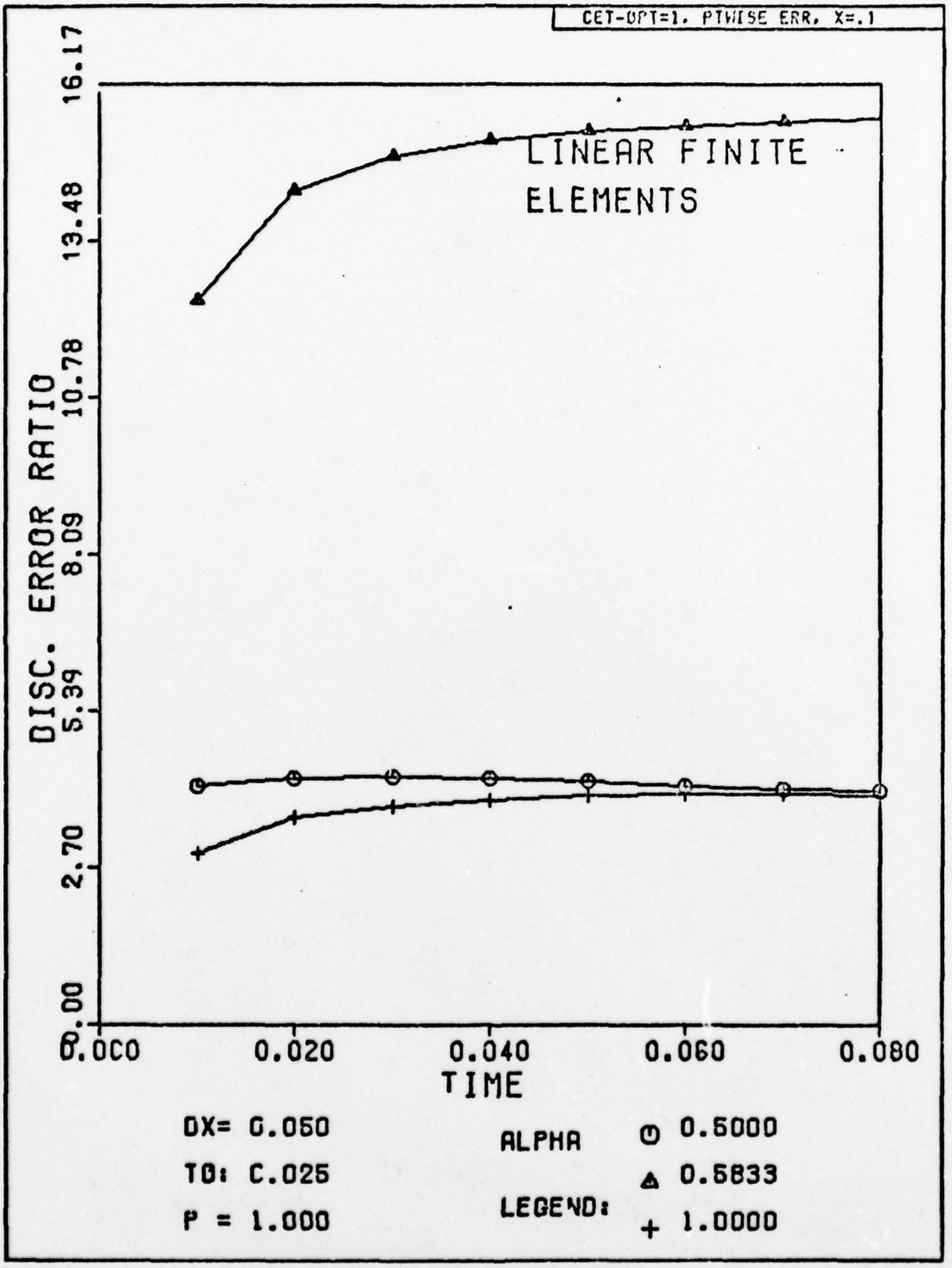

Fig. H-14. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

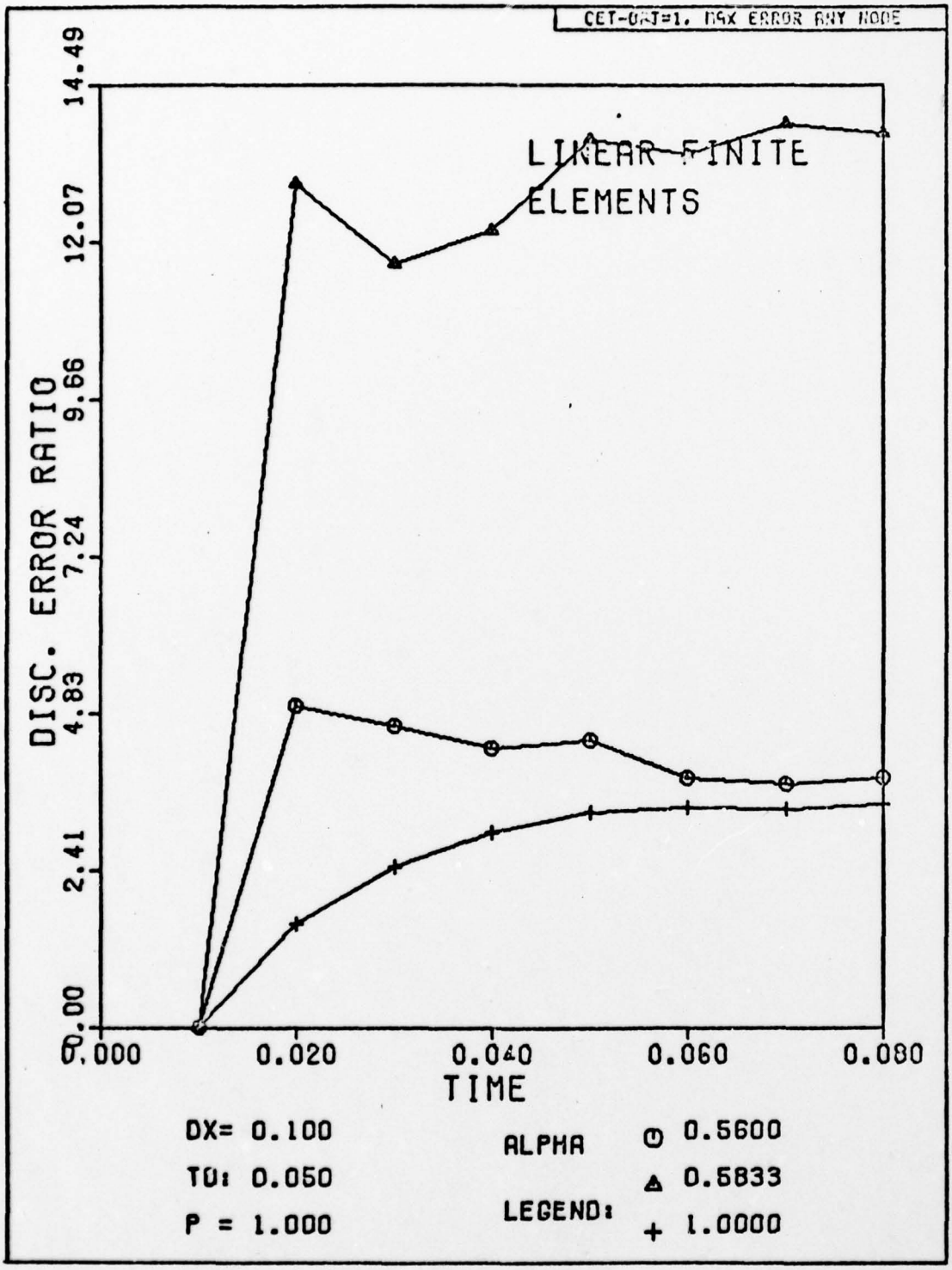

Fig. H-15. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

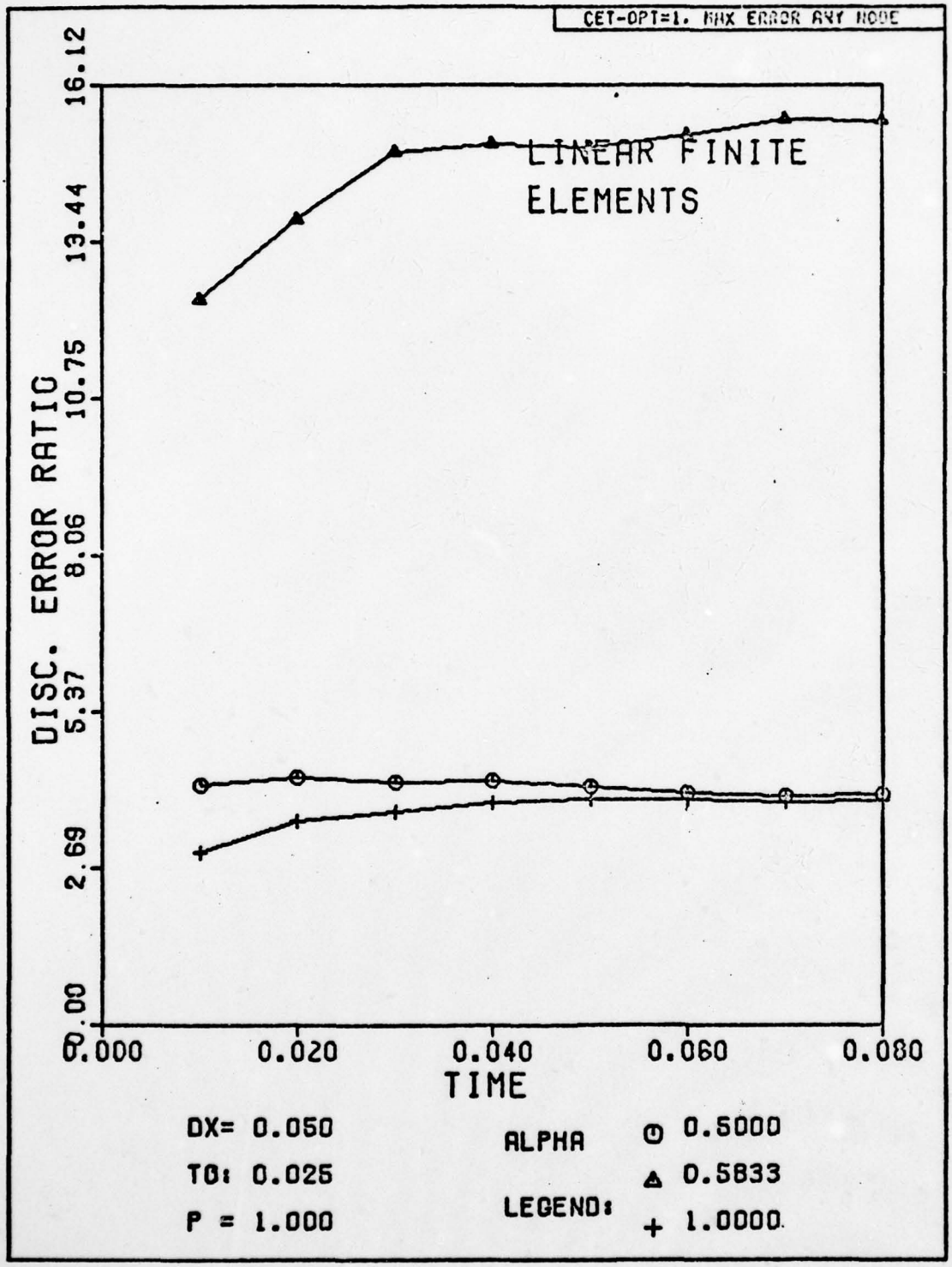

Fig. H-16. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

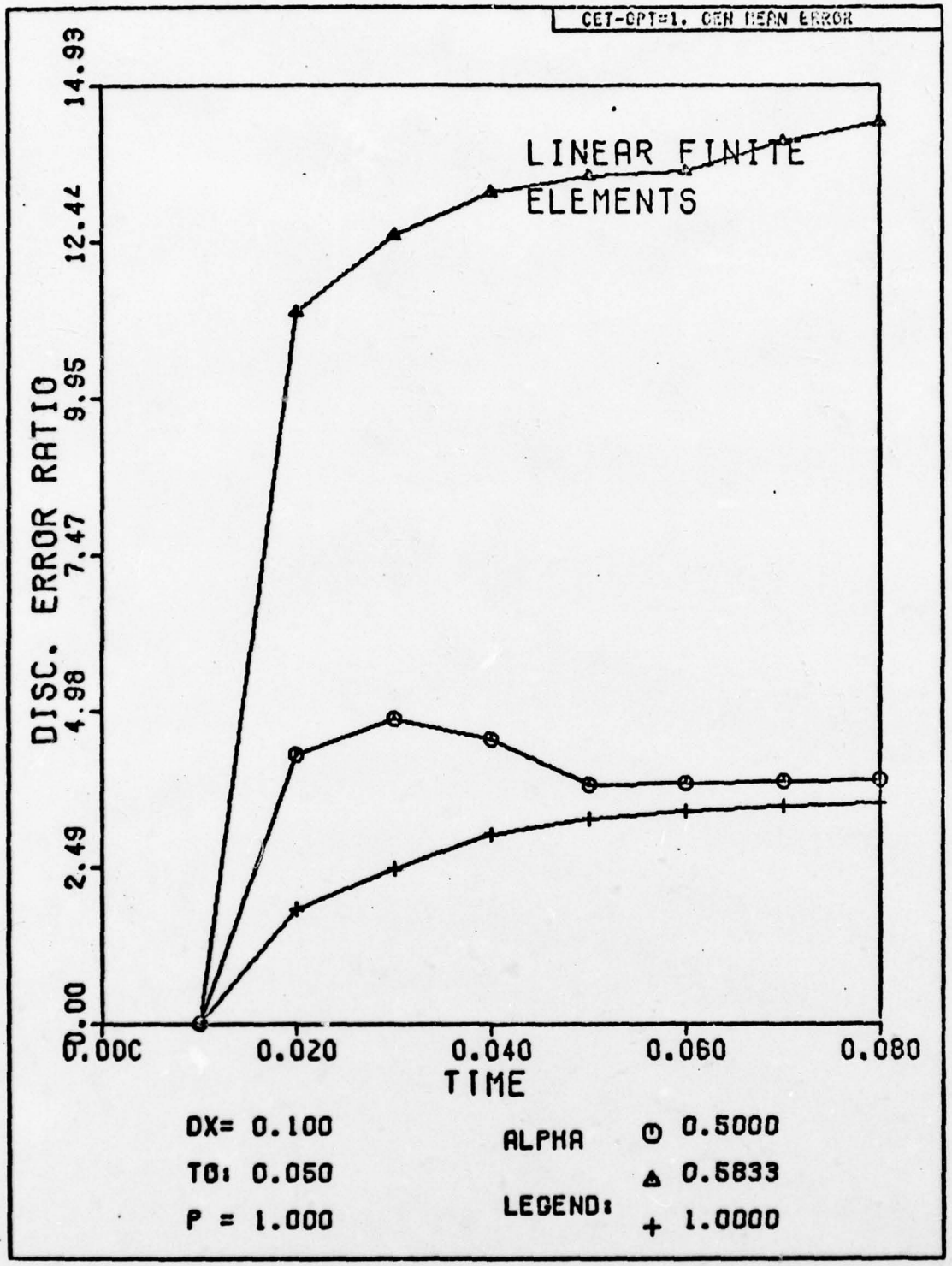

Fig. H-17. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

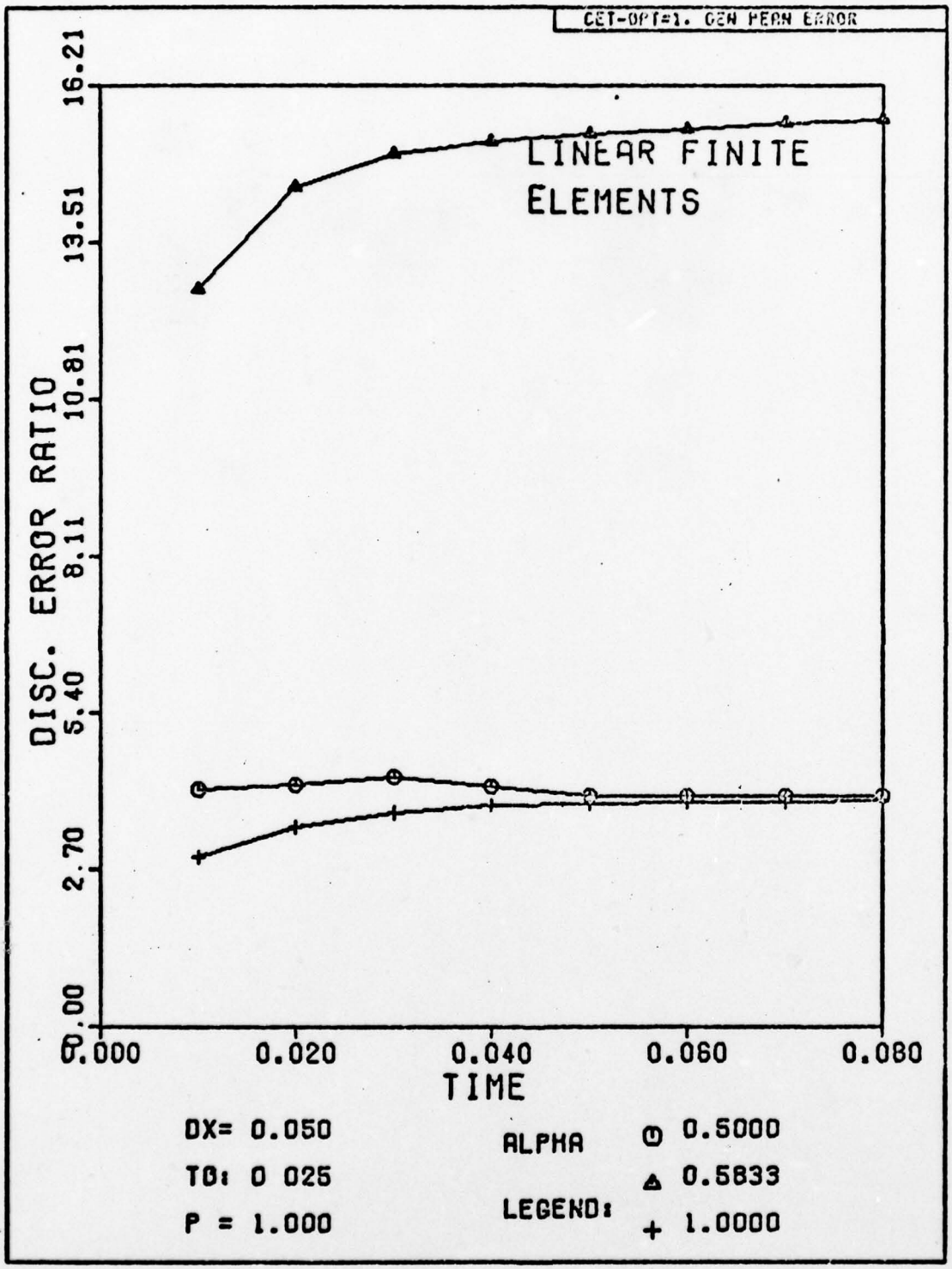

Fig. H-18. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

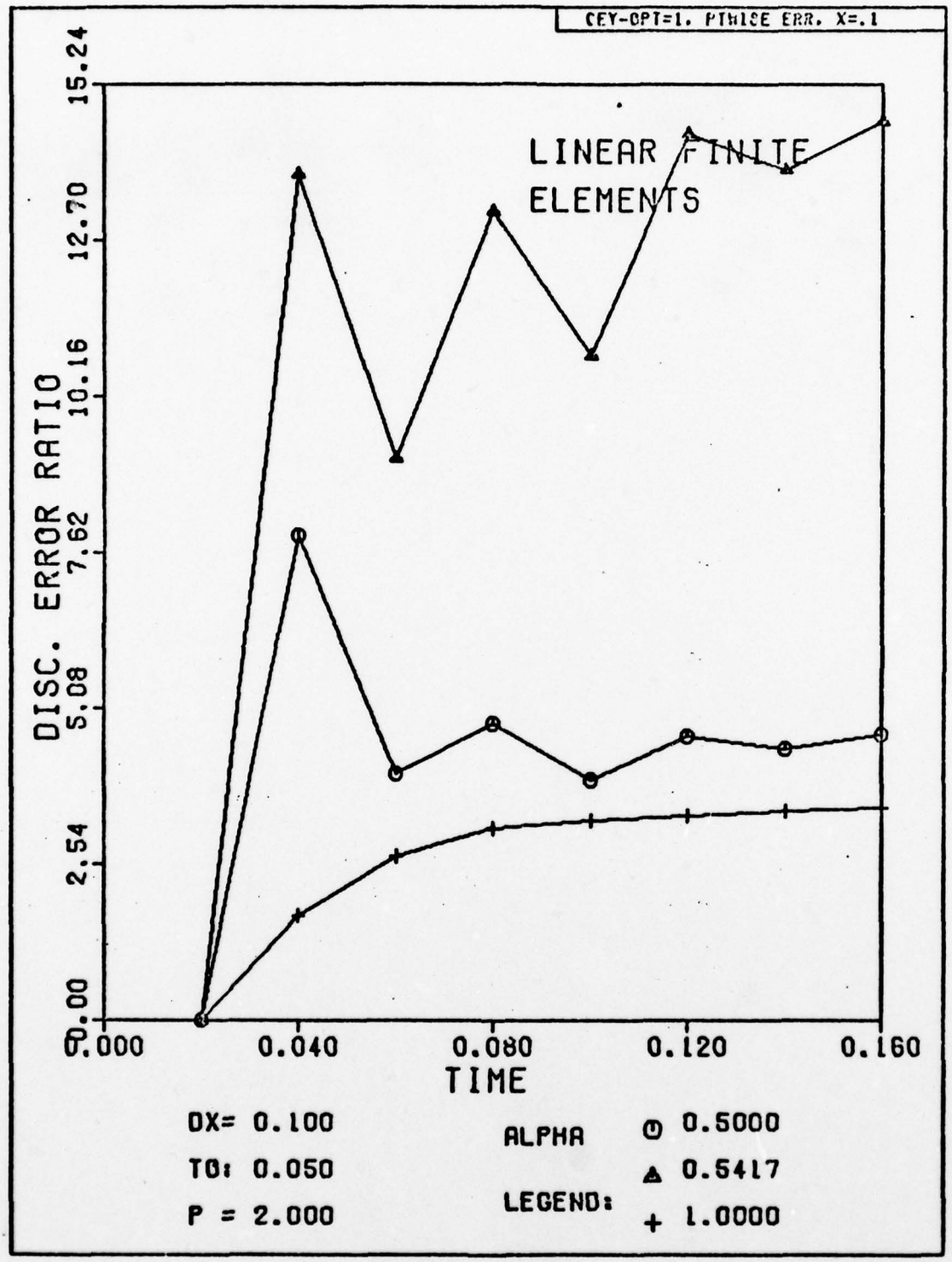

Fig. H-19. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

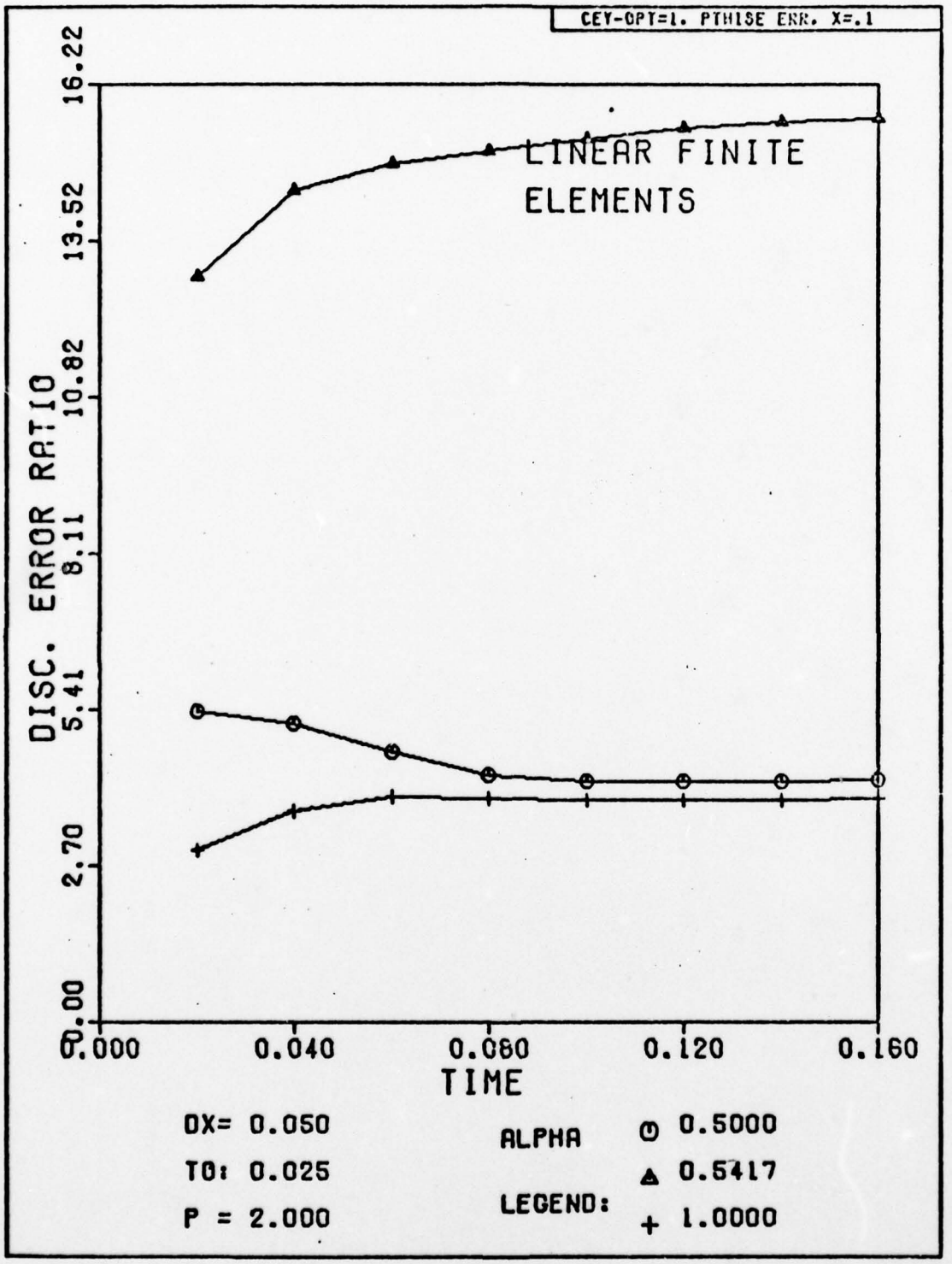

Fig. H-20. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

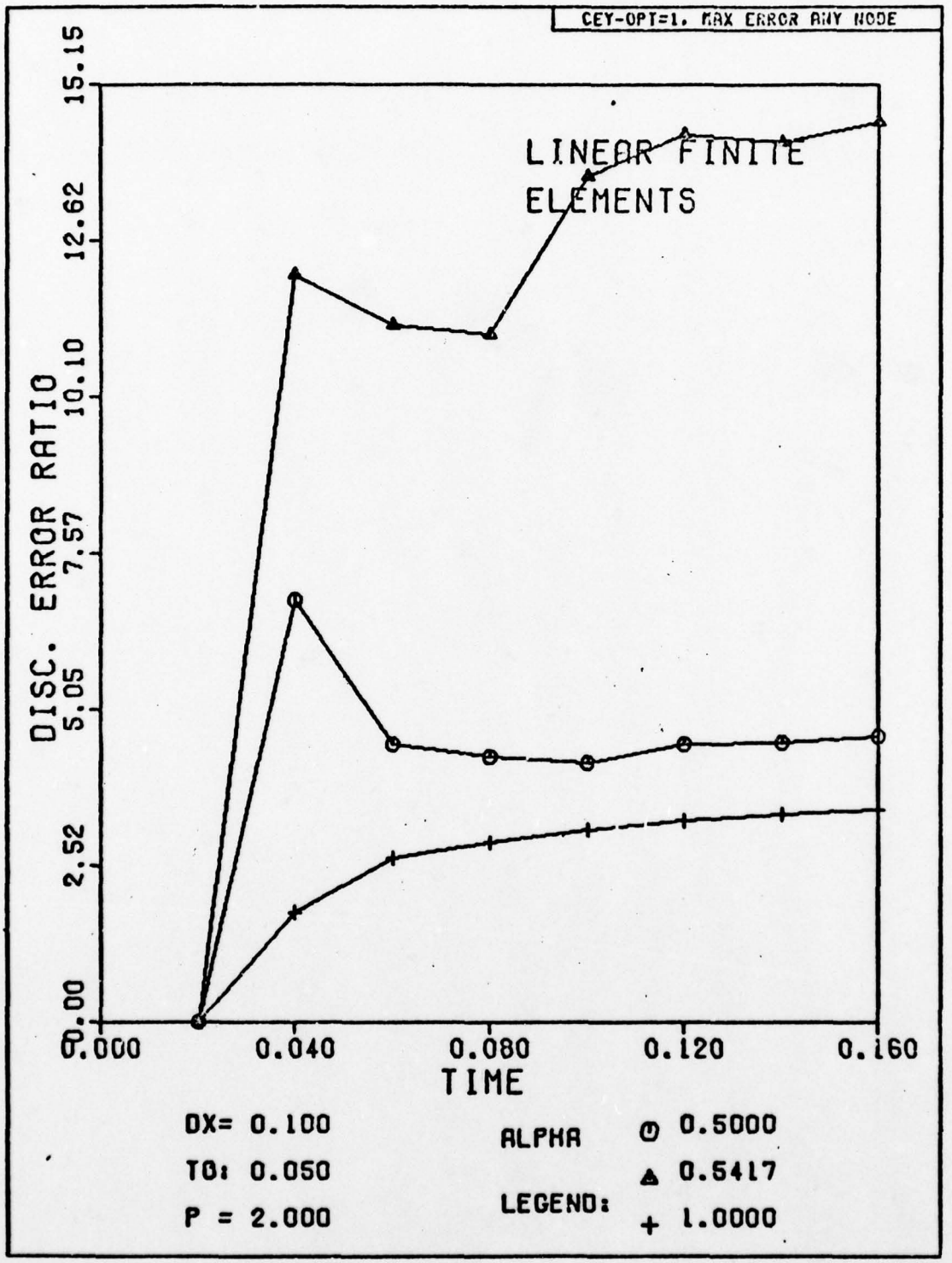

Fig. H-21. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

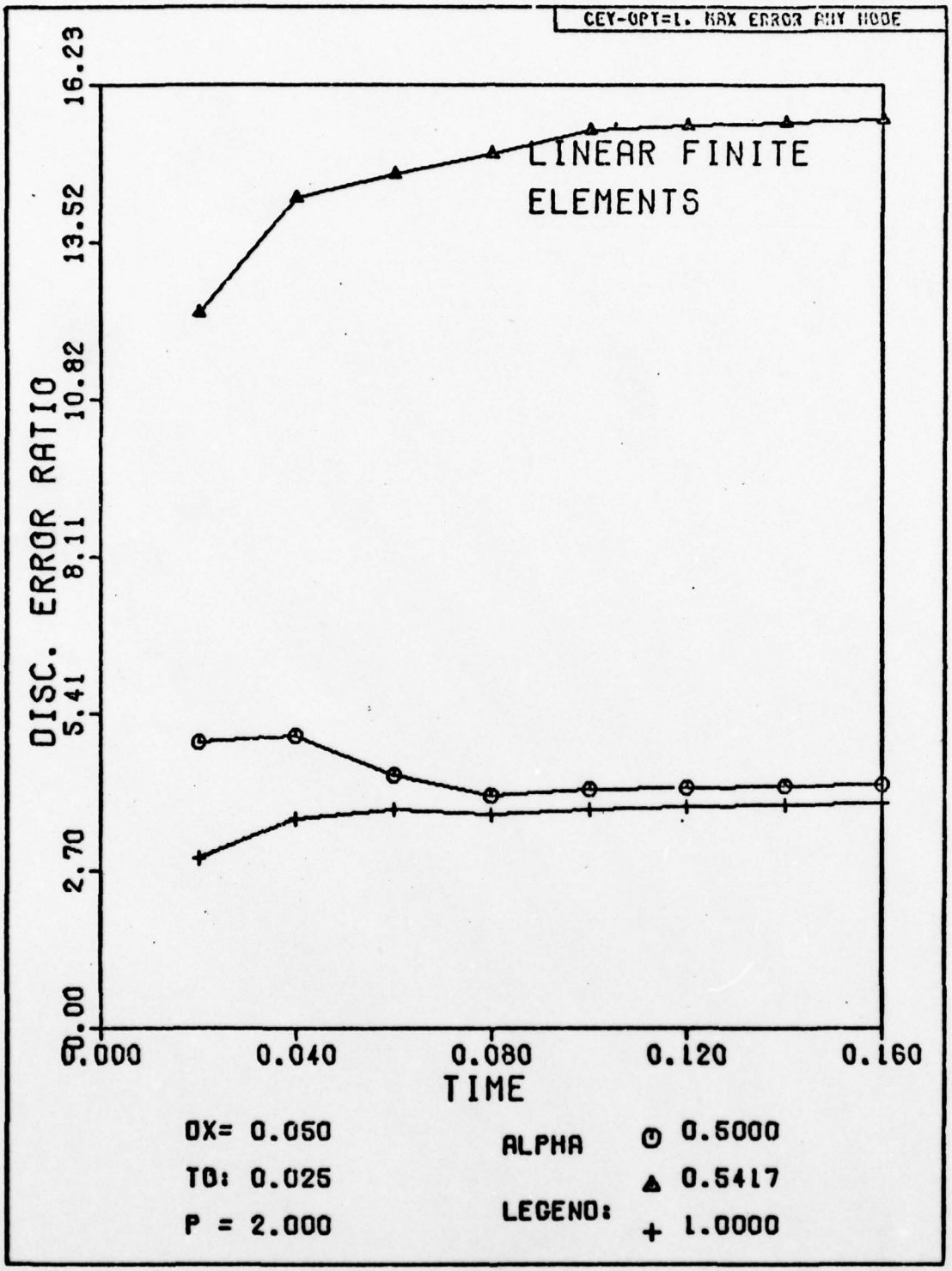

Fig. H-22. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

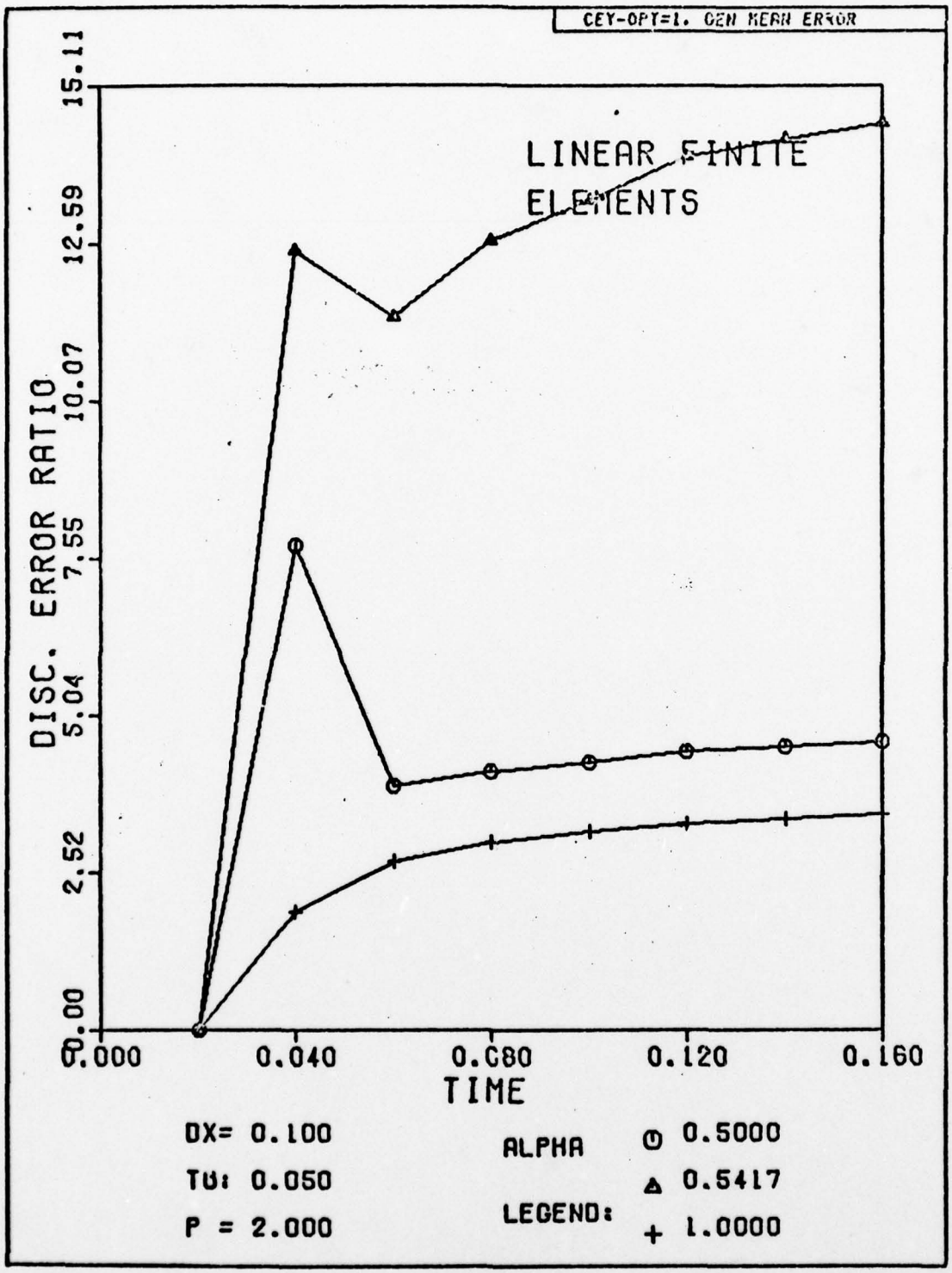

Fig. H-23. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

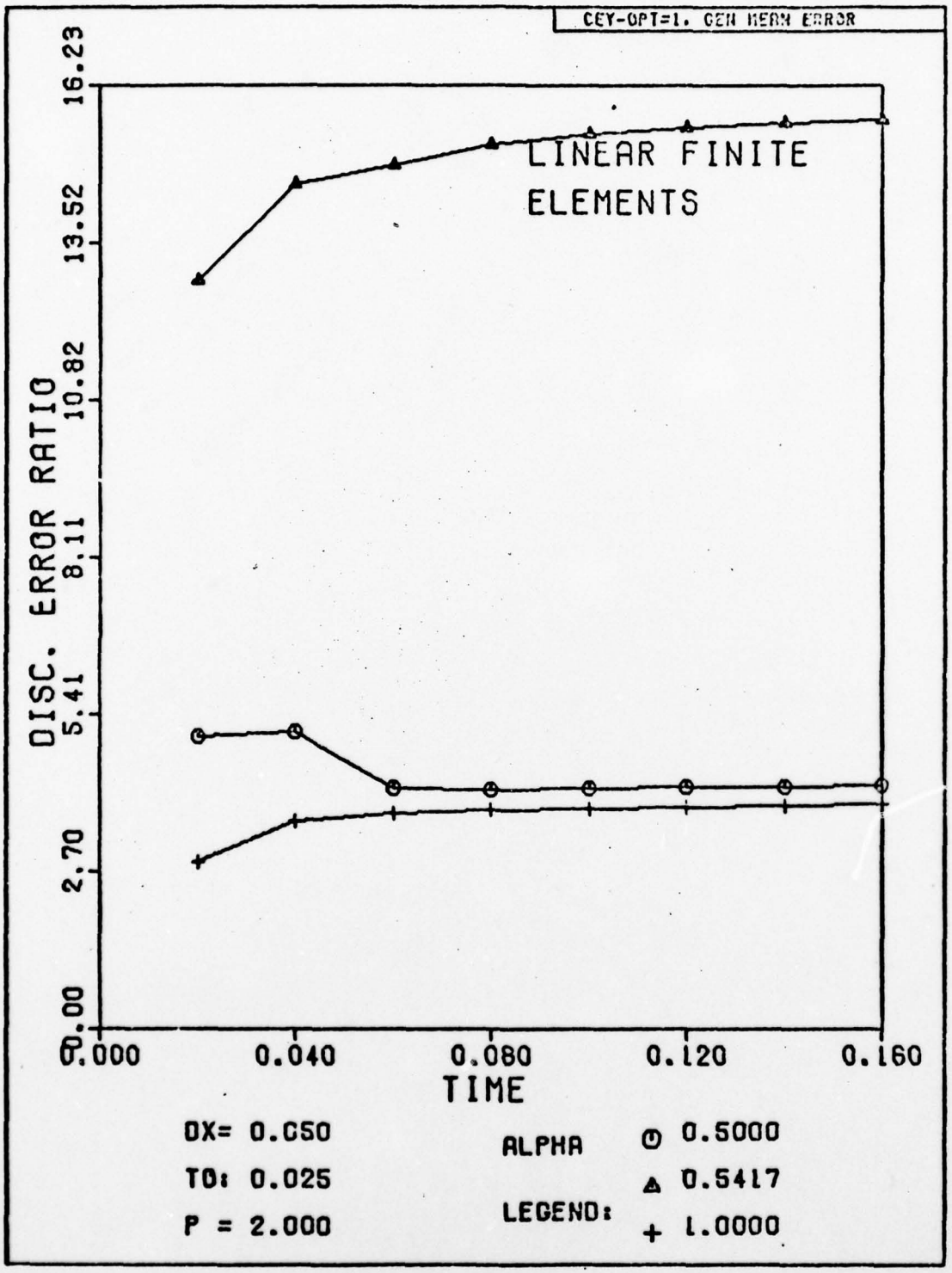

Fig. H-24. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

Section III

Results for the Problem Using Finite Elements and a Quadratic Interpolation Function

This section shows the graphical results for the solution of the problem by finite-elements, quadratic interpolation, with equivalent linear alpha values. Run identifiers are CET and CEY .

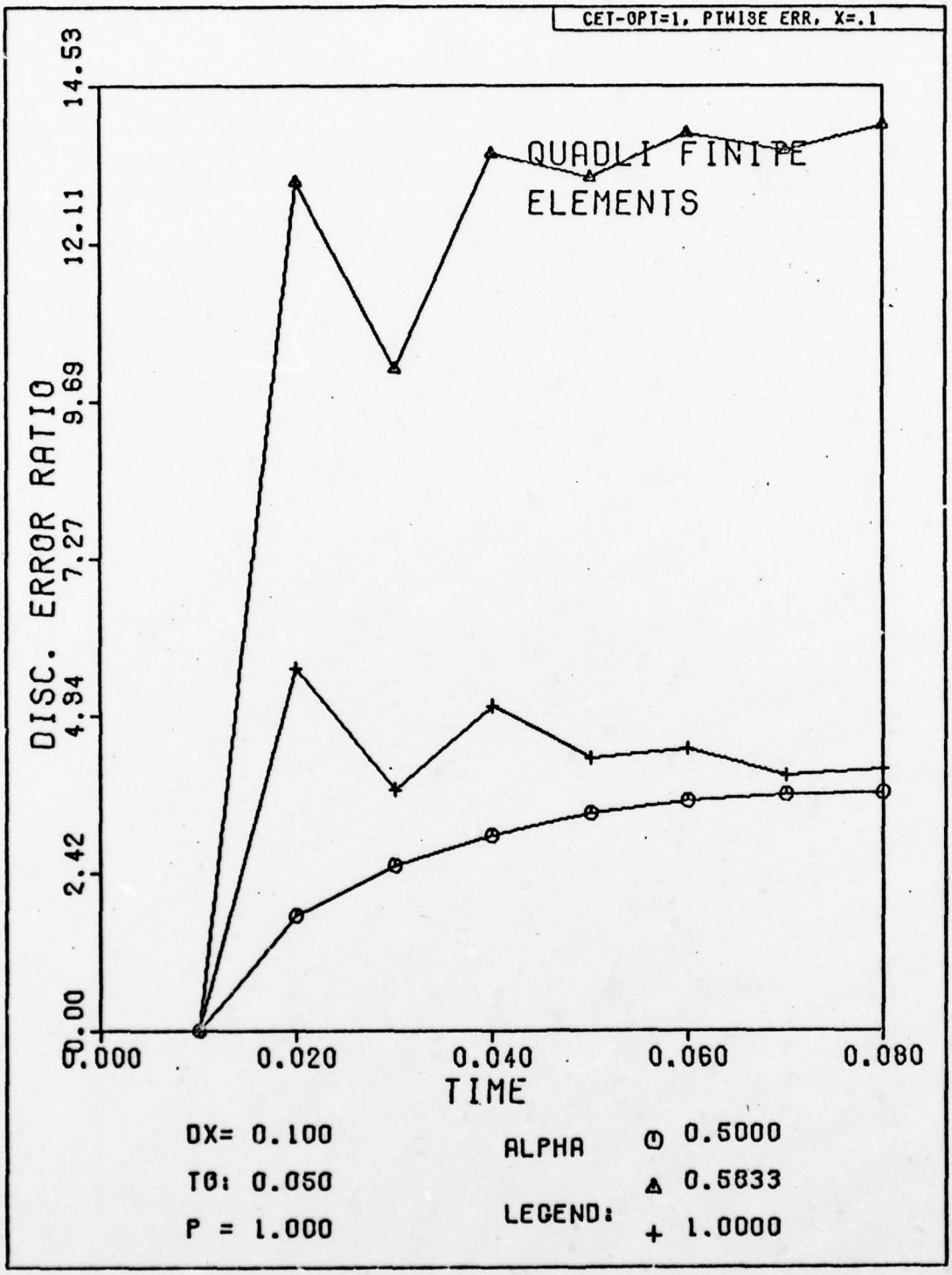

Fig. H-25. Discretization Error Ratio versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

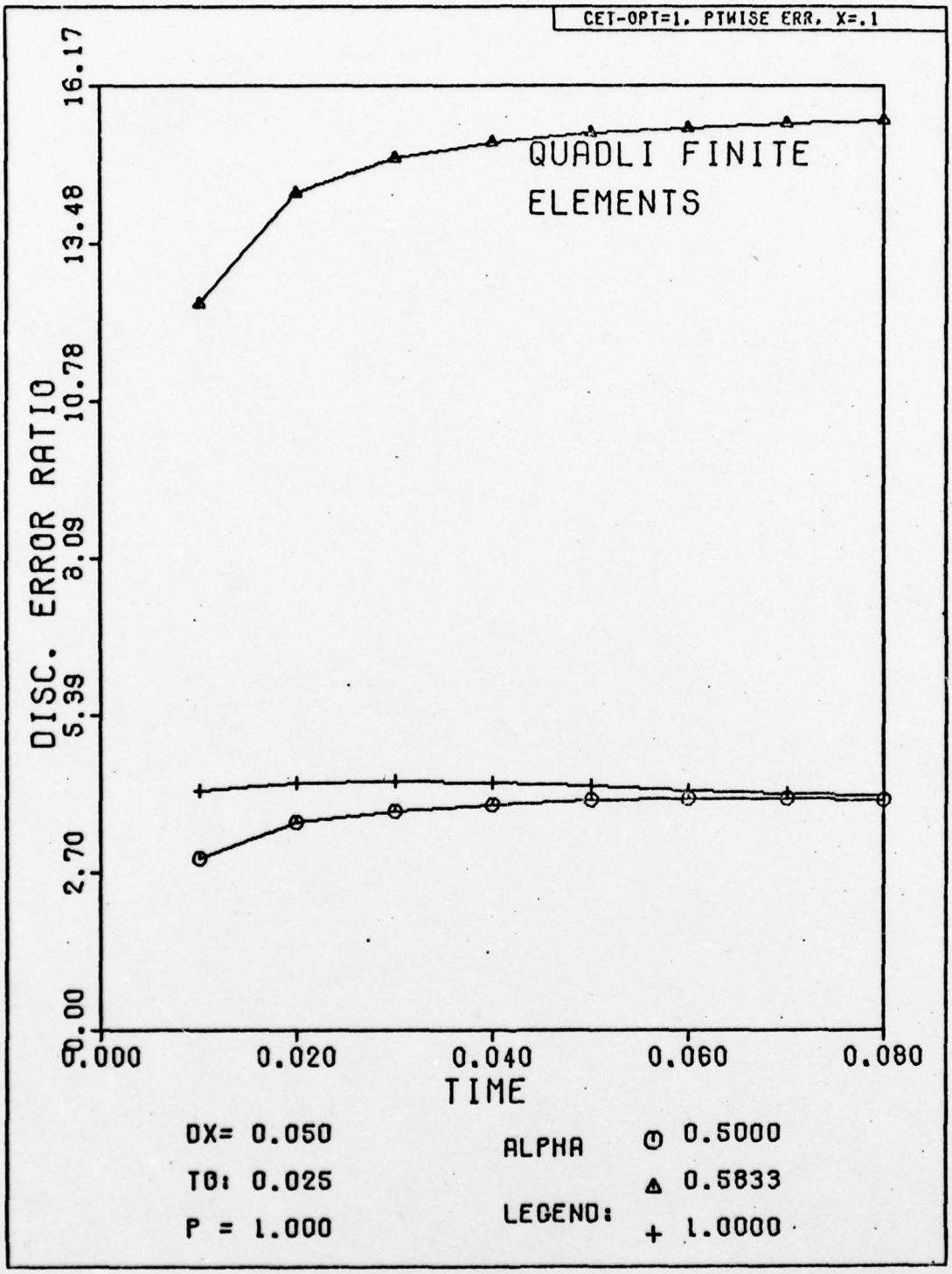

Fig. H-26. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

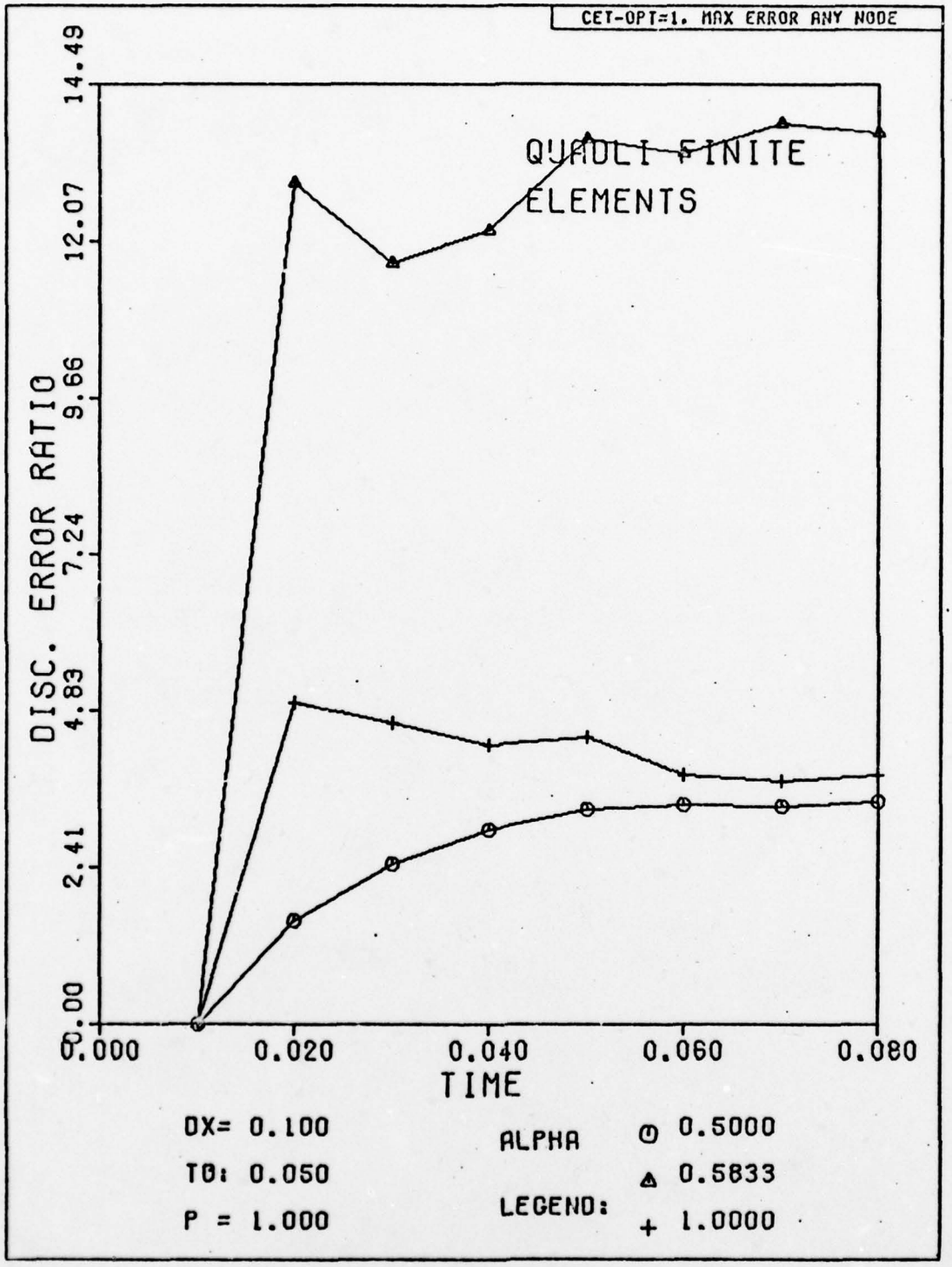

Fig. H-27. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

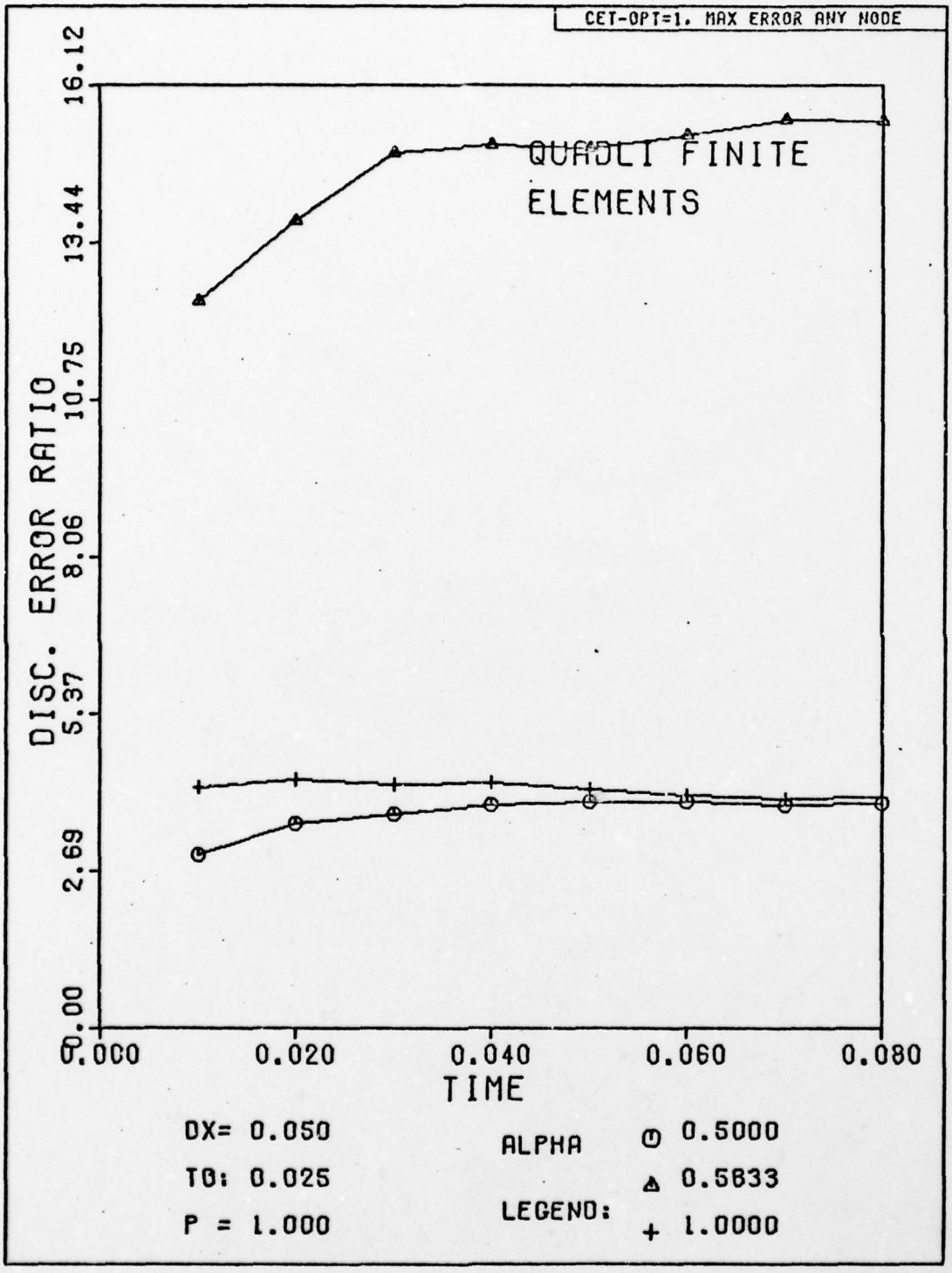

Fig. H-28. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

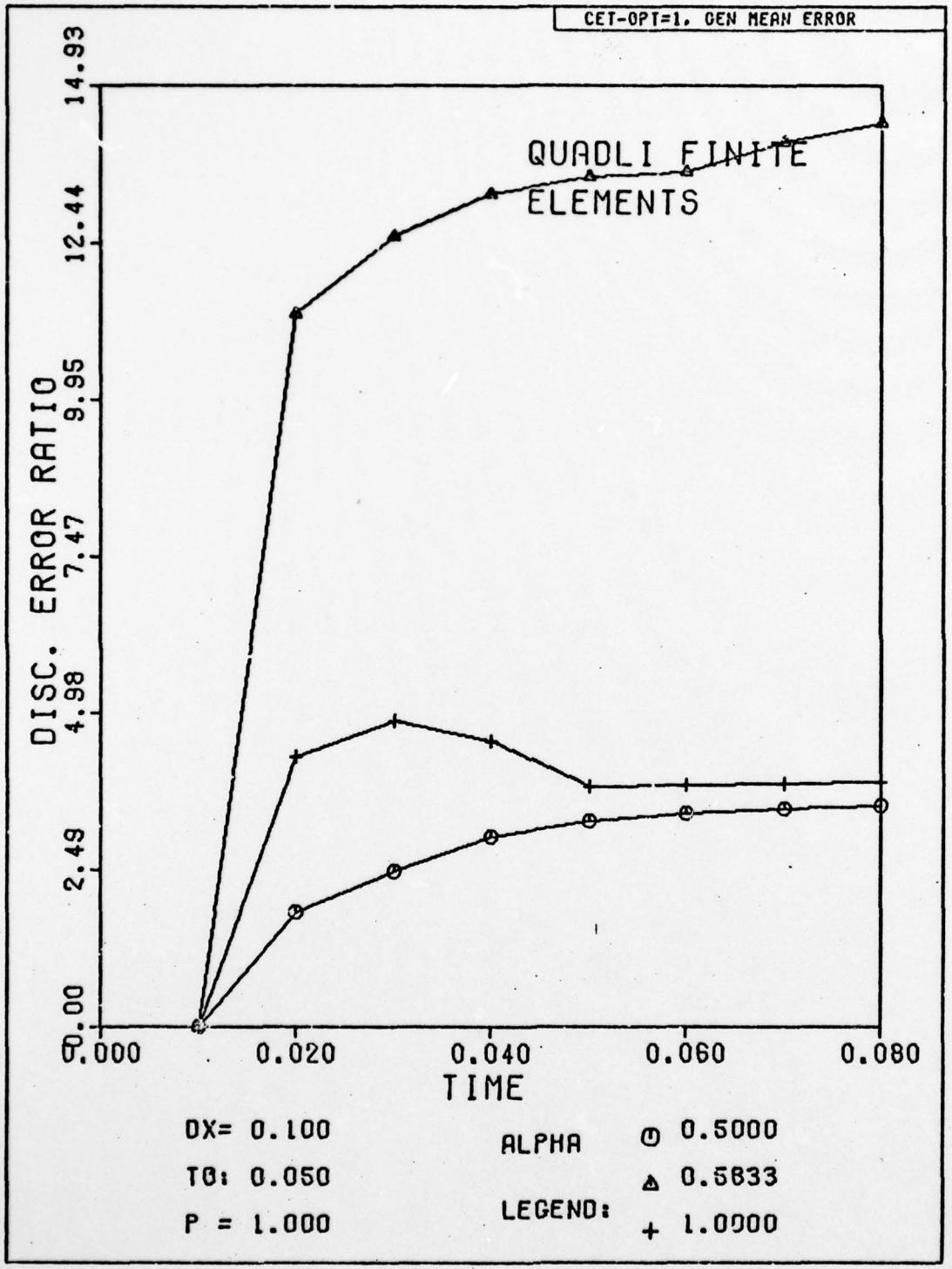

Fig. H-29. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

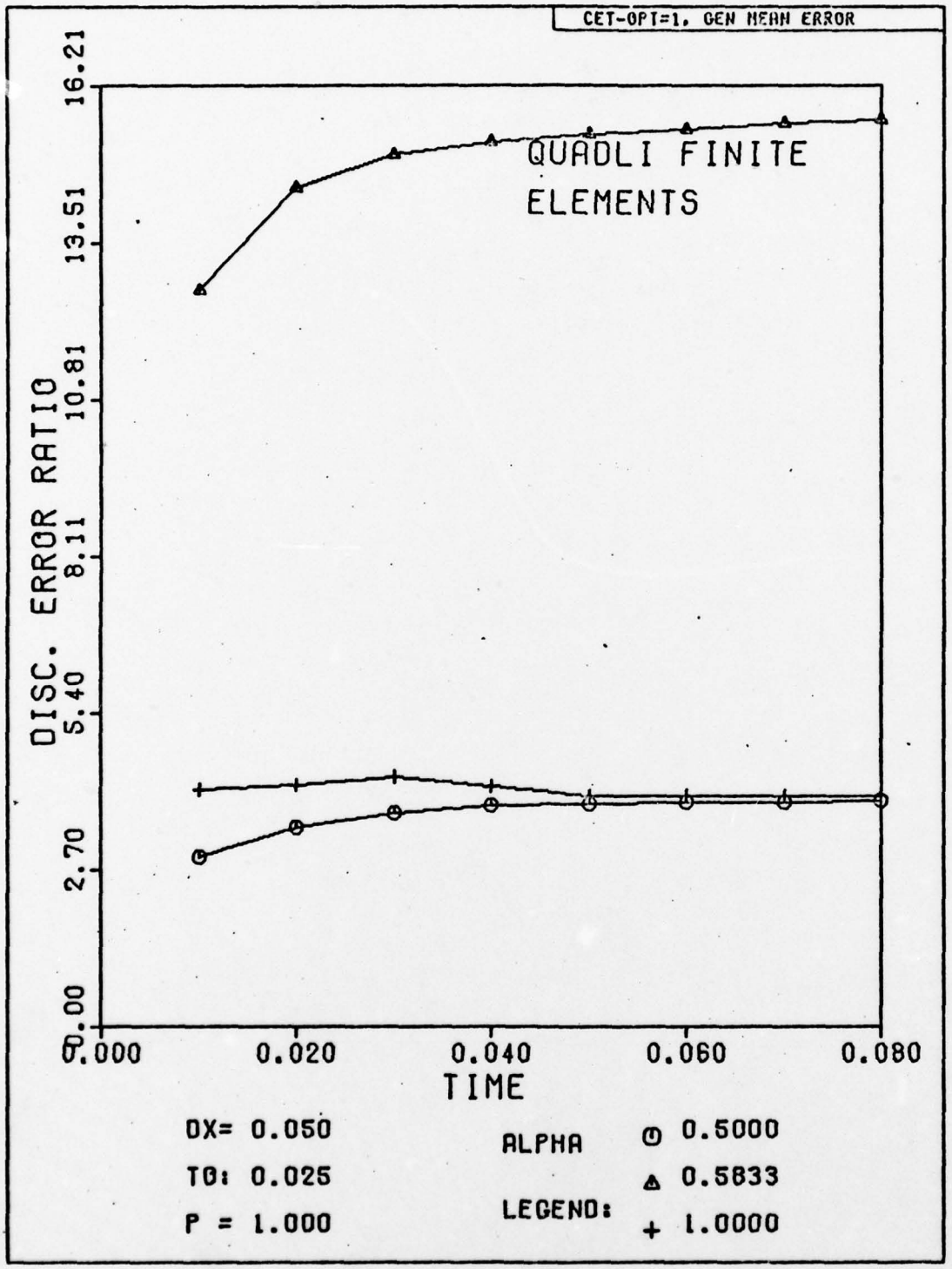

Fig. H-30. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

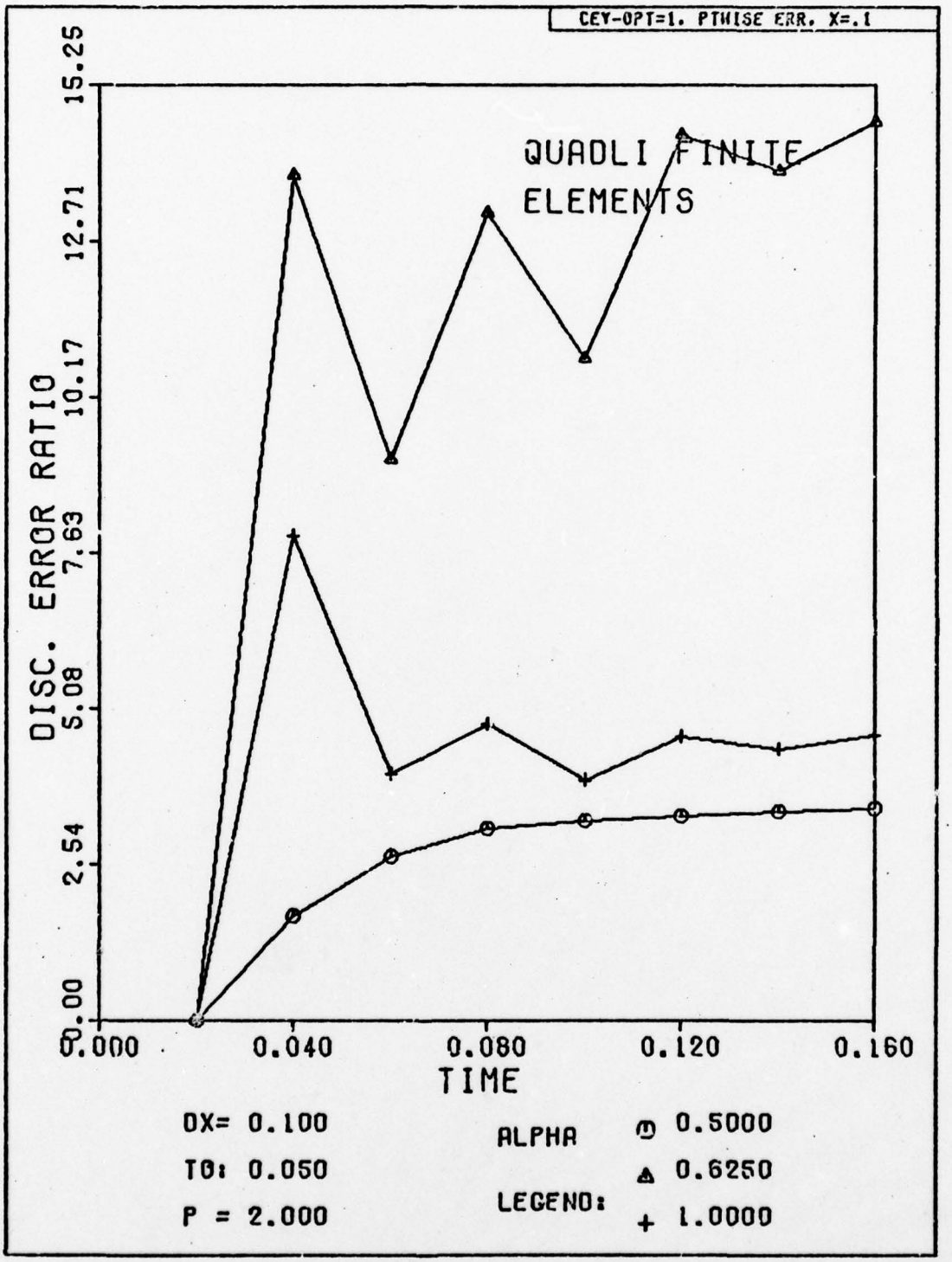

Fig. H-31. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

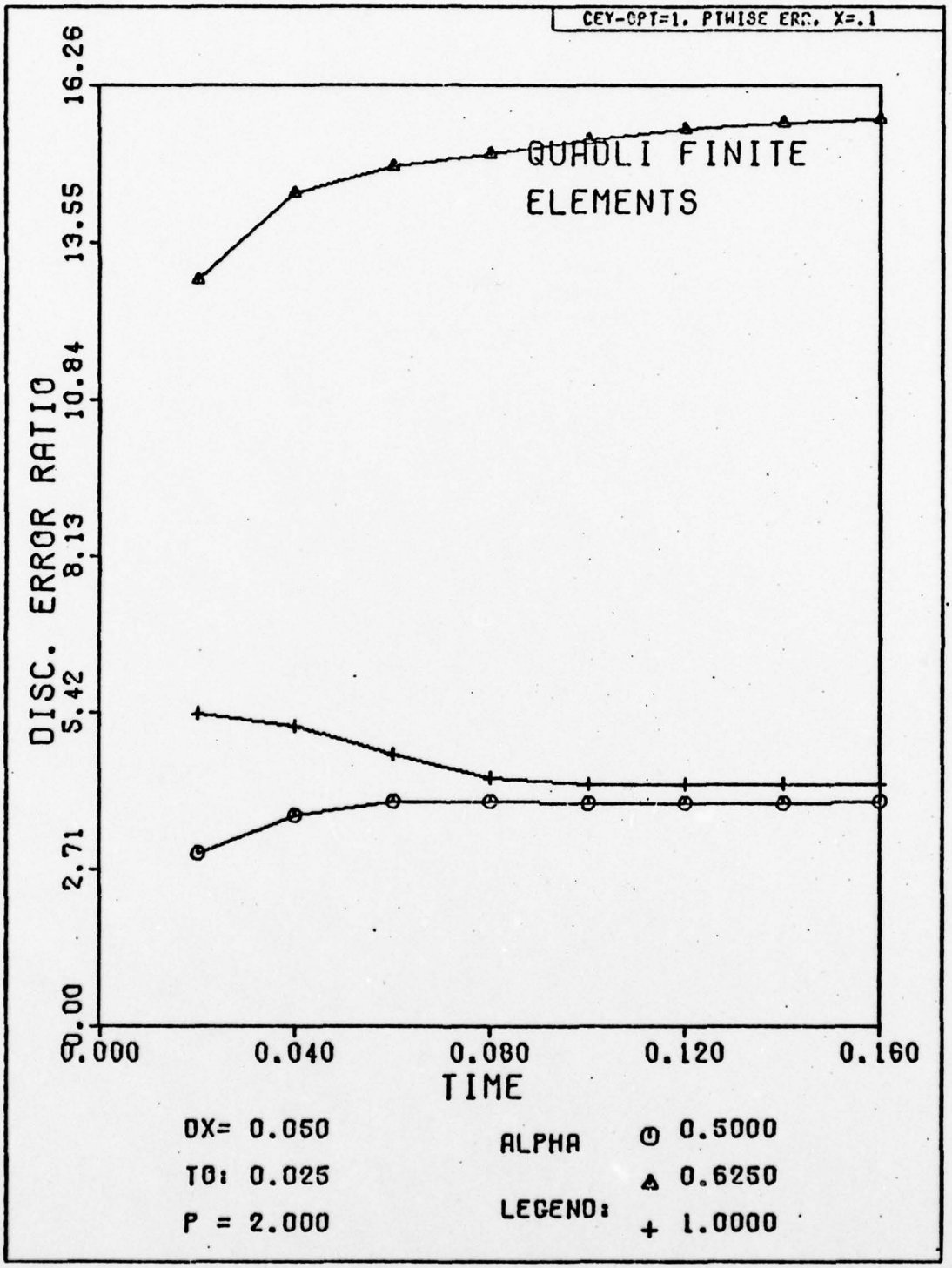

Fig. H-32. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

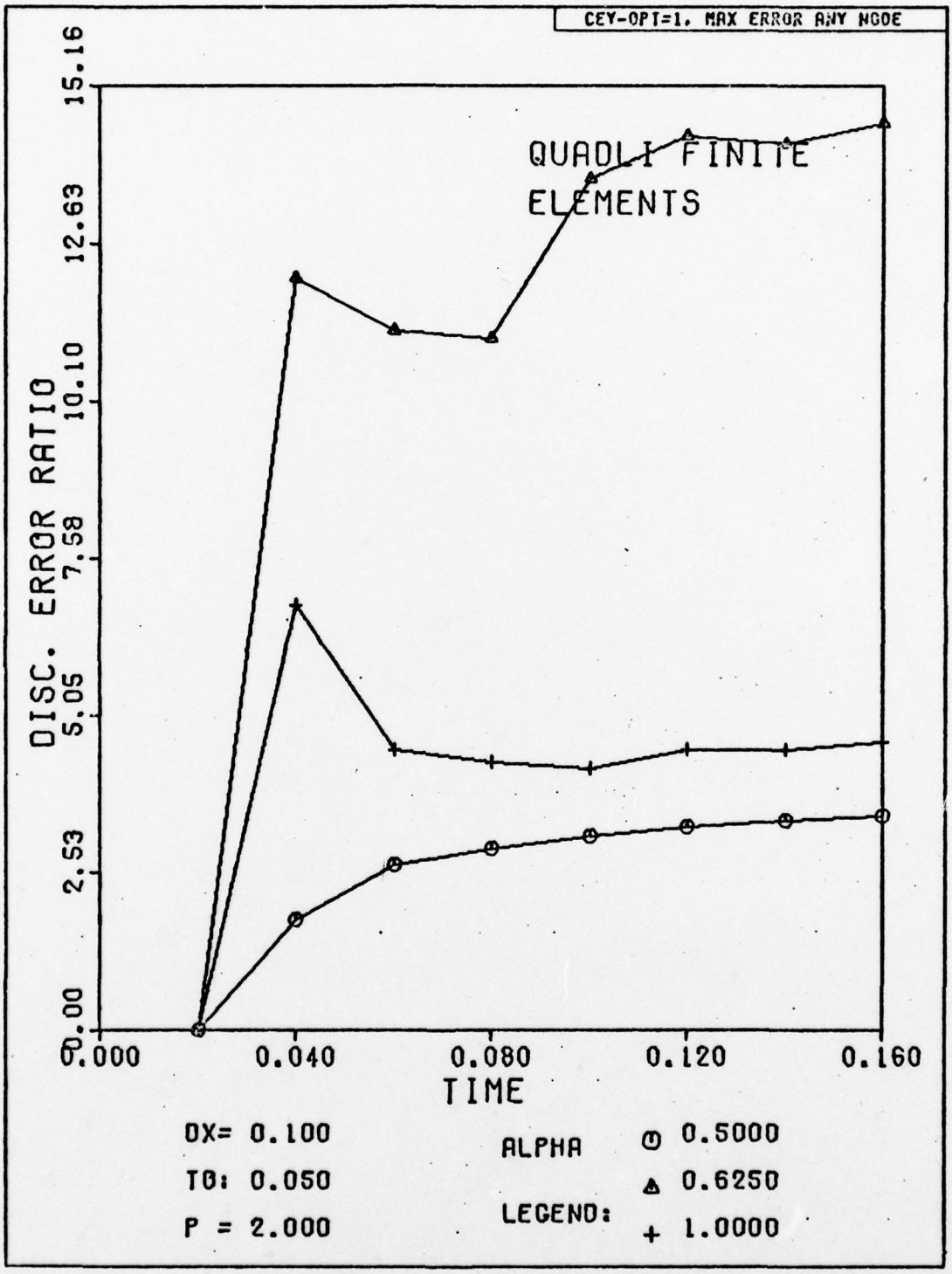

Fig. H-33. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

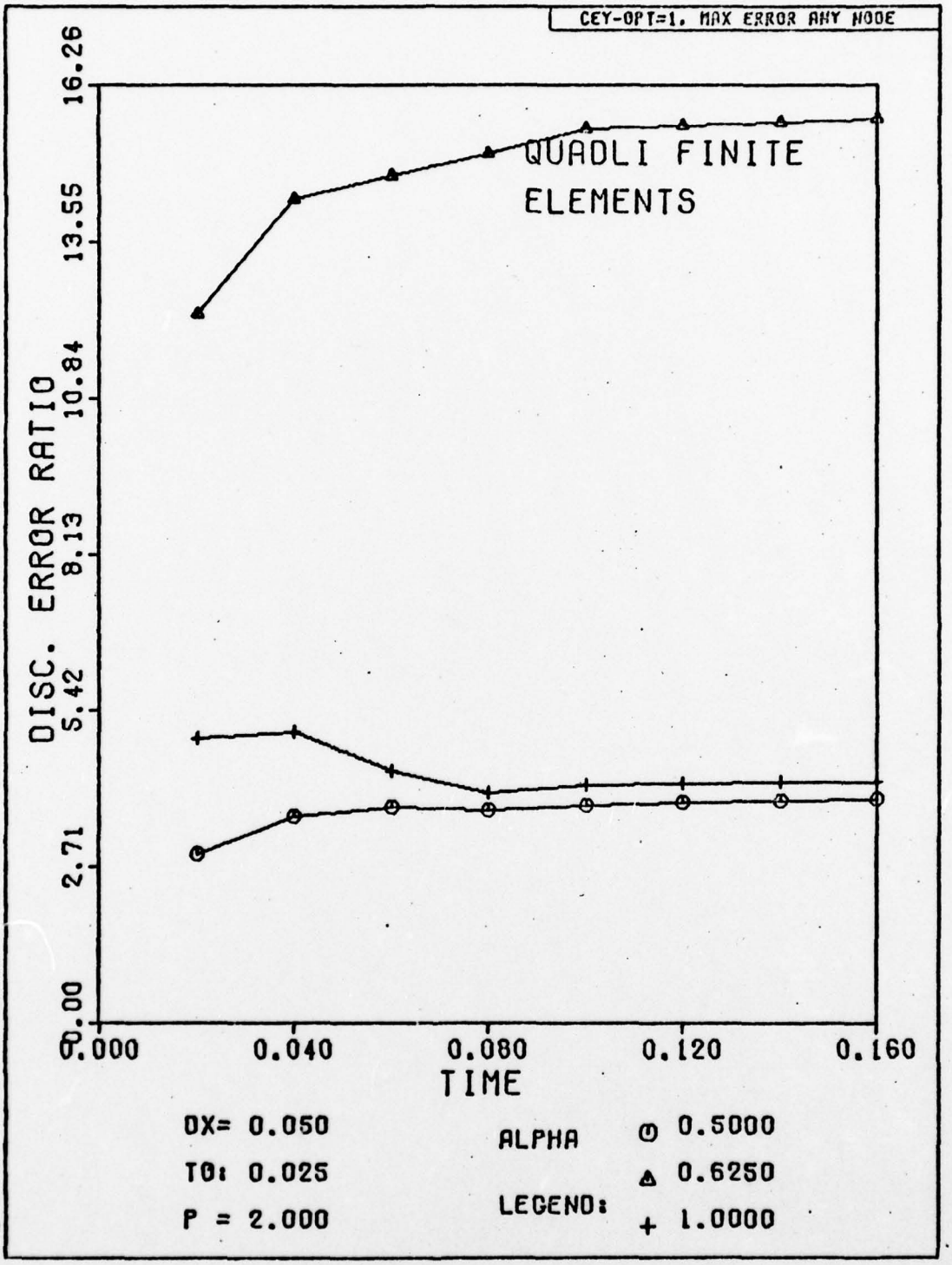

Fig. H-34. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

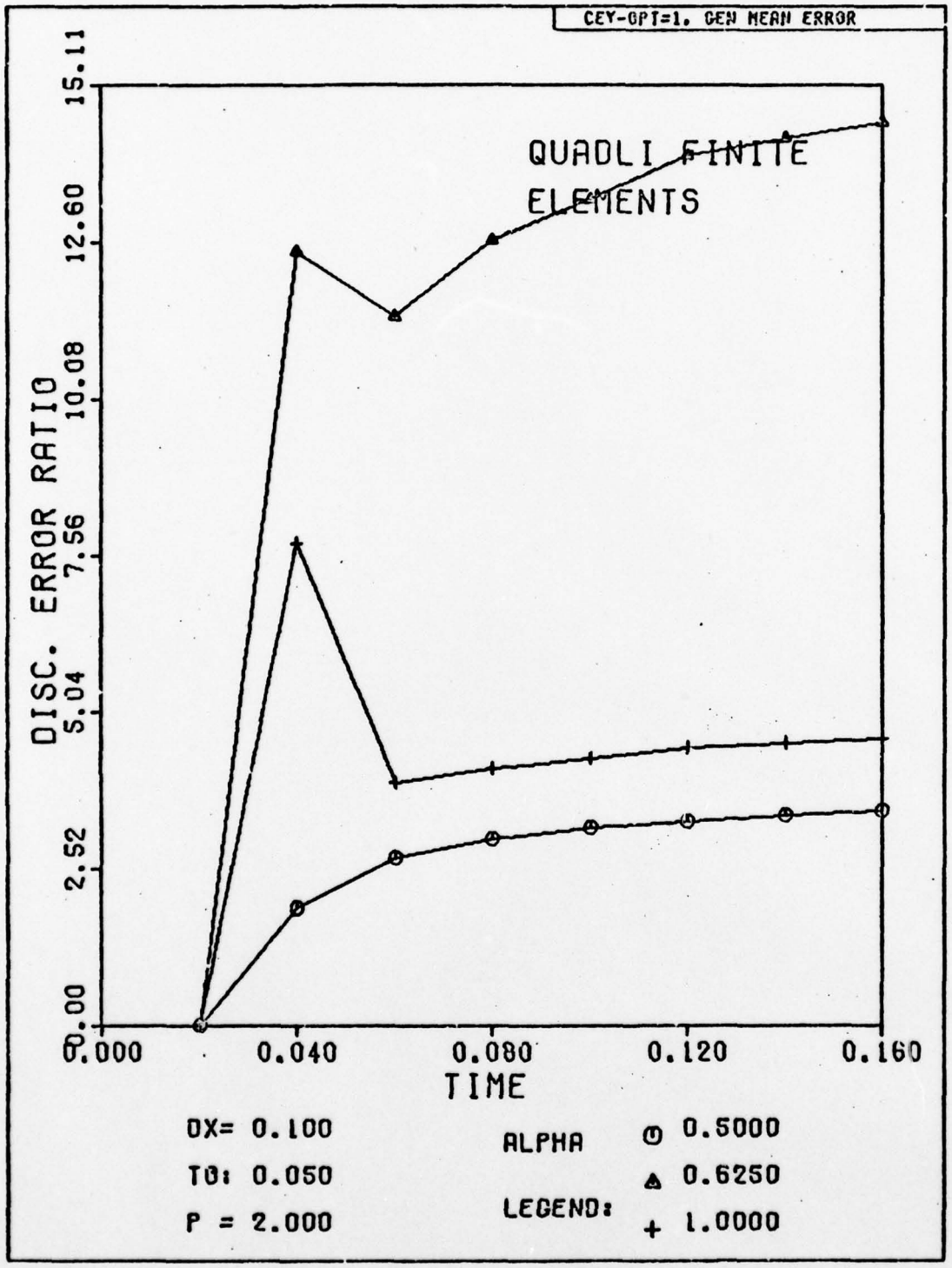

Fig. H-35. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

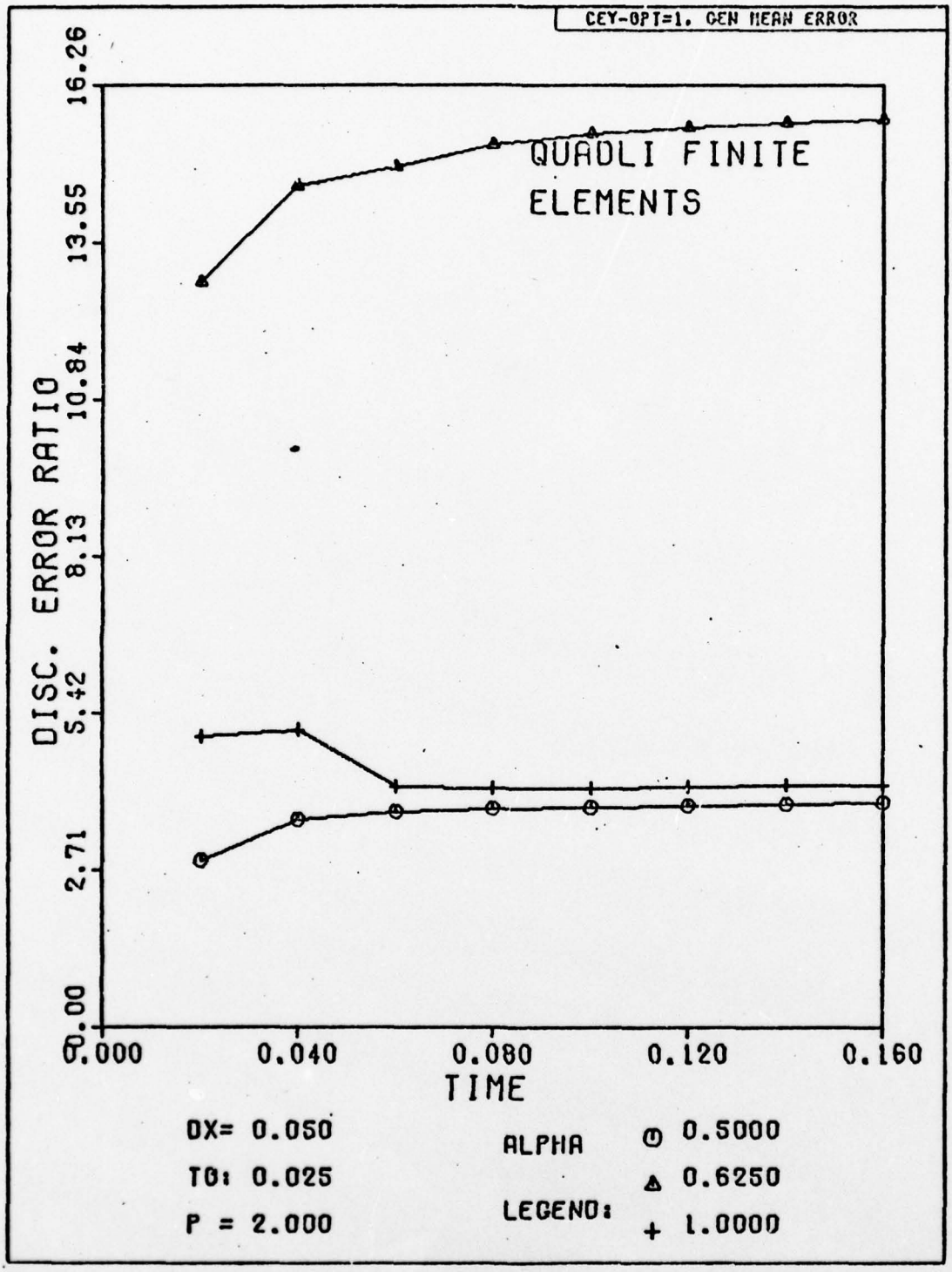

Fig. H-36. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

Section IV

Results for the Problem Using Finite Elements and a Quadratic Interpolation Function

This section shows the graphical results for the solution of the problem by finite-elements, quadratic interplation, direct application. Run identifier is CET .

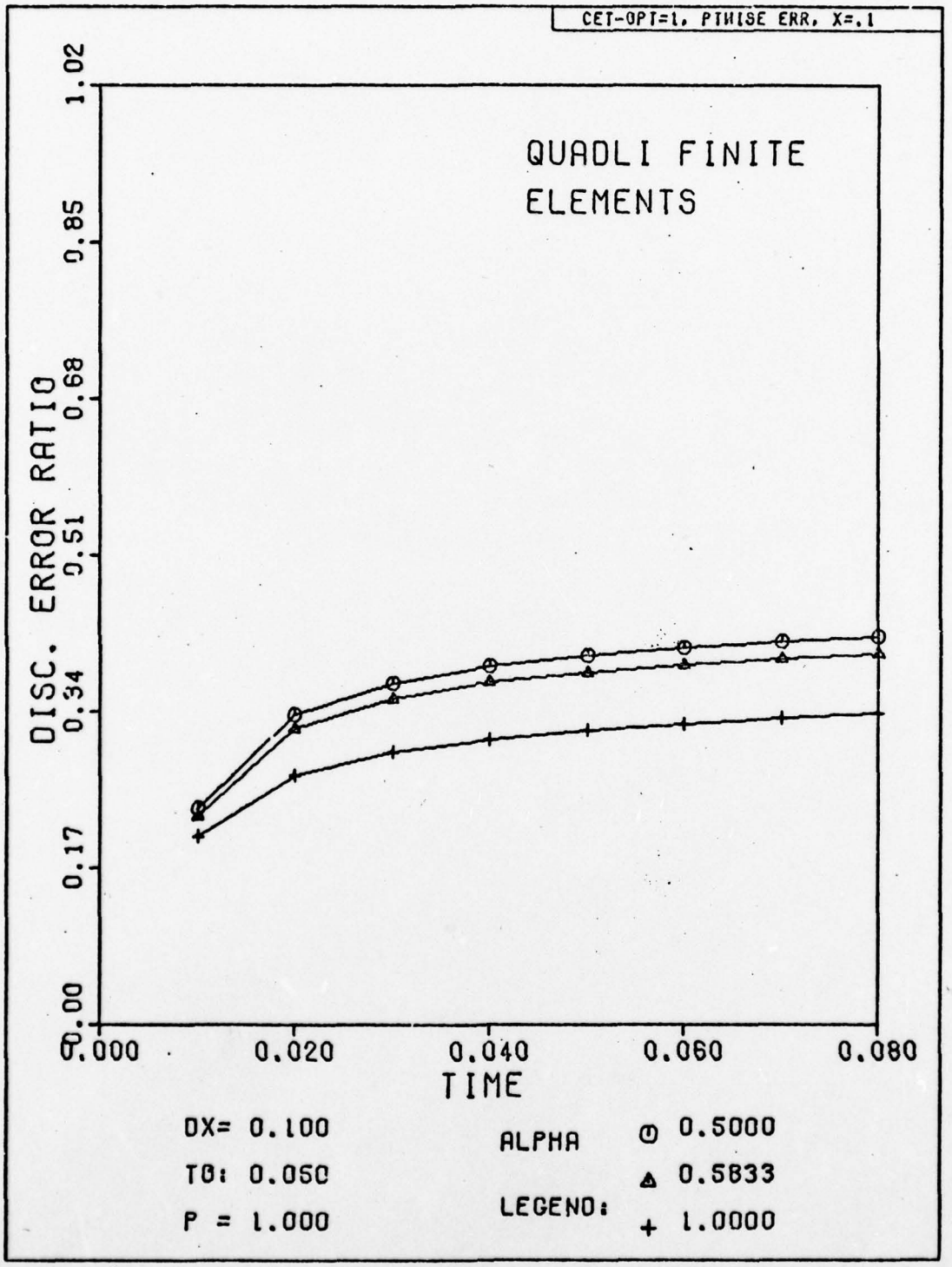

Fig. H-37. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

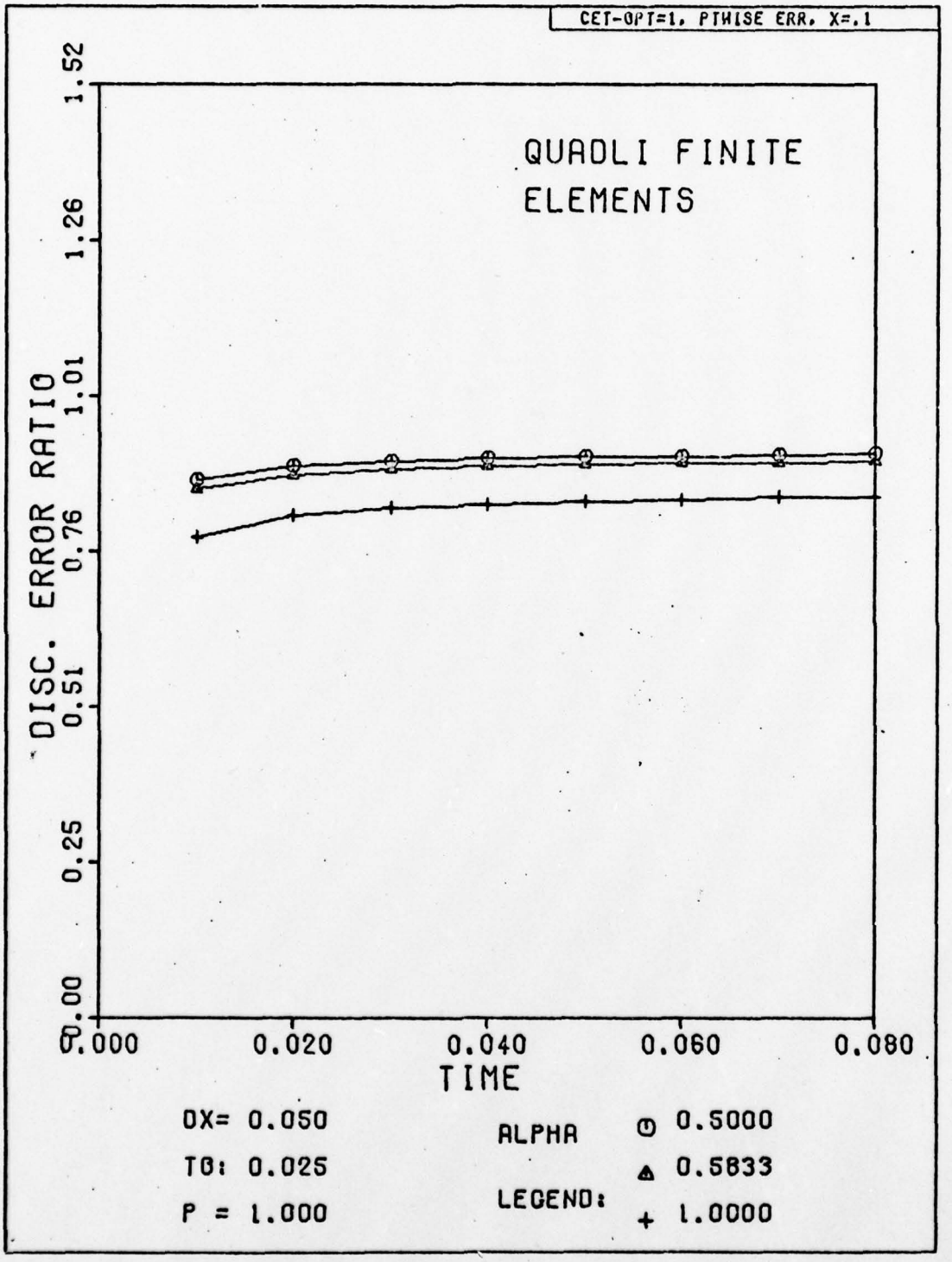

Fig. H-38. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

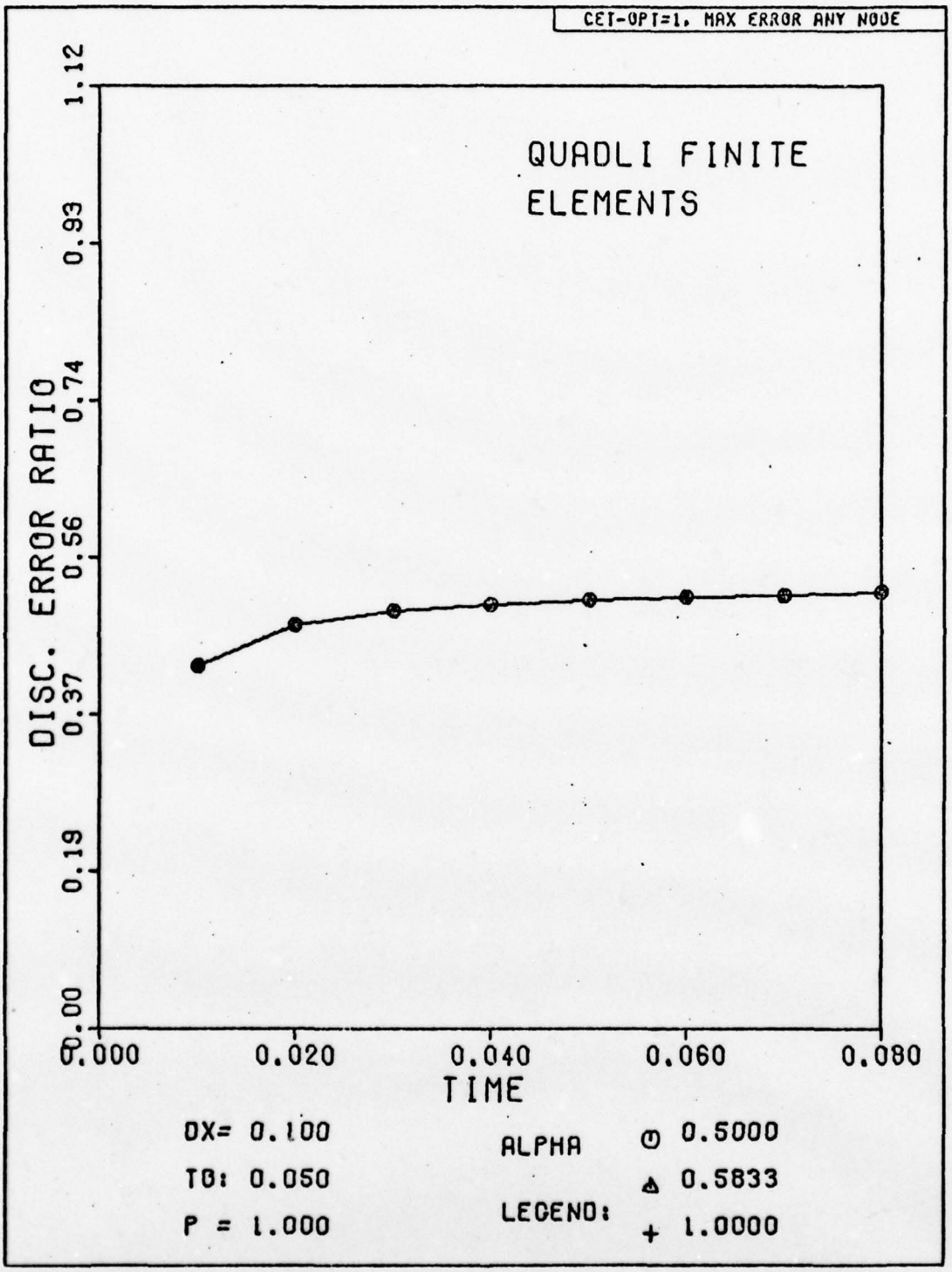

Fig. H-39. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

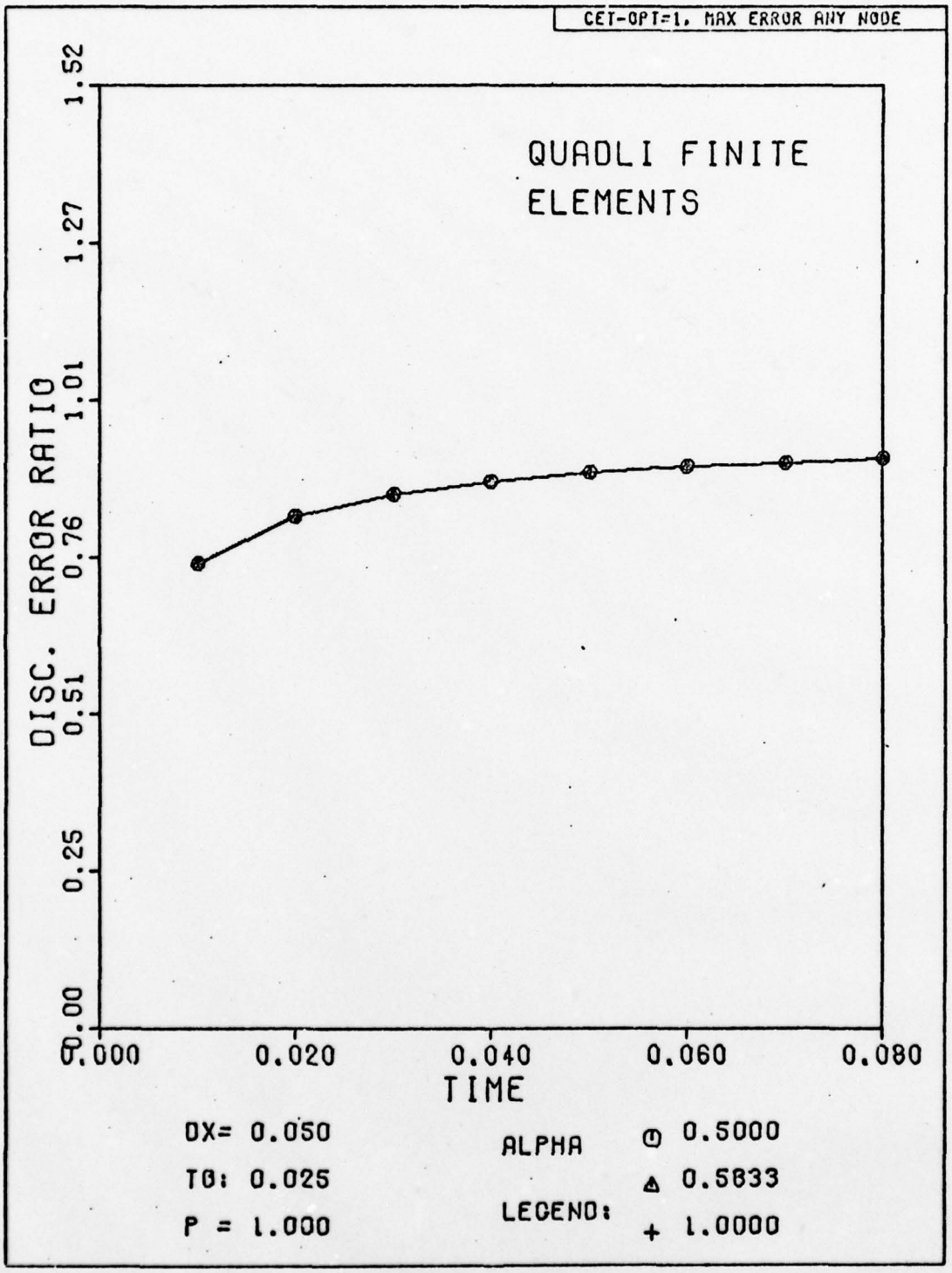

Fig. H-40. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

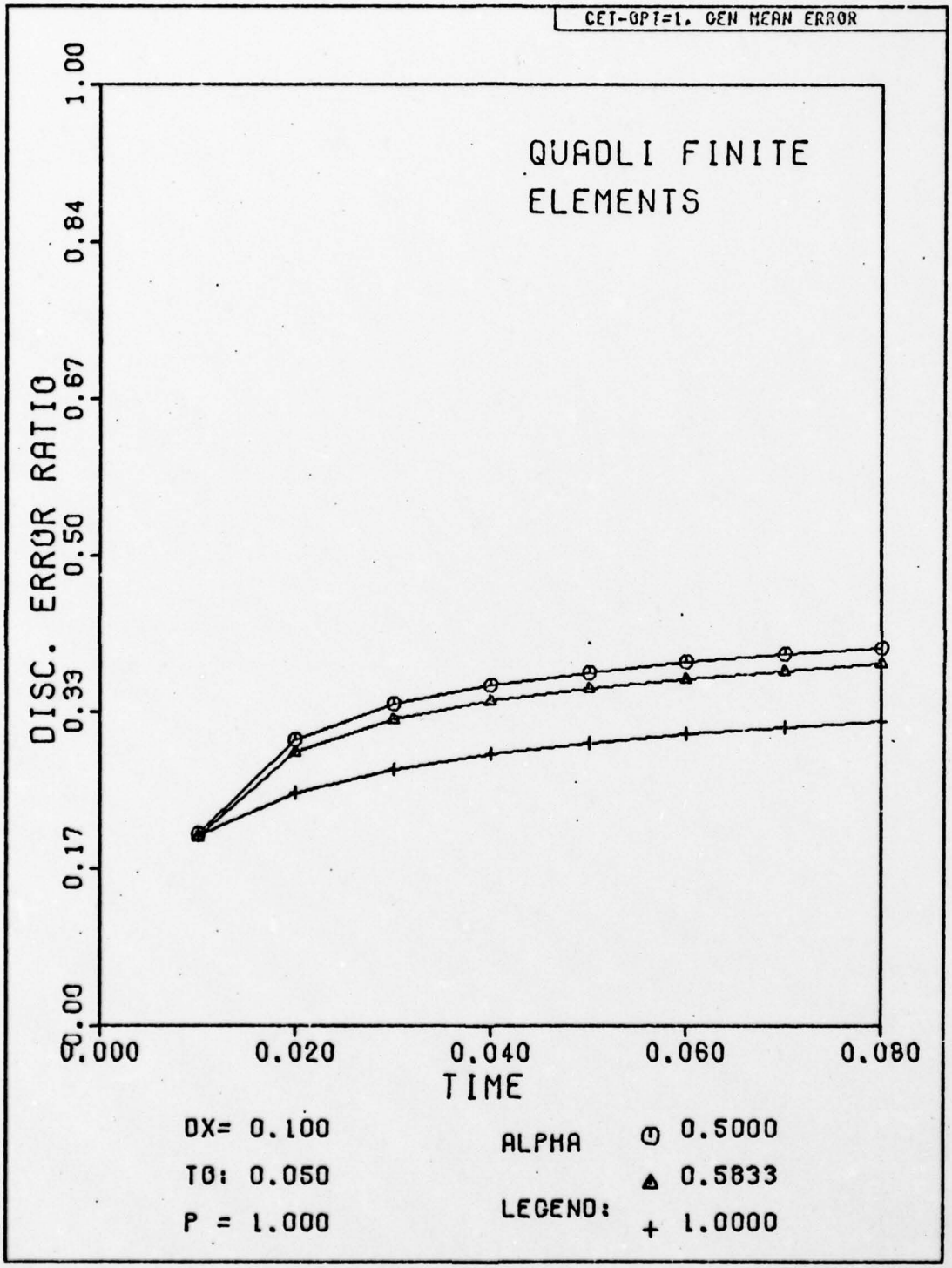

Fig. H-41. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

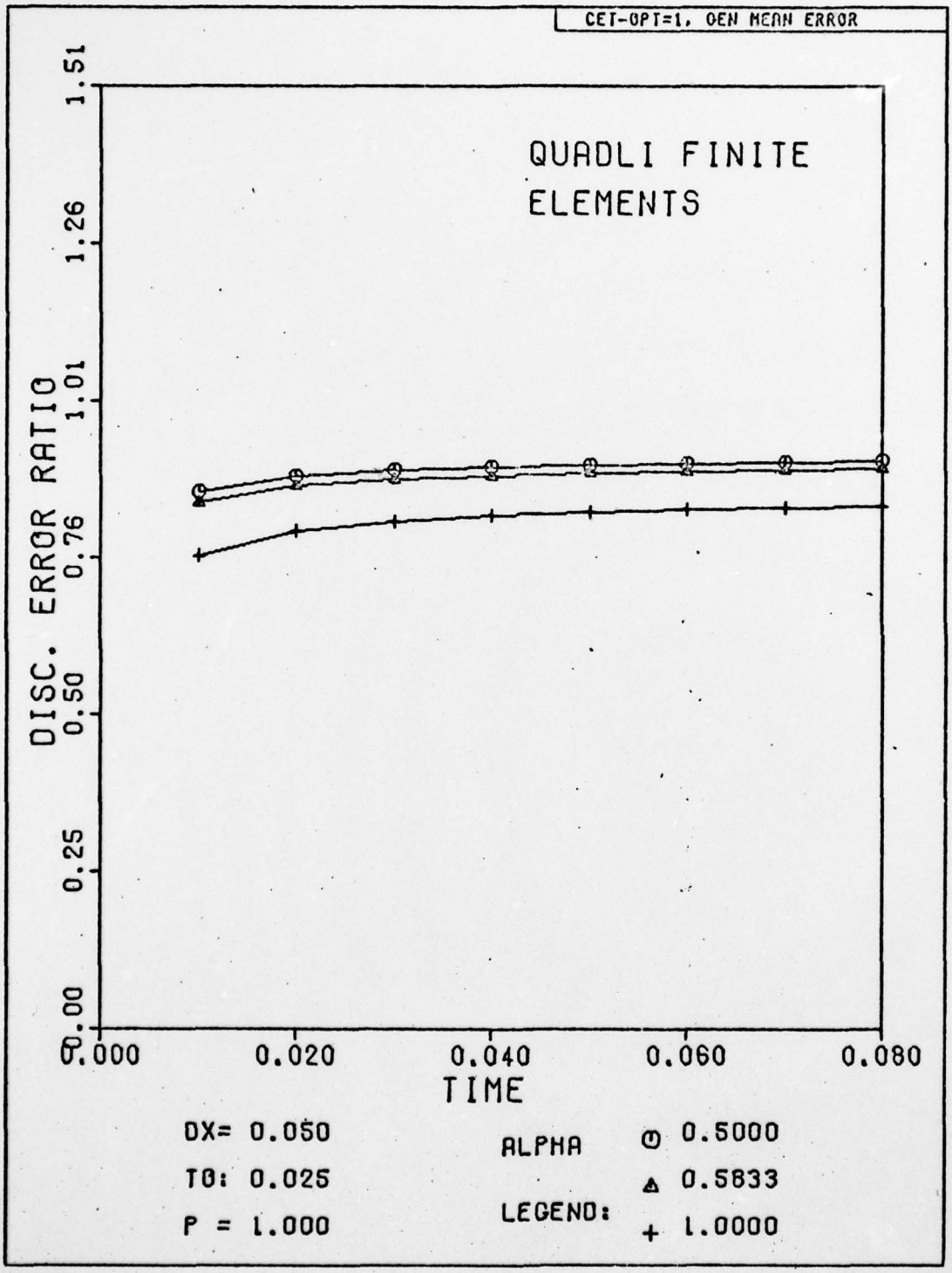

Fig. H-42. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

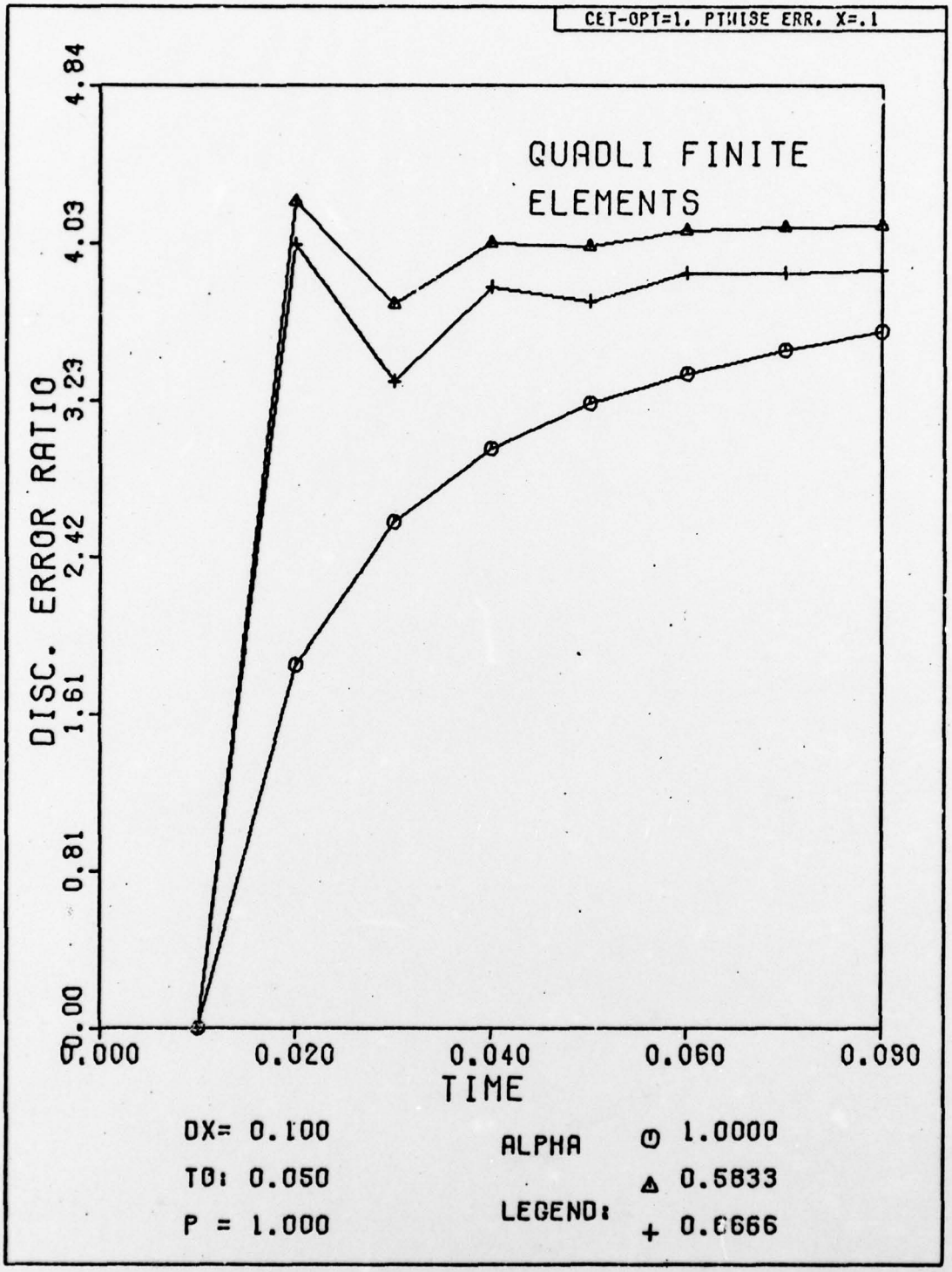

Fig. II-43. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

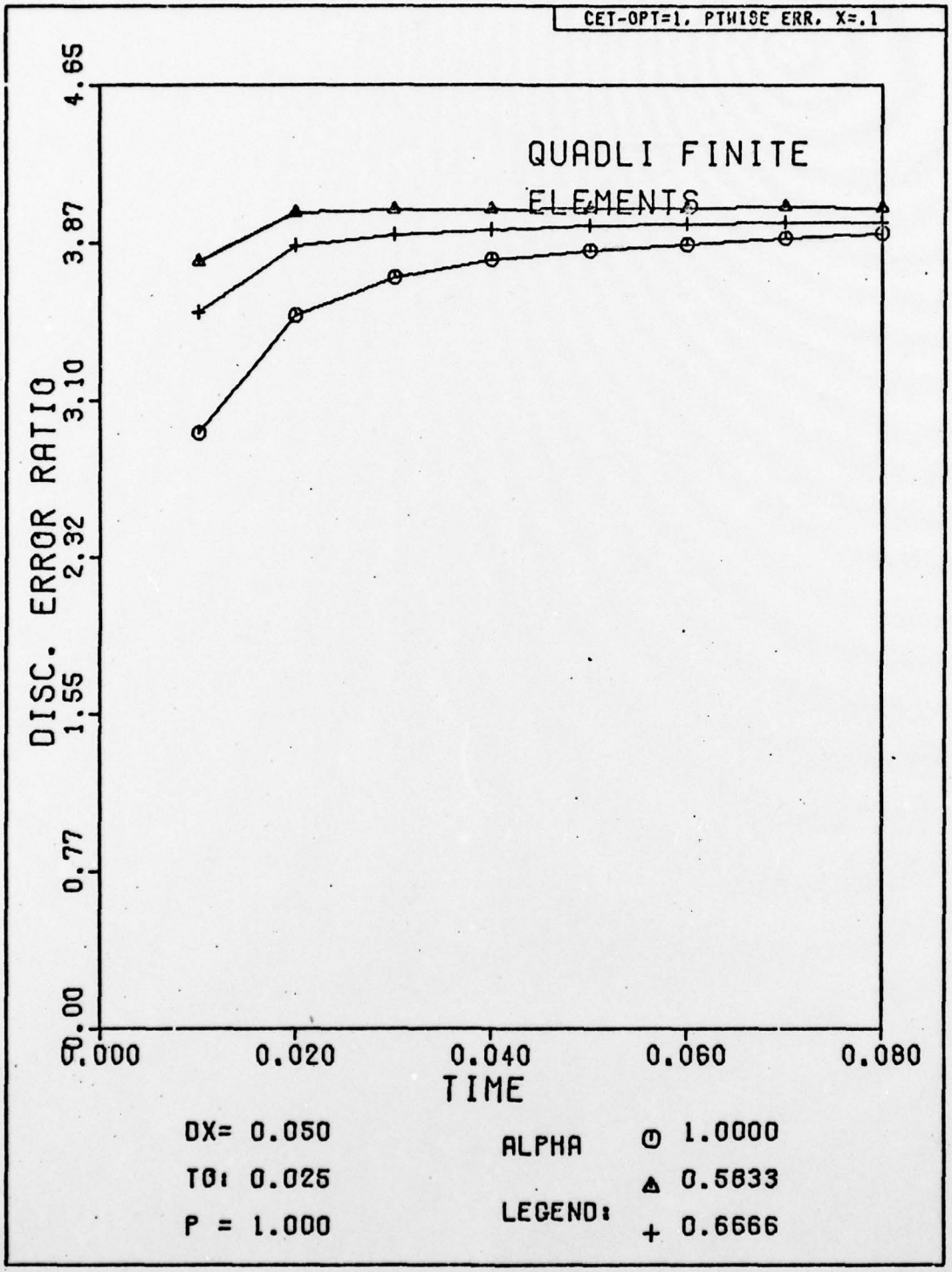

Fig. H-44. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

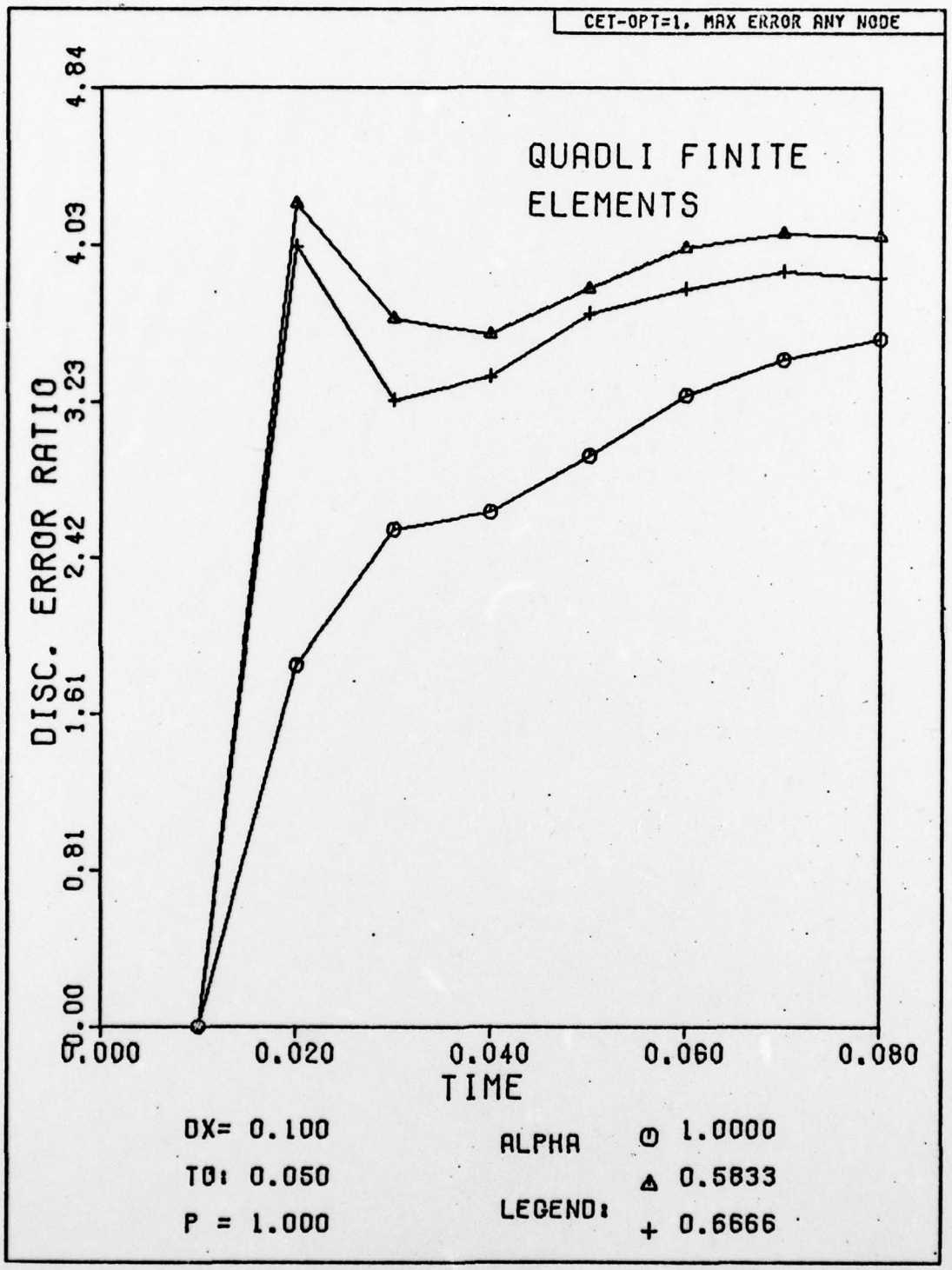

Fig. H-45. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

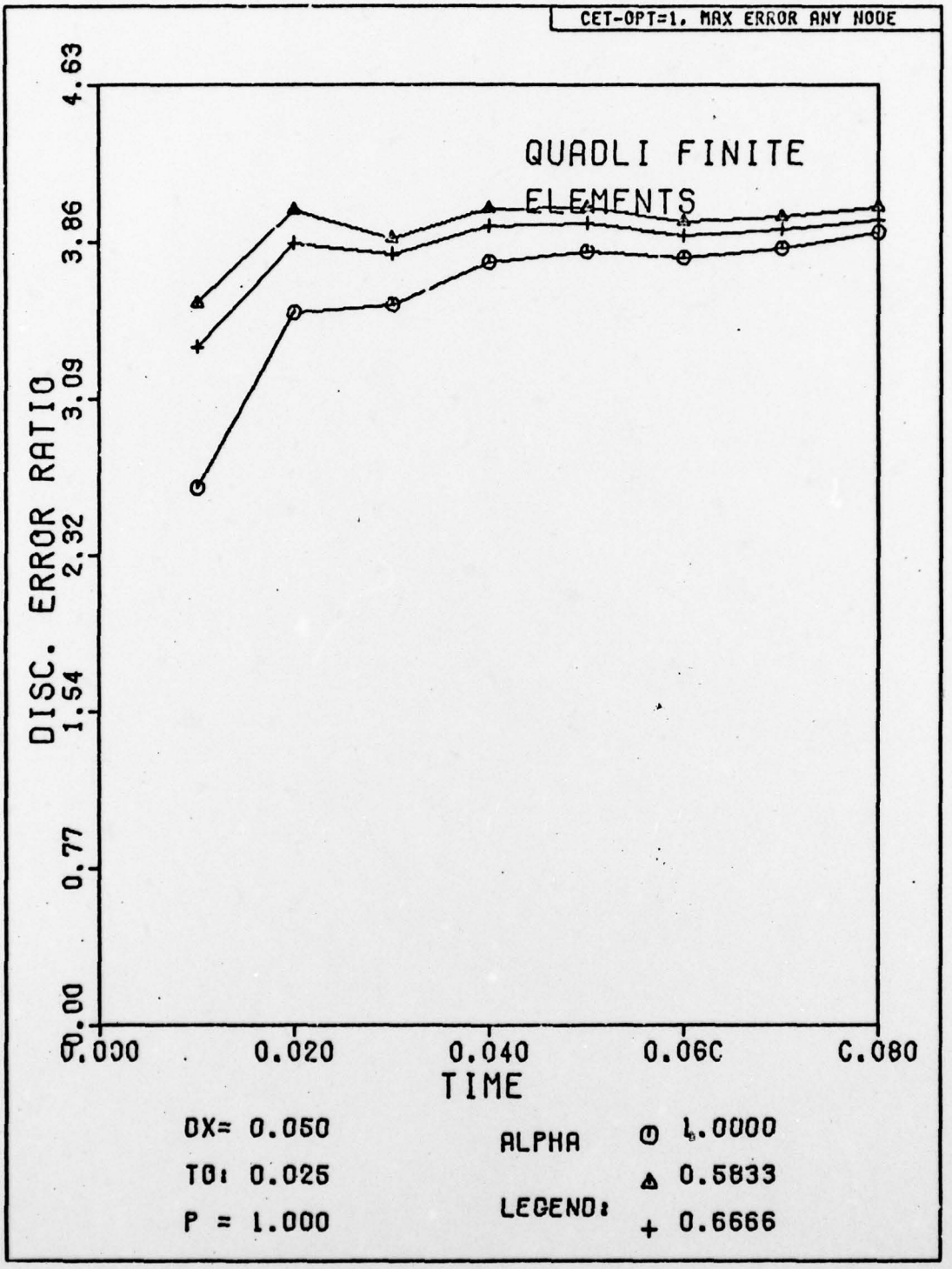

Fig. H-46. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

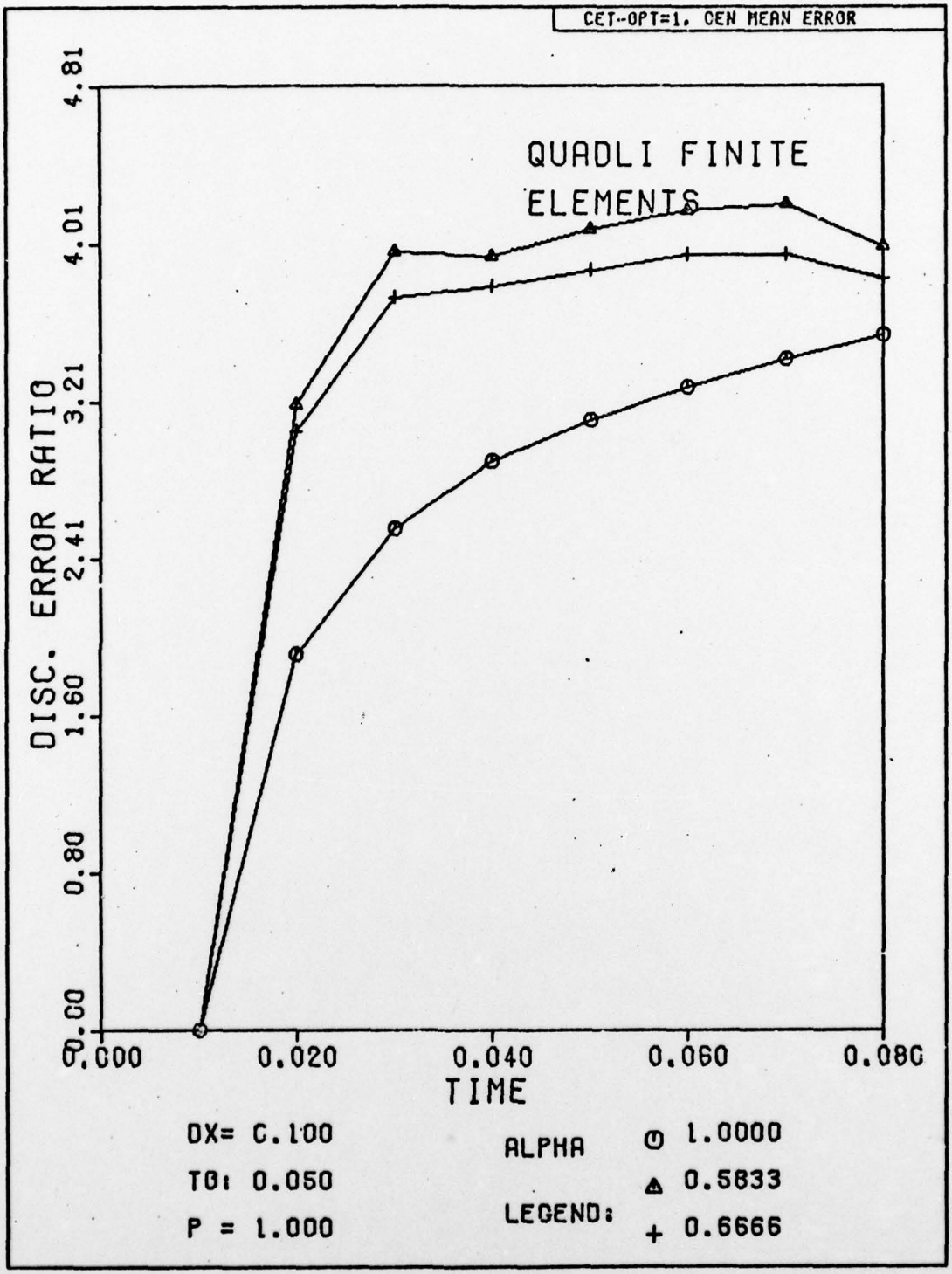

Fig. H-47. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

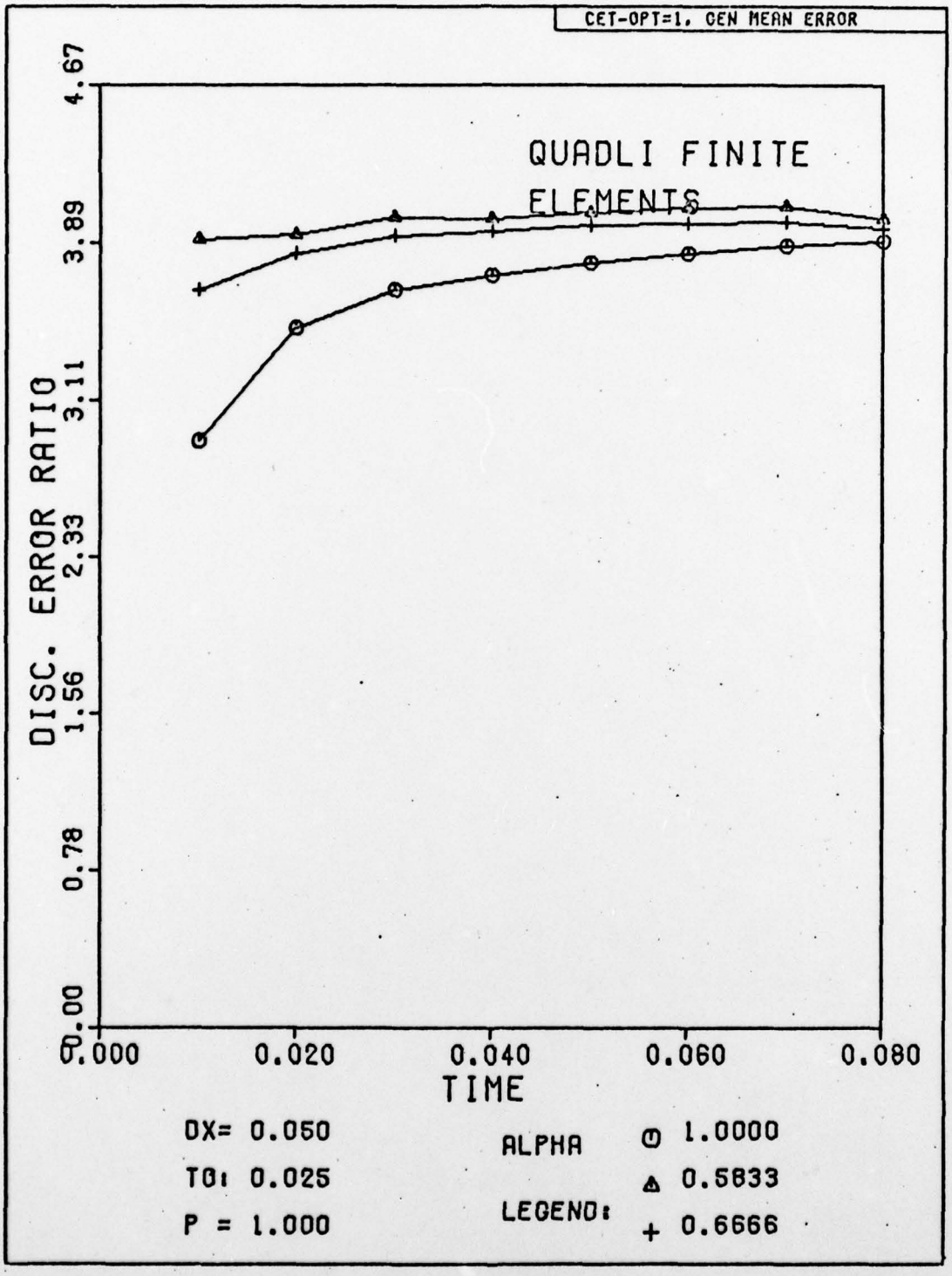

Fig. H-48. Discretization Error Ratio Versus Time for Selected Alphas. The exact solution has been substituted for the numerical solution at the first time step.

APPENDIX I

Alternative Formulation of the Time Response

The discussion leading to Equation (110) is concerned with the fact that the recurrence relation, Equation (97), of the finite-element formulation is based on finite-differencing the time variable, Equation (96). According to Zienkiewicz (Ref 13:334-336), a more stable if not more accurate finite-element solution could be obtained if the problem were discretized into finite-elements of time as well as space. This appendix constructs the recurrence relation based on this concept, and develops a solution equation. It is then shown that both recurrence relations are equal.

As before, the problem is an initial value problem with the initial normalized temperature defined as $\underline{u}_o$ at time, $\theta = 0$. The time interval goes from 0 to θ_n, where $\theta_n = \Delta\theta$. In analogy to Equation (36), an assumed interpolated form of $\underline{u}$ defined by its values at several time intervals may be written as

$$\underline{u} = \sum_{i=0}^{n} N_i(\theta)\underline{u}_i \qquad \text{(I-1)}$$

where $N_i(\theta)$ are appropriate shape functions or coefficients.

If a linear interpolation is employed, then only $n = 0$ and $n = 1$ need be considered. Therefore, in matrix form and following the procedures of Equation (49) and after

$$\{\underline{u}\} = [N_0, N_1] \begin{Bmatrix} \{\underline{u}\}_0 \\ \{\underline{u}\}_1 \end{Bmatrix} \tag{I-2}$$

where

$$N_0 = (\Delta\theta - \theta)/\Delta\theta \tag{I-3}$$

and

$$N_1 = \theta/\Delta\theta \tag{I-4}$$

Taking the time derivative yields

$$\frac{\partial\{\underline{u}\}}{\partial\theta} = \cdot\left[\frac{\partial N_0}{\partial\theta}, \frac{\partial N_1}{\partial\theta}\right] \begin{Bmatrix} \{\underline{u}\}_0 \\ \{\underline{u}\}_1 \end{Bmatrix} = \frac{1}{\Delta\theta}[-1,1] \begin{Bmatrix} \{\underline{u}\}_0 \\ \{\underline{u}\}_1 \end{Bmatrix} \tag{I-5}$$

If Equation (96) is multiplied by N_1 and integrated over time, the result is

$$\int_0^{\Delta\theta} \frac{\theta}{\Delta\theta} \left([\underline{K}][N_0, N_1] \begin{Bmatrix} \{\underline{u}\}_0 \\ \{\underline{u}\}_1 \end{Bmatrix} + [\underline{M}] \left[\frac{\partial N_0}{\partial\theta}, \frac{\partial N_1}{\partial\theta}\right] \begin{Bmatrix} \{\underline{u}\}_0 \\ \{\underline{u}\}_1 \end{Bmatrix} \right) d\theta = 0 \tag{I-6}$$

which, if Equations (I-2) and (I-5) are appropriately substituted, becomes

$$[\underline{K}]\left(\frac{1}{3}\{\underline{u}\}_0 + \frac{2}{3}\{\underline{u}\}_1\right) + \frac{1}{\Delta\theta}[\underline{M}]\left(-\{\underline{u}\}_0 + \{\underline{u}\}_1\right) \tag{I-7}$$

The solution, $\{\underline{u}\}_1$ is then found to be

$$\{\underline{u}\}_1 = \left(\frac{2}{3}[\underline{K}] + [\underline{M}] / \Delta\theta\right)^{-1}\left(\frac{1}{3}[\underline{K}] - [\underline{M}] / \Delta\theta\right)\{\underline{u}\}_0 \tag{I-8}$$

Equation (I-7) is the recurrence relation found by treating the time variable by finite-elements. It may be rewritten as

$$\left[\frac{1}{\Delta\theta}\ \underline{M} + \frac{2}{3}\underline{K}\right]\underline{u}_1^{(E)} = \left[\frac{1}{\Delta\theta}\ \underline{M} - \frac{1}{3}\underline{K}\right]\underline{u}_0^{(E)} \tag{I-9}$$

where the matrix brackets have been dropped for simplicity. This equation is similar to Equation (97) and may be written

$$\underline{A}'\,(\underline{u}^{(E)})^{k+1} = \underline{B}'\,(\underline{u}^{(E)})^{k} \tag{I-10}$$

in analogy to Equation (100), where

$$\underline{A}' = \underline{M} + \underline{K}\,(.6666)\,\Delta\theta \tag{I-11}$$

and

$$\underline{B}' = \underline{M} - \underline{K}\,(.3333)\,\Delta\theta \tag{I-12}$$

Therefore, Equation (I-10) may be written

$$(\underline{M} + \underline{K}\,(.6666)\Delta\theta)\;(\underline{u}^{(E)})^{k+1} = (\underline{M} - \underline{K}\,(.3333)\Delta\theta)\;(\underline{u}^{(E)})^{k} \qquad \text{(I-13)}$$

It is noted that Equation (I-13) equals Equation (97) if $\alpha = .6666$. In both the general linear and quadratic formulations, this value of alpha yields at most only second order accuracy. The one point of variation is for the Fourier modulus equal to $.5$; in this case, the linear alpha of $.6666$, equivalent to the quadratic alpha of $.5$, yields fourth order accuracy.

The important results of the derivations leading to Equation (I-13) are as follows:

(1) The determination of the optimum alpha value is independent of the treatment of time.

(2) The recurrence relation by quadratic interpolation equals the recurrence relation by linear interpolation, and both are inherently second order accurate in the general scheme of alpha values.

These findings were verified by Köhler and Pittr (Ref 6:625-630), who showed that even if a quadratic, parabolic, time interpolation was used, that is,

$$\{\underline{u}\} = [N_0(\theta), N_1(\theta), N_2(\theta)] \begin{Bmatrix} \{\underline{u}\}_0 \\ \{\underline{u}\}_1 \\ \{\underline{u}\}_2 \end{Bmatrix} \qquad \text{(I-14)}$$

no improvement over a linear time element was attained.

These results are significant if considering the idea of employing a high order Padé rational approximation to describe the temporal behavior of the solution. Varga (Ref 12:262-268) points out that the Crank-Nicolson method, as well as the forward difference and backward difference methods, are in fact, such approximations. The derivation proceeding from Equation (I-5), however, states that as long as the time domain is handled as in this variational approach, the previously achieved accuracy order will not be exceeded, which is logical, since the Padé approximation is merely a rational analog to its Taylor polynomial. This, however, does not preclude the more rapid achievement of that accuracy, inherent in operations by Padé approximation; that is, convergence will occur more quickly.

Vita

Joseph C. Geneczko was born on 7 August 1949 in Avoca, Pennsylvania. He graduated from Pittston Area High School, Duryea, Pennsylvania, in June 1967 and entered the University of Scranton shortly thereafter. In May 1971 he received a Bachelor of Science degree, cum laude, in physics and entered the Air Force Officer Training School at San Antonio, Texas. He received his commission, as a Distinguished Graduate, in September of that year and proceeded to Undergraduate Pilot Training at Laredo Air Force Base, Texas. He accomplished this program as a Distinguished Graduate and received the Academic Trophy, graduating in November 1972. For the next four and one half years, he served as pilot, instructor pilot, ground training officer, and flying safety officer at Langley Air Force Base, Virginia and at McChord Air Force Base, Washington. In June 1977 he entered the Air Force Institute of Technology, Dayton, Ohio.

Permanent Address: 263 Gedding Street
Avoca, Pennsylvania

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFIT/GNE/PH/78D-15	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) AN INVESTIGATION OF THE METHOD OF FINITE ELEMENTS WITH ACCURACY COMPARISONS TO THE METHOD OF FINITE DIFFERENCES FOR SOLUTION OF THE TRANSIENT HEAT CONDUCTION EQUATION USING OPTIMUM IMPLICIT FORMULATIONS		5. TYPE OF REPORT & PERIOD COVERED MS Thesis 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) GENECZKO, JOSEPH C.		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Air Force Institute of Technology (AFIT/EN) Wright-Patterson AFB OH 45433		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Materials Laboratory (AFML/MBC) Wright-Patterson AFB OH 45433		12. REPORT DATE December 1978 13. NUMBER OF PAGES 169
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Approved for public release IAW AFR 190-7

JOSEPH P. HIPPS, Major, USAF
Director of Information 19 Jan 79

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Numerical Analysis
Finite Differences
Finite Elements
Heat Conduction
Diffusion

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The one-dimensional transient heat conduction equation, with Dirichlet boundary conditions, is solved by the method of finite-elements, employing a quadratic interpolation function. The numerical solutions are investigated with respect to accuracy and stability, and compared to like results attained by the method of finite-differences, and the finite-element method with linear interpolation. The version of the finite-element method used was based on a variational principle which is stationary in time; the temporal behavior of the differential

(Continued on Reverse)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

BLOCK 20 (Cont'd)

equation is treated with a finite-difference approximation. This method is equivalent to the method of Galerkin, called the Method of Weighted Residuals. The inherent discontinuity between the initial condition and boundary conditions was accounted for by substituting the exact analytical solution at the first time step and numerically computating from there. An equivalency relationship between the two finite-element methods is shown to exist. The finite-difference version of the Crank-Nicolson method is found to be more accurate than the finite-element version; for the fully implicit method, the opposite is found to be true. In the optimum implicit method, both finite-element solutions are shown equivalent to the finite-difference solution for a Fourier modulus of one. For other values of this parameter, the finite-element solution is more accurate.