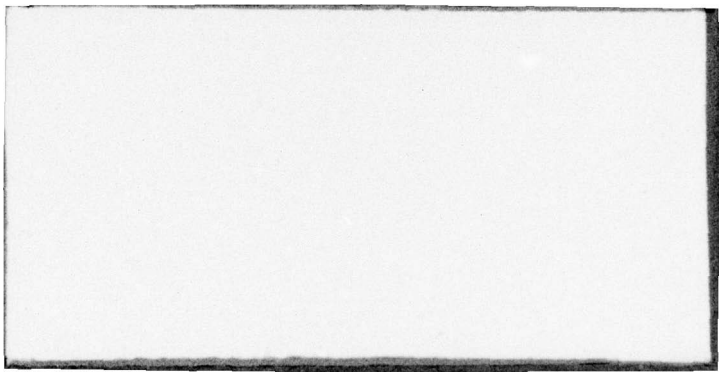
AD-A056 508 AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OHIO SCH--ETC F/G 20/13
AN INVESTIGATION OF THE NUMERICAL METHODS OF FINITE DIFFERENCES--ETC(U)
MAR 78 C R MARTIN

UNCLASSIFIED AFIT/GNE/PH/78M-6 NL

1 OF 4
AD
A056 508



AFIT/GNE/PH/78M-6

(1) LEVEL II

AD A056508

DDC
RECEIVED
JUL 21 1978
B

AD No. ____
DDC FILE COPY

(6) 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,

THESIS

(14) AFIT/GNE/PH/78M-6

(10) Charles R./Martin
Capt USAF

(9) Master's thesis,

(11) Mar 78

(12) 328p.

Approved for public release; distribution unlimited.

78 07 07 020

012225

AFIT/GNE/PH/78M-6

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

THESIS

Presented to the Faculty of the School of Engineering
of the Air Force Institute of Technology
Air Univeristy
in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

by

Charles R. Martin, B.S.
Capt USAF
Graduate Nuclear Engineering

March 1978

Preface

While the new theoretical treatment given here to the finite-element method resulted in discovering an optimum finite-element procedure which is actually a special case of a finite-difference method which was reported by Crandall, it lays the framework for a similar approach which can be used to give a truly useful numerical tool: a finite-element procedure for parabolic problems which results in a higher order of accuracy in the mean square sense over the entire solution domain.

I am deeply grateful to Drs. Bernard Kaplan and David Hardin, of the Air Force Institute of Technology, for their guidance and assistance in the development of this thesis, to Dr. J. Jones, also of the Air Force Institute of Technology, for his advice on special occasions, and to Dr. W. Kessler, of the Air Force Materials Laboratory, for sponsoring this research project. I wish also to express my gratitude to my wife, Susann, who helped in preparation of the final draft and gave invaluable moral support all during the project, and to Sharon Flores, whose expert typing skills added much to this thesis.

Contents

List of Tables

Abstract

The transient heat conduction equation, with Dirichlet and Neumann boundary conditions, is solved by the methods of finite-differences and finite-elements, and the numerical solutions are investigated with respect to accuracy and stability. A general six-point finite-difference expression is used for which there exists a high order accurate modification. 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 shown to be equivalent to the method of Galerkin. Several methods for treating accuracy and convergence problems which result from a discontinuity in the initial condition are investigated. The Crank-Nicolson method is a special case of both the finite-difference and finite-element methods. The finite-difference version of the Crank-Nicolson method is shown to be more accurate than the finite-element version, especially when a discontinuity exists between the initial condition and the boundary conditions. The high order accurate schemes for both finite-differences and finite-elements are shown to be equivalent for the case of linear elements derived from a stationary variational principle. Some of the results suggest the possiblity of finding a finite-element scheme which is highly accurate in a mean square sense over the entire soltuion domain.

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

I. Introduction

Background

Many problems arise in engineering practice which involve partial differential equations. Exact analytical methods, such as the classical separation of variables technique, often cannot be used to solve these equations. The purpose, then, of numerical techniques is to obtain reasonably accurate results for those problems for which exact methods cannot be used.

The most well known numerical technique is the finite-difference method. This method approximates the derivatives which occur in partial differential equations, reducing them to a set of algebraic equations which are then solved on a digital computer.

During the 1960's, there was widespread development of a new method, the finite-element method, for obtaining numerical solutions to differential equations. Most of these developments were in the field of solid-body mechanics, but more recently there have been applications of the finite-element method in the area of heat transfer. 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 will minimize this integral. As with the finite-difference method, application of the finite-element method results in a set of algebraic equations which can be solved on a digital computer.

One application of these numerical methods lies in the area of heat transfer. It is part of the responsibility of the Air Force Materials Laboratory to analyze the thermal response and ablation characteristics of rocket nozzles and re-entry vehicles. The transient heat conduction equation, a boundary value problem, is encountered in this analysis. Because of the irregular geometry involved, approximate numerical techniques, such as finite-differences and finite-elements, must be used.

One of the more accurate techniques of approximating the transient heat transfer equation is the Crank-Nicolson method (Ref 1). This technique was developed originally for use with the finite-difference method. There is a method which is related to the Crank-Nicolson scheme, the Crandall Method (Ref 2), which minimizes the truncation error in the finite-difference method. There is also a version of the Crank-Nicolson scheme which was developed for use with the finite-elements method. Obviously, if the error in the finite-element method could be minimized in a manner similar to the Crandall method, then the resulting solution would be more accurate. Finding this optimum method was the ultimate goal of this research project. The increased accuracy could be very beneficial to heat transfer engineers, provided it does not come at the expense of some more important factor, such as computer run time or stability.

Problem

The objective of this project was to solve a one-dimensional version of the transient heat conduction equation, with a known analytical solution, using modifications of the Crank-Nicolson versions of the finite-difference method and the finite-element method. The objective was to obtain the most accurate and stable solution to the differential equation. Accuracy is defined here as the deviation of the approximate solution from the exact analytical solution. Stability, on the other hand, refers to unwanted, numerically induced oscillations which may affect the solution. An unstable condition results when these oscillations grow without bound.

Scope

Initially, the analysis was only to include the following: (1) a comparison of the Crank-Nicolson version of the finite-difference method with the Crank-Nicolson version of the finite-element method with respect to accuracy and stability, and (2) an attempt to reduce the error in the finite-element method by experimentally varying a parameter. The analysis was not initially planned to include a theoretical treatment of the optimization process. However, during the investigation, the scope was broadened to include the explicit (Euler) and fully implicit methods, and also a theoretical treatment of the optimization process for finite-elements.

Assumptions

The fundamental assumptions which will hold throughout the text are: (1) that the physical properties of the material of interest are

constant with respect to time and spatial variables, (2) that the analysis in one dimension is representative of reality (for example, an insulated thin rod or a plane wall can be adequately represented with one spatial dimension under certain conditions), (3) that constant mesh spacing, Δx , is adequate for the user's needs, and finally, (4) that no heat generation takes place within the material of interest.

It is obvious that the restrictions are severe. One of the biggest difficulties encountered when using the finite-difference method occurs when either the arrangement of the nodes or the geometry is irregular. Some of these problems are simplified by the finite-element method. Unfortunately, these advantages may be negated by the restriction to constant material properties and constant mesh spacing. But the gains in accuracy are sure to find applications even under these restrictions. Also, it should be noted that this analysis does not necessarily hold for two or three spatial dimensions, although it is the opinion of the author that the method can be extended.

The restriction on heat generation here is merely for convenience, and the results of this paper would not be affected by the inclusion of a heat generation term so long as this term is constant with respect to the time and spatial variables.

II. Theory

The Physical Problem

General. The linear partial differential equation of second order in $u(w,v)$

$$Au_{ww} + Bu_{wv} + Cu_{vv} + Du_{w} + Eu_{v} + Fu = G \qquad (1)$$

where $A,B,\ldots,G$ are constants, or functions of w and v only, and where the subscript indicates differentiation with respect to the subscript, is classified as hyperbolic, parabolic, or elliptic type in a domain of the (w,v) plane depending on whether the values of the discriminant B^2-4AC are positive, zero, or negative, respectively, throughout the domain.

The heat conduction equation is parabolic in form and is no more than a mathematical statement of the first law of thermodynamics and Fourier's Law of heat transfer. The first law of thermodynamics states that the overall change in the internal energy of a system is equal to the heat which flows into the system plus the heat generated within the system less the heat which flows out of the system. If the derivative of each of these quantities is taken with respect to time, the result is the following energy balance:

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

where the dot indicates differentiation with respect to time.

The energy terms in Equation (2) are replaced by the appropriate rate equations. Some of the more common rate equations are for conduction, convection, radiation, heat storage, and heat generation. Conduction is defined as heat transfer through matter without net movement of the material, such as in a solid. The rate equation is given by Fourier's Law:

$$\dot{q} = -kA \frac{\partial T}{\partial x} \tag{3}$$

where

$\dot{q}$ = rate of heat flow in the x direction, Btu/hr

k = coefficient of thermal conductivity, Btu/hr-ft-°F

A = area normal to the x direction through which heat flows, H^2

T = temperature, °F

x = space variable, ft

Heat storage is described mathematically by

$$\dot{U}_{STOR} = \rho V c_p \frac{\partial T}{\partial t} \tag{4}$$

where

$\dot{U}_{STOR}$ = rate of heat storage, Btu/hr

ρ = density, lbm/ft^3

V = volume, ft^3

c_p = specific heat, Btu/lbm-°F

t = time, hr

The expressions for convection, radiation, and heat generation will not be needed in this report, but may be found in Myers (Ref 3:2-4).

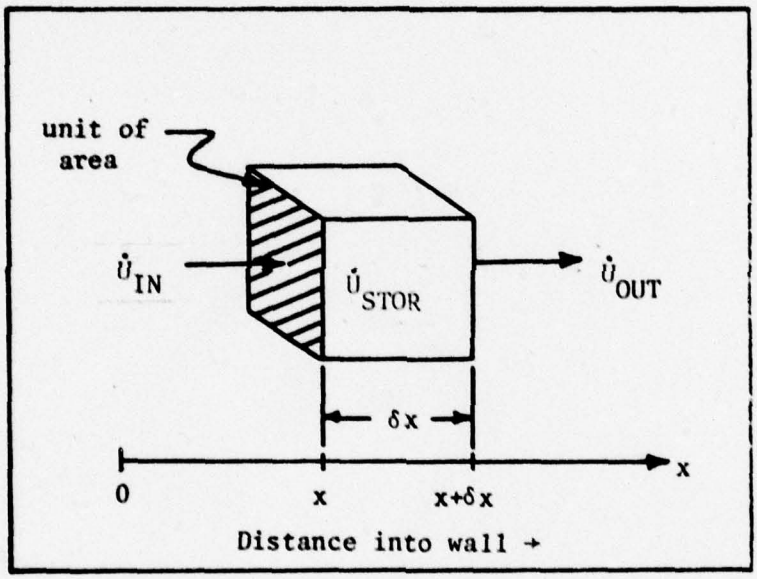


Figure 1. A Unit of Volume From a Wall with Large Dimensions in the y and z Directions.

From Figure 1, the energy balance equation for the given plane wall system can be seen to be

$$\rho c_p A \frac{\partial T}{\partial t} = \frac{kA\frac{\partial T}{\partial x}_{x+\delta x} - kA\frac{\partial T}{\partial x}_{x}}{\delta x} \qquad (5)$$

where it is assumed that no energy generation takes place within the unit of volume shown, and A represents the unit of area shown. In the limit as Δx approaches zero, the equation becomes the standard heat conduction equation which is parabolic in type:

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

If the material properties ρ , c_p and k are constant with respect to x , and a constant cross-sectional area is used, the equation becomes

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

This is the form of the heat equation which will be analyzed in this paper. The problem is not completely specified until the heat equation is associated with boundary conditions and initial conditions. Two types of boundary conditions will be investigated, Neumann and Dirichlet. If the function itself is specified on the boundary, it is termed a Dirichlet condition. If the derivative of the function is specified along the boundary, it is termed a Neumann condition. If the boundary condition contains values of the function and its derivative, it is referred to as a mixed boundary condition.

Primary Problem. The primary problem considered is that of a parallel sided plane wall, infinite in all directions normal to the direction of heat flow, which is heated until a steady state temperature

of 1000° F is reached at all points within the wall. The surfaces of the wall are then cooled instantaneously to 0° F and maintained at that temperature. This problem was chosen for its simplicity and also because it has been analyzed by the Crank-Nicolson and Crandall finite-difference methods (Ref 4:325). In order to provide more generality to the problem, it will be analyzed in normalized form. If T_B is the specified boundary temperature and T_i is the initial temperature, then the problem can be written

$$\frac{\partial T}{\partial t} = \alpha_m \frac{\partial^2 T}{\partial x^2} \tag{8}$$

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

$$T(0,t) = T(L,t) = T_B, \quad t > 0 \tag{10}$$

where

$\alpha_m = k/\rho c_p$ = the thermal diffusivity of the wall material

L = the thickness of the wall

T_i = 1000° F

T_B = 0° F

Let $u = (T-T_B)/T_o$, $\bar{x} = x/x_o$, and $\theta = t/t_o$ where T_o, x_o, and t_o are left unspecified for the time being. These dimensionless variables are then substituted into Equation (8) which is then multiplied by t_o/x_o^2 to yield

$$\frac{\partial u}{\partial \theta} = \frac{\alpha_m t_o}{x_o^2} \frac{\partial^2 u}{\bar{x}^2} \tag{11}$$

Since x_o and t_o are unspecified, the choice of $x_o = L$ and $t_o = L^2/\alpha_m$ yields the desired normalized partial differential equation:

$$\frac{\partial u}{\partial \theta} = \frac{\partial^2 u}{\partial \bar{x}^2} \tag{12}$$

If T_o is chosen as T_i, the boundary and initial conditions become, after substitution

$$u(\bar{x},\theta) = 1 \quad , \quad \theta = 0 \tag{13}$$

$$u(0,\theta) = u(1,\theta) = 0 \quad , \quad \theta > 0 \tag{14}$$

A schematic diagram of the normalized problem is shown in Figure 2. The most striking feature of the problem is the discontinuity between the initial condition, $u(\bar{x},0) = 1$, and the boundary conditions, $u(0,\theta) = u(1,\theta) = 0$. This discontinuity affects the solution most noticeably at points near the boundary and at early times, and it is

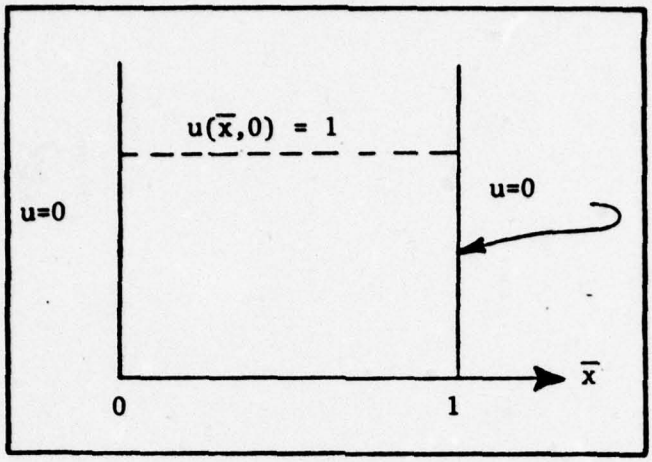


Figure 2. A Schematic Diagram of the Primary Problem.

the source of problems with convergence (to be discussed later) of both the finite-difference and finite-element higher order accurate schemes. Several attempts to reduce the effect of this discontinuity on the numerical solution of the problem will be discussed.

The exact solution of this problem can be obtained by a number of techniques. The method of separation of variables was used here, but the Laplace transform method yields the same form of solution. The problem is solved in Appendix A yielding an infinite series as the solution:

$$u(\bar{x},\theta) = \sum_{n=1}^{\infty} \frac{4}{(2n-1)\pi} \sin((2n-1)\pi\,\bar{x})\, e^{-((2n-1)\pi)^2\theta}, \quad \theta > 0 \qquad (15)$$

The solution is shown in graphical form in Figure 3.

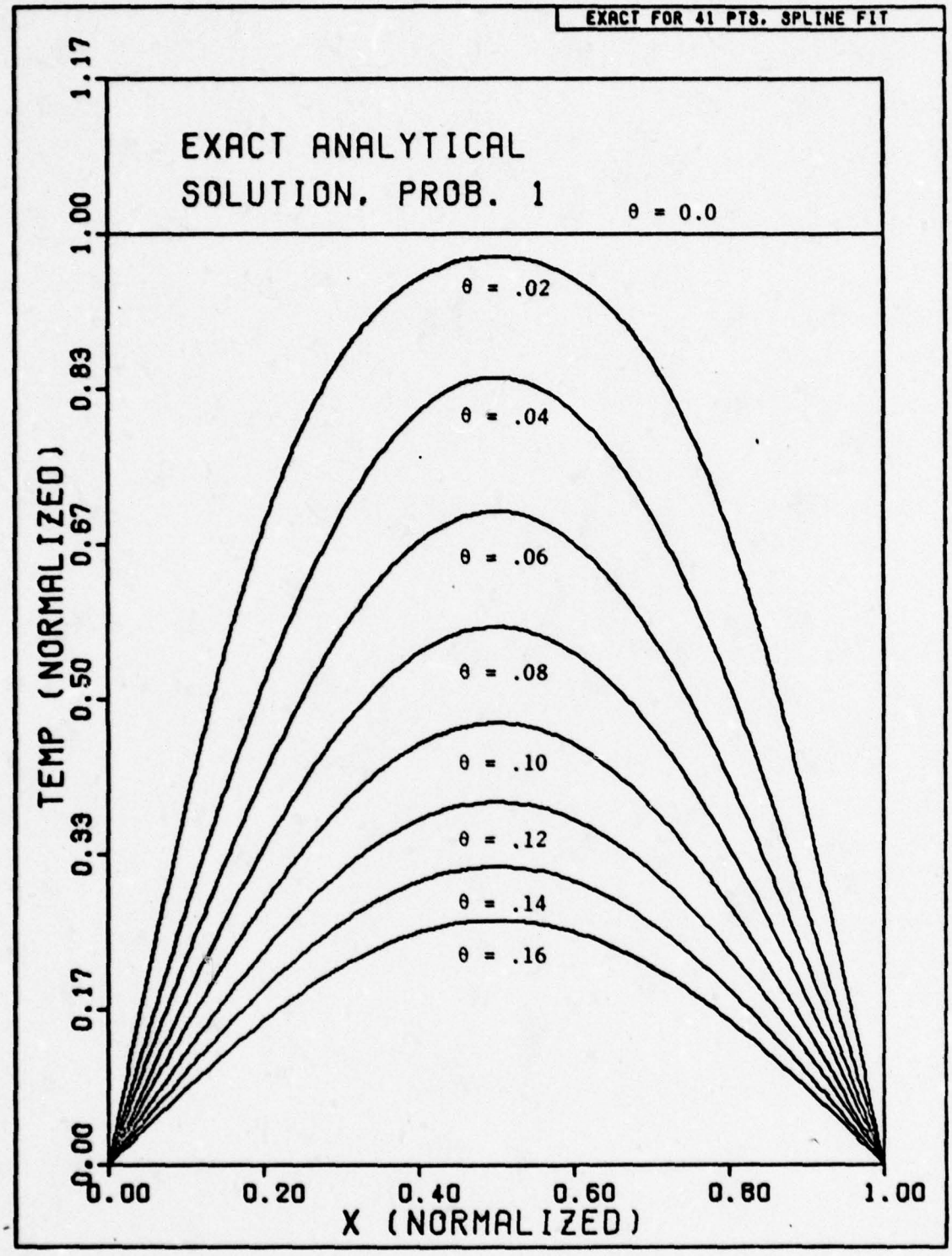


Figure 3. Analytic Solution of the Primary Problem.

Verification of the Primary Problem. The coefficients, $a_n = 4/(2n-1)\pi$, are bounded for all n since $a_n \to 0$ as $n \to \infty$; therefore, $|a_n| < M$, where M is some constant. Thus, whenever $\theta \geq \theta_o$, where θ_o is any positive constant

$$|a_n \sin((2n-1)\pi \bar{x}) \exp(-((2n-1)\pi)^2\theta| < M \exp(-((2n-1)\pi)^2\theta_o) \qquad (16)$$

Since the ratio test ensures the convergence of the infinite series with constant terms $M \exp(-((2n-1)\pi)^2\theta_o)$, the Weierstrass M-test provides a sufficient condition for the uniform convergence of Equation (15) when $0 \leq \bar{x} \leq 1$, $\theta \geq \theta_o > 0$ (Ref 5:28). Since the terms of the series in Equation (15) are continuous functions, the series converges to the continuous function $u(\bar{x},\theta)$ for $\theta > 0$, since θ_o is any positive constant.

Similarly, the series with terms $\exp(-((2n-1)\pi)^2\theta_o)$, or $(2n-1)\pi \exp(-((2n-1)\pi)^2\theta_o)$, also converges. Therefore, the series in Equation (15) can be differentiated once with respect to t and twice with respect to x, when $t > 0$, since, by the Weierstrass M-test, the series of derivatives converges uniformly. Further, since the terms of Equation (15) satisfy Equation (12), by superposition, the sum of the series, $u(\bar{x},\theta)$, also satisfies Equation (12).

To establish the convergence of the series in Equation (15) to the initial condition at $\theta = 0$, use will be made of Abel's test (Ref 5:228). First, $f(\bar{x})$ is defined as the odd periodic extension of the initial condition with a period of 2. Further, for each $\bar{x} \epsilon R$,

the series with terms $a_n \sin((2n-1)\pi \bar{x})$ converges to $f(\bar{x})$. At points where a discontinuity exists, $f(\bar{x})$ is defined as

$$f(\bar{x}) = \frac{f(\bar{x} + 0) + f(\bar{x} - 0)}{2} \tag{17}$$

where $f(\bar{x} + 0)$ and $f(\bar{x} - 0)$ are the right and left hand limits of f at $\bar{x}$. According to Abel's test, the series which results after multiplying each term of a convergent series by the corresponding elements of a sequence of functions of θ , which is bounded from above, as is $\exp(-((2n-1)\pi)^2\theta)$, is uniformly convergent for all $\theta \geq 0$.

This completes the formal verification of the solution to the primary problem.

Error in Truncation of the Infinite Series. The computer, although fast, is incapable of summing an infinite series. However, it would be of value to estimate the error which results when the infinite series solution is truncated after N terms. The following theorem, given by Thomas (Ref 6:660), gives the conditions necessary to estimate the error incurred by truncating an infinite series.

Theorem. If the series

$$\sum_{n=1}^{\infty} a_n \sin((2n-1)\pi \bar{x}) \exp(-((2n-1)\pi)^2\theta) \tag{18}$$

is strictly alternating, and if the n^{th} term tends to zero as $n \to \infty$, then each term is numerically less than or equal to its predecessor. The proof of this theorem is given by Thomas. One

further detail is required to get an error estimate. The terms of the series (18) must be regrouped into a strictly alternating series. This can be done if, and only if, (18) is uniformly convergent, which has previously been shown to be the case. Inspection of (18) indicates the second condition above is met, also. The regrouping is done by the computer by testing the sign of each term of the series sequentially and adding the current term to the previous term if the sign remains unchanged, or by beginning a new term if the sign changes. The error estimate is given by the last complete term of the new strictly alternating series. The series terminates when the computer reaches the limits of storage capacity prior to an underflow condition, that is when a number is so small that the computer cannot represent it in a memory word.

Of course, the attention to detail here is somewhat academic, since the sine and exponential functions used in the analytic solution are themselves approximated by built in computer functions which make use of truncated infinite series.

Secondary Problem. The secondary problem considered is not given as complete a treatment as the primary problem. The only difference between the two, in the unnormalized form, is that the right boundary temperature is not fixed for all $t > 0$, but rather that boundary is insulated. Therefore, the temperature gradient at the right boundary is zero. Normalization of the secondary problem is similar to that of the primary problem and yields

$$\frac{\partial u}{\partial \theta} = \frac{\partial^2 u}{\partial \bar{x}^2} \tag{19}$$

$$u(\bar{x},\theta) = 1 \quad , \quad \theta = 0 \tag{20}$$

$$u(0,\theta) = 0 \quad , \quad \theta > 0 \tag{21}$$

$$\left.\frac{\partial u}{\partial \bar{x}}\right|_{\bar{x}=1} = 0 \quad , \quad \theta > 0 \tag{22}$$

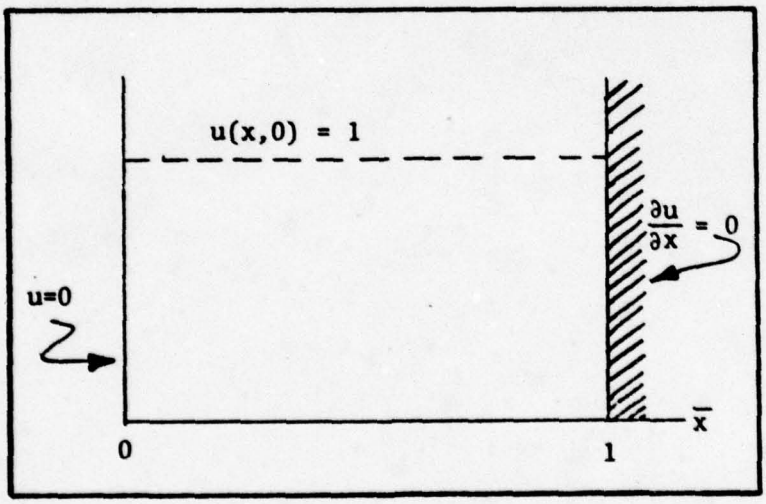


Figure 4. A Schematic Diagram of the Secondary Problem.

The normalized problem is shown schematically in Figure 4. Again, there is a discontinuity between the initial condition and the left boundary condition. The exact analytical solution of the secondary problem is given in Appendix B. The method of separation of variables was again used with the following result:

$$u(\bar{x},\theta) = \sum_{n=1}^{\infty} \frac{4}{(2n-1)\pi} \sin((2n-1)\frac{\pi}{2}) \, e^{-[(2n-1)\frac{\pi}{2}]^2\theta} \tag{23}$$

Verification of the solution is exactly analogous to that of the primary problem and will not be repeated here. The solution appears in graphical form in Figure 5.

Finite-Difference Formulation

General Expression. The finite-difference method makes use of differences to approximate derivatives. For example

$$\frac{du}{dx} \simeq \frac{u_b - u_a}{x_b - x_a} \tag{24}$$

could be used to approximate the derivative of u with respect to x taken at point b, point a, or halfway between a and b. If it represents the derivative at a in Figure 6, it is a forward difference expression. If the approximation represents the derivative at point b, it is a backward difference expression, and if it approximates the derivative at the point halfway between a and b, it is a central difference expression. The central difference approximation is the most accurate.

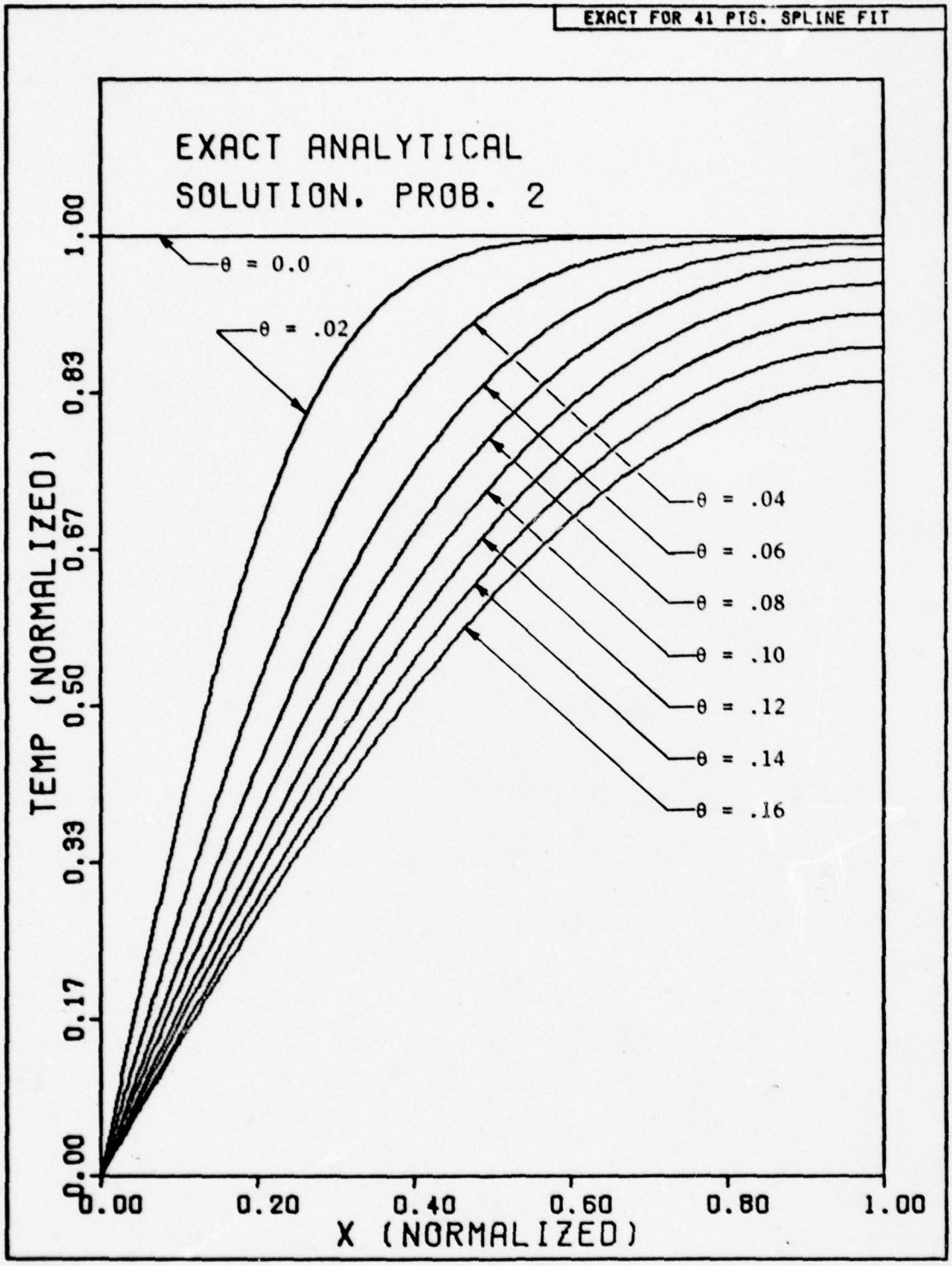


Figure 5. Analytic Solution of the Secondary Problem.

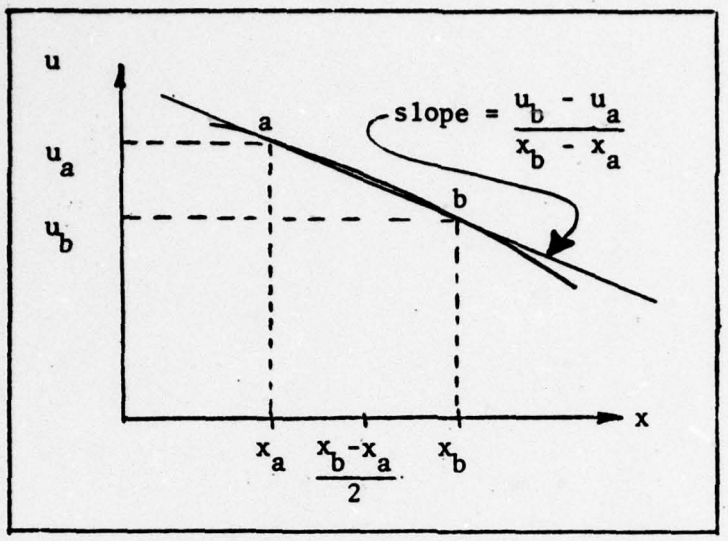


Figure 6. Finite-Difference Approximation of a Derivative.

These expressions can be used to replace the derivatives in the normalized heat equation. If the heat equation is forward differenced in time and central differenced in space, the resulting expression is

$$\frac{u_{i,k+1} - u_{i,k}}{\Delta\theta} = \frac{u_{i-1,k} - 2u_{i,k} + u_{i+1,k}}{\Delta x^2} \qquad (25)$$

where the subscript i refers to a nodal point on the space axis, and the subscript k refers to a point on the time axis in Figure 7. The bar over the x has been dropped for simplicity. This expression is called explicit since the temperature at the new time, $u_{i,k+1}$, is solved explicitly in terms of temperatures on the previous time level, k.

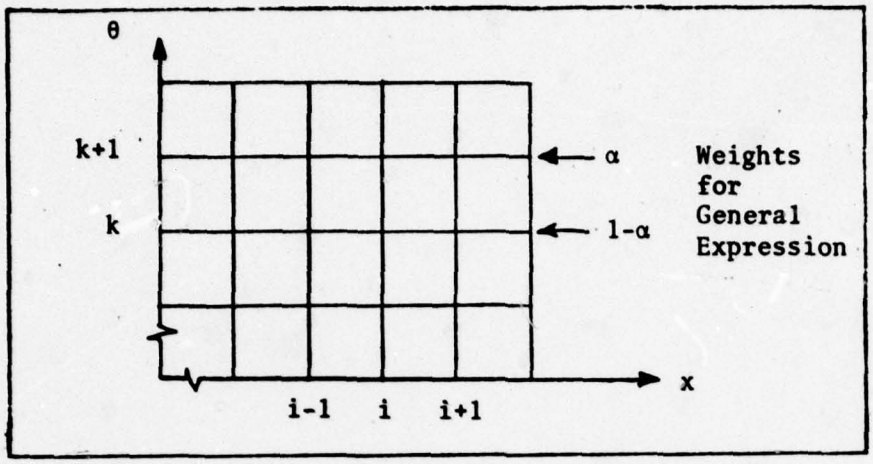


Figure 7. Space-Time Grid for the Heat Equation.

If the heat equation is backwards differenced in time and central differenced in space, the expression is

$$\frac{u_{i,k+1} - u_{i,k}}{\Delta\theta} = \frac{u_{i-1,k+1} - 2u_{i,k+1} + u_{i+1,k+1}}{\Delta x^2} \qquad (26)$$

This expression is termed implicit, since it requires the solution of a set of simultaneous equations at each new time level, shown here schematically as k+1 .

In 1947, Crank and Nicolson (Ref 1) proposed a scheme which can be considered as a central difference expression in time at time level $k + \frac{1}{2}$, and it makes use of central differences in space. The result is a more accurate difference equation:

$$\frac{u_{i,k+1} - u_{i,k}}{\Delta\theta} = \frac{1}{2} \frac{u_{i-1,k+1} - 2u_{i,k+1} + u_{i+1,k+1}}{\Delta x^2} + \frac{1}{2} \frac{u_{i-1,k} - 2u_{i,k} + u_{i+1,k}}{\Delta x^2} \tag{27}$$

This is also an implicit expression, as it requires the solution of a set of simultaneous equations. Note the equal weight on the new time level, $k+1$, and the old time level, k, on the right hand side of the equation.

In 1955, Crandall (Ref 2:318) proposed a generalized expression replacing the equal weights shown in Equation (27) with variable ones as shown in Figure 7. The resulting expression was

$$\frac{u_{i,k+1} - u_{i,k}}{\Delta\theta} = \alpha \frac{u_{i-1,k+1} - 2u_{i,k+1} + u_{i+1,k+1}}{\Delta x^2} + (1-\alpha) \frac{u_{i-1,k} - 2u_{i,k} + u_{i+1,k}}{\Delta x^2} \tag{28}$$

Crandall observed that for a certain value of the parameter α, the expression becomes very accurate. For this optimum value of α, the error incurred in approximating the time derivative very nearly equals the error in the approximation of the space derivative. The optimum value of α, shown in Figure 8, is given by

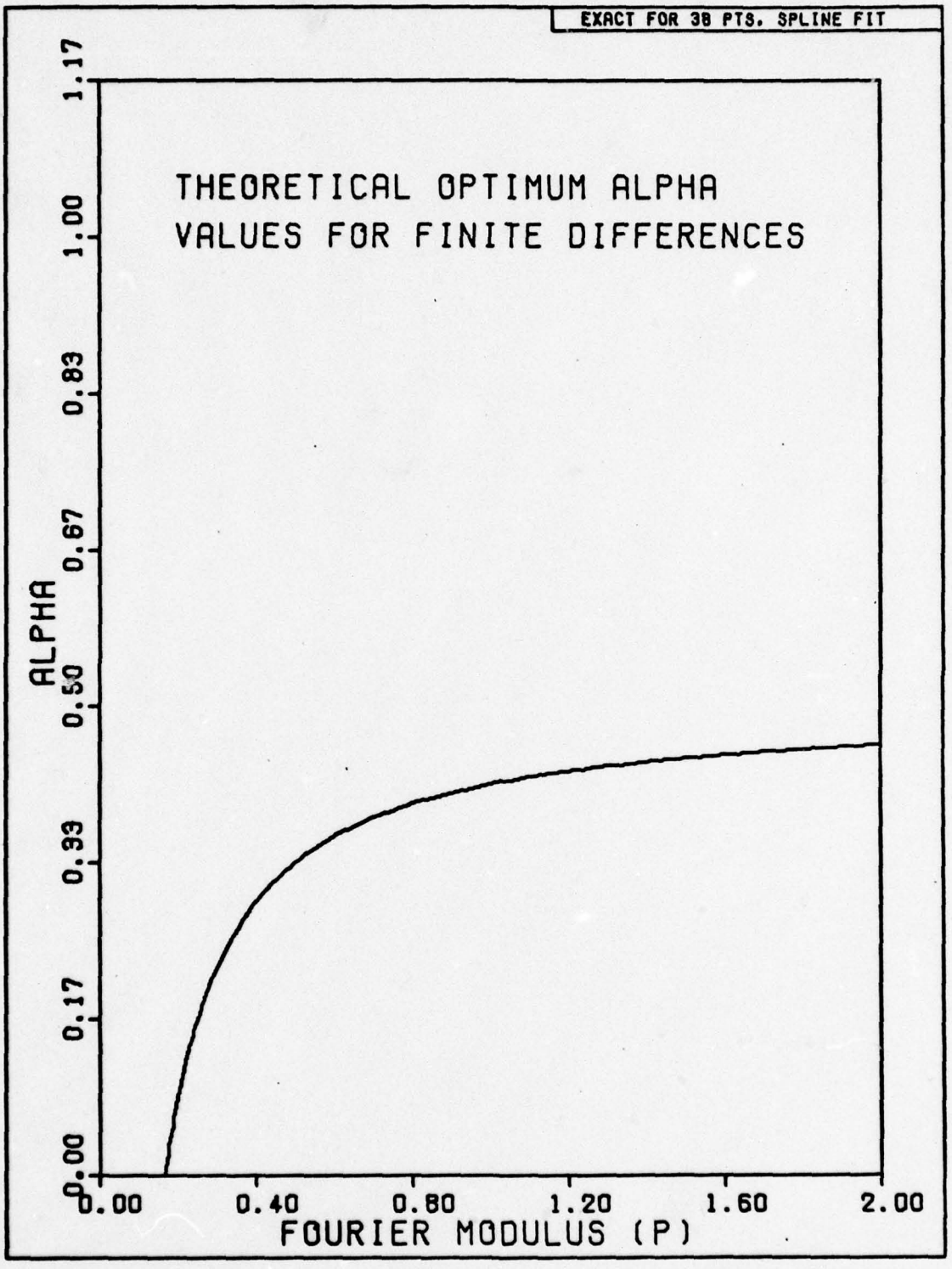


Figure 8. Theoretical Optimum Alpha Values for Finite Differences.

$$\alpha = \frac{1}{2}\left(1 - \frac{1}{6p}\right) \tag{29}$$

where $p = \Delta\theta/\Delta x^2$. The dimensionless parameter, p , is known in the literature as the Fourier modulus (Ref 19:938).

Application of the General Expression. To apply the general expression, Equation (28), to the primary problem, the space axis in Figure 7 is first divided into $N-1$ intervals. The value of the first node is governed by the boundary condition at $x=0$; whereas, the value of the Nth nodal point is governed by the boundary condition at $x=1$. As the subscript i is varied over the range of nodal values, the following set of equations results, written here in matrix form for the case $N=6$:

$$\begin{bmatrix} 0 & 1+2p\alpha & -p\alpha & 0 & 0 & 0 \\ 0 & -p\alpha & 1+2p\alpha & -p\alpha & 0 & 0 \\ 0 & 0 & -p\alpha & 1+2p\alpha & -p\alpha & 0 \\ 0 & 0 & 0 & -p\alpha & 1+2p\alpha & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{bmatrix}^{k+1}$$

$$= \begin{bmatrix} 0 & 1-2p(1-\alpha) & p(1-\alpha) & 0 & 0 & 0 \\ 0 & p(1-\alpha) & 1-2p(1-\alpha) & p(1-\alpha) & 0 & 0 \\ 0 & 0 & p(1-\alpha) & 1-2p(1-\alpha) & p(1-\alpha) & 0 \\ 0 & 0 & 0 & p(1-\alpha) & 1-2p(1-\alpha) & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{bmatrix}^k \qquad (30)$$

or in vector notation

$$\underline{A}\,\underline{u}^{k+1} = \underline{B}\,\underline{u}^{k} \qquad (31)$$

Equation (30) has four equations, but six unknowns. This difficulty is avoided, however, since the end nodes are not needed and can be discarded. The result is

$$\begin{bmatrix} 1+2p\alpha & -p\alpha & 0 & 0 \\ -p\alpha & 1+2p\alpha & -p\alpha & 0 \\ 0 & -p\alpha & 1+2p\alpha & -p\alpha \\ 0 & 0 & -p\alpha & 1+2p\alpha \end{bmatrix} \begin{bmatrix} u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix}^{k+1}$$

$$= \begin{bmatrix} 1-2p(1-\alpha) & p(1-\alpha) & 0 & 0 \\ p(1-\alpha) & 1-2p(1-\alpha) & p(1-\alpha) & 0 \\ 0 & p(1-\alpha) & 1-2p(1-\alpha) & p(1-\alpha) \\ 0 & 0 & p(1-\alpha) & 1-2p(1-\alpha) \end{bmatrix} \begin{bmatrix} u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix}^k \tag{32}$$

Note that, because of symmetry, only two columns of information will need to actually be stored in computer memory by a production computer code for each of the coefficient matrices. Additionally, the matrices are tridiagonal; thus, a matrix factorization scheme, the Thomas method, may be used to solve the matrix equation. The Thomas method (Ref 7:46-48) has the advantage of requiring fewer algebraic operatons than the more familiar Gauss reduction method. There are also iterative methods for solving these systems of equations.

To apply Equation (29) to the secondary problem, a fictitious point must be created where node N+1 would be if it existed (Ref 8:35). Applying Equation (28) to node N yields

$$-p\alpha u_{N-1,k+1} + (1+2p\alpha)\, u_{N,k+1} - p\alpha u_{N+1,k+1} \tag{33}$$

$$= p(1-\alpha)u_{N-1,k} + 1-2p(1-\alpha)u_{N,k} + p(1-\alpha)u_{N+1,k}$$

Additionally, the derivative boundary condition at $x=1$ must be approximated. Thus

$$\left.\frac{\partial u}{\partial x}\right|_{x=1} = 0 \quad , \quad \theta > 0 \tag{34}$$

becomes

$$\frac{u_{N+1,k} - u_{N-1,k}}{2\Delta x} = 0 \tag{35}$$

where a central difference approximation has been used. Solving Equation (35) for $u_{N+1,k}$ and substituting back into Equation (33) yields

$$-2p\alpha u_{N-1,k+1} + (1+2p\alpha)u_{N,k+1} = 2p(1-\alpha)u_{N-1,k} + (1-2p(1-\alpha))u_{N,k} \tag{36}$$

Again using $N=6$, the matrix equation now becomes

$$\begin{bmatrix} 1+2p\alpha & -p\alpha & 0 & 0 & 0 \\ -p\alpha & 1+2p\alpha & -p\alpha & 0 & 0 \\ 0 & -p\alpha & 1+2p\alpha & -p\alpha & 0 \\ 0 & 0 & -p\alpha & 1+2p\alpha & -p\alpha \\ 0 & 0 & 0 & -2p\alpha & 1+2p\alpha \end{bmatrix} \begin{bmatrix} u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{bmatrix}^{k+1}$$

$$\begin{bmatrix} 1-2p(1-\alpha) & p(1-\alpha) & 0 & 0 & 0 \\ p(1-\alpha) & 1-2p(1-\alpha) & p(1-\alpha) & 0 & 0 \\ 0 & p(1-\alpha) & 1-2p(1-\alpha) & p(1-\alpha) & 0 \\ 0 & 0 & p(1-\alpha) & 1-2p(1-\alpha) & p(1-\alpha) \\ 0 & 0 & 0 & 2p(1-\alpha) & 1-2p(1-\alpha) \end{bmatrix} \begin{bmatrix} u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{bmatrix}^{k} \qquad (37)$$

Linear Finite Element Formulation

Background. The finite-element method is fundamentally different in nature from the finite-difference method. The finite-difference method is designed to approximate the solution of a differential equation at a number of specified nodal points; this method gives no

information about the solution between these nodal points. The finite-element method does, however, assume a general form for the solution between these nodal points. The method will be introduced by way of the following simple illustration.

The differential equation

$$Du = \left(-\frac{d^2}{dx^2} + 1 \right) u = f \tag{38}$$

$$u(0) = u(L) = 0 \tag{39}$$

bears a special relationship to the functional

$$I(\tilde{u}) = \frac{1}{2} [D\tilde{u},\tilde{u}) - 2(f,\tilde{u})] \tag{40}$$

where the notation (g_1,g_2) represents the inner product of the two functions g_1 and g_2, or

$$(g_1,g_2) = \int_a^b g_1(x)\, g_2(x)\, dx \tag{41}$$

and $\tilde{u}(x) = u(x) + \xi v(x)$. The function $\tilde{u}$ is called a trial function, and the only restriction on the function v, called a variation of u, is that v must satisfy all Dirichlet boundary conditions, those written in terms of the function itself. Thus, v must have the properties

$$v(0) = v(L) = 0 \tag{42}$$

such that

$$\tilde{u}(0) = u(0) = \tilde{u}(L) = u(L) = 0 \tag{43}$$

Except for this restriction, $v(x)$ is left completely arbitrary. Now the relationship between Equation (38) and Equation (40) can be demonstrated. The functional $I(\tilde{u})$ has a minimum value at the point where $\tilde{u} = u$. That is, at the point where $\xi = 0$. Therefore, if the minimum value of the integral $I(\tilde{u})$ can be found, the value of $\tilde{u}$ which minimizes I is equal to the solution, u , of Equation (38).

In the finite-element method, the domain of the solution is divided into N elements. Within each of these elements, the temperature distribution is assumed to have a certain shape, for example, linear, as in Figure 9. A discussion of the notation and the meaning of certain terms would be helpful at this point.

The temperature distribution, $\tilde{u}^N$, where the superscript indicates the number of elements in the solution domain, is a member of the finite-dimensional space of functions, S^N ; the space, S^N , is a subspace of the infinite dimensional Hilbert space, H_E^1 , which, in addition to the general defining properties of a Hilbert space, is defined as the space of functions which are square integrable and whose first derivatives also are square integrable. Further, the subscript

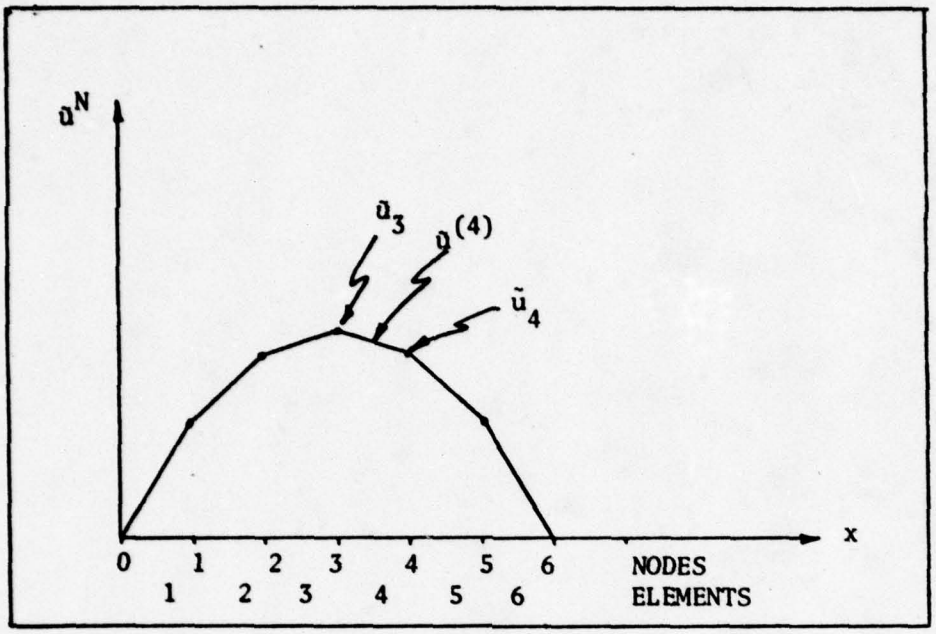


Figure 9. Finite-Element Arrangement for Solution of Eq. (38).

indicates that only the essential boundary conditions, defined below, of Equation (39) must be satisfied (Ref 9:5-11). In this particular case, both conditions are essential.

An essential boundary condition is one whose statement is in terms of derivatives below order s when the differential equation is of order $2s$. The reason for this distinction between essential conditions and natural conditions, which are defined to be those written in terms of derivatives or order s and higher, lies at the heart of the method. The inhomogeneous term, f, in Equation (38) will be used to illustrate convergence in the mean in hopes of clarifying the distinction made between essential and natural boundary conditions.

The eigenfunction expansion of f is made in terms of functions which necessarily satisfy the boundary conditions on the problem. However, the function f itself does not necessarily satisfy those conditions. The inifinte series expansion of f in terms of the eigenfunctions is said to converge in the mean square sense to f if

$$\lim_{m\to\infty} \int_0^L [f(x) - S_m(x)]^2 \, dx = 0 \tag{44}$$

where

$$S_m(x) = \sum_{n=1}^{m} a_n u_n(x) \tag{45}$$

and $u_n(x)$ represents the eigenfunctions of the differential operator (Ref 5:60-61). Thus, the limit of the expansion of eigenfunctions is contained in a space, H^0 in which only the functions themselves are required to be square integrable, outside the solution space for the problem, H_B^2 in which the functions along with their first and second derivatives must be square integrable and which must satisfy the boundary conditions on the problem, from which the eigenfunctions $u_n(x)$ are taken.

In a similar manner, the trial functions, $u(x)$, need only lie in H_E^1 , since Equation (40) can be integrated by parts to give

$$I(\tilde{u}) = \frac{1}{2}\int_0^L \left[\left(\frac{d\tilde{u}}{dx}\right)^2 + (\tilde{u}(x))^2 - 2f(x)\,\tilde{u}(x)\right] dx \qquad (46)$$

Thus, the functional to be minimized is expressed in terms of the first derivative. Any function, $\tilde{u}$, will be admissible in the minimizing process provided that it can be expressed as a limit in the mean of a sequence of functions, $\tilde{u}_N$, which lie in H_B^2 (Ref 9:11). Limit in the mean is the equivalent of Equation (44) where $f(x)$ is the limit in the mean of the sequence $S_m(x)$. It should be noted that, while the functions $\tilde{u}_N$ are required to lie in H_B^2, since only mean square convergence is asked, the limit $\tilde{u}$ will necessarily satisfy only the Dirichlet conditions and lie in the space H^1. Thus, there is the advantage that the functions, $\tilde{u}(x)$, can be constructed from piecewise linear functions.

The value $I(\tilde{u})$ of the functional I, then, is a limiting case for the sequence of values $I(\tilde{u}_N)$. The function $\tilde{u} = u$ which minimizes this functional, can be shown to satisfy any natural boundary conditions, since for any ξ and any $v(x)$ in H_E^1, $\tilde{u}(x)$ also lies in H_E^1 and

$$I(u) \leq I(\tilde{u}) \qquad (47)$$

where

$$\begin{aligned} I(\tilde{u}(x)) &= I(u(x) + \xi\, v(x)) \\ &= I(u) + 2\xi \int_0^L \left[\frac{du}{dx}\frac{dv}{dx} + uv - fv\right] dx \\ &\quad + \xi^2 \int_0^L \left[\left(\frac{dv}{dx}\right)^2 + v^2\right] dx \end{aligned} \qquad (48)$$

Since ξ can be positive or negative, the middle term on the right hand side of Equation (48) must be identically equal to zero. Thus,

$$\int_o^L \left[\frac{du}{dx}\frac{dv}{dx} + uv - fv\right]dx = 0 \qquad (49)$$

If the left hand side is integrated by parts, the result is

$$\int_o^L \left[\frac{du}{dx}\frac{dv}{dx} + uv - fv\right] dx$$

$$- \left(\frac{du}{dx} v\right)\Big]_o^L + \int_o^L \left[\left(-\frac{d^2u}{dx^2}\right) v + uv - fv\right] dx \qquad (50)$$

The right hand side of Equation (50) is equal to zero for any $v(x)$ in H_E^1 if the differential equation itself is satisfied and if, assuming only natural conditions were originally imposed on $u(x)$, $du/dx = 0$ at the boundaries. It is also easy to see why any essential conditions must be imposed on $v(x)$ and therefore upon $\tilde{u}(x) = u(x) + \xi v(x)$, since in the absence of a derivative boundary condition, the essential conditions imposed on v will be needed to force the first term on the right of Equation (50) to vanish.

In Figure 8, $\tilde{u}^N(x)$, which is a piecewise linear function defined in terms of the nodal values, can be substituted into I, and I in turn can be written as a sum of N functionals, each defined over a different portion of the domain. These portions are termed "elements." That portion of I defined on the ith element, $I^{(i)}$, can be

minimized by taking the derivative of $I^{(i)}$ with respect to each nodal value and setting the result equal to zero. The result is a system of algebraic equations which must be solved to obtain the nodal temperatures. Since the functional, I , is defined over the entire interval, the minimization process takes place in a mean square sense. This is an important distinction between finite-differences and finite elements. In the finite-element method, the errors are distributed over the entire domain and the solution can be, but is not necessarily, more accurate at points between the nodes than it is at the nodes, as shown in Figure 10.

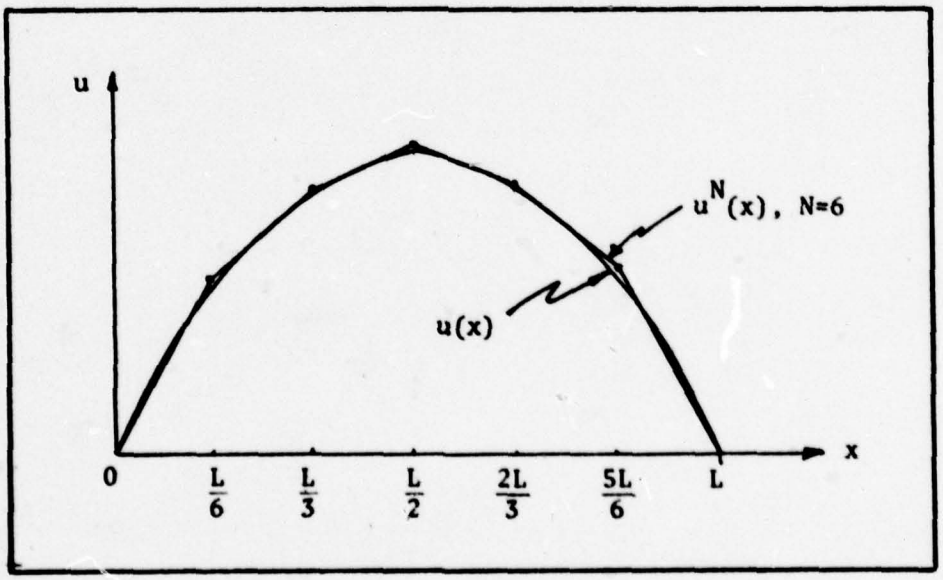


Figure 10. Example of a Finite-Element Solution.

Application of the Method. Now that the finite-element method has been introduced, the method will be applied to the heat conduction equation.

What is known as the finite-element method is actually a group of methods. The Rayleigh-Ritz method makes use of the associated variational principle as discussed in the previous section. There is another method, called the method of weighted residuals or the method of Faedo-Galerkin, which is nothing more than the discretization in time of the weak form of the problem:

$$\left(\left(\frac{\partial u}{\partial \theta} - \frac{\partial^2 u}{\partial x^2}\right), v\right) = 0 \tag{51}$$

where it should be recalled that θ and x are the normalized time and space variables. The weak form of the solution does not require that the associated functional, I, be minimized, but rather that the first variation of I must vanish (Ref 9:9). That is,

$$\frac{dI}{du} = 0 \tag{52}$$

The expression used in the Galerkin method is derived in Appendix E. The method that will be used here is based on a variational principle that was suggested by Meyer (Ref 3:399). Meyer was somewhat obscure in the justification for his variational statement. Indeed, if the approach he seems to suggest is used, the only conclusion which may be drawn is that no variational statement

exists. In fact, it was so thought by Strang and Fix (Ref 9:242) and Washizu (Ref 10). It turns out that Gurtin has derived a variational principle for the heat conduction problem which involves the use of convolutions (Ref 11:255). Meyer's variational principle can be justified, however, in the following way. If the space and time variables are separated, then the heat conduction equation can be considered to be the equivalent of an elliptic equation in x at each point in time, and a variational statement can be formulated by neglecting the time dependence. The variational principle, which is simply stated by Meyer, is derived by this method in Appendix E. The resulting functional, which is to be minimized, is

$$I(\tilde{u}) = \frac{1}{2} \int_0^1 \left[\frac{\partial (\tilde{u}^N)^2}{\partial \theta} + \left(\frac{\partial \tilde{u}^N}{\partial x} \right)^2 \right] dx \tag{53}$$

To simplify things, the integral is divided into two integrals:

$$I_2 = \frac{1}{2} \int_0^1 \frac{\partial}{\partial \theta} (\tilde{u}^N)^2 \, dx \tag{54}$$

and

$$I_1 = \frac{1}{2} \int_0^1 \left(\frac{\partial \tilde{u}^N}{\partial x} \right)^2 dx \tag{55}$$

The following formulation is developed using the matrix notation of Meyer (Ref 3:342-357).

I is to be differentiated with respect to the nodal temperatures and set equal to zero in order to find the minimum within the space S^N . If $\underline{\tilde{u}}^N$ is given by

$$\underline{\tilde{u}}^N = \begin{bmatrix} \tilde{u}_1 \\ \tilde{u}_2 \\ \cdot \\ \cdot \\ \cdot \\ \tilde{u}_{N-1} \\ \tilde{u}_N \end{bmatrix} \tag{56}$$

then the minimizing condition is

$$\frac{dI}{d\underline{\tilde{u}}^N} = \frac{dI_1}{d\underline{\tilde{u}}^N} + \frac{dI_2}{d\underline{\tilde{u}}^N} \tag{57}$$

The interval [o,1] is divided into N elements, as shown in Figure 11, which are considered individually. Instead of differentiating the elemental integrals with respect to each component of $\underline{\tilde{u}}^N$, a matrix, $\underline{D}^{(I)}$, is defined by

$$\underline{D}^{(i)} = \begin{bmatrix} 0 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} \begin{matrix} \\ \\ \leftarrow \text{ ith row} \\ \leftarrow \text{ i+1th row} \\ \\ \\ \\ \end{matrix} \qquad (58)$$

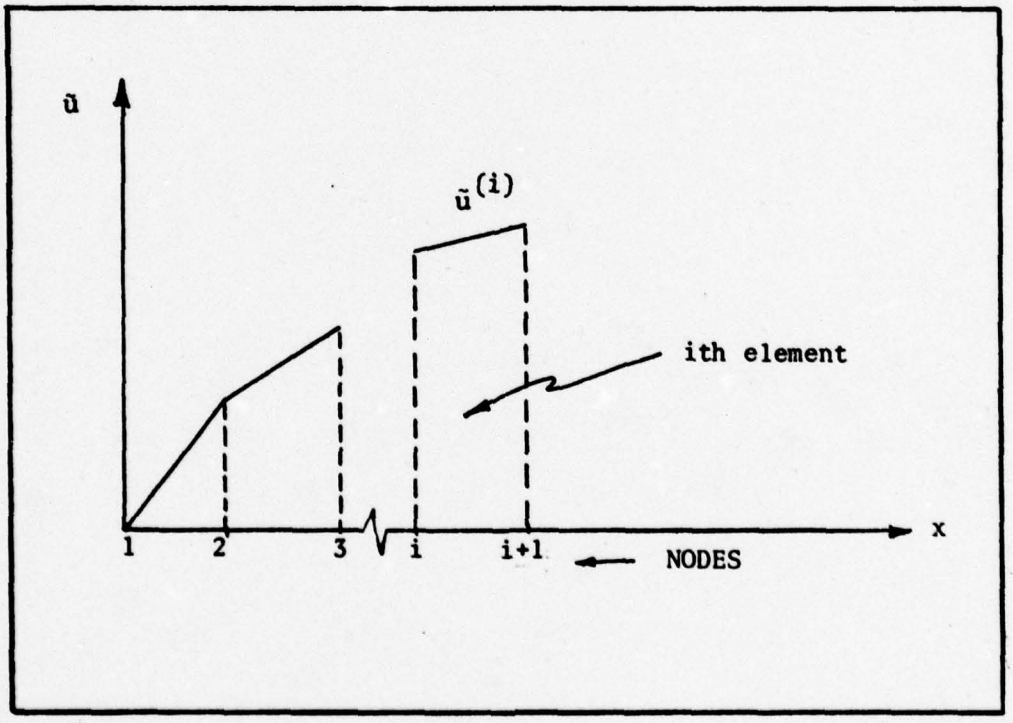


Figure 11. Arrangement of the Elements.

and used as follows:

$$\frac{dI_1}{d\underline{\tilde{u}}^N} = \sum_{i=1}^{N} \underline{D}^{(i)} \frac{dI_1^{(i)}}{d\underline{\tilde{u}}^{(i)}} \tag{59}$$

and

$$\frac{dI_2}{d\underline{\tilde{u}}^N} = \sum_{i=1}^{N} \underline{D}^{(i)} \frac{dI_2^{(i)}}{d\underline{\tilde{u}}^{(i)}} \tag{60}$$

where $I^{(i)}$ is the portion of I defined on the ith interval, $(i, i+1)$, and where

$$\underline{\tilde{u}}^{(i)} = \begin{bmatrix} \tilde{u}_i \\ \tilde{u}_{i+1} \end{bmatrix} \tag{61}$$

The temperature distribution within each element must be assumed. Since the trial functions, $\tilde{u}^N$, must be chosen from a subspace, S^N, of H_E^1, they must be continuous. Piecewise constant functions cannot be used. The next simplest choice is piecewise linear functions. Therefore.

$$\tilde{u}^{(i)} = c_1^{(i)} + c_2^{(i)} x \tag{62}$$

or

$$\tilde{u}^{(i)} = \underline{p}^T \underline{c}^{(i)} \tag{63}$$

where

$$\underline{p}^T = [1 \quad x] \tag{64}$$

and

$$\underline{c}^{(i)} = \begin{bmatrix} c_1^{(i)} \\ c_2^{(i)} \end{bmatrix} \tag{65}$$

The coefficients in Equation (62) can be eliminated by considering the set of equations formed by substituting Equation (62) into Equation (61) and eliminating $c_1^{(i)}$ and $c_2^{(i)}$. The result is

$$\tilde{u}^{(i)} = \underline{p}^T \underline{R}^{(i)} \underline{\tilde{u}}^{(i)} \tag{66}$$

where

$$\underline{R}^{(i)} \triangleq \frac{1}{x_i} \begin{bmatrix} x_{i+1} & -x_i \\ -1 & 1 \end{bmatrix} \tag{67}$$

and

$$\Delta x_i = x_{i+1} - x_i \tag{68}$$

The subscript on Δx can be dropped since the interval spacing is assumed to be constant. Next, the derivative of $\tilde{u}^{(i)}$ is taken with respect to x to give

$$\frac{\partial \tilde{u}^{(i)}}{\partial x} = \underline{p}_x^T \underline{R}^{(i)} \underline{\tilde{u}}^{(i)} \tag{69}$$

where

$$\underline{p}_x^T = \frac{\partial}{\partial x} (\underline{p}^T) = [0 \quad 1] \tag{70}$$

Equation (69) can then be substituted into $I_1^{(i)}$ to give

$$I_1^{(i)} = \frac{1}{2} \int_i^{i+1} (\underline{p}_x^T \underline{R}^{(i)} \underline{\tilde{u}}^{(i)})^2 \, dx \tag{71}$$

I_1 is then differentiated with respect to $\underline{\tilde{u}}^{(i)}$ to give

$$\frac{dI_1^{(i)}}{d\underline{\tilde{u}}^{(i)}} = \int_{x_i}^{x_{i+1}} (\underline{p}_x^T \underline{R}^{(i)} \underline{\tilde{u}}^{(i)}) \frac{d}{d\underline{\tilde{u}}^{(i)}} (\underline{p}_x^T R^{(i)} \underline{\tilde{u}}^{(i)}) \, dx \tag{72}$$

or

$$\frac{dI_1^{(i)}}{d\underline{\tilde{u}}^{(i)}} = \int_{x_i}^{x_{i+1}} (\underline{p}_x^T \underline{R}^{(i)} \underline{\tilde{u}}^{(i)}) (\underline{p}_x \underline{R}^{(i)})^T dx \tag{73}$$

Since $(\underline{p}_x \underline{R}^{(i)} \underline{\tilde{u}}^{(i)})$ is a scalar, its order is unimportant and Equation (73) can be rearranged as

$$\frac{dI_1^{(i)}}{d\underline{\tilde{u}}^{(i)}} = \int_{x_i}^{x_{i+1}} (\underline{p}_x^T \underline{R}^{(i)})^T (\underline{p}_x^T \underline{R}^{(i)} \underline{\tilde{u}}^{(i)}) dx \tag{74}$$

And since

$$(\underline{A}\ \underline{B})^T = \underline{B}^T \underline{A}^T \tag{75}$$

Equation (74) is equivalent to

$$\frac{dI_1^{(i)}}{d\underline{\tilde{u}}^{(i)}} = \int_{x_i}^{x_{i+1}} \underline{R}^{(i)^T} \underline{p}_x (\underline{p}_x^T \underline{R}^{(i)} \underline{\tilde{u}}^{(i)}) dx \tag{76}$$

The matrices $\underline{R}^{(i)}$ and $\underline{\tilde{u}}^{(i)}$ are independent of x and can be removed from the integral:

$$\frac{d\underline{I}_1^{(i)}}{d\underline{\tilde{u}}^{(i)}} = \underline{R}^{(i)^T} \left[\int_{x_i}^{x_{i+1}} \underline{p}_x \; \underline{p}_x^T \, dx \right] \underline{R}^{(i)} \; \underline{\tilde{u}}^{(i)} \tag{77}$$

or, if the integration and the matrix multiplication are carried out

$$\frac{d\underline{I}_1^{(i)}}{d\underline{\tilde{u}}^{(i)}} = \frac{1}{\Delta x} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \underline{\tilde{u}}^{(i)} \tag{78}$$

If $\underline{K}^{(i)}$, referred to in the literature as the element stiffness matrix, is defined by

$$\underline{K}^{(i)} \triangleq \frac{1}{\Delta x} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{79}$$

then, Equation (59) can be written

$$\frac{d\underline{I}_1}{d\underline{\tilde{u}}^N} = \sum_{i=1}^{N} \underline{D}^{(i)} \; \underline{K}^{(i)} \; \underline{\tilde{u}}^{(i)} \tag{80}$$

or

$$\frac{dI_1}{d\underline{\tilde{u}}^N} = \sum_{i=1}^{N} \underline{D}^{(i)} \underline{K}^{(i)T} \underline{\tilde{u}}^N \tag{81}$$

where $\underline{\tilde{u}}^N$ was defined by Equation (56). If $\underline{K}$, the global stiffness matrix, is defined by

$$\underline{K} \triangleq \sum_{i=1}^{N} \underline{D}^{(i)} \underline{K}^{(i)} \underline{D}^{(i)T} \tag{82}$$

and substituted into Equation (81), then

$$\frac{dI_1}{d\underline{\tilde{u}}^N} = \underline{K}\, \underline{\tilde{u}}^N \tag{83}$$

A similar treatment is given to Equation (60). First, since

$$I_2^{(i)} = \frac{1}{2} \int_{x_i}^{x_{i+1}} \frac{\partial}{\partial\theta} (\tilde{u}^{(i)})^2 \, dx \tag{84}$$

Leibnitz' rule for differentiation of integrals can be used to give

$$I_2^{(i)} = \frac{1}{2} \frac{d}{d\theta} \int_{x_i}^{x_{i+1}} (\tilde{u}^{(i)})^2 \, dx \tag{85}$$

or

$$I_2^{(i)} = \frac{1}{2} \frac{d}{d\theta} \int_{x_i}^{x_{i+1}} (\underline{P}^T \underline{R}^{(i)} \underline{\tilde{u}}^{(i)})^2 \, dx \tag{86}$$

The derivative of this expression is taken with respect to $\underline{u}^{(i)}$ to yield

$$\frac{dI_2^{(i)}}{d\underline{u}^{(i)}} = \frac{d}{d\theta} \frac{\Delta x}{6} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \underline{\tilde{u}}^{(i)} \tag{87}$$

The matrix

$$\underline{M}^{(i)} \triangleq \frac{\Delta x}{6} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \tag{88}$$

is referred to in the finite-element literature as an elemental mass matrix. This definition and Equation (87) are then substituted into Equation (60):

$$\frac{dI_2}{d\underline{\tilde{u}}^N} = \sum_{i=1}^{N} \underline{D}^{(i)} \frac{d}{d\theta} (\underline{M}^{(i)} \underline{\tilde{u}}^{(i)}) = \sum_{i=1}^{N} \underline{D}^{(i)} \frac{d}{d\theta} (\underline{M}^{(i)} \underline{D}^{(i)T} \underline{\tilde{u}}^N) \tag{89}$$

or, since $\underline{M}^{(i)}$ and $\underline{D}^{(i)T}$ are independent of θ

$$\frac{dI_2}{d\underline{\tilde{u}}^n} = \sum_{i=1}^{N} \underline{D}^{(i)} \underline{M}^{(i)} \underline{D}^{(i)T} \frac{d}{d\theta} (\underline{\tilde{u}}^N) \tag{90}$$

If a global mass matrix is defined by

$$\underline{M} \triangleq \sum_{i=1}^{N} \underline{D}^{(i)} \underline{M}^{(i)} \underline{D}^{(i)T} \tag{91}$$

then Equation (90) can be rewritten as

$$\frac{dI_2}{d\underline{\tilde{u}}^N} = \underline{M} \frac{d}{d\theta} (\underline{\tilde{u}}^N) \tag{92}$$

Minimization (actually, the second derivative would have to be checked to prove that minimization occurs) is forced by setting the derivatives in Equations (83) and (92) equal to zero:

$$\frac{dI}{d\underline{\tilde{u}}^N} = \frac{dI_1}{d\underline{\tilde{u}}^N} + \frac{dI_2}{d\underline{\tilde{u}}^N} = \underline{K}\,\underline{\tilde{u}}^N + \underline{M} \frac{d}{d\theta} (\underline{\tilde{u}}^N) = 0 \tag{93}$$

or

$$\underline{M} \frac{d}{d\theta} (\underline{u}^N) = - \underline{K}\,\underline{u}^N \tag{94}$$

The tilde has been dropped since the solution of this set of equations yields the closest approximation in S^N to u .

Equation (94) represents a system of ordinary differential equations. These equations are written in terms of the nodal temperatures, but it must be remembered that the solution is assumed to vary linearly between these values, for it is $u^N(x)$ which is being forced close to $u(x,\theta)$, not just the nodal values of u^N .

The system of equations (94) could be solved directly for a given number of elements, N . If N is large, the system must be solved using finite-differences to approximate the time derivative, or one of the methods suggested by Strang and Fix may be used (Ref 9:244). Three common methods which use finite-differences are the Euler method, the Crank-Nicolson method, and the fully implicit method. In the Euler method, forward differences are used to approximate the time derivative in Equation (94) as

$$\frac{d}{d\theta}(\underline{u}^N)^k = \frac{(\underline{u}^N)^{k+1} - (\underline{u}^N)^k}{\Delta\theta} \tag{95}$$

where the superscript, k , indicates the present time level and $k+1$ indicates the next time level, after a lapse of $\Delta\theta$. It is important to note that, while Equation (95) defines $(\underline{u}^N)^{k+1}$ explicitly in terms of temperatures on the previous time level, the resulting scheme after substitution of Equation (95) into Equation (94) is not truly explicit:

$$\underline{M}\,(\underline{u}^{N})^{k+1} = (\underline{M} - \underline{K}\,\Delta\theta)\,(\underline{u}^{N})^{k} \tag{96}$$

This conclusion is a result of the fact that $\underline{M}$ cannot be inverted without destroying its tridiagonal character. Thus, this "explicit" scheme has no advantage over implicit schemes in the number of operations required to obtain a solution.

A second method is the Crank-Nicolson scheme, which approximates the time derivative in Equation (94) as a central difference at the point $k+1/2$:

$$\frac{d}{d\theta}(\underline{u}^{N})^{k+\frac{1}{2}} = \frac{1}{2}\left(\frac{d}{d\theta}(\underline{u}^{N})^{k+1} + \frac{d}{d\theta}(\underline{u}^{N})^{k}\right) = \frac{(\underline{u}^{N})^{k+1} - (\underline{u}^{N})^{k}}{\Delta\theta} \tag{97}$$

Substitution of Equation (97) into Equation (94) gives

$$\left(\underline{M} + \underline{K}\,\frac{\Delta\theta}{2}\right)(\underline{u}^{N})^{k+1} = \left(\underline{M} - \underline{K}\,\frac{\Delta\theta}{2}\right)(\underline{u}^{N})^{k} \tag{98}$$

In the third method, a backward difference in time, at time level $k+1$, gives

$$\frac{d}{d\theta}(\underline{u}^{N})^{k+1} = \frac{(\underline{u}^{N})^{k+1} - (\underline{u}^{N})^{k}}{\Delta\theta} \tag{99}$$

With this approximation, Equation (94) becomes

$$(\underline{M} + \underline{K}\,\Delta\theta)\ (\underline{u}^N)^{k+1} \quad = \quad \underline{M}\ (\underline{u}^N)^k \tag{100}$$

A general scheme which can generate any of these methods is given by

$$(\underline{M} + \underline{K}\,\alpha\Delta\theta)\ (\underline{u}^N)^{k+1} \quad = \quad (\underline{M} - \underline{K}(1-\alpha)\Delta\theta)\ (\underline{u}^N)^k \tag{101}$$

The system of equations, (101), will satisfy any natural boundary conditions imposed upon it, automatically. However, all essential conditions must be forced. In the case of problem one, the normalized surface temperature, u_1 , is equal to zero except for the initial point in time, at which a step change in temperature occurs. The first and last equation in the system (101) can be modified as follows: The matrices $\underline{A}$ and $\underline{B}$ are defined by

$$\underline{A} \quad = \quad \underline{M} + \underline{K}\,\alpha\Delta\theta \tag{102}$$

and

$$\underline{B} \quad = \quad \underline{M} - \underline{K}\,(1-\alpha)\Delta\theta \tag{103}$$

These matrices are substituted into Equation (101), then the parameter p , called the Fourier modulus, is defined by

$$p = \frac{\Delta\theta}{\Delta x^2} \tag{104}$$

and using matrix notation for the case where the interval has been divided into four elements, Equation (101) becomes

$$\begin{bmatrix} 2+6\alpha p & 1-6\alpha p & 0 & 0 & 0 \\ 1-6\alpha p & 4+12\alpha p & 1-6\alpha p & 0 & 0 \\ 0 & 1-6\alpha p & 4+12\alpha p & 1-6\alpha p & 0 \\ 0 & 0 & 1-6\alpha p & 4+12\alpha p & 1-6\alpha p \\ 0 & 0 & 0 & 1-6\alpha p & 2+6\alpha p \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix}^{k+1}$$

$$= \begin{bmatrix} 2-6(1-\alpha)p & 1+6(1-\alpha)p & 0 & 0 & 0 \\ 1+6(1-\alpha)p & 4-12(1-\alpha)p & 1+6(1-\alpha)p & 0 & 0 \\ 0 & 1+6(1-\alpha)p & 4-12(1-\alpha)p & 1+6(1-\alpha)p & 0 \\ 0 & 0 & 1+6(1-\alpha)p & 4-12(1-\alpha)p & 1+6(1-\alpha)p \\ 0 & 0 & 0 & 1+6(1-\alpha)p & 2-6(1-\alpha)p \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix}^{k} \tag{105}$$

Because of the constant temperature condition on the boundaries, the matrices become

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & a_{22} & a_{23} & 0 & 0 \\ 0 & a_{32} & a_{33} & a_{34} & 0 \\ 0 & 0 & a_{43} & a_{44} & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix}^{k+1}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & b_{22} & b_{23} & 0 & 0 \\ 0 & b_{32} & b_{33} & b_{34} & 0 \\ 0 & 0 & b_{43} & b_{44} & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_B \\ u_2 \\ u_3 \\ u_4 \\ u_B \end{bmatrix} + \begin{bmatrix} 0 \\ b_{21}\, u_B \\ 0 \\ b_{45}\, u_B \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ a_{21}\, u_B \\ 0 \\ a_{45}\, u_B \\ 0 \end{bmatrix} \qquad (106)$$

where the lower case letters represent the value at the indicated position in $\underline{A}$ or $\underline{B}$ and u_B is the imposed boundary temperature. Since, except for the first instant in time, $u_B = 0$, the two column matrices on the right hand side of Equation (106) may be dropped.

For the second problem, the above modification is needed only for the left boundary condition, since the Neumann condition on the right is a natural condition for the method and is satisfied automatically.

Optimum Implicit Condition. For this system of equations, there is an optimum choice for the parameter, α, shown in Figure 12. If α is chosen according to the formula (derived in Appendix D)

$$\alpha = \frac{1}{2}\left(1 + \frac{1}{6p}\right) \tag{107}$$

then the resulting expression in Equation (106) is fourth order accurate at the nodes. That is, the truncation error at the nodes is proportional to $(\Delta x)^4$. The Euler, Crank-Nicolson, and fully implicit schemes are only second order accurate.

Error Analysis

General. One of the objectives of this project was to compare the accuracy of the Crank-Nicolson version of the finite-element method with its counterpart in finite-differences. There are, however, several fundamental differences between finite-elements and finite-differences. First, in the finite-element method, there are two ways to improve the accuracy of the approximate solution. The first is to decrease the size of the interval between the nodal points, Δx. This procedure has a counterpart in finite-differences. In fact, in finite differences, if the limit as $\Delta x \to 0$ is taken for the difference equation, and $p = \Delta\theta/(\Delta x)^2$, then the difference equation should converge to the differential equation in the limit. For the finite-element method, there is an additional procedure for improving the accuracy of the solution. Since the temperature distribution is assumed to have a certain shape within each element,

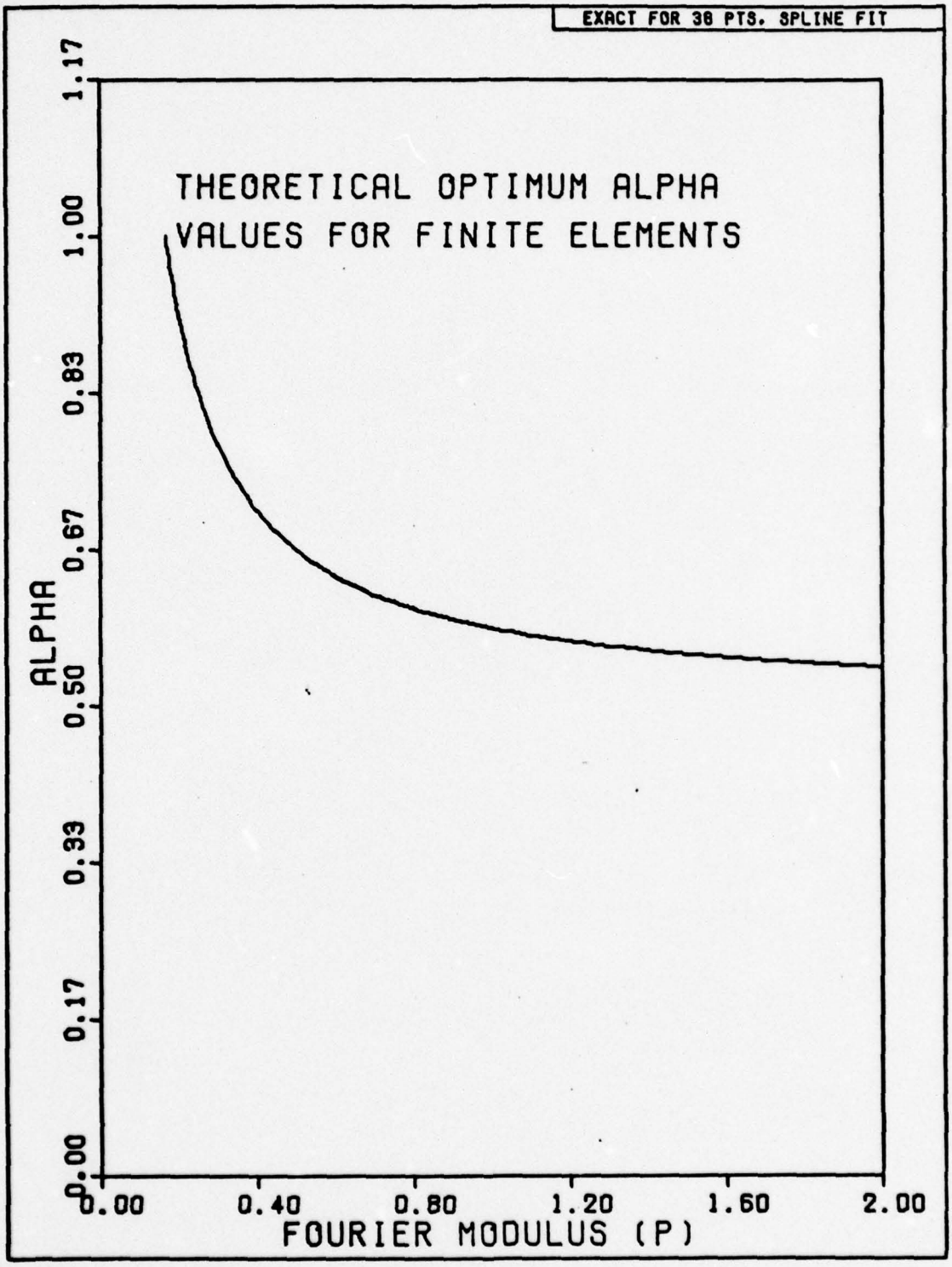


Figure 12. Theoretical Optimum Alpha Values for Finite-Elements.

in the case of a polynomial approximation, a higher order polynomial will give better accuracy. This procedure has no counterpart in finite-differences, in which only the nodes are considered.

Secondly, the error in the finite-element method is most naturally measured over the entire interval rather than in the pointwise sense. The method was intended to minimize the error between the approximate solution and the exact solution in a mean square sense. This fact must be considered when comparing the error in the two methods.

In order to compare linear finite-elements to ordinary finite-differences, a common yardstick must be used. Three error norms were chosen which could be applied to both methods, and one is also used which applies only to finite-elements.

Finite-Difference Error Analysis. The first measure of the error for finite-differences is the pointwise error. The estimate for this error is given by a truncation error analysis. Actually, the pointwise error is a sum of two types of error. The total pointwise error will be referred to as the discretization error. This error is composed of round-off error, which results from carrying only a finite number of significant figures in all of the computations, and truncation error, which results from elimination of the higher order derivatives when Taylor's series is used to approximate a differential equation. Crandall suggests that discretization error can be decreased by using a smaller nodal spacing, Δx , in the difference equation, but warns that if the time interval is also decreased, to maintain the same stability, then more computations

will be required to cover the time domain. Thus, when the number of time steps is large, round-off error, while usually much smaller in magnitude than the truncation error, could become significant. One solution to this problem is to increase the number of significant figures carried in the computation (Ref 12:170). Thus, while the discretization error is actually what is being measured, it is reasonable to consider this error as roughly equivalent to the truncation error. The truncation error is derived in Appendix C for the general expression, Equation (32), and results in the following estimate:

$$e_t(i\Delta x, k\Delta\theta) = c_1 \left[\frac{p}{2} - \alpha p - \frac{1}{12}\right] \Delta x^2 + O(\Delta x)^4 + \ldots \tag{108}$$

where

$$c_1 = \left.\frac{\partial^2 u}{\partial \theta^2}\right|_{\substack{x = i\Delta x \\ \theta = k\Delta\theta}}$$

and $O(\Delta x)^4$ indicates that the first term which is neglected is proportional to $(\Delta x)^4$. This error estimate applies to all nodal points in a Dirichlet problem. In the secondary problem, the Neumann condition

$$\frac{\partial u}{\partial x} = 0 \quad , \quad x = 1 \tag{109}$$

can be analyzed by using the type of truncation error analysis suggested in Appendix C. The truncation error at the right boundary for Equation (37) is thus given by

$$e_t(N\Delta x, k\Delta\theta) = (2\alpha p + \frac{1}{3})\ \Delta x\ \frac{\partial^3 u}{\partial x^3}\Bigg|_{\substack{x = N\Delta x \\ \theta = k\Delta\theta}} \tag{110}$$

where the condition $\partial u/\partial x = 0$ at $x = 1$ must be applied to cancel the lower order terms. This result agrees with Crandall (Ref 12:266) who indicates an order of Δx is lost for each degree in the derivative boundary condition.

The error estimates given above apply to all of the error measures used for the finite-difference method. The first error measurement is called pointwise error. This is simply the total discretization error at a given nodal point. It is defined as the difference between the exact analytical solution and the discrete approximate solution at that point. If the round-off error is assumed to be negligible, this error is equivalent to the truncation error.

The second error measure is defined to be the maximum error between the exact solution and the finite-difference solution taken at any node. In the finite-element literature, the continuous analog of this error measure is called the Tchebycheff norm. Because it

is defined here at nodal points only, it is estimated by the truncation error estimates previously derived.

The last error measure for finite-differences is called here a generalized mean error. Actually, it is nothing more than the sum of the absolute values of the discretization errors at each of the non-zero nodes. This generalized mean was devised to compare the error at the first interior node, $x = .1$, which is measured in the pointwise sense, and the error at all the nodes. The idea here was to investigate whether or not a higher or lower order of convergence would be seen for the first interior node compared to the convergence at all of the nodes. There was still another reason for this error measure which will be discussed in the next section.

In Equation (108), it is interesting to note the order of accuracy for each of the four finite-difference schemes which have been discussed: explicit, Crandall, Crank-Nicolson, and fully-implicit. For the explicit scheme, $\alpha = 0$; therefore

$$e_t(i\Delta x, k\Delta\theta) = \left(\frac{P}{2} - \frac{1}{12}\right)(\Delta x)^2 \left.\frac{\partial^2 u}{\partial x^2}\right|_{\substack{x = i\Delta x \\ \theta = k\Delta\theta}} + O(\Delta x)^4 + \ldots \qquad (111)$$

or second order accurate. For the Crandall method, $\alpha = (1-1/(6p))/2$, and

$$e_t(i\Delta x, k\Delta\theta) = 0(\Delta x)^4 + \ldots \tag{112}$$

For the Crank-Nicolson method, $\alpha = 1/2$, so

$$e_t(i\Delta x, k\Delta\theta) = \left(\frac{-1}{12}\right)(\Delta x)^2 \left.\frac{\partial^2 u}{\partial x^2}\right|_{\substack{x = i\Delta x \\ \theta = k\Delta\theta}} + 0(\Delta x)^4 + \cdots \tag{113}$$

And last, for the pure implicit method, $\alpha = 1$, and therefore,

$$e_t(i\Delta x, k\Delta\theta) = \left(\frac{-p}{2} - \frac{1}{12}\right)(\Delta x)^2 \left.\frac{\partial^2 u}{\partial x^2}\right|_{\substack{x = i\Delta x \\ \theta = k\Delta\theta}} + 0(\Delta x)^4 + \ldots \tag{114}$$

It is interesting to note that, with the exception of the Crandall method which is fourth order accurate, all of the above schemes are second order accurate with respect to truncation error.

The next section will deal with error estimates for finite-elements.

Finite-Element Error Analysis. As in the finite-difference method, the first error measure which will be considered will be the pointwise error. This error measure is alien to the finite-element method because the concept behind the method is to minimize the error everywhere in the region, not just at the nodes. However, error in the pointwise sense can be estimated by treating the individual equations in the system (106) as if they were simple difference equations. In

Appendix D, the truncation error is derived for the ith equation in this system. The result is

$$e_t(i\Delta x, k\Delta\theta) = \left[\frac{p}{2} - \alpha p + \frac{1}{12}\right](\Delta x)^2 \left.\frac{\partial^2 u}{\partial x^2}\right|_{\substack{x = i\Delta x \\ \theta = k\Delta\theta}} + O(\Delta x)^4 + \cdots \qquad (115)$$

This error estimate applies to all nodes for a Dirichlet problem.

For the secondary problem, the truncation error for the Neumann boundary condition is given by

$$e_t(i\Delta x, N\Delta\theta) = 6\alpha p\, \Delta x \left.\frac{\partial^3 u}{\partial x^3}\right|_{\substack{x = N\Delta x \\ \theta = k\Delta\theta}} + \cdots \qquad (116)$$

This order of accuracy is the same as the predicted order of accuracy in the L_2 norm for a derivative boundary condition (Ref 9:118).

The error estimates given in Equations (115) and (116) apply to the pointwise error, the discrete Tchebycheff norm (maximum error at any node), and the generalized mean error (as defined for the finite-difference method). However, as mentioned previously, there are other methods for measuring the error in the finite-element method. The method used here is called the L_2 norm or sometimes the H^o norm. It is defined as follows (Ref 9:5):

$$||u-u^N||_o = [\int_o^1 (u - u^N)^2 \, dx]^{1/2} \tag{117}$$

Since the exact analytical solution for $u(x,\theta)$ is given in terms of an infinite series, some special questions on the existence of $||u-u^N||_o$ must be answered. The computation is involved; therefore, it has been placed in Appendix G. Strang and Fix develop a theorem which bounds the error in the L_2 norm for a piecewise linear finite-element space by (Ref 9:250)

$$||u - u^N||_o \leq C (\Delta x)^2 \tag{118}$$

where the constant C is constant only with respect to the spatial domain, but can be a different value at each point in time. This order of accuracy applies only when the full admissible space of functions, u, is H_E^1, and S^N, the space of all functions u^N, is a piecewise linear subspace of H_E^1.

Stability Analysis

General. Since only a finite number of significant figures can be carried out by the computer in a calculation, every time a calculation is performed, there is the chance of introducing an error. This error is called round-off error. If a series of finite-difference or finite-element computations was carried out using an infinite number of significant figures, the result would be the exact solution of the set of equations. If the magnitude of the difference between this

exact numerical solution and the truncated numerical solution, which would be generated by a computer using a finite number of significant figures, grows exponentially as the calculation proceeds, then the numerical scheme is termed unstable.

There are several methods for treating stability, but by far the best method is one which deals with the complete numerical scheme including the boundary conditions. The method used here is an adaptation of one used by Crandall (Ref 12:382). Crandall, however, indirectly addresses this question using the method of separation of variables to isolate the spatial dependence from the temporal in the case of a parabolic problem. He then analyzes the eigenvalues of the separated spatial system of equations. An eigenvalue greater than one in value will cause steady, unbounded growth in its associated spatial eigenfunction, assuming that eigenfunction was excited by the initial conditions of the problem or was introduced by round-off errors during the computation. In other words, the spatial mode associated with the eigenvalue will be amplified and be of the same sign after each computation. If the value of an eigenvalue lies between zero and one, it will cause steady decay of the corresponding spatial mode. The amplitude of that mode will be smaller in magnitude and of the same sign after each computation. If an eigenvalue has a magnitude between zero and minus one, its associated eigenfunction will be smaller in amplitude, but alternate in sign with each step in the computation. Under these conditions, the solution is said to undergo stable oscillations. Finally, if an eigenvalue has a value less than minus one, then its eigenfunction will undergo unstable oscillations where

the amplitude will increase in magnitude, but have an alternating sign with each step.

Here, the error in the solution is the thing of interest. Round-off errors should not grow exponentially for a stable solution. If Equations (32) and (106) are written in terms of error vectors where the vector $\underline{e}_0$ represents an error introduced by, say, round-off, and $\underline{e}_1$ represents the new error vector after solution of the set of equations, then the following equation can be written:

$$\underline{A}\,\underline{e}_1 = \underline{B}\,\underline{e}_0 \tag{119}$$

or

$$\underline{e}_1 = \underline{C}\,\underline{e}_0 \tag{120}$$

where $\underline{C} = (\underline{A}^{-1}\,\underline{B})$. If $\underline{e}_0$ is expanded in terms of the eigenvectors, $\underline{\phi}_i$, of $\underline{C}$ then

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

where c_i is a constant and $\underline{\phi}_i$ is the ith eigenvector of $\underline{C}$. With the definition of an eigenvalue, Equation (119) can be written

$$\underline{e}_1 = \sum_{i=1}^{n} c_i\, \lambda_i\, \underline{\phi}_i \tag{122}$$

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

$$\underline{e_2} = \underline{C}\,\underline{e_1} = \sum_{i=1}^{n} c_i \lambda_i^2 \underline{\phi_i} \tag{123}$$

and after k computations

$$\underline{e_k} = \sum_{i=1}^{n} c_i \lambda_i^k \underline{\phi_i} \tag{124}$$

This demonstrates that the values of the eignevalues of the iteration matrix, $\underline{C}$, determine the growth or decay of errors just as the eigenvalues of the separated spatial system did before. Separating the variables in the difference equation is not necessary in this case, however. All that is required is to solve for the eigenvalues of the iteration matrix $\underline{C}$.

The stability analysis, as discussed here, will be applied only to the primary problem.

Analysis of the Finite-Difference Formulation. Equation (32) can be written in vector notation as

$$\underline{A}\,\underline{u}^{k+1} = \underline{B}\,\underline{u}^{k} \tag{125}$$

Both $\underline{A}$ and $\underline{B}$ are tridiagonal of the form

$$\begin{bmatrix} b & a & & & \\ a & b & a & & \\ & \circ & \circ & \circ & \\ & & a & b & a \\ & & & a & b \end{bmatrix} \tag{126}$$

Smith (Ref 8:65) gives the eigenvalues of a matrix of this type as

$$\lambda_n = b + 2\,a \cos\left(\frac{n\pi}{N+1}\right) \quad , \quad n = 1,2,\cdots,N \tag{127}$$

where N is the order of the matrix. The iteration matrix, $\underline{C} = \underline{A}^{-1}\,\underline{B}$, in Equation (125) has eigenvalues

$$(\lambda_C)_n = \frac{(\lambda_B)_n}{(\lambda_A)_n} \tag{128}$$

or from Equation (32)

$$(\lambda_C)_n = \frac{(1 + 2\,p\alpha) + 2(-p\alpha)\cos\dfrac{n\pi}{N+1}}{(1 + 2p(1-\alpha)) + 2(p(1-\alpha))\cos\dfrac{n\pi}{N+1}} \tag{129}$$

The analysis will be done for the limiting case where $N \to \infty$, that is, for a large system of equations. The eigenvalues for a system of 40 equations are very close to the limiting values. The eigenvalues of such a system are bounded above by one. The thing of interest here,

then, is the minimum eigenvalue, which is given, for a given p and a given α, by Equation (129) when $n = N$. In the limit, then

$$\lim_{N\to\infty} \cos\left(\frac{N\pi}{N+1}\right) = -1 \tag{130}$$

Thus, the eigenvalue of interest is given by

$$(\lambda_C)_\infty = \frac{1 + 4p\alpha}{1 - 4p(1-\alpha)} \tag{131}$$

which will be termed the critical eigenvalue and where the subscript indicates that $n = N = \infty$. Since α gives the "degree of implicitness" of Equation (32), the entire family of finite-difference expressions can be investigated. Figure 13 shows graphically the results of this analysis for the pure-implicit, Crank-Nicolson, Crandall, and explicit formulations. Table 1 gives the values for the Fourier modulus, p, which will cause oscillation or instability in each scheme. The values given here are the same as those given by Crandall using the separation of variables technique (Ref 2:319).

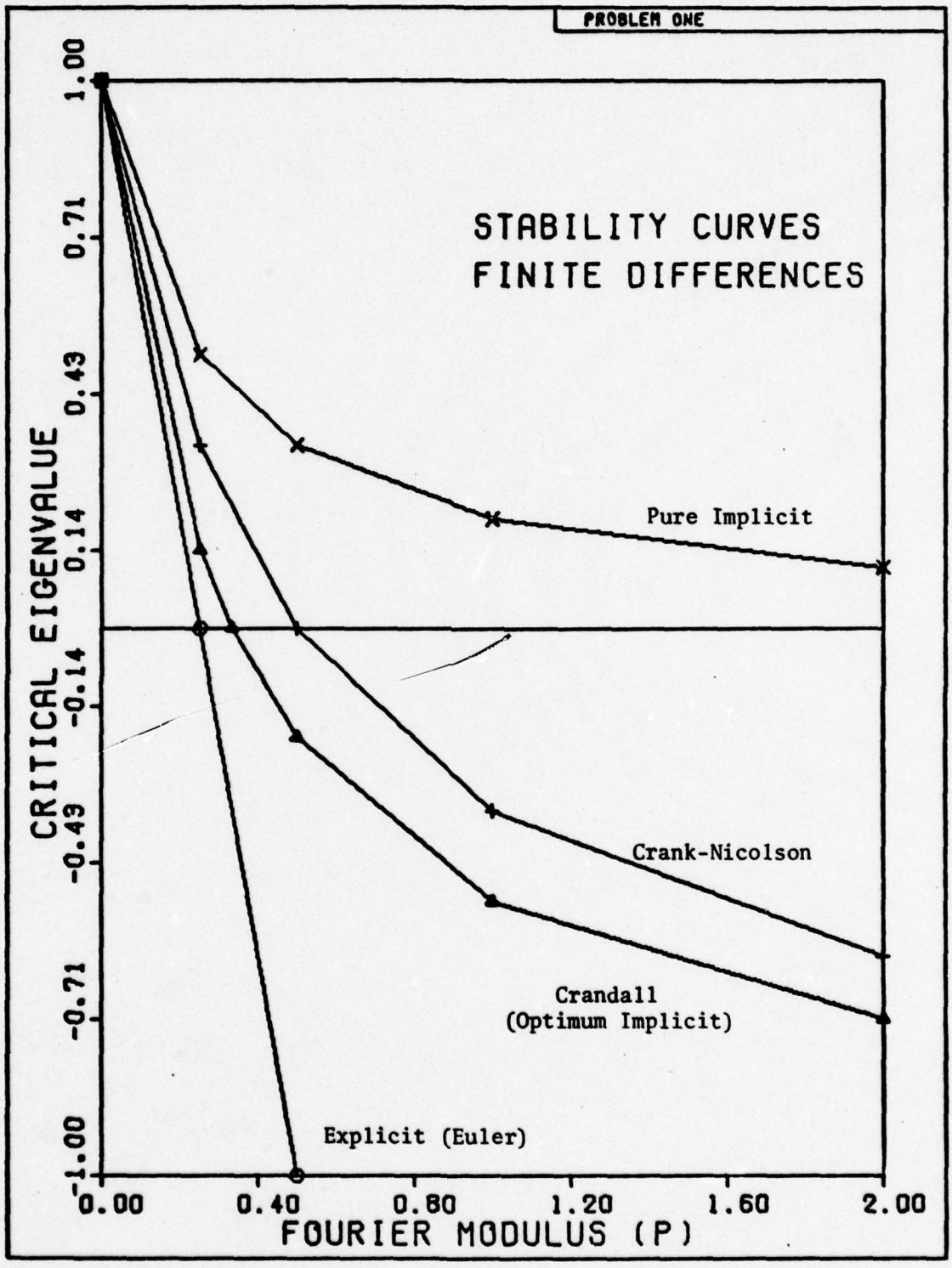


Figure 13. Stability Curves for Finite-Differences.

TABLE I

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

Analysis of the Finite-Element Formulation. Equation (106), where $u_B = 0$, can similarly be written as

$$\underline{A}\,\underline{u}^{k+1} = \underline{B}\,\underline{u}^{k} \tag{132}$$

In this formulation, the first and last equations in this system can be dropped. The number of unknowns can be reduced by two by eliminating the known boundary values, u_B. All of these changes can be made without altering the numerical solution. If Equation (132) represents the new system after these changes have been made, then the resulting system (132) can be analyzed exactly as was done for the

finite-difference case. Again, both $\underline{A}$ and $\underline{B}$ are tridiagonal of the same form as Equation (126). If the same rationale which was previously used in the finite-difference case is applied here, then the critical eigenvalue of the iteration matrix $\underline{C} = \underline{A}^{-1}\,\underline{B}$ is given by

$$(\lambda_C)_\infty = \frac{1 + 12\,p\alpha}{1 - 12p(1-\alpha)} \tag{133}$$

Figure 14 shows the relationship between the various finite-element schemes. Table II gives the oscillation and instability limits in terms of the Fourier modulus, p . A discussion of the stability curves and limits will be reserved for the results section.

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

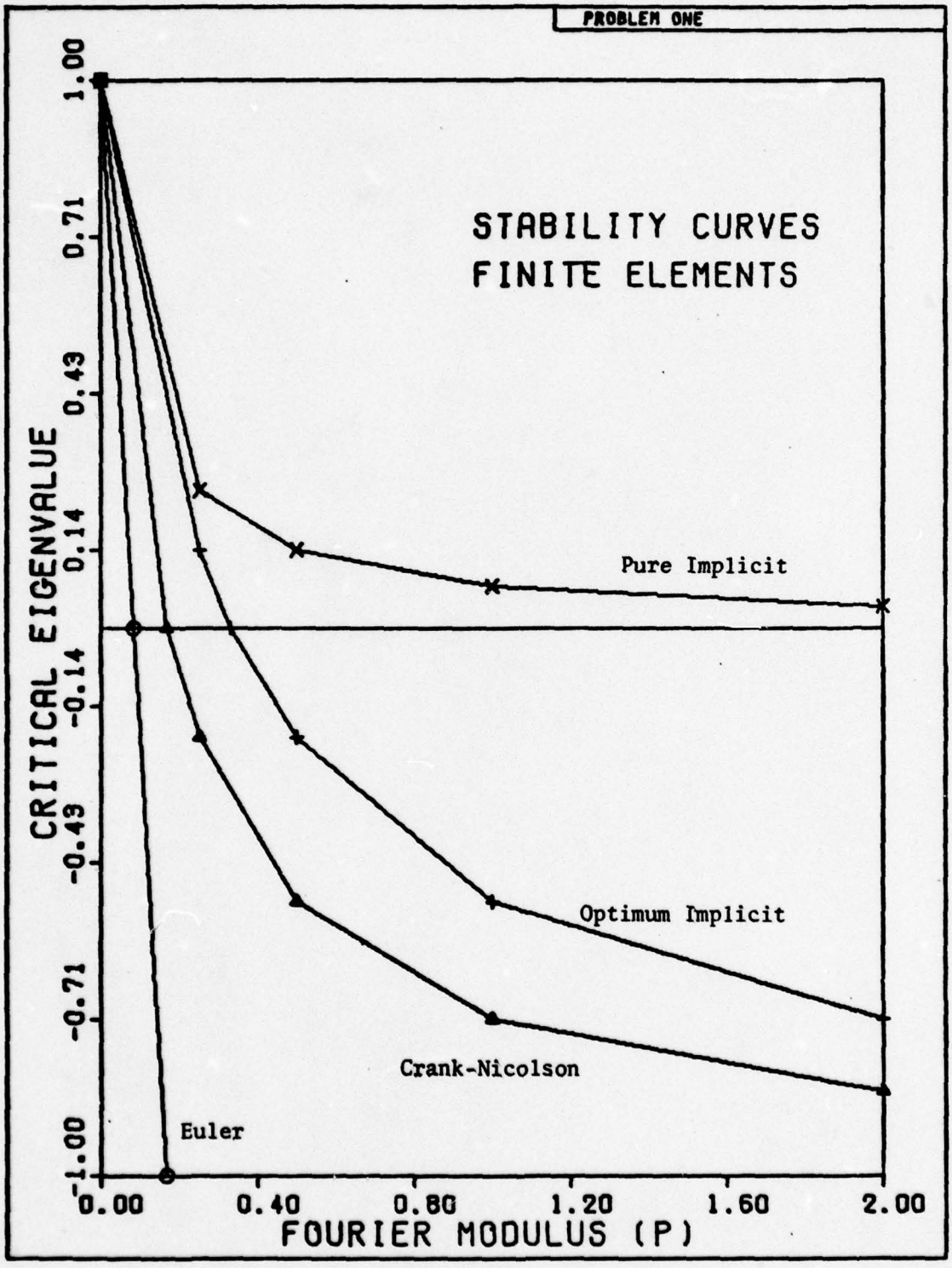


Figure 14. Stability Curves for Finite Elements.

III. Procedure

Approach

Initial Phase. The first step in this investigation began with the programming of the finite-difference method. This phase was relatively easy since the necessary background for this method had been obtained in previous course work. Concurrently, a literature search was initiated along with a study of conduction heat transfer, and a study of the basic principles of the finite-element method. Meyers (Ref 3) was the text for the initial study of the finite-element method. His treatment was on an introductory level; a more complete treatment of the calculus of variations was found in Forray (Ref 14).

Second Phase. This phase began with the programming of the finite-element method. The study of the finite-element method was continued during this phase, as indeed it continued throughout the entire project. Initial results collected during this phase were encouraging in that they indicated that there may be a point of minimum error for the method. The difficulty lay in how to predict this condition. The solution of the problem is a function of x, θ, p, Δx, $\Delta\theta$, and α. Not only that, but the error could be measured in a number of ways. This meant that there was an enormous amount of data to be analyzed and a number of ways to analyze it. As the sifting of this data took place, various methods of presentation of the data were considered. It was decided that a method used by Campbell, Kaplan, and Moore (Ref 4:325) would provide

the most information concerning convergence of the numerical solution to the true solution as Δx becomes smaller. It should be mentioned that, as Δx gets very small, round-off error begins to dominate the solution such that there exists a value of Δx, for a given number of significant figures carried in calculation, beyond which the approximate solution gets less accurate. Campbell, et. al., defined the discretization error ratio as the ratio of the 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. For the L_2 norm in finite-elements, it is somewhat of a misnomer since the assumed solution is not discrete, but continuous; however, the terminology is retained and is understood to be the error ratio over the entire spatial domain after the number of elements has been doubled. At about this point in time, it was discovered that by making a simple truncation error analysis of the finite-element equations, as if they were simple difference equations, and setting the coefficient of the $0(\Delta x)^2$ term in the resulting expression equal to zero, that an $0(\Delta x)^4$ scheme could be obtained. This was very significant since the optimum value of the parameter α could be predicted rather than estimated from empirical results. Indeed, such a large amount of data would be needed to find the optimum condition as a function of all the variables in the problem, that the search was sure to be very time consuming and perhaps even inconclusive. It turns out that the optimum value of α is a function of the Fourier modulus, p, only, and is invariant with respect to the other five variables, at least in theory.

Third Phase. The third phase of the investigation began with a study of error in both the finite-difference and finite-element methods. More emphasis was placed on finite-elements for which the error norms are more involved. Next, the question of stability was investigated. Of the methods considered, Von Neumann, Dusinberre, and matrix eigenvalue, one method stood out as being straightforward and, in fact, more precise. This was the matrix eigenvalue method. Of course, the other methods are useful if the eigenvalues of the iteration matrix are too difficult to obtain.

Also during this phase, the results of the primary problem were analyzed. Once the optimum value of α was determined by theory, computer calculations, which investigated the behavior of the solution in the neighborhood of the optimum value of α, were begun. The initial results of this investigation were disappointing since they showed little improvement in the discretization error ratio, or rate of convergence, for the predicted optimum value of α. The source of the difficulty, it was discovered, was the discontinuity between the initial condition and the boundary conditions. This problem is discussed by Smith (Ref 8:48-49) and by Campbell, Kaplan, and Moore (Ref 4:325-326). Smith suggests two methods for handling this problem. The first involves a transformation of the independent variables from (x,θ) to (X,Θ) where

$$X = x(\theta)^{-1/2} \tag{134}$$

$$\Theta = \theta^{1/2} \tag{135}$$

The result of this transformation is an expansion of the origin (0,0) onto the positive side of the new X axis while the old x axis is concentrated at a point at infinity on the X axis. The jump condition at the origin has been transformed into a continuous change defined over the positive side of the X axis. Smith suggests, also, that an alternative would be to calculate an analytical solution which is continuous in the neighborhood of the jump condition. A third approach, subdividing the space time grid for the first time step as suggested by Campbell, et. al., is the one chosen for this investigation. This last method was chosen because of the ease with which it could be incorporated into the computer programs written during phases one and two.

It was felt that the merits of this investigation lay not in overcoming the problem with the discontinuity, but rather in predicting and verifying the existence of an optimum value of the parameter α for the finite-element method. For this reason, much of the investigation was concerned with a modification of both the primary and secondary problems. The modification was made by simply substituting the exact analytical solution after the first time step. In effect, this transformed the original problem into a new problem in which no discontinuity existed between the initial condition and the boundary conditions. After this transformation, the discretization error ratio was found to approach the predicted values at the nodes.

There was another method which was used in an attempt to reduce the effect of the discontinuity. It was pointed out that, since the constant initial temperature in problem one was represented by setting

u_B equal to zero in Equation (106), that this method underestimates the initial temperature distribution. Since the coefficients of b_{21} and b_{45} in the first column matrix in Equation (106) should more properly be equal to unity for the first time step, it was thought that a calculation made using that modification would give better convergence rates that the original version. This modification was used only for the first time step after which the original scheme was used. The results are shown in Figures 15 and 16 for two different points in time. It can be seen that the effect is to overestimate the temperature distribution. This modification was not pursued any further.

Fourth Phase. The last phase of the project began with a reprogramming of the two methods to solve the secondary problem. After that was completed, the next step, of course, was to analyze the results from the second problem for both finite-elements and finite-differences. The same types of problems which were encountered with the primary problem were also encountered with the secondary problem and were handled in the same way.

Computer System and Programs

Computer. The computer system used for this project is one designed by the Control Data Corporation. It consists of input and output devices, peripheral processors, and two central processors. The central processors are a CDC 6613 and a CDC CYBER 74 which operate in parallel. Each has 131,000 60-bit words of central memory. Magnetic disc and drum storage were used as temporary storage devices.

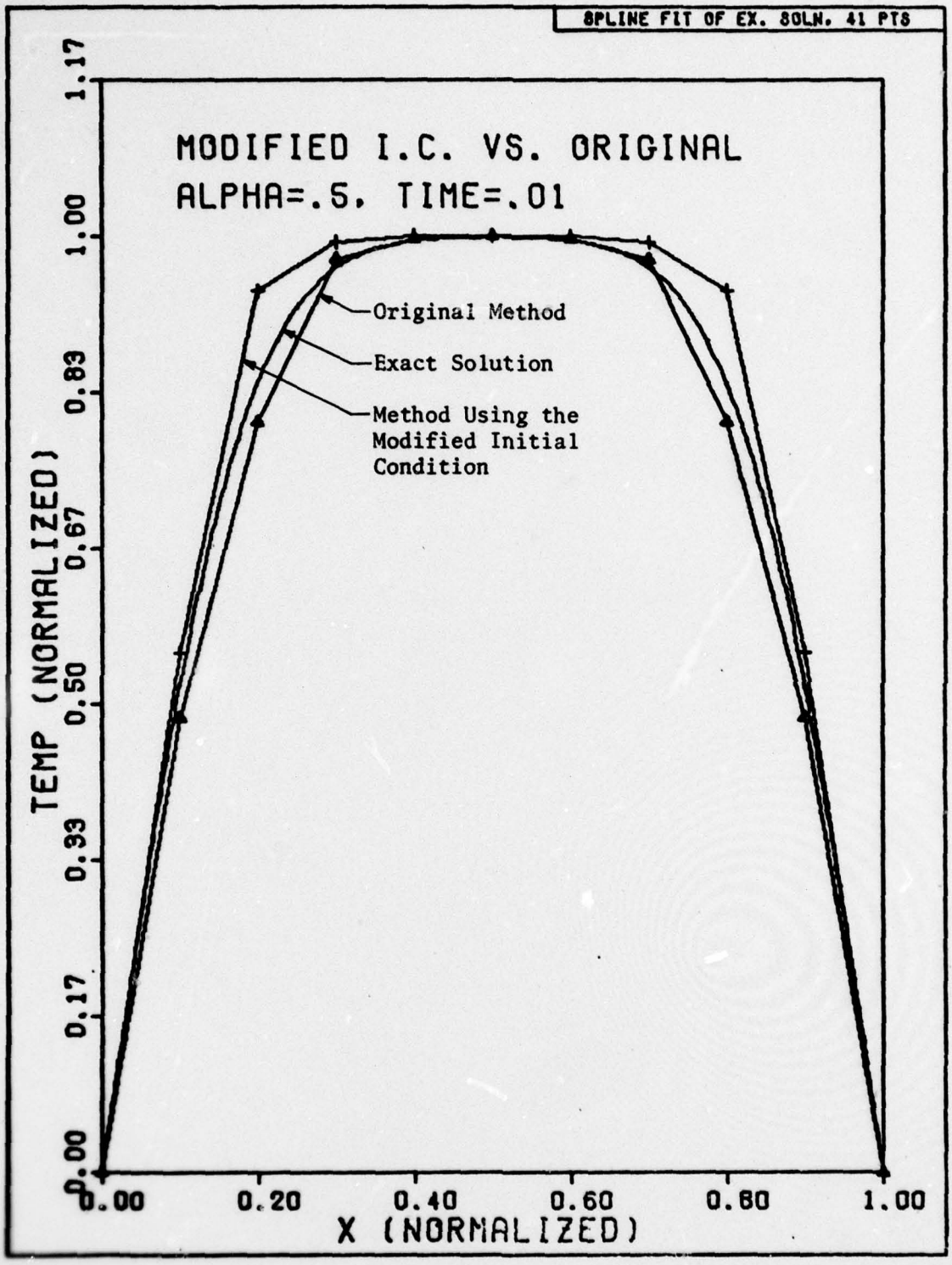


Fig. 15. The Exact Analytical Solution Compared to the Numerical Solution Using the Original and Modified Approximations for the Initial Conditions at θ = .01 for the Crank-Nicolson Version of the Finite-Element Formulation.

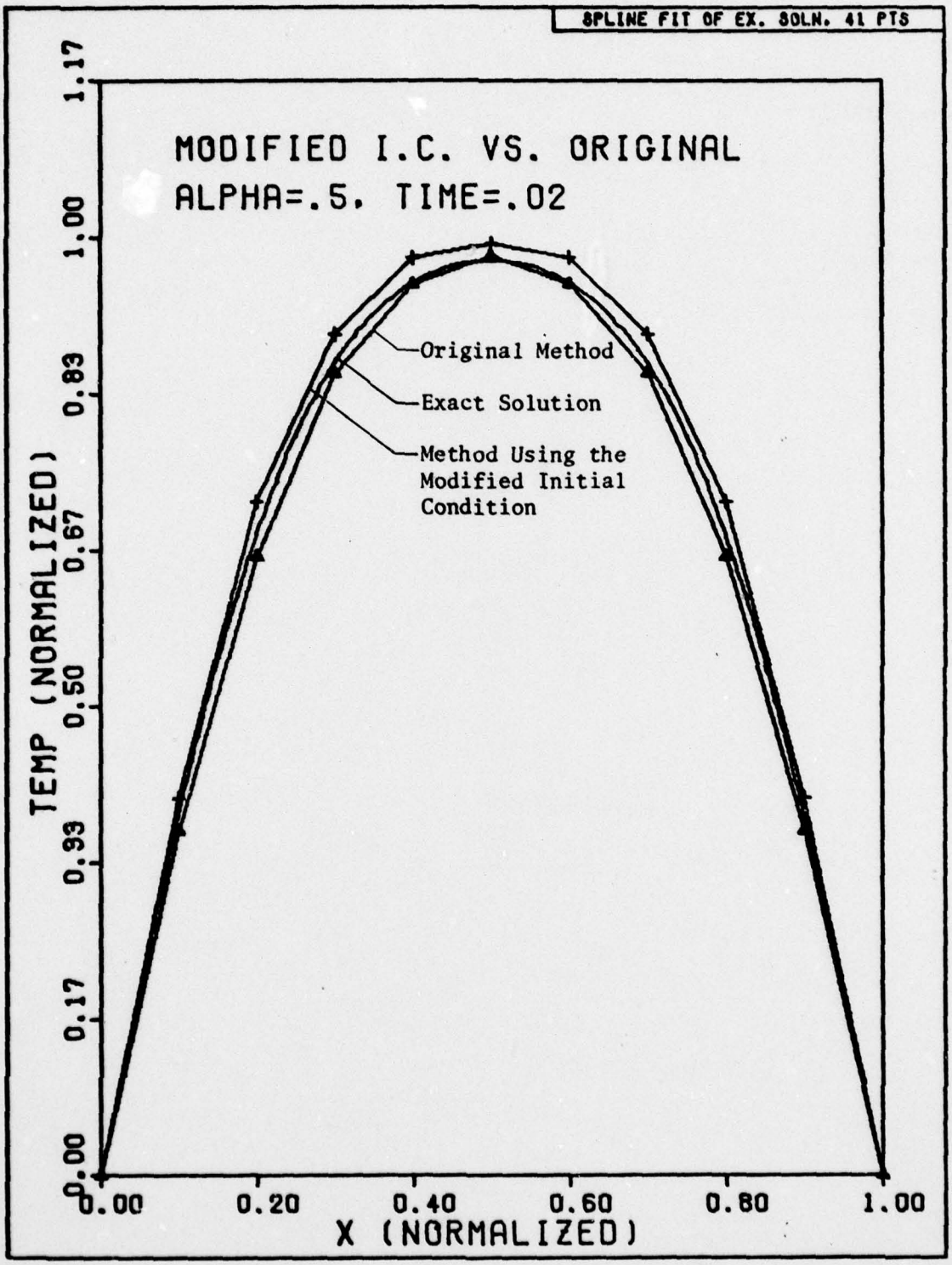


Fig. 16. The Exact Analytical Solution Compared to the Numerical Solution Using the Original and Modified Approximations for the Initial Conditions at θ = .02 for the Crank-Nicolson Version of the Finite-Element Formulation.

Computer Programs. Four major programs were written during this investigation. Two finite-difference programs and two finite-element programs. There were also numerous other programs which were written for stability analysis and plotting purposes. The programs for the secondary problem were primarily modifications of the ones written for the primary problem. The language which was used for all the programs was FORTRAN EXTENDED.

Cost and computer run time are certainly of interest in most computer work, but because the thrust of this investigation was in the direction of investigating the behavior and accuracy of the finite-element solution as a function of the parameter α, there was no effort to compare cost and run time for various options. The main reason for this was due to the fact that the programs which were written were designed for generality and make use of mass data storage on temporary files. For these reasons, an accurate comparison of run time and cost could not be made. However, except for the explicit finite-difference scheme, all of the methods considered were implicit with tridiagonal matrixes and thus should have been roughly comparable with respect to computation time. Also, modifications would need to be made for production codes. Certainly, the user would need to store only two columns of data to represent the symmetric tridiagonal matrixes which are generated by the two methods as used here. Flow charts are given in Figures 17 and 18 for the primary problem using finite-differences and finite-elements, respectively.

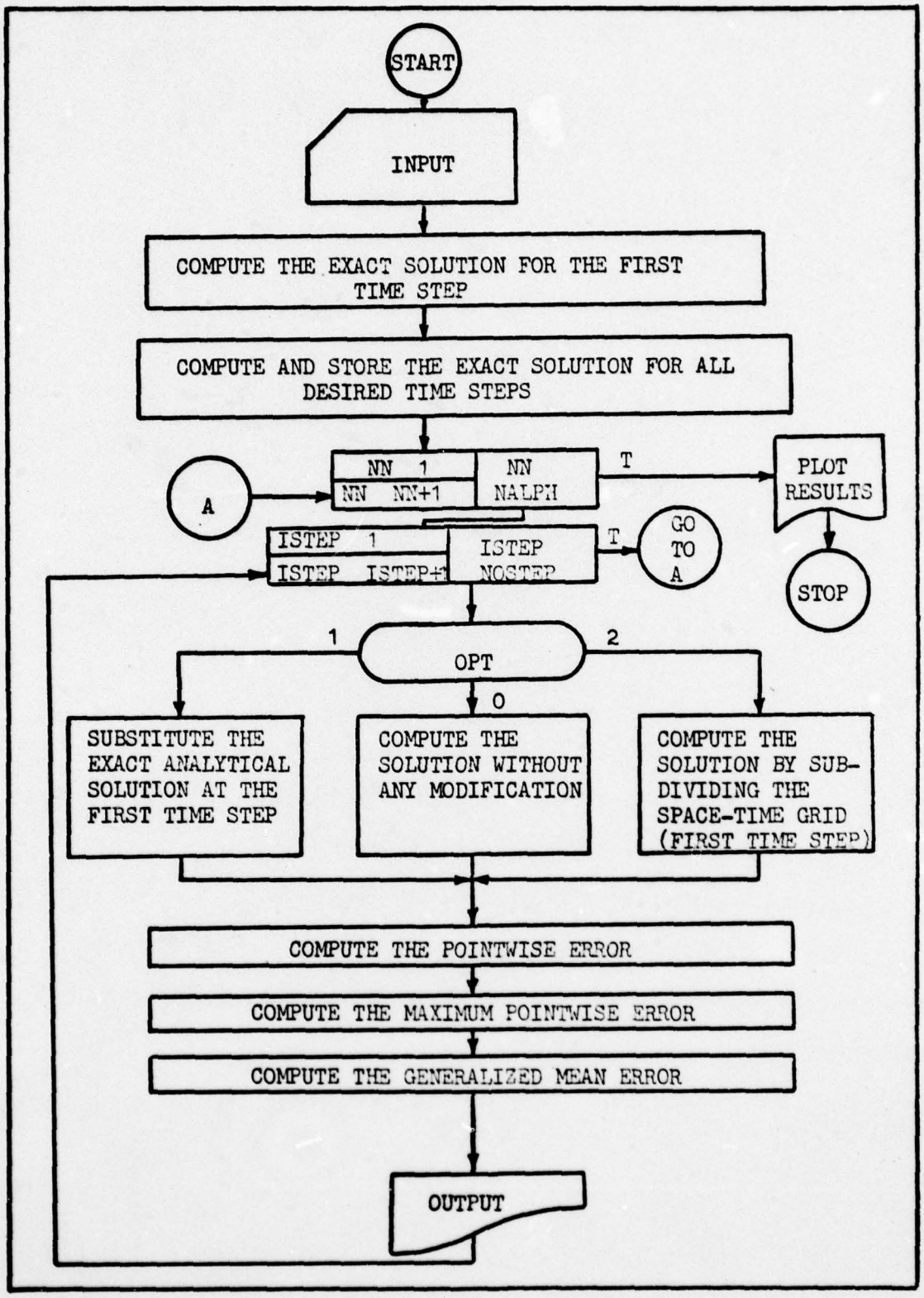


Figure 17. Finite-Difference Computer Program Flow Diagram.

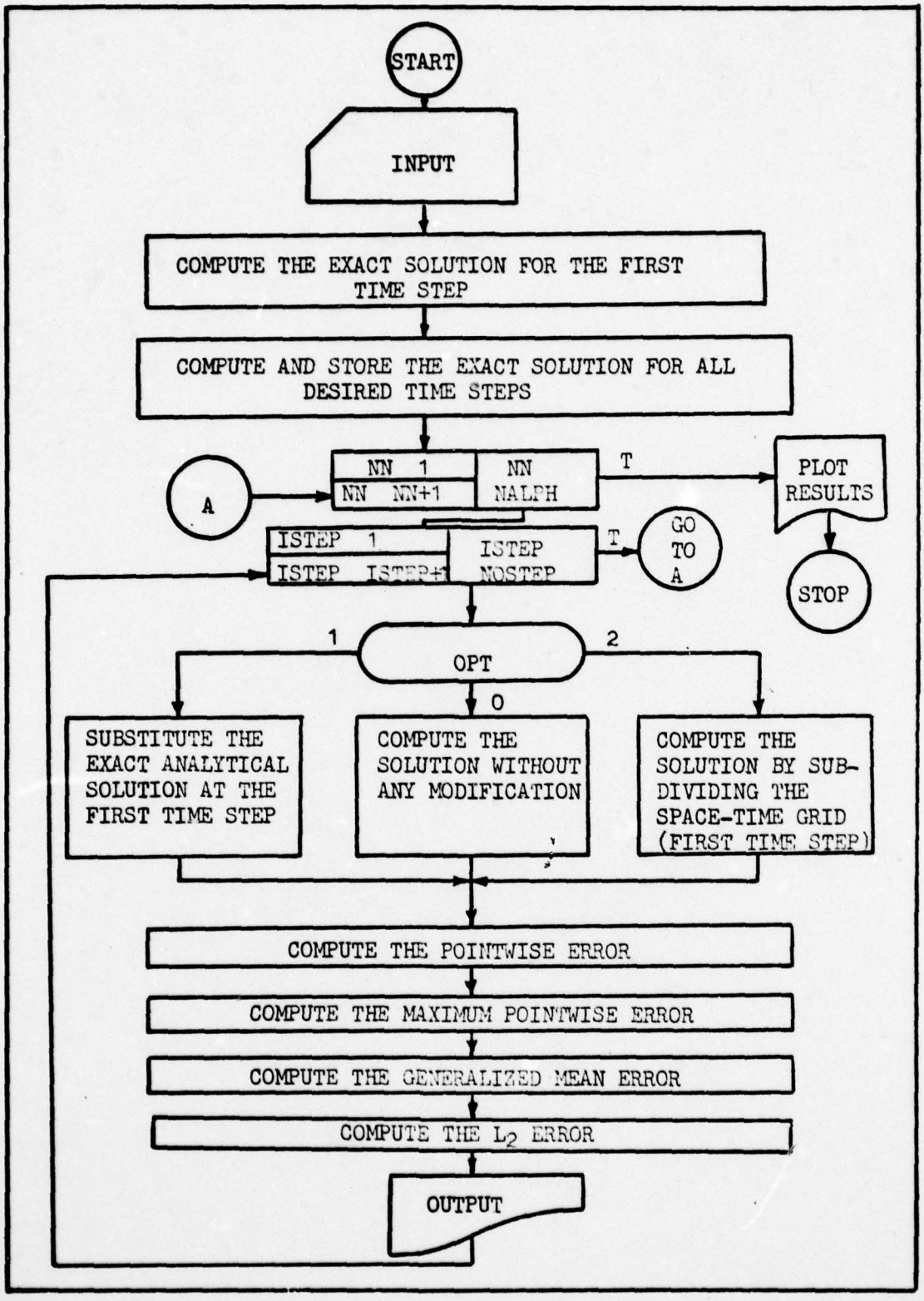


Figure 18. Finite-Element Computer Program Flow Chart.

IV. Results

Stability Analysis

The results of the stability analysis have already been presented in graphical and tabular form in Chapter II. It will be pointed out here, however, that it is very interesting to note that the stability curve for the optimum implicit finite-difference scheme proposed by Crandall coincides exactly with the optimum implicit finite-element scheme proposed here. Further, it is interesting to note that, while the Crandall method is less stable than the Crank-Nicolson method for finite-differences, the optimum implicit method suggested for finite-elements is more stable than the Crank-Nicolson method for finite-elements. In general, the finite-element method is more restrictive with respect to stability; that is, the oscillation limits and instability limits occur for smaller values of the Fourier modulus, $p = \Delta\theta/(\Delta x)^2$.

Error Analysis

The results for the error analysis occupy the bulk of the discussion in this chapter. A very large number of plots have also been generated during the investigation and have been placed in Appendix H for reference. These plots are offered as proof for the theories set forth in this paper and only the trends will be identified. Specific comments on each plot will not necessarily be made.

First, one of the objectives of this research was to compare the finite-element method with the finite-difference method with respect

to accuracy. Tables III through VI give the results of this analysis for two points in time for the case where the space domain is divided into 20 intervals. Surely, a more complete picture could be presented, and the data is available; but the real thrust of this project was in verifying the new theories developed.

Table III

Error Comparisons for the Various Methods
for θ = .04 and p = 1.0 in the Primary Problem.

Method*	Pointwise Error, x = .1	Maximum Error at Any Node	Generalized Mean Error	L2 Error Norm (FE)
CNM-FD	7.7494×10^{-6}	1.4877×10^{-3}	6.0711×10^{-3}	———
CNM-FE	1.7696×10^{-2}	2.6208×10^{-3}	1.8231×10^{-2}	3.1957×10^{-3}
OIM-FD	8.7106×10^{-4}	1.5579×10^{-3}	1.2018×10^{-2}	———
OIM-FE	8.7106×10^{-4}	1.5579×10^{-3}	1.2018×10^{-2}	2.6011×10^{-3}
FIM-FD	-5.60758×10^{-3}	7.1078×10^{-3}	3.6781×10^{-2}	———
FIM-FE	-3.6498×10^{-3}	4.5849×10^{-3}	2.4720×10^{-2}	2.5395×10^{-3}

*CNM = Crank-Nicolson Method
OIM = Optimum Implicit Method
FIM = Fully Implicit Method
FD = Finite-Differences
FE = Finite-Elements

Table IV

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

Method*	Pointwise Error, x - .1	Maximum Error at Any Node	Generalized Mean Error	L_2 Error Norm (FE)
CNM-FD	6.0493×10^{-5}	3.0606×10^{-4}	1.6801×10^{-3}	———
CNM-FE	7.1111×10^{-3}	2.0911×10^{-3}	1.3683×10^{-2}	2.3863×10^{-3}
OIM-FD	3.8710×10^{-4}	1.1952×10^{-3}	7.6779×10^{-3}	———
OIM-FE	3.8710×10^{-4}	1.1952×10^{-3}	7.6779×10^{-3}	1.7356×10^{-3}
FIM-FD	-1.9654×10^{-3}	4.8803×10^{-3}	3.4197×10^{-2}	———
FIM-FE	-1.2764×10^{-3}	3.1811×10^{-3}	2.2256×10^{-2}	1.6441×10^{-3}

*See Table III

Table V

Error Comparisons for the Various Methods for $\theta = .04$ and $p = 1.0$ in the Primary Problem and where the Exact Analytic Solution has been substituted at the First Time Step.

Method*	Pointwise Error, x = .1	Maximum Error at Any Node	Generalized Mean Error	L_2 Error Norm (FE)
CNM-FD	-6.8464×10^{-4}	9.4432×10^{-4}	5.1987×10^{-3}	———
CNM-FE	9.3977×10^{-4}	1.2464×10^{-3}	6.4337×10^{-3}	1.9695×10^{-3}
OIM-FD	1.3598×10^{-4}	1.6546×10^{-4}	1.0025×10^{-3}	———
OIM-FE	1.3598×10^{-4}	1.6546×10^{-4}	1.0025×10^{-3}	1.4031×10^{-3}
FIM-FD	-5.9604×10^{-3}	7.8714×10^{-3}	4.1075×10^{-2}	———
FIM-FE	-4.1337×10^{-3}	5.5091×10^{-3}	2.8901×10^{-2}	2.7448×10^{-3}

*See Table III

Table VI

Error Comparisons for the Various Methods for $\theta = .08$ in the Primary Problem and where the Exact Analytical Solution has been substituted at the First Time Step.

Method*	Pointwise Error, x = .1	Maximum Error at Any Node	Generalized Mean Error	L_2 Error Norm (FE)
CNM-FD	-3.0351×10^{-4}	8.5043×10^{-4}	5.6708×10^{-3}	———
CNM-FE	3.2920×10^{-4}	8.8078×10^{-4}	5.9831×10^{-3}	1.5442×10^{-3}
OIM-FD	1.4103×10^{-5}	2.1640×10^{-5}	1.5280×10^{-4}	———
OIM-FE	1.4103×10^{-5}	2.1640×10^{-5}	1.5280×10^{-4}	9.3405×10^{-4}
FIM-FD	-2.2702×10^{-3}	5.8835×10^{-3}	4.0493×10^{-2}	———
FIM-FE	-1.6021×10^{-3}	4.2338×10^{-3}	2.8906×10^{-2}	2.3705×10^{-3}

*See Table III

The most striking feature of Tables III through VI is that the optimum implicit methods for both finite-differences and finite-elements give identical accuracy. In all four tables, the finite-element method shows the optimum implicit method to be more accurate than the Crank-Nicolson or the fully implicit methods; while, in the finite-difference method, Tables III and IV (discontinuous initial conditions) show greater accuracy for the Crank-Nicolson method, but Tables V and VI, in which the initial and boundary conditions have been matched, show greater accuracy for the optimum implicit method. The Crank-Nicolson method appears to be invariably more accurate for finite-differences than for linear finite-elements. For the fully implicit method, the finite-element version was the most accurate in each case.

The large number of plots generated during this investigation will be analyzed in the order in which they appear in Appendix H. This appendix is divided into four sections. The first deals with the primary problem as solved by finite-differences. The second deals with the primary problem as solved by finite-elements. The third presents the results of the secondary problem as solved by finite-differences. And the fourth deals with the secondary problem as solved by finite-elements. The introductory remarks in the appendix explain the rationale for the use of the various error norms and graphical formats. A special note is in order, however, on the interpretation of the discretization error ratio. If the error for some norm is given by

$$e = \xi(\Delta x)^2 \tag{136}$$

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

$$\frac{e_1}{e_{\frac{1}{2}}} = \frac{\xi_1(\Delta x)^2}{\xi_{\frac{1}{2}}(\frac{\Delta x}{2})^2} \tilde{=} 4 \tag{137}$$

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

$$e = \xi(\Delta x)^4 \tag{138}$$

is

$$\frac{e_1}{e_{\frac{1}{2}}} = \frac{\xi_1(\Delta x)^4}{\xi_{\frac{1}{2}}(\frac{\Delta x}{2})^4} \simeq 16 \qquad (139)$$

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

Section I Results. This section gives the results of the primary problem as solved by the finite-difference method. The first set of graphs in this sections, Figures H-1 through H-15, show the order of convergence for the case $p = 0.5$ when the exact analytical solution has been substituted at the first time step. The first six of these show the discretization error ratio (DER) plotted against a number of values of the parameter α. The peak occurs at or near the value of α predicted by Crandall ($\alpha = .33333$). Because of transient effects on the peak shape and position, the DER was also plotted against time in Figures H-7 through H-12 for three methods: optimum implicit, Crank-Nicolson, and fully implicit. In each error norm, the optimum implicit and the Crank-Nicolson methods approach $0(\Delta x)^2$ in accuracy, while the optimum implicit method approaches $0(\Delta x)^4$ in accuracy. The behavior with respect to time cannot be explained. In Figures H-13 through H-15, the absolute magnitude of the error is seen to dip to a minimum in the neighborhood of the predicted optimum value of α for each error norm.

The next three sets of graphs all deal with results for the case $p = 1.0$. In Figures H-16 through H-21, the numerical solution has not been modified at all. Because of the discontinuity, the convergence rate is not very much enhanced for the optimum value of α ($\alpha = 0.416667$) . These plots also show how instability has begun to affect the solution for small values of α . In Figures H-22 through H-36, the numerical solution has been replaced by the exact analytical solution, and again, the results approach the predicted behavior, especially for the finer discretization. In Figures H-34 through H-36, the absolute magnitude of each error norm can be seen to fall to a minimum in the neighborhood of the predicted optimum value. Figures H-37 through H-42 show that the order of convergence can be improved over the unmodified solution by subdividing the space-time grid for the first time step as discussed previously. The improvement is not dramatic, however. More improvement could be made, perhaps, by using one of the other methods considered in previous chapters.

In the last set of graphs in this section, Figures H-43 through H-57, the exact solution has been substituted at the first time step and the peaks in the DER versus α curves are well defined and fall in the neighborhood of $\alpha = 0.45833$, the predicted optimum for $p = 2.0$. Again, the DER versus time curves show that the DER approaches the predicted values for each of the methods. Finally, Figures H-55 through H-57 indicate that the minimum absolute error occurs near the predicted optimum value of α .

Section II Results. This section deals with the primary problem as solved by finite-elements. The first set of graphs in this section was generated for the case $p = 0.5$. In Figures H-58 through H-77, the exact analytical solution has been substituted at the first time step to eliminate the problem with the discontinuity. Just as in the finite-difference method, each error norm is seen to approximate the predicted behavior. This is perhaps best demonstrated in Figures H-66 through H-73. The L_2 norm in Figures H-64 and H-65 is seen to have a very broad peak, but in terms of order of convergence, there seems to be hardly any advantage for one method over another. Also, as in the finite-difference case, Figures H-74 through H-76 show that error in the pointwise sense has a minimum in the neighborhood of the predicted optimum value, $\alpha = 0.66667$. Figure H-77 shows that the minimum in the L_2 norm is a function of time and occurs at a point close to the fully implicit method.

The next four sets of graphs all give results for $p = 1.0$ where the predicted optimum value of α is 0.58333. In Figures H-78 through H-85, no modifications have been made to correct for the effects of the initial discontinuity. The order of convergence seems at times to have a peak and at times not to have a peak. Any peaks which do appear are only transient and occur for values of α much higher than the predicted optimum value. For Figures H-86 through H-105, the exact analytical solution has been substituted at the first time step. As before when this approach was taken, the predicted results are seen. Also, in Figure H-105, the L_2 norm is seen to have a fairly well defined minimum, but at a value of α which is displaced to the

right of the optimum value. In the next set of plots, Figures H-106 through H-113, an attempt was made to reduce the effect of the initial discontinuity by a subdivision of the space-time grid for the first time step. Any improvement in the order of convergence is only transient, or at best, not very significant. Another method, as suggested by Figures 13 and 14 in Chapter III, is shown in Figures H-114 through H-121. While this method looked promising when it was constructed, the results were very discouraging. These figures show virtually no advantage for any of the stable methods in any error norm.

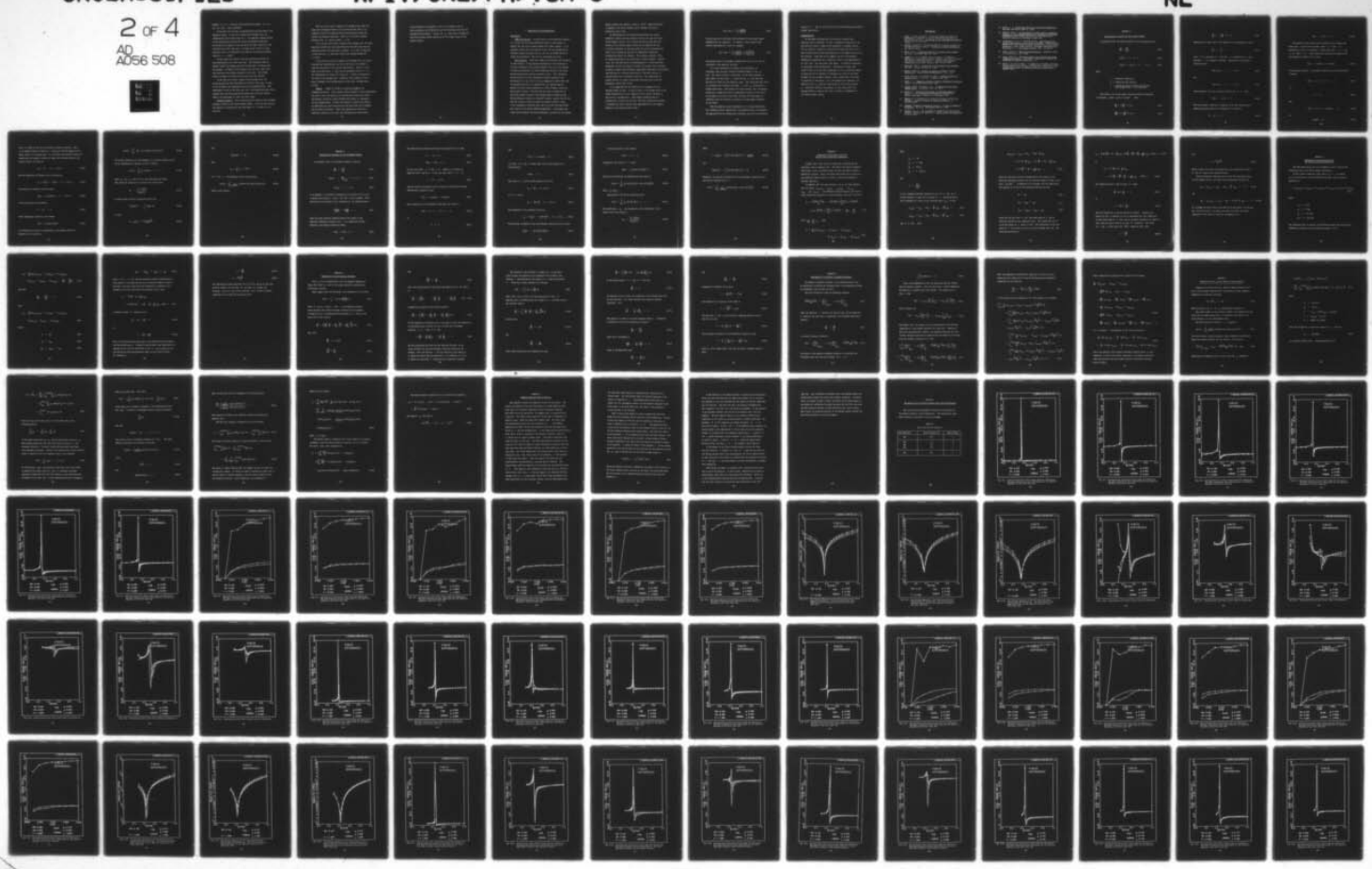
The last set of graphs in this section gives results for the case where $p = 2.0$. In these plots, Figures H-122 through H-141, the exact solution has once again been substituted for the numerical solution at the first time step. As before when this was done, the DER has its peak at the predicted value of α , 0.541667 , and the DER versus time curves indicate that, in the norms which measure pointwise or related error, the predicted convergence rate is approached for each of the three indicated methods. Further, and perhaps even most importantly, the L_2 error norm in Figures H-128 and H-129 shows a peak structure which is very similar to that observed for the pointwise error norms. This is a very encouraging development which will be more thoroughly discussed in the next chapter. Finally, Figures H-138 through H-141 show that the minimum point in each of the error norms is very well defined.

Section III Results. This section deals with the secondary problem as solved by finite-differences. The analysis in this section and the next was limited by time; therefore, only one value of the Fourier

AD-A056 508 AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OHIO SCH--ETC F/G 20/13
AN INVESTIGATION OF THE NUMERICAL METHODS OF FINITE DIFFERENCES--ETC(U)
MAR 78 C R MARTIN

UNCLASSIFIED AFIT/GNE/PH/78M-6 NL

2 OF 4
AD
AD56 508



modulus, $p = 1.0$, and only two discretization schemes, $\Delta x = 0.1$ and $\Delta x = 0.05$, were considered.

In the first set of plots, no modifications have been made to the numerical scheme. In this set, Figures H-142 through H-144, the instabilities in the lower values of α are apparent, perhaps excited by the discontinuity between the initial and boundary conditions since they do not appear in the case where the exact solution is substituted at the first time step. In Figure H-142, there is a peak structure, but it is displaced to the right of the predicted optimum value of α, 0.41667.

In the second set of plots, the exact analytical solution has been substituted at the first time step. This procedure shows the sharper results which appear when the discontinuity is removed from the problem. In Figures H-145 through H-153, the same trends which were observed in the primary problem are seen here. The minimum points in Figures H-152 and H-153 are also very well defined.

In the last set of plots, Figures H-154 through H-156, the space-time grid has been subdivided for the first time step, in order to try to improve the convergence rate for the optimum method. Some improvement is seen in the first two, but in the generalized mean, which computes the error from nodal points spread over the entire spatial domain, the improvement is seen to be only slight.

Section IV Results. This section gives the results of the secondary problem as solved by the finite-element method. As in the last section, the analysis is restricted to the cases where $p = 1.0$ and $\Delta x = 0.1$ and 0.05.

The first set of plots, Figures H-157 through H-160, shows the usual poor convergence behavior of the solution when there is no attempt to correct the problem with the discontinuity between the initial and boundary conditions. There is a very poorly defined minimum in the L_2 norm at around $\alpha = 0.85$.

In the second set of plots, Figures H-161 through H-172, the exact analytical solution has been substituted for the first time step and predicted results once more begin to appear. It is very interesting to note that the L_2 norm has a fairly well defined minimum for $\alpha \tilde{=} 0.675$.

In the last set of plots, Figures H-173 through H-176, the space-time grid was subdivided for the first time step in an attempt to alleviate the effect of the initial discontinuity. There is a nice peak in the DER versus α curve for the pointwise error at $x = 0.1$, but unfortunately it occurs for a value of α which is displaced to the right of the optimum value. Otherwise, there appears to be no great advantage in the optimum method in this case, although there is some.

Summary. A number of trends has appeared throughout the foregoing discussion. First, whether finite-elements or finite-differences are used to solve the problem, the discontinuity between the initial and boundary conditions causes problems with respect to convergence for the optimum method. Second, the attempt to correct this problem by subdividing the space-time grid for the first time step is helpful, but far from satisfactory. Third, the substitution of the exact analytical solution at the first step eliminates the discontinuity

in the problem by transforming it into a new problem, and for this new problem the difficulties with the convergence rate for the optimum method disappear. Fourth, the L_2 norm shows a minimum for the finite-element method especially for the larger values of the Fourier modulus.

V. Conclusions and Recommendations

Conclusions

Stability Analysis. From the results of the stability analysis, it is clear that the finite-difference method is more stable in general than the finite-element method with linear elements. It is certainly curious that the stability curve for the optimum finite-difference method coincides with the one for the optimum method in linear finite-elements. More will be said about this later.

Error Analysis. There are a number of conclusions and inferences in this section. If the problem considered has a discontinuity of the type considered in the primary problem, then of all the methods of solution considered, the finite-difference version of the Crank-Nicolson method gives the most accurate results. This conclusion may not hold true if some acceptable alternate handling of the discontinuity is used, such as the one suggested by Smith (Ref 8:49). For problems which have no singularities, the optimum implicit method, by either finite-differences or finite-elements, gives the most accurate results. The fact that the error is identical for the finite-element version of the optimum implicit method and the finite-difference version, or Crandall's method, coupled with the fact that the stability curves for these two methods coincide, leads to the inescapable conclusion that they are in fact the same method, only derived by two widely different approaches. Said another way, linear finite-elements and finite-differences coincide for the optimum

implicit method and, moreover, share an $0(\Delta x)^6$ scheme as derived in Appendix D for finite-elements and by Crandall for finite-differences (Ref 2:320).

In considering only the Crank-Nicolson method, the finite-difference version was more accurate in every case than the linear finite-element version. But it should be borne in mind that the accuracy of the finite-element version can be enhanced not only by further refinement of the space mesh, but also by increasing the order of the polynomials used for the approximate temperature distributions within the elements. The previous conclusion, then, is largely qualified by the restriction to linear elements. However, since 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 approximate temperature distributions, then there may be a serious restriction on the total increase in accuracy which can be attained in this way. This problem may be solved, however, by using a variational principle such as the one proposed by Gurtin (Ref 11:255).

It is suggested that the problems with convergence for the optimum schemes are due, at least in part, to the weight given to the higher eigenfunctions in the expansion of the initial condition in the case of a discontinuity. These higher eigenfunctions are automatically filtered out and lost, since the Crank-Nicolson method is nothing more than a Padé rational approximation for the temporal behavior of the solution (Ref 17:266):

$$\exp(-\lambda\Delta\theta) \cong \frac{1 - \lambda(\Delta\theta/2)}{1 + \lambda(\Delta\theta/2)} \tag{140}$$

thus giving rise to difficulty with problems for which these higher eigenfunctions are important. If, however, a more accurate Padé rational approximation is used, for example

$$\exp(-\lambda\Delta\theta) \cong \frac{1 - \lambda(\Delta\theta/2) + \lambda^2(\Delta\theta^2/12)}{1 + \lambda(\Delta\theta/2) + \lambda^2(\Delta\theta^2/12)} \tag{141}$$

then perhaps fewer of the higher eigenfunctions will be lost and can contribute to the numerical solution.

It is very interesting to surmise on the possibility of attaining a more accurate solution for finite-elements in the L_2 norm. The results indicate a trend that, as the Fourier modulus p gets larger, the DER versus α curves for the L_2 norm look more and more like those for the pointwise norms. This would mean that there is a minimum error condition for large p in the L_2 norm which is perhaps predictable. This would be of great benefit, since an engineer using such a scheme would have an approximate solution at every point on the domain, not just at the nodes, and know with some degree of confidence that the solution is accurate to a high degree everywhere on the domain.

From the numerous plots in Appendix H, it is clearly shown that, for a problem without singularities, the predicted rates of convergence are approached and the minimum error condition occurs for the predicted

values of α . This is true for both the finite-difference and finite-element formulations.

Recommendations

As with many investigations, this project uncovered more questions than it answered. For one, it would be interesting to see the result of using a higher-order polynomial, or perhaps splines, for the elemental temperature distribution coupled with the use of a more accurate Padé rational approximation for the temporal behavior of the solution. Varga (Ref 18:70) suggests the use of Tchebycheff semidiscrete approximations to spread the error in the approximation of $\exp(-\lambda\Delta\theta)$ over the entire time domain. It would be interesting to investigate these modifications. Additionally, further work could be done in two and three dimensions to try to extend the method. It would also be of interest to investigate the error at the right boundary for the secondary problem, since a lower order of accuracy is predicted at that point. Last, and probably most important, the behavior of the L_2 norm for large values of the Fourier modulus, p , should be carefully investigated, as this could lead to an optimum method in terms of this norm, one which is fundamental to the finite-element concept.

Bibliography

1. Crank, J. and P. Nicolson. "A Practical Method for Numerical Evaluation of Solutions of Partial Differential Equations of Heat-Conduction Type." Proceedings of the Cambridge Philosophical Society, 43: 50-67 (1947).

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

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

4. Campbell, R. C., B. Kaplan, and A. H. Moore. "A Numerical Comparison of the Crandall and the Crank-Nicolson Implicit Methods for Solving a Diffusion Equation." Journal of Heat Transfer, Transactions of the ASME, Series C, 88: 324-326 (1966).

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

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

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

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

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

10. Washizu, K. Variational Principles in Continuum Mechanics. Report No. 62-2, Department of Aeronautical Engineering. Seattle, Washington: University of Washington, 1962.

11. Gurtin, M. E. "Variational Principles for Linear Initial Value Problems.: Quarterly Journal of Applied Mathematics, 22: 252-256 (1964).

12. Crandall, Stephen H. Engineering Analysis: A Survey of Numerical Procedures. New York: McGraw-Hill Book Co., 1956.

13. Douglas, Jim, Jr. "On the Numerical Integration of Quasi-Linear Parabolic Differential Equations." Pacific Journal of Mathematics, 6: 35-42 (1956).

14. Forray, J. J. Variational Calculus in Science and Engineering. New York: McGraw-Hill Book Co., 1968.

15. Campbell, Bart C. An Investigation of the Accuracy of Numerical Solutions of the Diffusion Equation for Transient Heat Transfer. Unpublished thesis, Wright-Patterson Air Force Base, Ohio: Air Force Institute of Technology, August 1964.

16. Yalamanchile, R. V. S. and Shih-Chi Chu. Application of the Finite Element Method to Heat Transfer Problems. Technical Report RE TR 71-41. Rock Island, Illionis: Research Directorate Weapons Laboratory at Rock Island Research, Development and Engineering Directorate, U. S. Army Weapons Command, 1971. AD 726 371

17. Varga, Richard S. Matrix Iterative Analysis. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1962.

18. Varga, Richard S. Functional Analysis and Approximation Theory in Numerical Analysis. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics, 1971.

19. Moore, A. H. and B. Kaplan. "A Comparison of Crandall and Crank-Nicolson Methods for Solving a Transient Heat Conduction Problem." International Journal for Numerical Methods in Engineering, 9: 938-943 (1975).

Appendix A

The Analytical Solution for the Primary Problem

In normalized form, the primary problem in this investigation was

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

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

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

where

u = normalized temperatue

θ = normalized time variable

x = normalized spatial variable (the bar has been dropped from x for simplicity)

The solution will be developed, using the method of separation of variables. Assume $u(x,\theta) = X(x)\Theta(\theta)$. Then

$$\frac{\partial u}{\partial x} = \Theta \frac{dX}{dx} \triangleq \Theta X' \tag{A-4}$$

$$\frac{\partial^2 u}{\partial x^2} = \Theta \frac{\partial^2 X}{dx^2} \triangleq \Theta X'' \tag{A-5}$$

$$\frac{\partial u}{\partial \theta} = X \frac{d\Theta}{d\theta} \triangleq X \Theta' \qquad \text{(A-6)}$$

Substitution of these results into Equations (A-1) through (A-3) gives

$$\frac{X''}{X} = \frac{\Theta'}{\Theta} = \gamma \qquad \text{(A-7)}$$

since X is a function of x only and Θ is a function of θ only, and where γ is a separation constant. Separation of the boundary conditions gives

$$X(0)\,\Theta(\theta) = 0 \quad , \quad \theta > 0 \qquad \text{(A-8)}$$

and

$$X(1)\,\Theta(\theta) = 0 \quad , \quad \theta > 0 \qquad \text{(A-9)}$$

Since Equations (A-8) and (A-9) must hold for any $\theta > 0$, then

$$X(0) = X(1) = 0 \qquad \text{(A-10)}$$

With the boundary conditions of Equation (A-10) the Sturm-Liouville problem contained within (A-1) through (A-3) is given by

$$X'' - \gamma X = 0 \qquad \text{(A-11)}$$

$$X(0) = X(1) = 0 \tag{A-12}$$

The solution of the Sturm-Liouville problem must be broken into three cases. In the first two cases, where $\gamma > 0$ and $\gamma = 0$ respectively, there is no solution. In the third case, $\gamma < 0$, let $\gamma = -\alpha^2$ where $\alpha \neq 0$, then solving the characteristic equation for the differential equation (A-11) gives

$$X(x) = A \cos(\alpha x) + B \sin(\alpha x) \tag{A-13}$$

as the general solution. The boundary conditions are then substituted to obtain

$$X(0) = A \cdot 1 + B \cdot 0 = 0 \tag{A-14}$$

or

$$A = 0 \tag{A-15}$$

and

$$X(1) = (0) \cdot 1 + B \sin(\alpha) = 0 \tag{A-16}$$

or

$$\sin(\alpha) = 0 \tag{A-17}$$

since B cannot be zero if a non-trivial solution is desired. There is an infinite number of values of α which will satisfy Equation (A-17) which, after it is recalled that $\alpha \neq 0$ and that only positive values are needed since the negative values only repeat the solutions given by the positive values, are given by

$$\alpha_n = n\pi \quad , \quad n = 1,2,3,\cdots \tag{A-18}$$

Now the eigenvalues of Equation (A-11) are given by

$$\gamma_n = -(\alpha_n)^2 = -(n\pi)^2 \quad , \quad n = 1,2,3,\ldots \tag{A-19}$$

The solutions of Equation (A-13) are then

$$X_n(x) = B_n \sin(n\pi x) \quad , \quad n = 1,2,3,\ldots \tag{A-20}$$

In the second part of the problem

$$\theta' - \gamma\theta = 0 \tag{A-21}$$

After integrating, Equation (A-21) becomes

$$\theta(\theta) = C_n \exp(-(n\pi)^2\theta) \tag{A-22}$$

By invoking the principle of superposition, the general solution of Equation (A-1) is given by

$$u(x,\theta) = \sum_{n=1}^{\infty} (B_n \cdot C_n) \sin(n\pi x) \exp(-(n\pi)^2\theta) \qquad \text{(A-23)}$$

The initial conditions are then expanded in an infinite Fourier series of the eigenfunctions, Equation (A-20), as follows:

$$u(x,0) = \sum_{n=1}^{\infty} A_n \sin(n\pi x) \cdot 1 = 1 \qquad \text{(A-24)}$$

where $A_n = B_n \cdot C_n$, and zero has been substituted for theta. This should be recognized as a Fourier sine series where

$$A_n = \frac{(1 \cdot X_n(x))}{||X_n(x)||^2} \qquad \text{(A-25)}$$

in inner product notation introduced earlier and

$$||X_n(x)||^2 = \int_0^1 X_n^2(x)\, dx \qquad \text{(A-26)}$$

or since

$$(1 \cdot X_n(x)) = \frac{(1 - (-1)^n)}{n\pi} \qquad \text{(A-27)}$$

and

$$||X_n(x)||^2 = \frac{1}{2} \tag{A-28}$$

then

$$A_n = \frac{2}{n\pi} \left(1 - (-1)^n\right) \tag{A-29}$$

If $n = 2m - 1$, then Equation (A-23) can be written

$$u(x,\theta) = \sum_{m=1}^{\infty} \frac{4}{(2m-1)\pi} \sin((2m-1)\pi x) \exp(-((2m-1)\pi)^2\theta) \tag{A-30}$$

which is the solution.

Appendix B

The Analytical Solution for the Secondary Problem

In normalized form, the secondary problem is given by

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

$$u(0,\theta) = \left.\frac{\partial u}{\partial x}\right|_{x=1} = 0, \quad \theta > 0 \tag{B-2}$$

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

As in Appendix A, the method of separation of variables will be used. A product type solution, $u(x,\theta) = X(x)\,\Theta(\theta)$ will be assumed. After substitution into Equations (B-1) through (B-3), the problem becomes

$$\frac{X''(x)}{X(x)} = \frac{\Theta'(\theta)}{\Theta(\theta)} = \gamma \tag{B-4}$$

where the prime indicates differentiation with respect to the indicated independent variable, and γ is a separation constant. Similarly, the boundary conditions become

$$X(0) = X'(1) = 0 \tag{B-5}$$

The Sturm-Liouville problem associated with Equation (B-1) is then

$$X'' - \gamma X = 0 \tag{B-6}$$

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

For the cases where $\gamma > 0$ and $\gamma = 0$, there is no solution to Equations (B-6) and (B-7). In the last case, where $\gamma < 0$, if

$$\gamma = -\alpha^2 \quad , \quad \alpha \neq 0 \tag{B-8}$$

then the solution of Equation (B-6) is given by solving the resulting characteristic equation to give

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

After substitution of the boundary conditions, the result is

$$X(0) = A \cdot 1 + B \cdot 0 = 0 \tag{B-10}$$

or

$$A = 0 \tag{B-11}$$

and

$$X'(1) = \alpha B \cos(\alpha) = 0 \qquad \text{(B-12)}$$

or, since $\alpha \neq 0$ and B cannot equal zero if the solution is to be non-trivial

$$\cos(\alpha) = 0 \qquad \text{(B-13)}$$

The values of α which satisfy Equation (B-13) are

$$\alpha_n = n\frac{\pi}{2} \quad , \quad n = 1,3,5,... \qquad \text{(B-14)}$$

or if $n = (2m-1)$

$$\alpha_m = (2m-1)\frac{\pi}{2} \quad , \quad m = 1,2,3,... \qquad \text{(B-15)}$$

The eigenvalues of the problem are given by

$$\gamma_m = -(\alpha_m)^2 = -((2m-1)\frac{\pi}{2})^2 \quad , \quad m = 1,2,3,... \qquad \text{(B-16)}$$

The solutions of Equation (B-6) with boundary conditions (B-7) are then

$$X_m(x) = B_m \sin((2m-1)\frac{\pi}{2} x) \qquad \text{(B-17)}$$

In the second part of the problem

$$\theta'(\theta) - \gamma\,\theta = 0 \tag{B-18}$$

integration with respect to θ yields

$$\theta_m(\theta) = C_m \exp(-((2m-1)\tfrac{\pi}{2})^2\,\theta) \tag{B-19}$$

The use of the principle of superposition then leads to

$$u(x,\theta) = \sum_{m=1}^{\infty} A_m \sin((2m-1)\tfrac{\pi}{2}\,x)\, \exp(-((2m-1)\tfrac{\pi}{2})^2\theta) \tag{B-20}$$

where $A_m = B_m \cdot C_m$.

Application of the initial conditions gives

$$u(x,0) = \sum_{m=1}^{\infty} A_m \sin((2m-1)\tfrac{\pi}{2}\,x) \cdot 1 = 1 \tag{B-21}$$

The coefficients, A_m , are recognized as the coefficients of the Fourier sine series given by

$$A_m = \frac{(1 \cdot X_m(x))}{||X_m(x)||^2} \tag{B-22}$$

where

$$(1 \cdot X_m(x)) = \int_0^1 \sin\left((2m-1)\frac{\pi}{2}x\right) dx = \frac{2}{(2m-1)\pi} \qquad \text{(B-23)}$$

and

$$||X_m(x)||^2 = \int_0^1 \sin^2\left((2m-1)\frac{\pi}{2}x\right) dx = \frac{1}{2} \qquad \text{(B-24)}$$

Therefore, the solution of Equation (B-1) with boundary conditions (B-2) and initial condition (B-3) is

$$u(x,\theta) = \sum_{m=1}^{\infty} \frac{4}{(2m-1)\pi} \sin\left((2m-1)\frac{\pi}{2}x\right) \exp\left(-\left((2m-1)\frac{\pi}{2}\right)^2\theta\right) \qquad \text{(B-25)}$$

Appendix C

Derivation of the Truncation Error for the Finite-Difference Formulation

Crandall (Ref 2:319) outlines a method for investigating the truncation error of Equation (28). The details are given by Campbell (Ref 15:89), with a few minor errors, for the case which includes a diffusivity constant. Here, the correct derivation will be given for the normalized equation which, of course, carries the diffusivity constant implicitly.

In Equation (28), the exact solution $u(x,\theta)$ has been expanded over six points $u_{i-1,k+1}$, $u_{i,k+1}$, $u_{i+1,k+1}$, $u_{i-1,k}$, $u_{i,k}$, and $u_{i+1,k}$. The difference between Equation (28) and the exact differential equation is called truncation error and is given by

$$e_t = \left(\frac{u_{i,k+1} - u_{i,k}}{\Delta\theta}\right) - \alpha\left(\frac{u_{i-1,k+1} - 2u_{i,k+1} + u_{i+1,k+1}}{\Delta x^2}\right)$$

$$- (1-\alpha)\left(\frac{u_{i-1,k} - 2u_{i,k} + u_{i+1,k}}{\Delta x^2}\right) - \left(\frac{\partial u}{\partial \theta} - \frac{\partial^2 u}{\partial x^2}\right) \qquad \text{(C-1)}$$

Since $\frac{\partial u}{\partial \theta} - \frac{\partial^2 u}{\partial x^2} = 0$, then

$$e_t = \frac{1}{\Delta\theta}\left[A_1 u_{i-1,k+1} + A_2 u_{i,k+1} + A_1 u_{i+1,k+1}\right.$$

$$\left. - B_1 u_{i-1,k} - B_2 u_{i,k} - B_1 u_{i+1,k}\right] \qquad \text{(C-2)}$$

where

$$A_1 = -p\alpha$$

$$A_2 = 1 + 2p\alpha$$

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

$$B_2 = 1 - 2p(1-\alpha)$$

and

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

It is a standard shorthand convention to let $\Delta x = h$ and $\Delta\theta = k$. If this notation is used, the six points of u, mentioned before, can be expanded in a Taylor series centered about $u_{i,k}$ to give

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

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

and, if $k = ph^2$, then

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

$$\pm h^3 (p + \frac{1}{6})\, u_{3x} + h^4 (\frac{p^2}{2} + \frac{p}{2} + \frac{1}{24})\, u_{4x}$$

$$\pm h^5 (\frac{p^2}{2} + \frac{p}{6} + \frac{1}{120})\, u_{5x} + \ldots \qquad \text{(C-5)}$$

where the subscripts indicate differentiation with respect to the indicated independent variable for the indicated number of times at the point $(i\Delta x, k\Delta\theta)$. If Equations (C-3) through (C-5) are substituted into Equation (C-2) and if the following relations are used

$$u_\theta = u_{2x} \qquad \text{(C-6)}$$

$$u_{2\theta} = u_{\theta 2x} = u_{4x} \qquad \text{(C-7)}$$

$$u_{3\theta} = u_{2\theta 2x} = u_{\theta 4x} = u_{6x} \qquad \text{(C-8)}$$

along with the fact that $k = ph^2$, then each power of h may be collected together in the truncation error. The coefficients of all of the odd powers of h cancel to zero. The coefficients of the even powers of h also cancel to zero up to and including order two. The resulting expression is

$$e_t = \frac{1}{k}\left[h^4\left(\frac{p^2}{2} - \alpha p^2 - \frac{p}{12}\right) u_{4x} + h^6\left(\frac{p^3}{6} - \frac{\alpha p^3}{2} - \frac{\alpha p^2}{12} - \frac{p}{360}\right) u_{6x} + 0(h)^8 + \ldots\right] \quad \text{(C-9)}$$

or

$$e_t = h^2\left(\frac{p}{2} - \alpha p - \frac{1}{12}\right) u_{4x}$$

$$+ h^4\left(\frac{p^2}{6} - \frac{\alpha p^2}{2} - \frac{\alpha p}{12} - \frac{1}{360}\right) u_{6x} + 0(h^6) + \ldots \quad \text{(C-10)}$$

The truncation error is then of order h^2 unless

$$\left(\frac{p}{2} - \alpha p - \frac{1}{12}\right) = 0 \quad \text{(C-11)}$$

or

$$\alpha = \frac{1}{2}\left(1 - \frac{1}{6p}\right) \quad \text{(C-12)}$$

This last expression is the one derived by Crandall. Crandall also points out that, if Equation (C-12) is substituted into the coefficient of the fourth power of h term, and the resulting expression set equal to zero, then the result would be an order h^6 expression. The values of α and p which yield this $0(h)^6$ scheme are (Ref 2:320)

$$\alpha = \frac{\sqrt{5}}{10} \quad \text{(C-13)}$$

and

$$p = \frac{3 - \sqrt{5}}{6} \qquad \text{(C-14)}$$

which is merely the point of intersection of the coefficients of the h^2 and h^4 terms in the truncation error.

Dirichlet boundary conditions do not alter this conclusion since, for the boundary points, the Taylor series expansions are given by

$$u_B = u_{i-1,k} = u_{i,k} - hu_x + \frac{h^2}{2} u_{2x} + \ldots = 0 \qquad \text{(C-15)}$$

and

$$u_B = u_{i-1,k+1} = u_{i,k} - hu_x + h^2 (p + \tfrac{1}{2}) u_{2x} + \ldots = 0 \qquad \text{(C-16)}$$

So, although the exact value of the function at the points $\{(i-1)\Delta x, k\Delta\theta)$ and $((i-1)\Delta x, (k+1)\Delta\theta)$ are known, this does not affect the Taylor expansions or the result of their use in Equation (C-2).

Appendix D

Derivation of the Truncation Error for the Finite-Element Formulation

The same method which was used in Appendix C will be used for the truncation error in the finite-element formulation.

If the system of equations (106), where $u_B = 0$, is treated as if it were merely a set of difference equations, then the general expression is

$$A_1 u_{i-1,k+1} + A_2 u_{i,k+1} + A_1 u_{i+1,k+1} = B_1 u_{i-1,k} + B_2 u_{i,k} + B_1 u_{i+1,k} \tag{D-1}$$

where

$$A_1 = 1 = 6\alpha p$$

$$A_2 = 4 + 12\alpha p$$

$$B_1 = 1 + 6(1-\alpha)p$$

$$B_2 = 4 - 12(1-\alpha)p$$

The truncation error is given by the difference between the exact partial differential equation and the difference equation (D-1):

$$e_t = \frac{1}{\Delta\theta}\left[A_1 u_{i-1,k+1} + A_2 u_{i,k+1} + A_1 u_{i+1,k+1}\right.$$

$$\left. - B_1 u_{i-1,k} - B_2 u_{i,k} - B_1 u_{i+1,k} - \left(\frac{\partial u}{\partial \theta} - \frac{\partial^2 u}{\partial x^2}\right)\right] \qquad \text{(D-2)}$$

But since

$$\frac{\partial u}{\partial \theta} - \frac{\partial^2 u}{\partial x^2} = 0 \qquad \text{(D-3)}$$

then

$$e_t = \frac{1}{\Delta\theta}\left[A_1 u_{i-1,k+1} + A_2 u_{i,k+1} + A_1 u_{i+1,k+1}\right.$$

$$\left. - B_1 u_{i-1,k} - B_2 u_{i,k} - B_1 u_{i+1,k}\right] \qquad \text{(D-4)}$$

Now if

$$k = ph^2 \qquad \text{(D-5)}$$

$$u_\theta = u_{2x} \qquad \text{(D-6)}$$

$$u_{2\theta} = u_{\theta 2x} = u_{4x} \qquad \text{(D-7)}$$

$$u_{3\theta} = u_{2\theta 2x} = u_{\theta 4x} = u_{6x} \qquad \text{(D-8)}$$

where $k = \Delta\theta$, $h = \Delta x$ and the subscripts indicate differentiation with respect to the subscript and for the indicated number of times at the point $(i\Delta x, k\Delta\theta)$, then the Taylor expansions in Equations (C-3) through (C-5) can be substituted into Equation (D-4) to give

$$e_t = h^2 \left(\frac{p}{2} - \alpha p + \frac{1}{12}\right) u_{4x}$$

$$+ h^4 \left(p^2(1-3\alpha) + p\left(\frac{1}{2} - \frac{\alpha}{2}\right) + \frac{1}{15}\right) u_{6x} + 0(h)^6 + \ldots \qquad \text{(D-9)}$$

To obtain an order h^4 expression, let

$$\frac{p}{2} - \alpha p + \frac{1}{12} = 0 \qquad \text{(D-10)}$$

or

$$\alpha = \frac{1}{2}\left(1 + \frac{1}{6p}\right) \qquad \text{(D-11)}$$

This is the finite-element equivalent of the method derived by Crandall for finite-differences. It should be noted further that substitution of Equation (D-11) into the coefficient of the h^4 term in Equation (D-9) and setting the resulting expression equal to zero, gives an order h^6 expression:

$$\alpha = \frac{\sqrt{5}}{10} \tag{D-12}$$

$$p = \frac{3 + \sqrt{5}}{6} \tag{D-13}$$

The similarities between Equations (D-11), (D-12), and (D-13) and those given by Crandall for the order h^4 and order h^6 schemes are striking. Also, as in the finite-difference case, Dirichlet boundary conditions do not affect the truncation error.

Appendix E

Derivation of the Variational Statement

First, it will be demonstrated that the approach suggested by Meyer (Ref 3:399) will lead to the wrong conclusion concerning the variational principle.

Meyer suggests that the functional to be minimized is of the form

$$I(\tilde{u}) = \int_0^1 F\left(x,\theta,\tilde{u},\frac{\partial\tilde{u}}{\partial x},\frac{\partial\tilde{u}}{\partial\theta}\right) dx \qquad \text{(E-1)}$$

where $\tilde{u} = u(x,\theta) + \xi v(x,\theta)$, and v is an arbitrary function which satisfies the essential boundary conditions on the problem. If Equation (E-1) is differentiated with respect to ξ then, by the chain rule of the calculus

$$\frac{\partial I}{\partial\xi} = \int_0^1 \left[\frac{\partial F}{\partial\tilde{u}}\ \frac{\partial\tilde{u}}{\partial\xi} + \frac{\partial F}{\partial\tilde{u}_x}\ \frac{\partial\tilde{u}_x}{\partial\xi} + \frac{\partial F}{\partial\tilde{u}_\theta}\ \frac{\partial\tilde{u}_\theta}{\partial\xi}\right] dx \qquad \text{(E-2)}$$

Now, since

$$\frac{\partial\tilde{u}}{\partial\xi} = v\ (x,\theta) \qquad \text{(E-3)}$$

$$\frac{\partial\tilde{u}_x}{\partial\xi} = \frac{\partial v}{\partial x} \qquad \text{(E-4)}$$

and

$$\frac{\partial \tilde{u}_\theta}{\partial \xi} = \frac{\partial v}{\partial \theta} \tag{E-5}$$

then, after substitution of these back into Equation (E-2), the result is

$$\frac{\partial I}{\partial \xi} = \int_o^1 [\frac{\partial F}{\partial \tilde{u}} v + \frac{\partial F}{\partial \tilde{u}_x} \frac{\partial v}{\partial x} + \frac{\partial F}{\partial \tilde{u}_\theta} \frac{\partial v}{\partial \theta}] dx \tag{E-6}$$

Integration of Equation (E-6) by parts leads to

$$\frac{\partial I}{\partial \xi} = \int_o^1 [\frac{\partial F}{\partial \tilde{u}} v + \frac{\partial F}{\partial \tilde{u}_\theta} \frac{\partial v}{\partial \theta} - v \frac{d}{dx} (\frac{\partial F}{\partial \tilde{u}_x})] dx \tag{E-7}$$

For the expression in Equation (E-7) to be equal to zero, the expression in the brackets must be equal to zero, or since for the minimum condition, $\xi = 0$, then $\tilde{u} = u$ and

$$v (\frac{\partial F}{\partial u} - \frac{d}{dx} (\frac{\partial F}{\partial u_x})) + \frac{\partial v}{\partial \theta} \frac{\partial F}{\partial u_\theta} = 0 \tag{E-8}$$

But this expression must hold for any arbitrary function $v(x,\theta)$ which satisfies the essential boundary conditions imposed on the problem. Since the function v and its derivative with respect to θ cannot be removed from the expression, it is worthless as a tool to obtain the functional F needed for the variational statement.

The solution to this problem is a simple one. At any given point in time, the problem can be considered to be elliptic, thus removing θ and derivatives with respect to θ from the functional F. After this is done, Equation (E-1) becomes

$$I(\tilde{u}) = \int_0^1 (F\ (x, \tilde{u}, \frac{d\tilde{u}}{dx})\ dx \qquad \text{(E-9)}$$

where $\tilde{u}(x) = u(x) + \xi v(x)$ for some fixed point in time. If Equation (E-9) is differentiated with respect to ξ the chain rule of the calculus leads to

$$\frac{\partial I}{\partial \xi} = \int_0^1 [\frac{\partial F}{\partial \tilde{u}} \frac{\partial \tilde{u}}{\partial \xi} + \frac{\partial F}{\partial \tilde{u}_x} \frac{\partial \tilde{u}_x}{\partial \xi}]\ dx \qquad \text{(E-10)}$$

And since now

$$\frac{\partial \tilde{u}}{\partial \xi} = v(x) \qquad \text{(E-11)}$$

and

$$\frac{\partial \tilde{u}_x}{\partial \xi} = \frac{\partial v}{\partial x} \qquad \text{(E-12)}$$

then, after substitution and integration by parts

$$\frac{\partial I}{\partial \xi} = \int_0^1 [\frac{\partial F}{\partial \tilde{u}} \; v(x) \; - \; v(x) \; \frac{\partial}{\partial x} \; (\frac{\partial F}{\partial \tilde{u}_x})] \; dx \qquad \text{(E-13)}$$

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

$$\frac{\partial I}{\partial \xi} = 0 \qquad \text{(E-14)}$$

For Equation (E-14) to hold, the expression in the brackets must hold for any arbitrary $v(x)$ which satisfies the essential boundary conditions. Thus

$$\frac{\partial F}{\partial u} - \frac{\partial}{\partial x} \; (\frac{\partial F}{\partial u_x}) = 0 \qquad \text{(E-15)}$$

This equation is known as the Euler-Lagrange equation. A comparison of Equation (E-15) and the differential equation

$$\frac{\partial u}{\partial \theta} = \frac{\partial^2 u}{\partial x^2} \qquad \text{(E-16)}$$

which can be rearranged as

$$(\frac{\partial u}{\partial \theta}) - \frac{\partial}{\partial x} \; (\frac{\partial u}{\partial x}) = 0 \qquad \text{(E-17)}$$

leads to the observation that

$$\frac{\partial F}{\partial u} = \frac{\partial}{\partial \theta} \; (u) \qquad \text{(E-18)}$$

and

$$\frac{\partial F}{\partial u_x} = \frac{\partial u}{\partial x} \tag{E-19}$$

Integration of Equation (E-18) gives

$$F = \frac{\partial}{\partial \theta} \left(\frac{u^2}{2}\right) + f(u_x) \tag{E-20}$$

and integration of Equation (E-19) leads to

$$F = \frac{(u_x)^2}{2} + g(u) \tag{E-21}$$

The functions f and g can be found by comparing Equations (E-20) and (E-21) so that

$$F = \frac{1}{2} \left[\frac{\partial}{\partial \theta} (u^2) + \left(\frac{\partial u}{\partial x}\right)^2\right] \tag{E-22}$$

The variational statement of the differential equation is then

$$I = \frac{1}{2} \int_0^1 \left[\frac{\partial}{\partial \theta} (u^2) + \left(\frac{\partial u}{\partial x}\right)^2\right] dx \tag{E-23}$$

which is, after normalization, the same variational statement given by Meyer.

Appendix F

Development of the Method of Weighted Residuals

The method of weighted residuals, or the Galerkin method, will be developed by following the treatment given a two-dimensional problem by Yalamanchili and Chu (Ref 16:10-17).

The error incurred by using the difference equation

$$\frac{u_{i,k+1} - u_{i,k}}{\Delta\theta} = \alpha\frac{\partial^2 u}{\partial x^2}\bigg|_{i,k+1} + (1-\alpha)\frac{\partial^2 u}{\partial x^2}\bigg|_{i,k} \qquad \text{(F-1)}$$

where the subscript i indicates the spatial node, and the subscript k indicates the time step, to approximate the following differential equation

$$\frac{\partial u}{\partial \theta} = \frac{\partial^2 u}{\partial x^2} \qquad \text{(F-2)}$$

is called a residual, defined by

$$R(x) = \alpha\frac{\partial^2 u}{\partial x^2}\bigg|_{i,k+1} + (1-\alpha)\frac{\partial^2 u}{\partial x^2}\bigg|_{i,k} - \left(\frac{u_{i,k+1} - u_{i,k}}{\Delta\theta}\right) \qquad \text{(F-3)}$$

The object in the method of weighted residuals is to minimize and distribute this error over the interval $(0,1)$, or

$$\int_0^1 R(x)\, W(x)\, dx = 0 \qquad \text{(F-4)}$$

First, the neighborhood of the ith nodal point must be divided into two adjacent elements. Then, for this case, a linear temperature distribution is assumed within each element. For element one at time $\theta = k\Delta\theta$

$$u(x) = u_{i-1,k}\left(\frac{x_i - x}{x_i - x_{i-1}}\right) + u_{i,k}\left(\frac{x - x_{i-1}}{x_i - x_{i-1}}\right) \qquad \text{(F-5)}$$

and for element two

$$u(x) = u_{i,k}\left(\frac{x_{i+1} - x}{x_{i+1} - x_i}\right) + u_{i+1,k}\left(\frac{x - x_i}{x_{i+1} - x_i}\right) \qquad \text{(F-6)}$$

The weights $W(x)$ are chosen to be the coefficients of the discrete temperatures in the element equations (F-5) and (F-6). Outside of these two one-dimensional elements, the weighting functions are zero. If these weighting functions are substituted into Equation (F-4) along with the residual, Equation (F-3), then

$$\int_{x_{i+1}}^{x_i} \left(\alpha \frac{\partial^2 u}{\partial x^2}\bigg|_{k+1} + (1-\alpha)\frac{\partial^2 u}{\partial x^2}\bigg|_{k} - \left(\frac{u_{i,k+1} - u_{i,k}}{\Delta\theta}\right)\right)\left(\frac{x - x_{i-1}}{x_i - x_{i-1}}\right)$$

$$+ \int_{x_i}^{x_{i+1}} \left(\alpha \frac{\partial^2 u}{\partial x^2}\bigg|_{k+1} + (1-\alpha)\frac{\partial^2 u}{\partial x^2}\bigg|_{k} - \left(\frac{u_{i,k+1} - u_{i,k}}{\Delta\theta}\right)\right)\left(\frac{x_{i+1} - x}{x_{i+1} - x_i}\right) \qquad \text{(F-7)}$$

Next, the temperature distributions, Equations (F-5) and (F-6), are substituted into Equation (F-7) and the following finite-difference expression for the Laplacian

$$\frac{\partial^2 u}{\partial x^2} \cong \frac{u_{i-1} - 2u_i + u_{i+1}}{\Delta x^2} \qquad \text{(F-8)}$$

is also substituted into Equation (F-7), then Equation (F-7) becomes

$$\int_{x_{i-1}}^{x_i} \frac{\alpha}{\Delta x^2} (u_{i-1,k+1} - 2u_{i,k+1} + u_{i+1,k+1}) \left(\frac{x - x_{i-1}}{\Delta x}\right) dx$$

$$+ \int_{x_{i-1}}^{x_i} \frac{(1-\alpha)}{\Delta x^2} (u_{i-1,k} - 2u_{i,k} + u_{i+1,k}) \left(\frac{x - x_{i-1}}{\Delta x}\right) dx$$

$$- \int_{x_{i-1}}^{x_i} \frac{1}{\Delta\theta\Delta x} (x - x_{i-1}) \left[u_{i-1,k+1} \left(\frac{x_i - x}{\Delta x}\right) + u_{i,k+1} \left(\frac{x - x_{i-1}}{\Delta x}\right)\right.$$

$$\left. - u_{i-1,k} \left(\frac{x_i - x}{\Delta x}\right) - u_{i,k} \left(\frac{x - x_{i-1}}{\Delta x}\right)\right] dx$$

$$+ \int_{x_i}^{x_{i+1}} \frac{\alpha}{\Delta x^2} (u_{i-1,k+1} - 2u_{i,k+1} + u_{i+1,k+1}) \left(\frac{x_{i+1} - x}{\Delta x}\right) dx$$

$$+ \int_{x_i}^{x_{i+1}} \frac{(1-\alpha)}{\Delta x^2} (u_{i-1,k} - 2u_{i,k} + u_{i+1,k}) \left(\frac{x_{i+1} - x}{\Delta x}\right) dx$$

$$- \int_{x_i}^{x_{i+1}} \frac{1}{\Delta\theta\Delta x} (x_{i+1} - x) \left[u_{i,k+1} \left(\frac{x_{i+1} - x}{\Delta x}\right) + u_{i+1,k+1} \left(\frac{x - x_i}{\Delta x}\right)\right.$$

$$\left. - u_{i,k} \left(\frac{x_{i+1} - x}{\Delta x}\right) - u_{i+1,k} \left(\frac{x - x_i}{\Delta x}\right)\right] dx = 0 \qquad \text{(F-9)}$$

After integration and simplification, Equation (F-9) becomes

$$\frac{\alpha}{2\Delta x}\,(u_{i-1,k+1} - 2u_{i,k+1} + u_{i+1,k+1})$$

$$+ \frac{(1-\alpha)}{2\Delta x}\,(u_{i-1,k} - 2u_{i,k} + u_{i+1,k})$$

$$- \frac{\Delta x}{6\Delta\theta}\,u_{i-1,k+1} - \frac{\Delta x}{3\Delta\theta}\,u_{i,k+1} + \frac{\Delta x}{6\Delta\theta}\,u_{i-1,k} + \frac{\Delta x}{3\Delta\theta}\,u_{i,k}$$

$$+ \frac{\alpha}{2\Delta x}\,(u_{i-1,k+1} - 2u_{i,k+1} + u_{i+1,k+1})$$

$$+ \frac{(1-\alpha)}{2\Delta x}\,(u_{i-1,k} - 2u_{i,k+1} + u_{i+1,k}) \qquad \text{(F-10)}$$

$$- \frac{\Delta x}{6\Delta\theta}\,u_{i,k+1} - \frac{\Delta x}{6\Delta\theta}\,u_{i+1,k+1} + \frac{\Delta x}{3\Delta\theta}\,u_{i,k} + \frac{\Delta x}{6\Delta\theta}\,u_{i+1,k} = 0$$

If $p = \Delta\theta/(\Delta x)^2$, then Equation (F-10) can be rewritten as

$$(\tfrac{1}{6} - \alpha p)\,u_{i-1,k+1} + (\tfrac{2}{3} - 2\alpha p)\,u_{i,k} + (\tfrac{1}{6} - \alpha p)\,u_{i+1,k+1} \qquad \text{(F-11)}$$

$$= (\tfrac{1}{6} + (1-\alpha)p)\,u_{i-1,k} + (\tfrac{2}{3} - 2(1-\alpha)p)\,u_{i,k} + (\tfrac{1}{6} + (1-\alpha)p)\,u_{i+1,k}$$

This is the general finite-element difference equation which, on close inspection, is seen to be exactly equivalent to the general expression which was derived by the Ritz method using the variational principle given by Meyers.

Appendix G

Derivation of the L_2 Error Norm for Finite-Elements

Strang and Fix (Ref 9:49) show, using the famous "Nitsche trick," that a finite-element approximation, with piecewise linear elemental temperature distribution, satisfies

$$||u - u^N||_o \leq C'h^2 \tag{G-1}$$

where the norm is the H^o or L_2 norm defined in Chapter II.

This error measure is more closely related to the underlying nature of the finite-element method than is a pointwise error bound; it is often referred to as the displacement error.

The exact analytical solution, u , is given by

$$u(x,\theta) = \sum_{n=1}^{\infty} \frac{4}{(2n-1)\pi} \sin((2n-1)\pi x) \exp(-((2n-1)\pi)^2\theta) \tag{G-2}$$

The finite-element solution assumes a linear temperature distribution within the elements defined over the interval $(i\Delta x,(i+1)\Delta x)$:

$$u^N(x)\Big|_{\theta = k\Delta\theta} = u_{i,k} + \left(\frac{u_{i+1,k} - u_{i,k}}{\Delta x}\right)(x - i\Delta x) \tag{G-3}$$

Substitution of Equations (G-2) and (G-3) into the L_2 norm gives

$$||u-u^N||_o = [\int_o^1 (u(x) - u^N(x))^2 \, dx]^{\frac{1}{2}}$$

$$= [\sum_{i=0}^{N-1} \int_{i\Delta x}^{(i+1)\Delta x} ([\sum_{n=1}^{\infty} a_n \sin(b_n x) \exp(-b_n^2\theta)] - (C + D_x))^2 \, dx]^{\frac{1}{2}} \quad \text{(G-4)}$$

where

$$a_n = 4/(2n-1)\pi$$

$$b_n = (2n-1)\pi$$

$$C = u_{i,k} - (i\Delta x(u_{i+1,k} - u_{i,k})/\Delta x)$$

$$D = (u_{i+1,k} - u_{i,k})/\Delta x$$

$$N = \text{number of elements}$$

Since the integration is taken with respect to x, then let

$$E_n = \exp(-b_n^2\theta) \quad \text{(G-5)}$$

as a further simplification. Expanding Equation (G-4)

$$||u - u^N||_o = \left[\sum_{i=0}^{N-1} \left(\int_{i\Delta x}^{(i+1)\Delta x} \left(\sum_{n=1}^{\infty} a_n \sin(b_n x)\, E_n\right)^2 dx\right.\right.$$

$$- 2\int_{i\Delta x}^{(i+1)\Delta x} \left(\sum_{n=1}^{\infty} a_n \sin(b_n x)\, E_n\right)(C + D_x)\, dx$$

$$\left.\left. + \int_{i\,x}^{(i+1)\,x} (C + D_x)^2\, dx\right)\right]^{\frac{1}{2}} \qquad \text{(G-6)}$$

This last step can be taken since it is well known that in the following equation

$$\sum_{n=1}^{\infty} w_n = \left(\sum_{n=1}^{\infty} s_n\right)\left(\sum_{n=1}^{\infty} t_n\right) \qquad \text{(G-7)}$$

if the series whose terms are s_n and the series whose terms are t_n both converge absolutely, then the series whose terms are w_n also converges absolutely. The exact solution has previously been shown to be uniformly convergent. Further, the integrals which involve infinite series in Equation (G-6) are integrable since, in the equation

$$S(x) = \sum_{n=1}^{\infty} v_n(x) \quad , \quad a \le x \le b \qquad \text{(G-8)}$$

if the functions $v_n(x)$ are continuous (they are), and if the series in Equation (G-8) whose terms are $v_n(x)$ is uniformly convergent (previously demonstrated for the first integral on the right-hand side of Equation (G-6), then $S(x)$ is also continuous and can be integrated

term by term (Ref 5:28). The series

$$P(x) = \left[\sum_{n=1}^{\infty} (a_n \sin(b_n x) E_n) \right] (C + Dx) = \sum_{n=1}^{\infty} P_n(x) \qquad \text{(G-9)}$$

can be shown to be uniformly convergent by the "Weierstrass M-test" (Ref 5:28). If there is a convergent series of positive constants

$$\sum_{n=1}^{\infty} M_n \qquad \text{(G-10)}$$

such that

$$|P_n(x)| \leq M_n \quad , \quad a \leq x \leq b$$

then series (G-10) is uniformly convergent on (a,b) . The terms defined by Equation (G-9) satisfy the relation

$$|P_n(x)| \leq \left| \frac{4}{(2n-1)\pi} \exp(-((2n-1)\pi)^2\theta) \right| \qquad \text{(G-11)}$$

since

$$u^N(x) \leq 1 \qquad \text{(G-12)}$$

and

$$|\sin((2n-1)\pi x)| \leq 1 \qquad \text{(G-13)}$$

Then, by the ratio test for convergence of an infinite series

$$\lim_{n\to\infty} \left| \frac{\frac{4}{(2n+1)\pi} \exp(-(2n+1)^2\pi^2\theta)}{\frac{4}{(2n-1)\pi} \exp(-(2n-1)^2\pi^2\theta)} \right| = 0 \qquad \text{(G-14)}$$

which implies convergence and, therefore, uniform convergence for Equation (G-9).

Now the first integral in Equation (G-6) can be written

$$I_1 = \int_{i\Delta x}^{(i+1)\Delta x} \left(\sum_{n=1}^{\infty} a_n \sin(b_n x)\, E_n \right)^2 dx = \int_{i\Delta x}^{(i+1)\Delta x} \sum_{n=1}^{\infty} f_n^2(x)\, dx \qquad \text{(G-15)}$$

This series, previously shown to be square-integrable, can be written

$$\int_{i\Delta x}^{(i+1)\Delta x} \sum_{n=1}^{\infty} f_n^2(x)\, dx = \sum_{n=1}^{\infty} \int_{i\Delta x}^{(i+1)\Delta x} f_n^2(x)\, dx$$

$$+ \; 2 \sum_{n=1}^{\infty} \left[\sum_{m=n+1}^{\infty} \int_{i\Delta x}^{(i+1)\Delta x} f_n(x)\, f_m(x)\, dx \right] \qquad \text{(G-16)}$$

The series is simply truncated when the computer reaches its limit for storing small numbers. No effort was made to estimate the error, but it could be done by a method analogous to the one used in Chapter II for the analytical solution. After integration, the expression in

Equation (G-16) becomes

$$I_1 \cong \sum_{n=1}^{K} a_n^2 E_n^2 \left[\frac{\Delta x^2}{2} - \frac{1}{4b_n}\left(\sin 2b_n(i+1)\Delta x - \sin 2b_n\, i\Delta x\right)\right]$$

$$+ \sum_{n=1}^{K-1} \sum_{m=n+1}^{K} 2a_n E_n a_m E_m \left[\frac{1}{(b_n - b_m)}\left(\sin((b_n - b_m)(i+1)\Delta x\right)\right.$$

$$\left. - \sin((b_n - b_m) i\Delta x)\right) - \frac{1}{(b_n + b_m)}\left(\sin((b_n + b_m)(i+1)\Delta x\right)$$

$$\left. - \sin((b_n + b_m) i\Delta x)\right] \qquad \text{(G-17)}$$

where K is large.

The second integral in Equation (G-6) can be shown to be termwise integrable, using the same rationale as used for the first integral. The result, then, after integration is

$$I_2 = 2C \sum_{n=1}^{K} \frac{a_n E_n}{b_n}\left(\cos(b_n(i+1)\Delta x) - \cos(b_n i\Delta x)\right)$$

$$-2D \sum_{n-1}^{K} \frac{a_n E_n}{b_n^2}\left[\left(\sin(b_n(i+1)\Delta x) - \sin(b_n i\Delta x)\right)\right.$$

$$\left. - (b_n(i+1)\Delta x \cos(b_n(i+1)\Delta x] - b_n i\Delta x \cos(b_n i\Delta x))\right] \qquad \text{(G-18)}$$

The third integral in Equation (G-6) is trivial and is given by

$$I_3 = C^2 [(i+1\Delta x - i\Delta x] + CD [((i+1)\Delta x)^2 - (i\Delta x)^2]$$

$$+ \frac{D^2}{3} [((i+1)\Delta x)^3 - (i\Delta x)^3] \qquad \text{(G-19)}$$

The complete L_2 norm then is

$$||u-u^N||_o = [I_1 + I_2 + I_3]^{\frac{1}{2}} \qquad \text{(G-20)}$$

Appendix H

Computer-Generated Plots of Results

This appendix contains the graphical results of the project. The results are handled in this way since there is a large number of plots which need to be carefully organized in order to prevent confusion. Each three-letter run identifier, for example, CER, is associated with a data set. The results from a given set of data may be displayed in a number of ways. There are three basic formats used. The first shows the discretization error ratio as a function of α . This format emphasizes the effect on the discretization error ratio for small deviations in α near the optimum value. It also shows how the discretization error ratio, which is discussed in the chapter on results, varies as α varies over its range of stable values. The results from three time steps are given in each case. The time steps shown are usually the last three out of a total of eight in a given calculation, since the first few time steps show some oscillatory behavior, but then smooth out in later time steps. The second format shows the discretization error ratio as a function of time. Only three values of the parameter α were followed in time using this format. These were usually the values for the Crank-Nicolson, optimum-implicit, and pure-implicit methods. This second format shows the behavior of the solution for the predicted values of α . This format is very informative, since the peak in a discretization error ratio versus α curve may appear to be displaced from the optimum value of α in some cases, and to show a high convergence rate. These high peaks are only transient, however, and are superimposed upon

the true peak, whose behavior is perhaps better indicated by the second format. The third format shows the absolute magnitude of the error as a function of α . The discretization error ratio is a useful tool for convergence studies, but the absolute magnitude of the error shows, in the most direct way, the effect of the parameter α on the accuracy of the solution.

Each of these three formats is used to display the four error norms for finite-elements and the three for finite-differences. Because of the large amount of data that could be generated, the pointwise error is measured only at the point $x = 0.1$. The generalized mean is the sum of the absolute values of the pointwise errors at nine (or ten for the secondary problem) evenly spaced nodes. This shows the effect of the parameter α on the pointwise error over the whole interval. The error norm termed "maximum error any node", or more properly called a discrete Tchebycheff norm, may be the error measure of greatest interest to the engineer. It shows the effect of the parameter α on the maximum deviation at any node between the true solution and the numerical solution. The L_2 norm is defined only for the finite-element method by

$$||u-u^N||_0 = \left[\int_0^1 (u-u^N)^2 \, dx \right]^{\frac{1}{2}} \qquad \text{(H-1)}$$

Since the numerical solution is defined for all points on the interval in the finite-element method, not just at the nodes, this error norm shows how the accuracy of the complete numerical solution varies with the parameter α .

In the analysis of the primary problem, to obtain the discretization error, the error obtained when the domain was divided into 10 intervals was compared to the error obtained when the domain was further subdivided into 20 intervals. Then the error for a 20-interval arrangement was also compared to the error for a 40-interval arrangement. In the analysis of the secondary problem, only the 10 to 20-interval error ratio was computed. The more complete analysis in the primary problem allows inspection of the results as the space domain is further subdivided. The parameter DX in the legend for the graphs represents Δx . P is the Fourier modulus $\Delta\theta/(\Delta x)^2$ and α is the parameter which measures the weight placed on the temperatures in the new time step in the numerical scheme, or in other words, the "degree of implicitness." A value of 0 for α would correspond to Euler's method, or for finite-differences, an explicit scheme. A value of .5 for α gives the famous Crank-Nicolson scheme, and when α is 1 , the scheme is fully implicit.

In the graphs for the finite-element method, results are shown only for values of α between 0.4 and 1.0 , since for the values of the Fourier modulus used in the investigation, the finite-element method is more restrictive with respect to stability, and the solutions for values of α less than .5 are often subject to severe oscillations or perhaps even instability.

Each section of graphs is introduced with a descriptive note and a key to aid in the analysis. In each section, reference will be made to three options under which the calculations were performed. Option one is the straightforward approach which has no modifications. In option two, the exact analytical solution has been substituted at the first

time step. This transforms the problem into a new problem without the discontinuity between the initial and boundary conditions. In option three, an attempt is made to reduce the effect of the discontinuity in the problem by subdividing the space-time grid for the first time step. The most extensive analysis is made with option one, since it shows most clearly the predicted behavior of the optimum implicit methods for both finite-difference and finite-elements.

Section I

The Results for the Primary Problem Using Finite-Differences

This section shows the graphical results for the solution of the primary problem by finite-differences. The following key shows which options are included in this set of graphs.

Table H-I

Key to the Plots in Section I

Run Identifier	Fourier Modulus (p)	Option Number
CDD	0.5	1
CDE	1.0	0
CDF	1.0	1
CDG	1.0	2
CDH	2.0	1

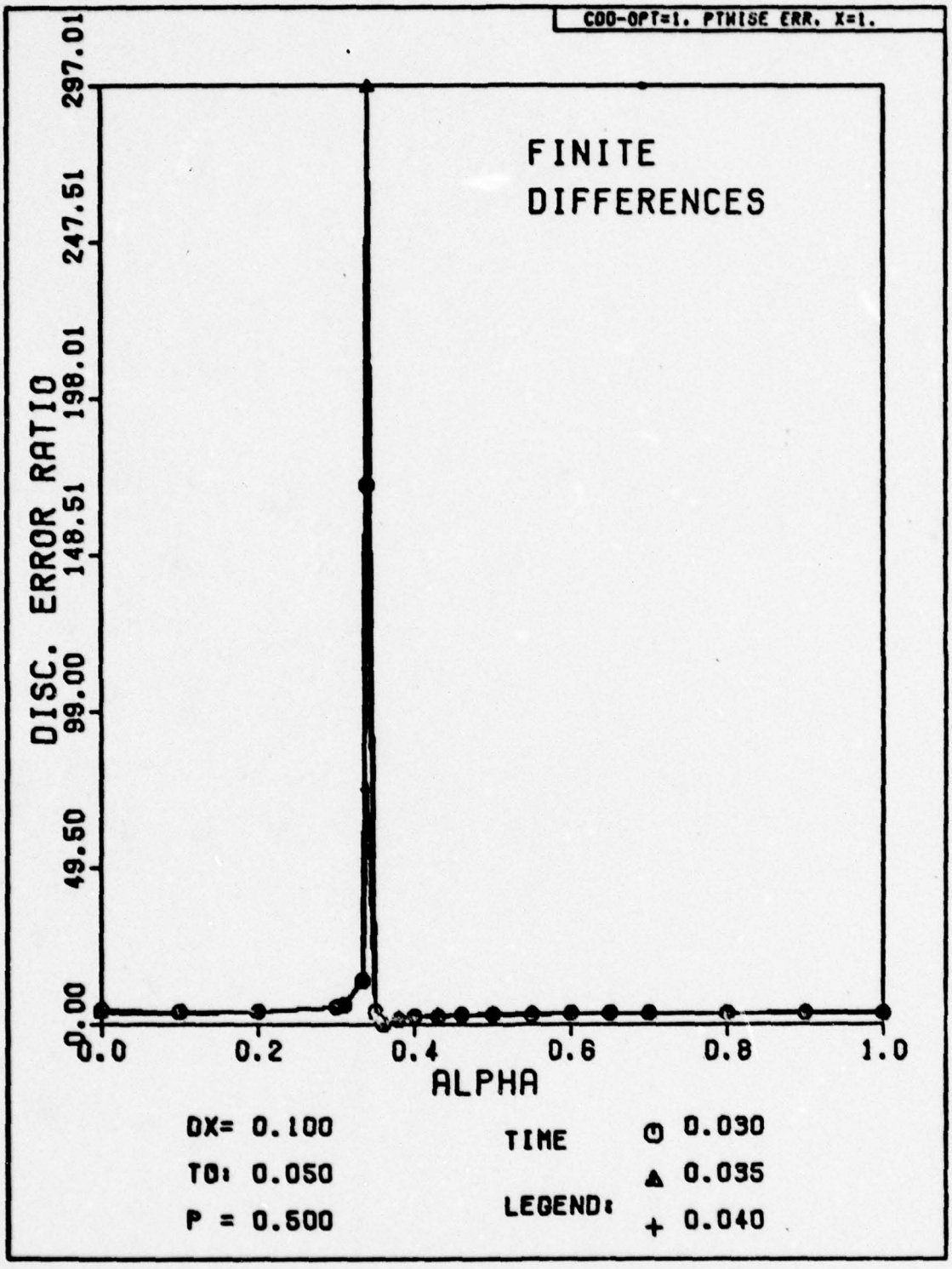


Fig. H-1. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution for the first time step.

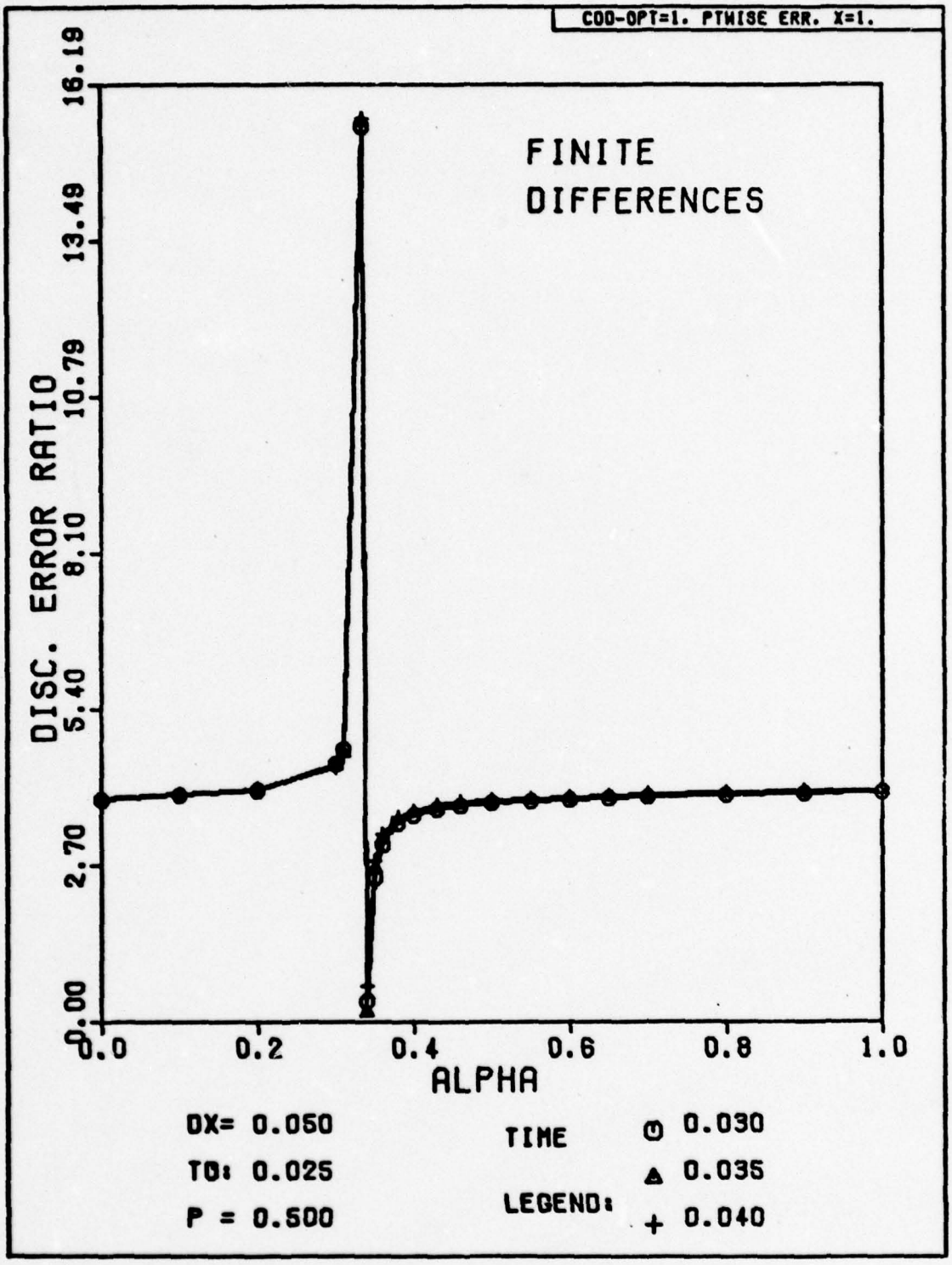


Fig. H-2. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

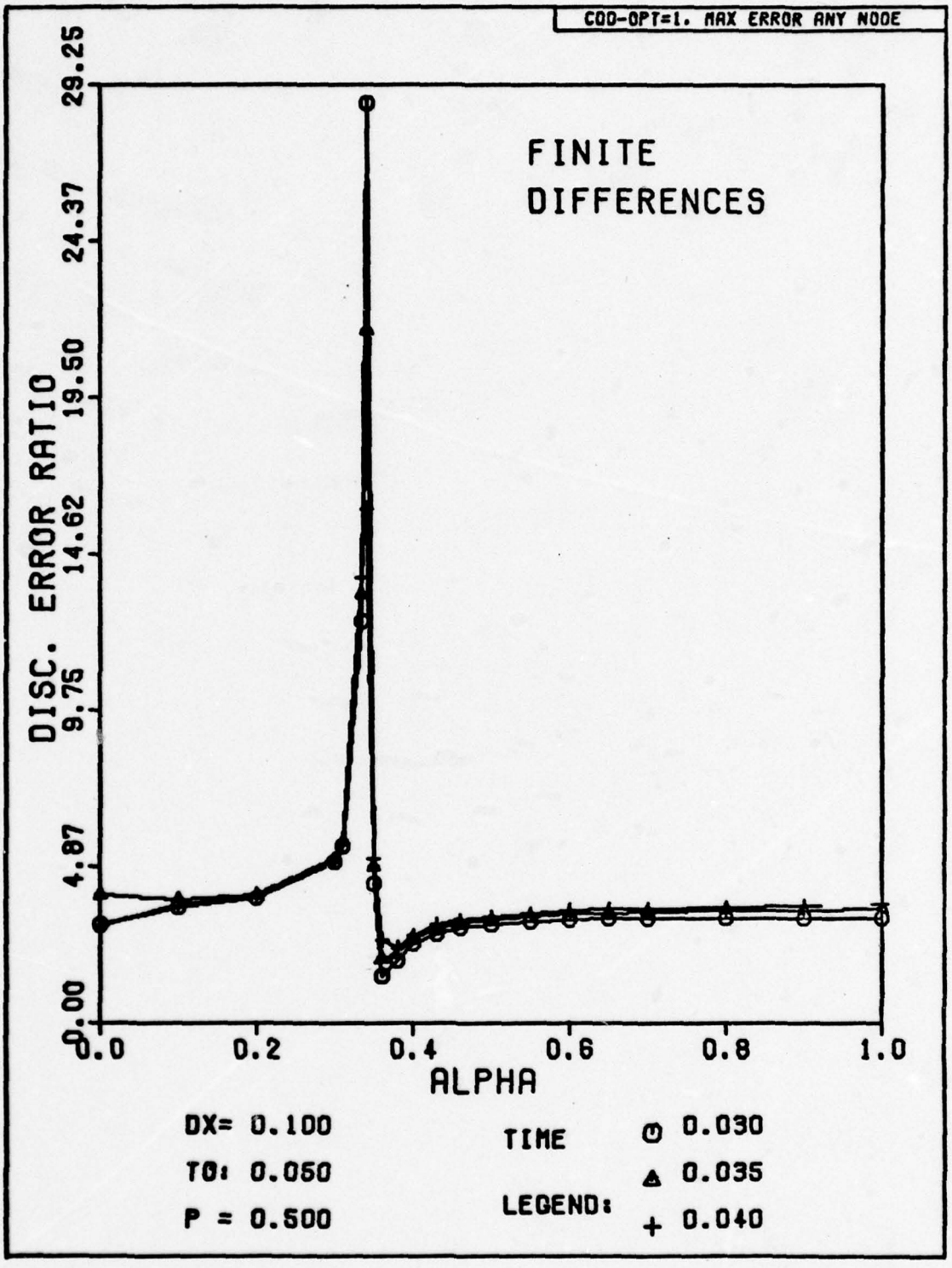


Fig. H-3. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

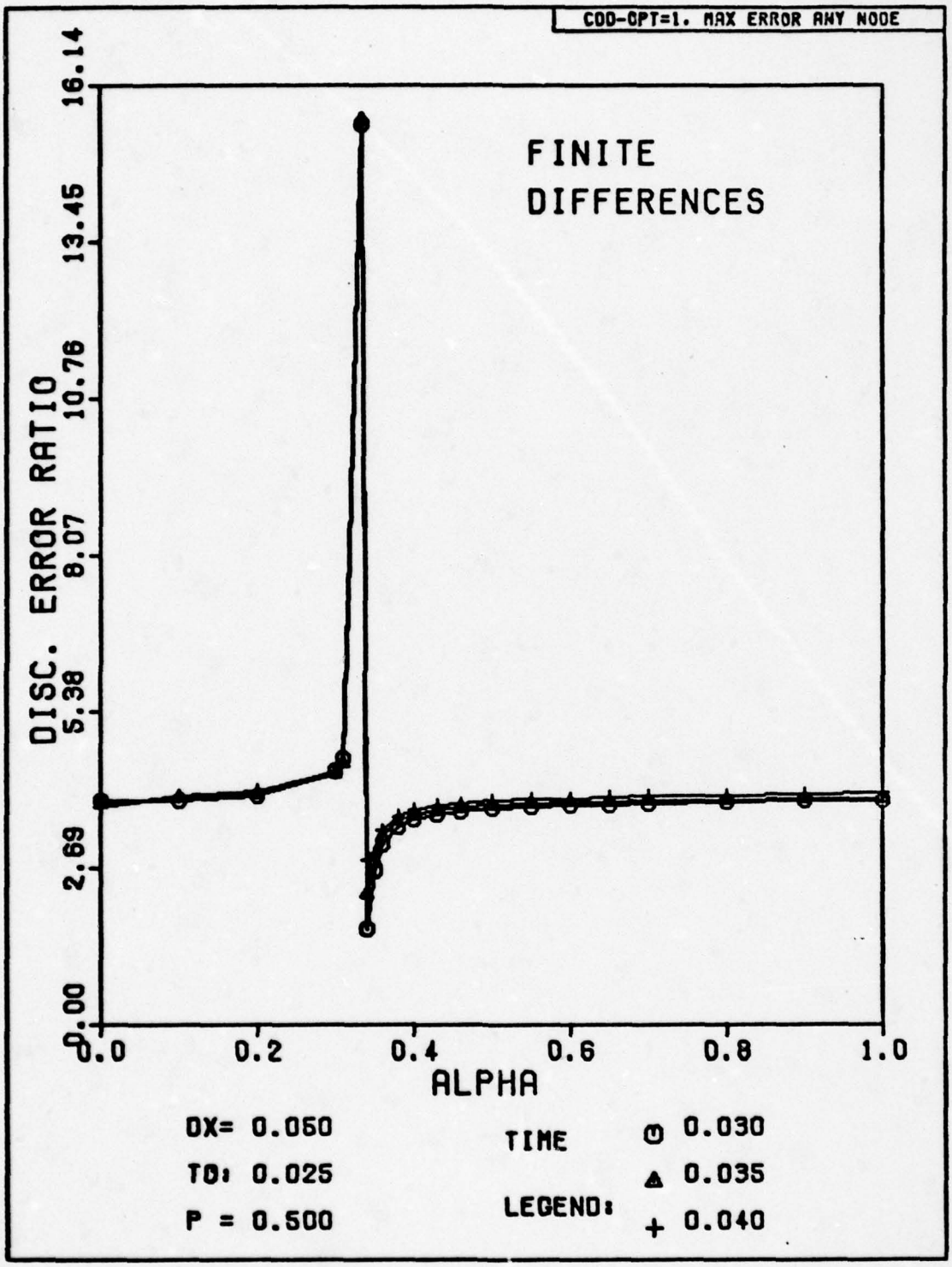


Fig. H-4. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

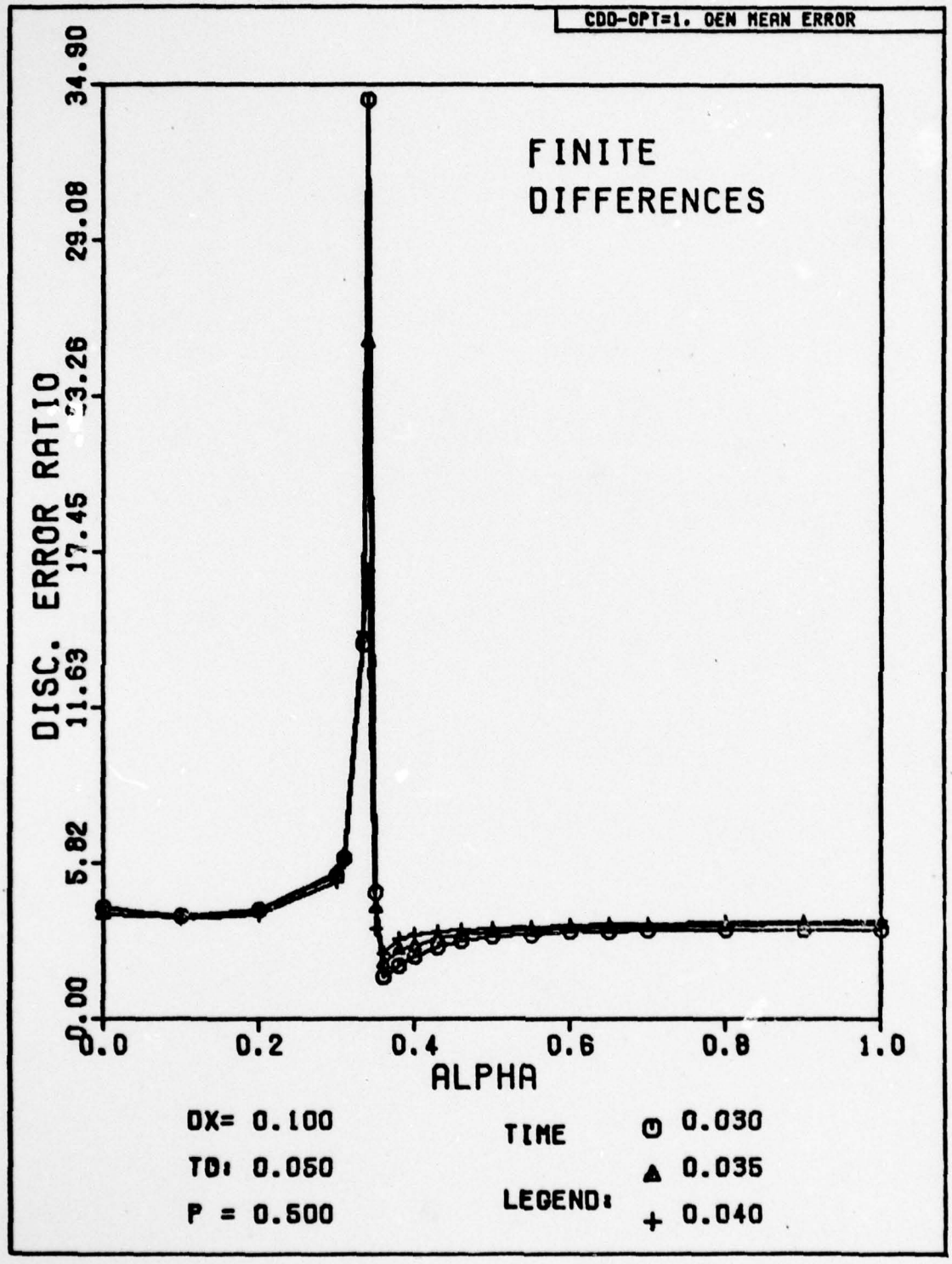


Fig. H-5. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

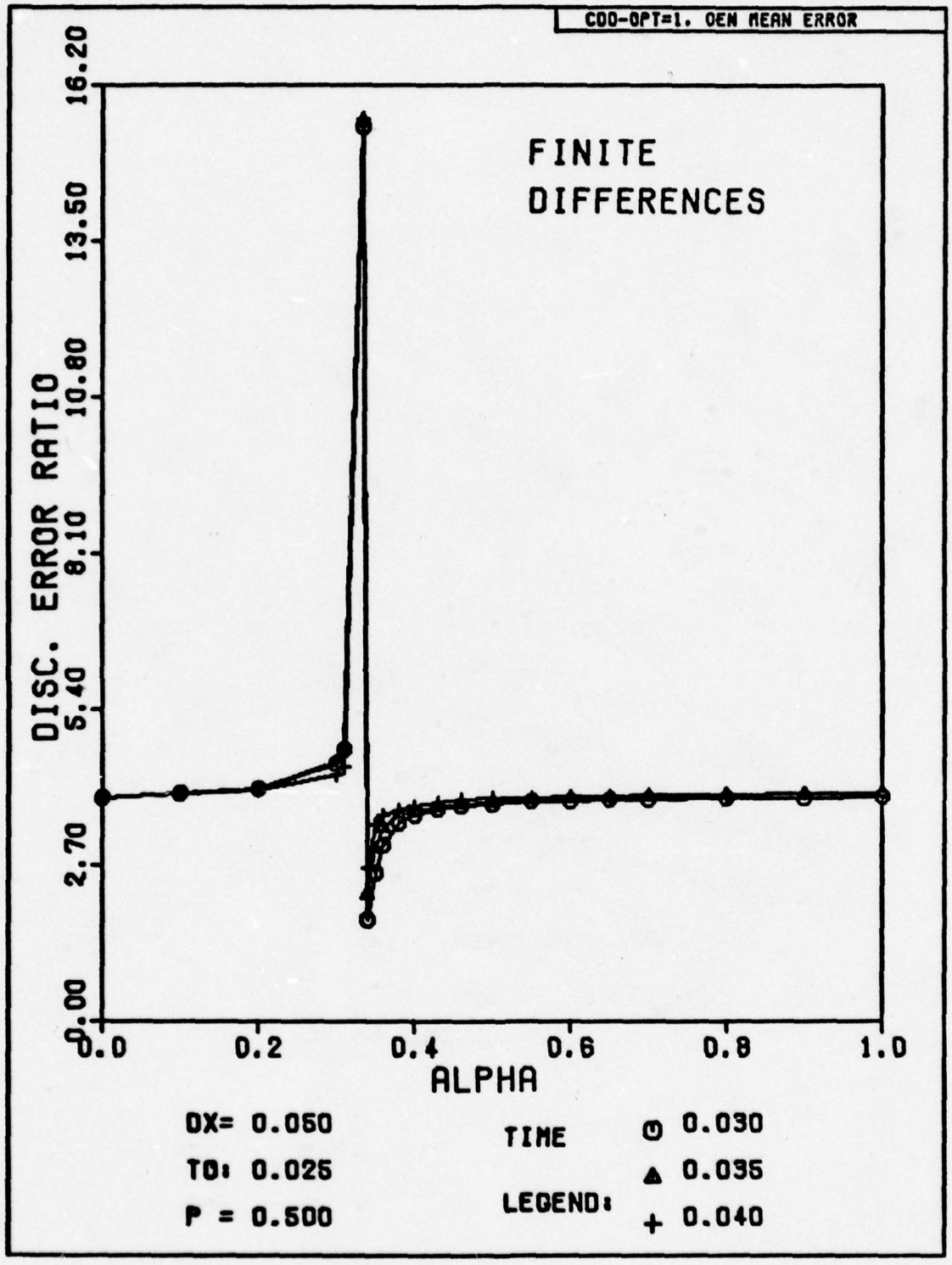


Fig. H-6. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

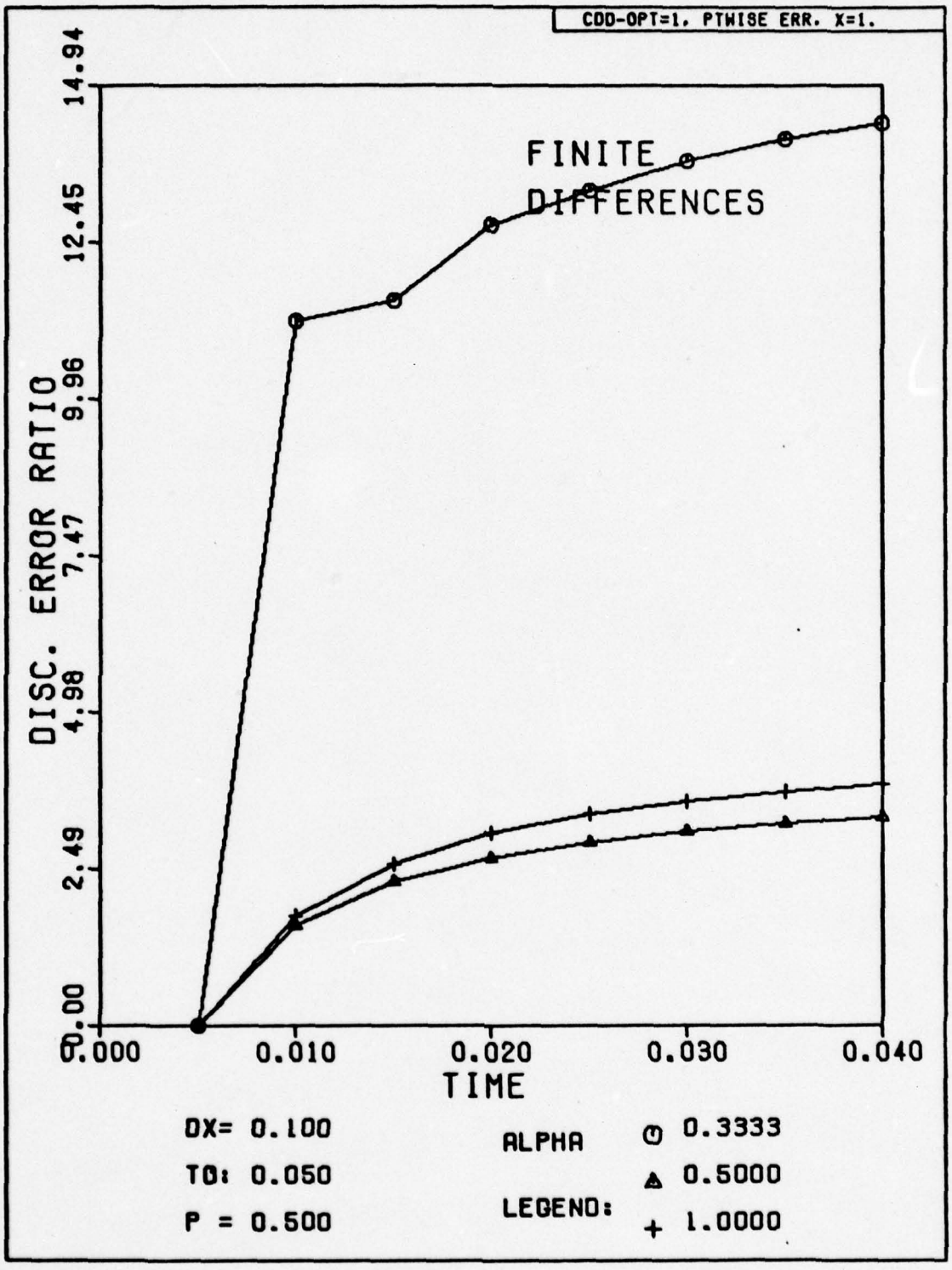


Fig. H-7. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

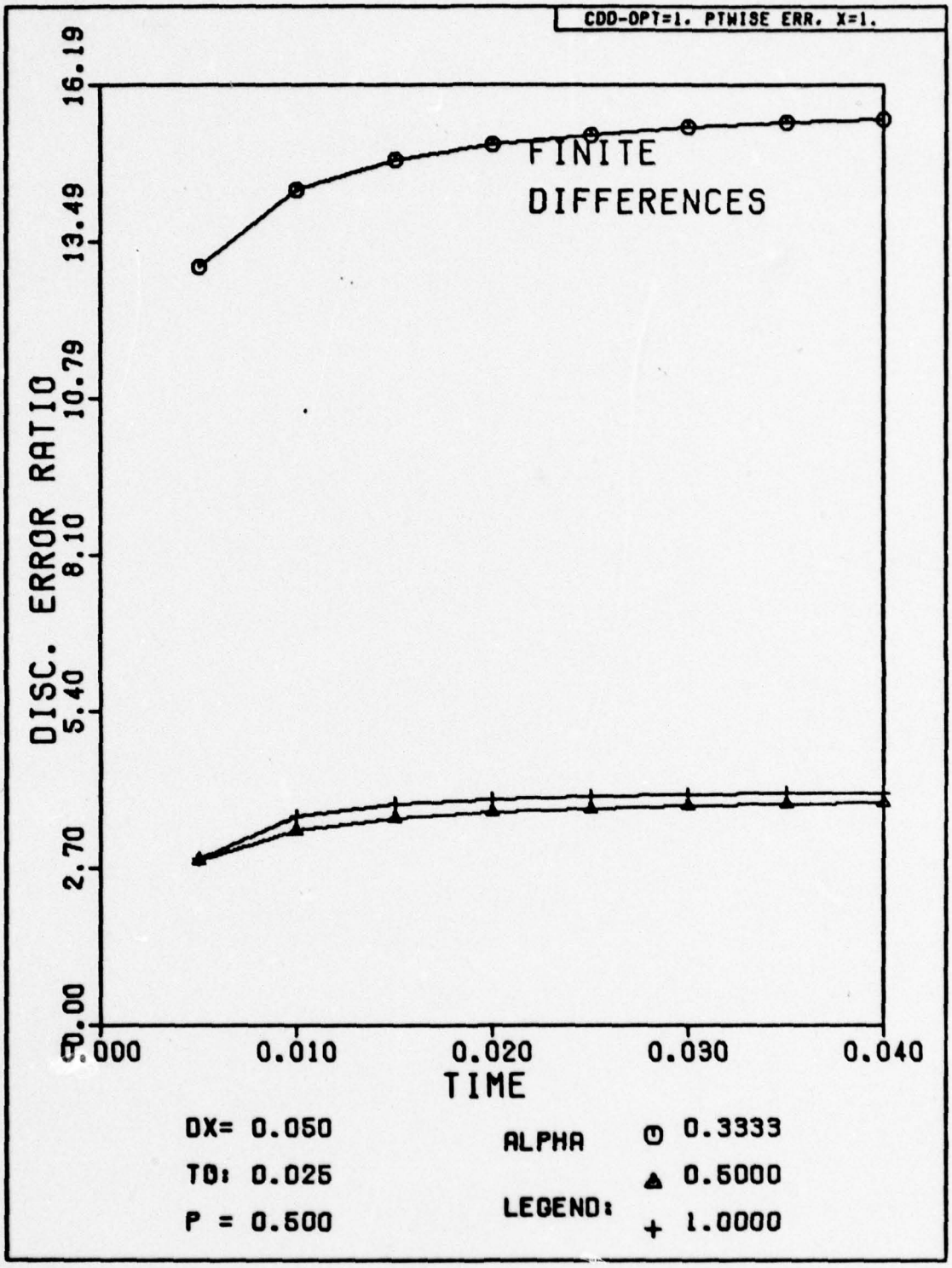


Fig. H-8. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

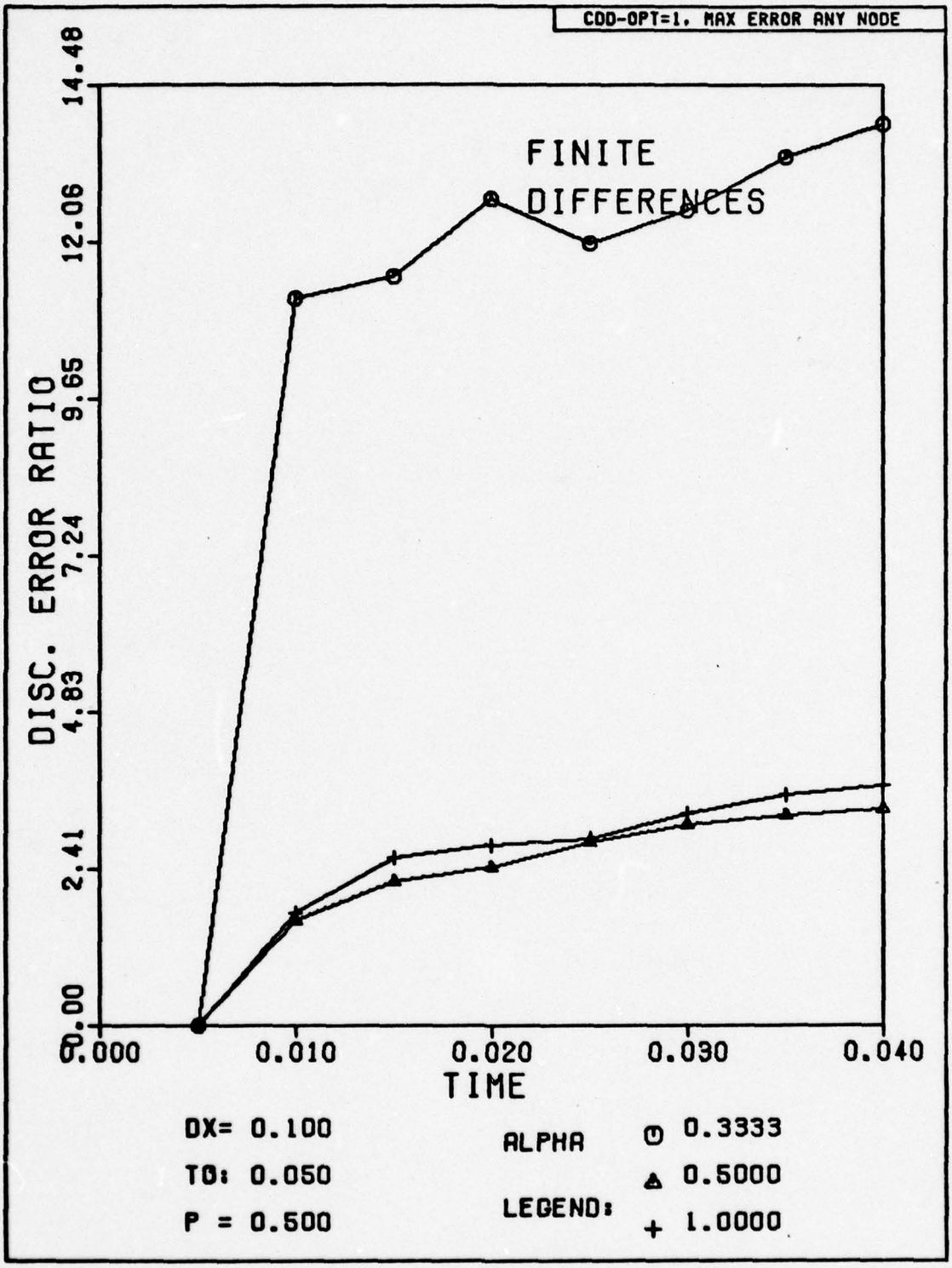


Fig. H-9. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

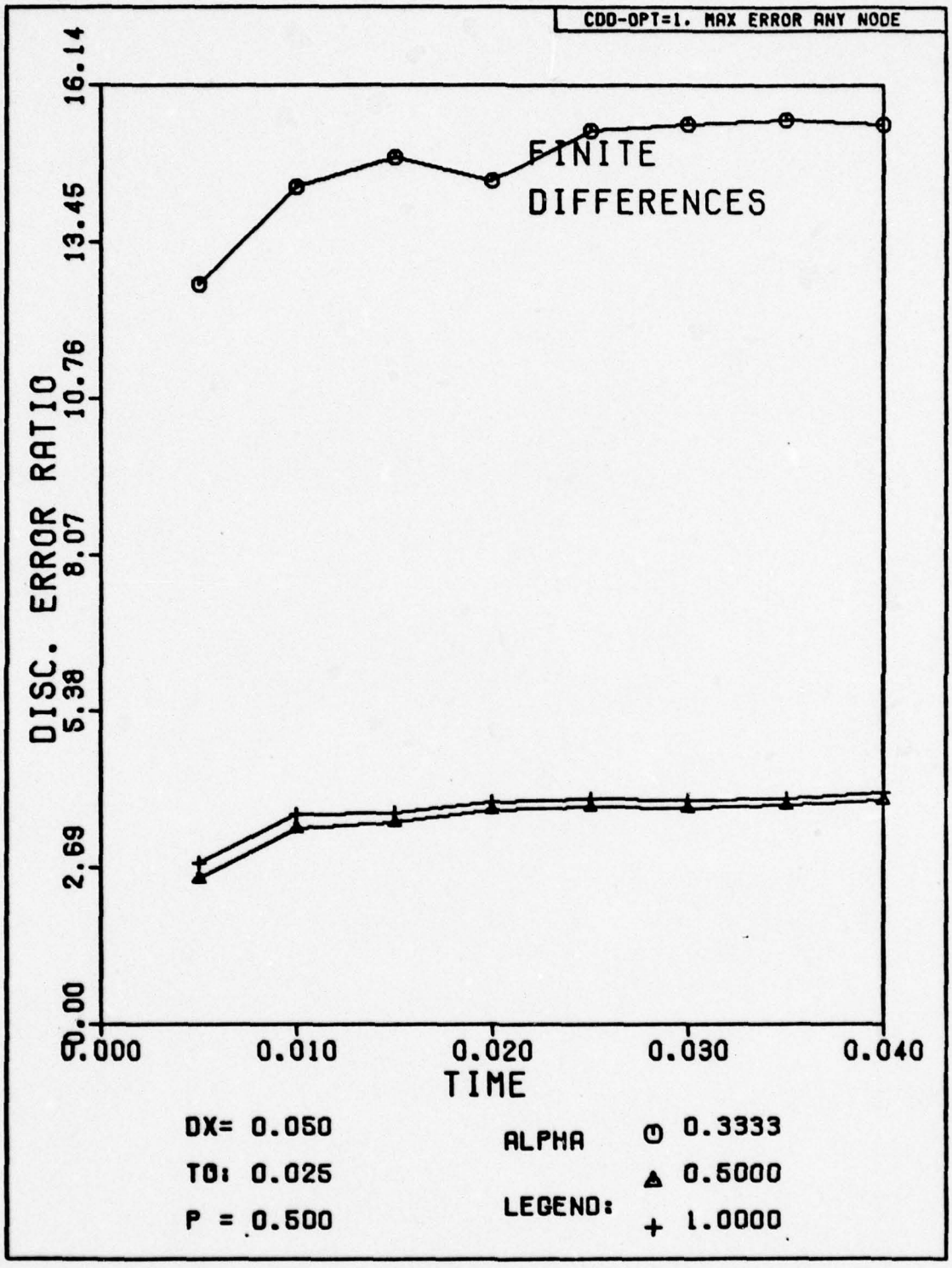


Fig. H-10. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

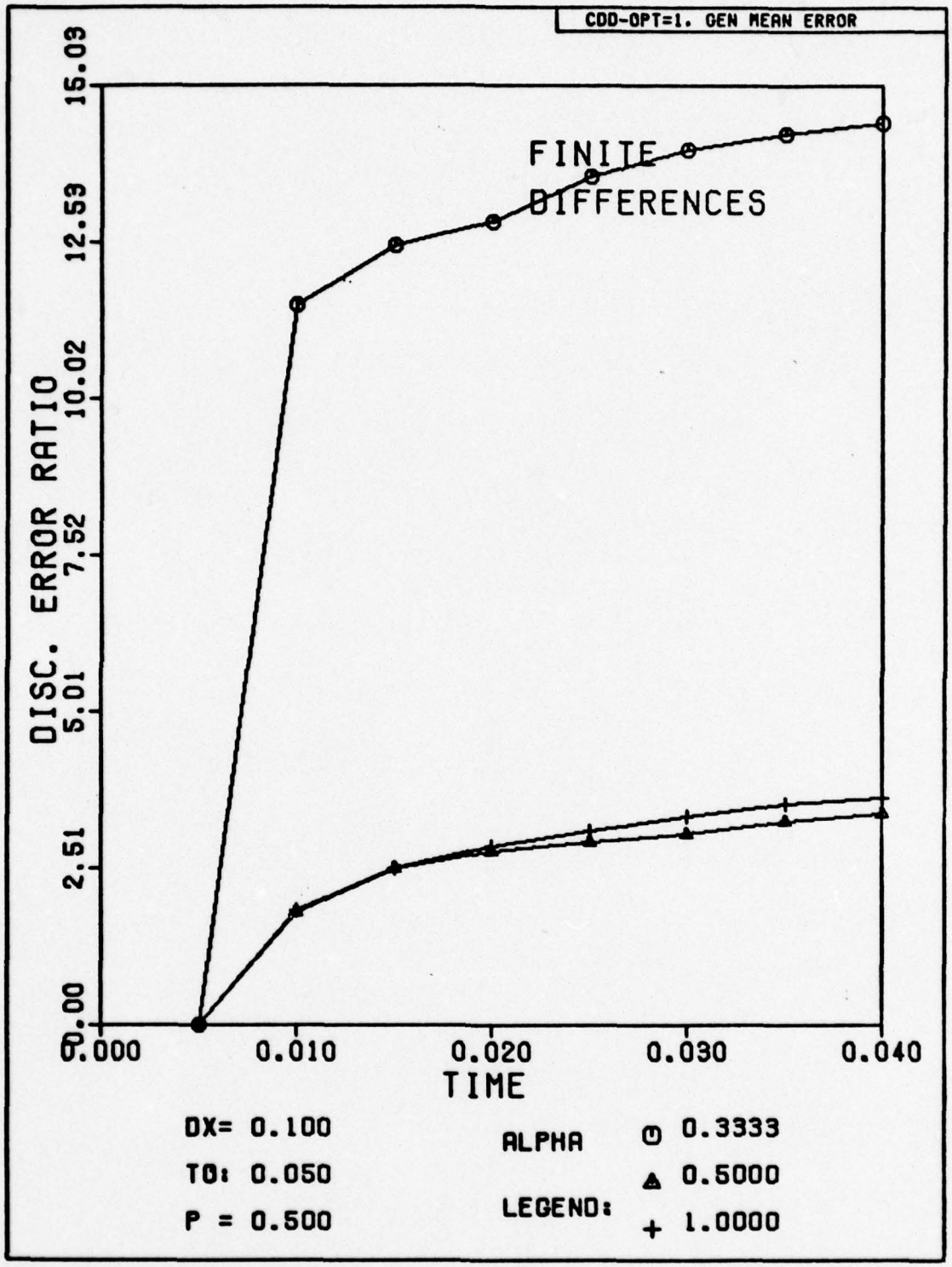


Fig. H-11. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

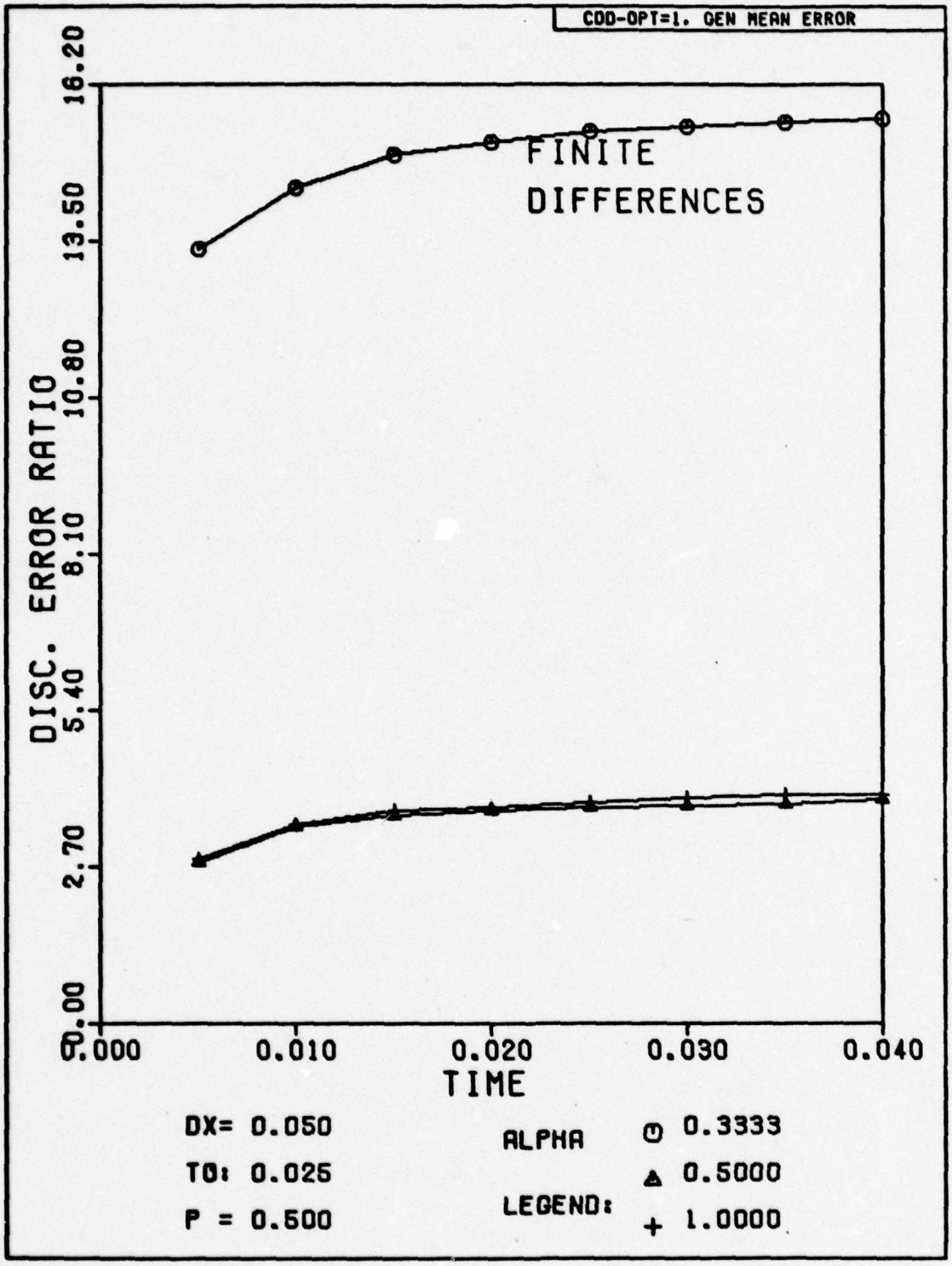


Fig. H-12. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

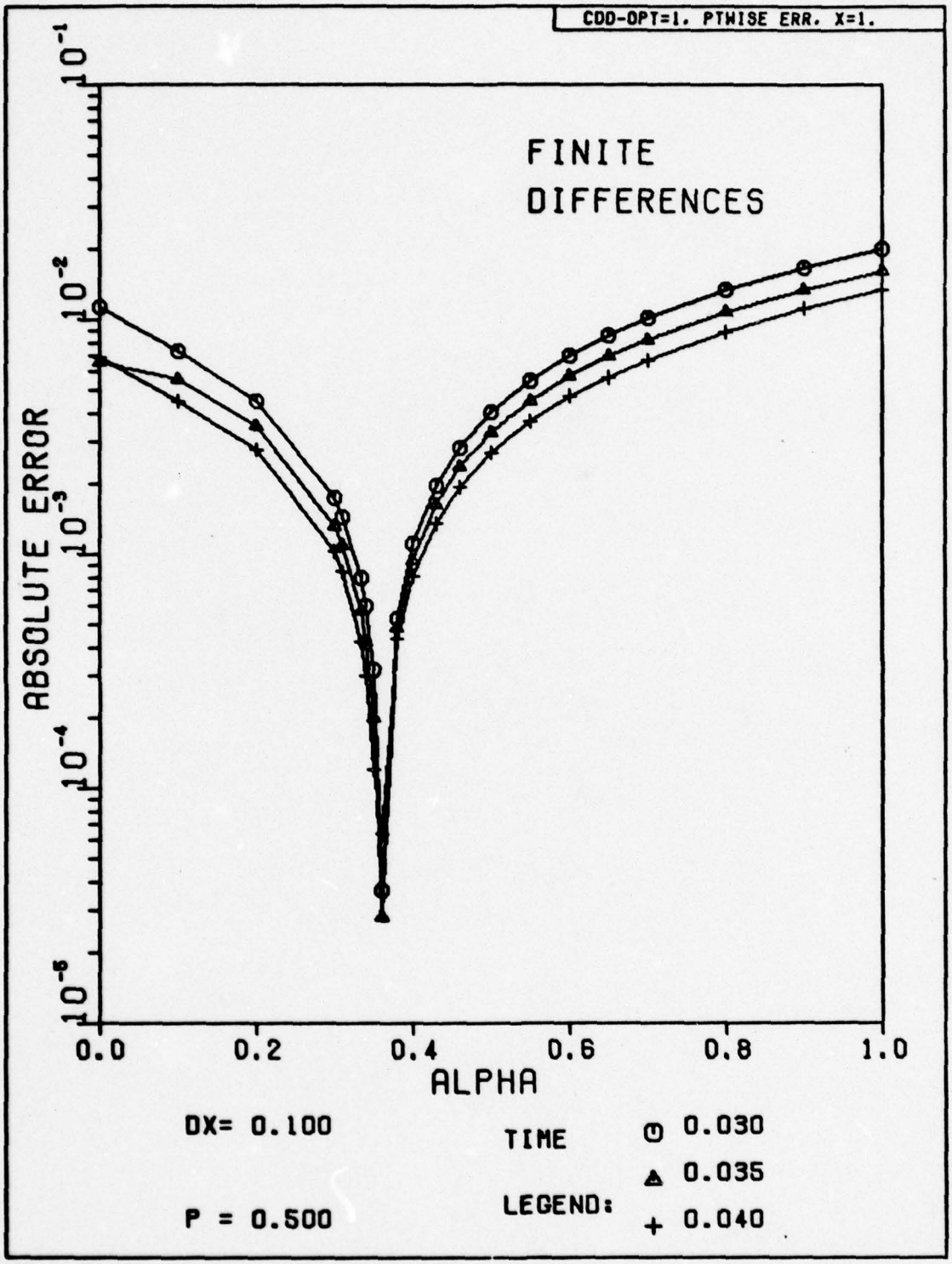


Fig. H-13. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

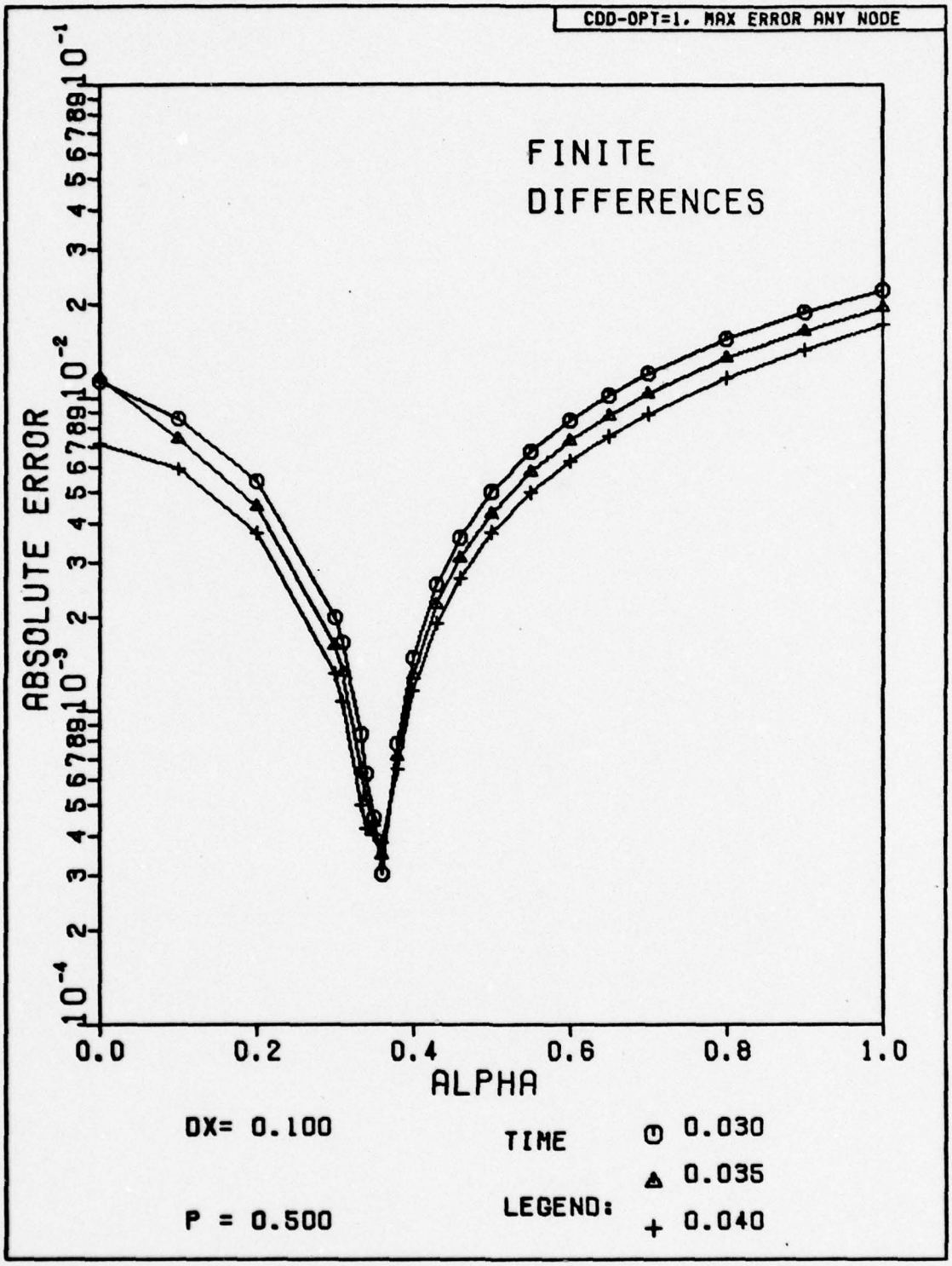


Fig. H-14. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

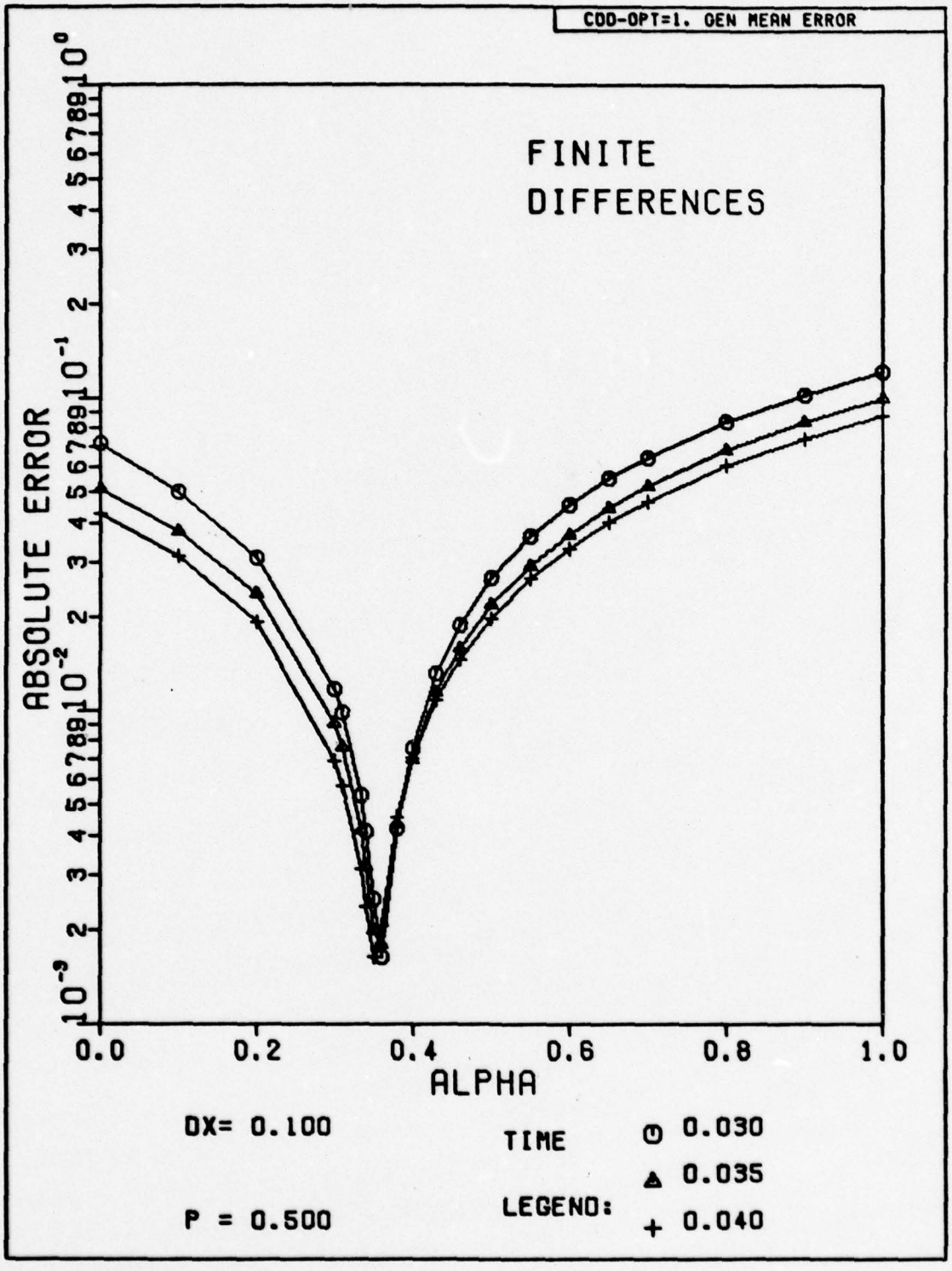


Fig. H-15. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

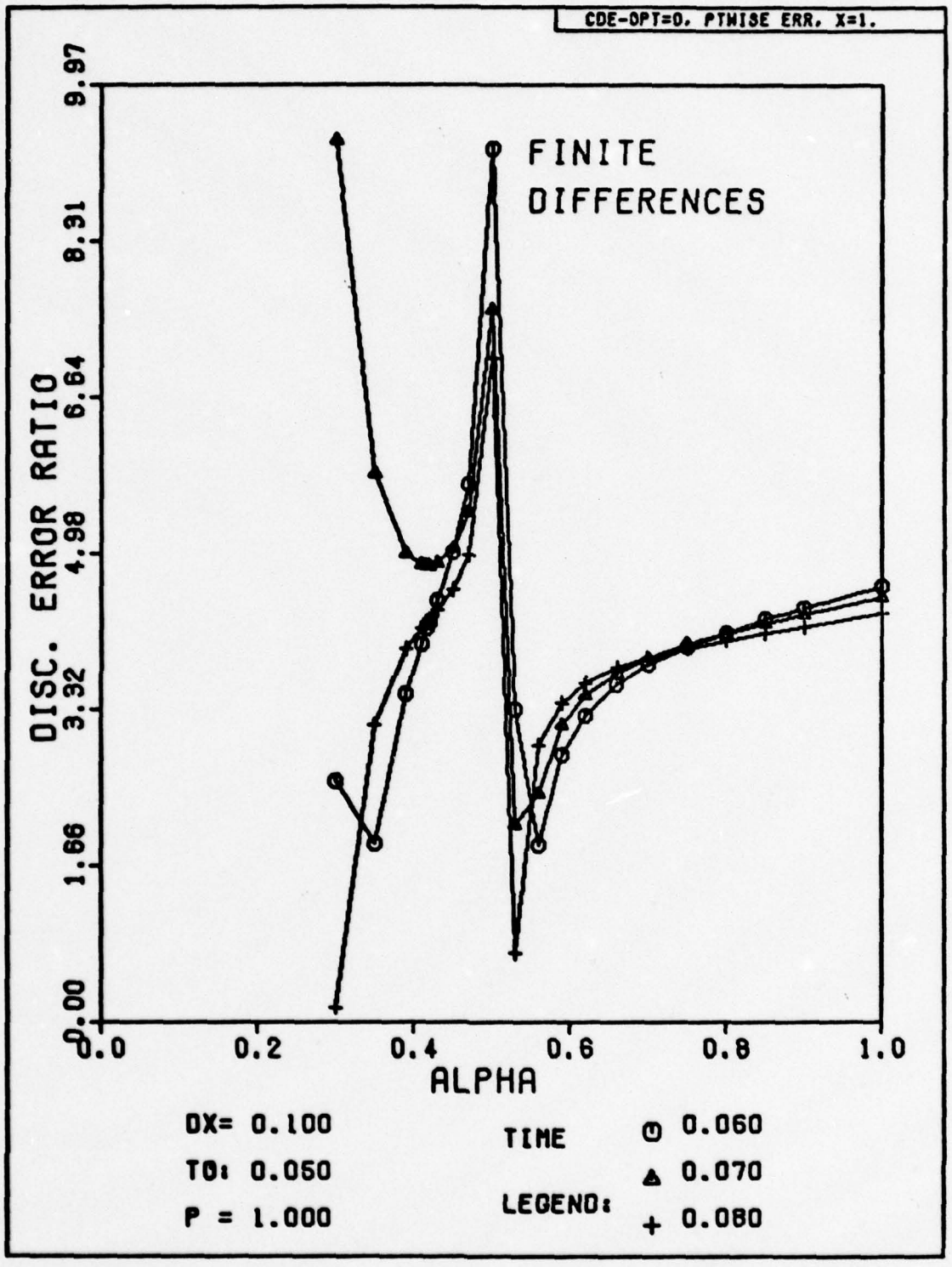


Fig. H-16. Discretization Error Ratio Versus Alpha for Problem One.

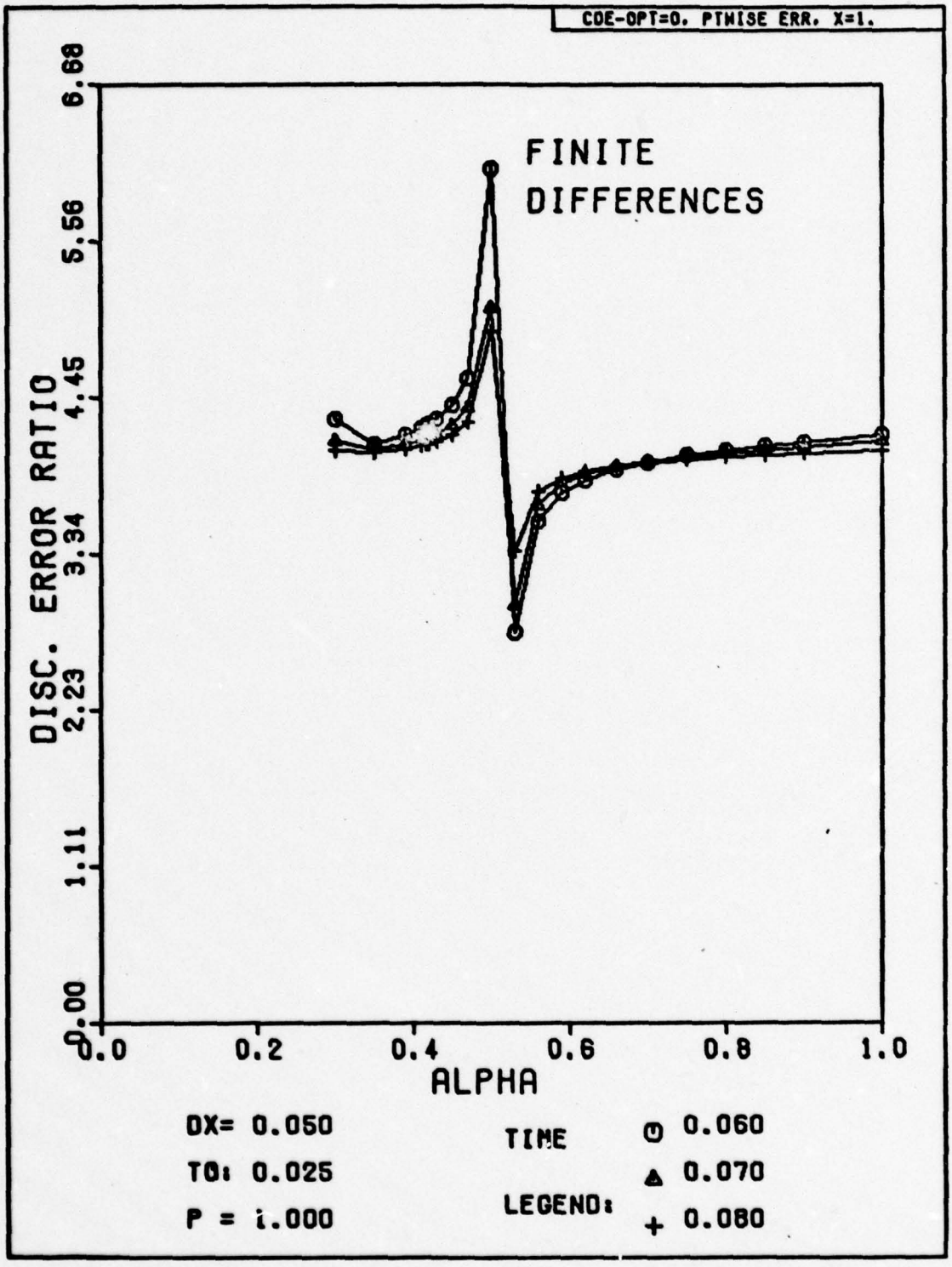


Fig. H-17. Discretization Error Ratio Versus Alpha for Problem One.

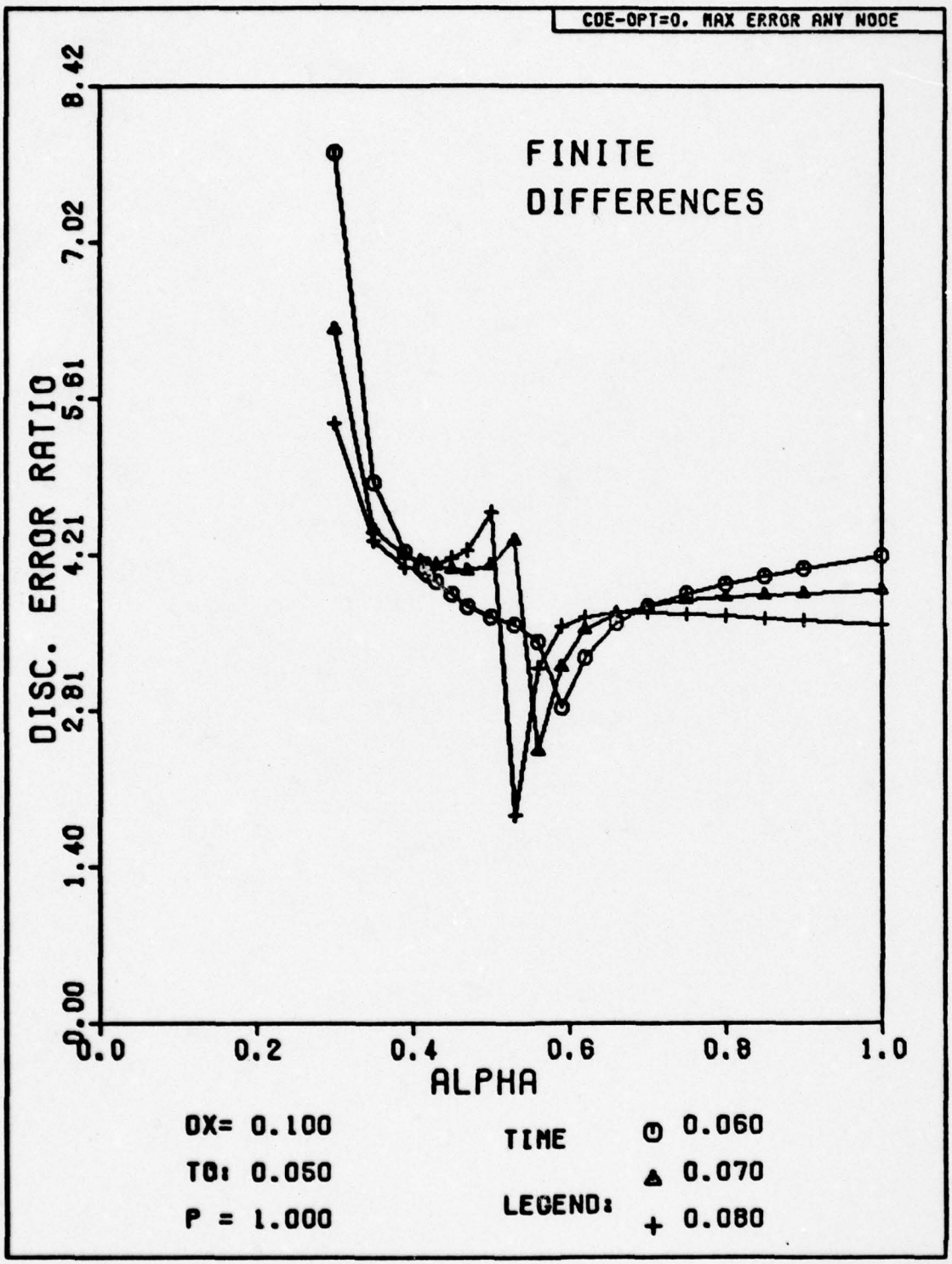


Fig. H-18. Discretization Error Ratio Versus Alpha for Problem One.

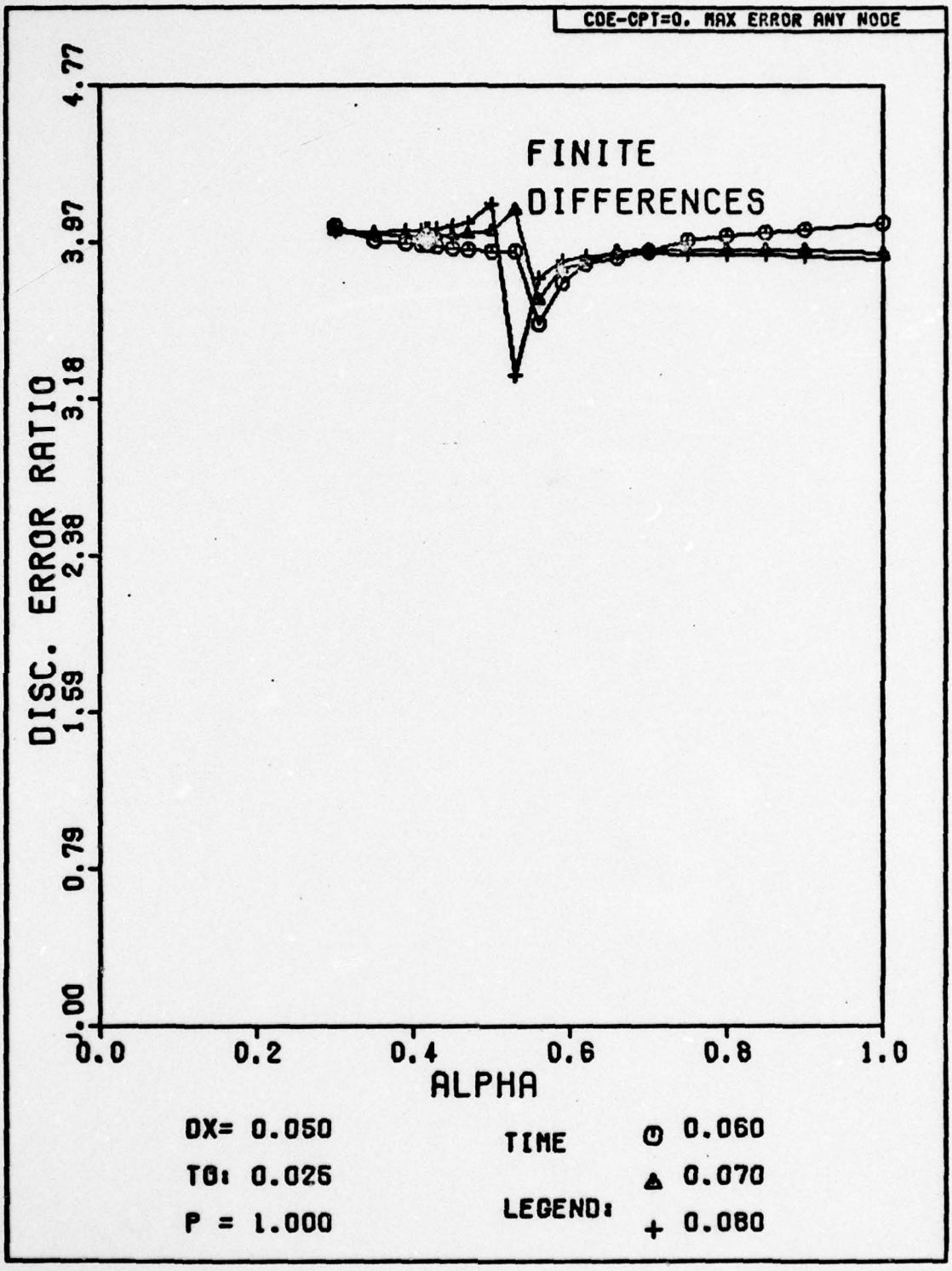


Fig. H-19. Discretization Error Ratio Versus Alpha for Problem One.

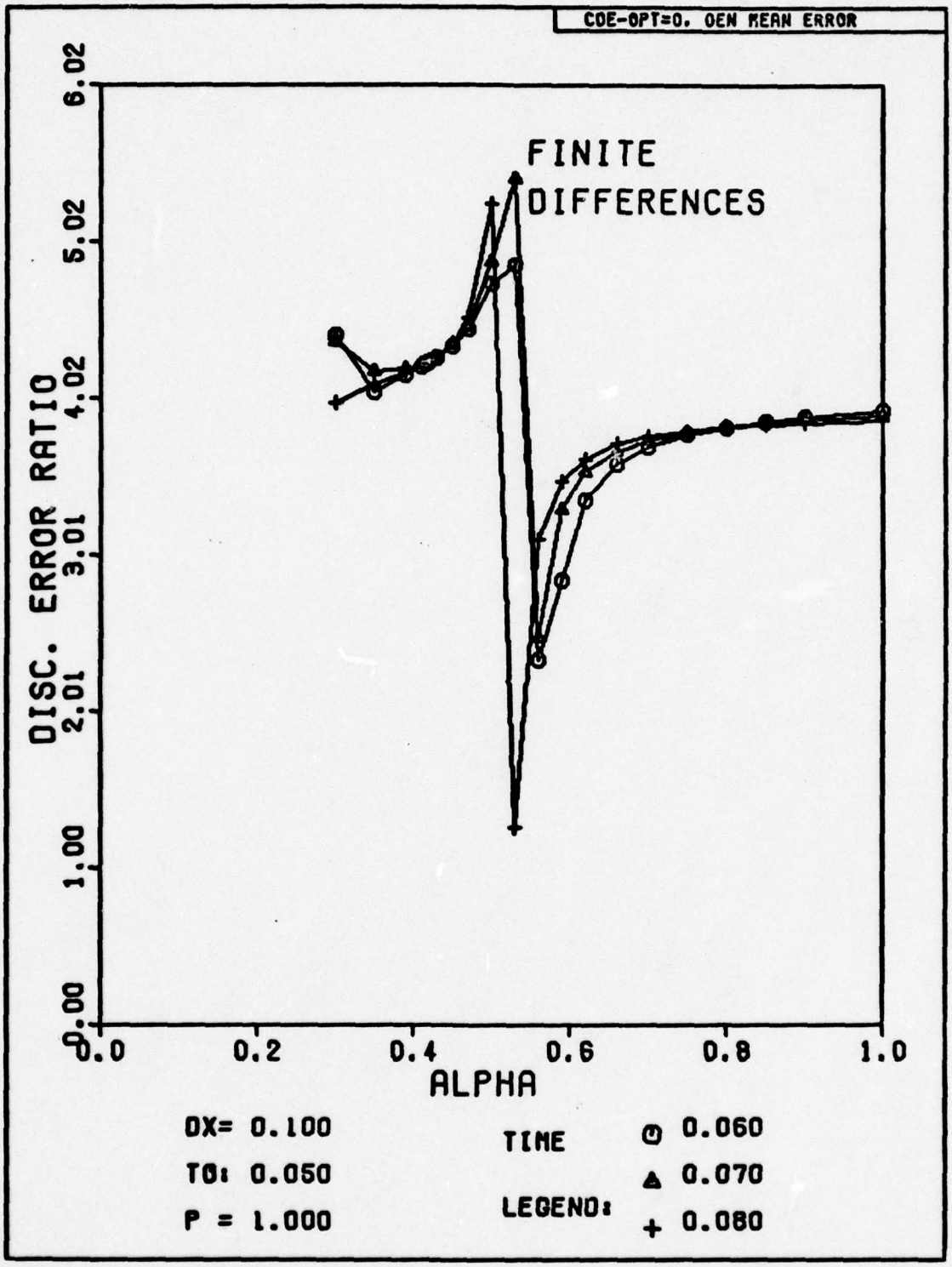


Fig. H-20. Discretization Error Ratio Versus Alpha for Problem One.

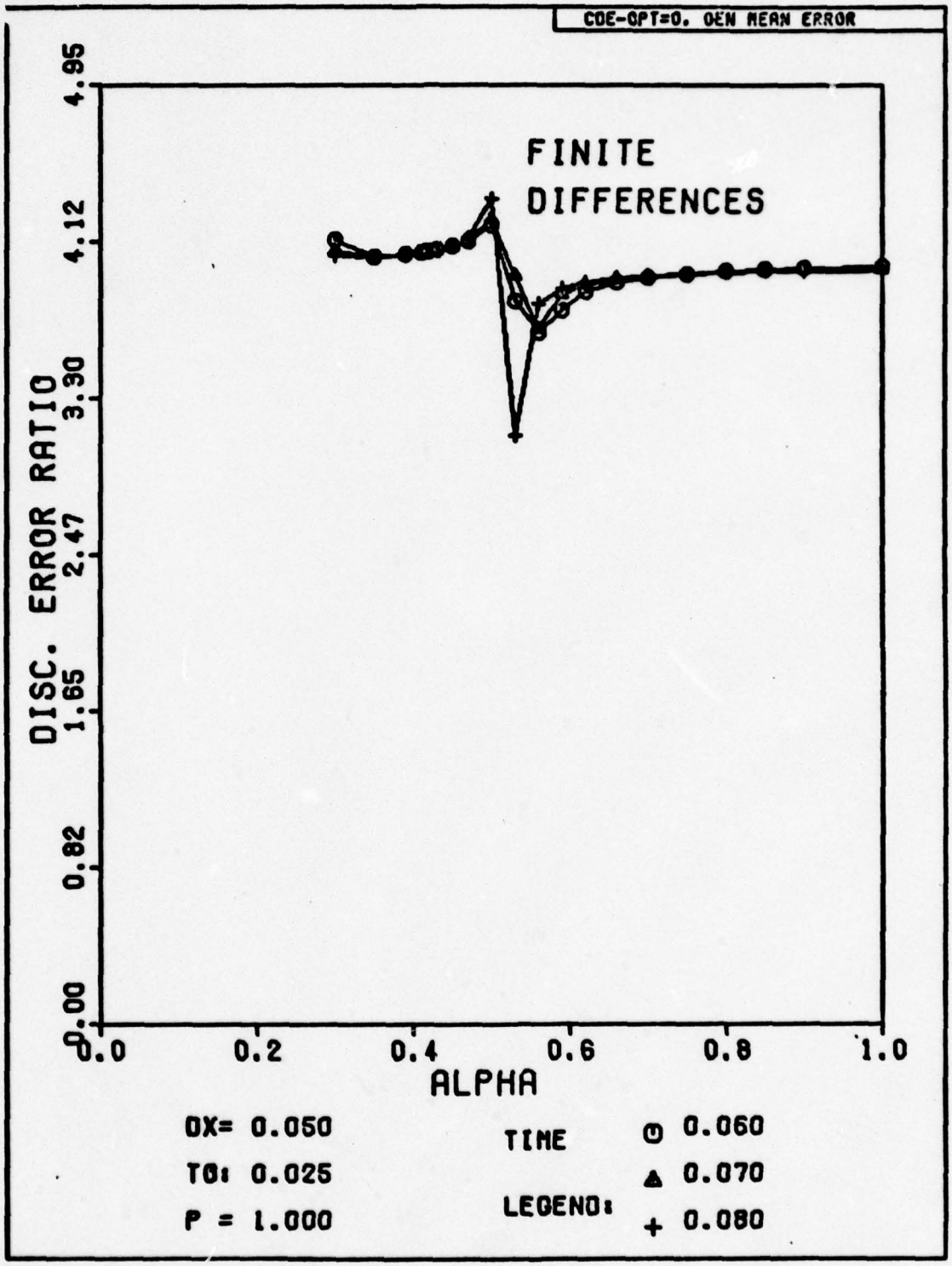


Fig. H-21. Discretization Error Ratio Versus Alpha for Problem One.

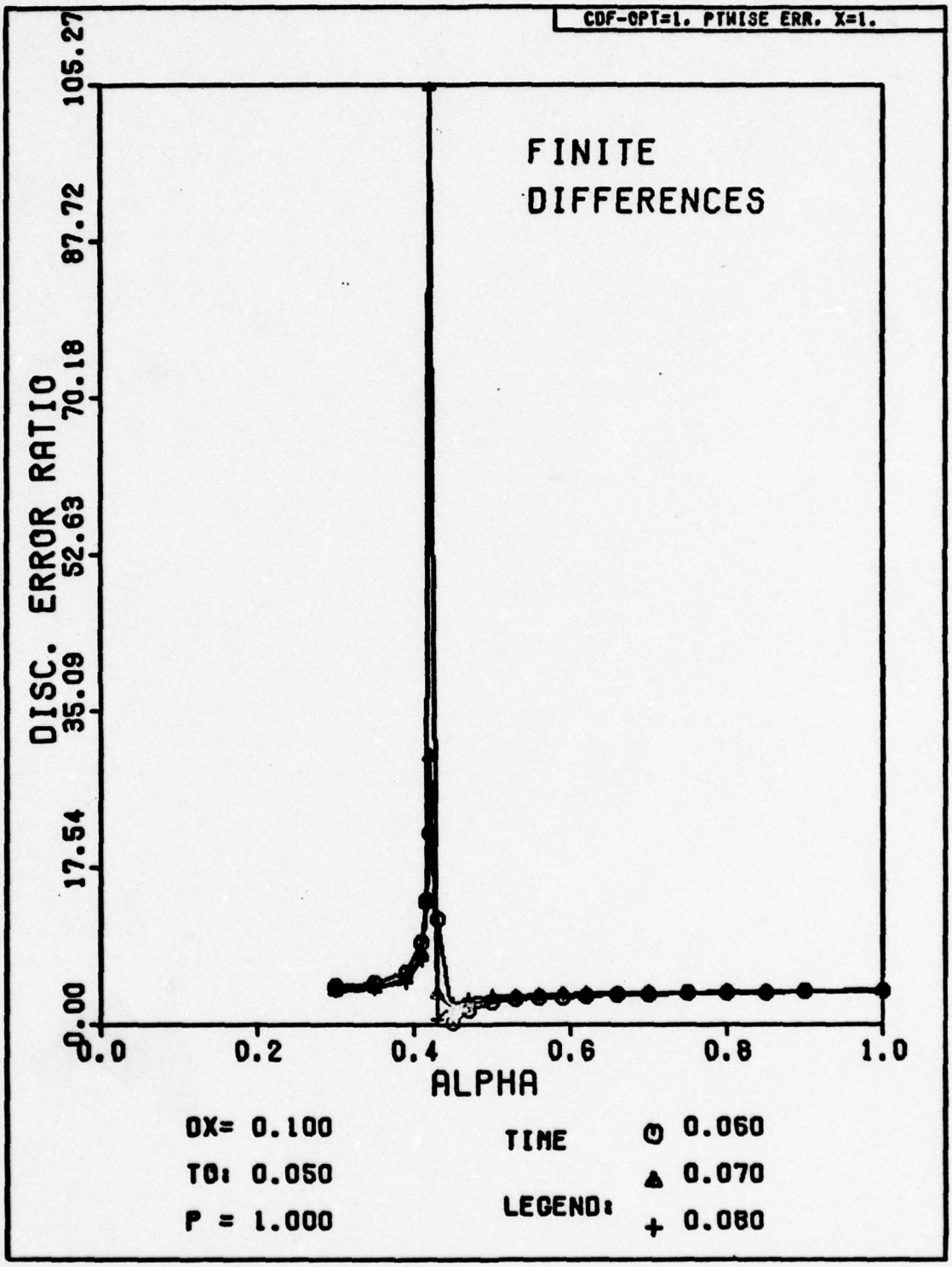
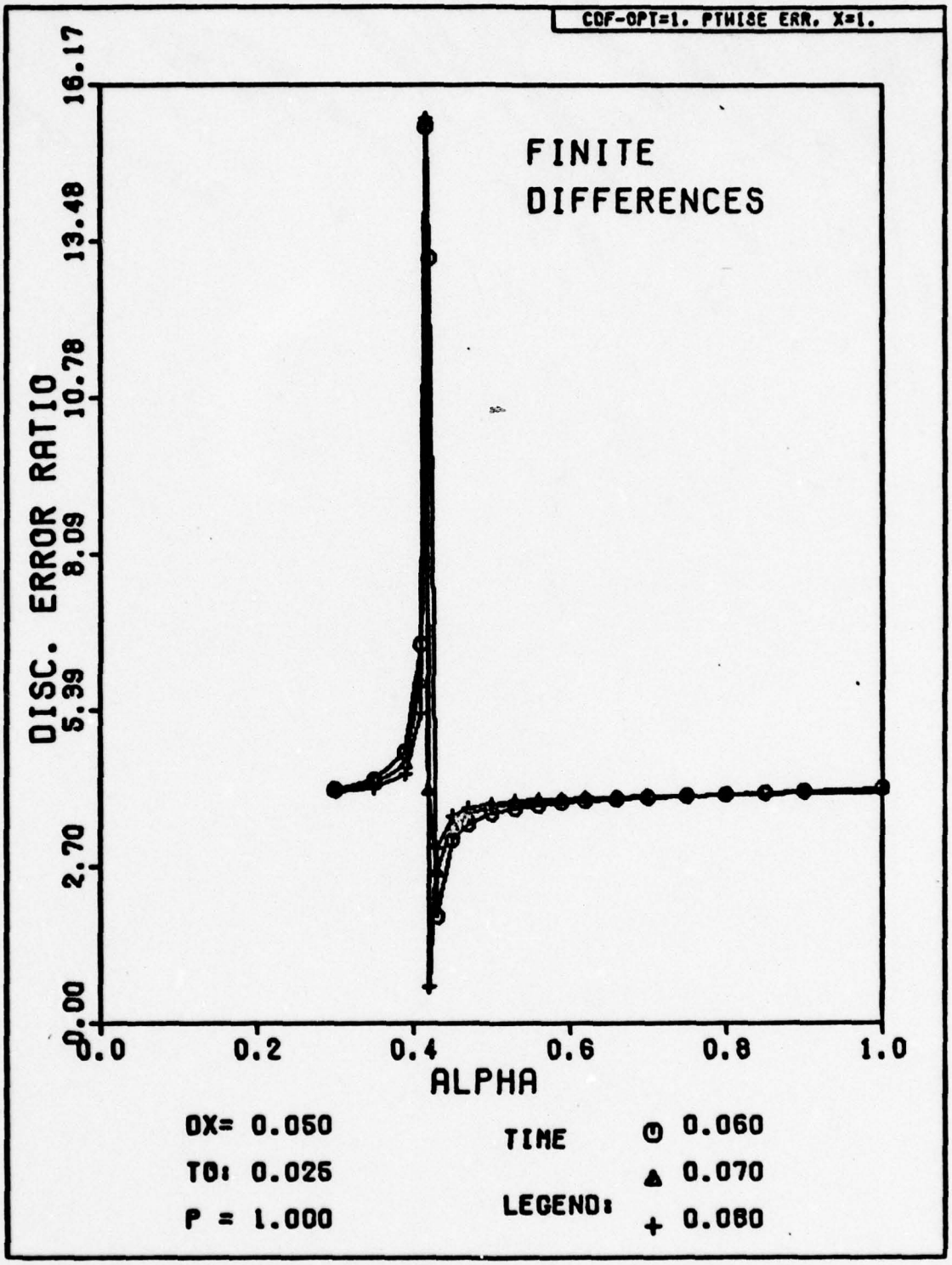


Fig. H-22. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.



H-23. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

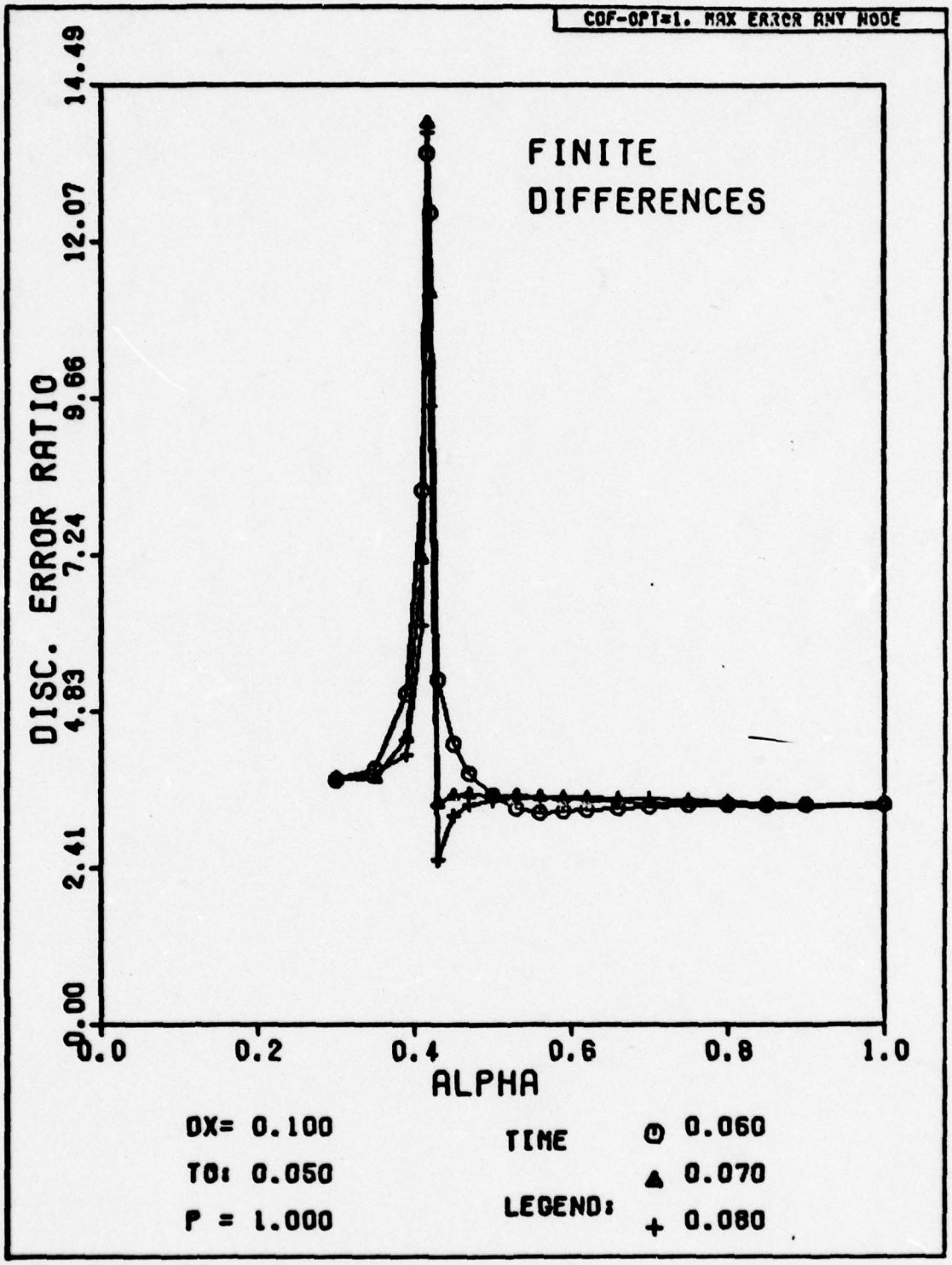


Fig. H-24. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

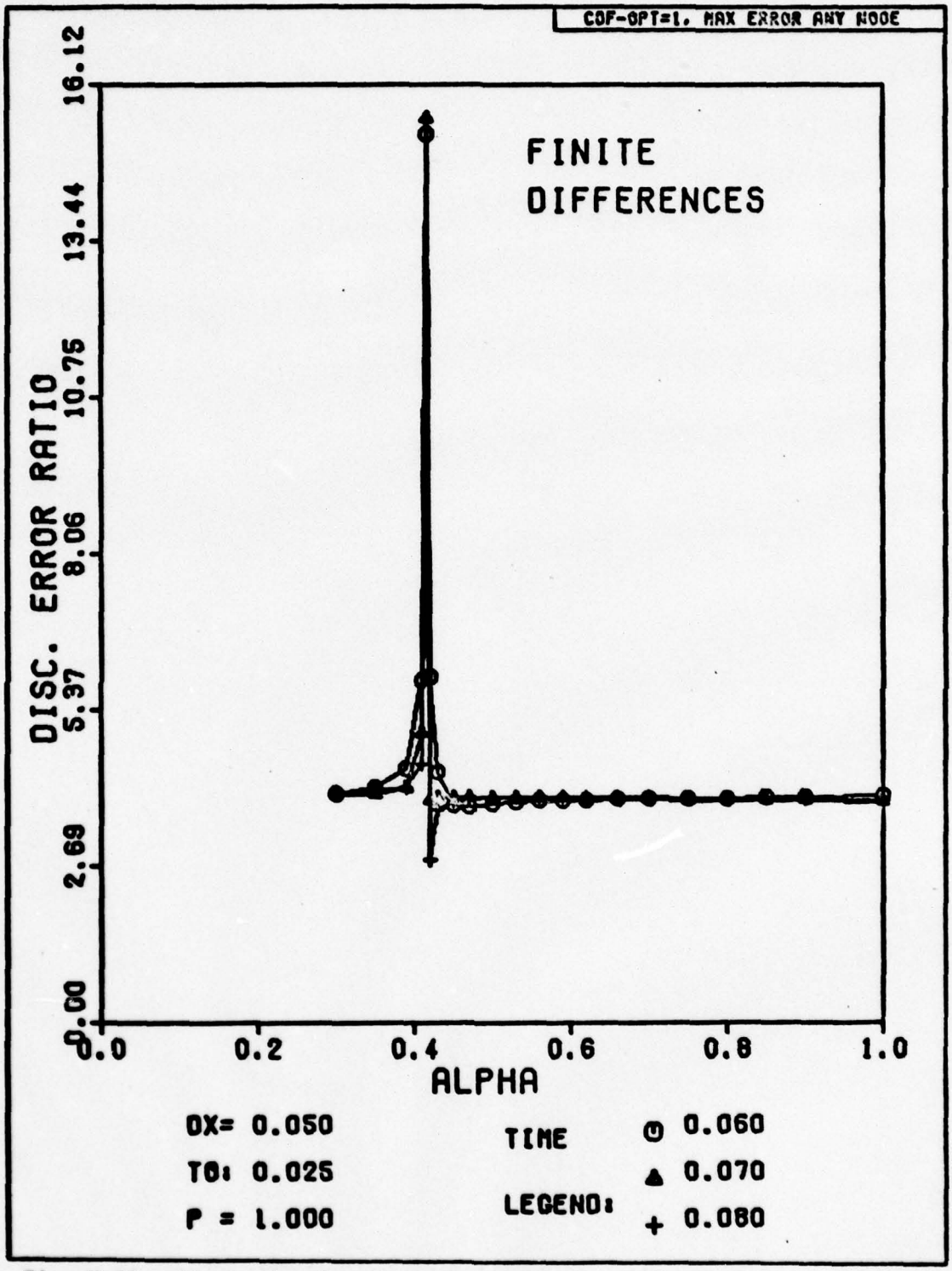


Fig. H-25. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

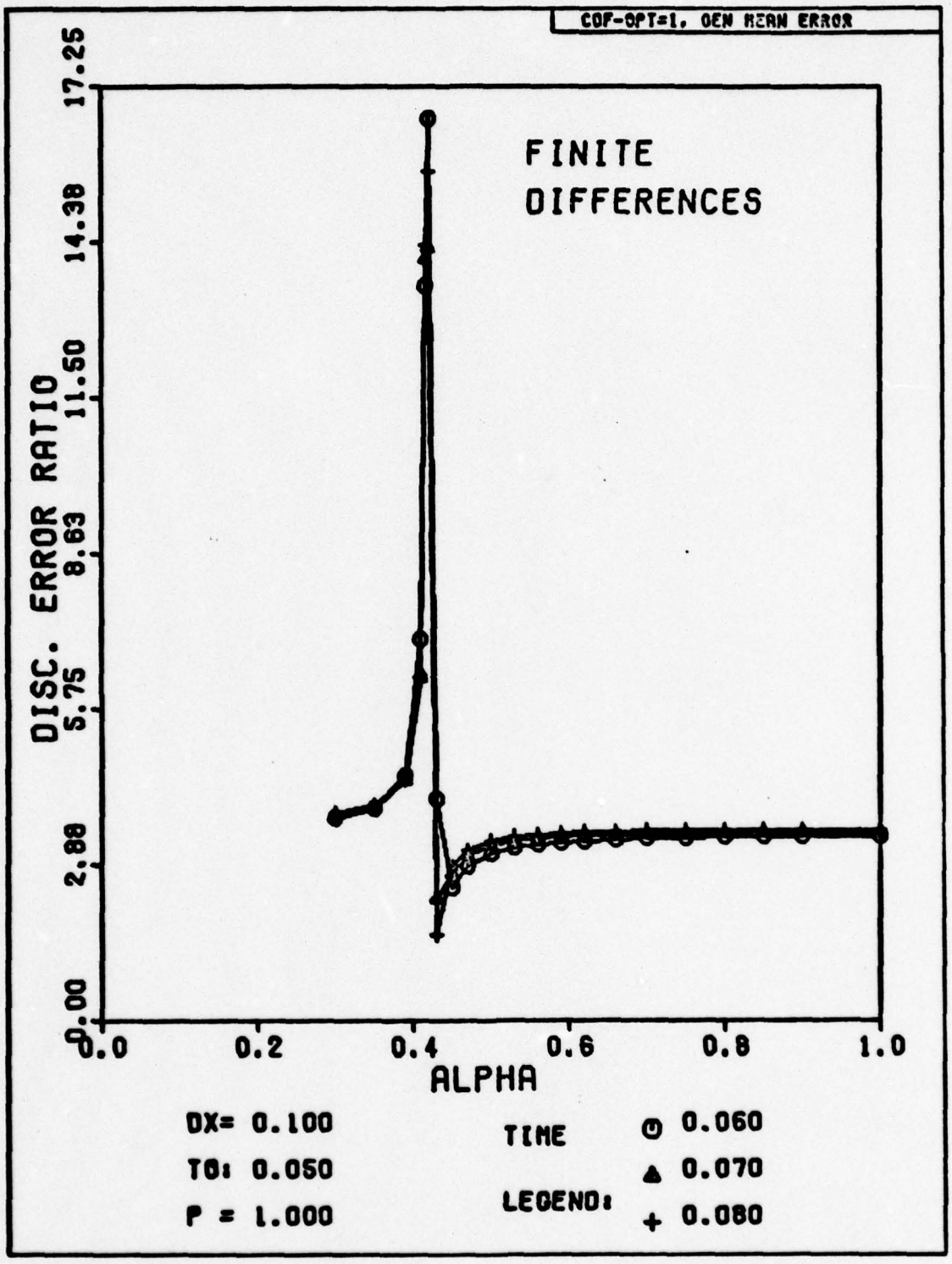
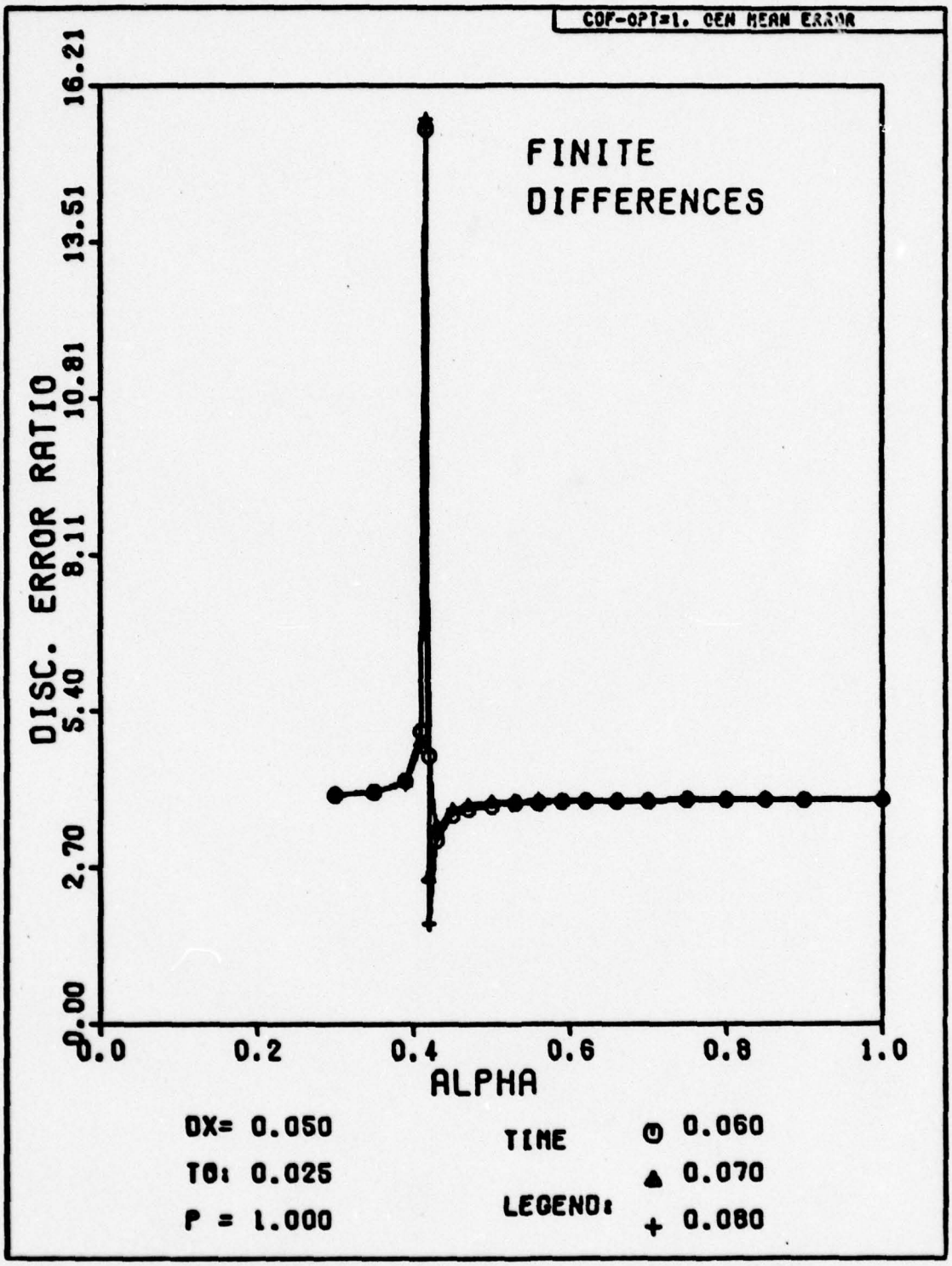


Fig. H-26. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.



F ig. H-27. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

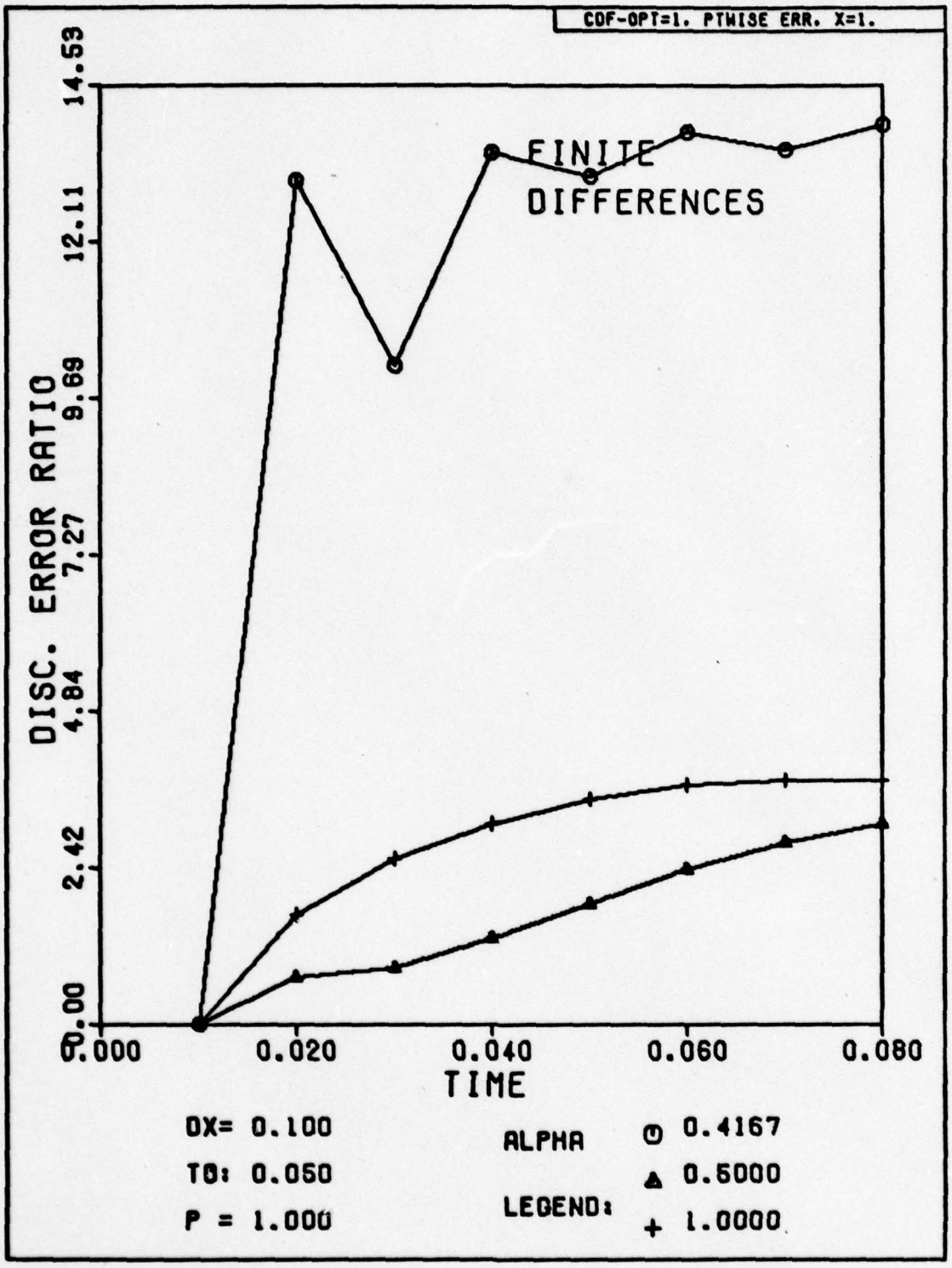


Fig. H-28. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

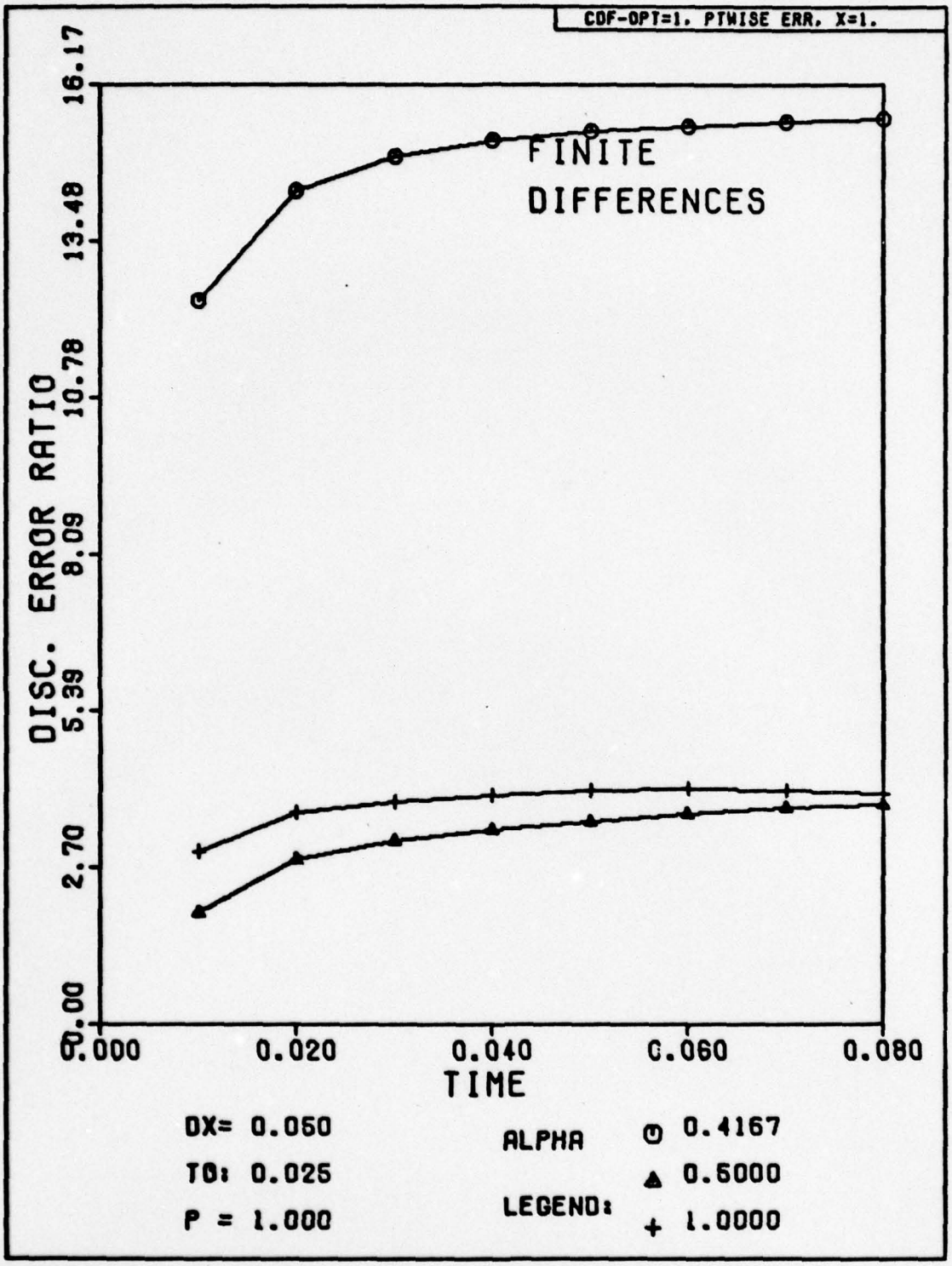


Fig. H-29. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

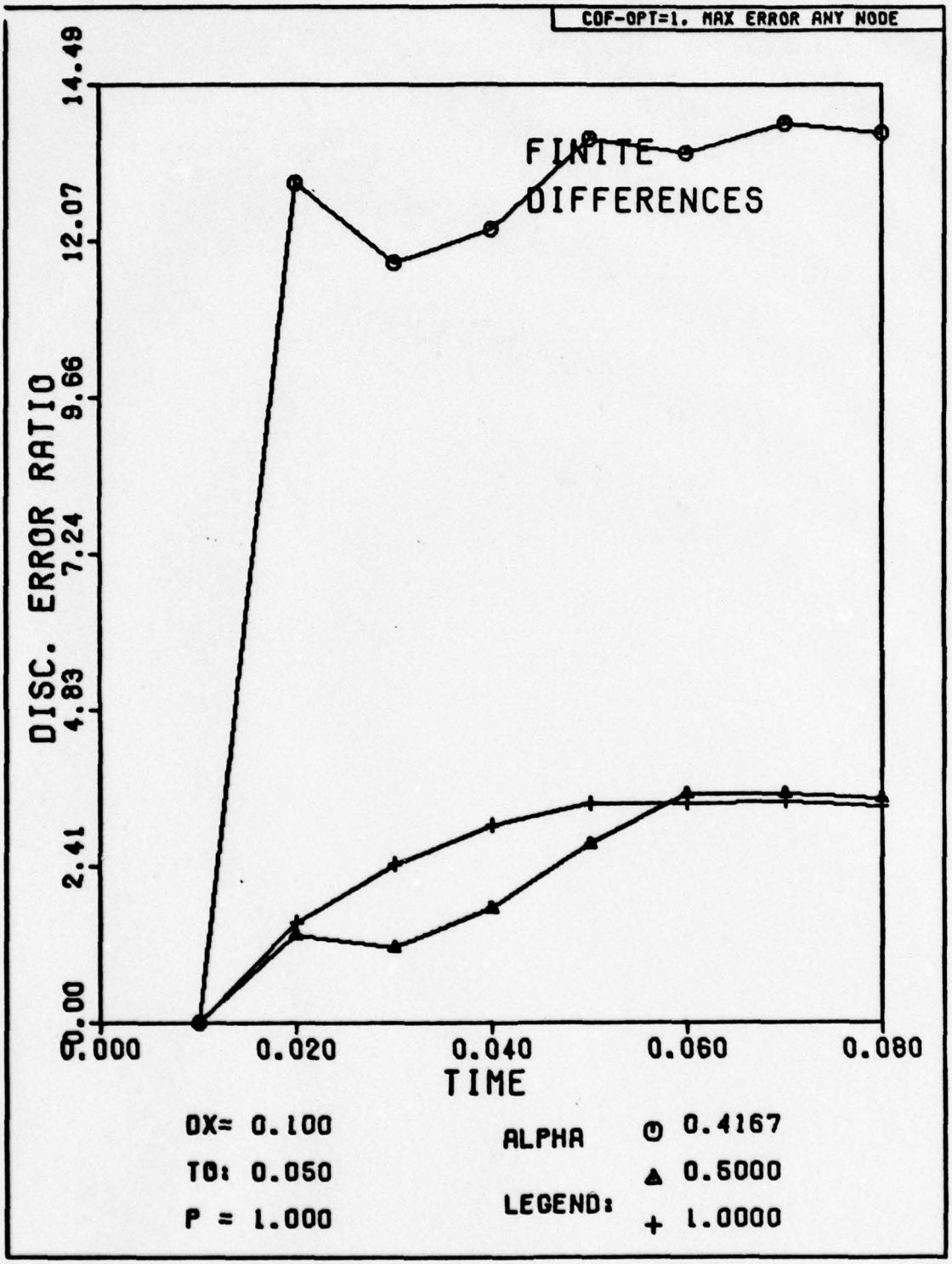


Fig. H-30. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

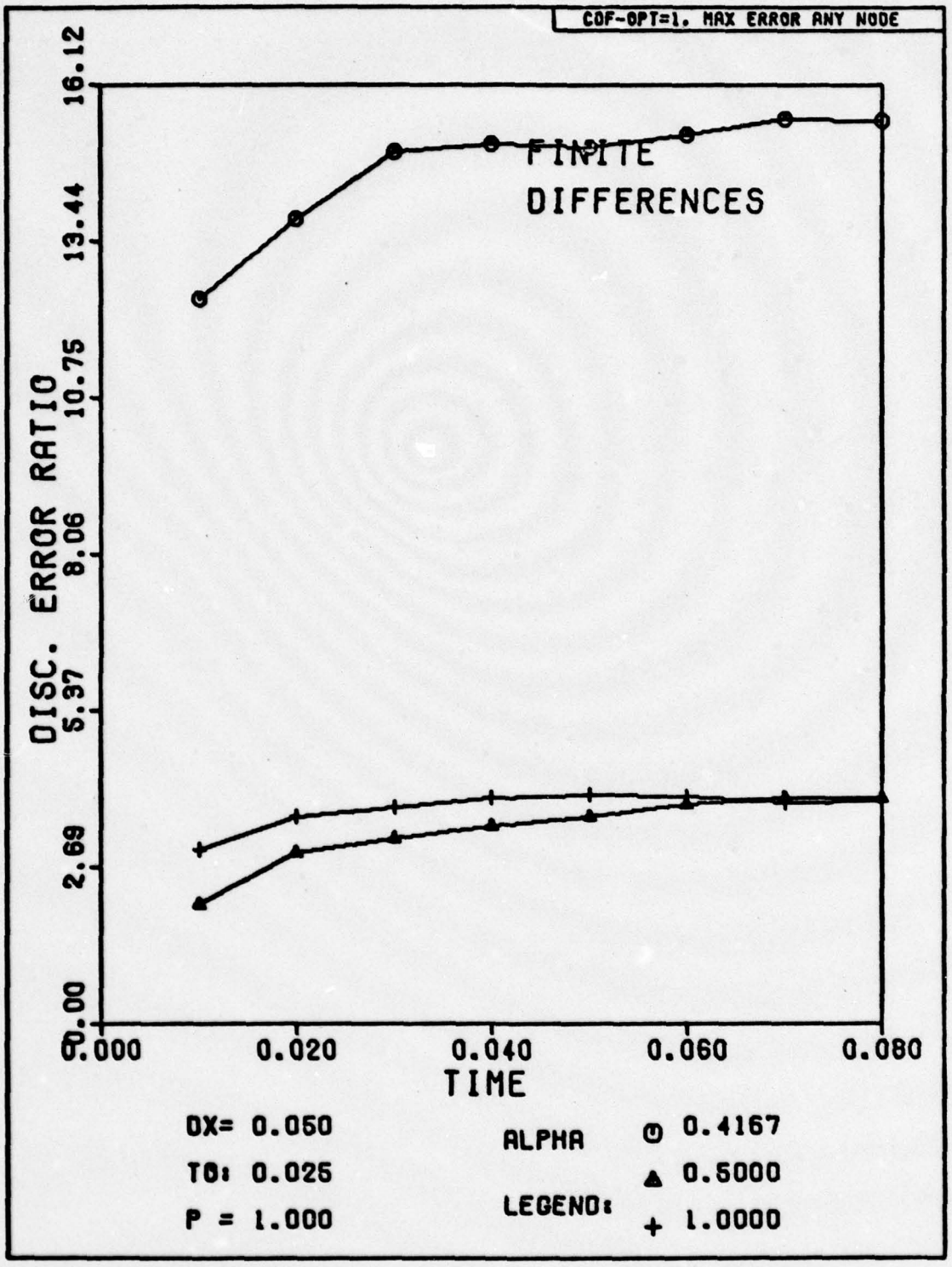


Fig. H-31. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

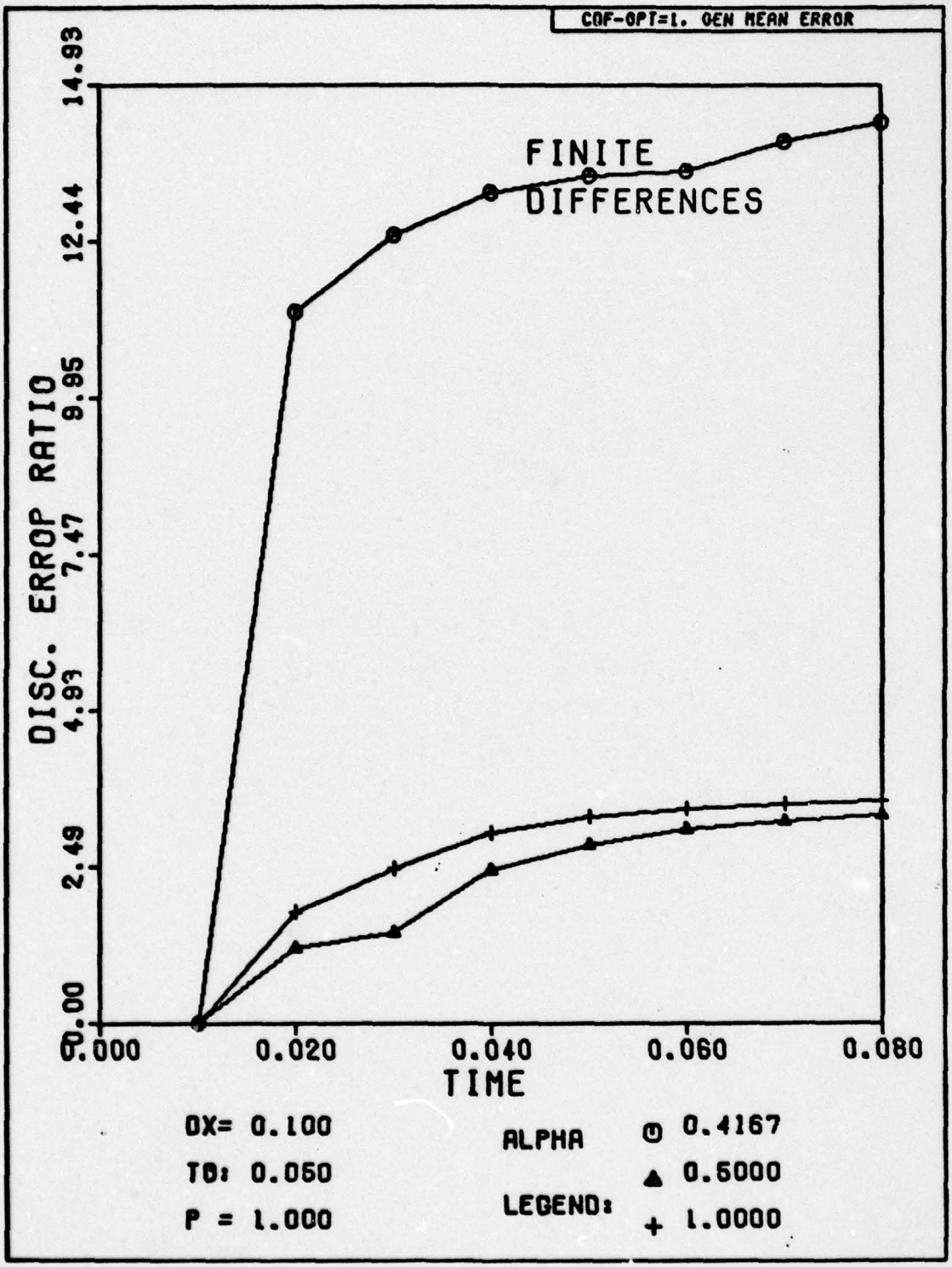


Fig. H-32. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

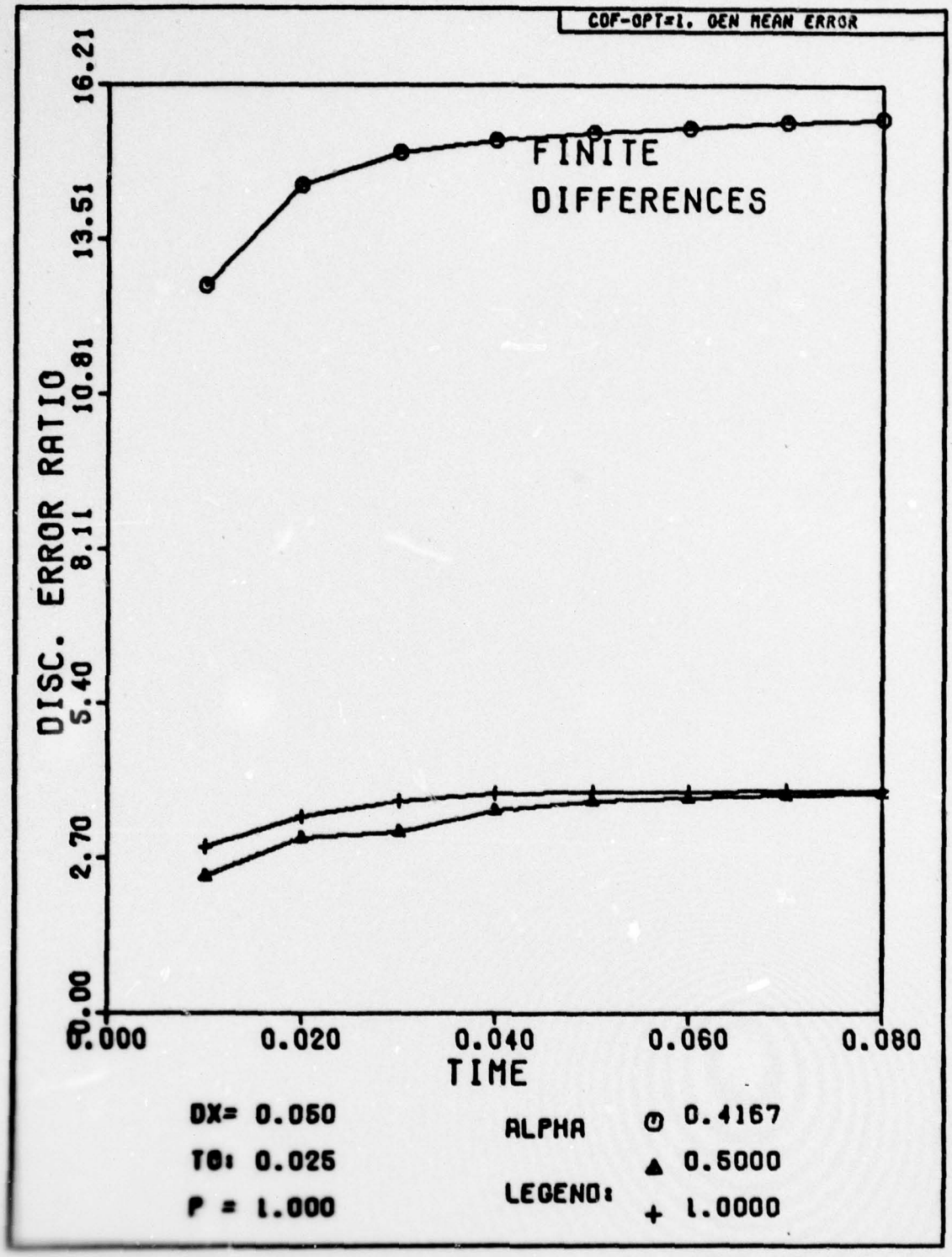


Fig. 8-33. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

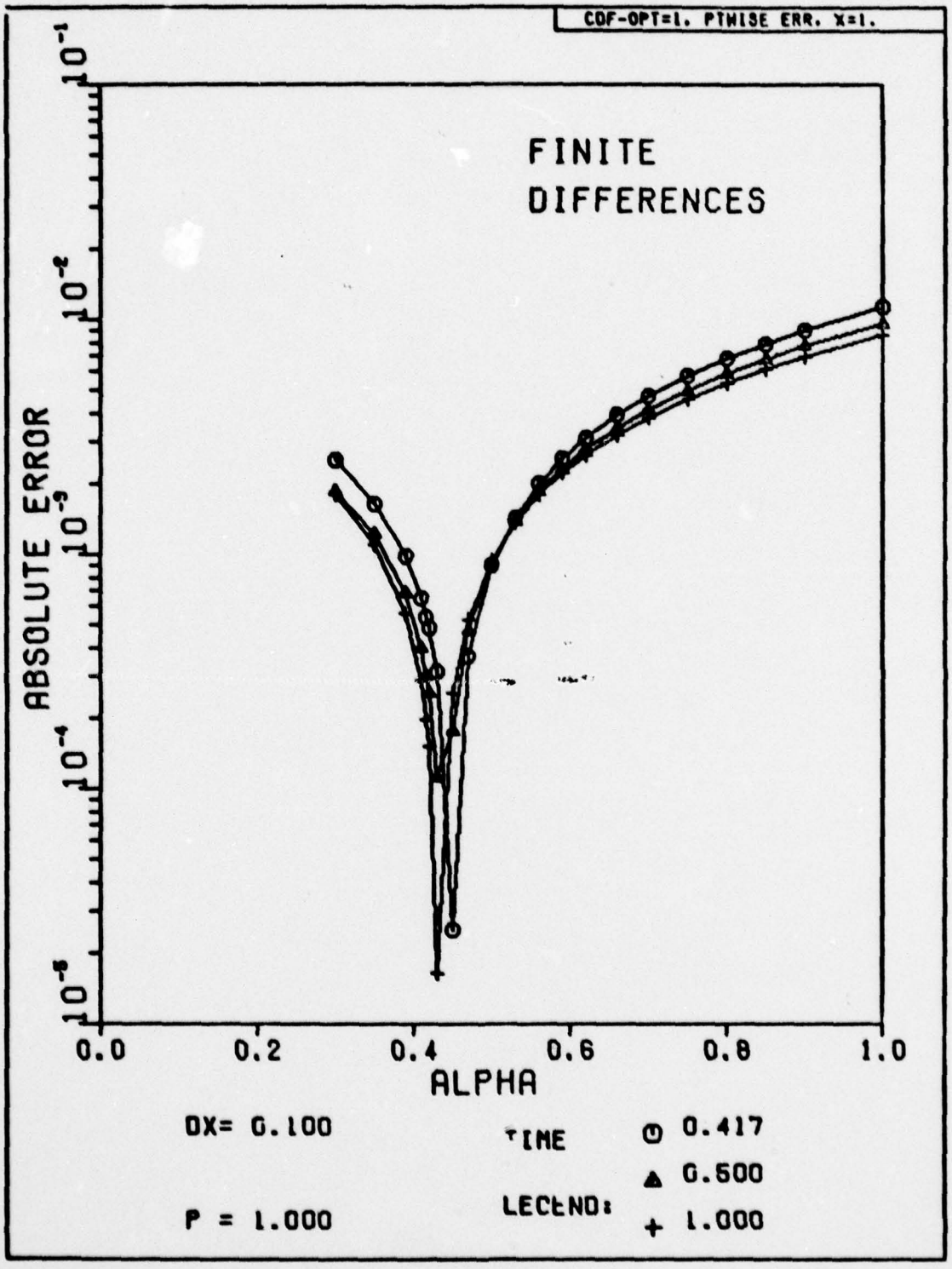


Fig. H-34. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

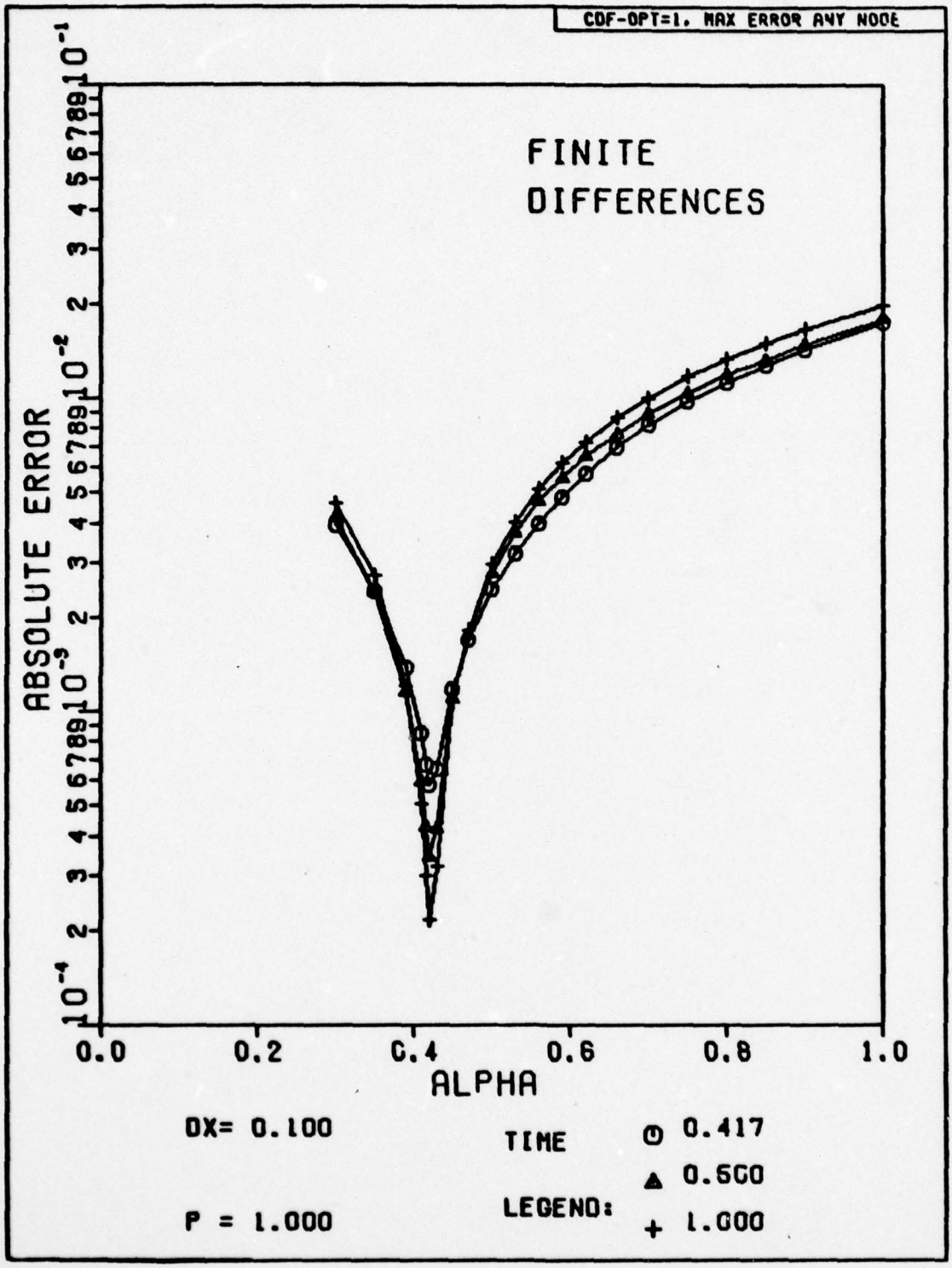


Fig. H-35. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

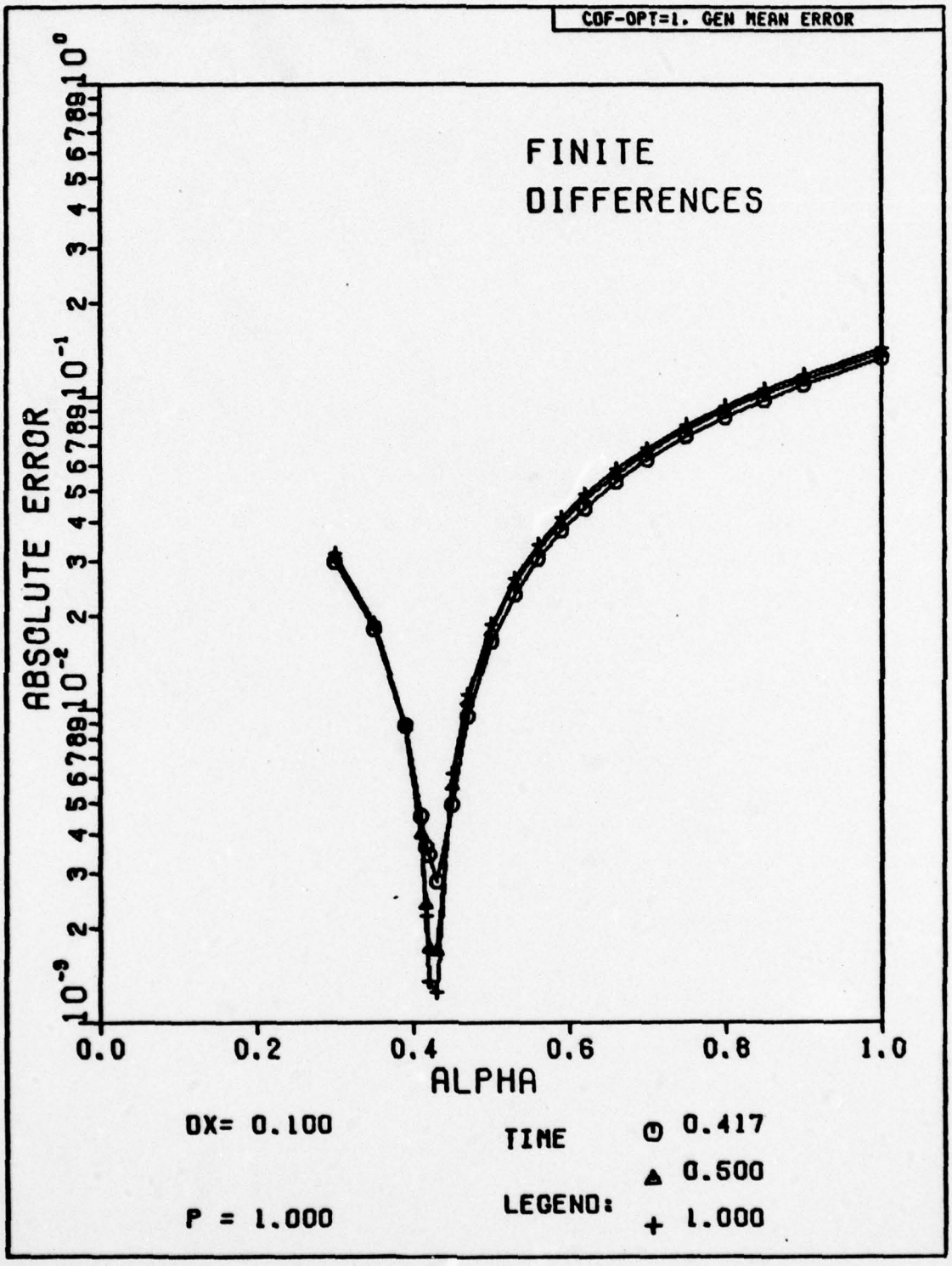


Fig. H-36. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

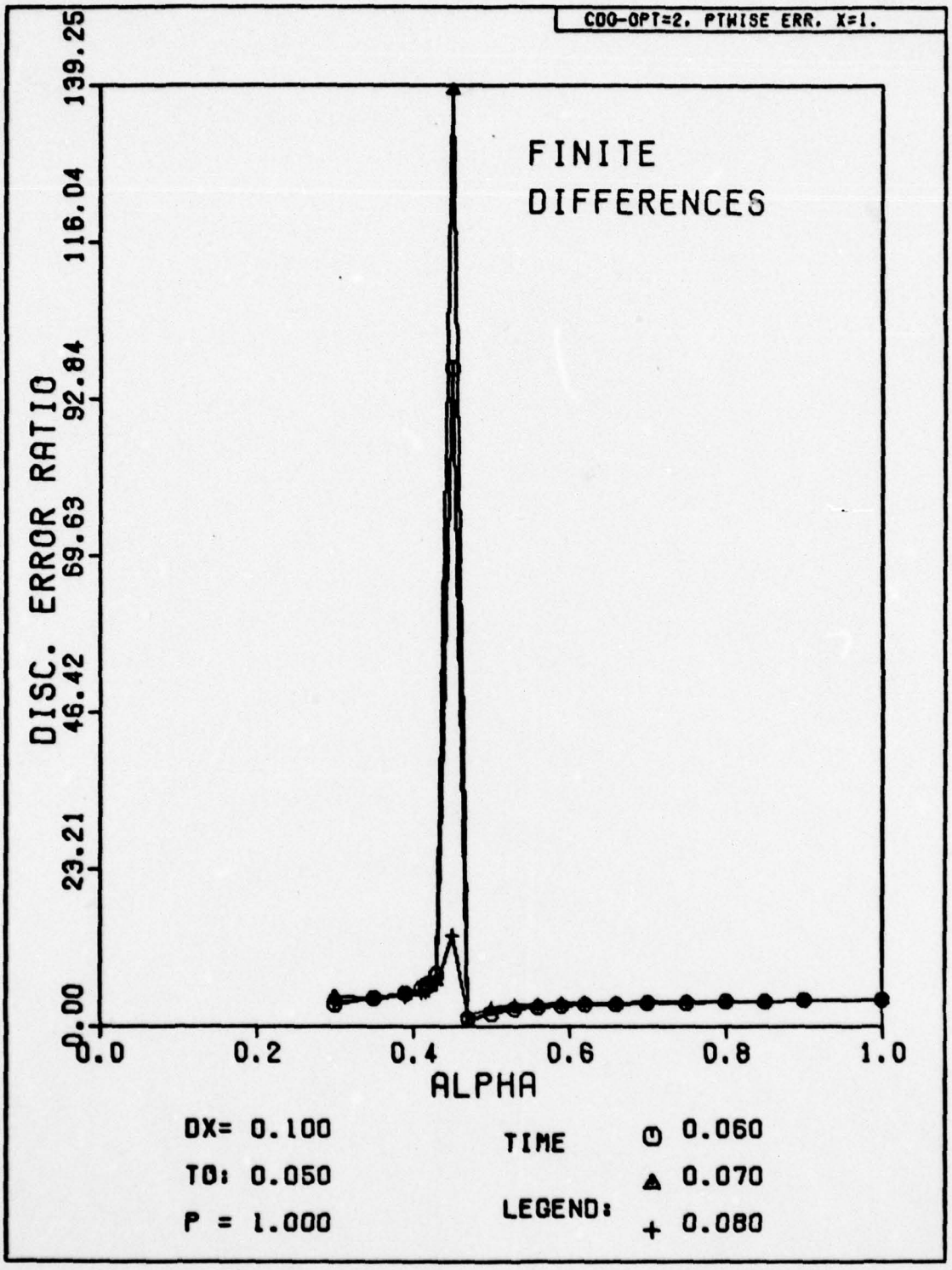


Fig. H-37. Discretization Error Ratio Versus Alpha for Problem One. The space-time grid has been subdivided for the first time step to obtain a more accurate solution.

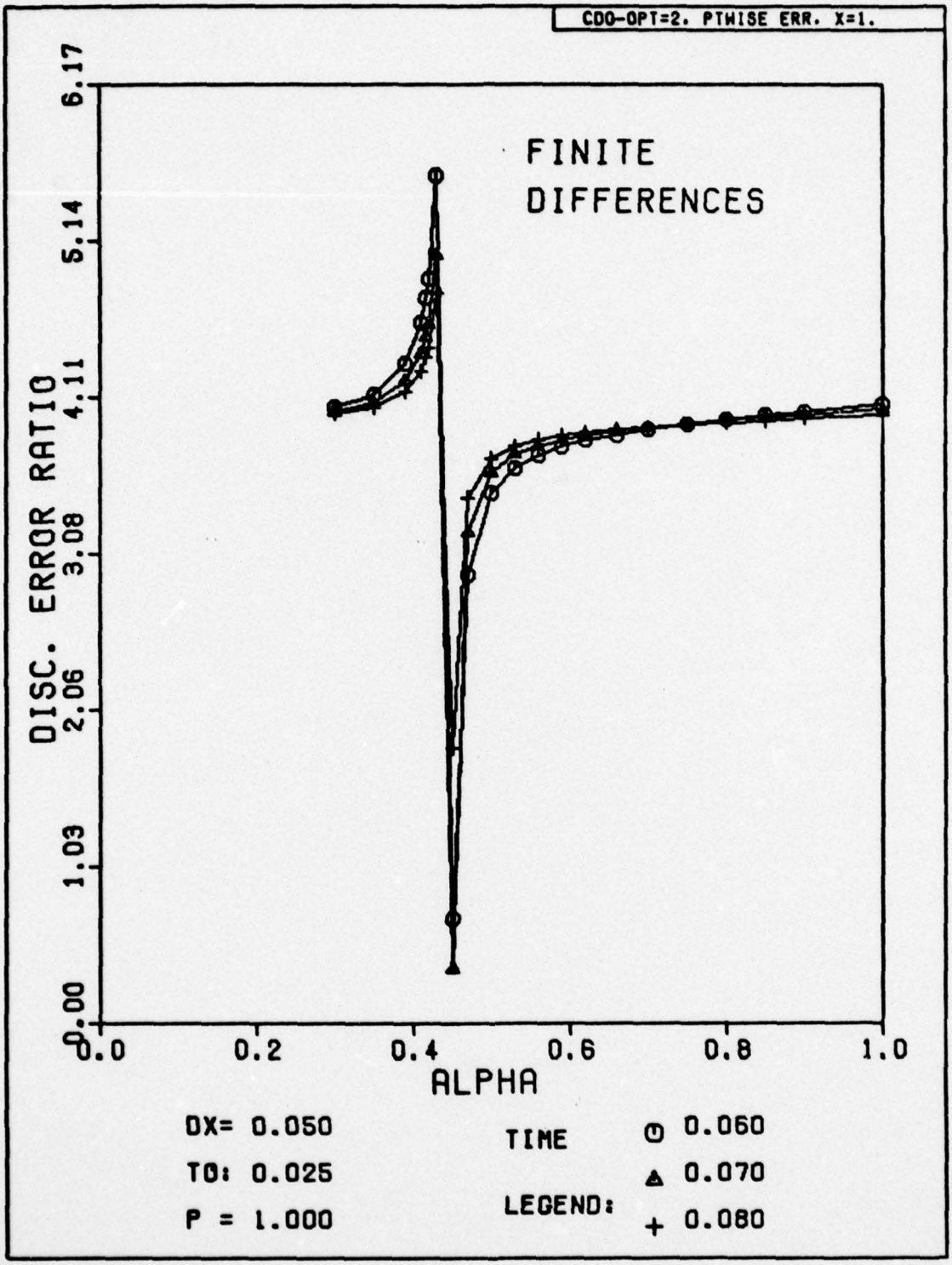


Fig. H-38. Discretization Error Ratio Versus Alpha for Problem One. The space-time grid has been subdivided for the first time step to obtain a more accurate solution.

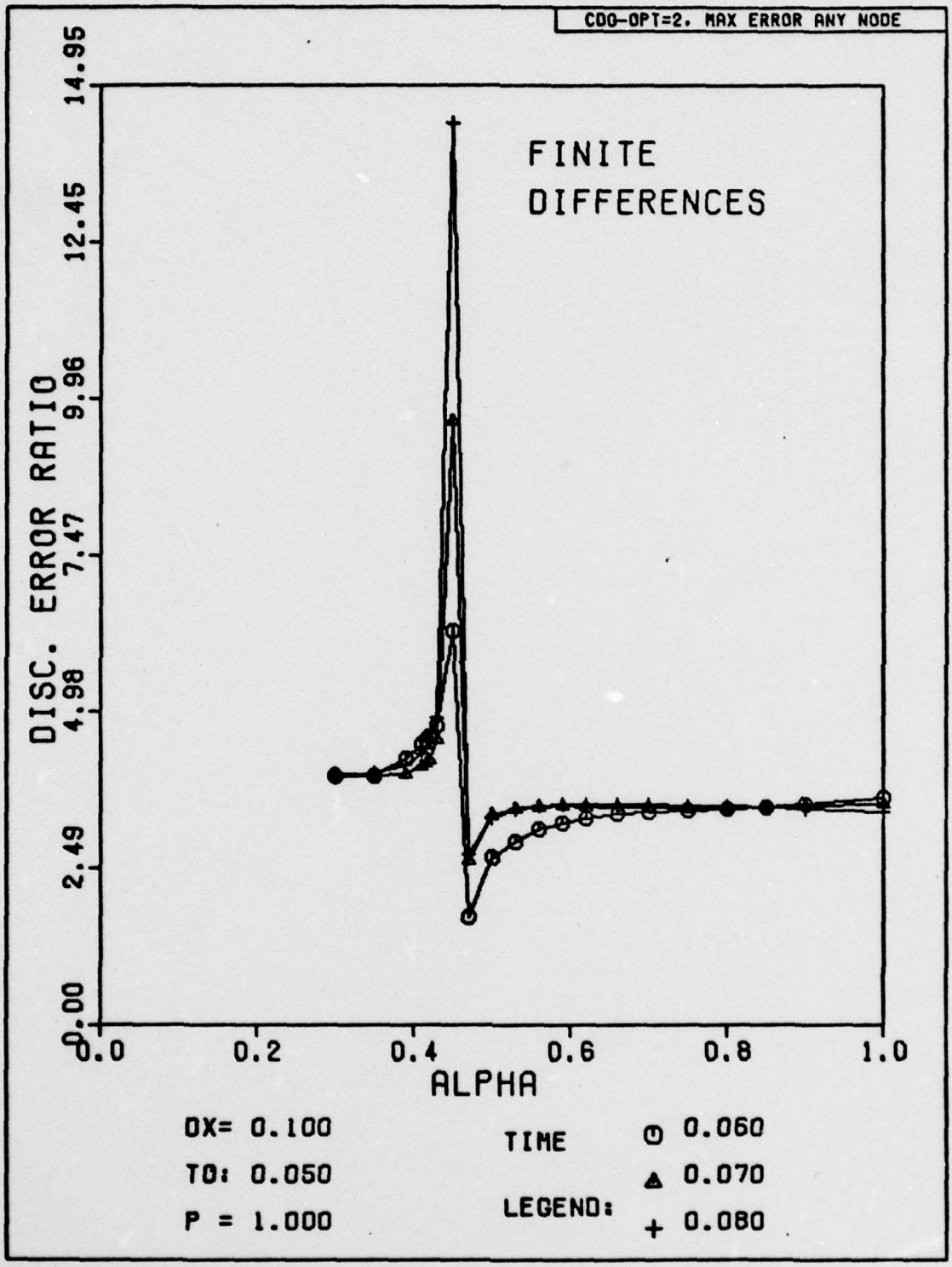


Fig. H-39. Discretization Error Ratio Versus Alpha for Problem One. The space-time grid has been subdivided for the first time step to obtain a more accurate solution.

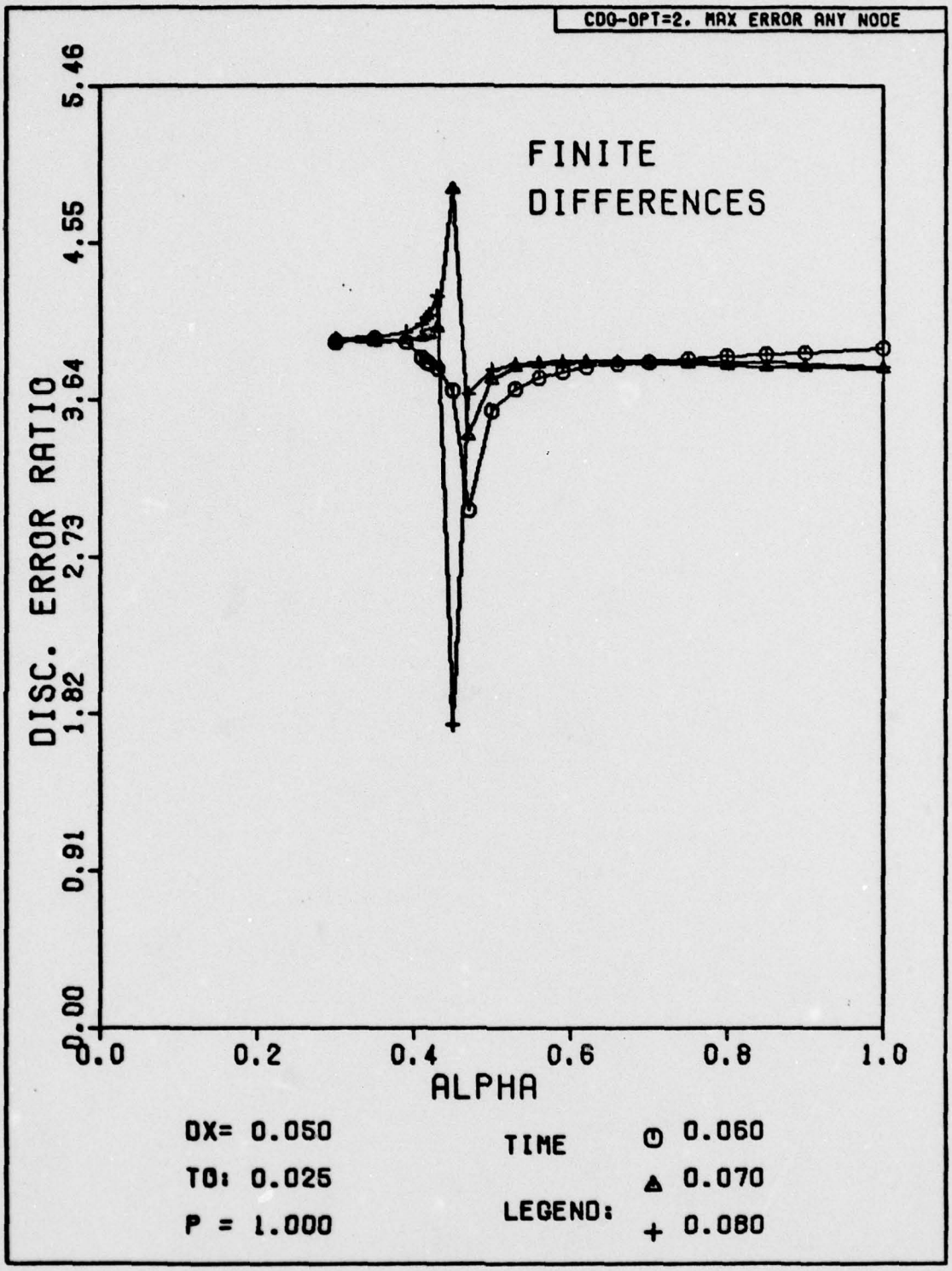


Fig. H-40. Discretization Error Ratio Versus Alpha for Problem One. The space-time grid has been subdivided for the first time step to obtain a more accurate solution.

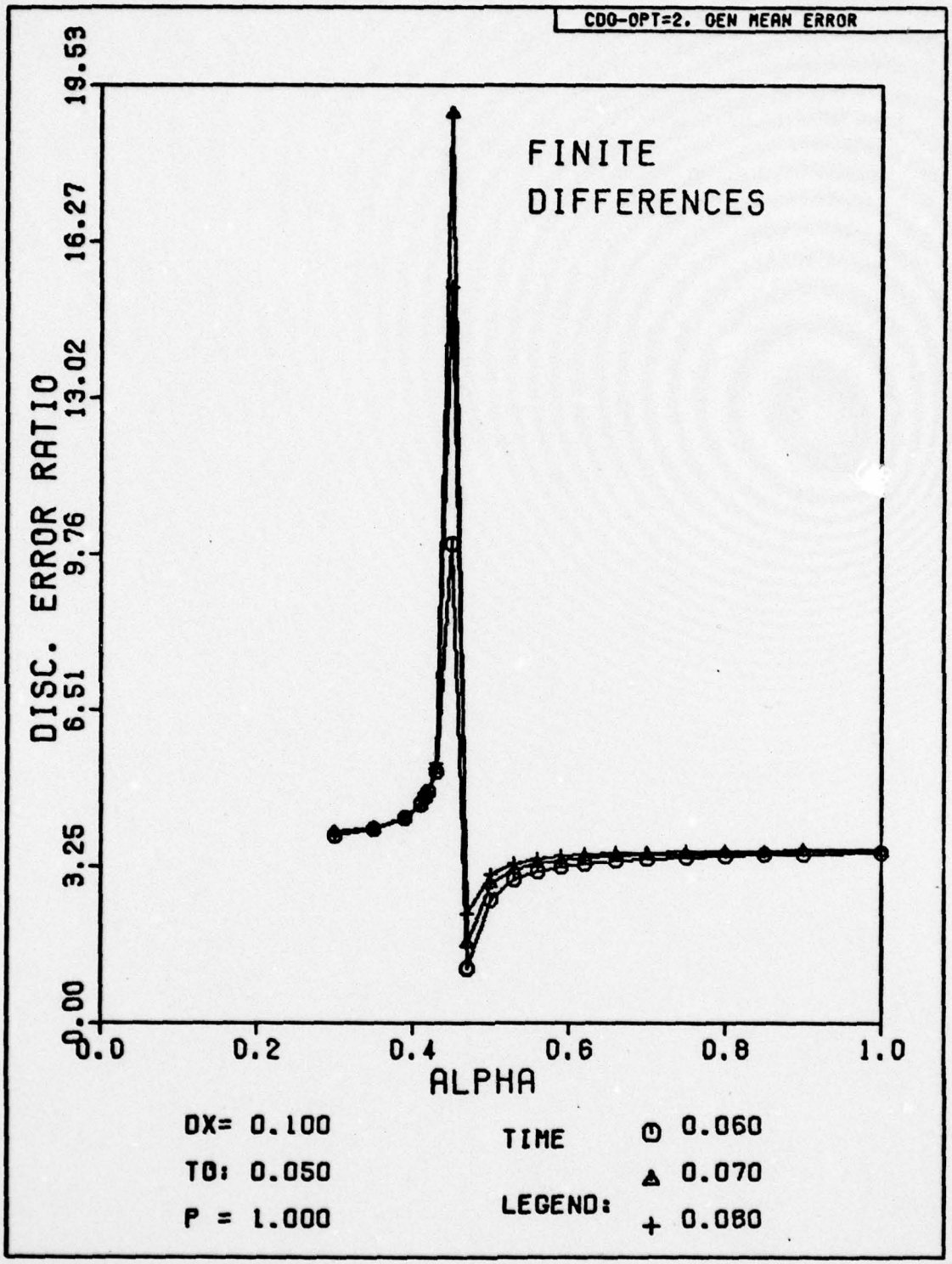


Fig. H-41. Discretization Error Ratio Versus Alpha for Problem One. The space-time grid has been subdivided for the first time step to obtain a more accurate solution.

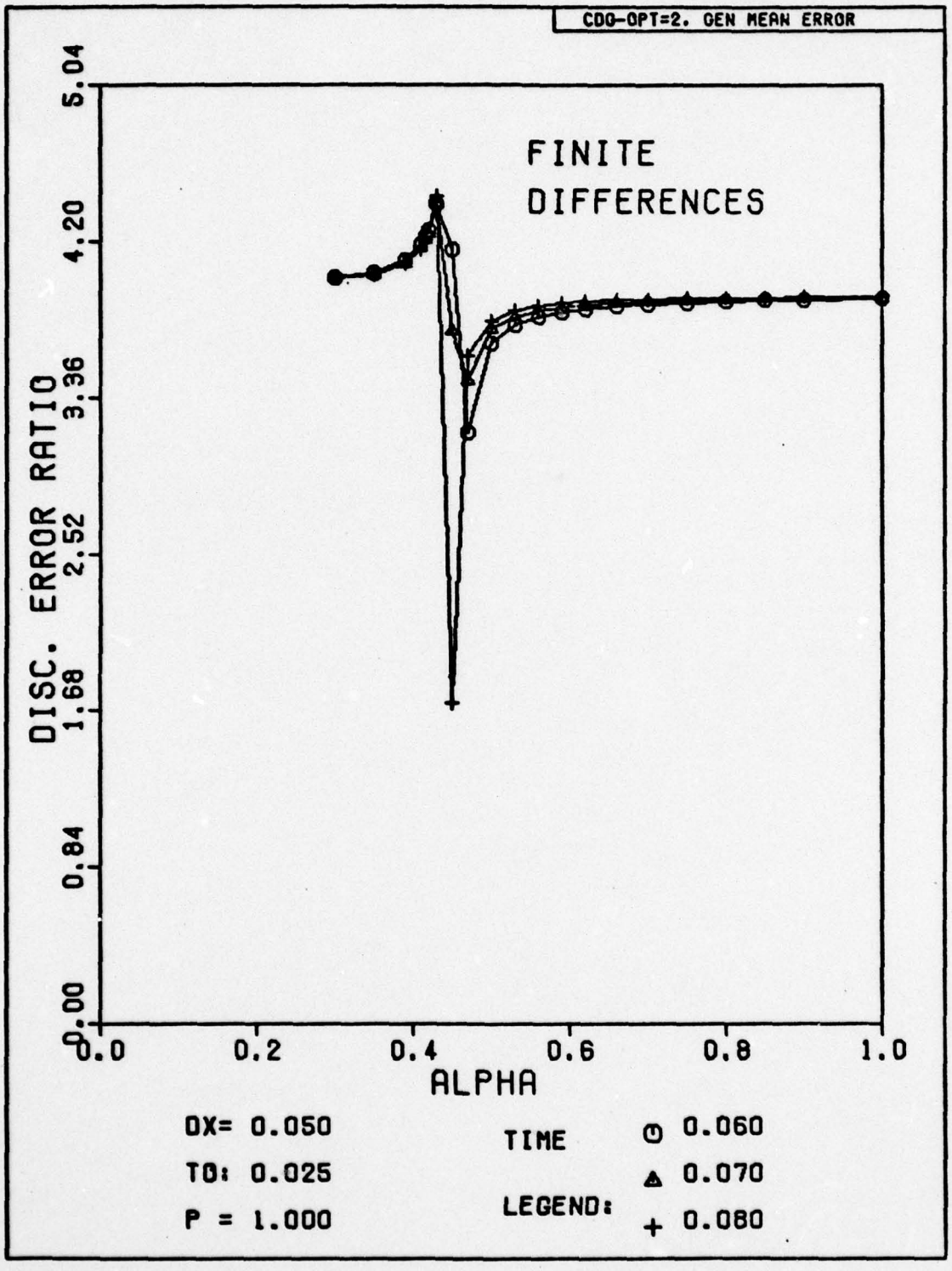


Fig. H-42. Discretization Error Ratio Versus Alpha for Problem One. The space-time grid has been subdivided for the first time step to obtain a more accurate solution.

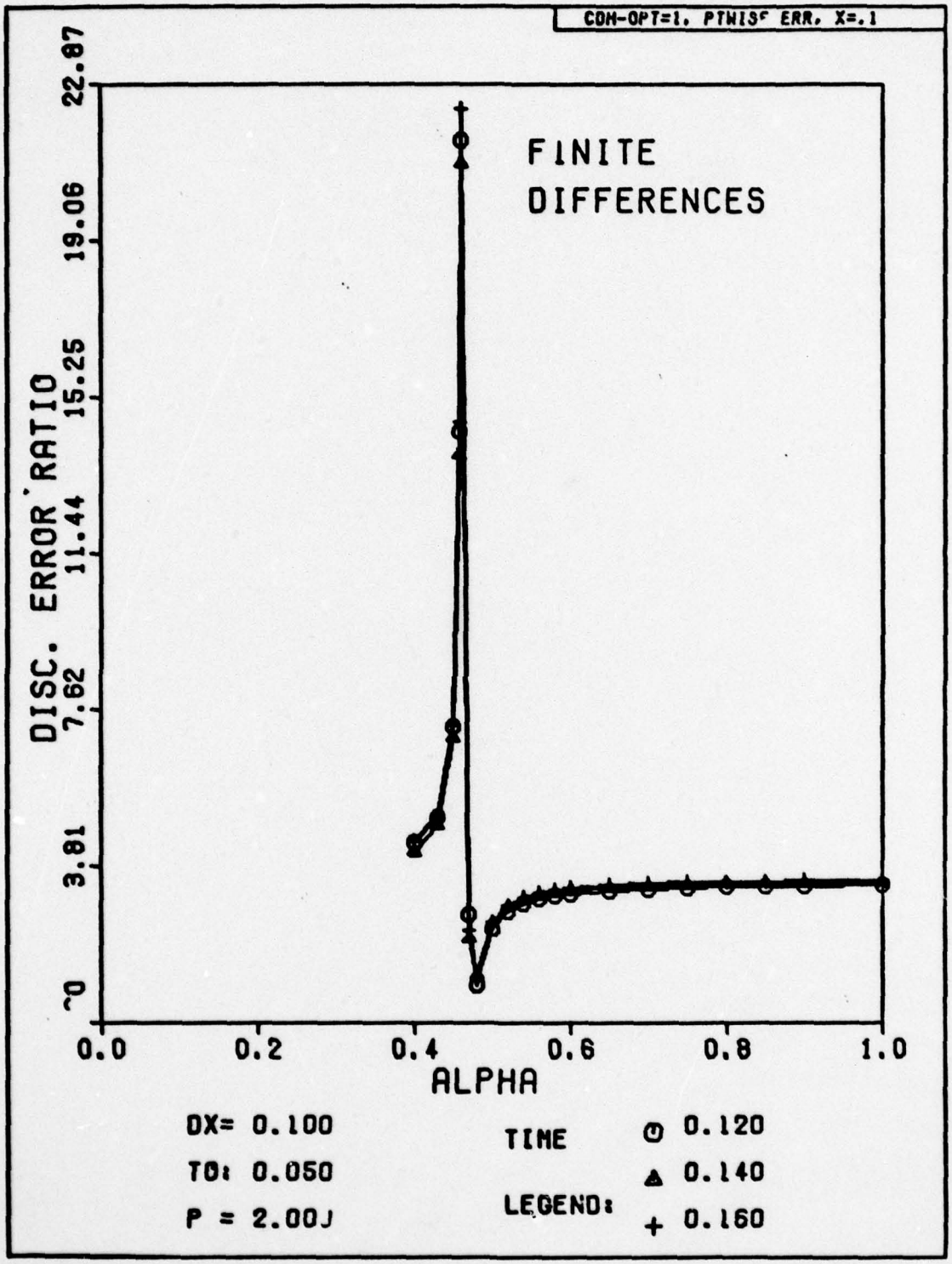


Fig. H-43. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

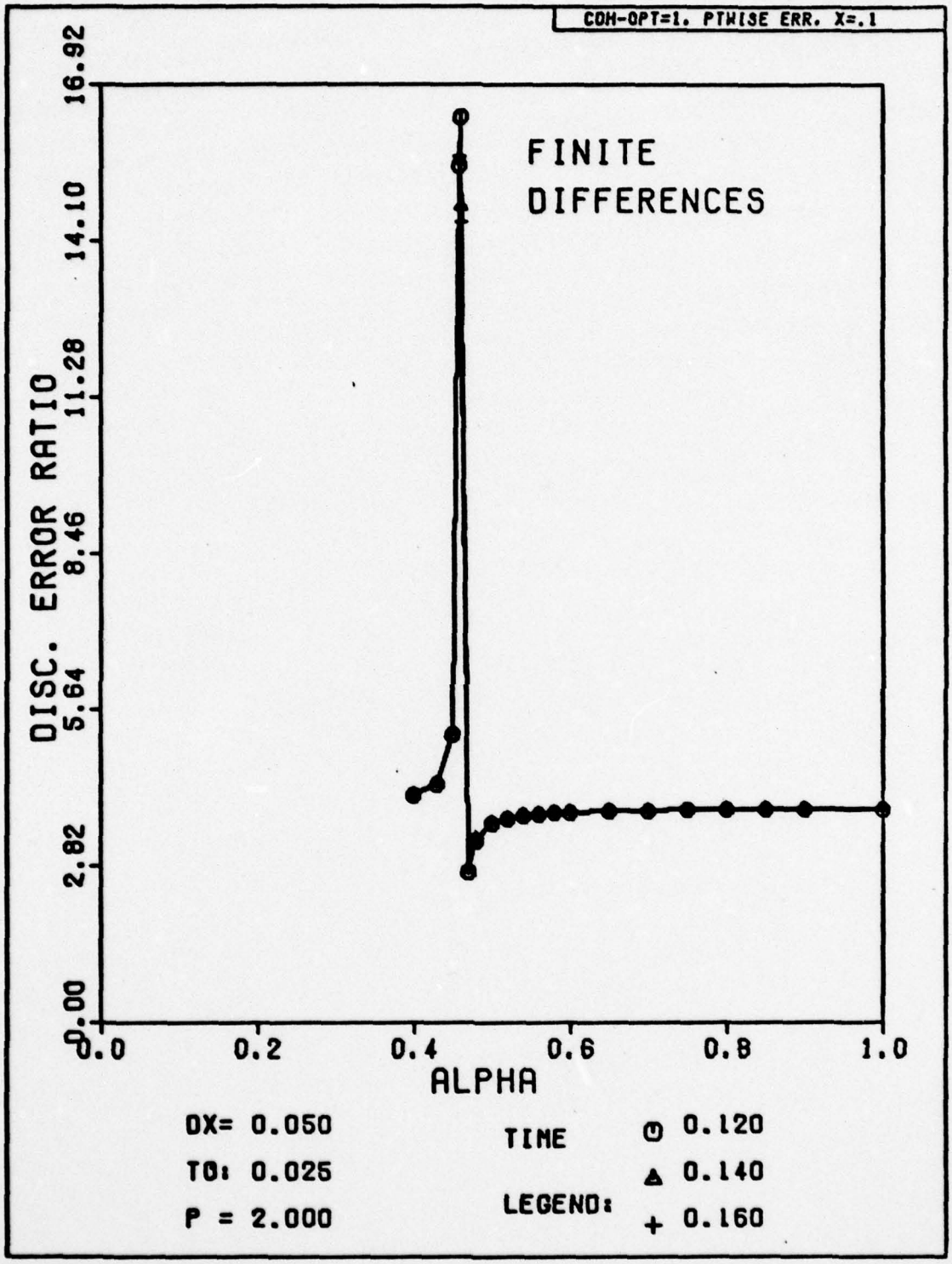


Fig. H-44. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

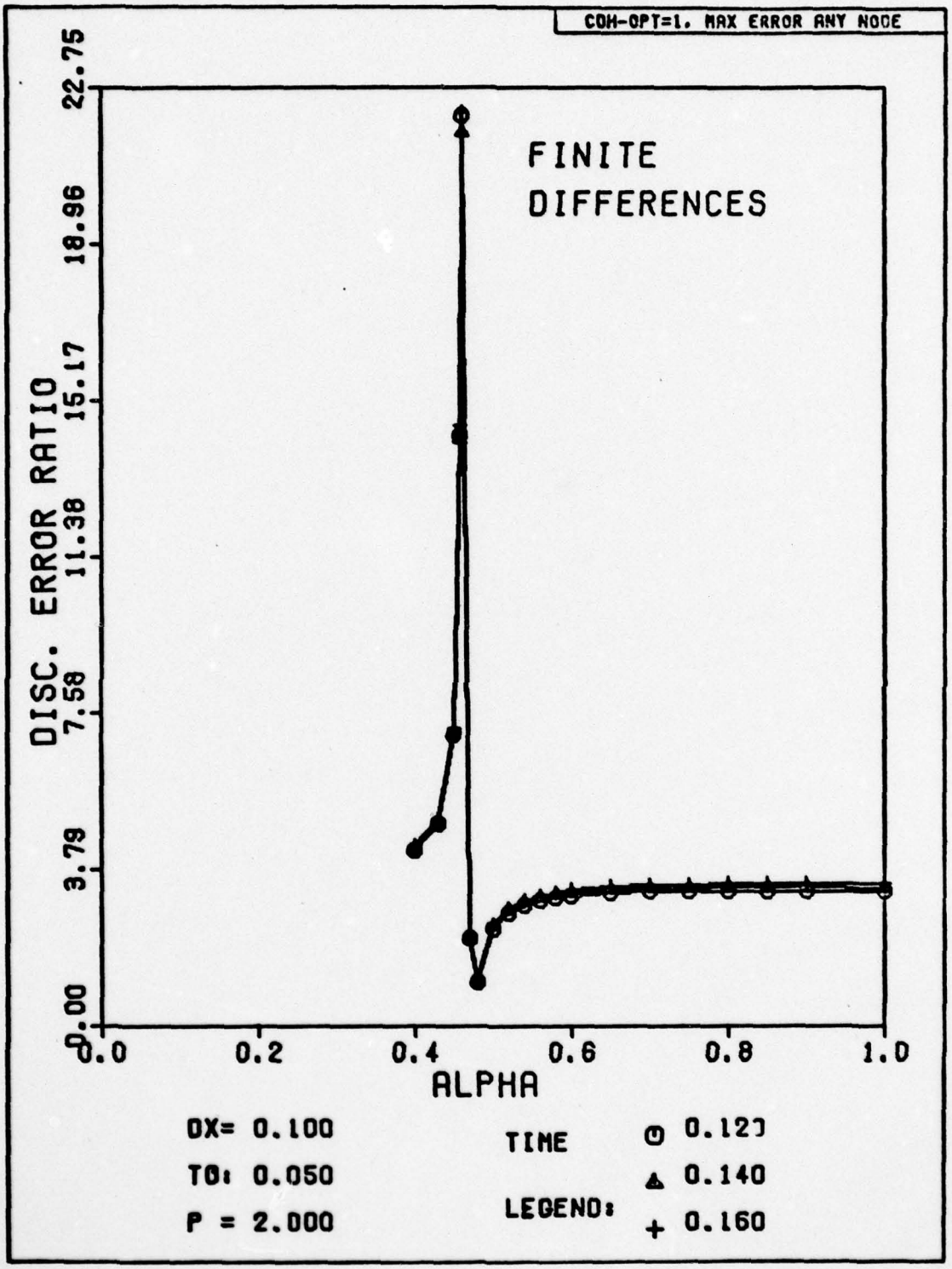


Fig. H-45. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

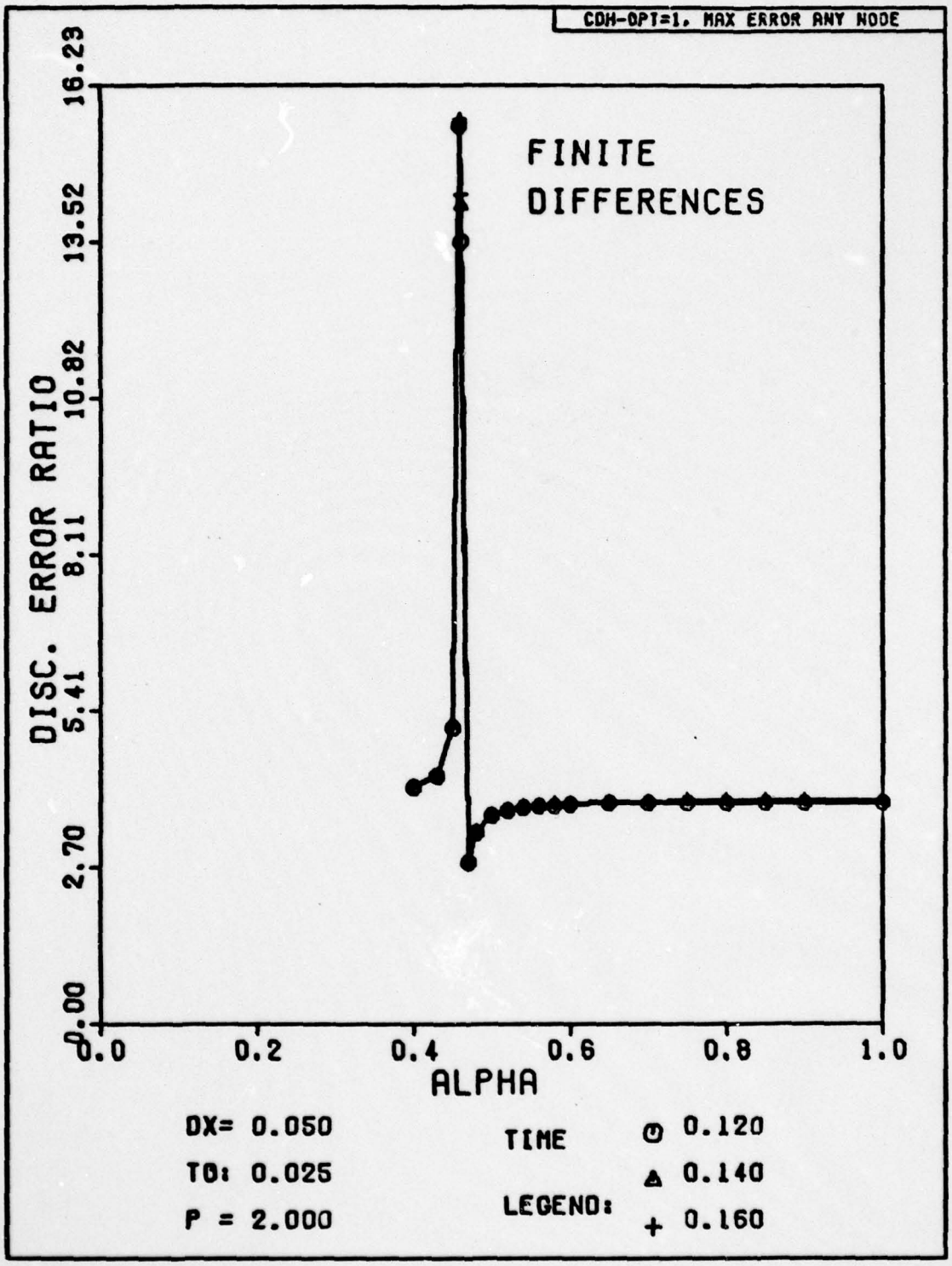
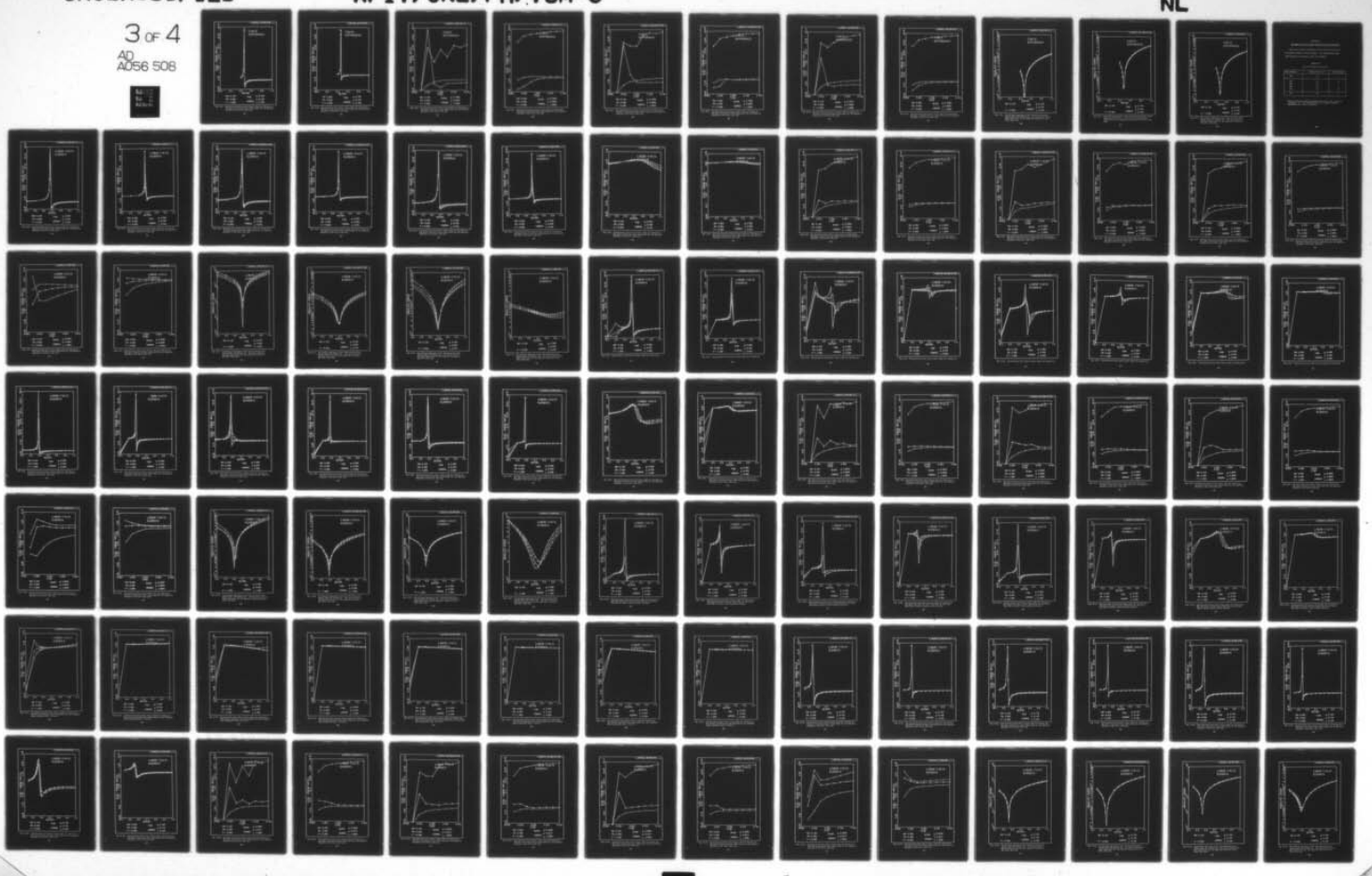


Fig. H-46. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

AD-A056 508 AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OHIO SCH--ETC F/G 20/13
AN INVESTIGATION OF THE NUMERICAL METHODS OF FINITE DIFFERENCES--ETC(U)
MAR 78 C R MARTIN

UNCLASSIFIED AFIT/GNE/PH/78M-6 NL

3 OF 4
AD A056 508



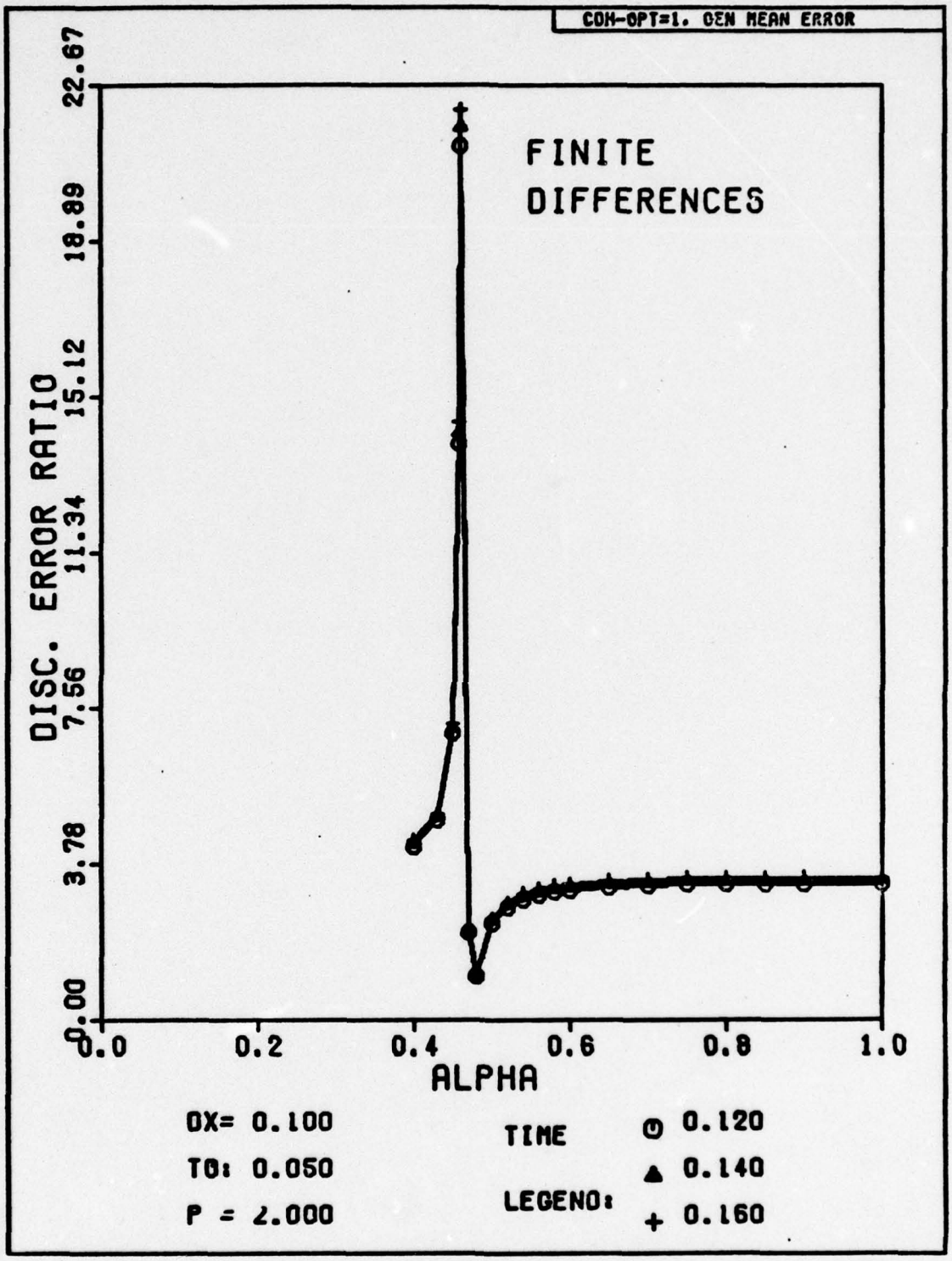


Fig. H-47. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

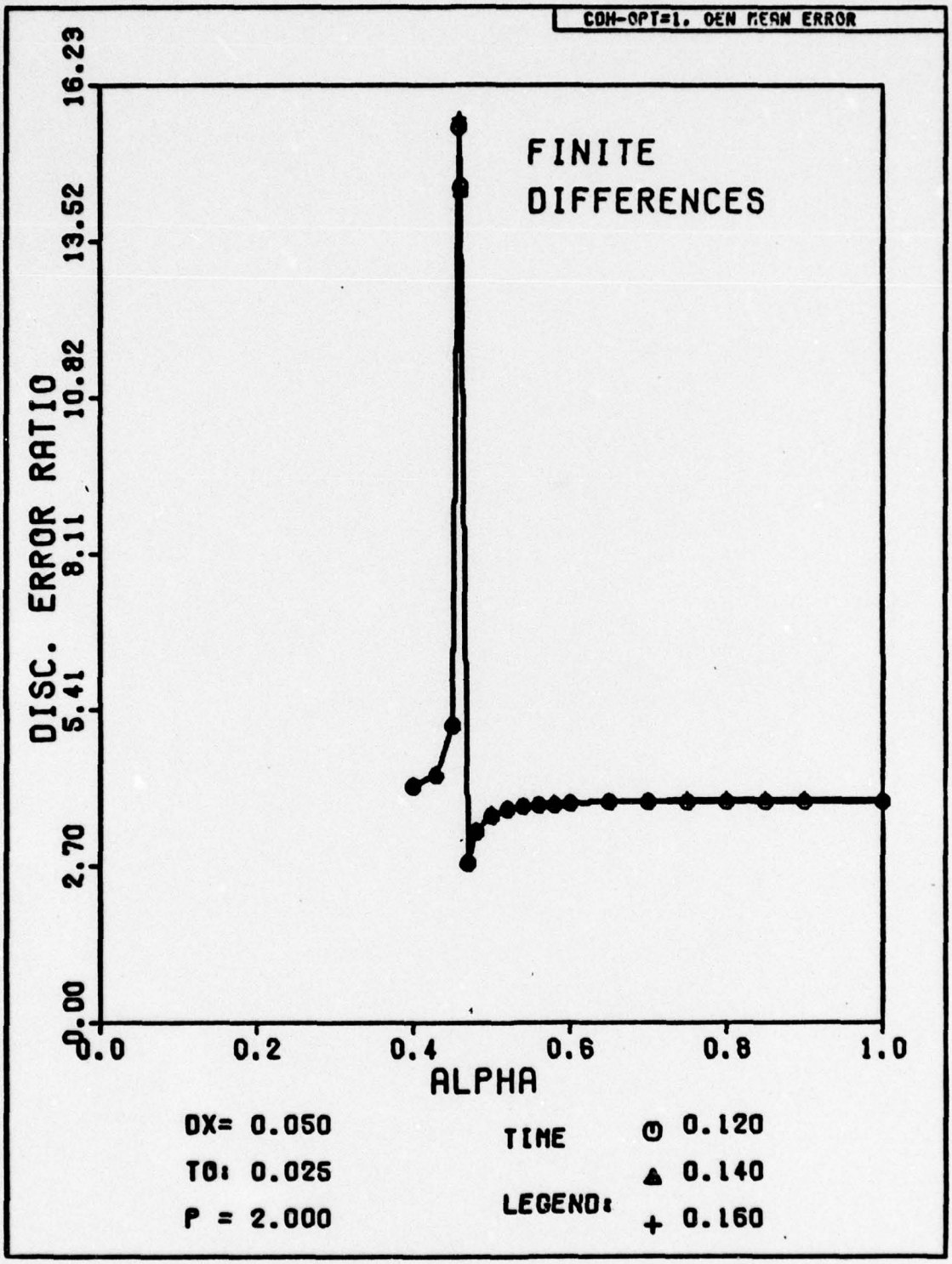


Fig. H-48. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

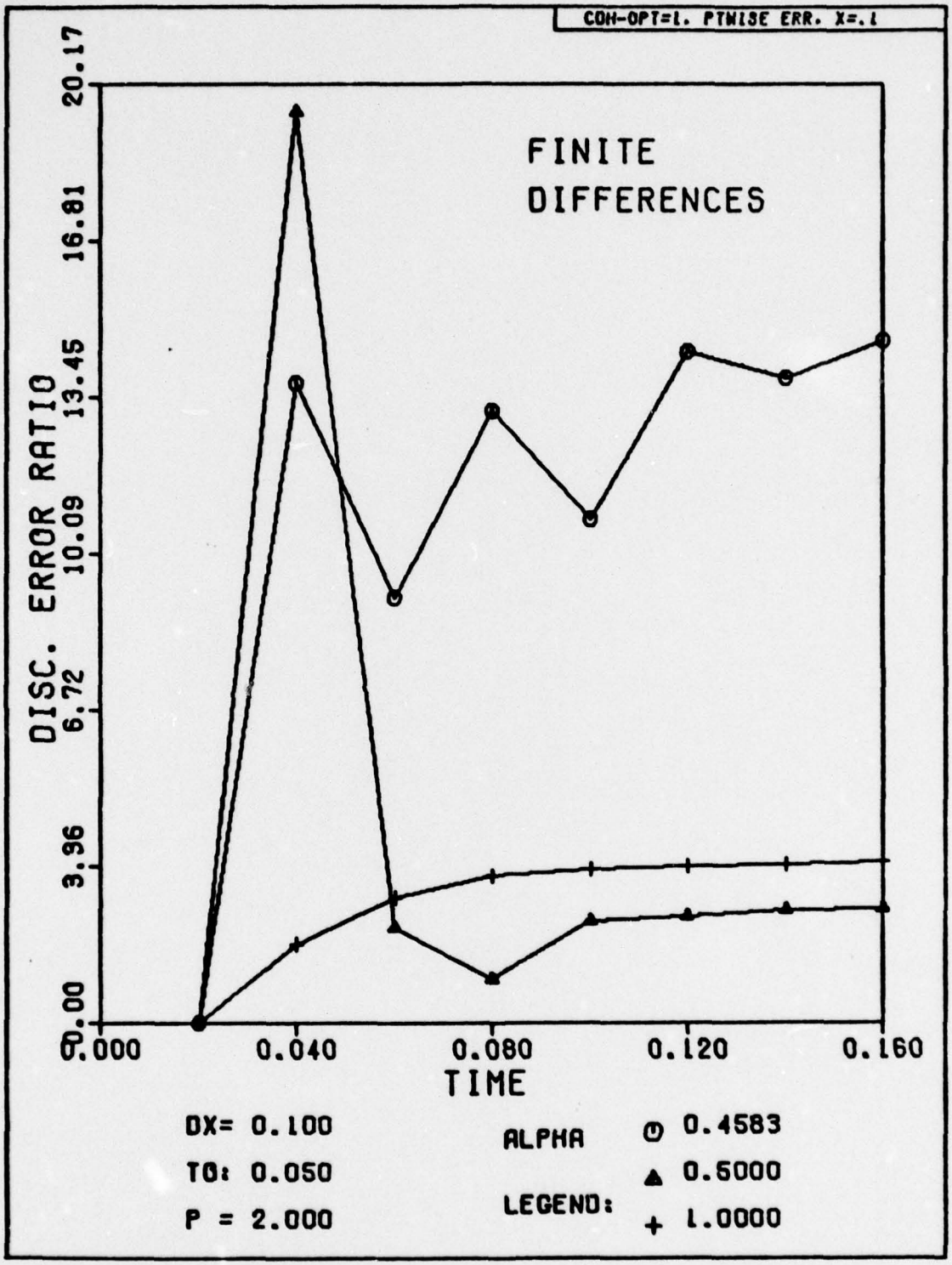


Fig. H-49. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

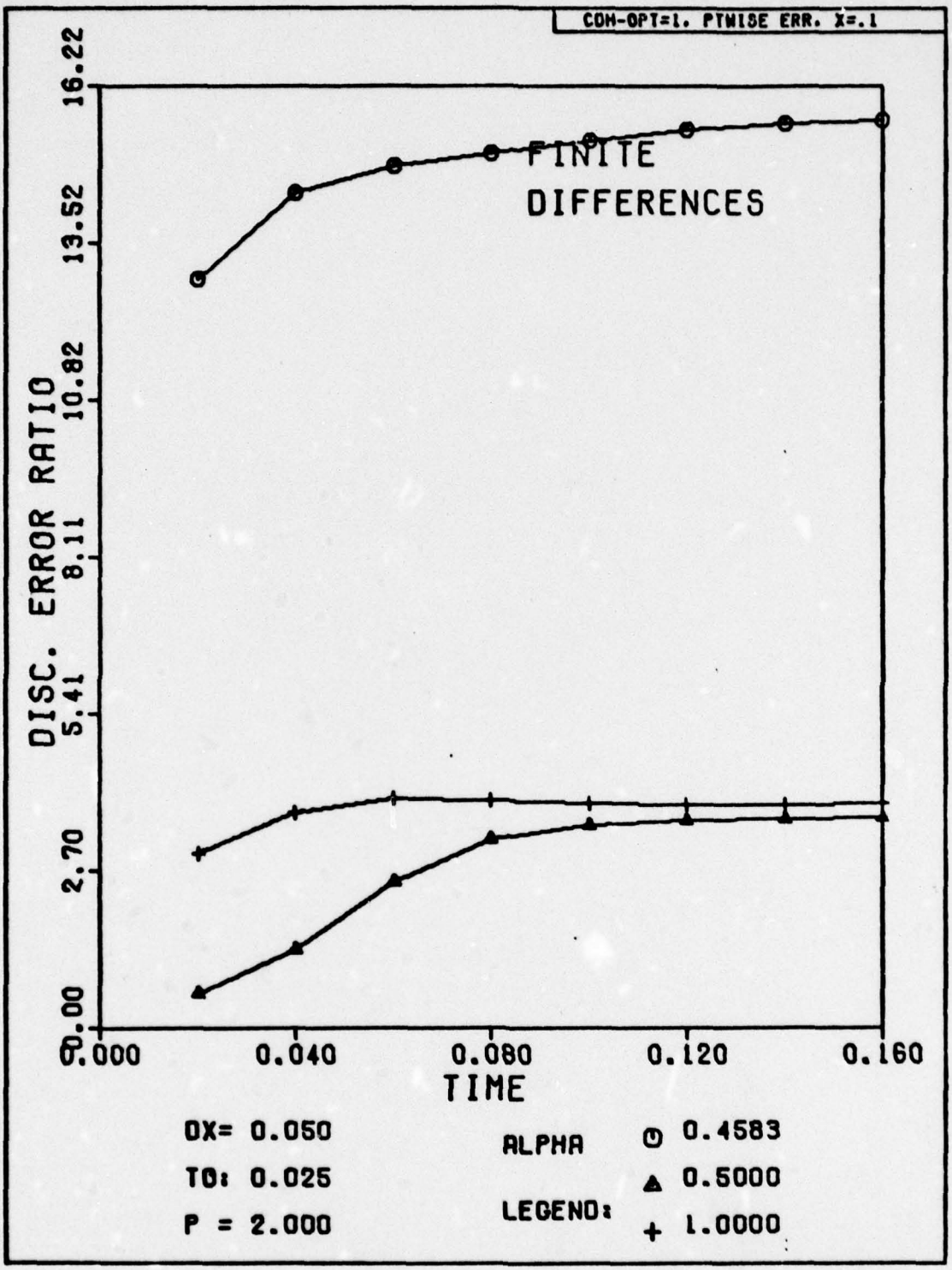


Fig. H-50. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

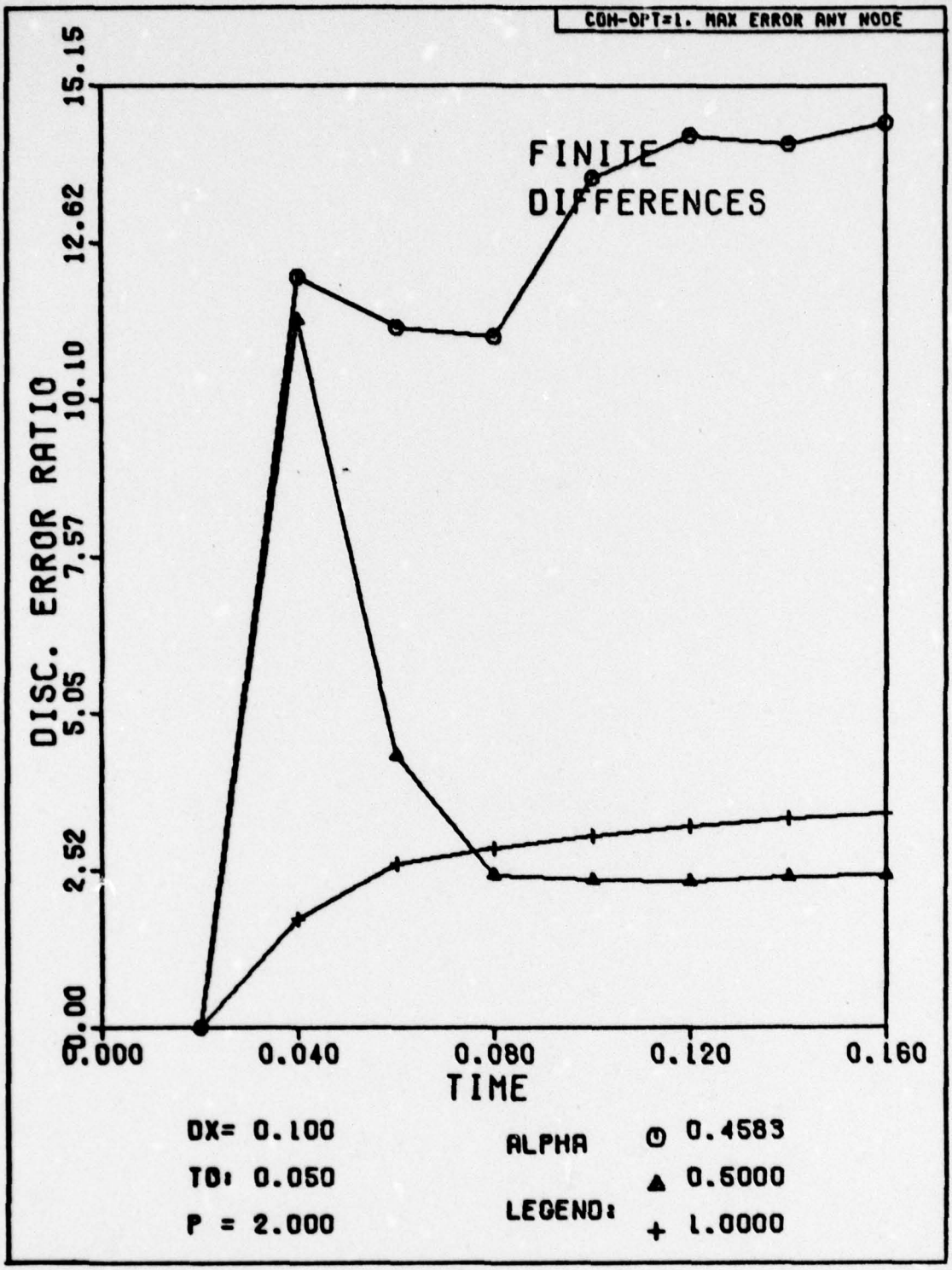


Fig. H-51. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

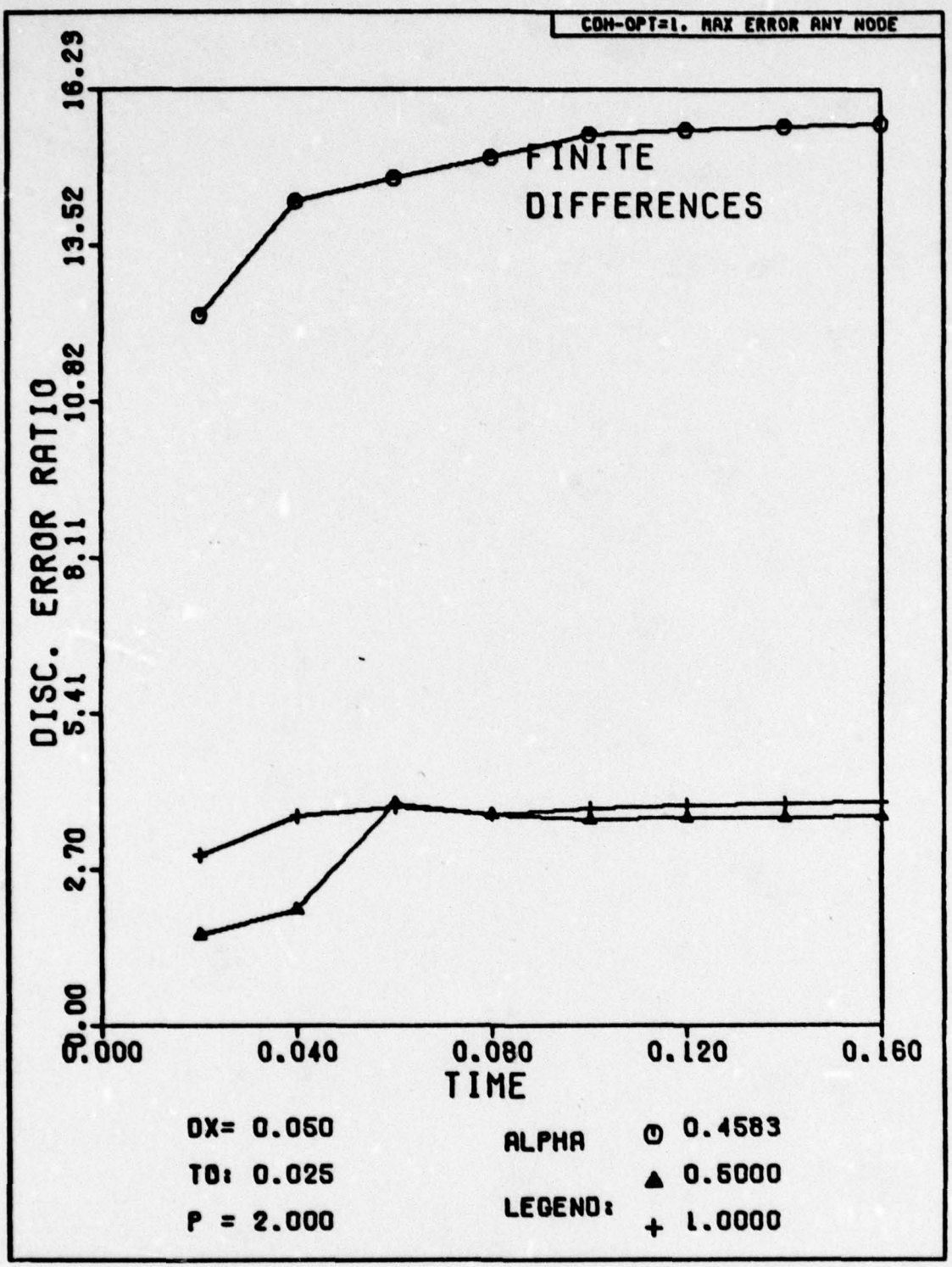


Fig. H-52. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

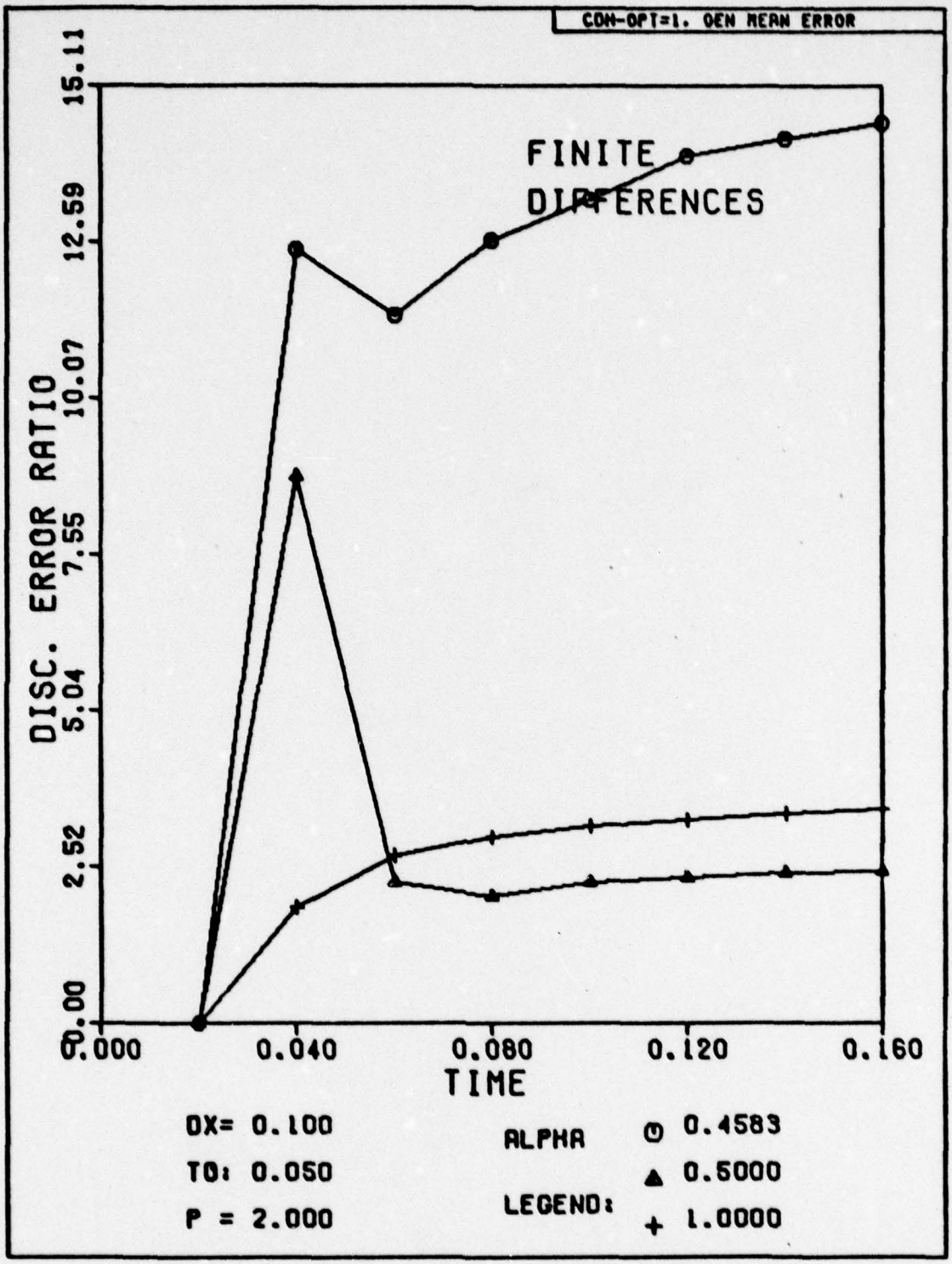


Fig. H-53. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

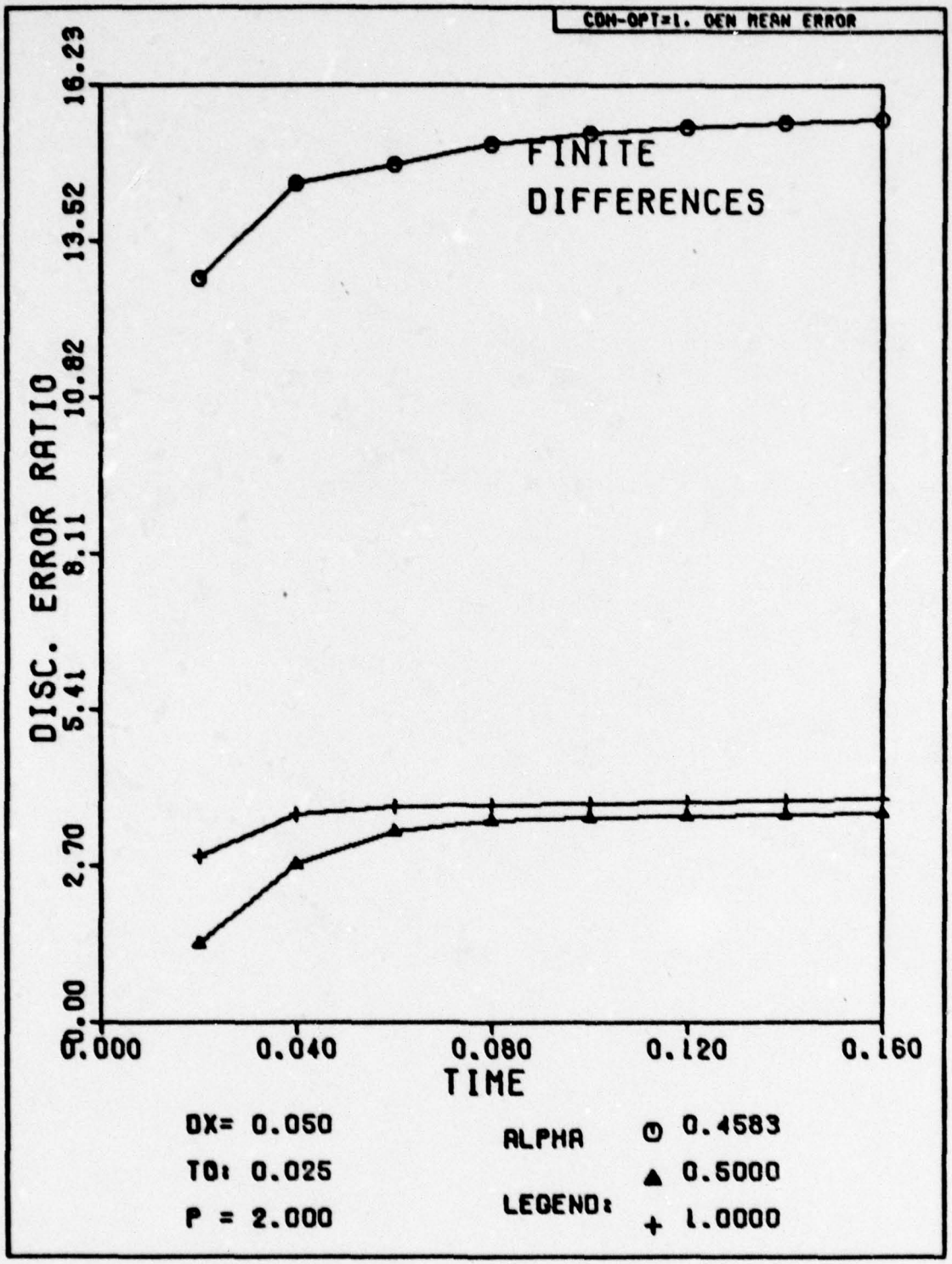


Fig. H-54. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

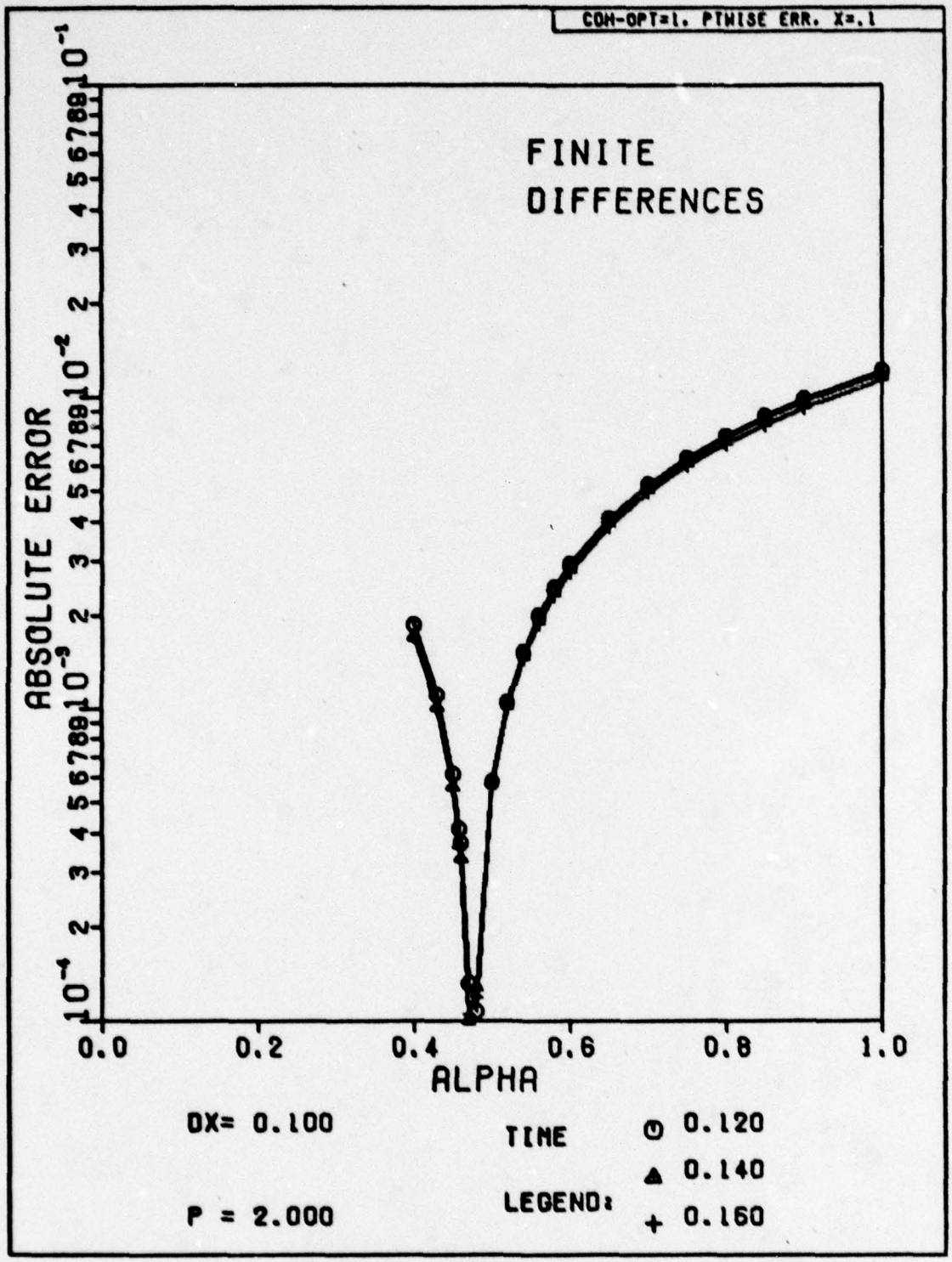


Fig. H-55. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

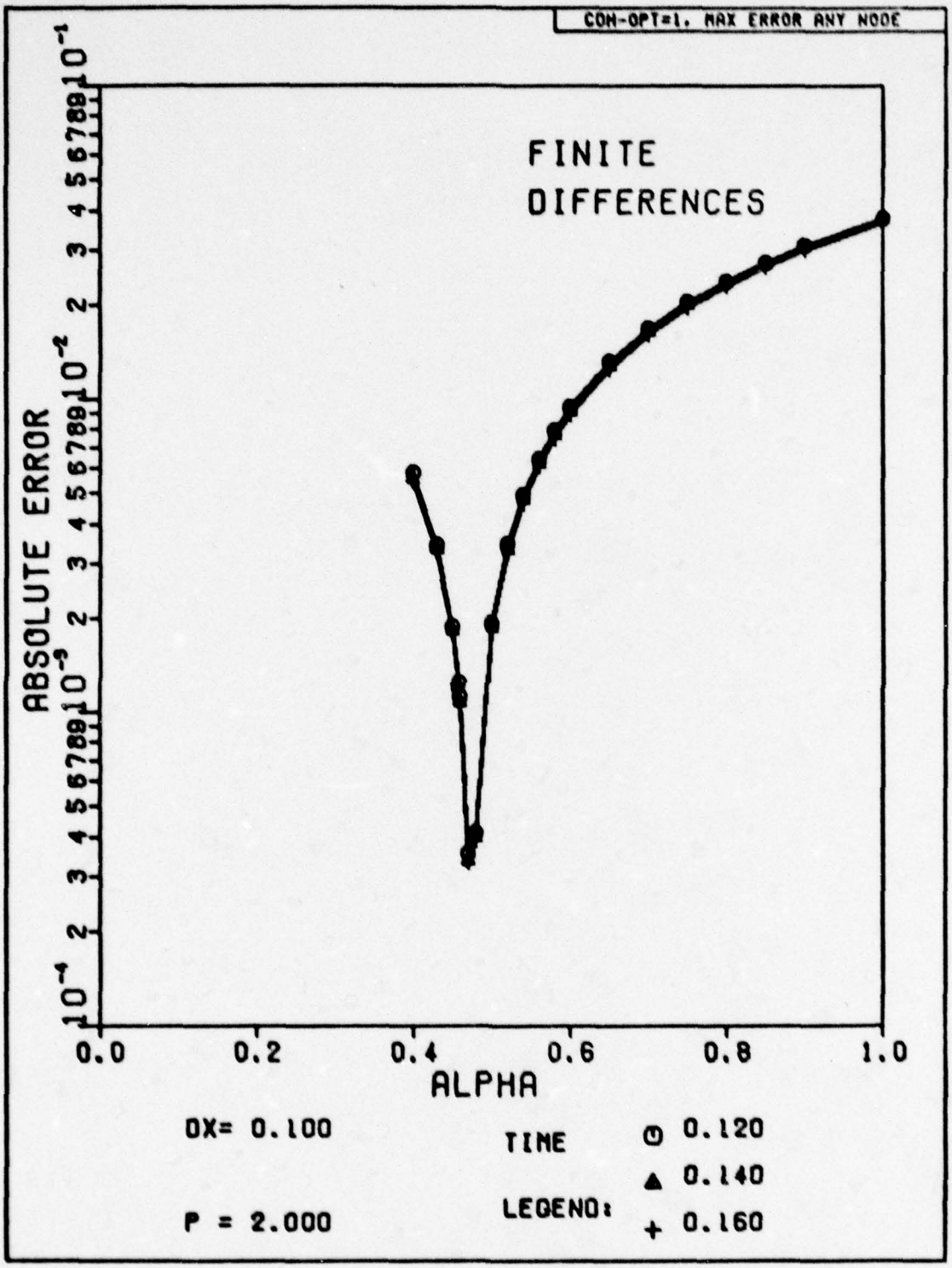


Fig. H-56. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

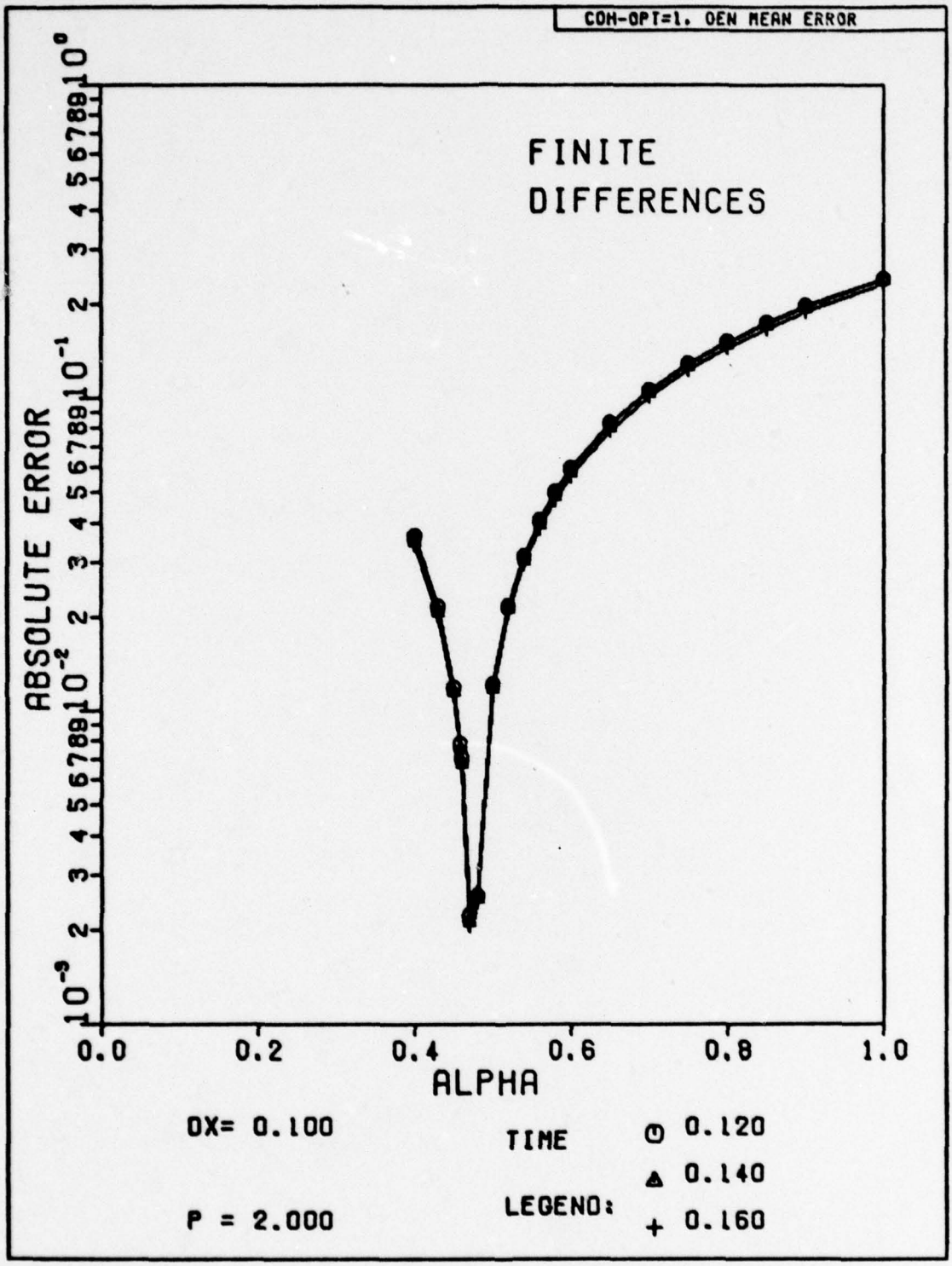


Fig. H-57. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

Section II

The Results for the Primary Problem Using Finite-Elements

This section shows the graphical results for the solution of the primary problem by finite-elements. The following key shows which options are included in this set of graphs.

Table H-II

Key to the Plots in Section II

Run Identifier	Fourier Modulus (p)	Option Number
CER	0.5	1
CES	1.0	0
CET	1.0	1
CEU	1.0	2
CEV	1.0	3*
CEY	2.0	1

* This run was made using the modification used in Figs. 13 and 14 in Chapter III. This modification was an attempt to better approximate the initial conditions.

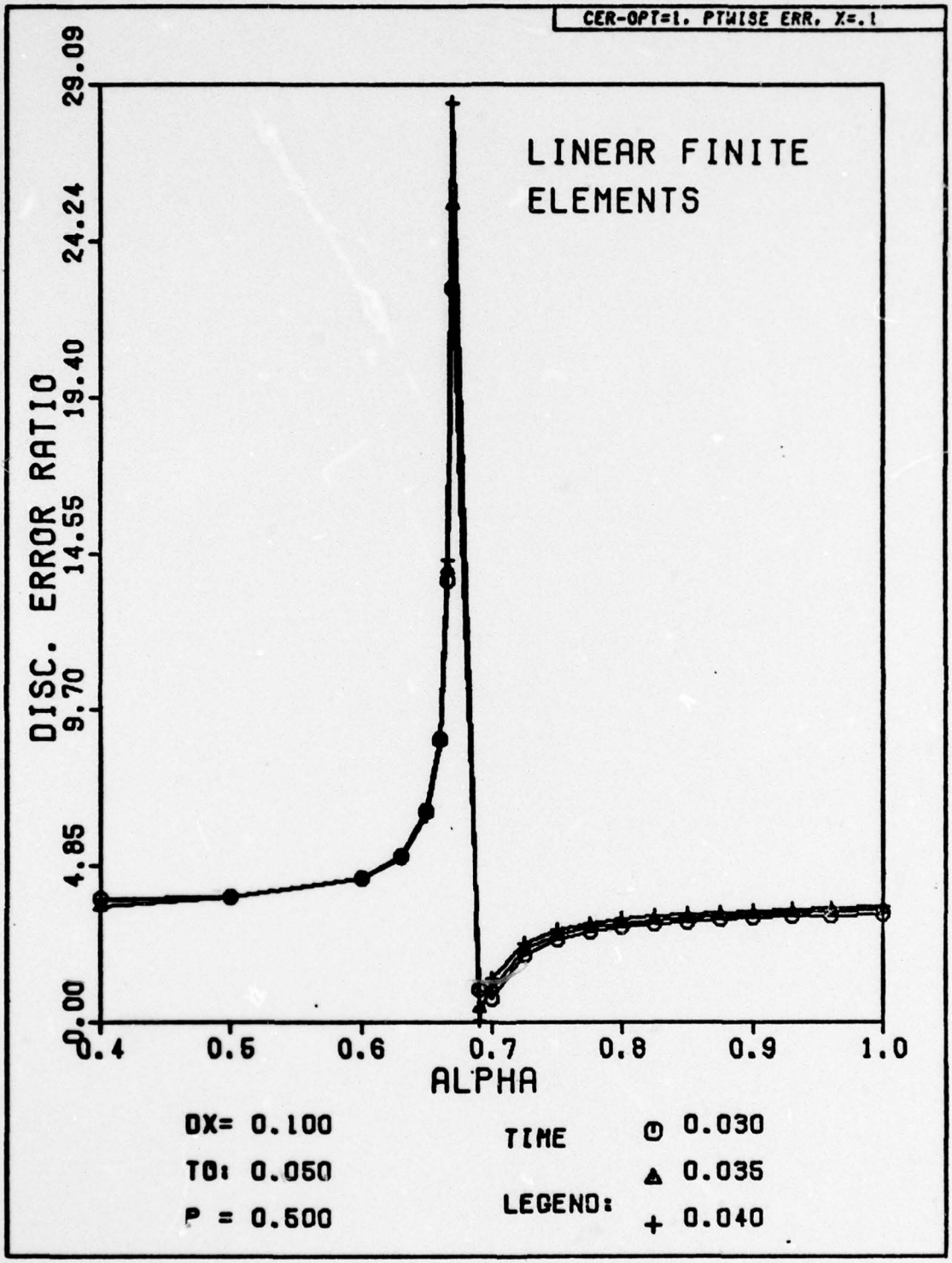


Fig. H-58. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

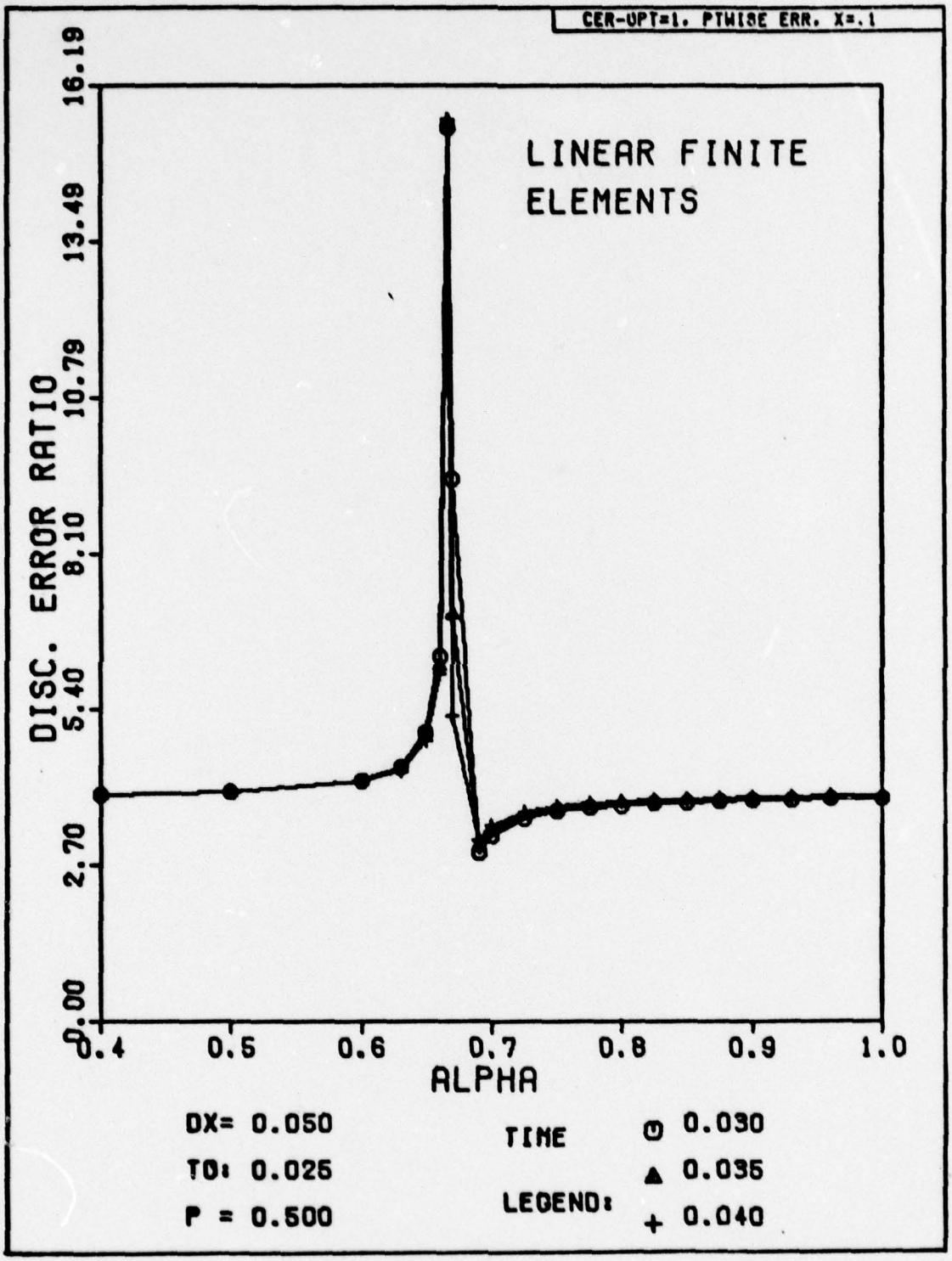


Fig. H-59. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

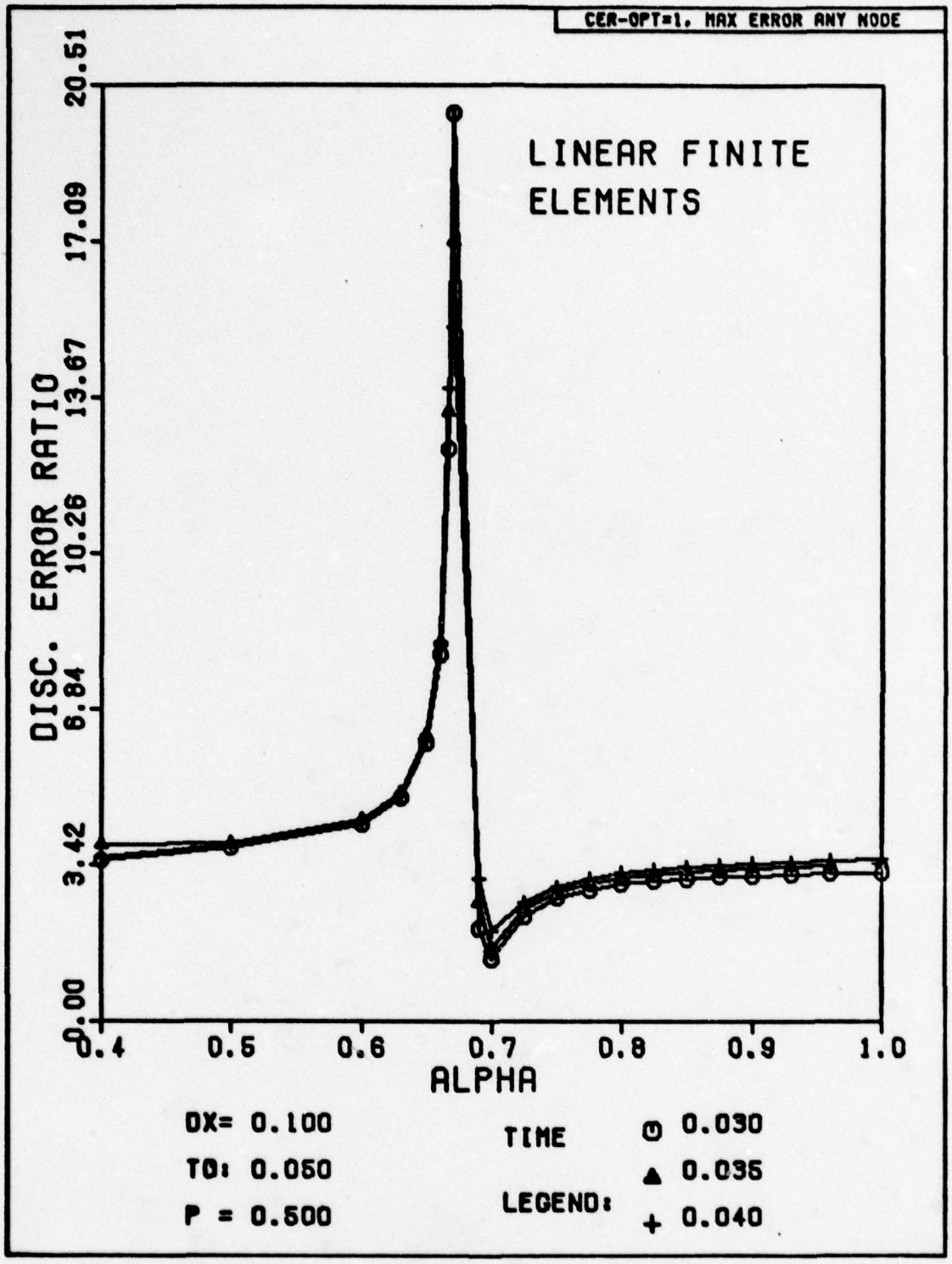


Fig. H-60. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

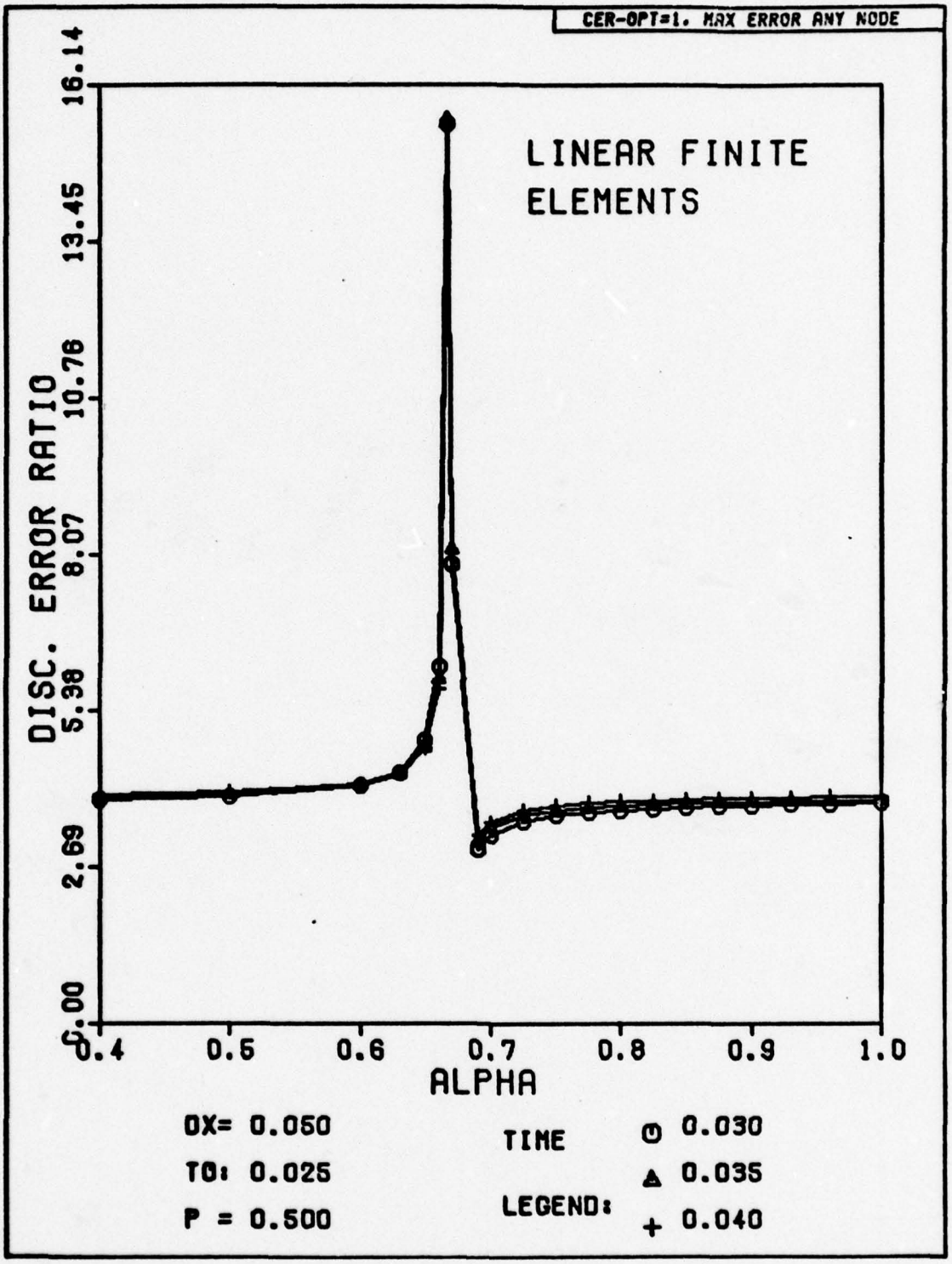


Fig. H-61. Discretization Error Ration Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

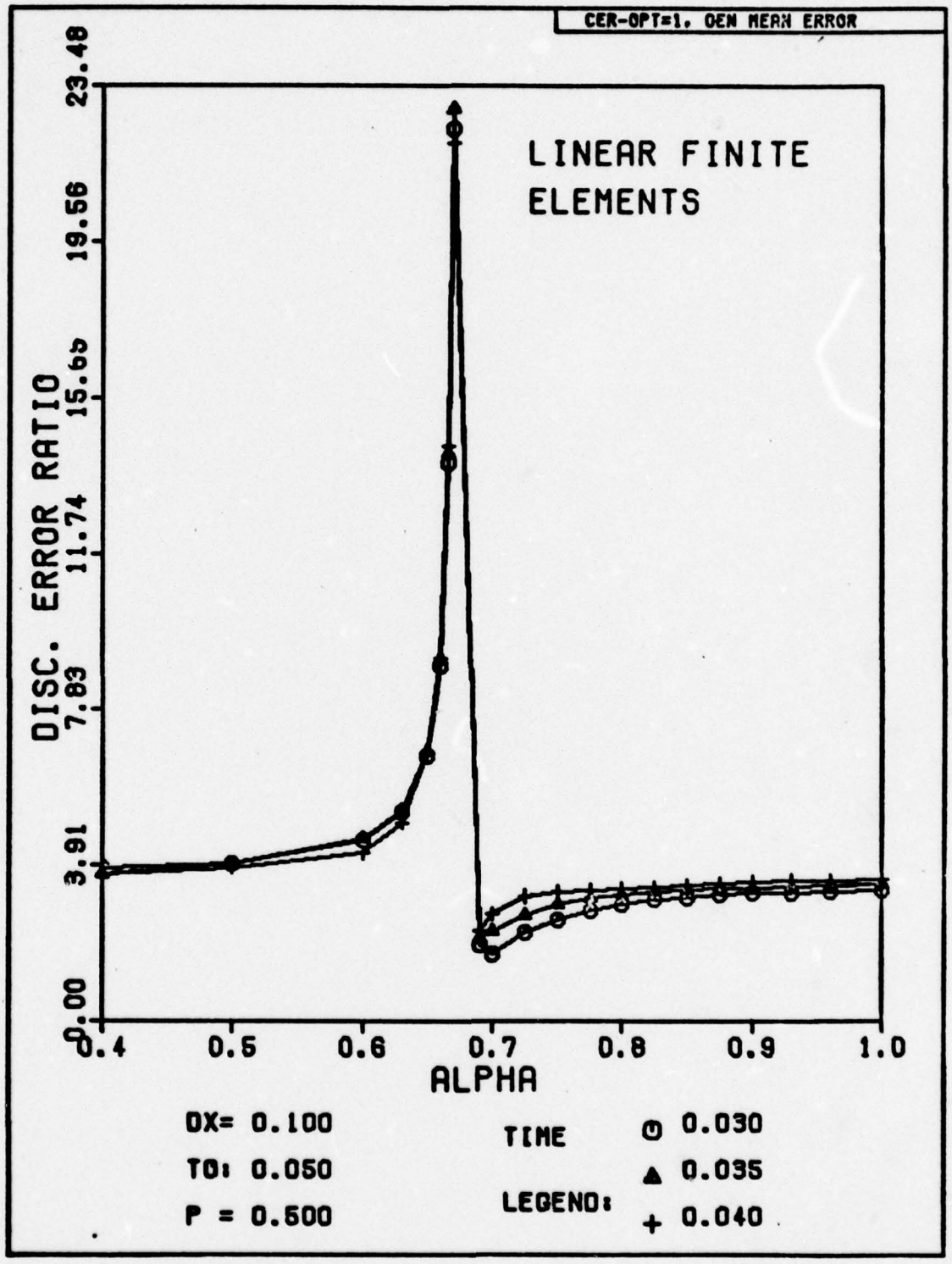


Fig. H-62. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

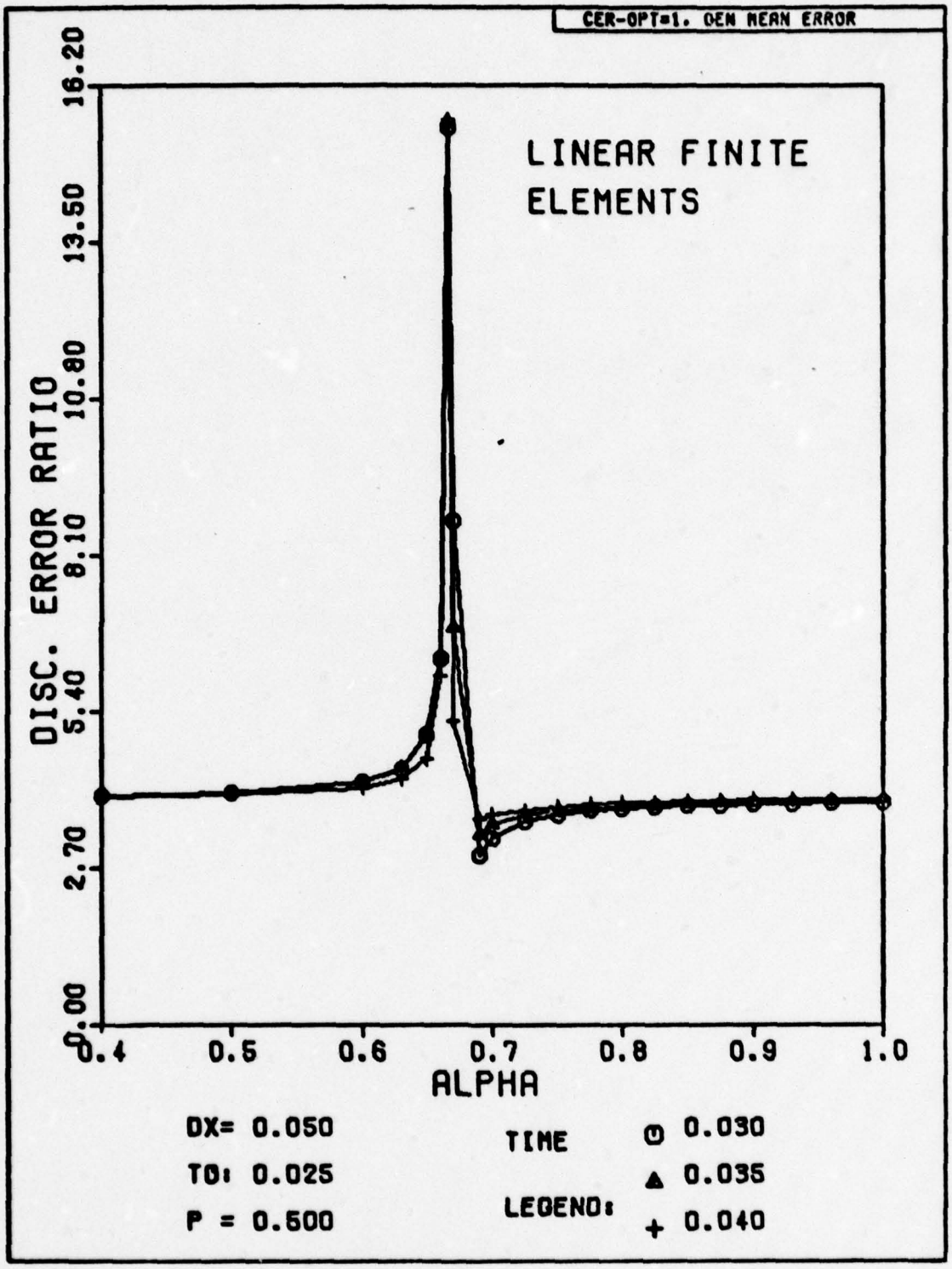


Fig. H-63. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

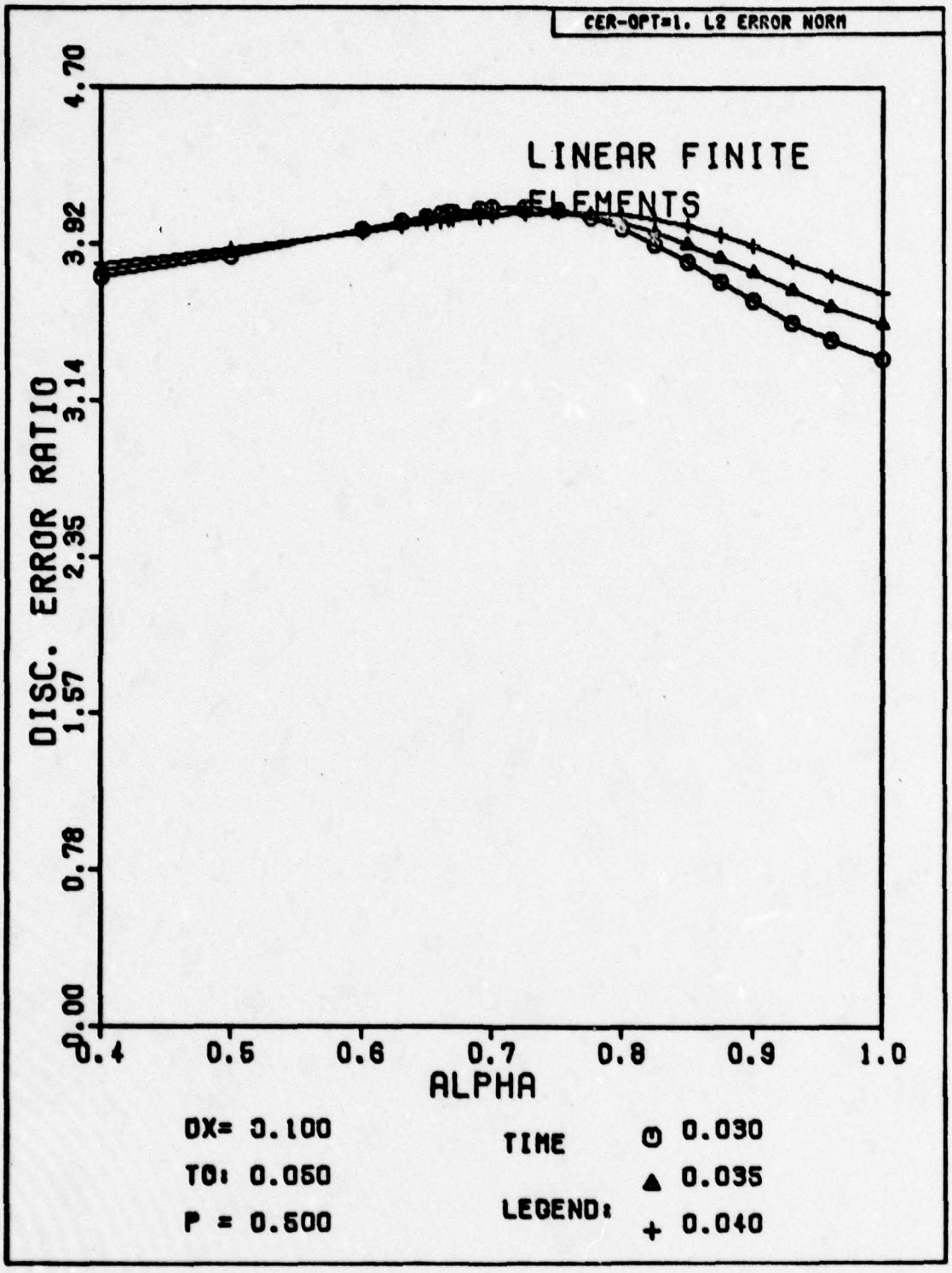


Fig. H-64. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

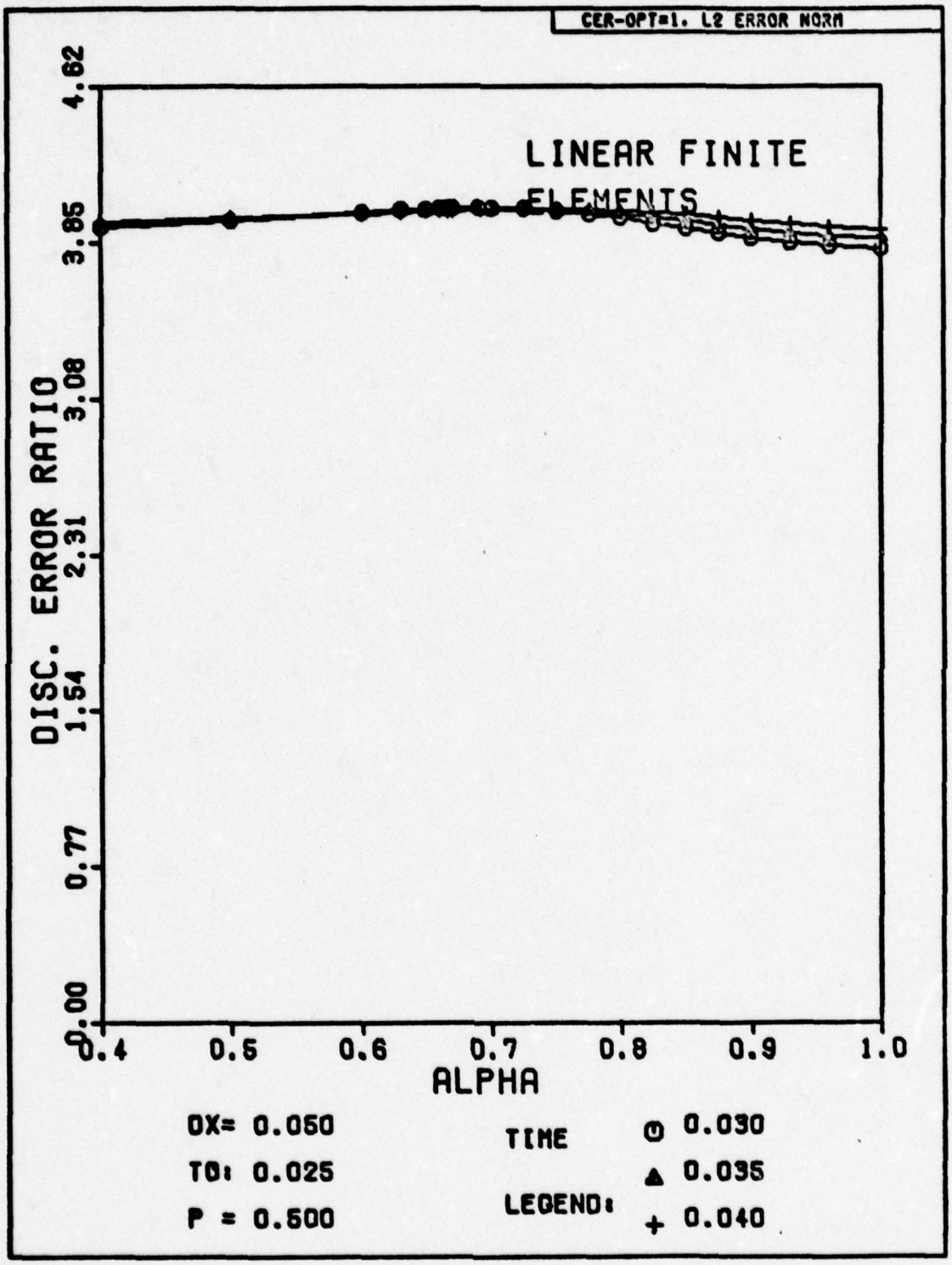


Fig. H-65. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

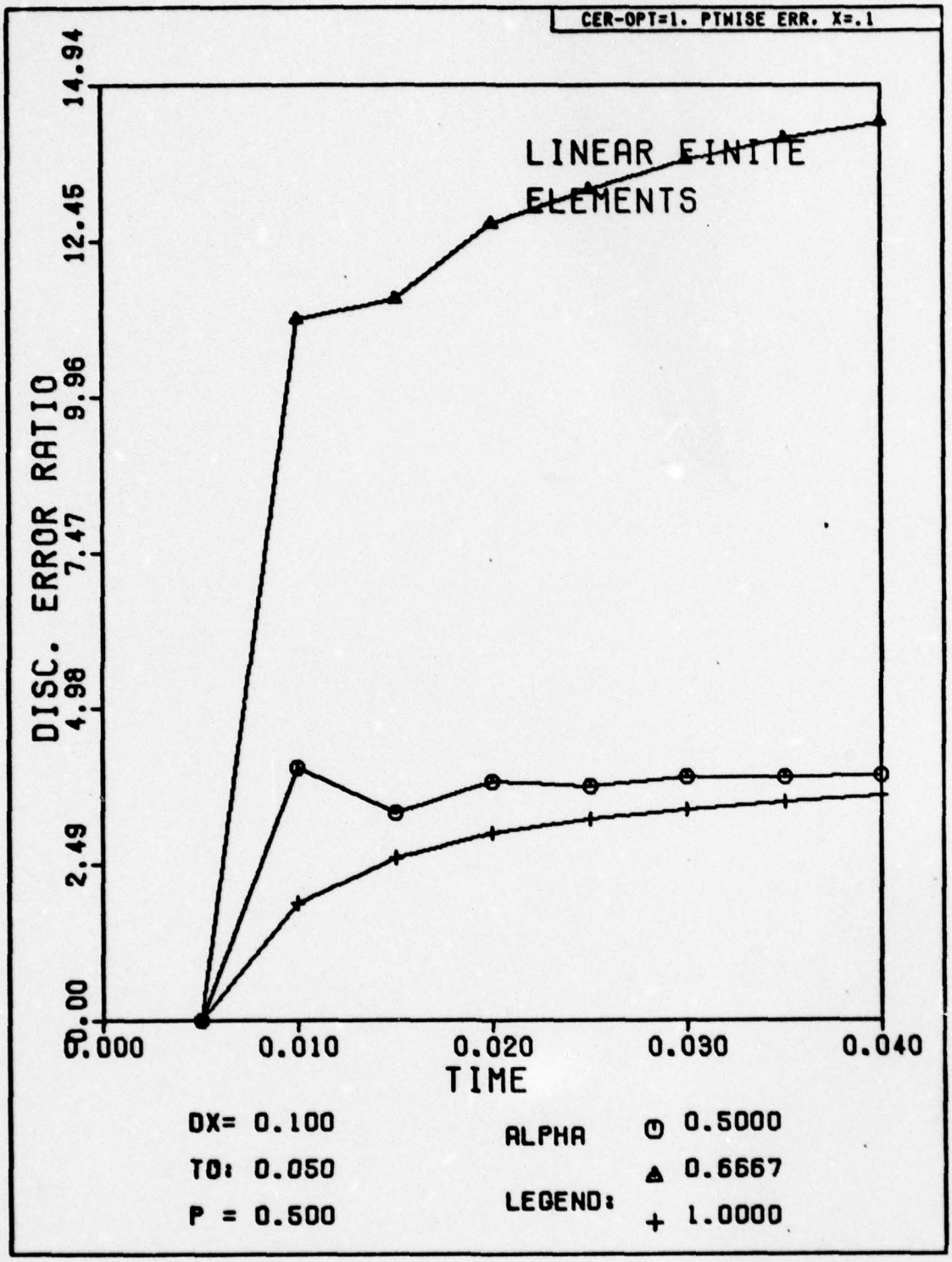


Fig. H-66. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

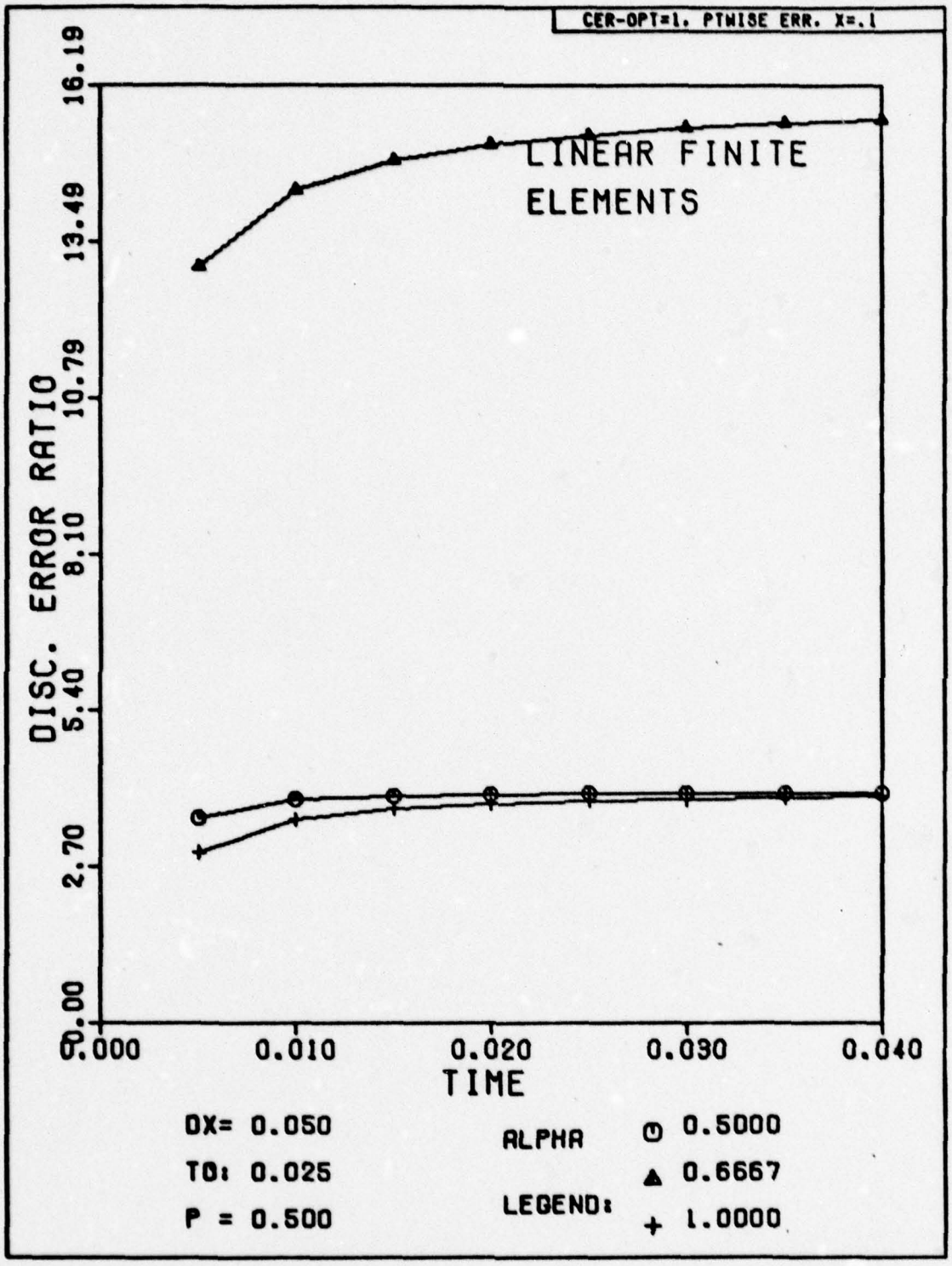


Fig. H-67. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

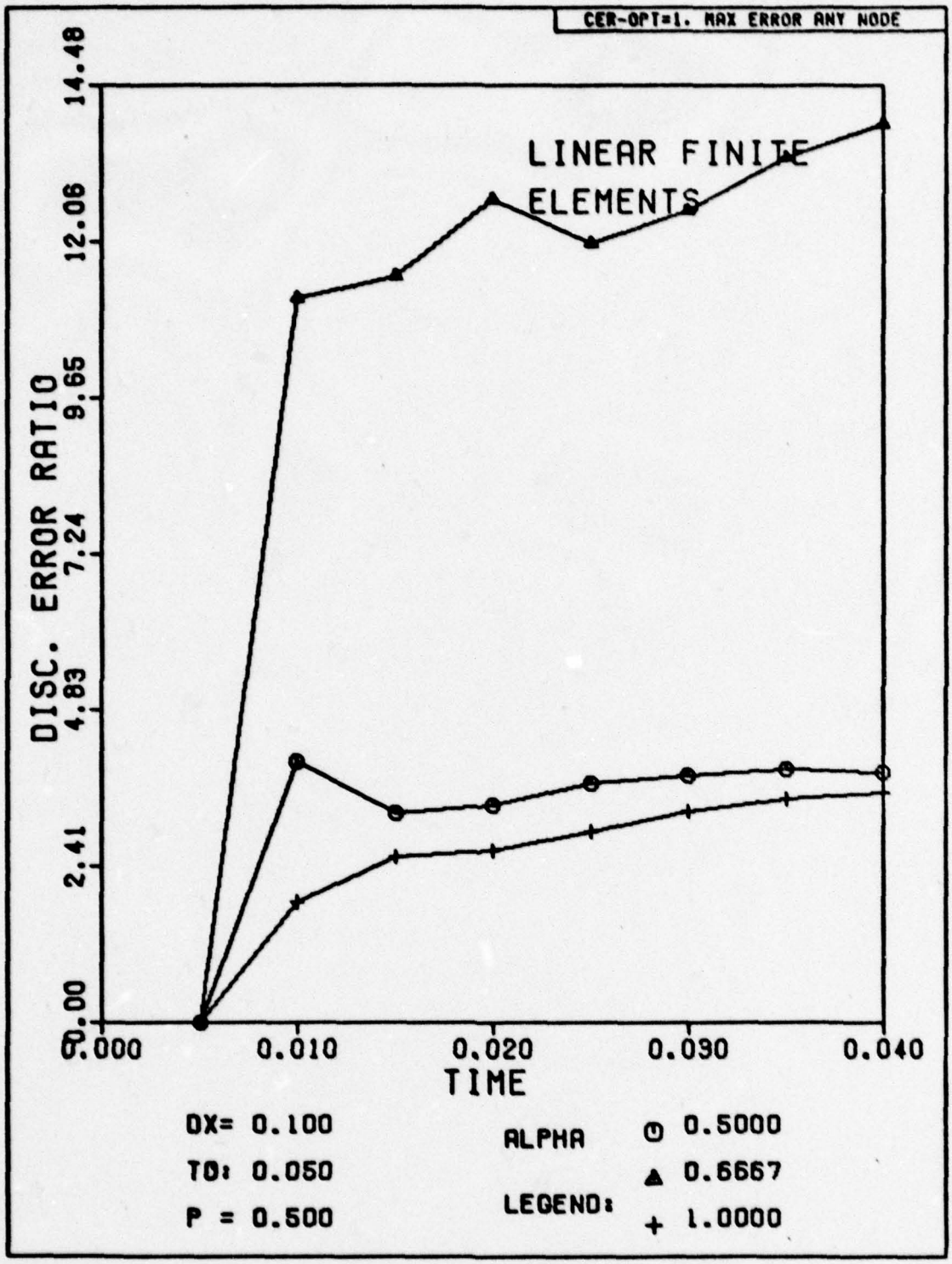


Fig. H-68. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

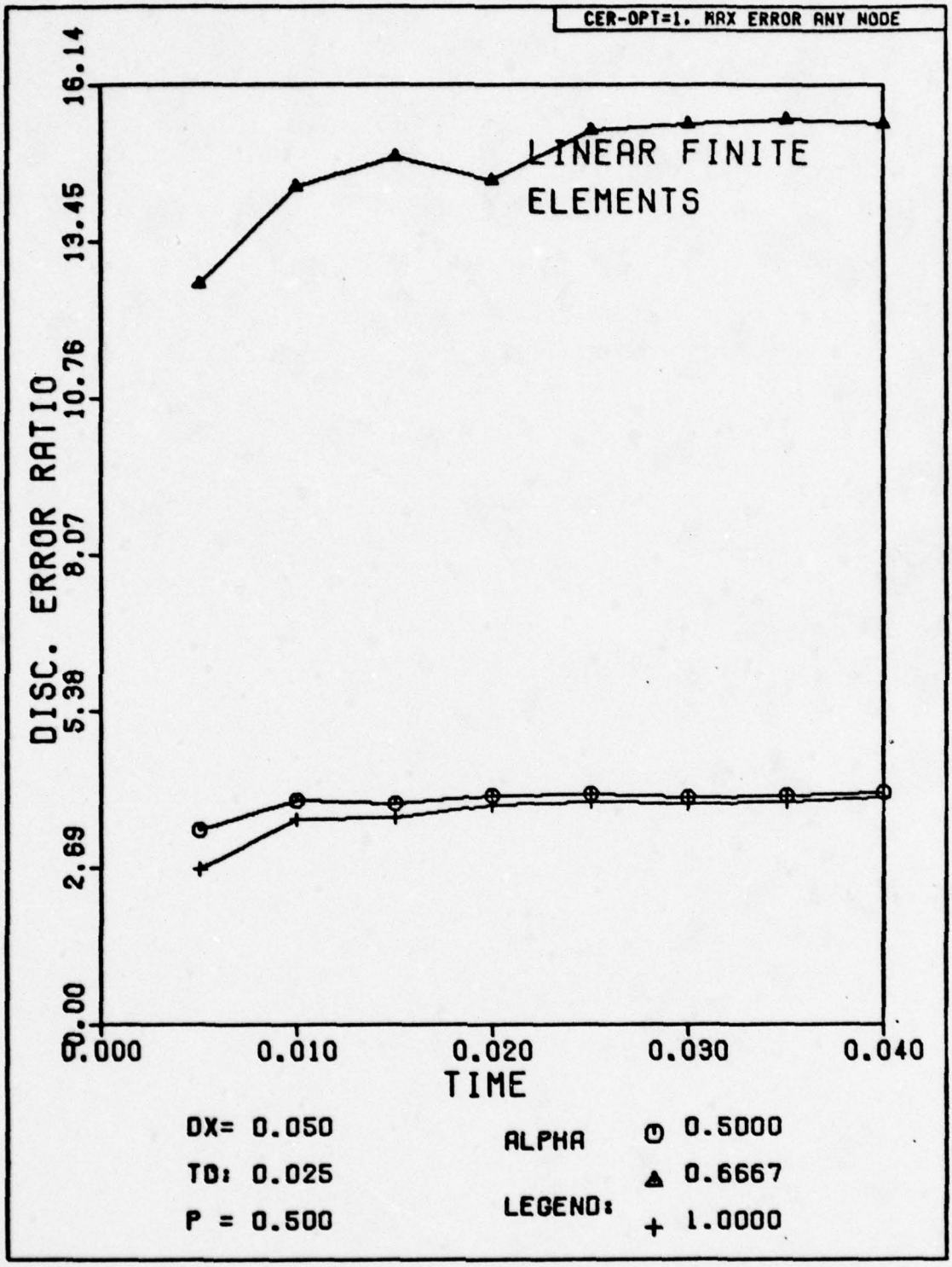


Fig. H-69. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

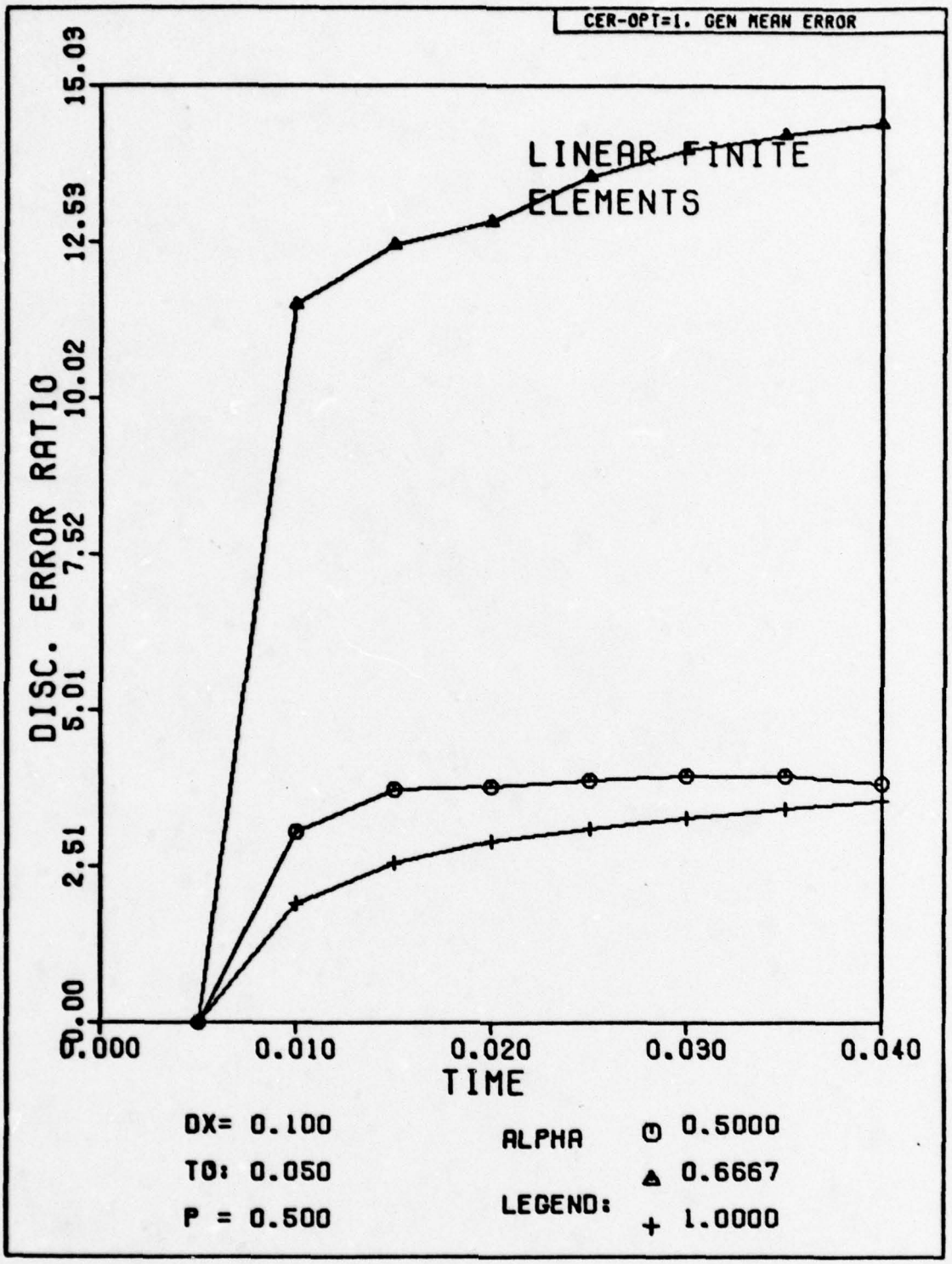


Fig. H-70. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

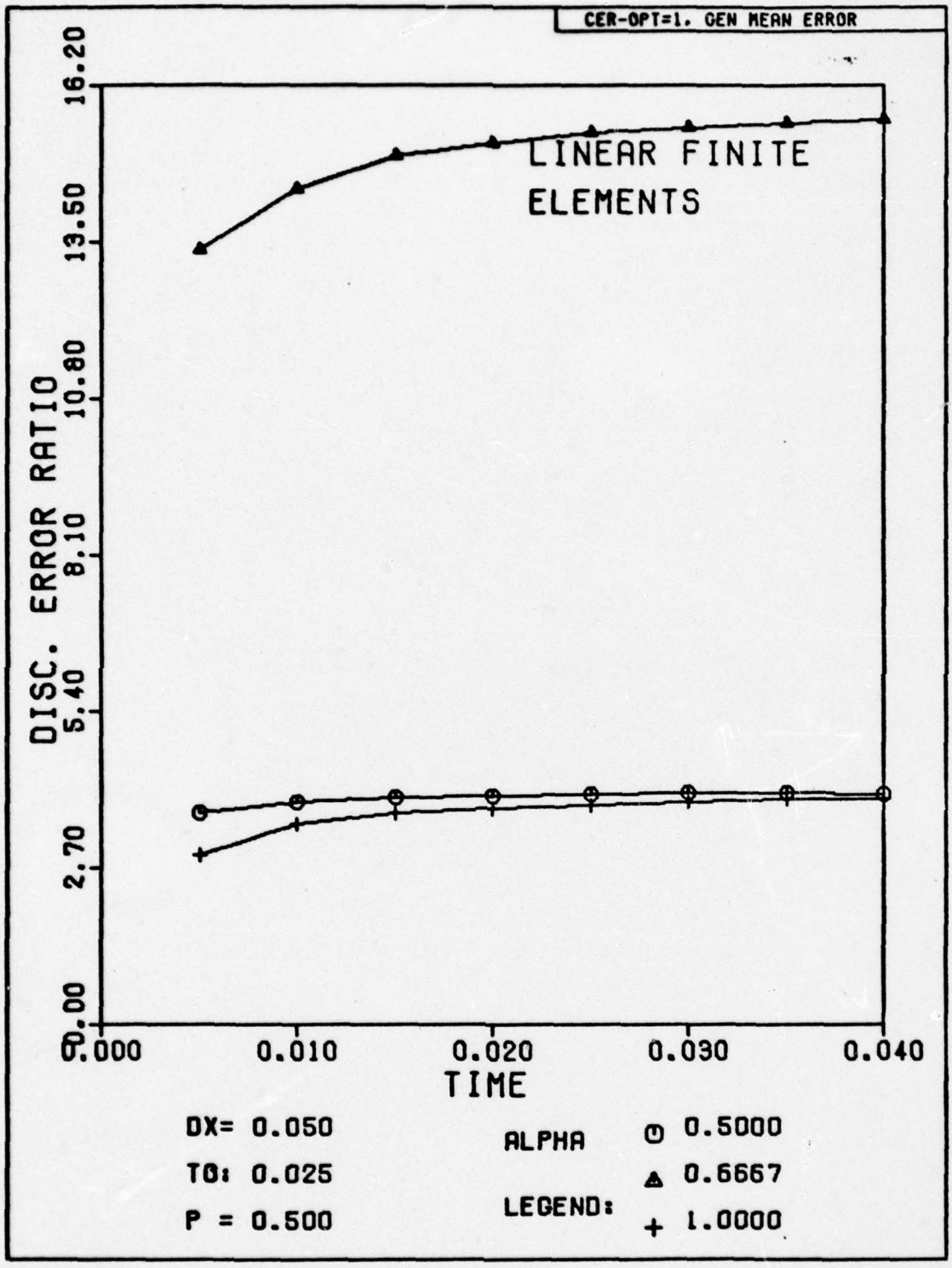


Fig. H-71. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

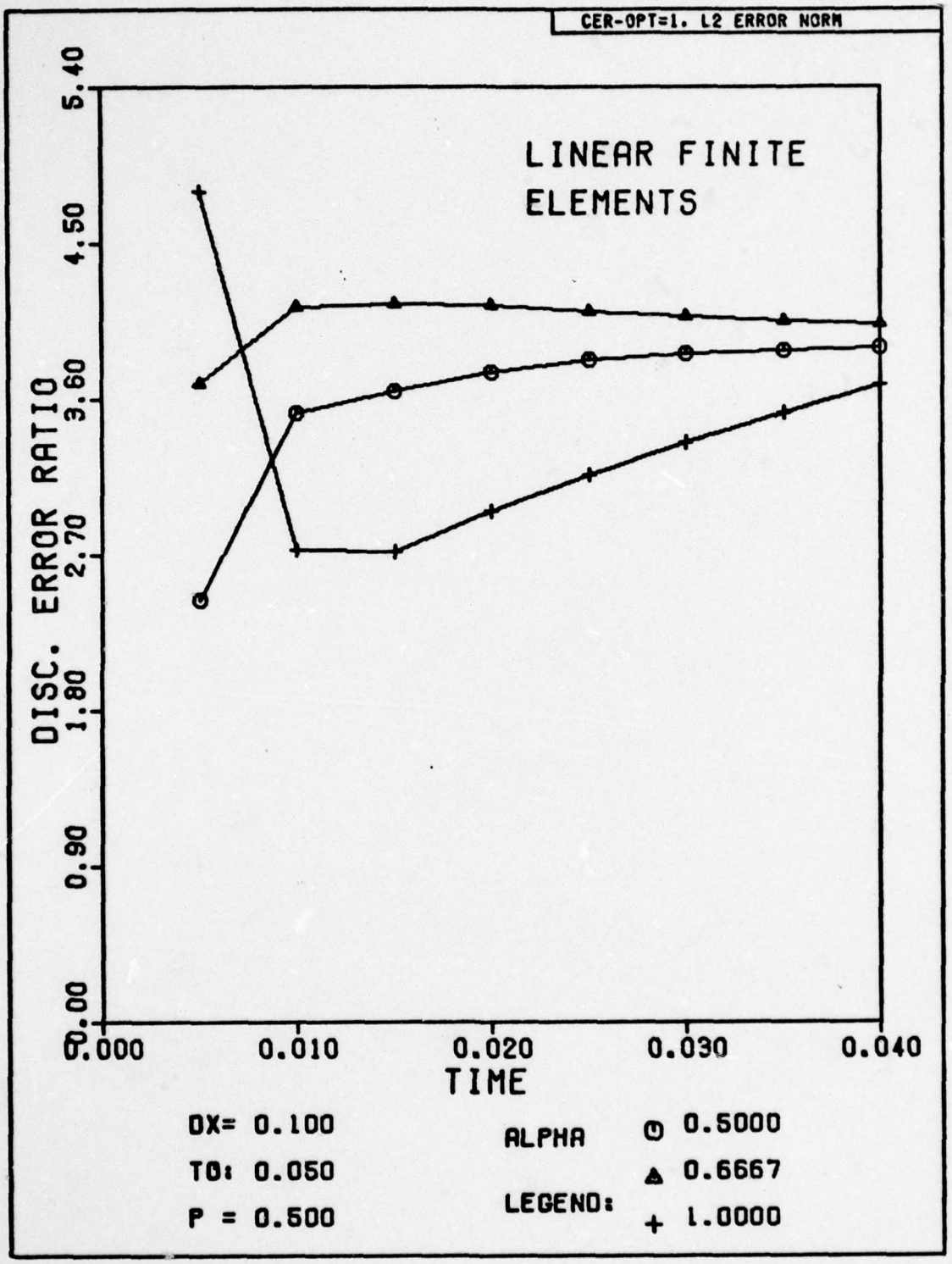


Fig. H-72. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

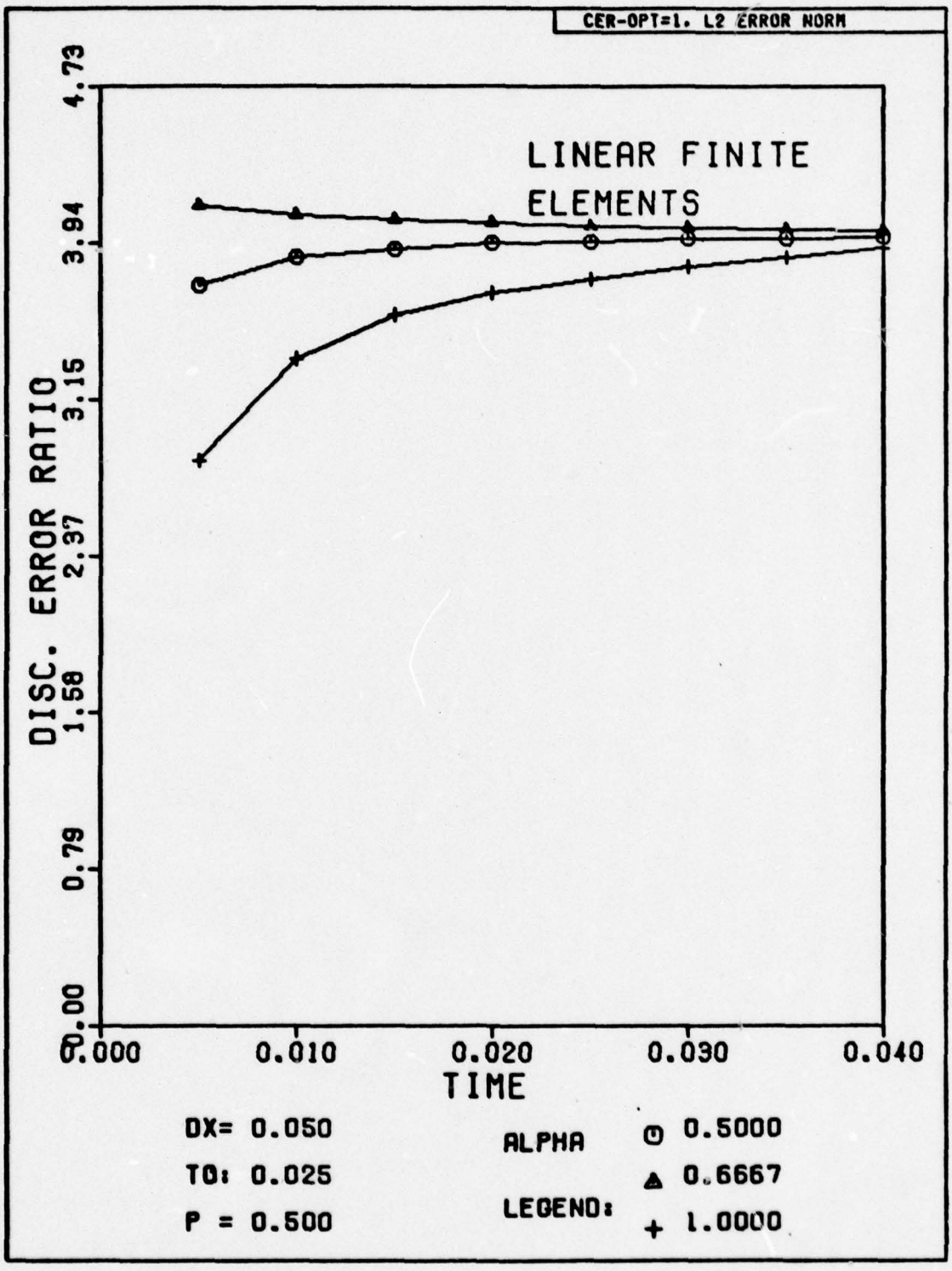


Fig. H-73. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

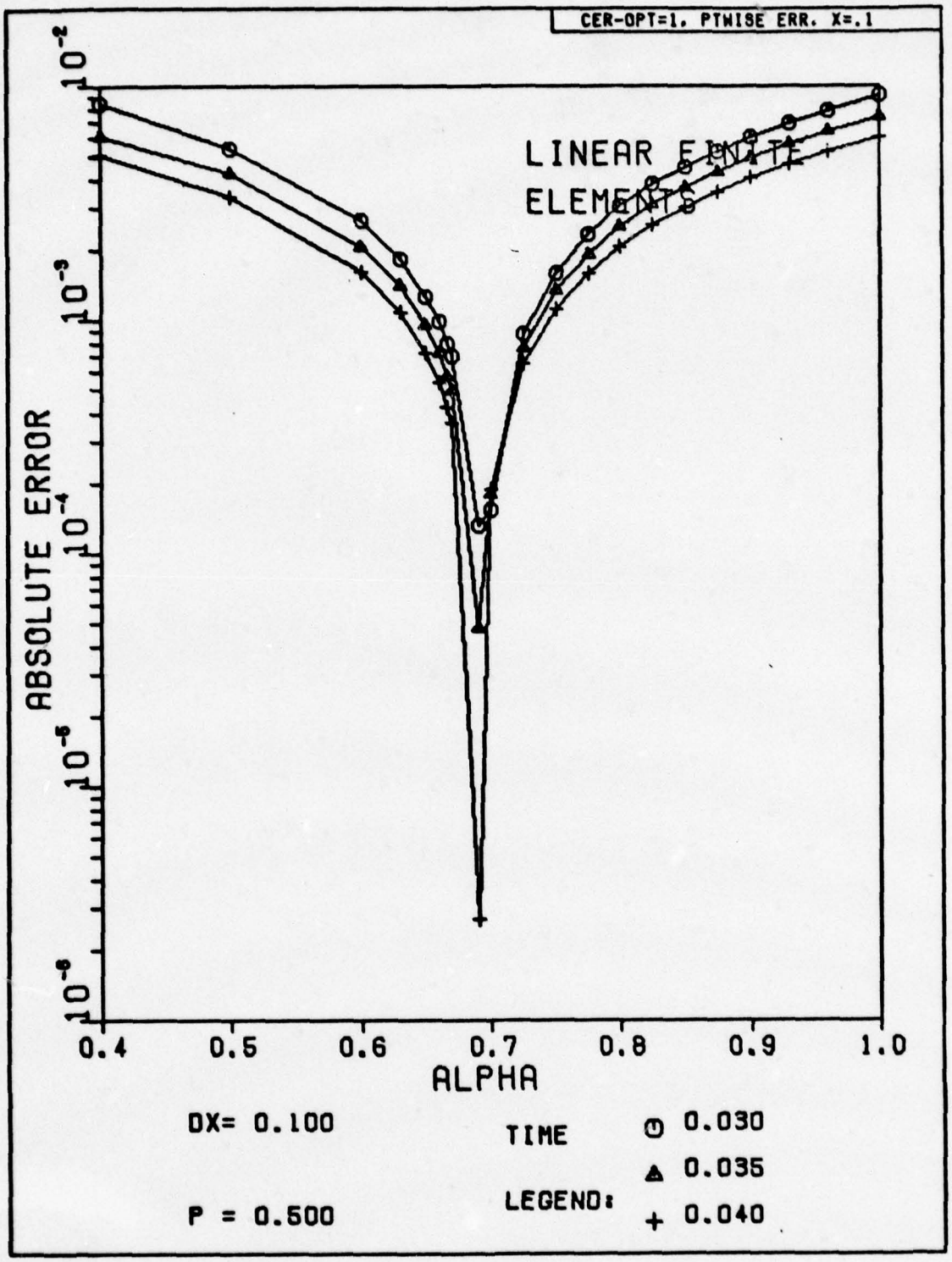


Fig. H-74. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

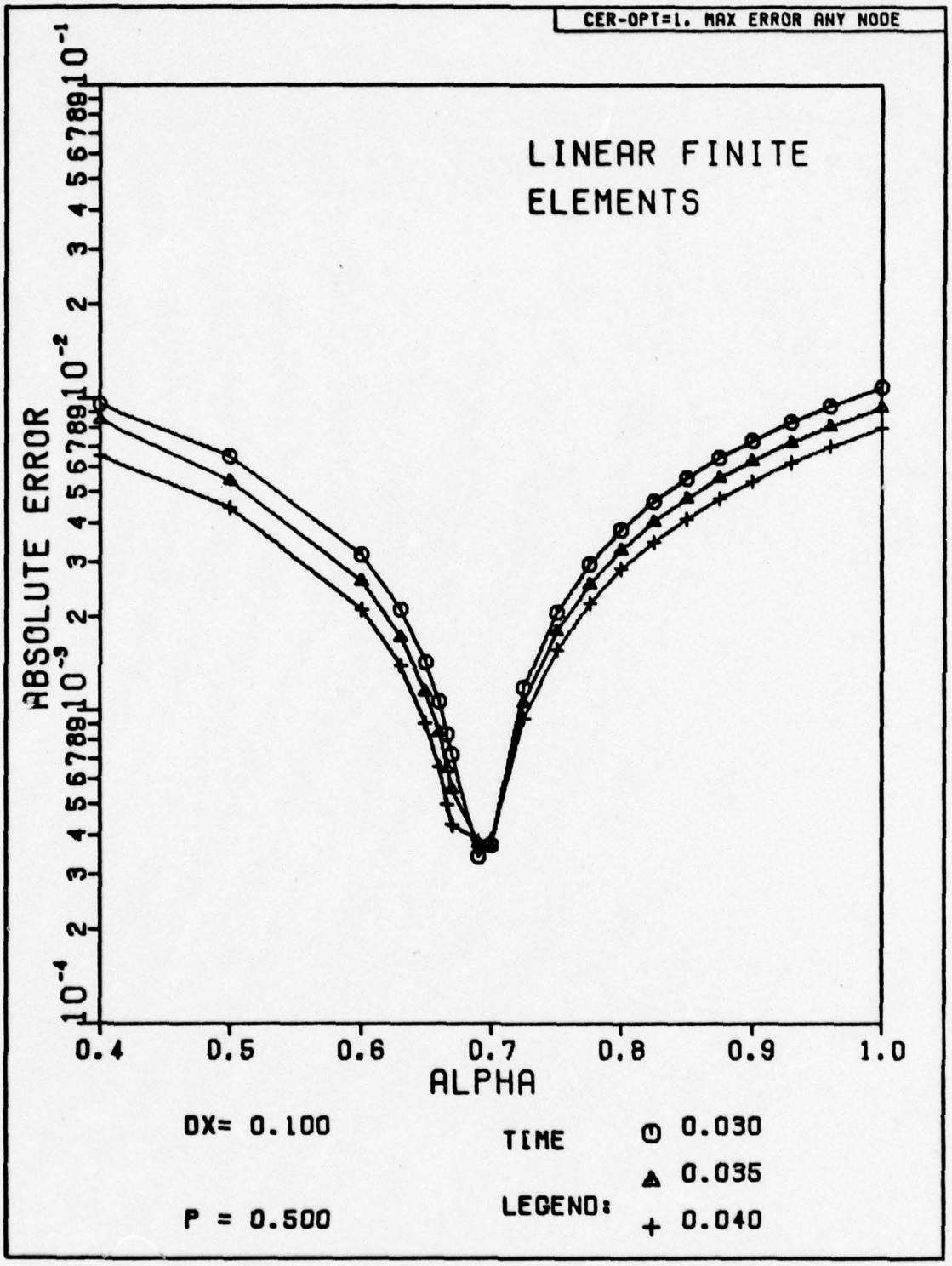


Fig. H-75. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

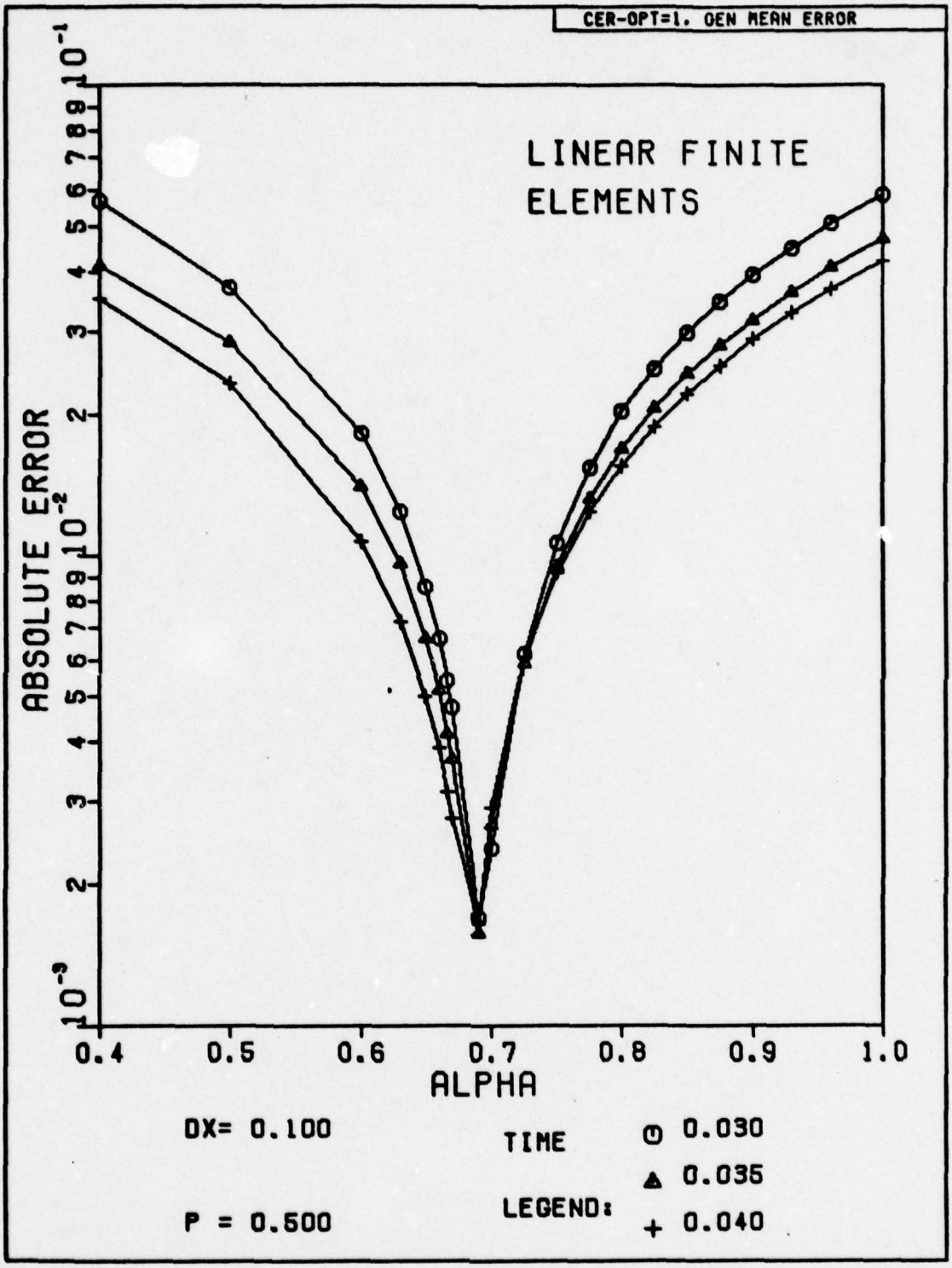


Fig. H-76. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

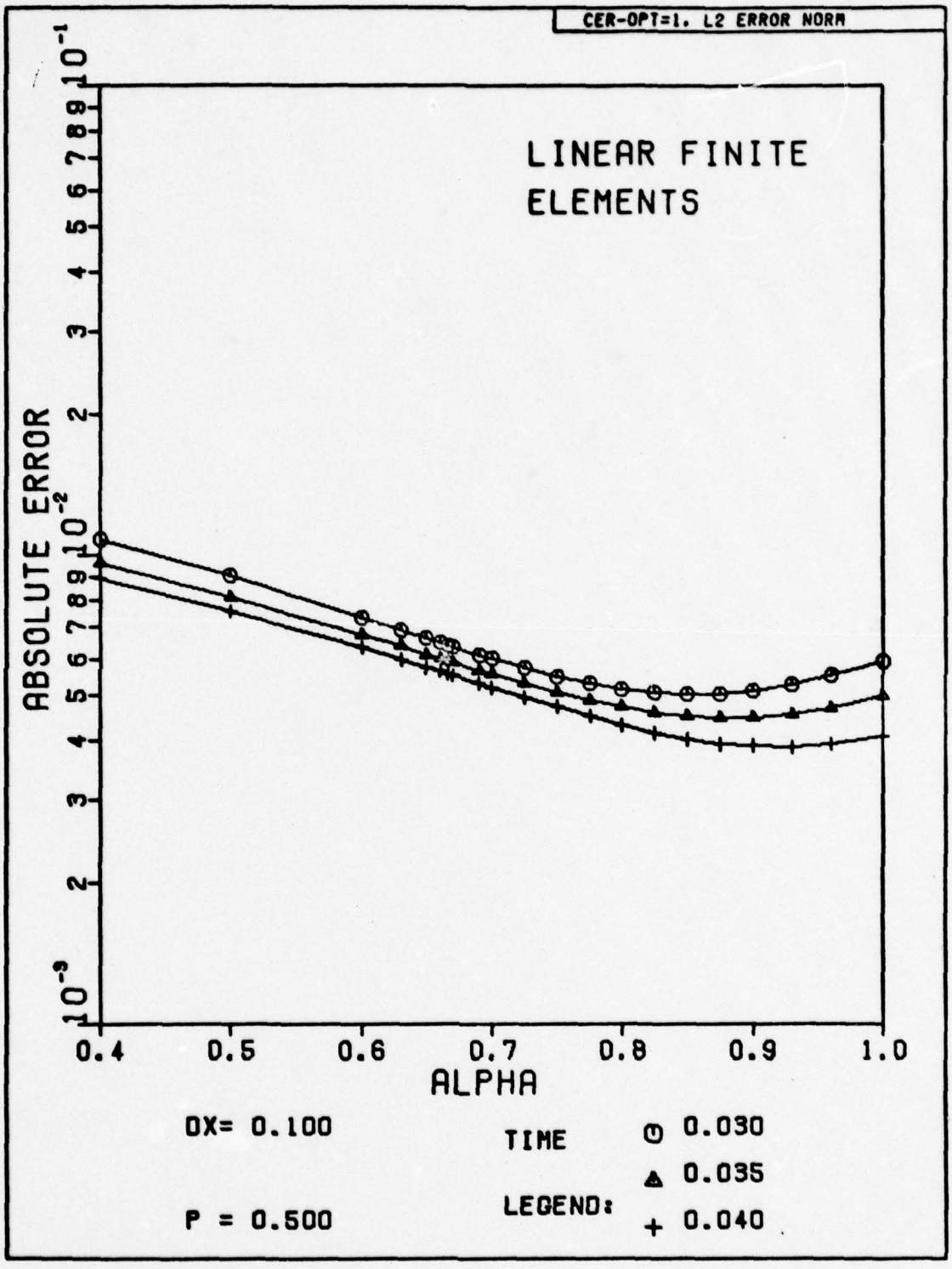


Fig. H-77. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

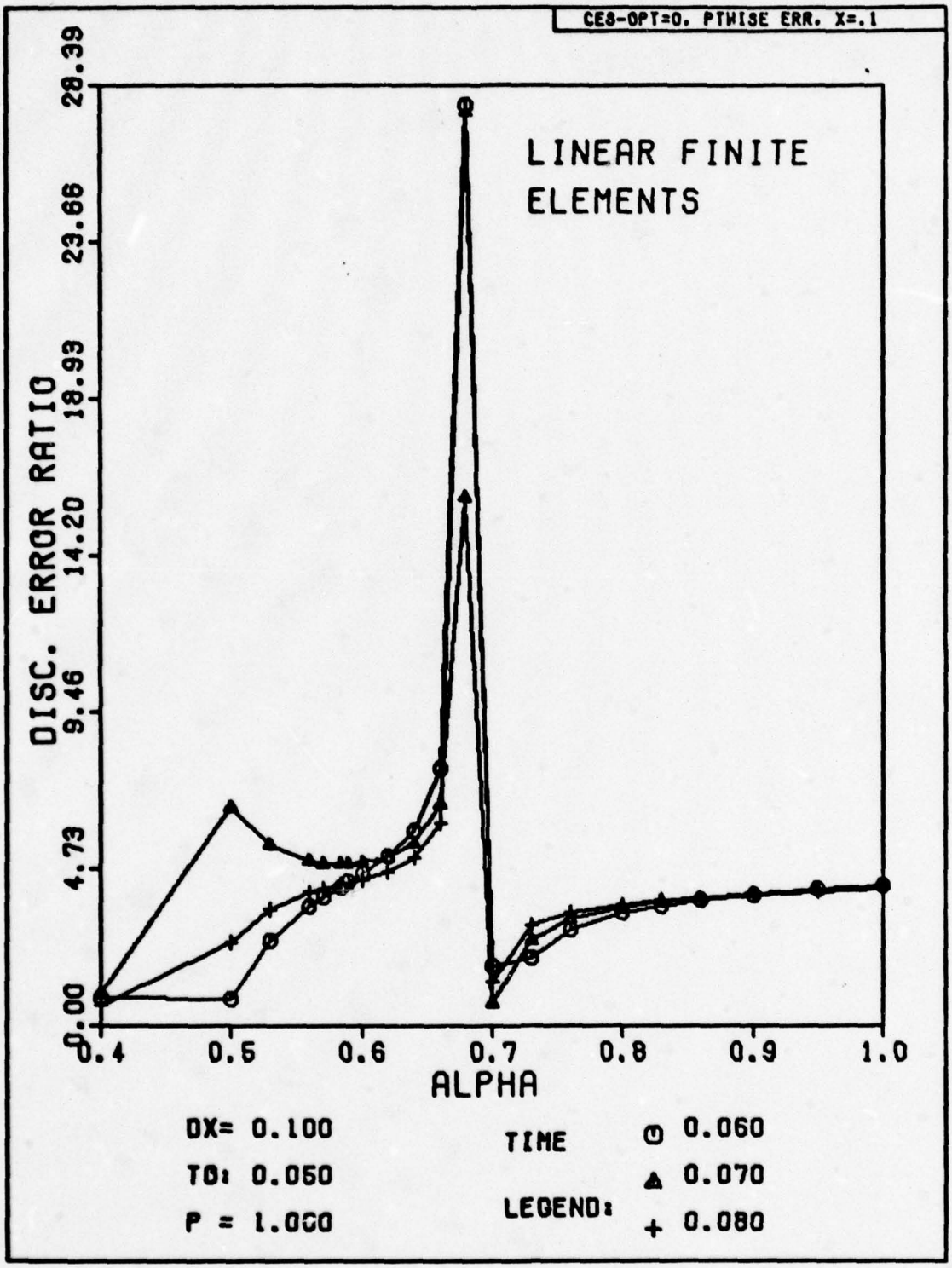


Fig. H.78. Discretization Error Ratio Versus Alpha for Problem One.

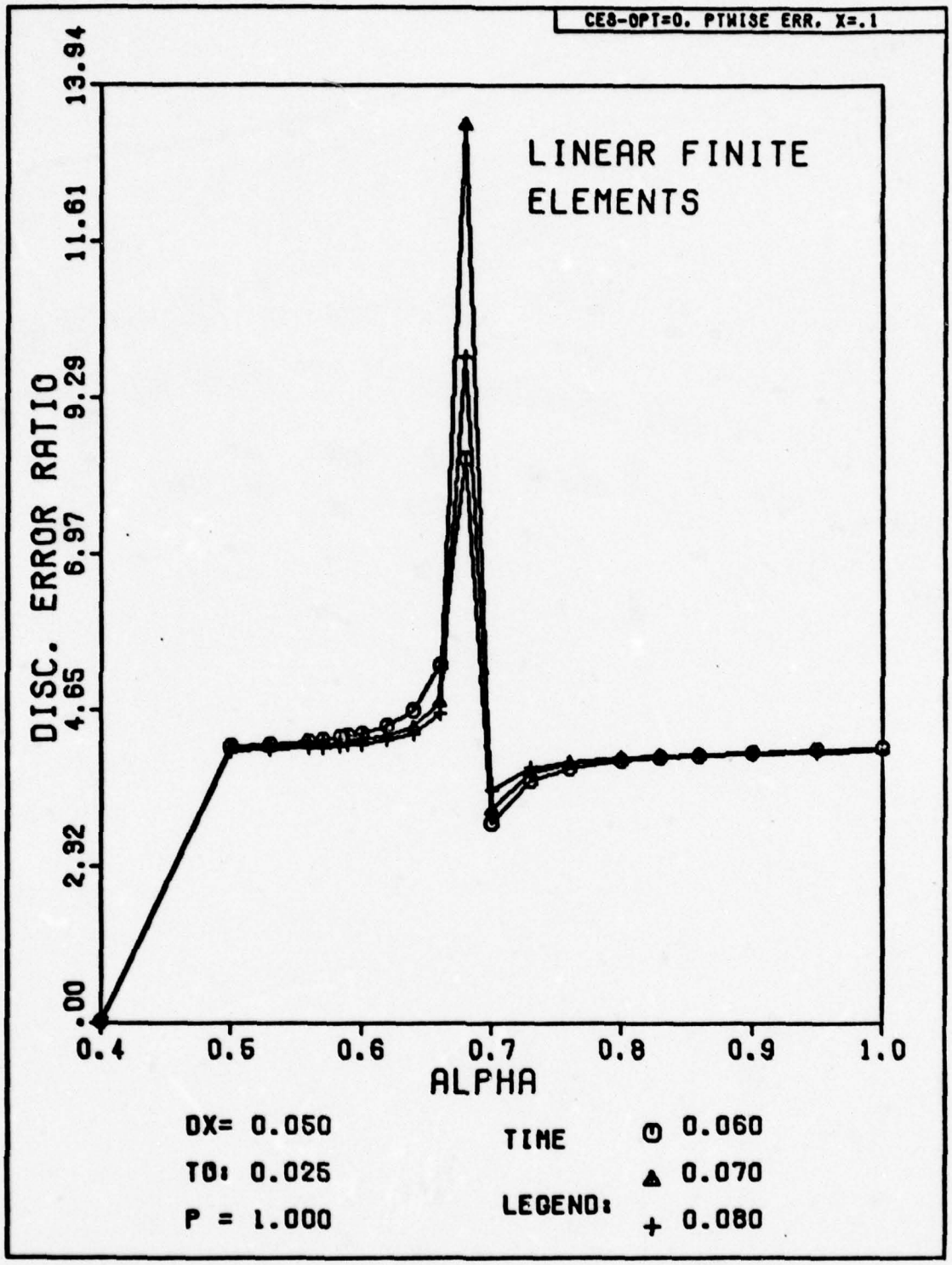


Fig. H-79. Discretization Error Ratio Versus Alpha for Problem One.

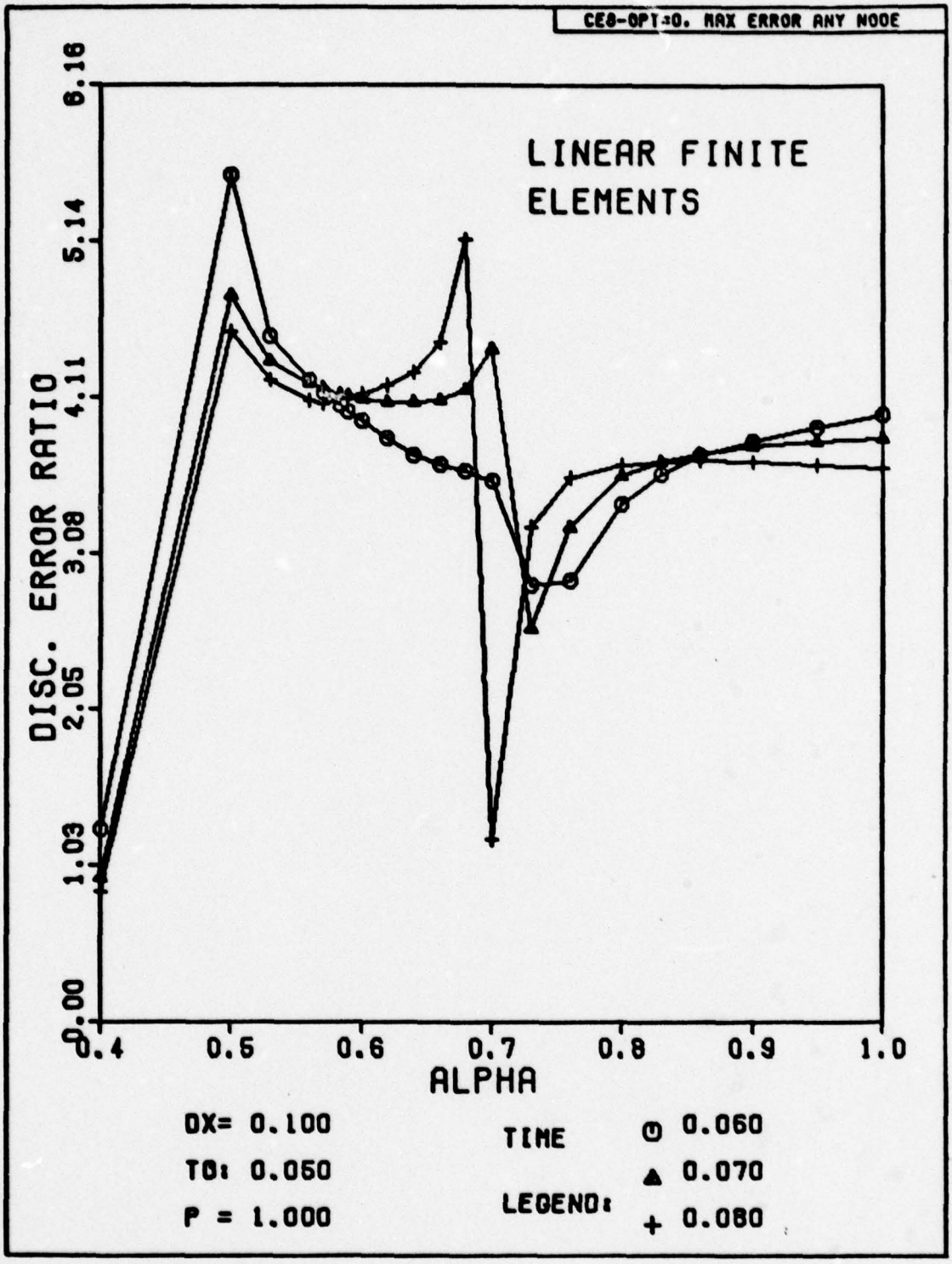


Fig. H-80. Discretization Error Ratio Versus Alpha for Problem One.

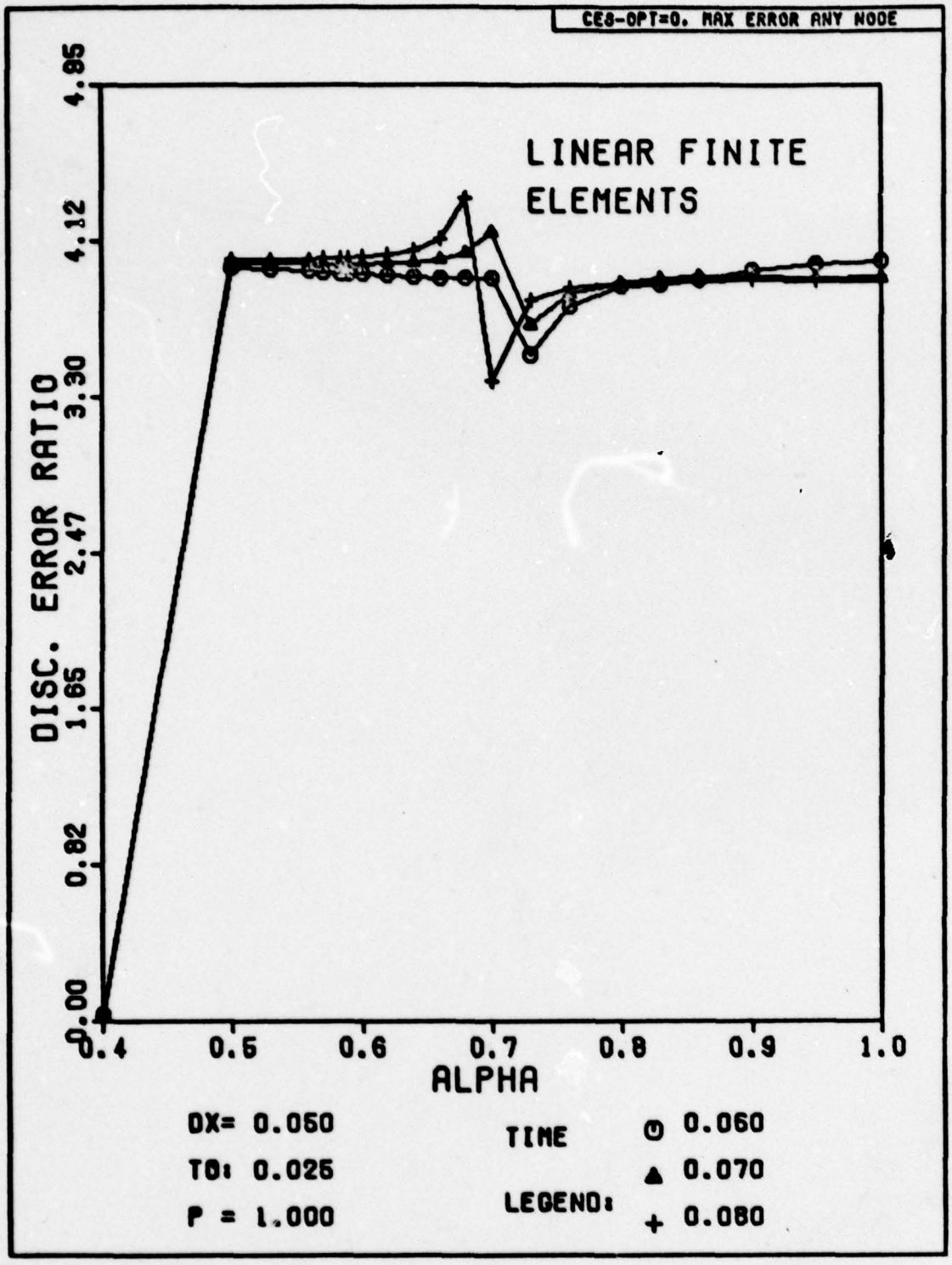


Fig. H-81. Discretization Error Ratio Versus Alpha for Problem One.

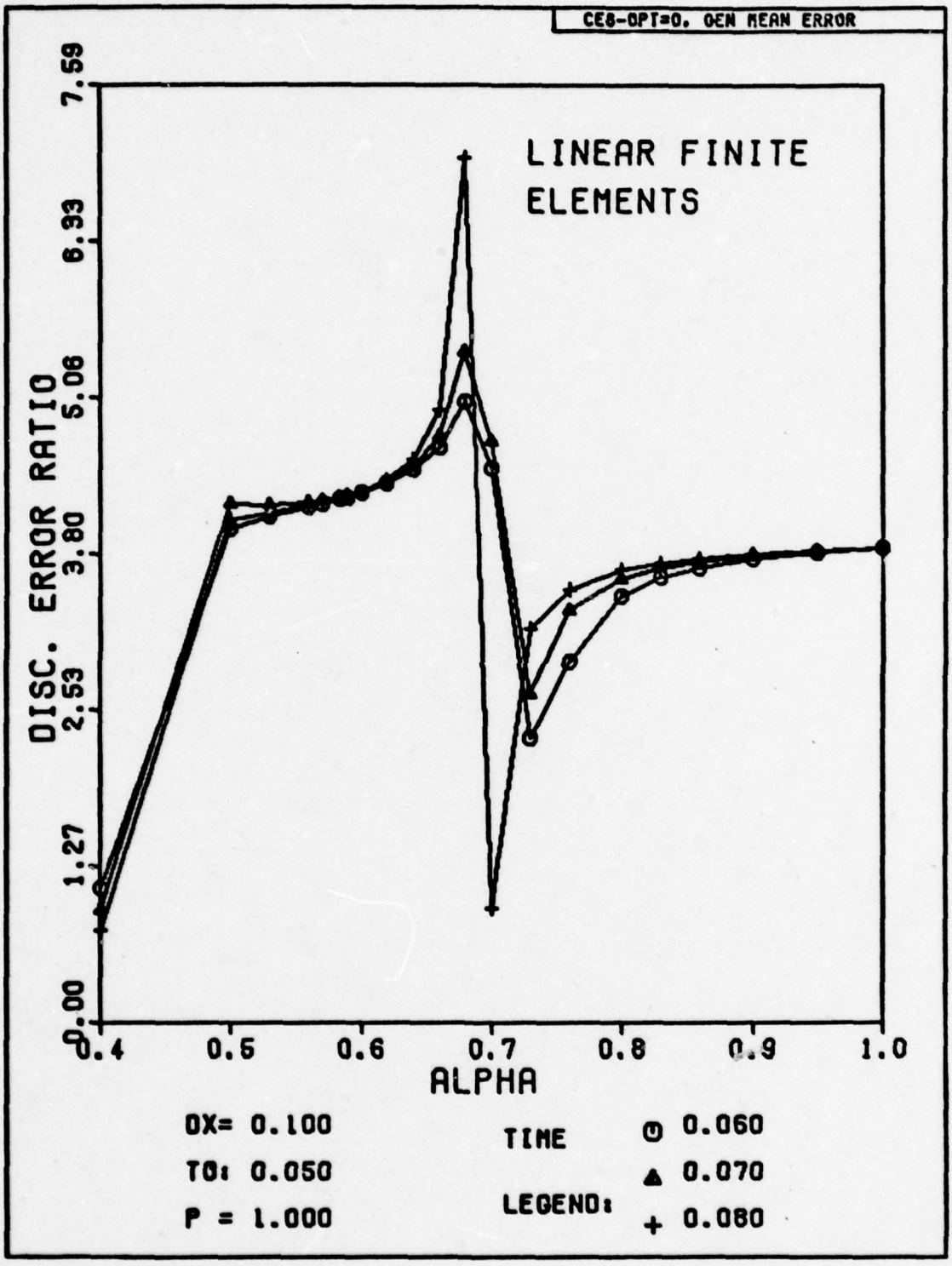


Fig. H-82. Discretization Error Ratio Versus Alpha for Problem One.

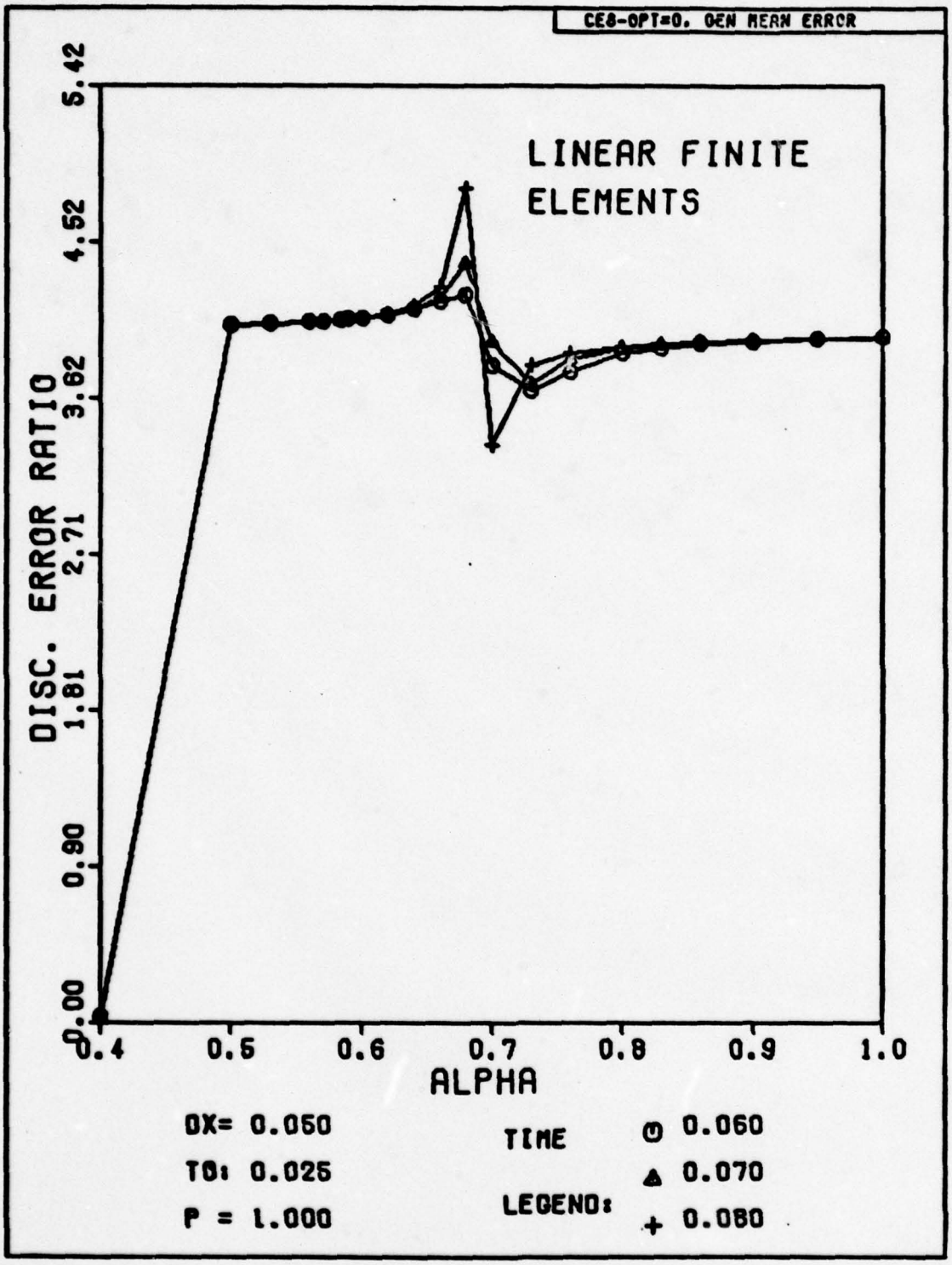


Fig. H-83. Discretization Error Ratio Versus Alpha for Problem One.

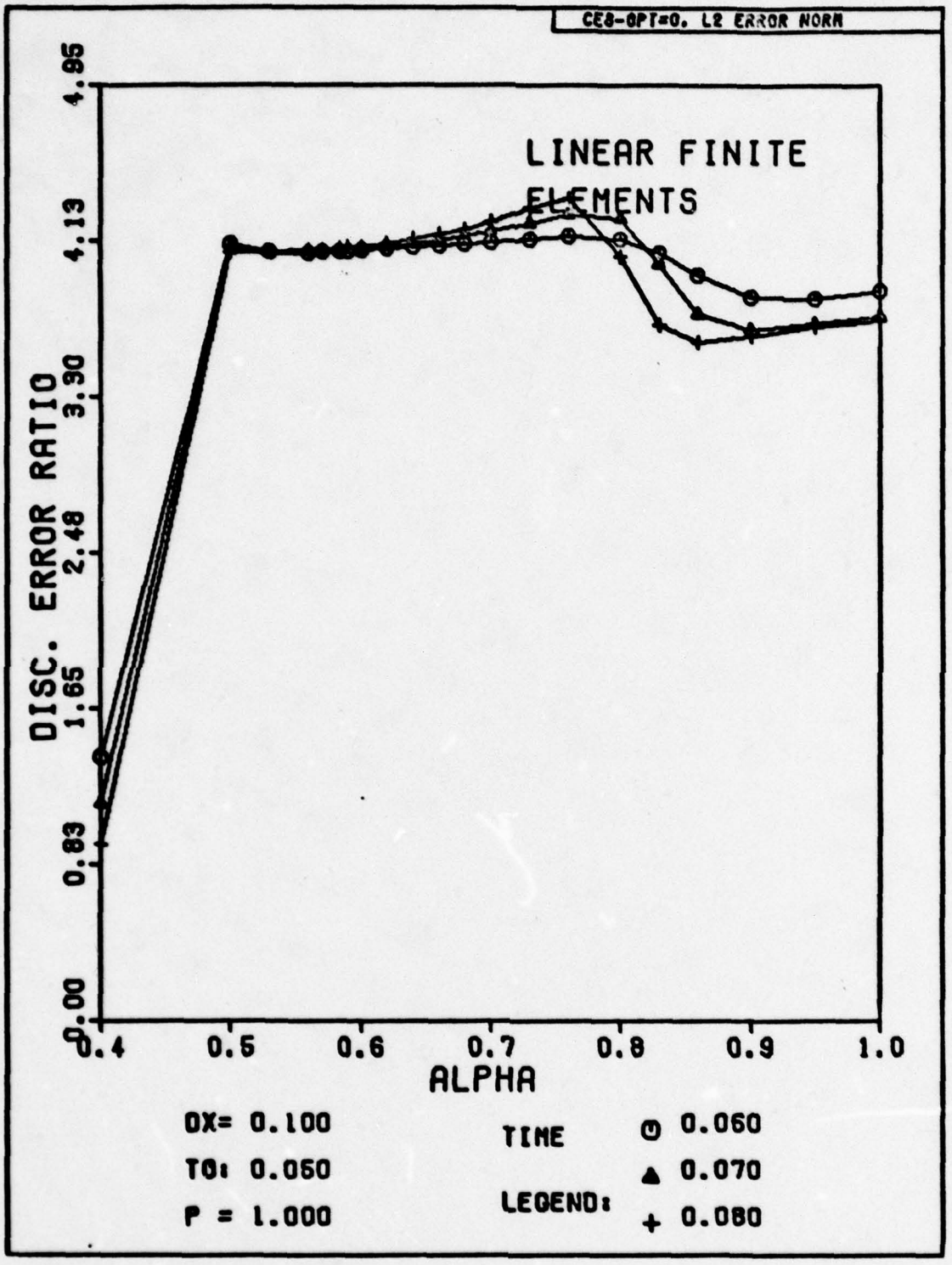


Fig. H-84. Discretization Error Ratio Versus Alpha for Problem One.

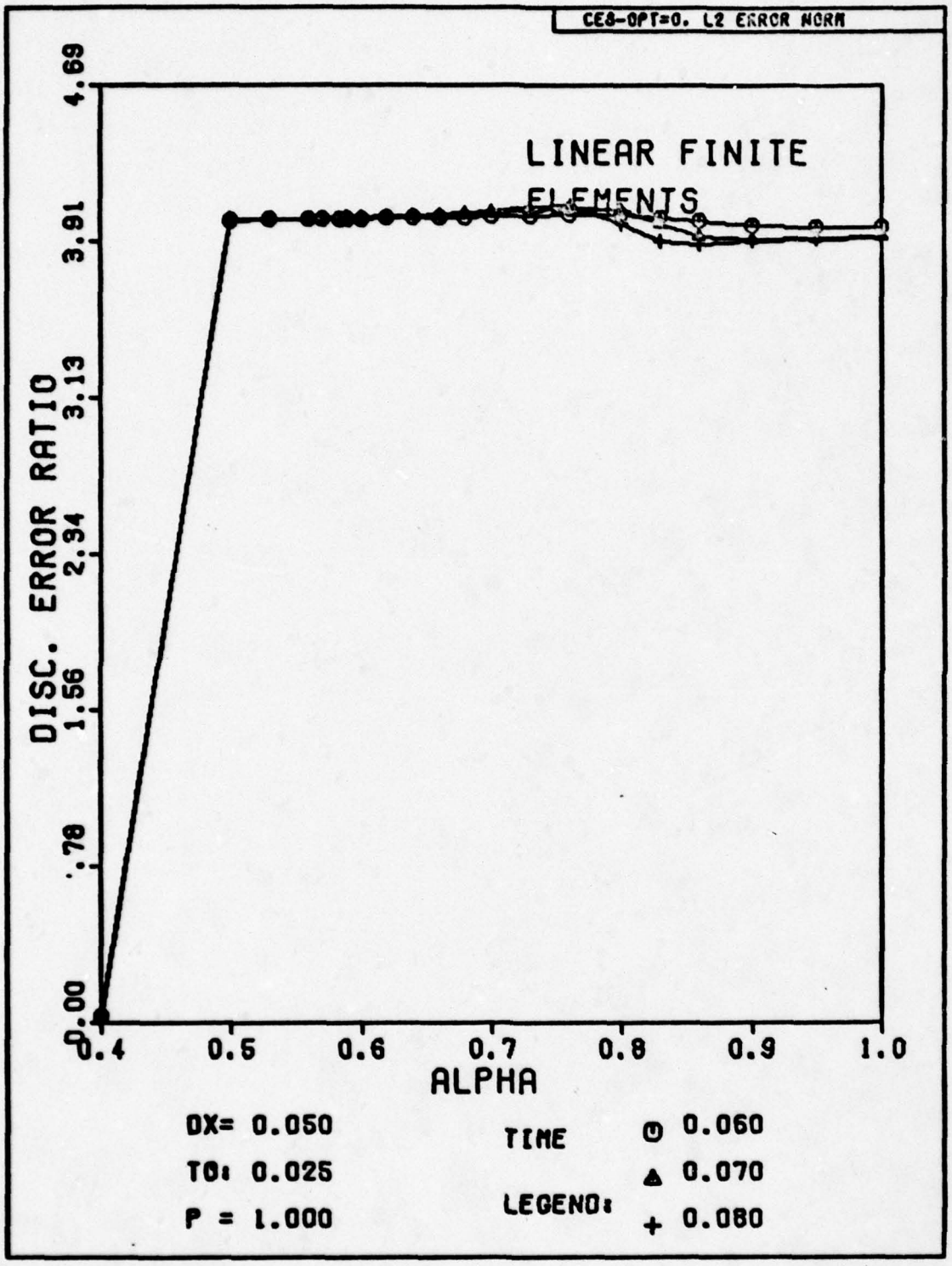


Fig. H-85. Discretization Error Ratio Versus Alpha for Problem One.

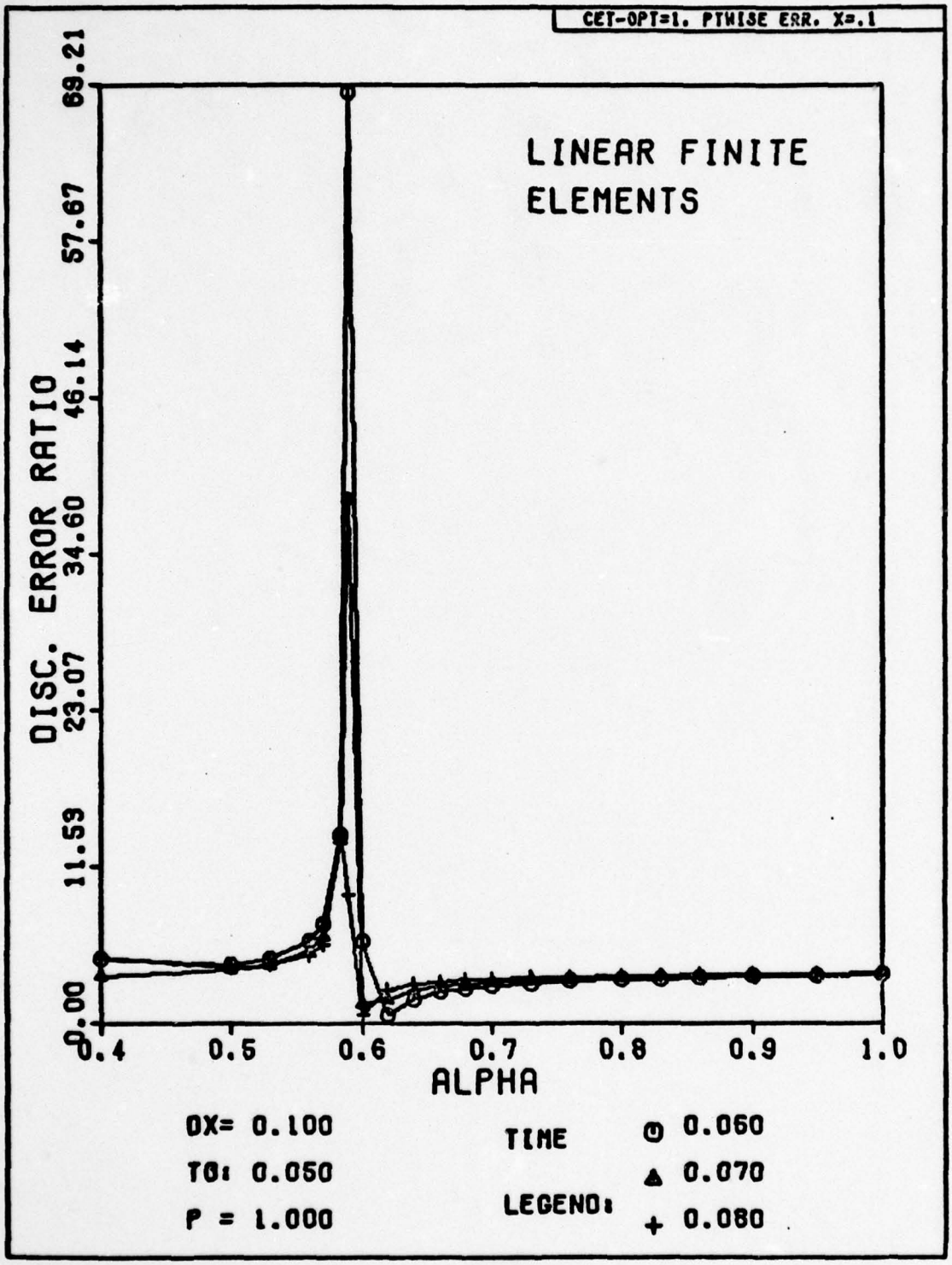


Fig. H-86. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

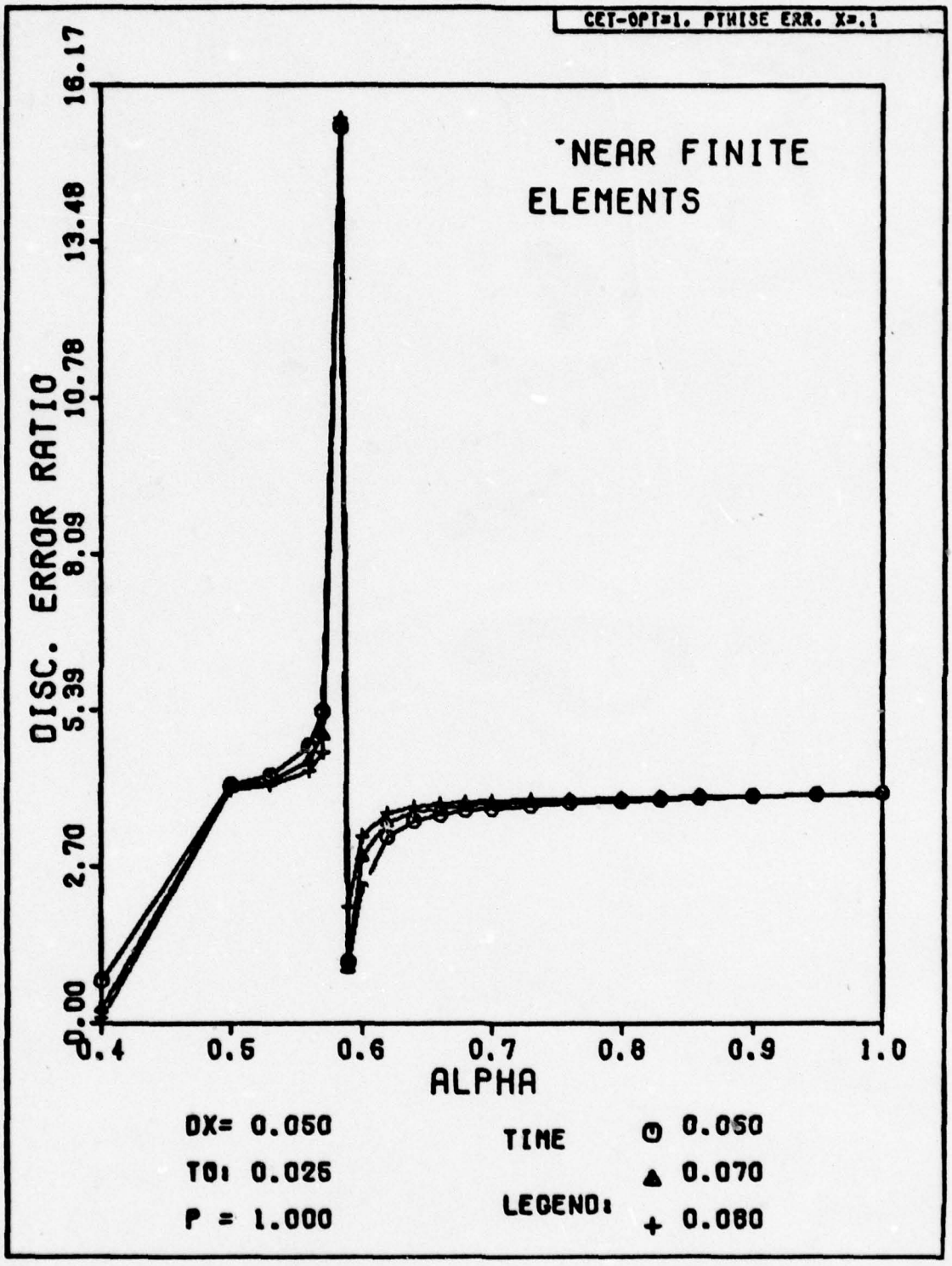


Fig. H-87. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

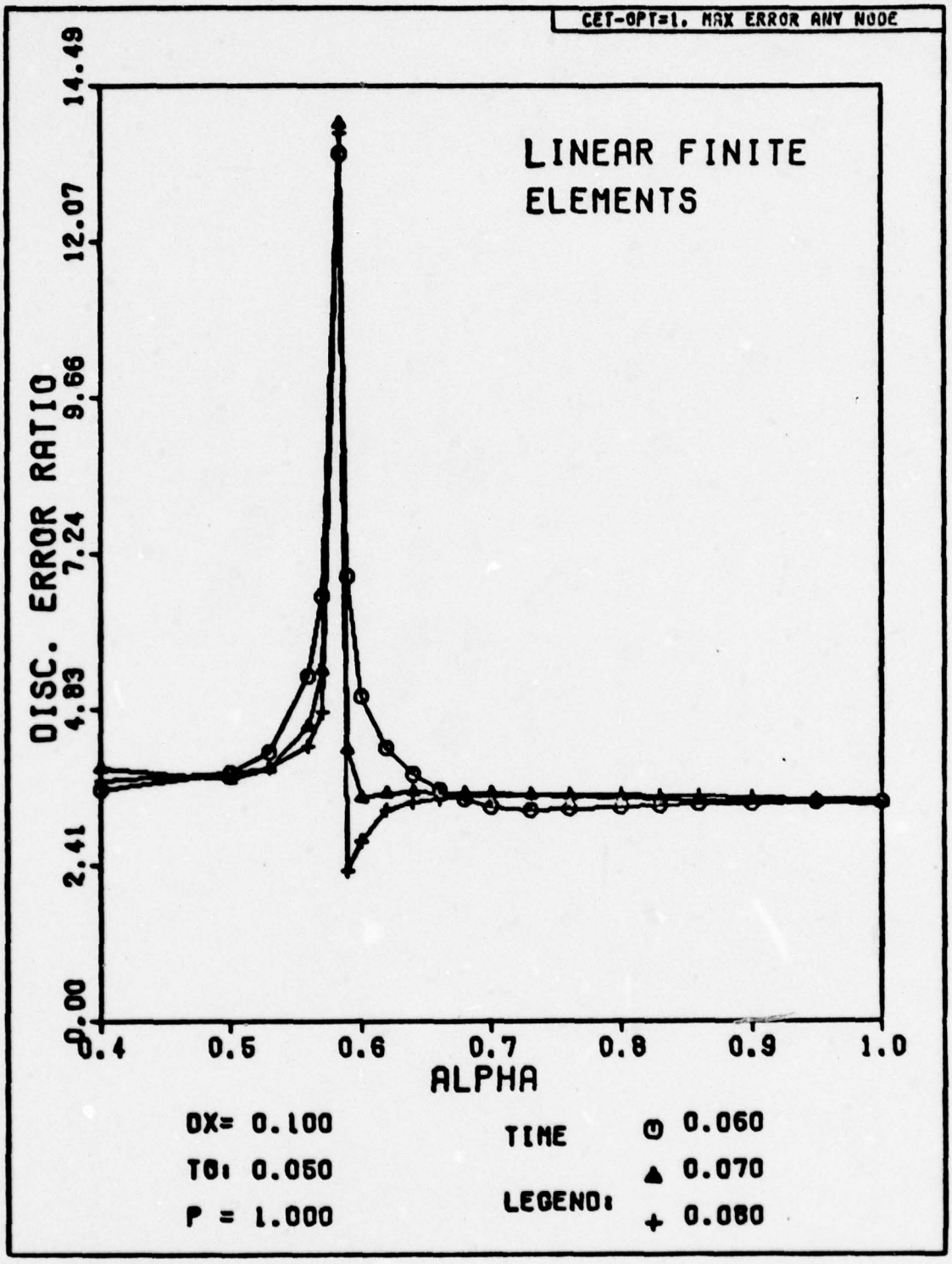


Fig. H-88. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

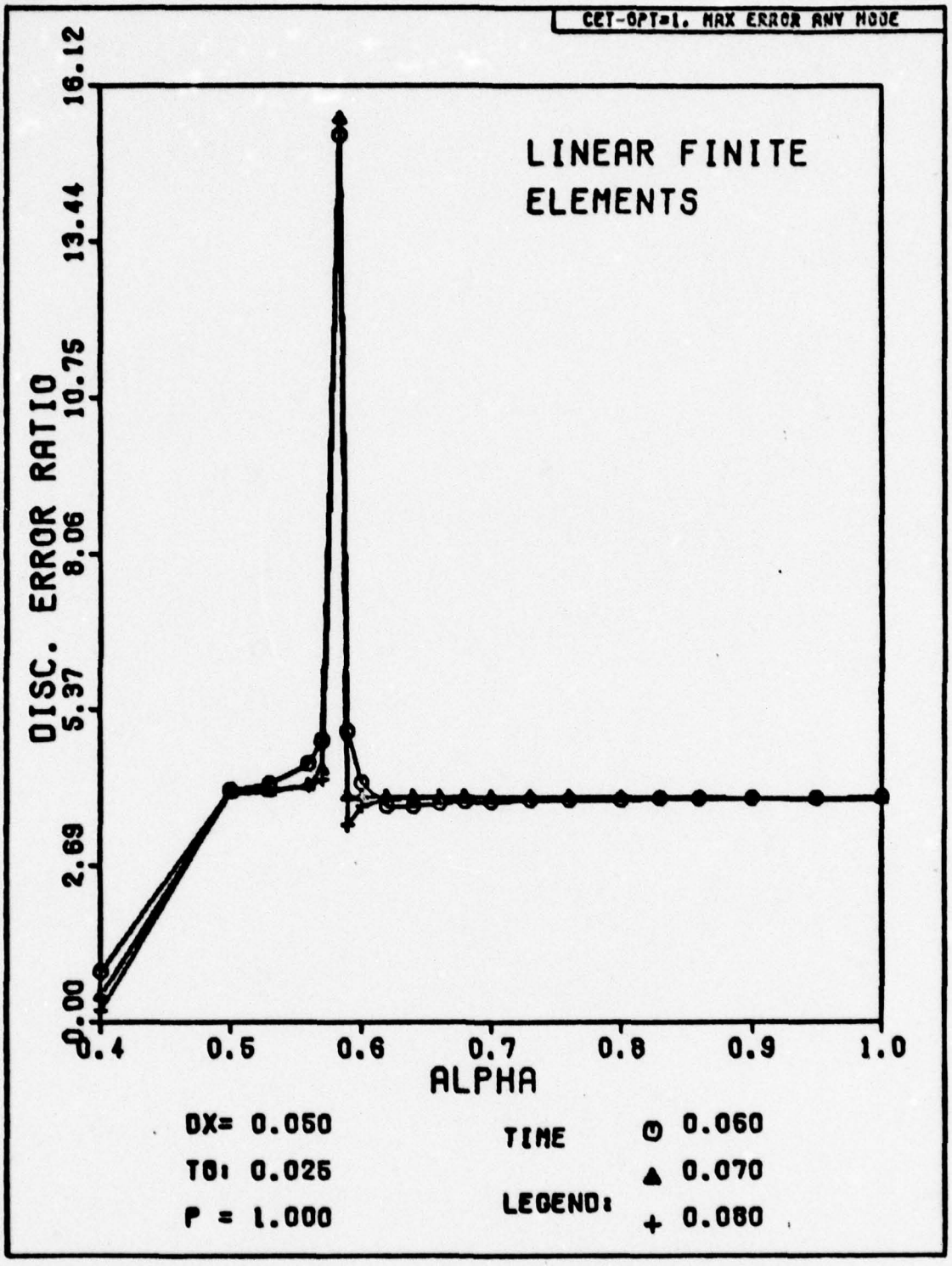


Fig. H-89. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

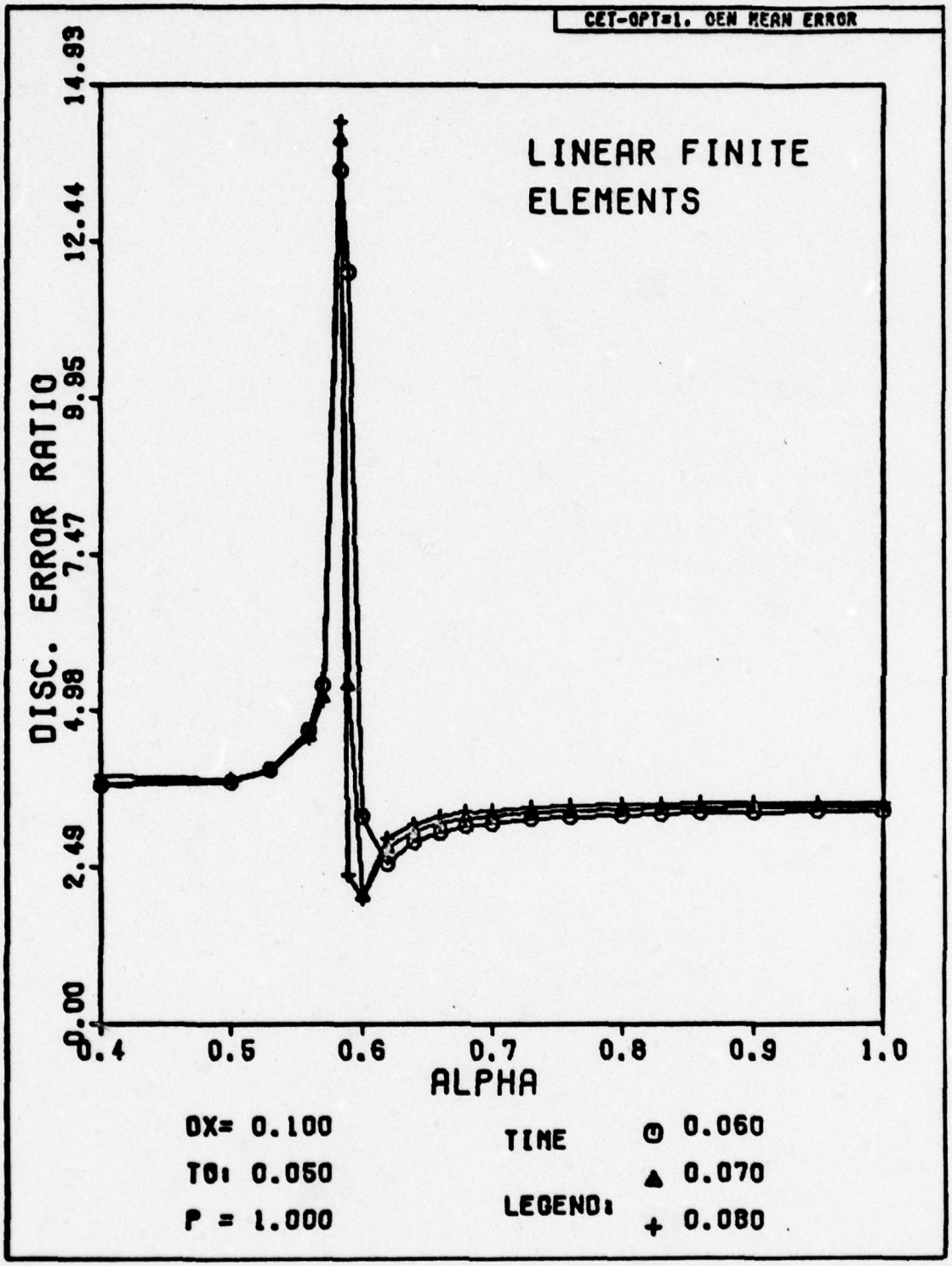


Fig. H-90. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

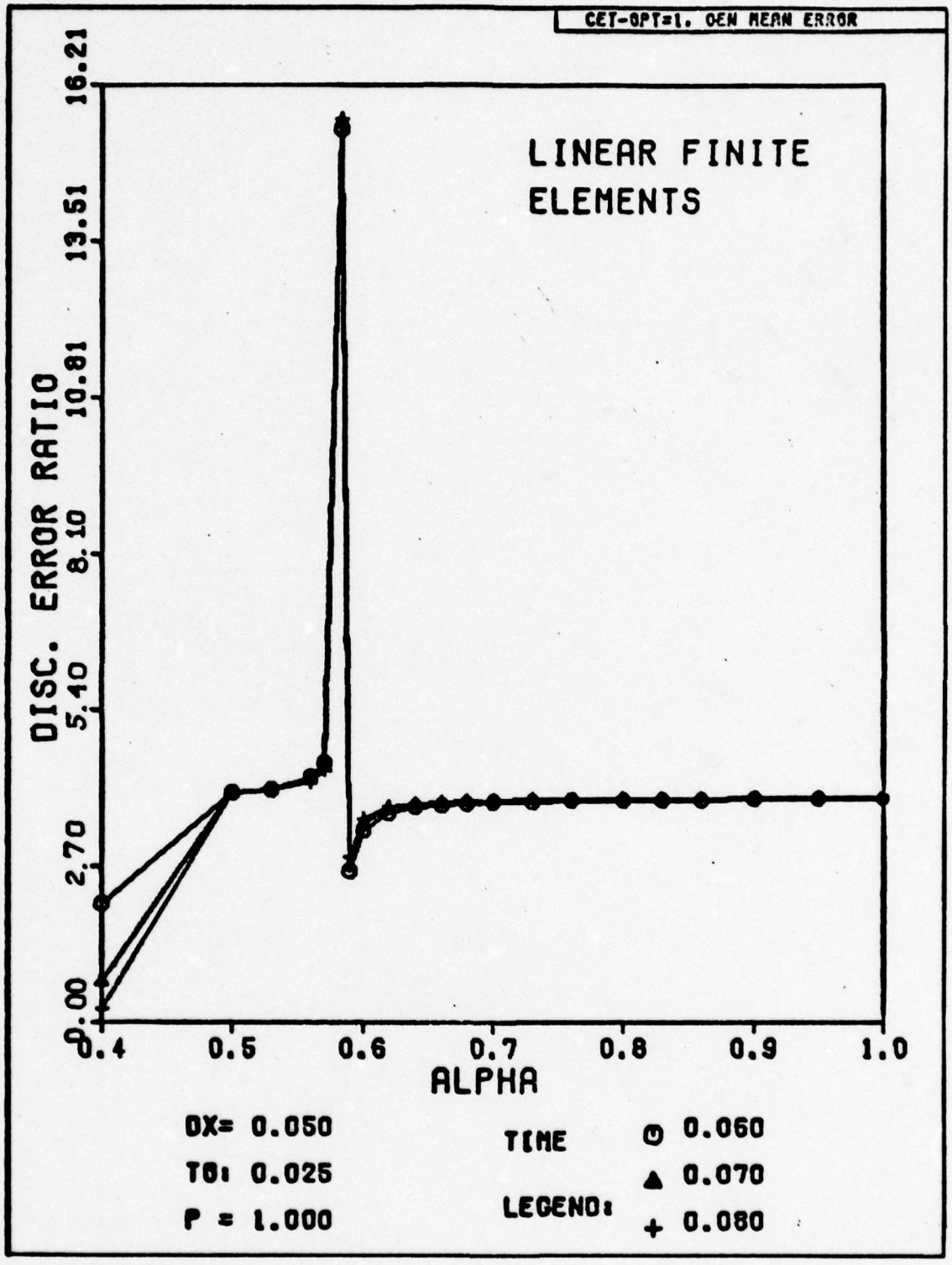


Fig. H-91. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

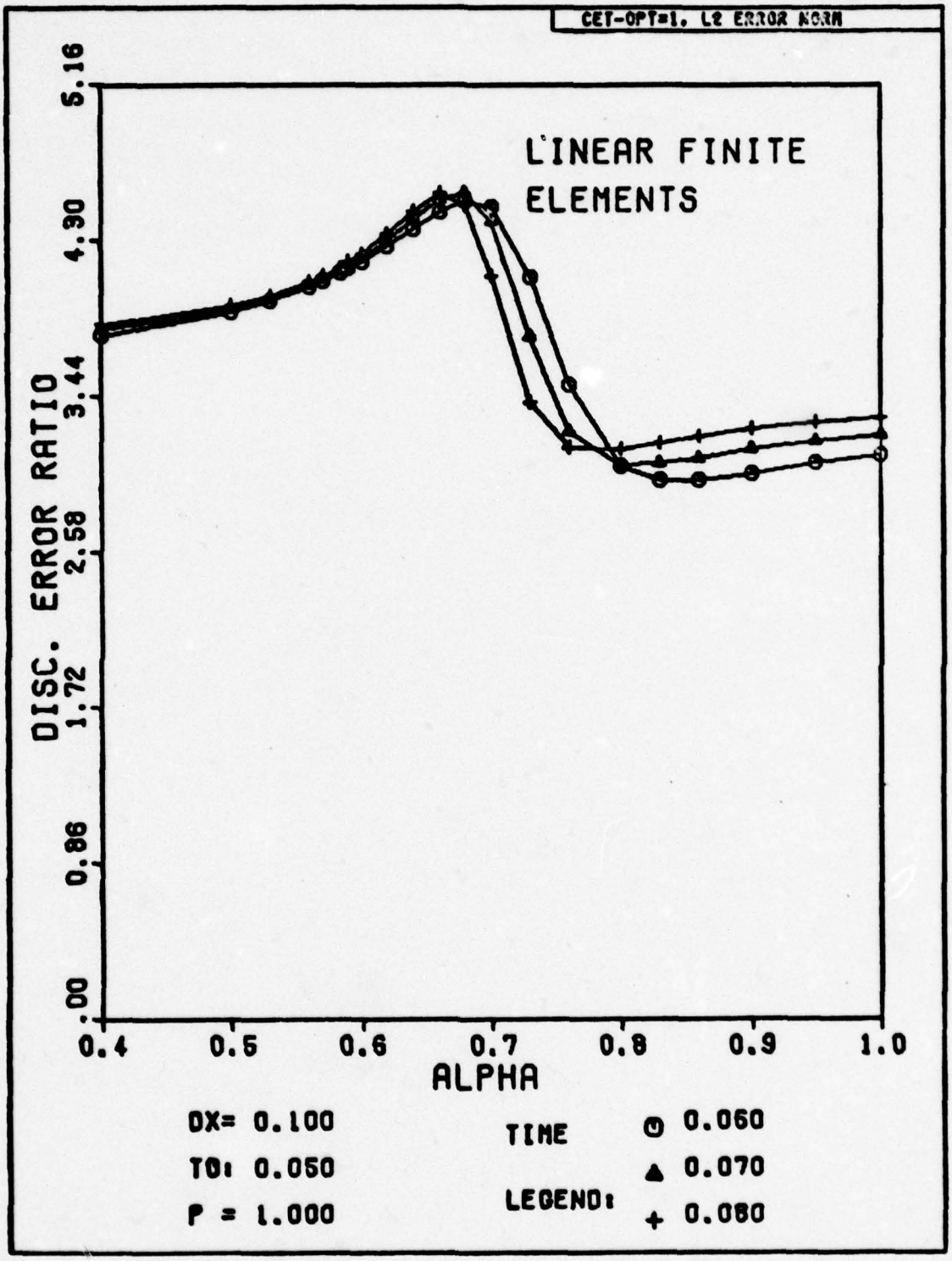


Fig. H-92. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

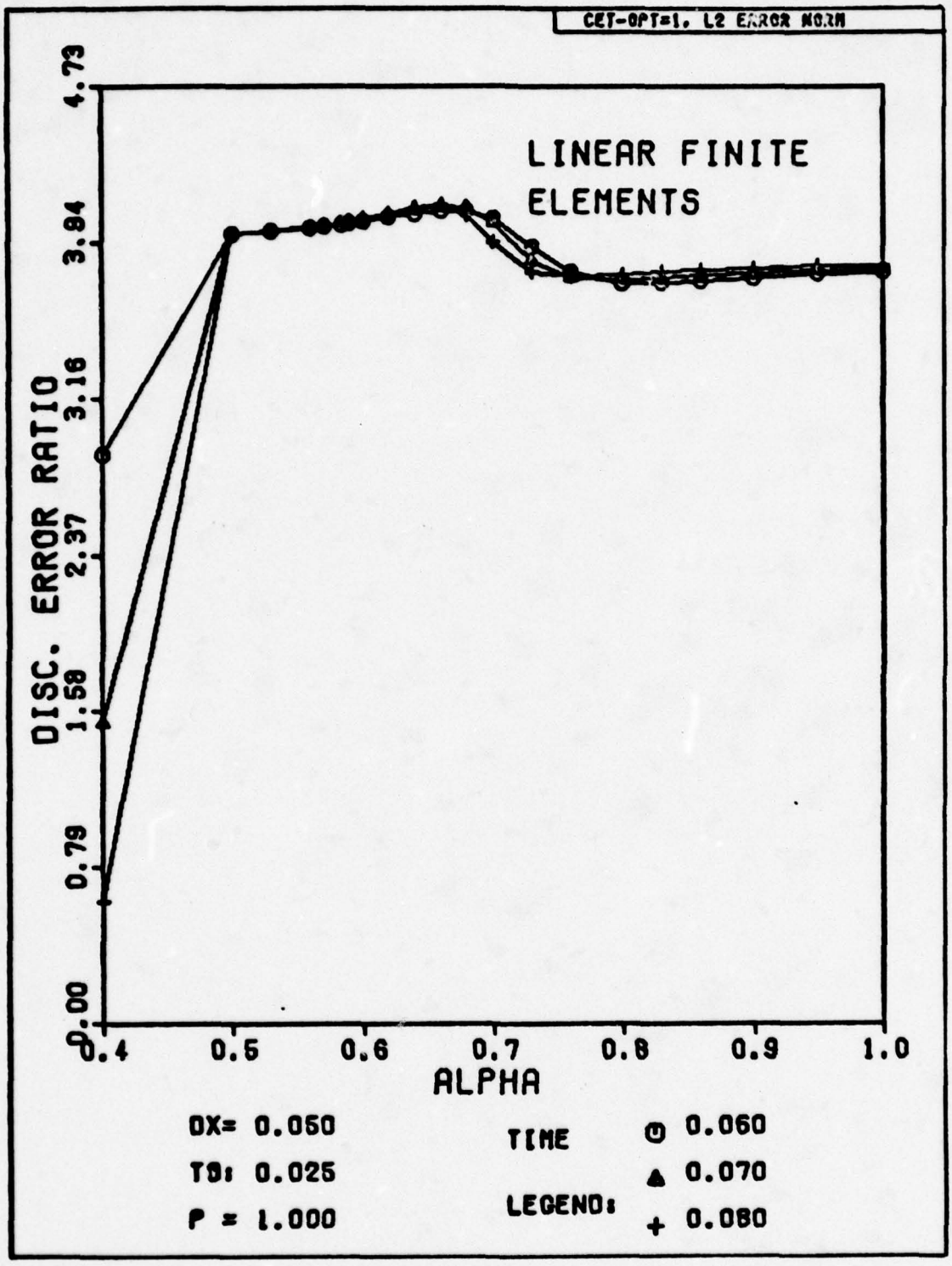


Fig. H-93. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

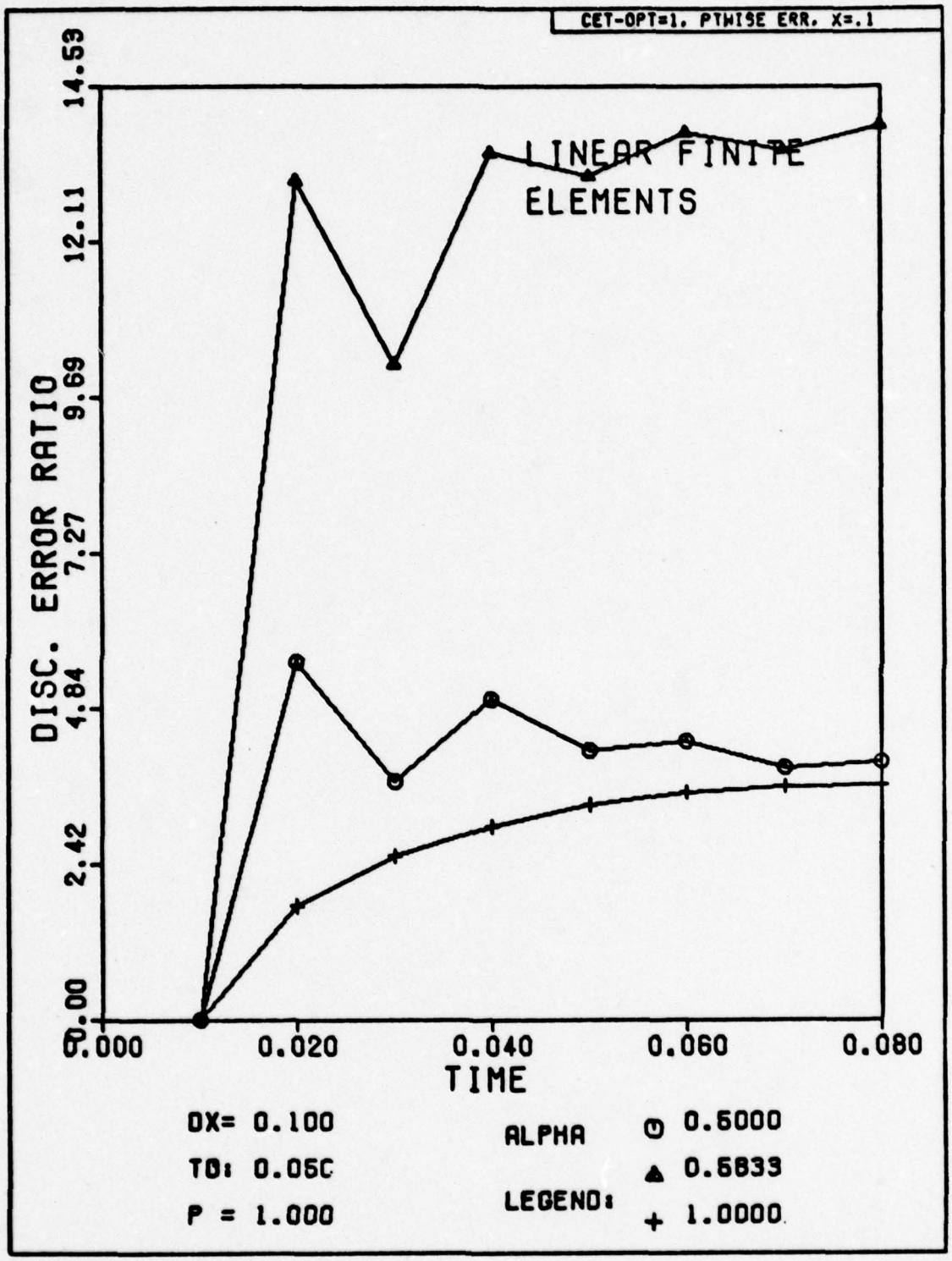


Fig. H-94. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

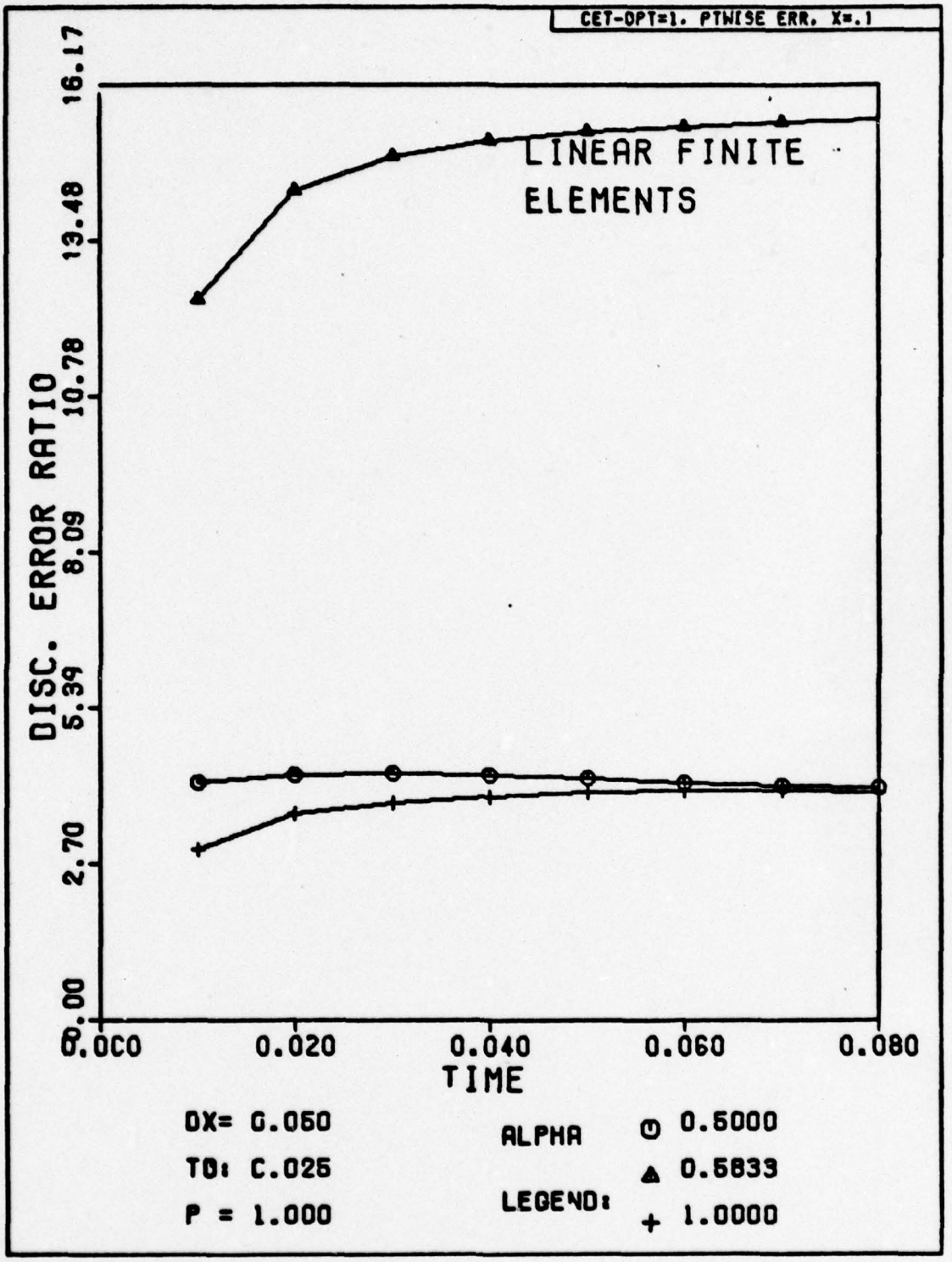


Fig. H-95. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

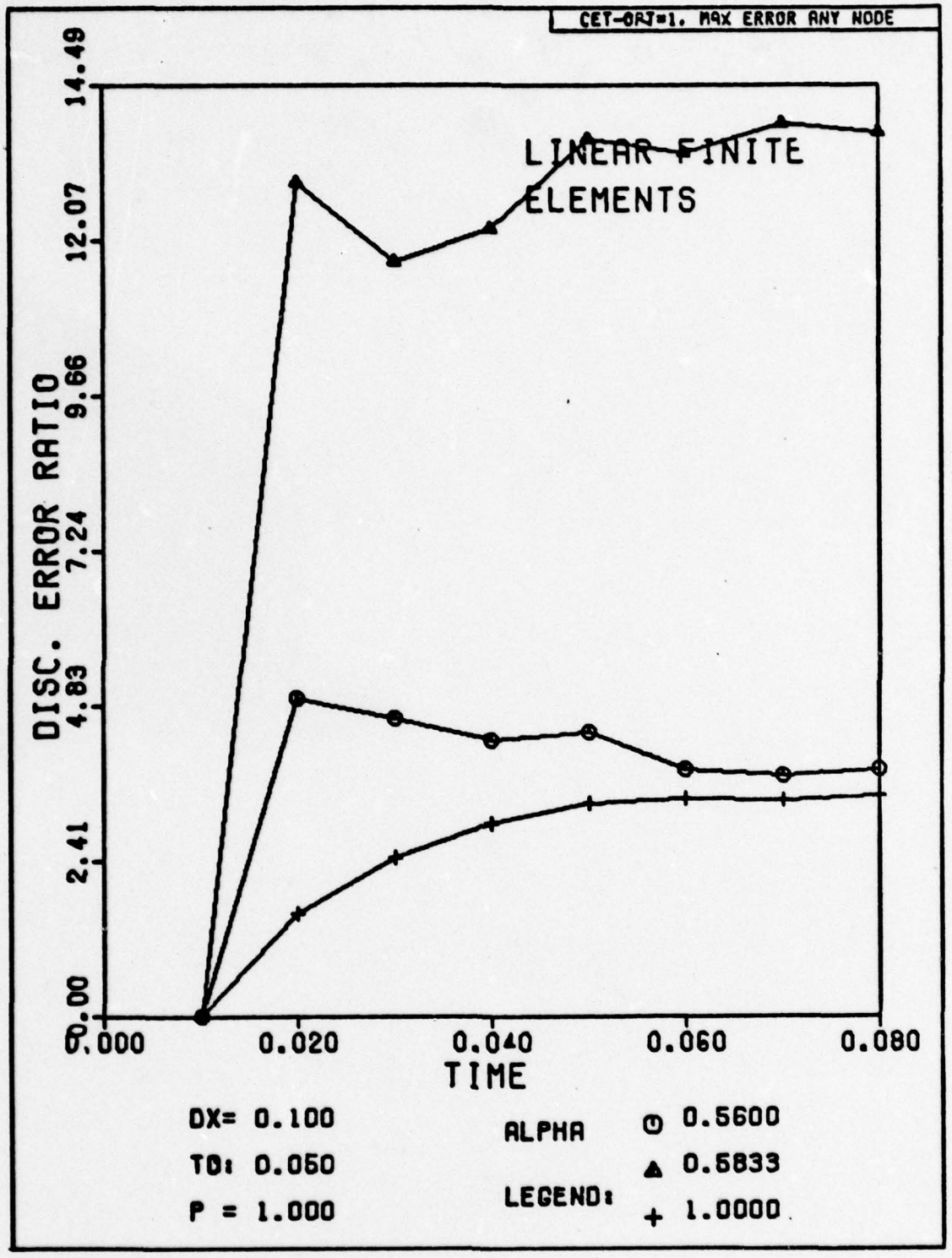


Fig. H-96. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

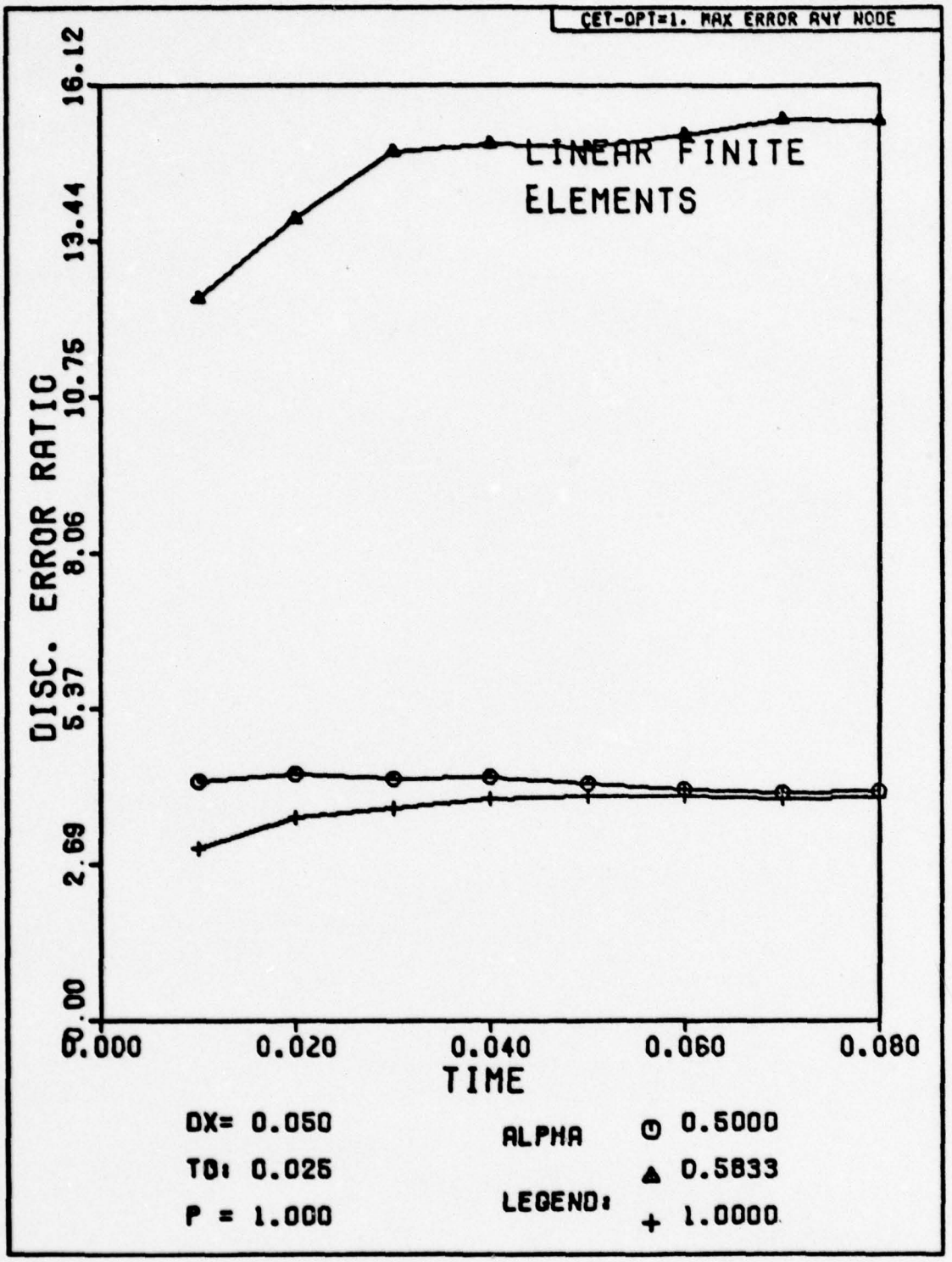


Fig. H-97. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

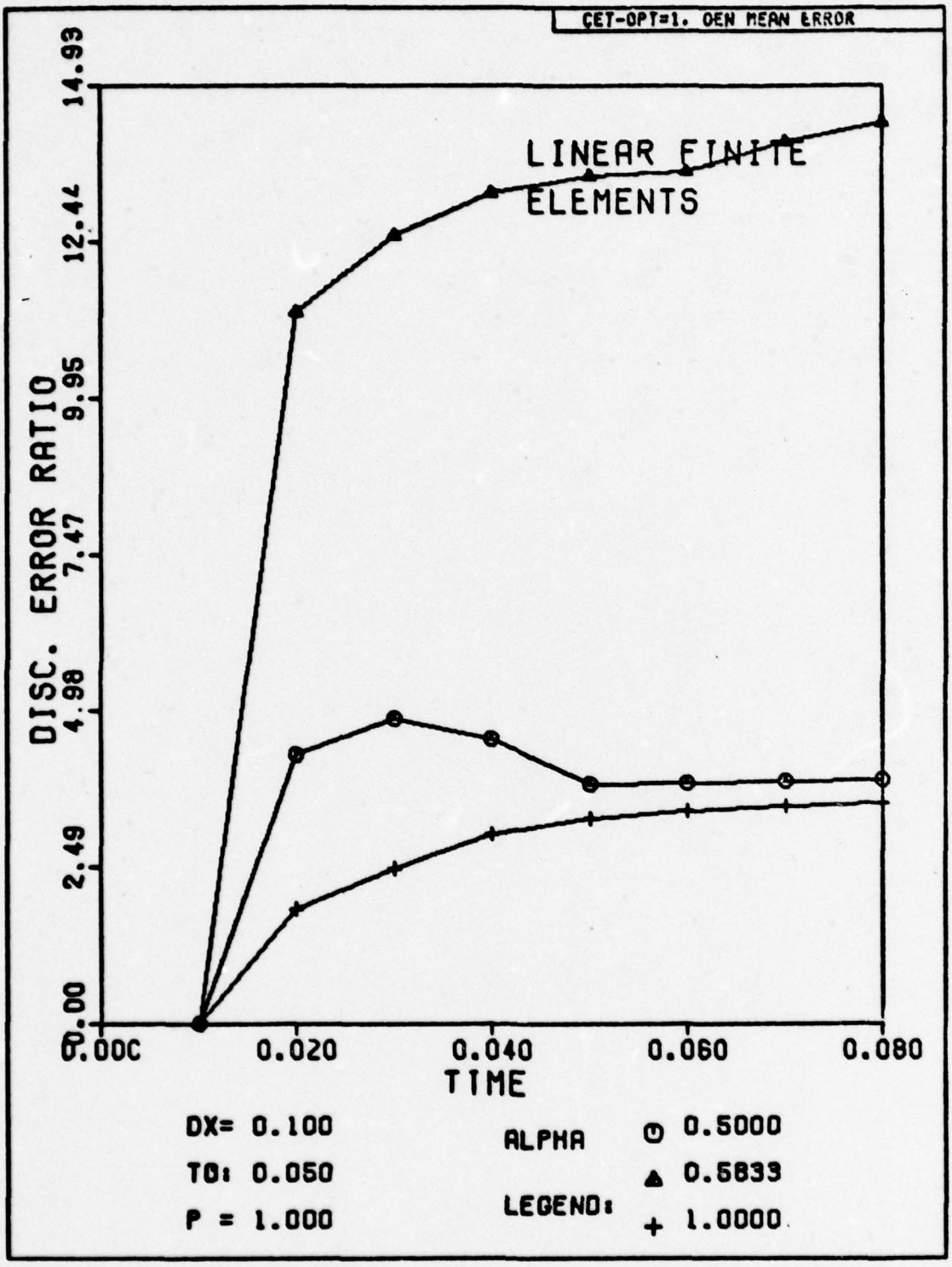


Fig. H-98. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

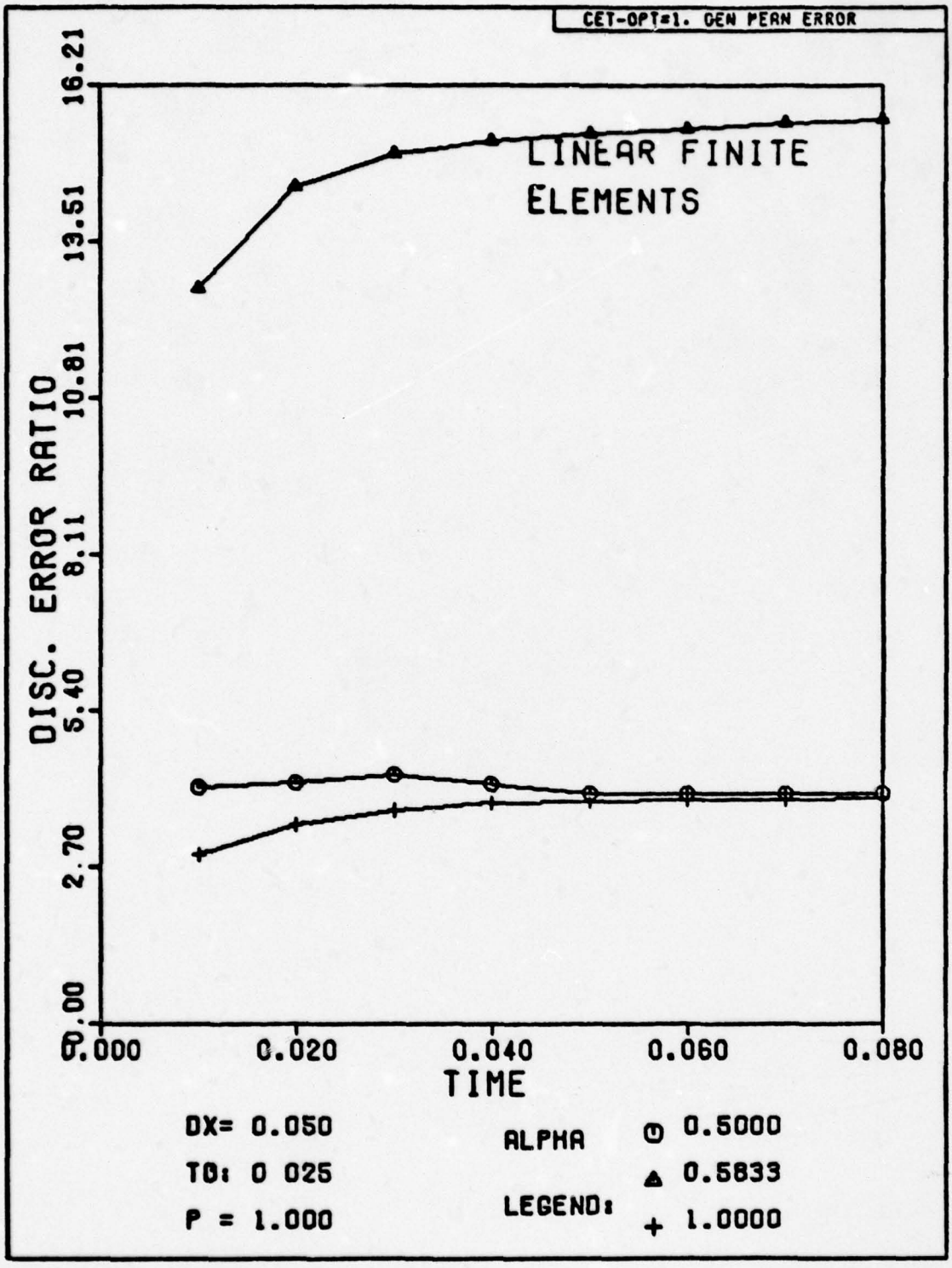


Fig. H-99. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

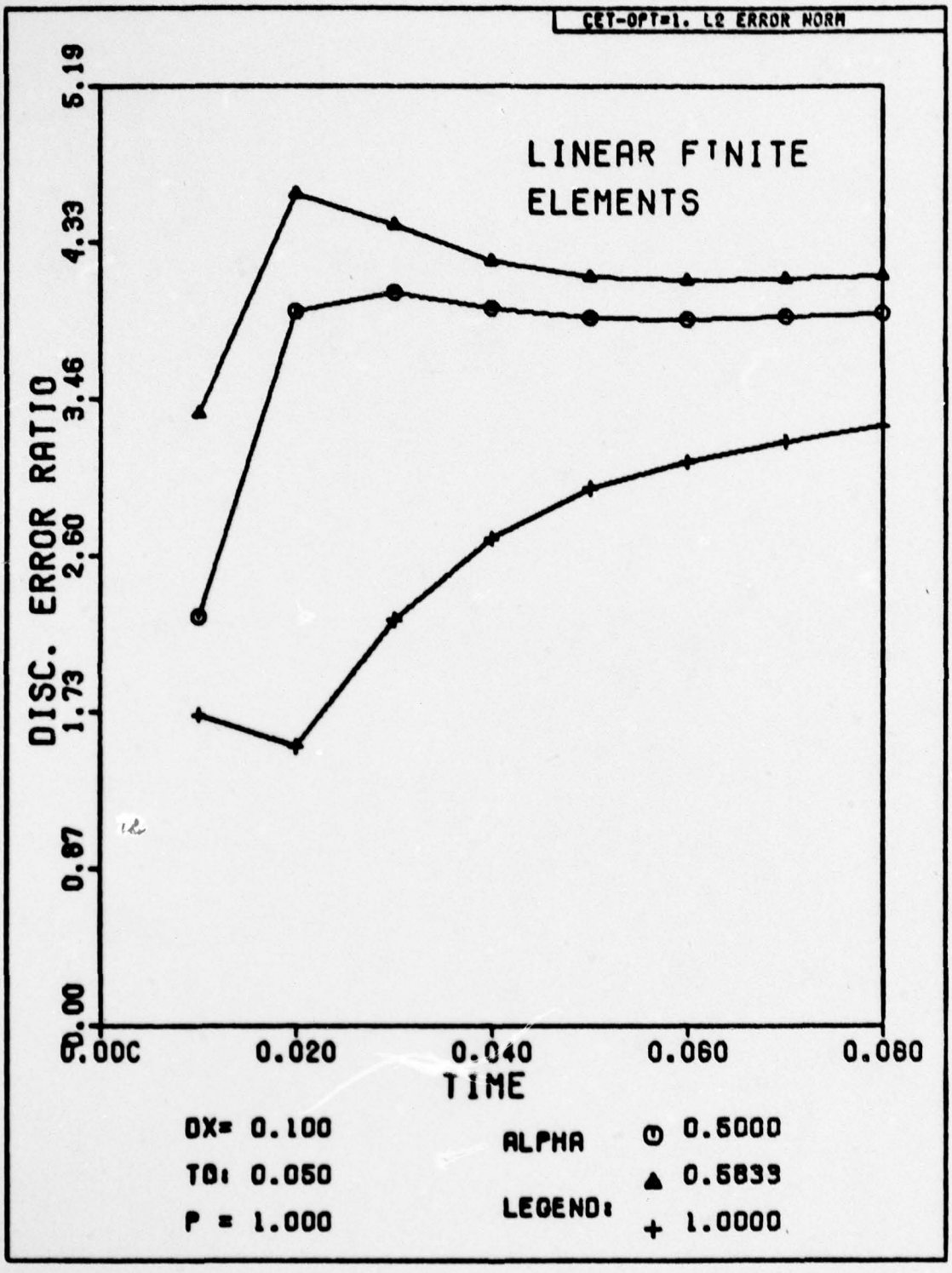


Fig. H-100. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

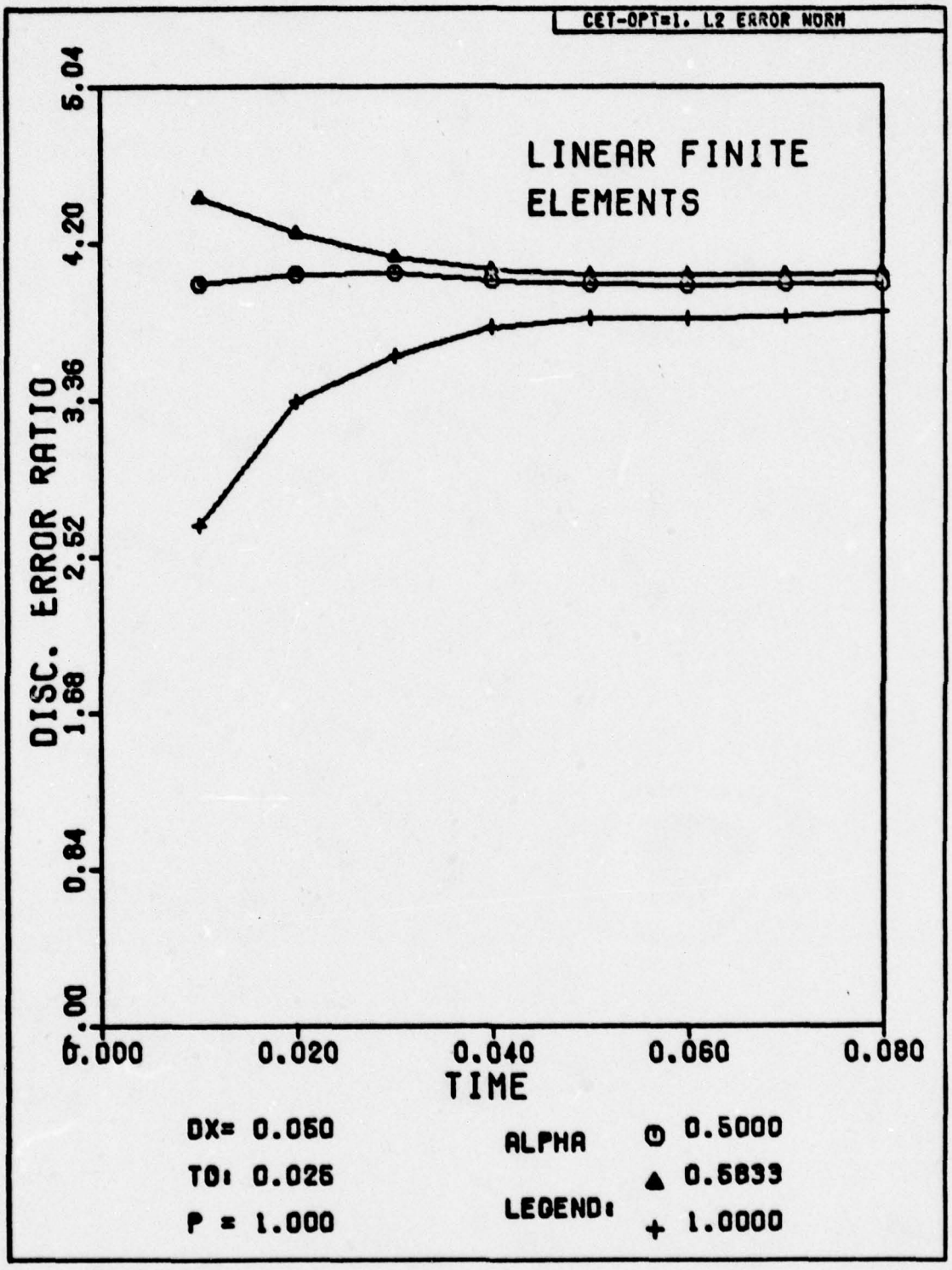


Fig. H-101. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

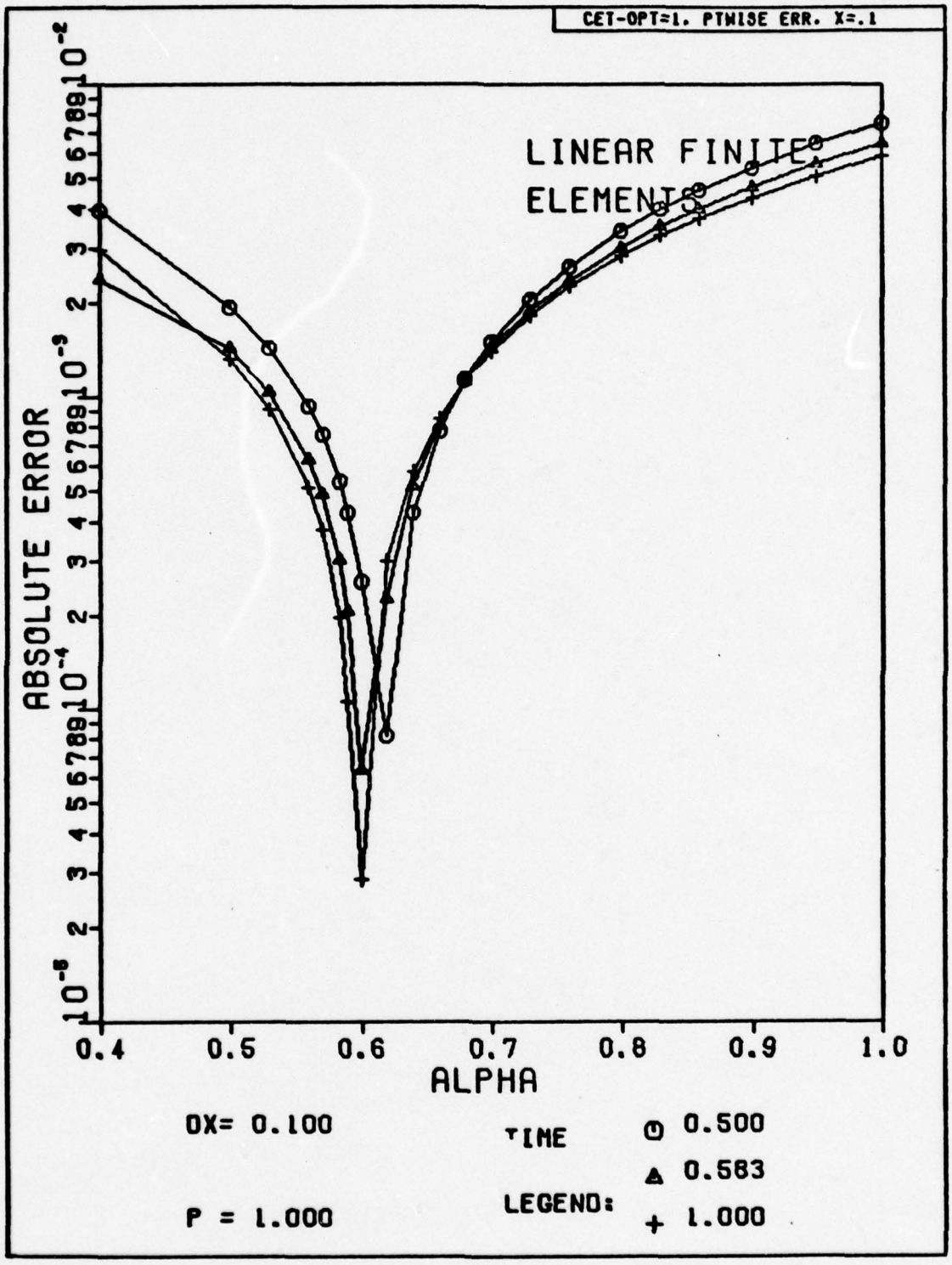


Fig. H-102. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

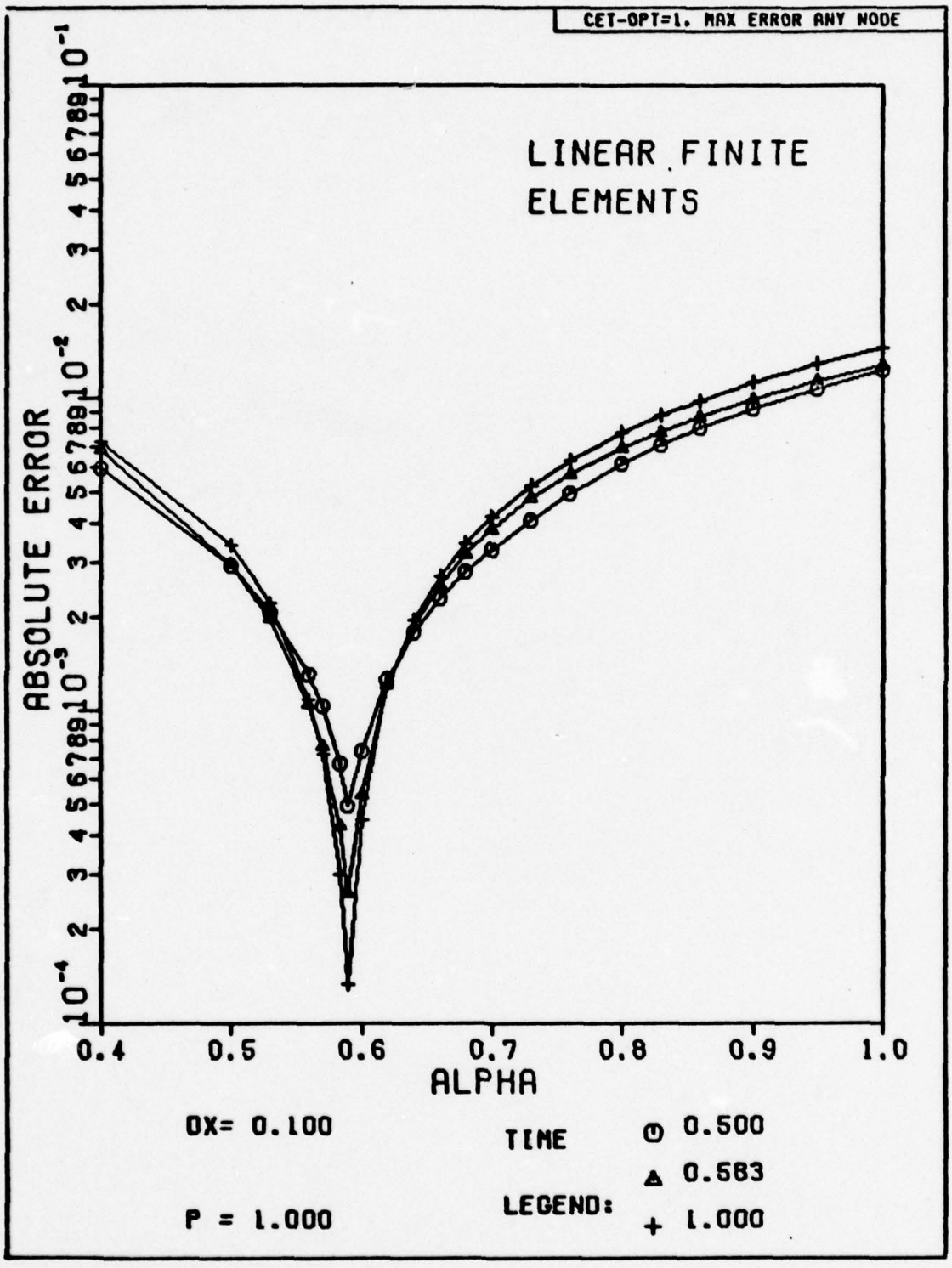


Fig. H-103. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

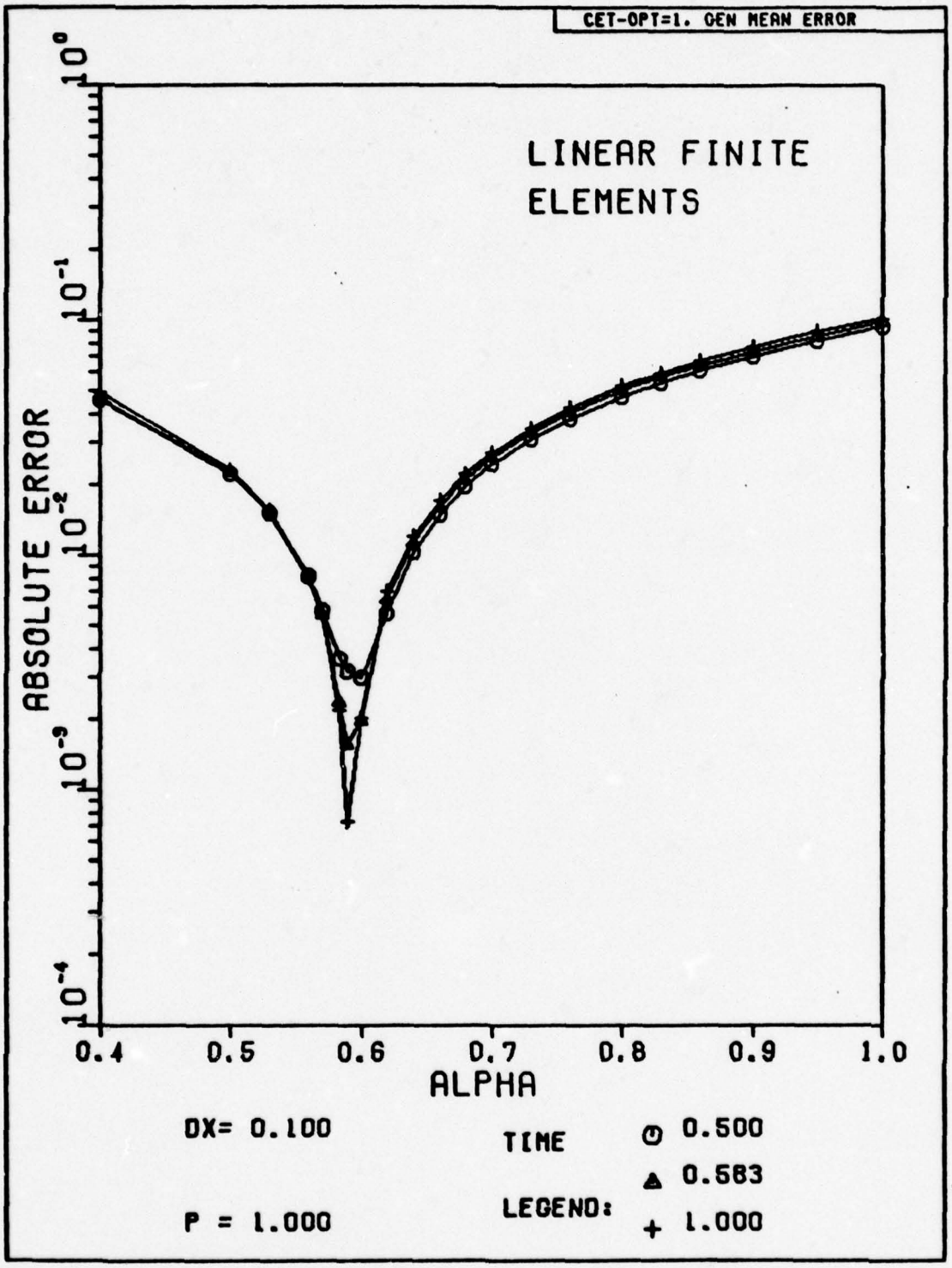


Fig. H-104. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

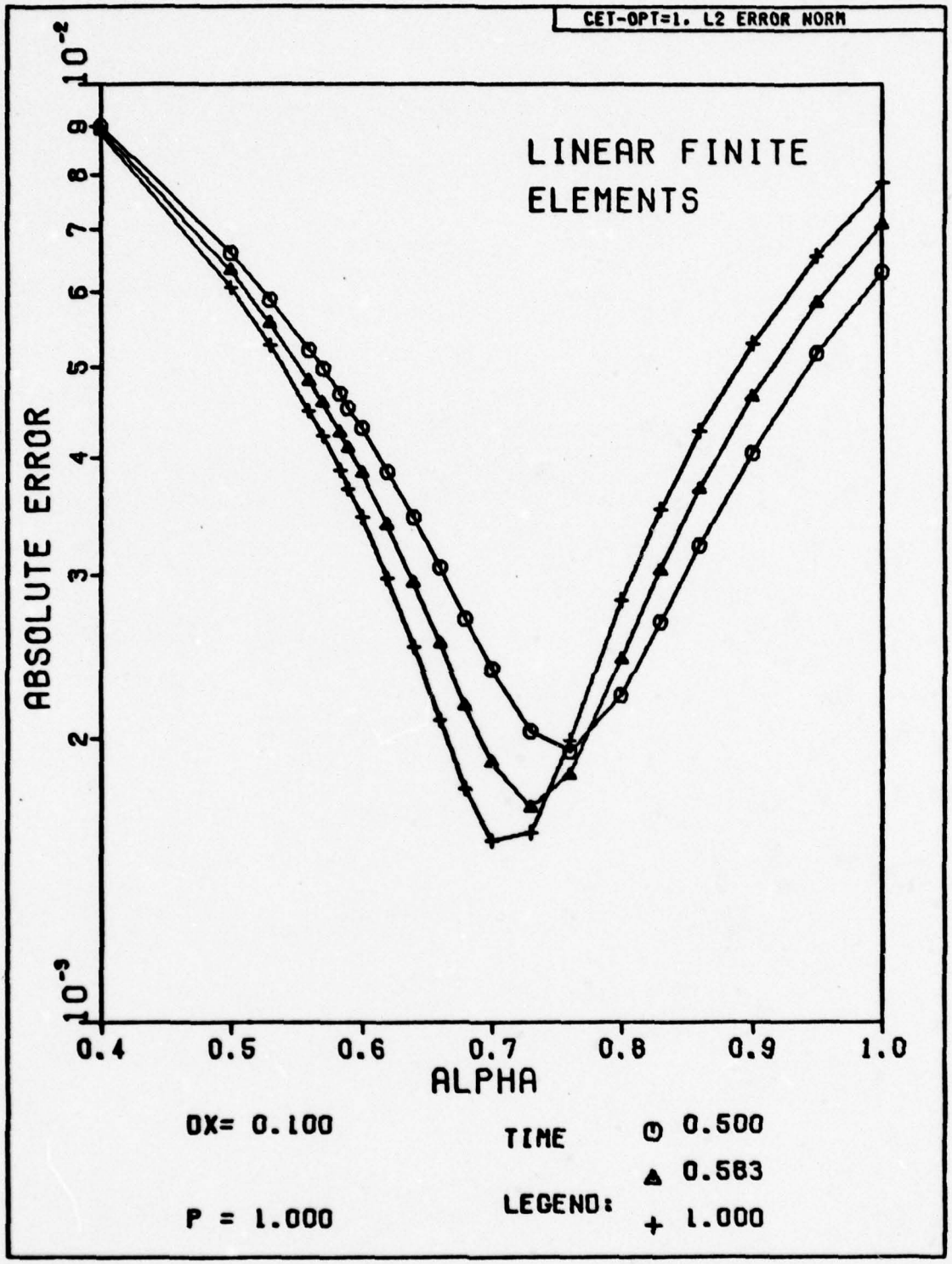


Fig. H-105. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

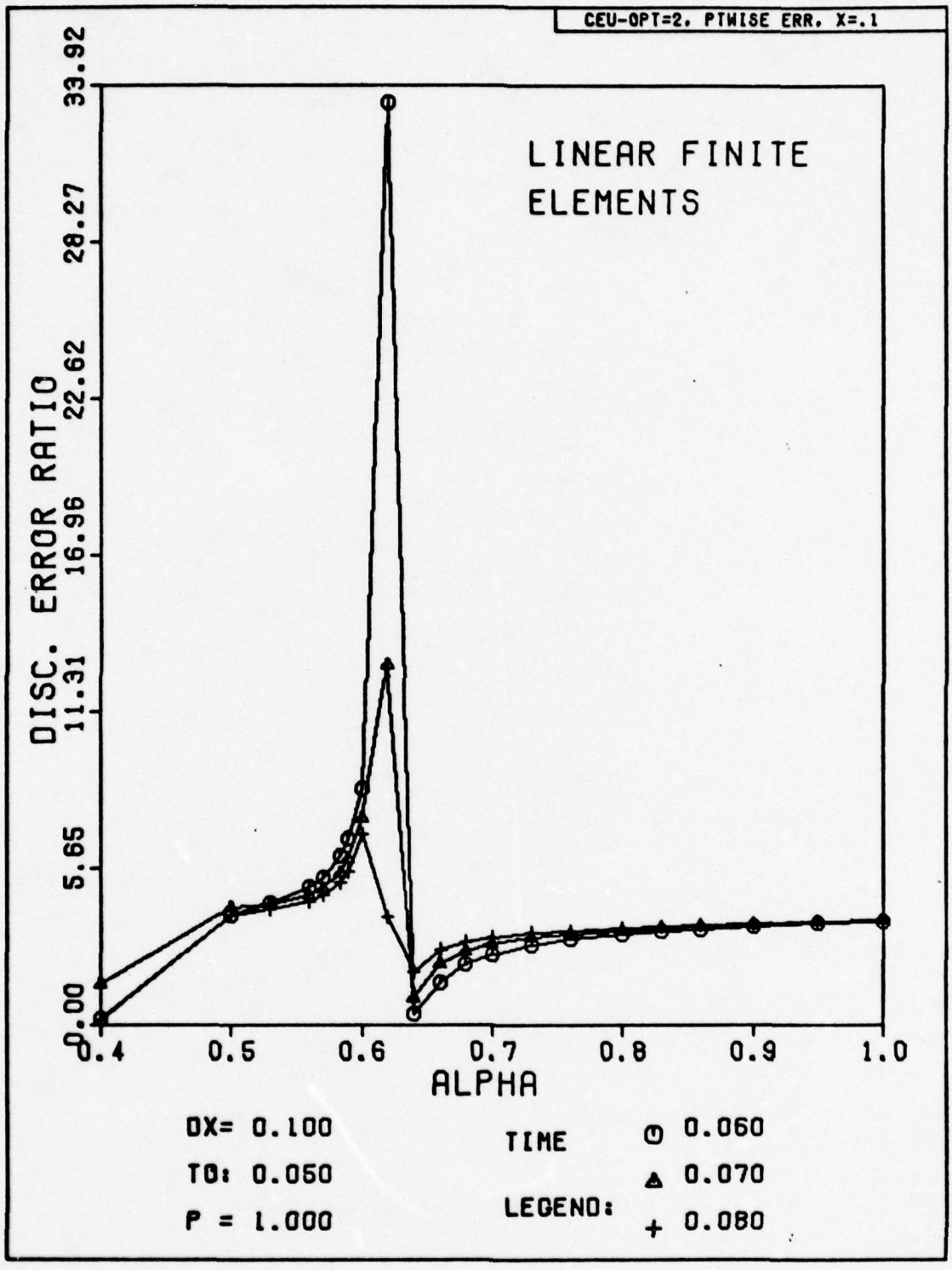


Fig. H-106. Discretization Error Ratio Versus Alpha for Problem One. The space-time grid has been subdivided for the first time step to obtain a more accurate solution.

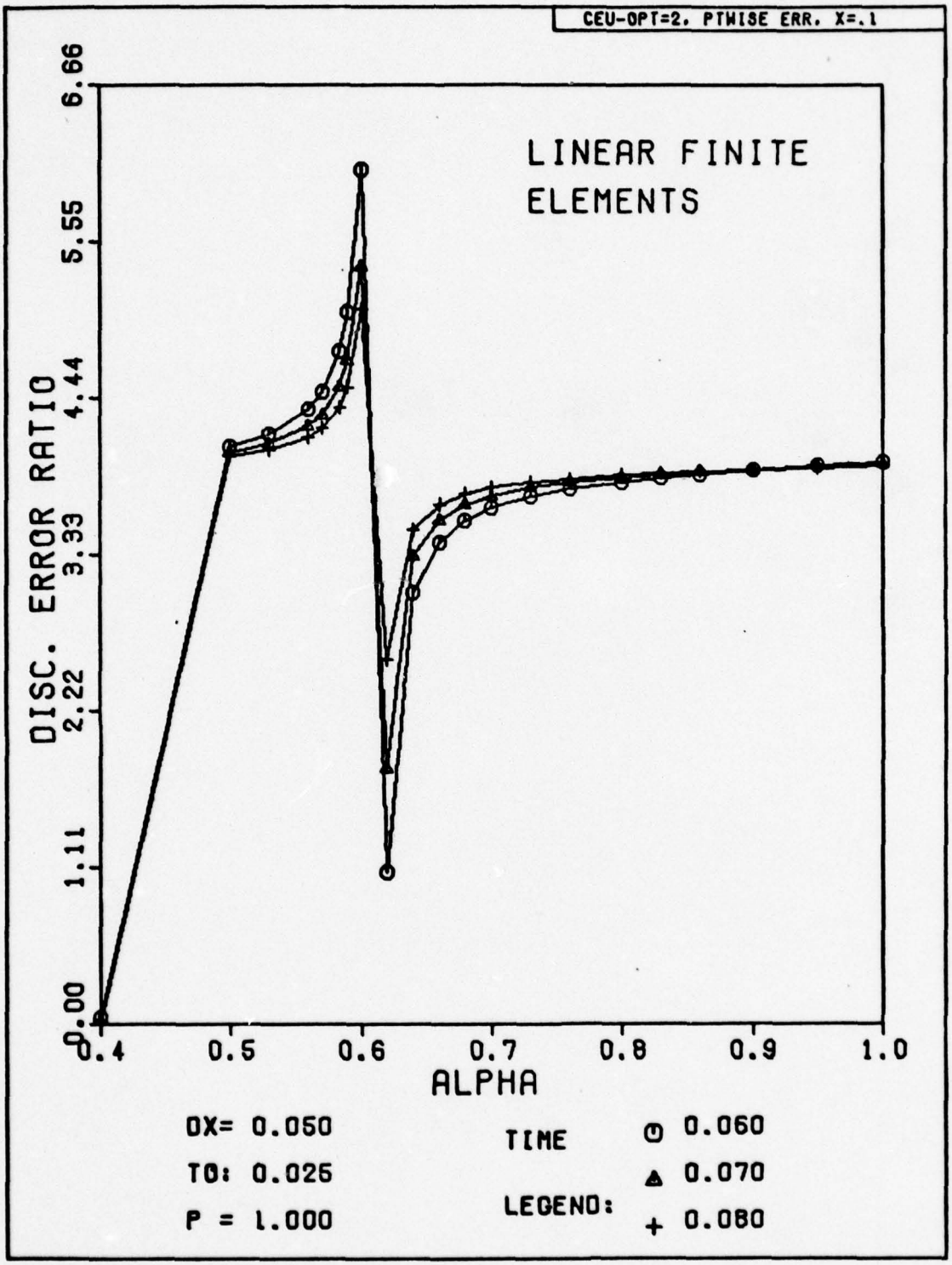


Fig. H-107. Discretization Error Ratio Versus Alpha for Problem One. The space-time grid has been subdivided for the first. time step to obtain a more accurate solution.

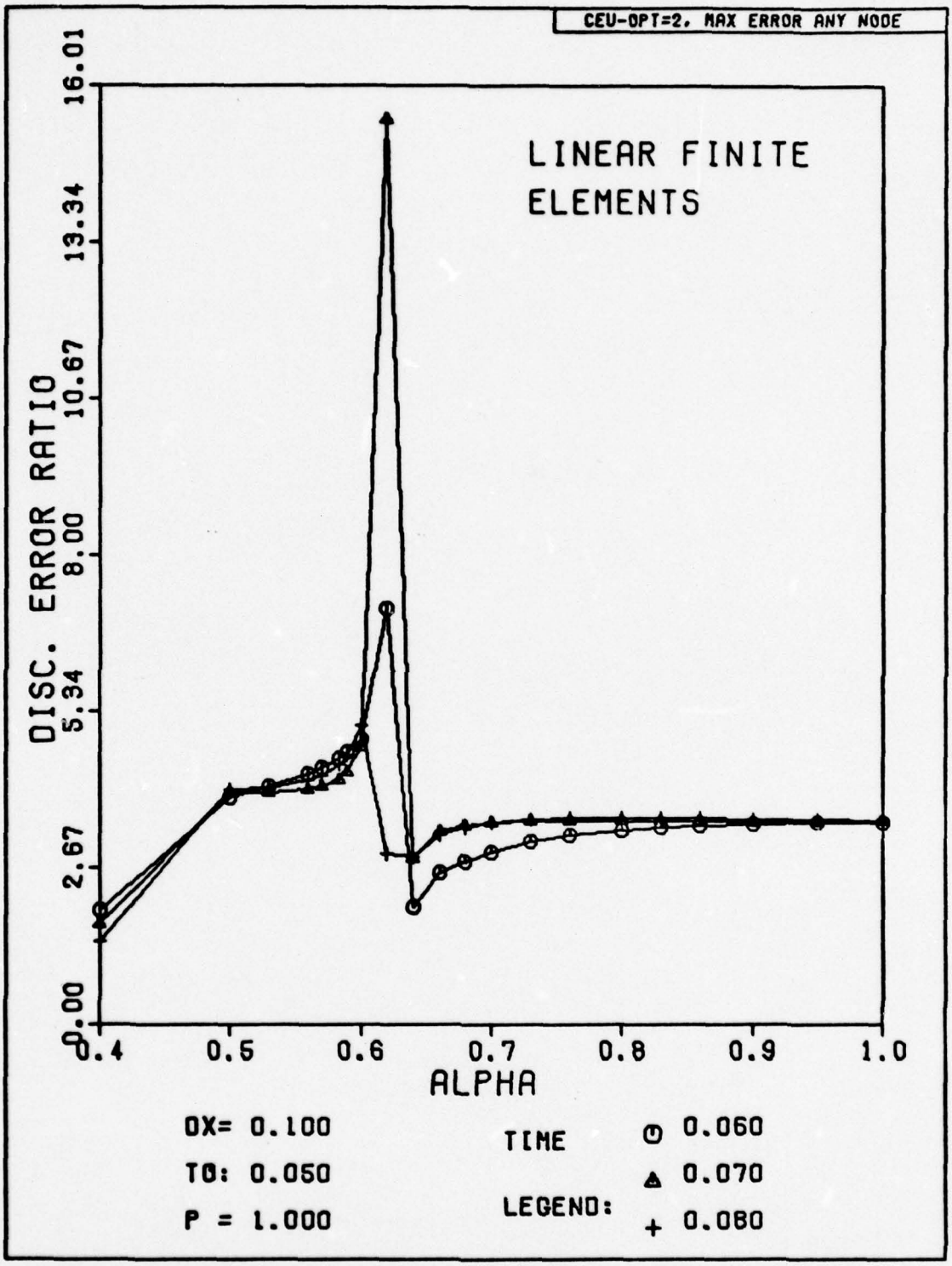


Fig. H-108. Discretization Error Ratio Versus Alpha for Problem One. The space-time grid has been subdivided for the first time step to obtain a more accurate solution.

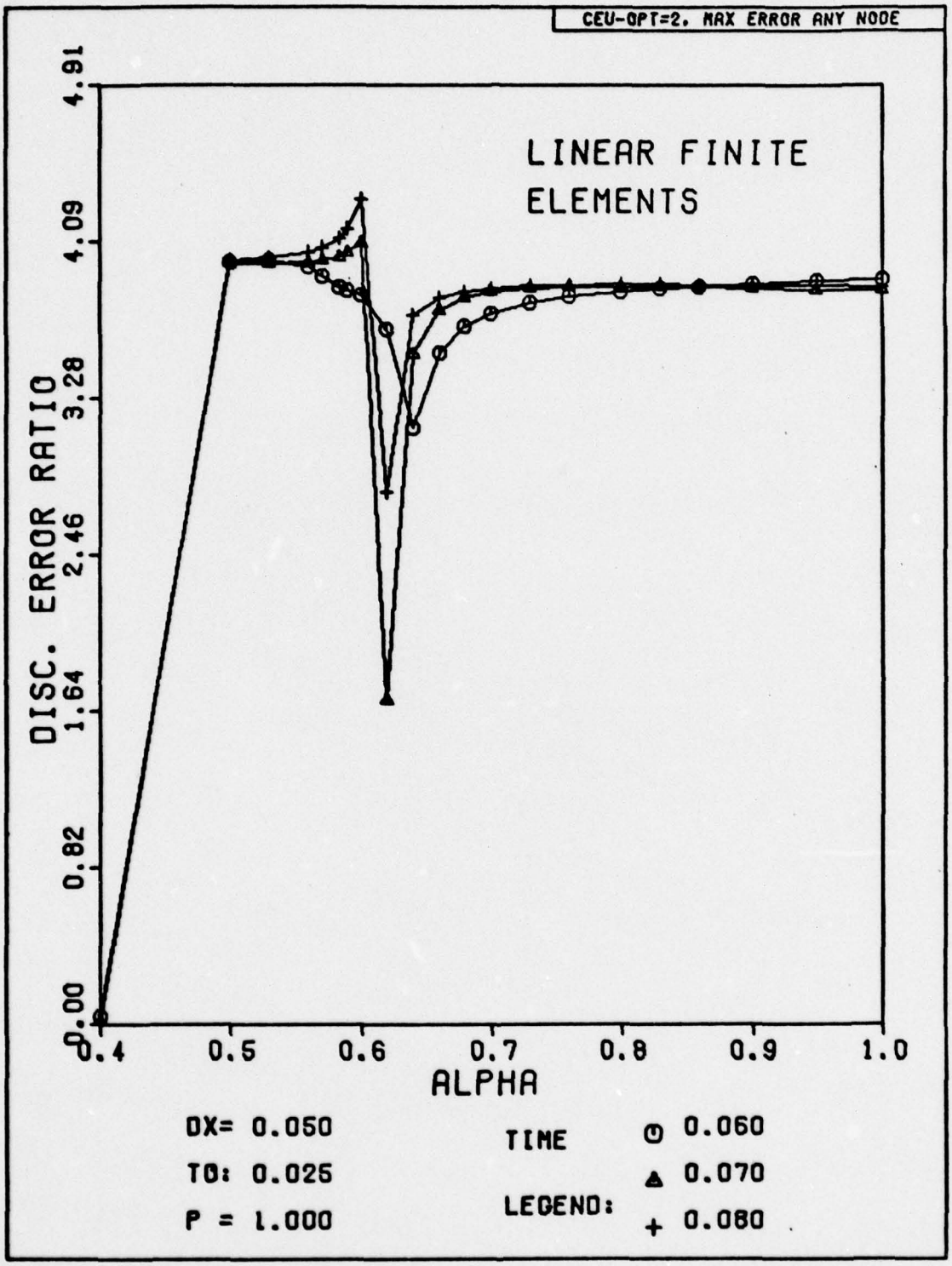


Fig. H-109. Discretization Error Ratio Versus Alpha for Problem One. The space-time grid has been subdivided for the first time step to obtain a more accurate solution.

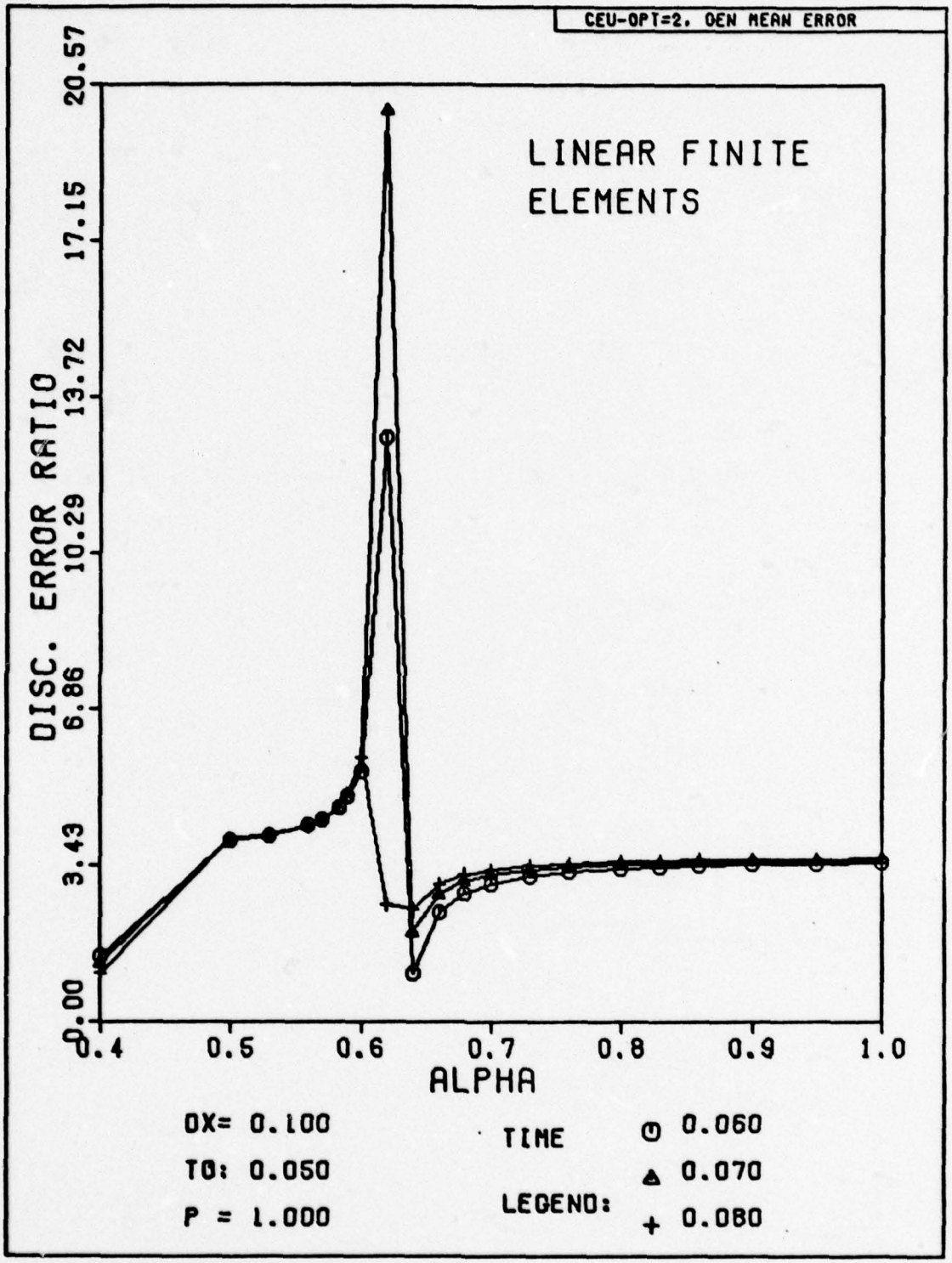


Fig. H-110. Discretization Error Ratio Versus Alpha for Problem One. The space-time grid has been subdivided for the first time step to obtain a more accurate solution.

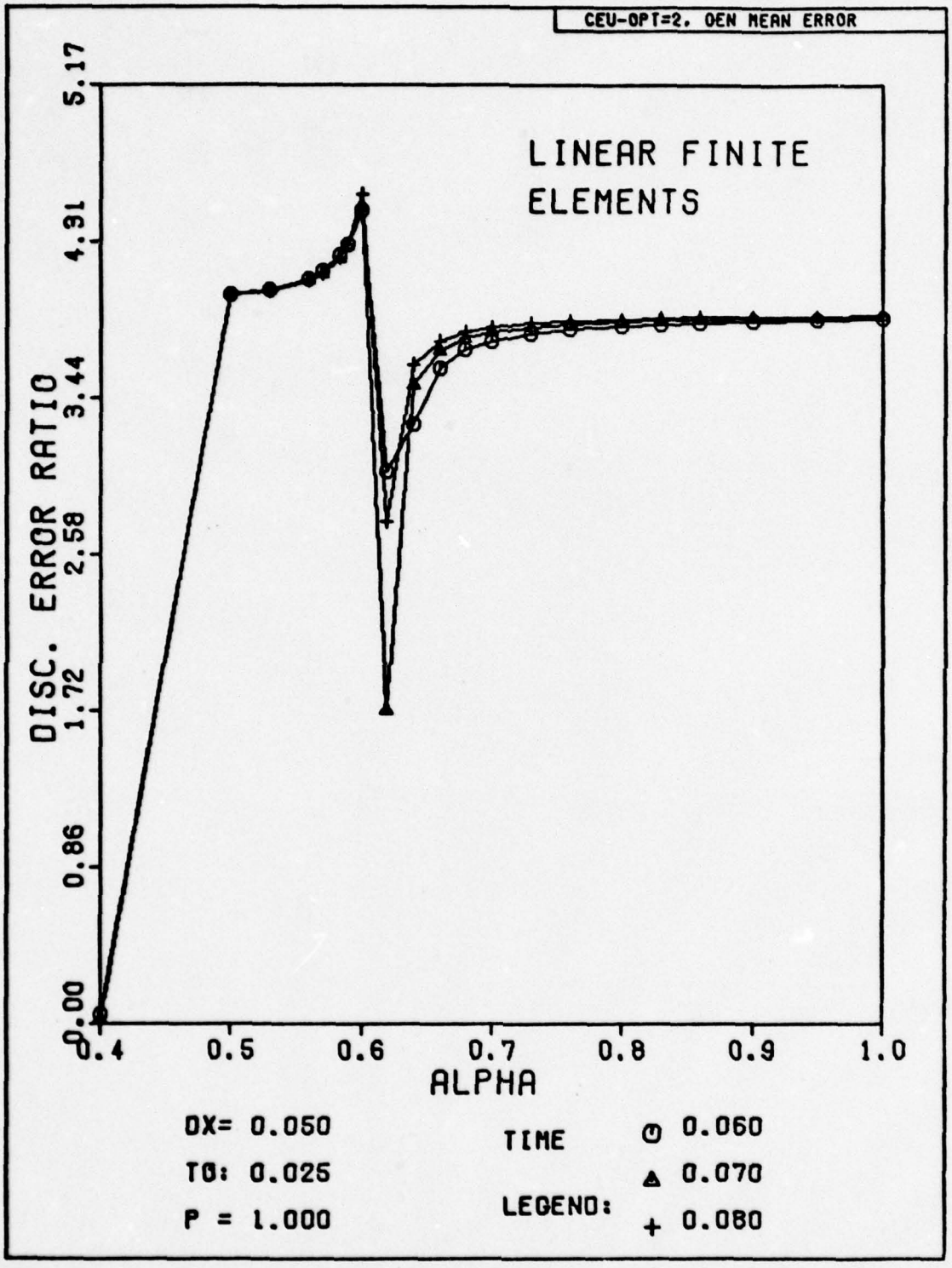


Fig. H-111. Discretization Error Ratio Versus Alpha for Problem One. The space-time grid has been subdivided for the first time step to obtain a more accurate solution.

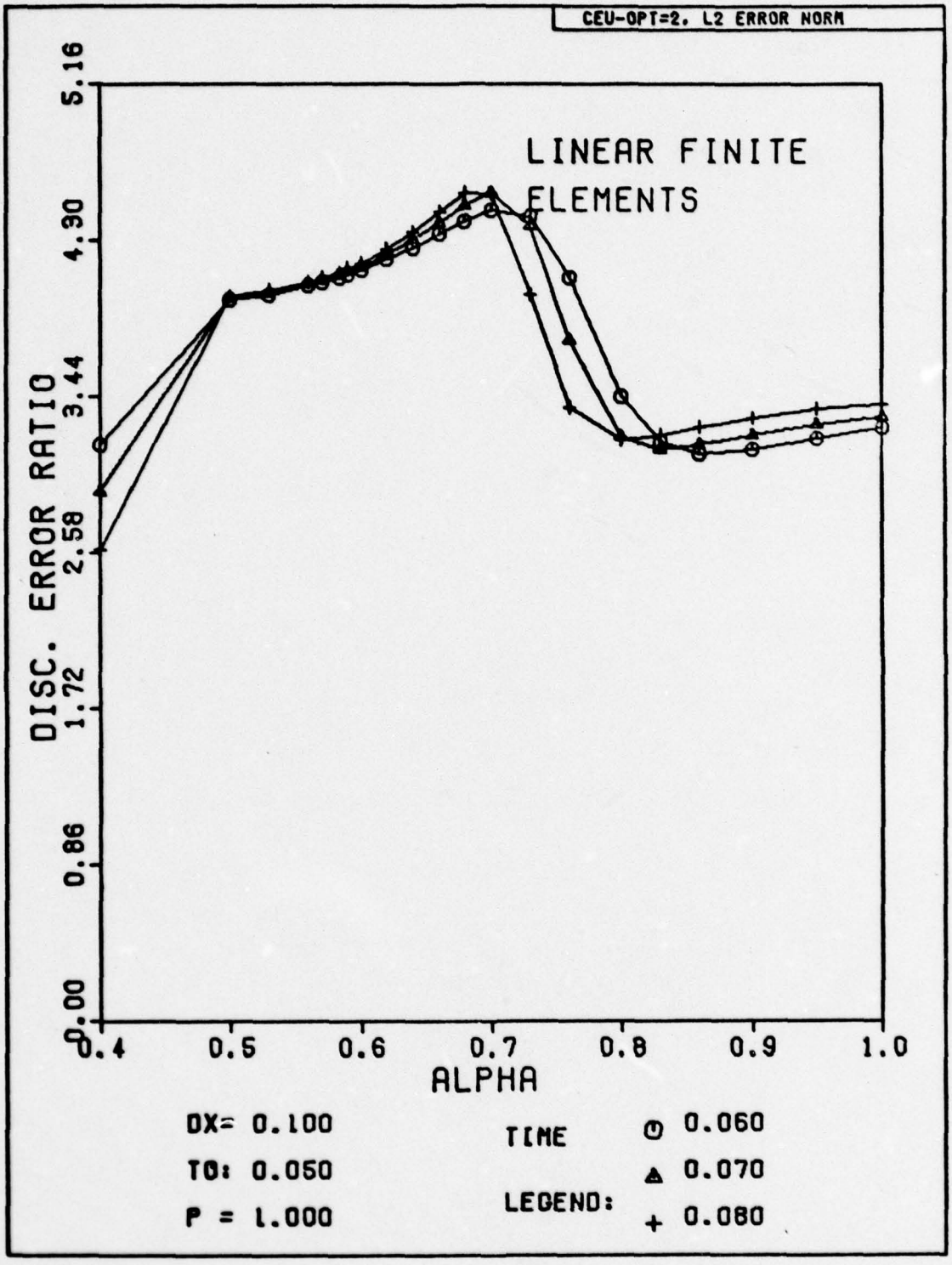


Fig. H-112. Discretization Error Ratio Versus Alpha for Problem One. The space-time grid has been subdivided for the first time to obtain a more accurate solution.

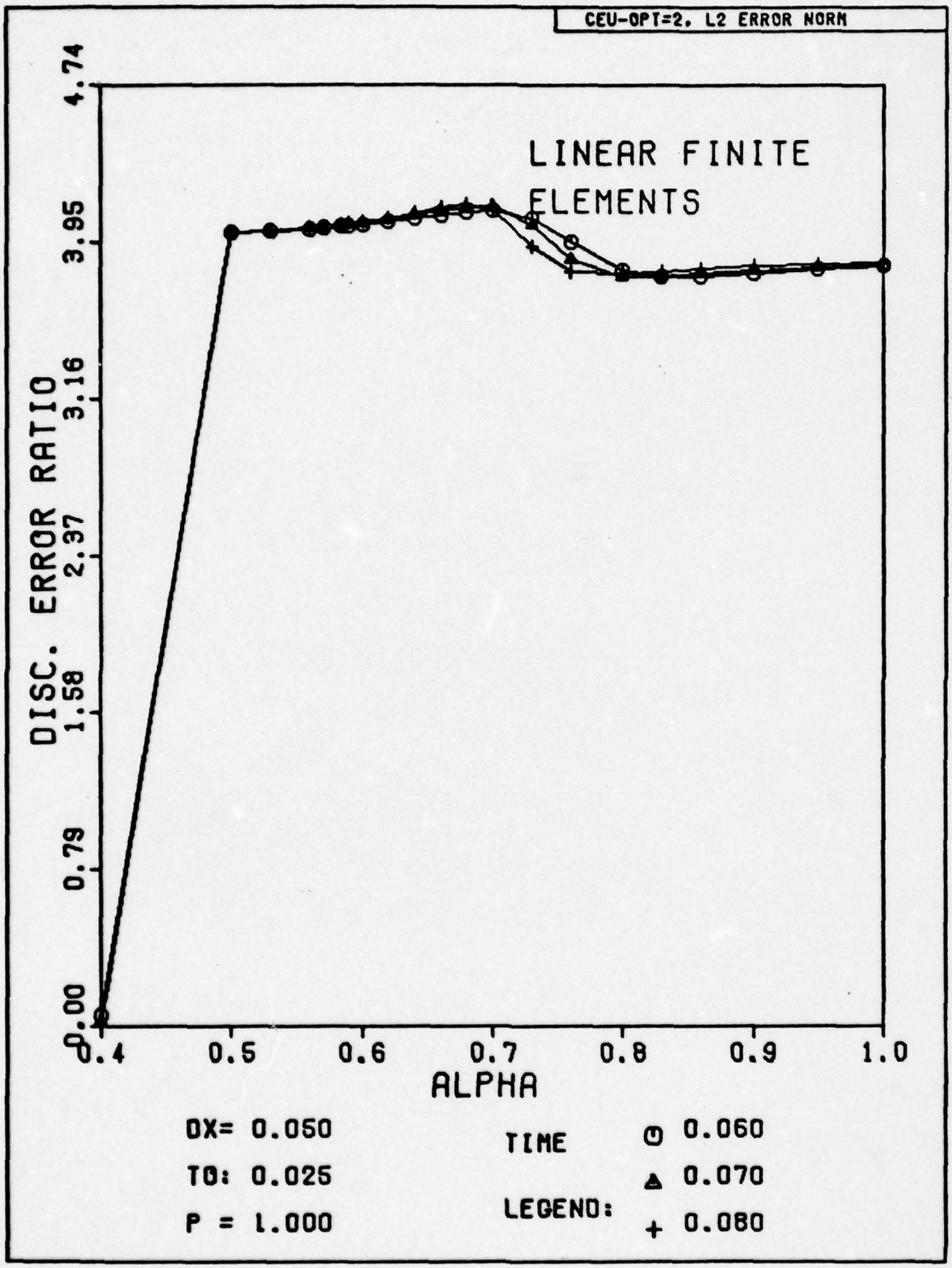


Fig. H-113. Discretization Error Ratio Versus Alpha for Problem One. The space-time grid has been subdivided for the first time step to obtain a more accurate solution.

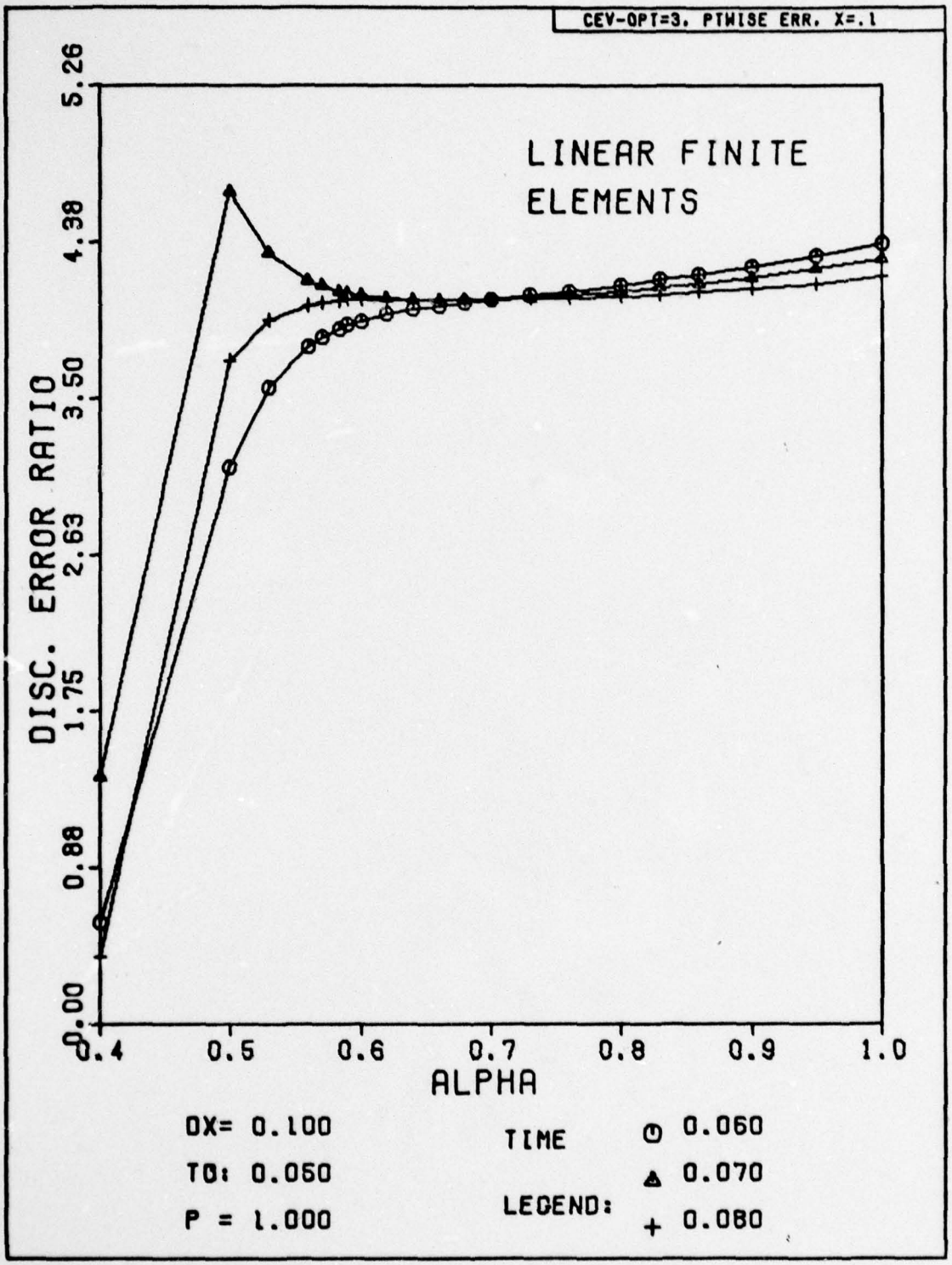


Fig. H-114. Discretization Error Ratio Versus Alpha for Problem One. The initial conditions have been modified by the technique used to generate Figs. 15 and 16.

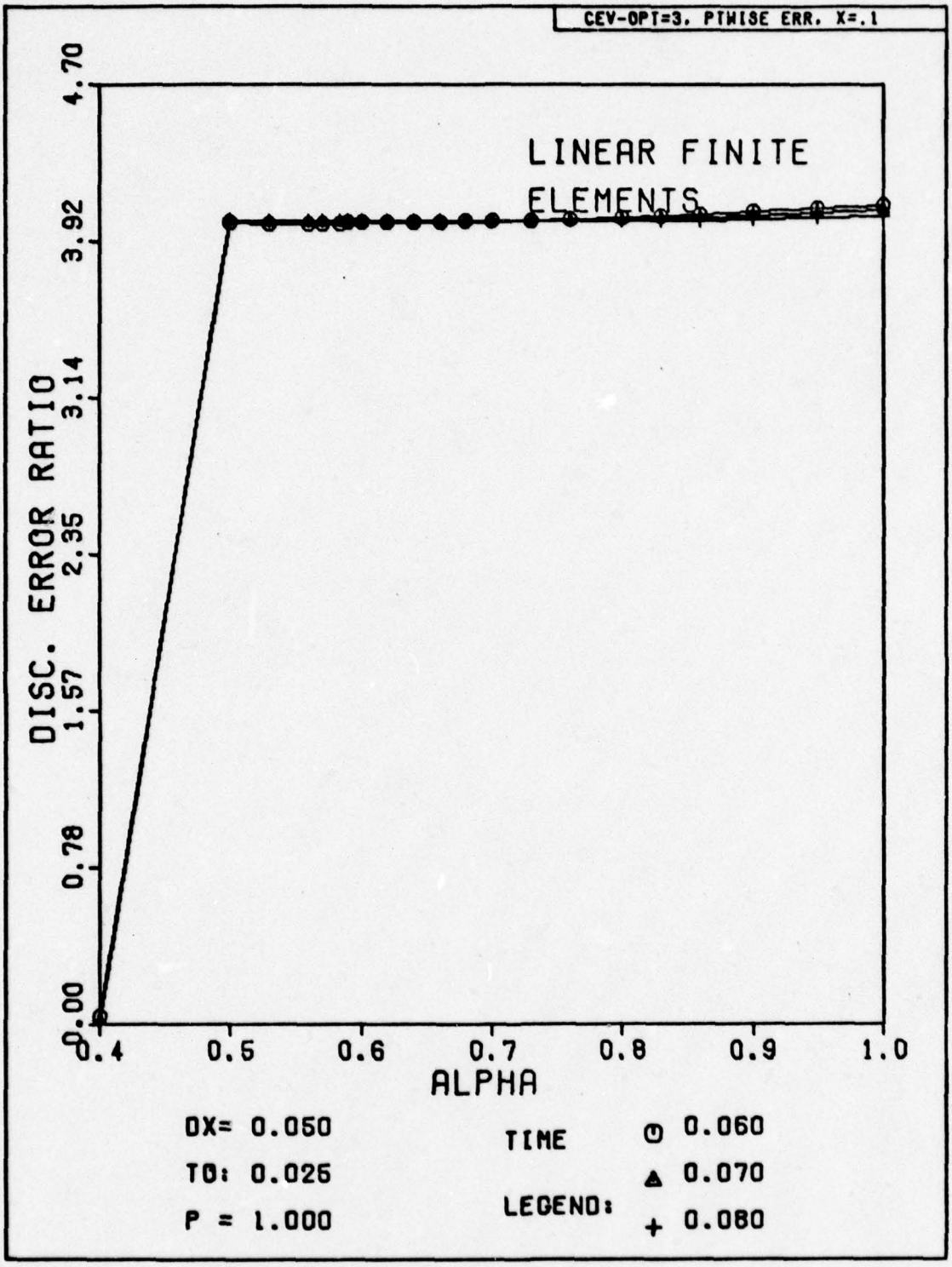


Fig. H-115. Discretization Error Ratio Versus Alpha for Problem One. The initial conditions have been modified by the technique used to generate Figs. 15 and 16.

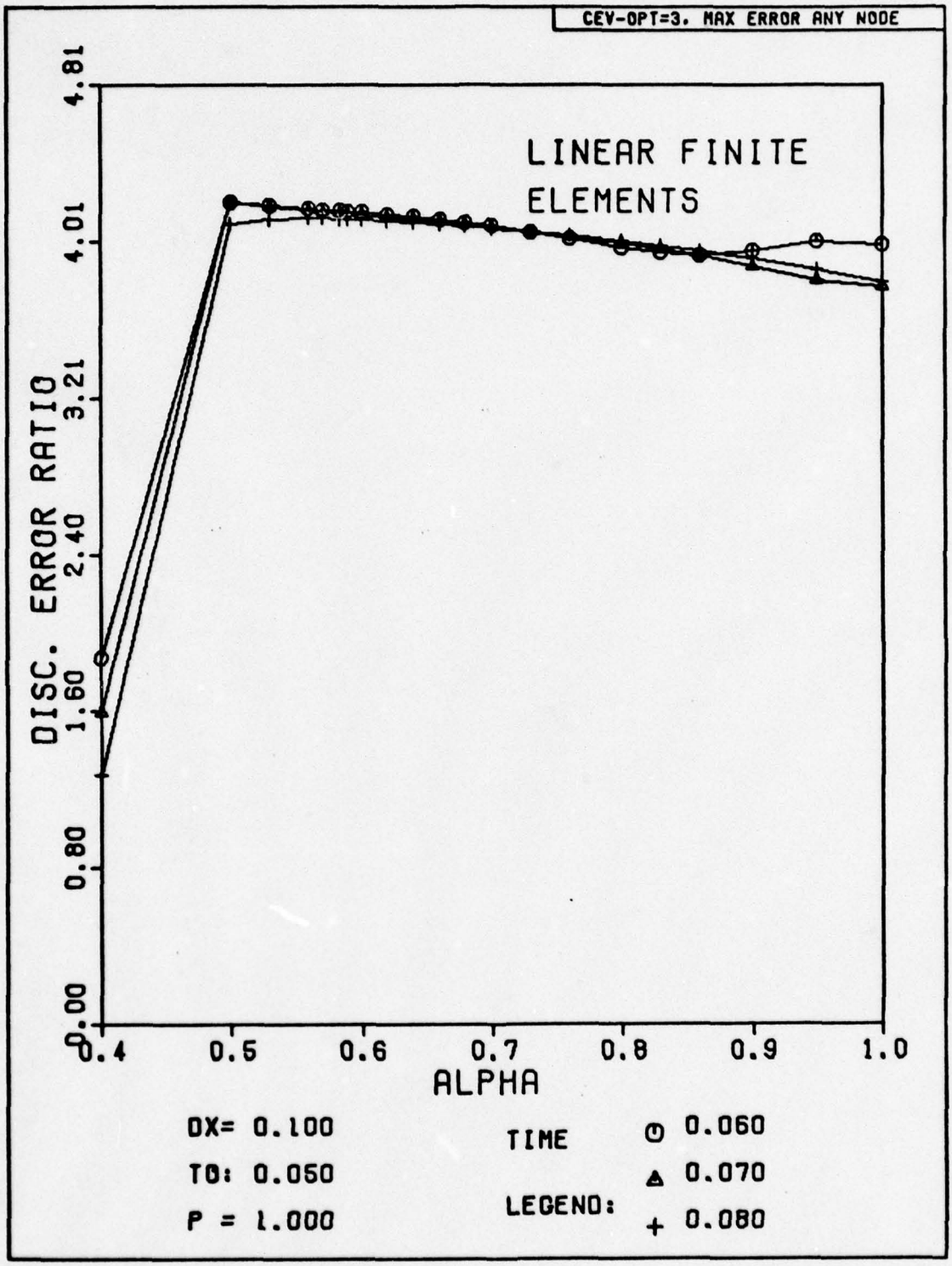


Fig. H-116. Discretization Error Ratio Versus Alpha for Problem One. The initial conditions have been modified by the technique used to generate Figs. 15 and 16.

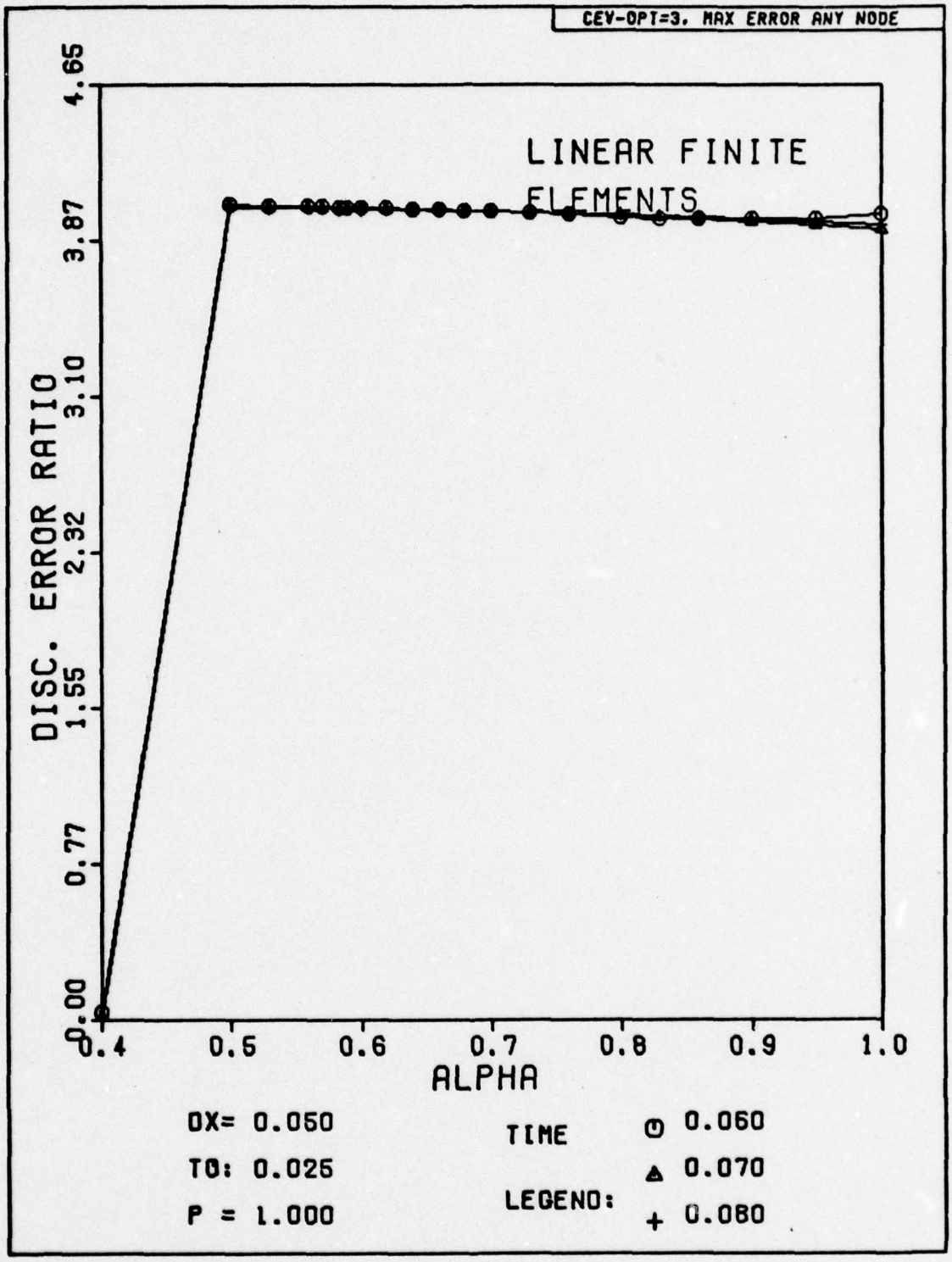


Fig. H-117. Discretization Error Ratio Versus Alpha for Problem One. The initial conditions ahve been modified by the technique used to generate Figs. 15 and 16.

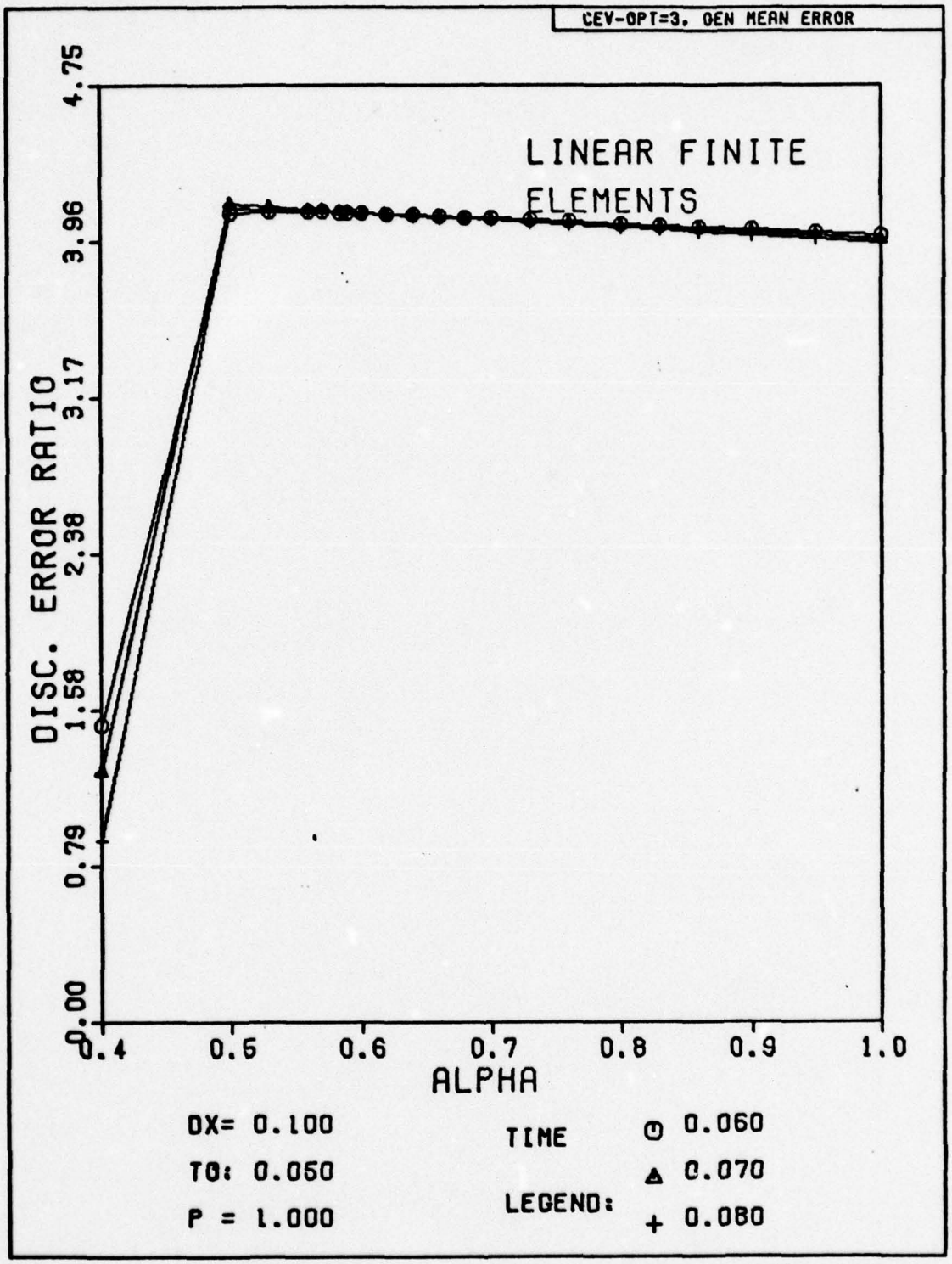


Fig. H-118. Discretization Error Ratio Versus Alpha for Problem One. The initial conditions have been modified by the technique used to generate Figs. 15 and 16.

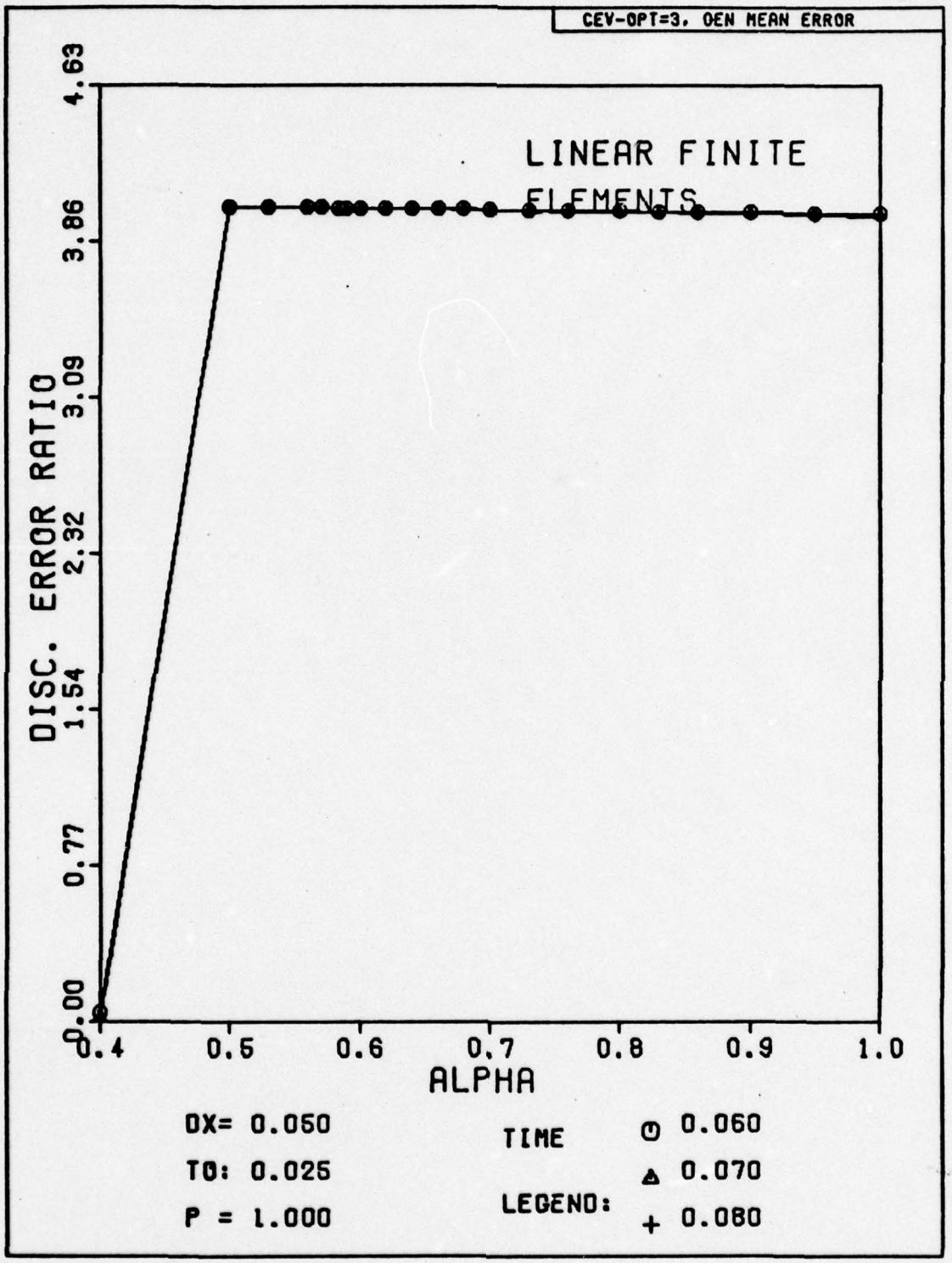


Fig. H-119. Discretization Error Ratio Versus Alpha for Problem One. The initial conditions have been modified by the technique used to generate Figs. 15 and 16.

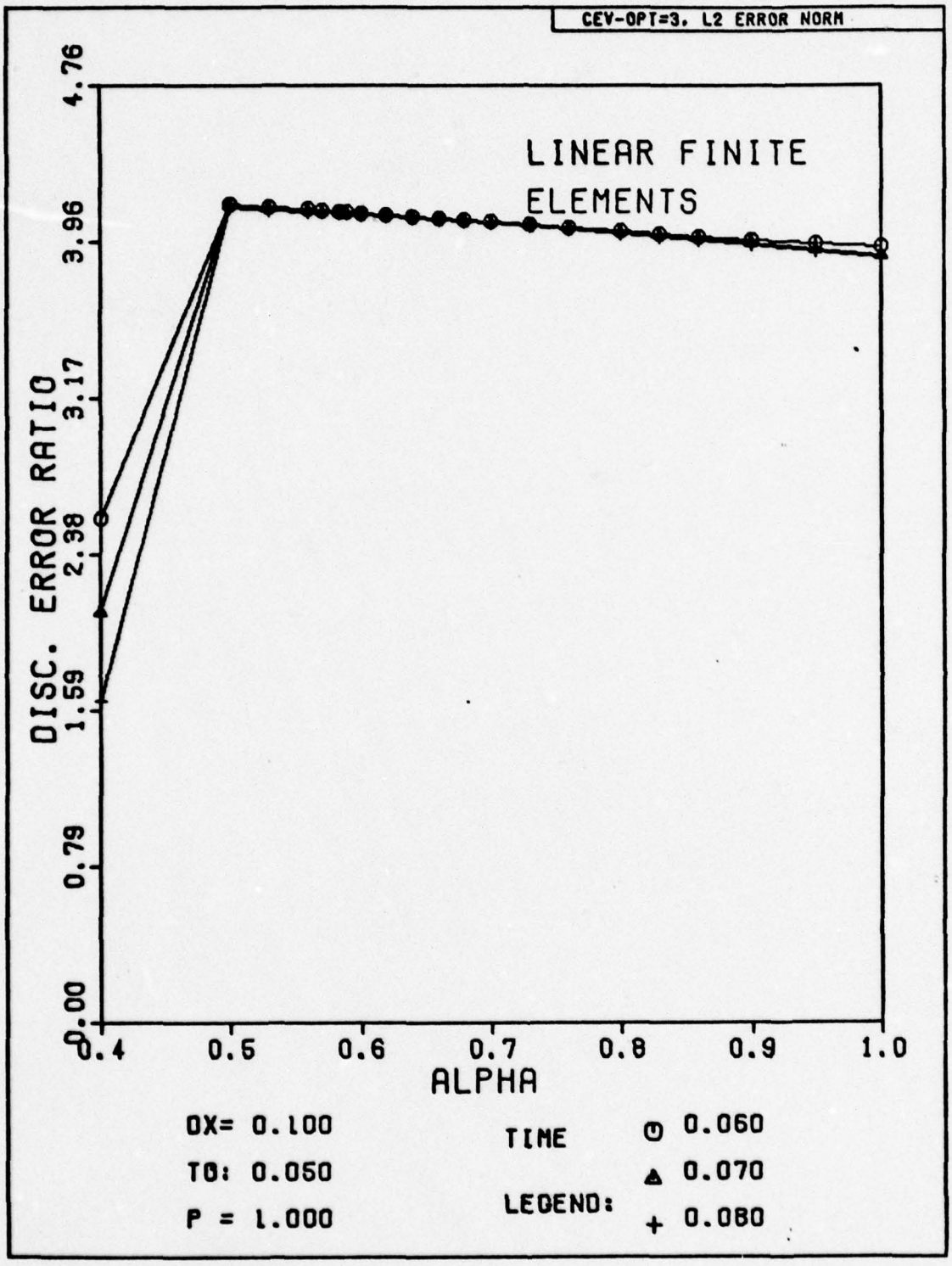


Fig. H-120. Discretization Error Ratio Versus Alpha for Problem One. The initial conditions have been modified by the technique used to generate Figs. 15 and 16.

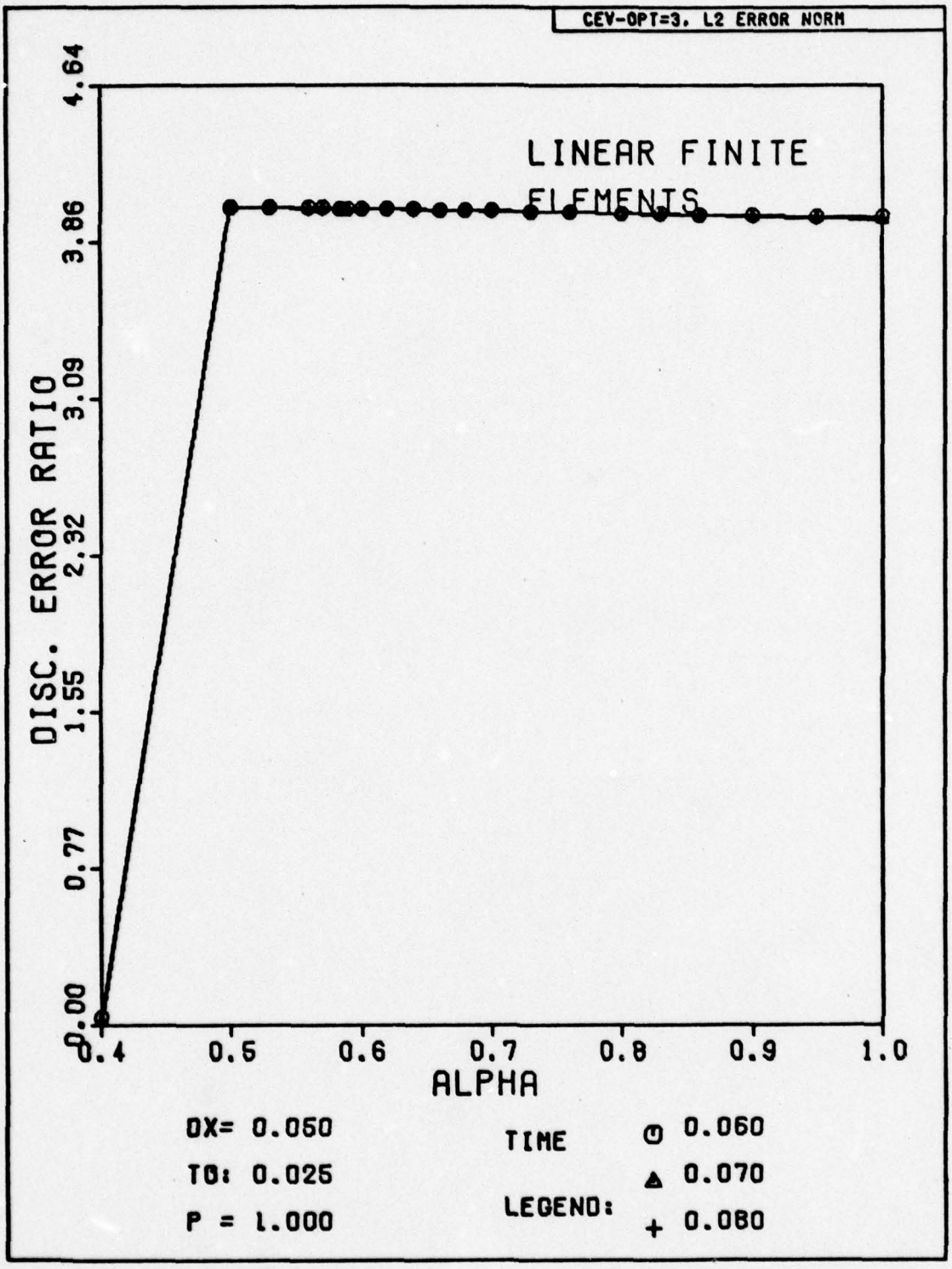


Fig. H-121. Discretization Error Ratio Versus Alpha for Problem One. The initial condition has been modified by the technique used to generate Figs. 15 and 16.

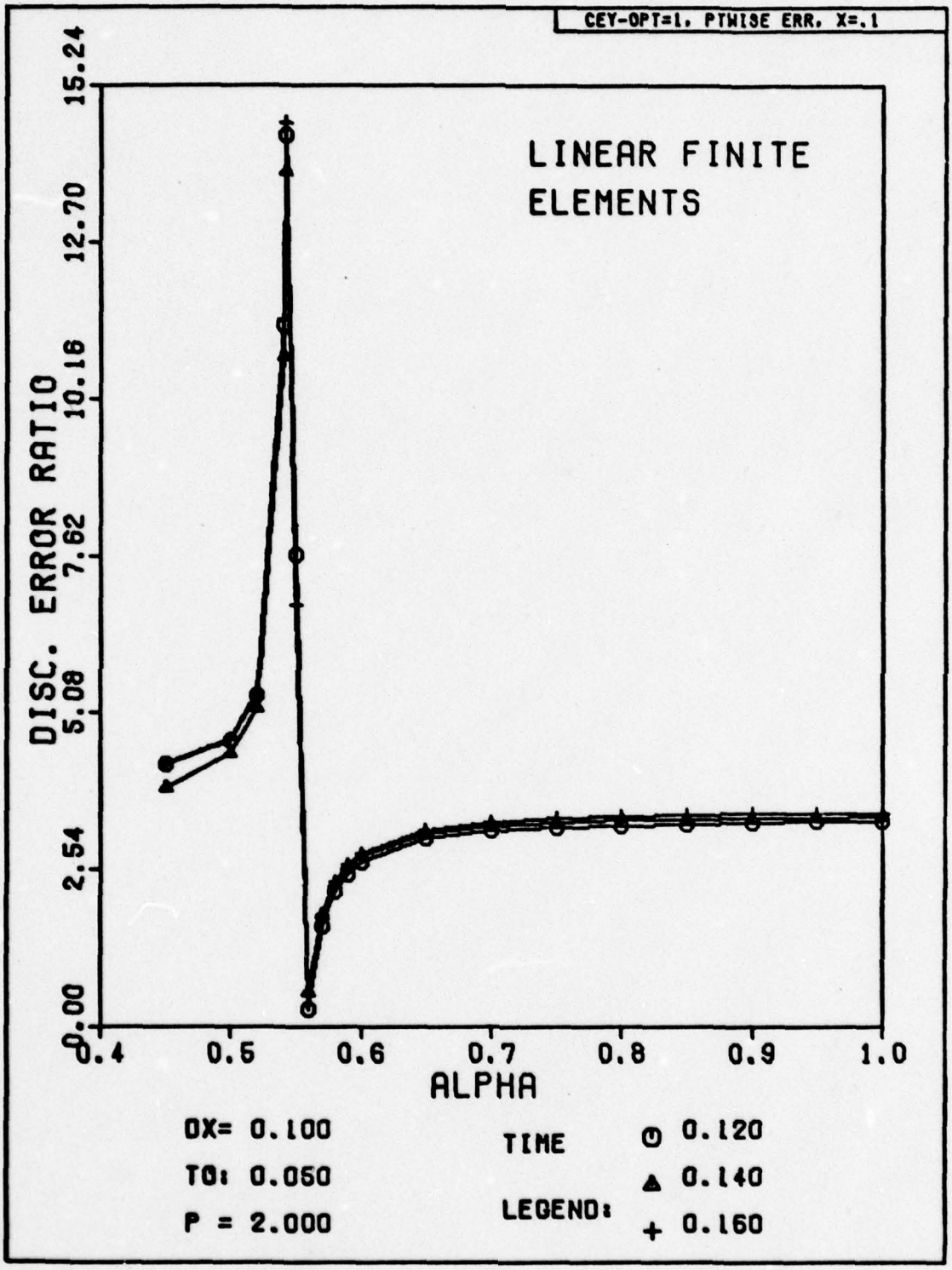


Fig. H-122. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

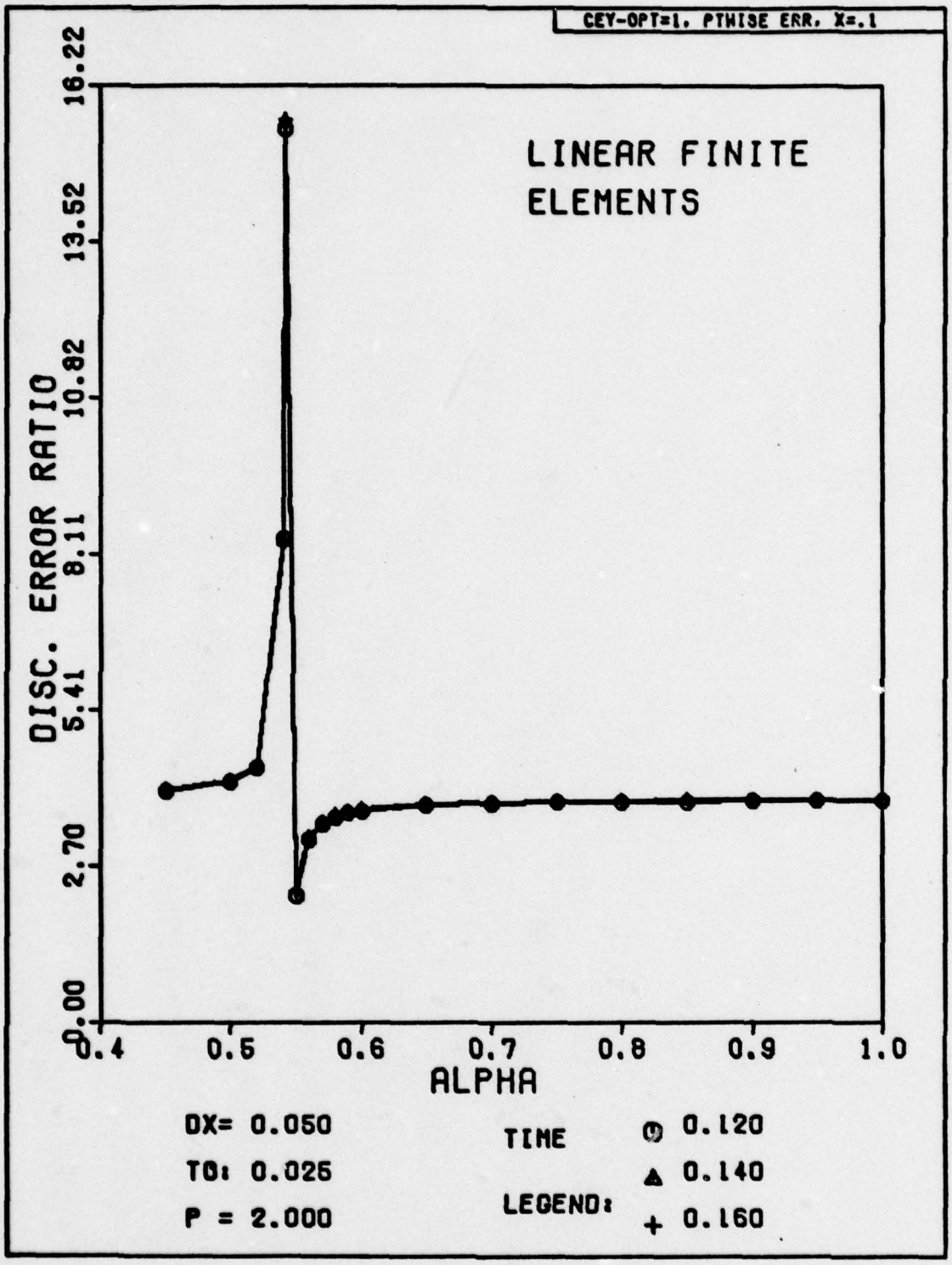


Fig. H-123. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

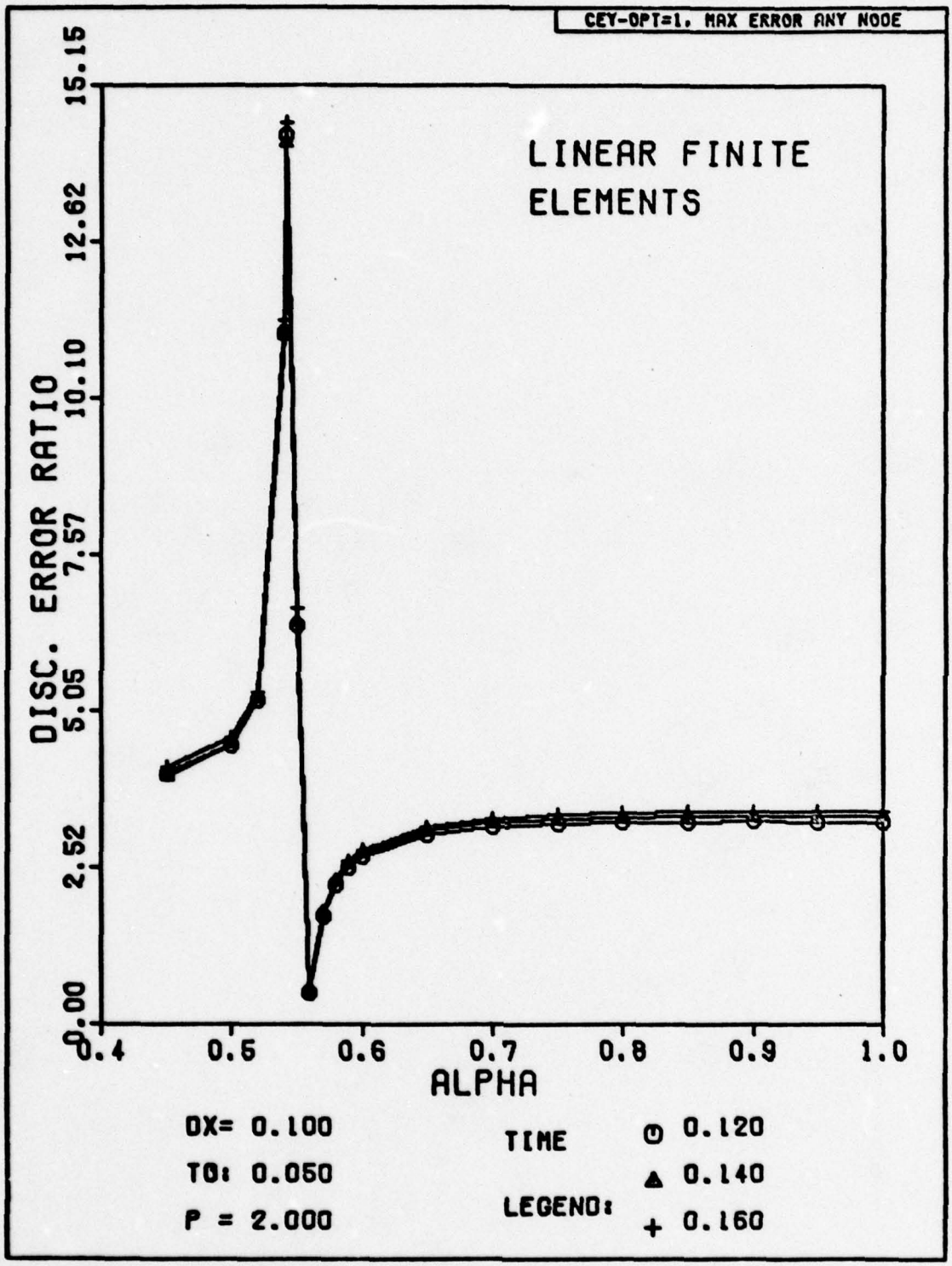


Fig. H-124. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

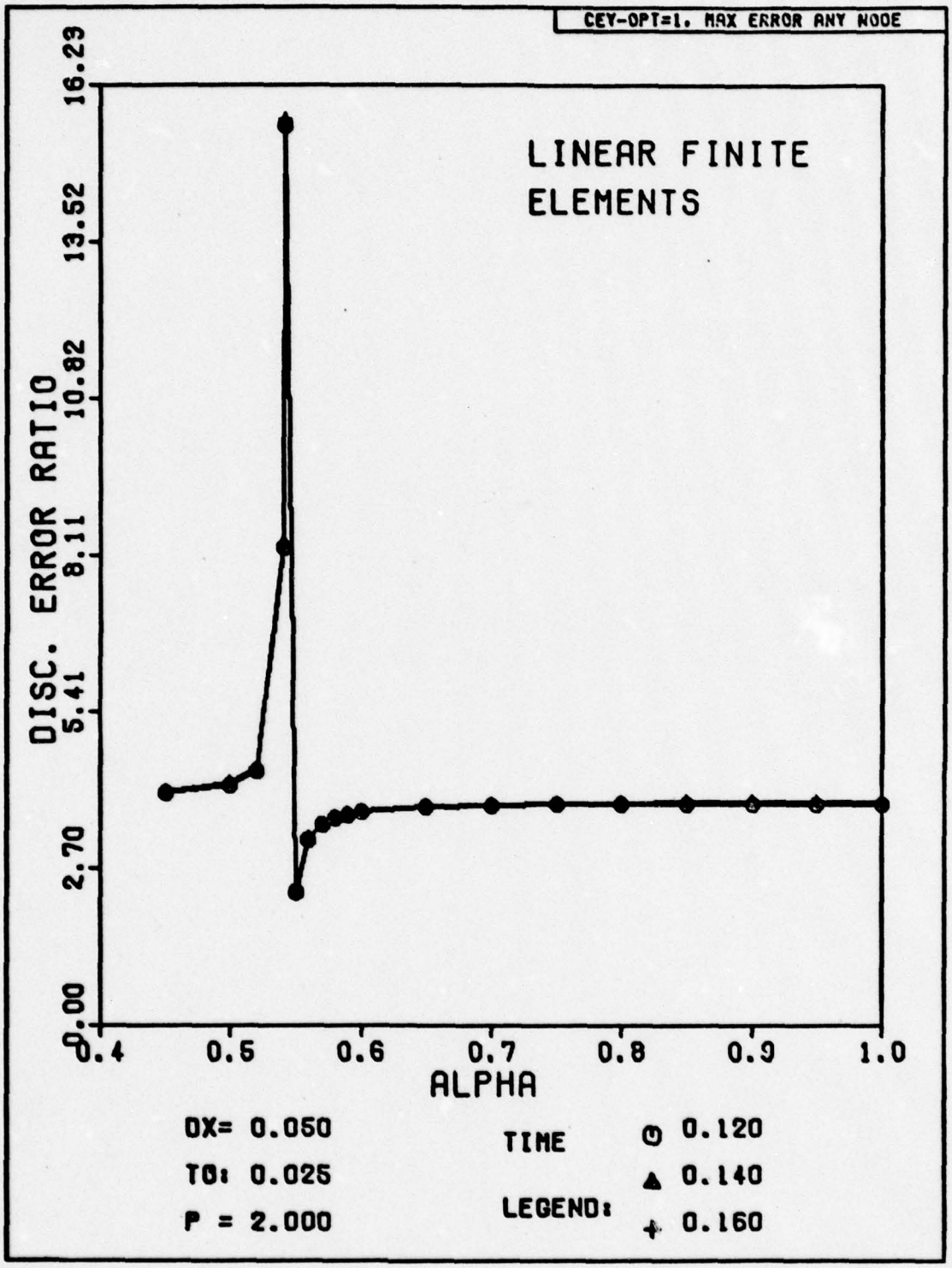


Fig. H-125. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

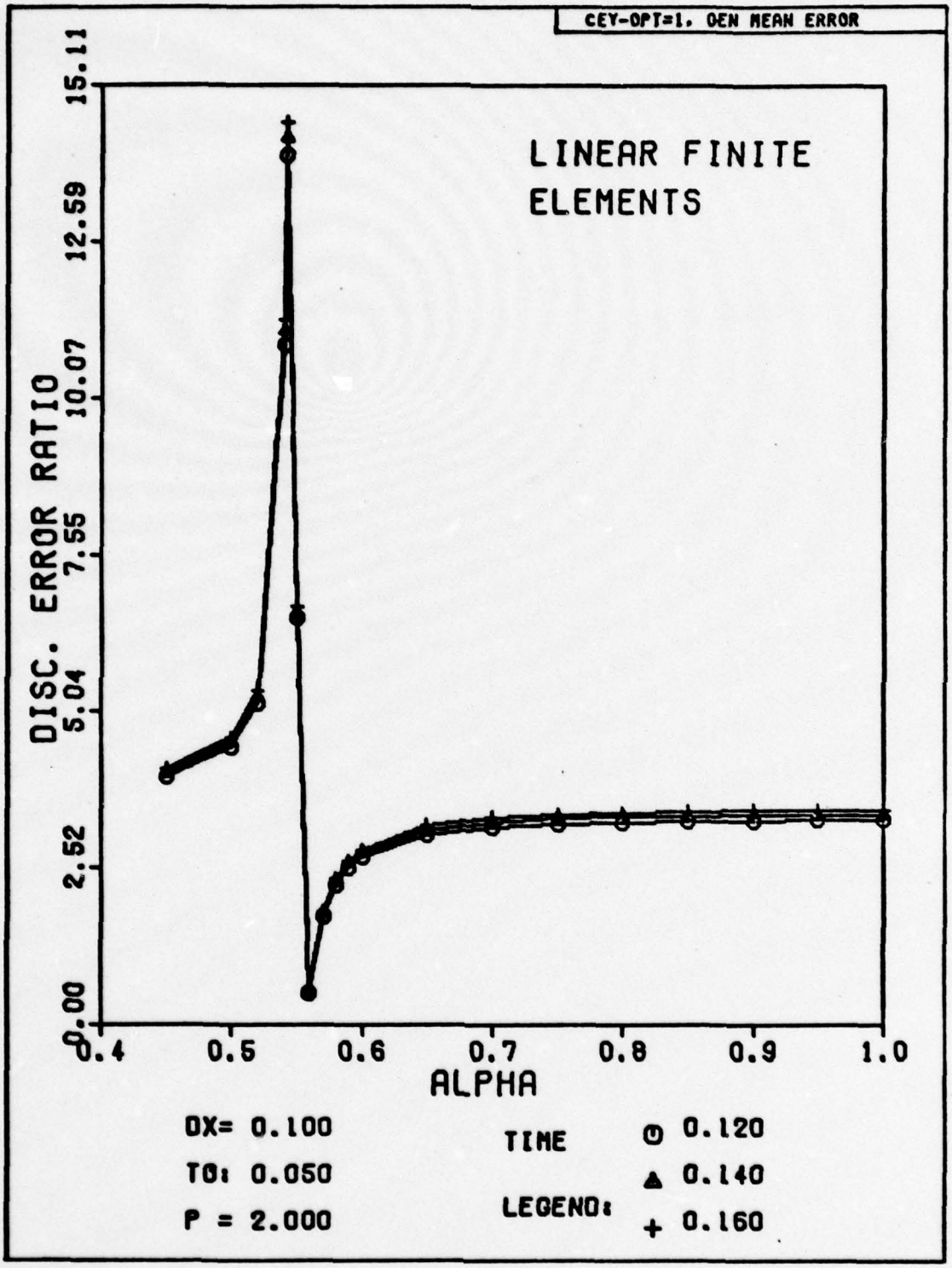


Fig. H-126. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

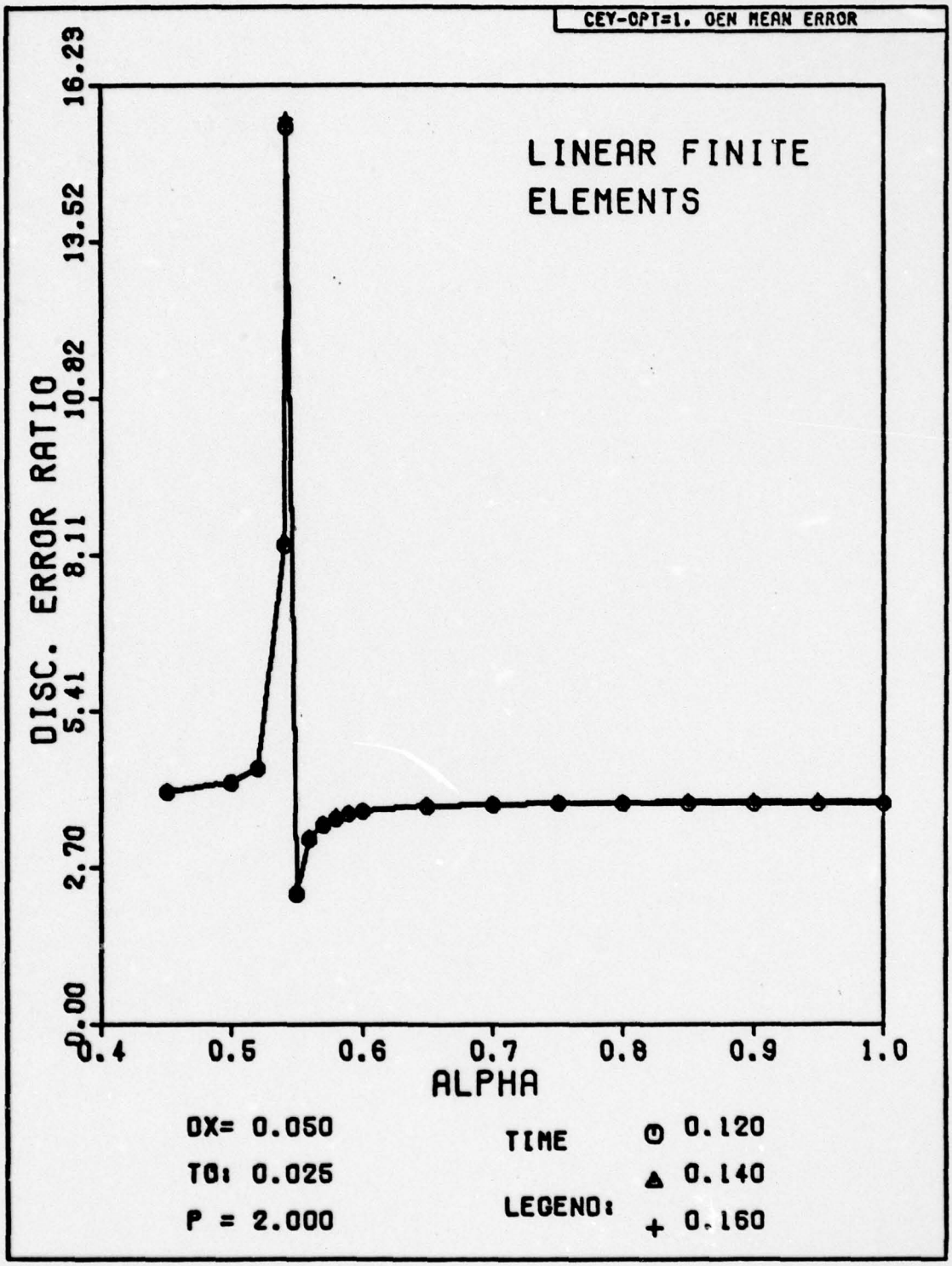


Fig. H-127. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

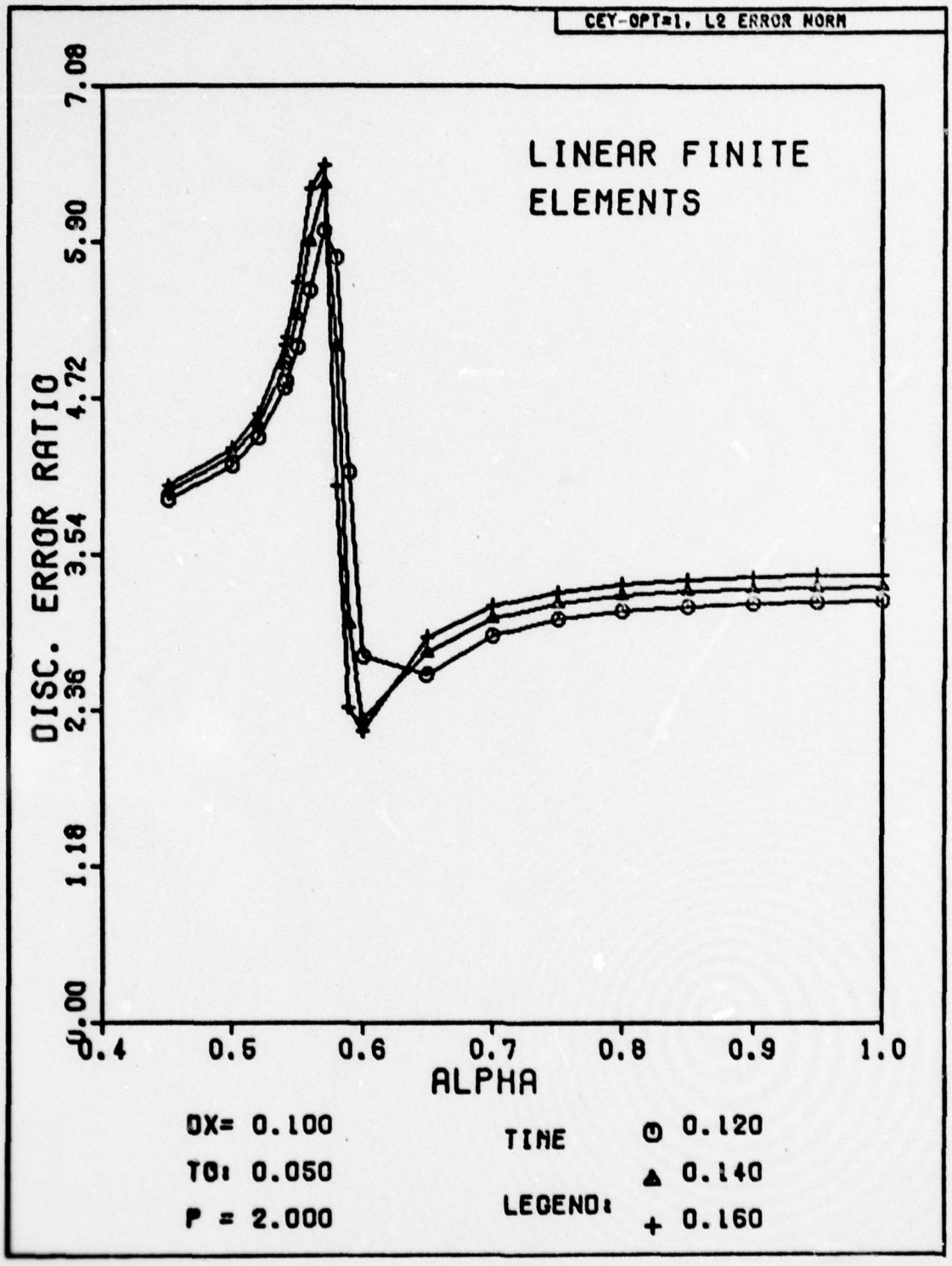


Fig. H-128. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

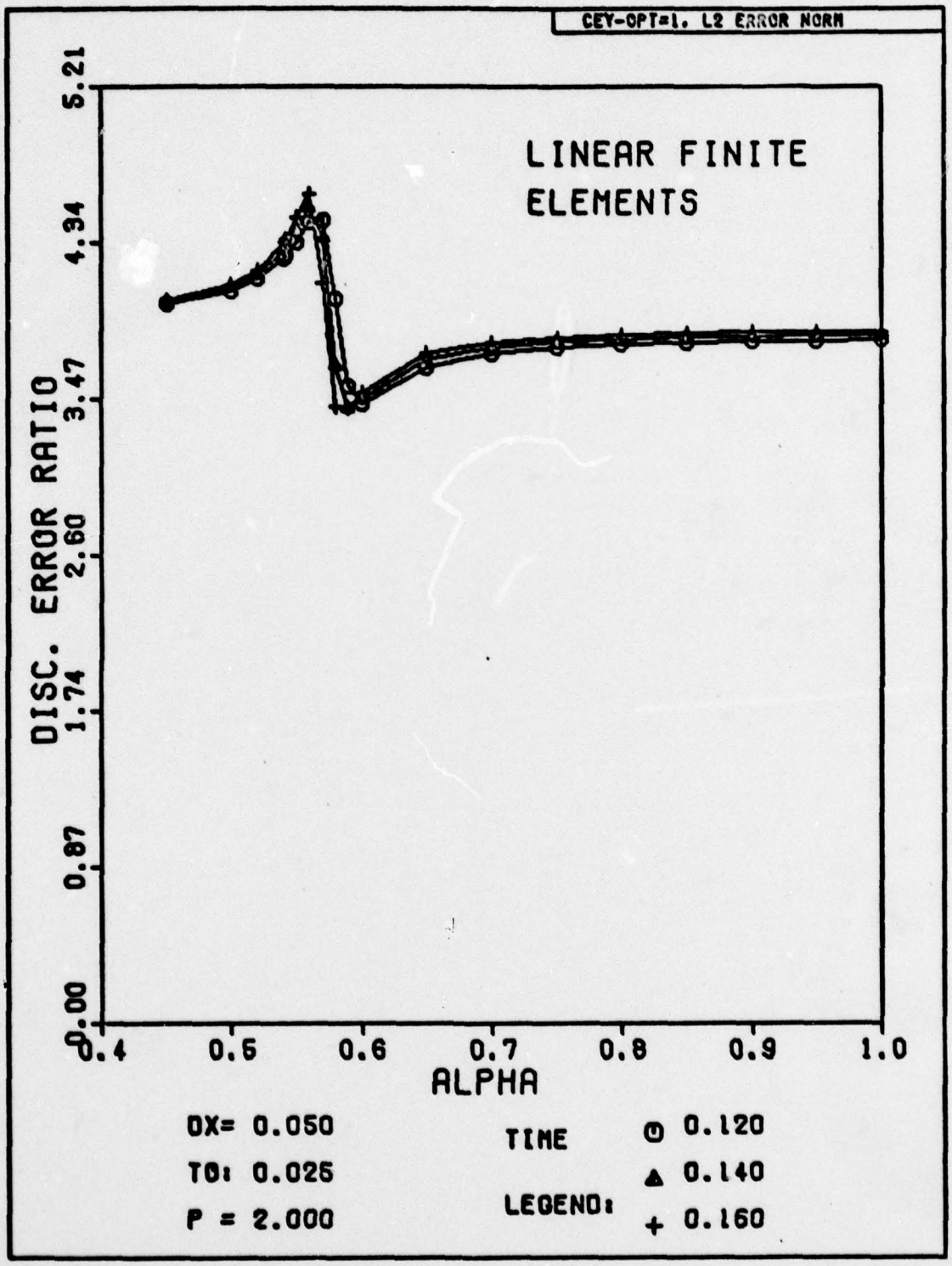


Fig. H-129. Discretization Error Ratio Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

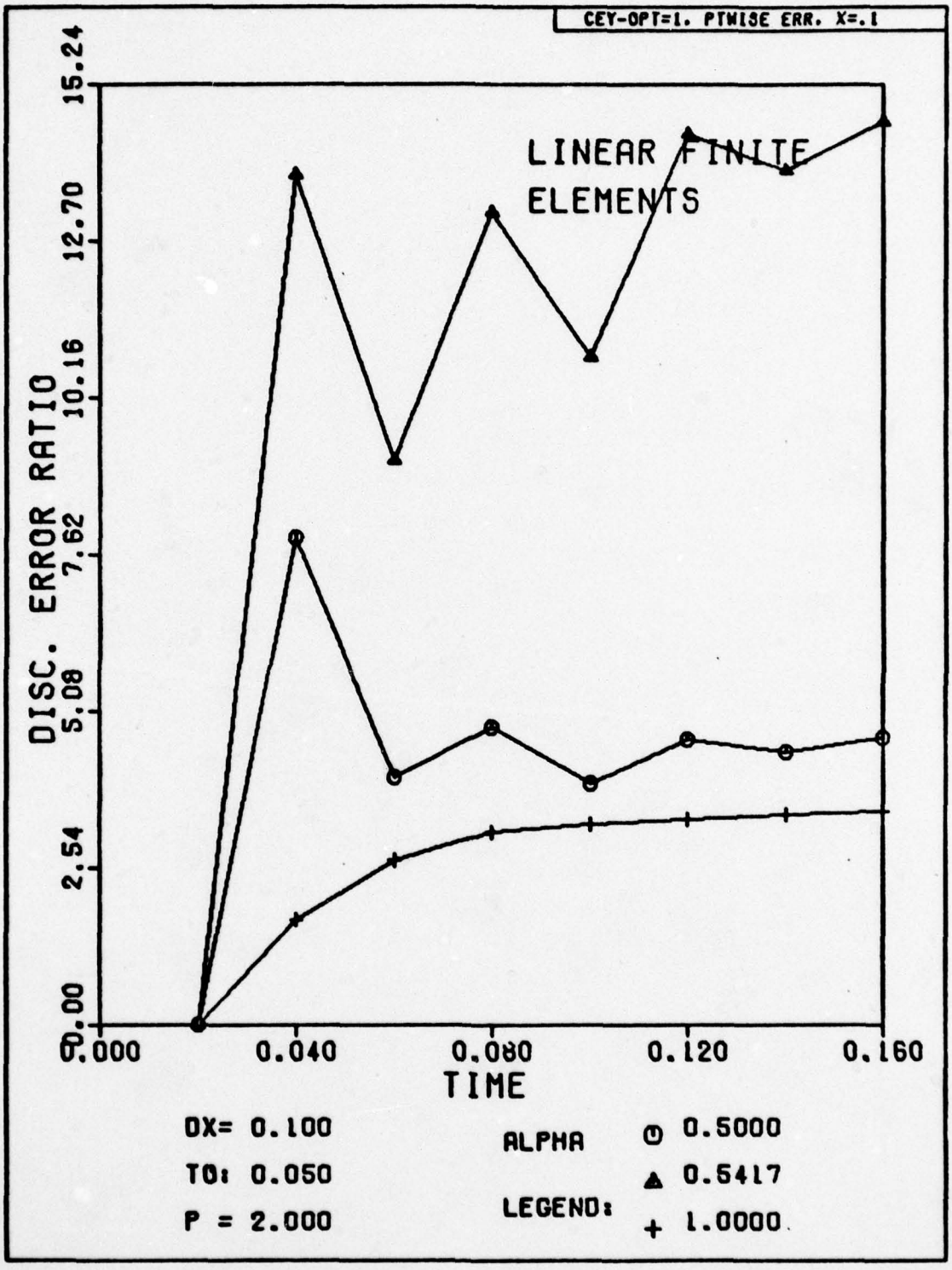


Fig. H-130. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

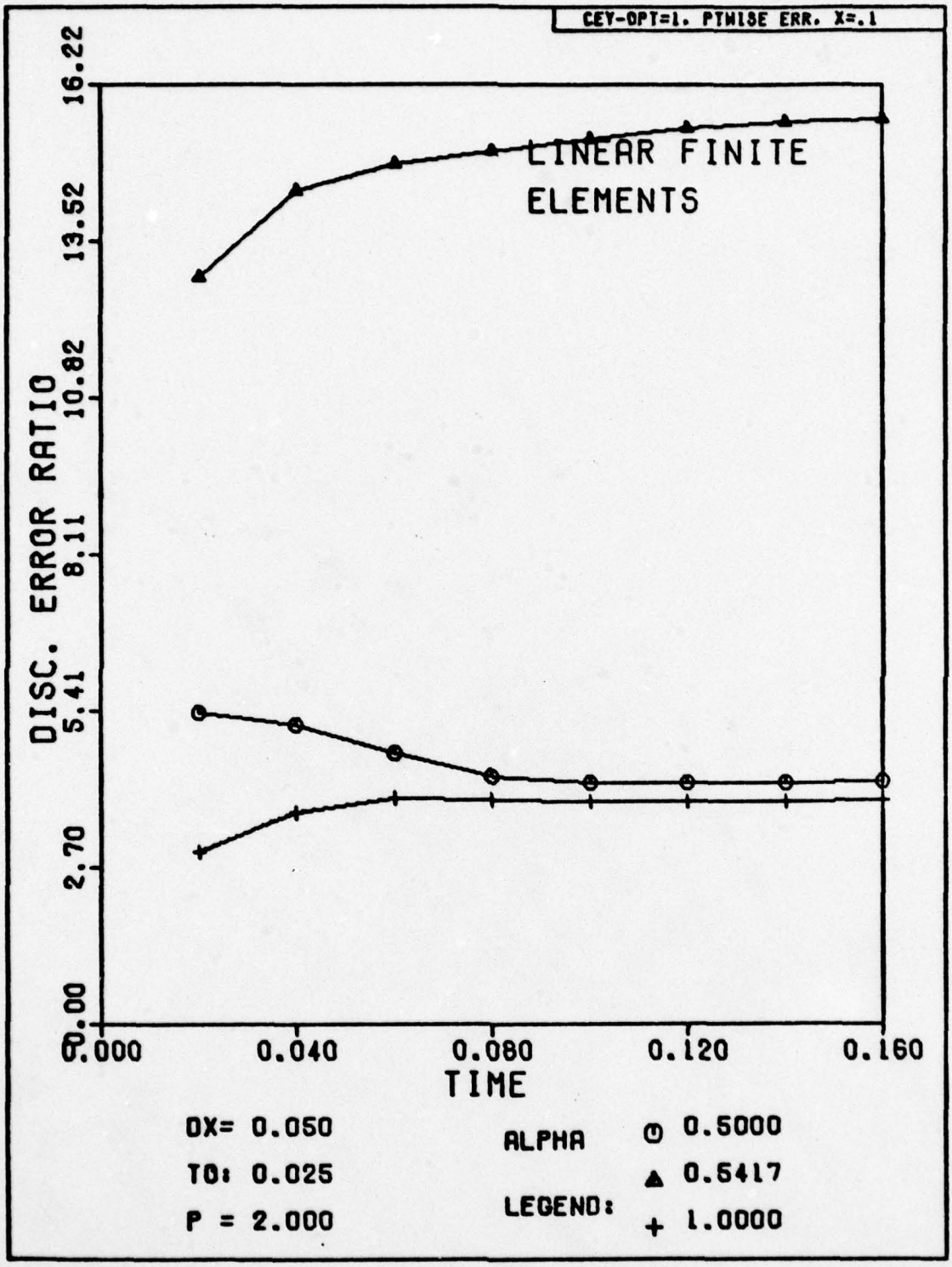


Fig. H-131. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

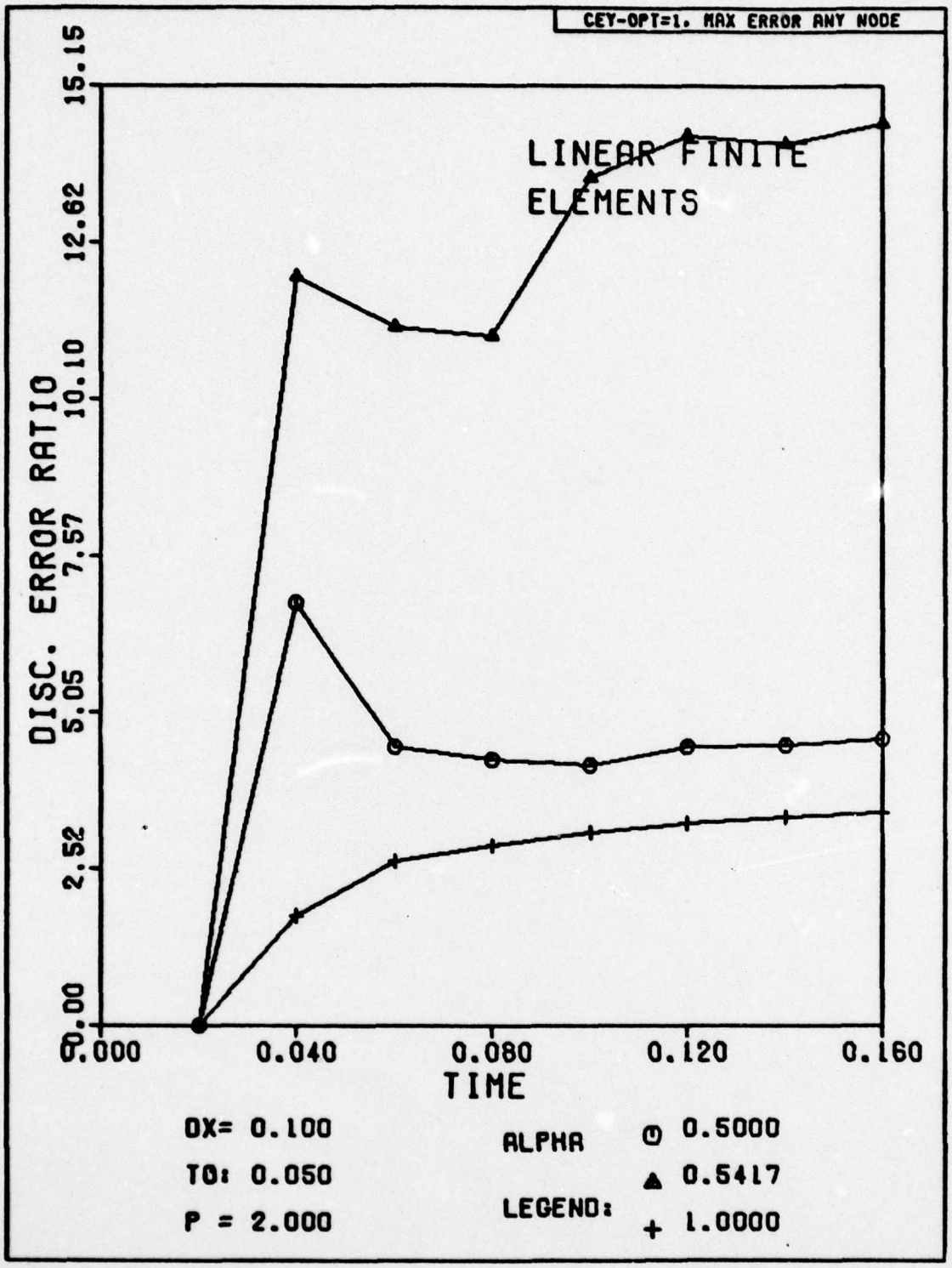


Fig. H-132. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

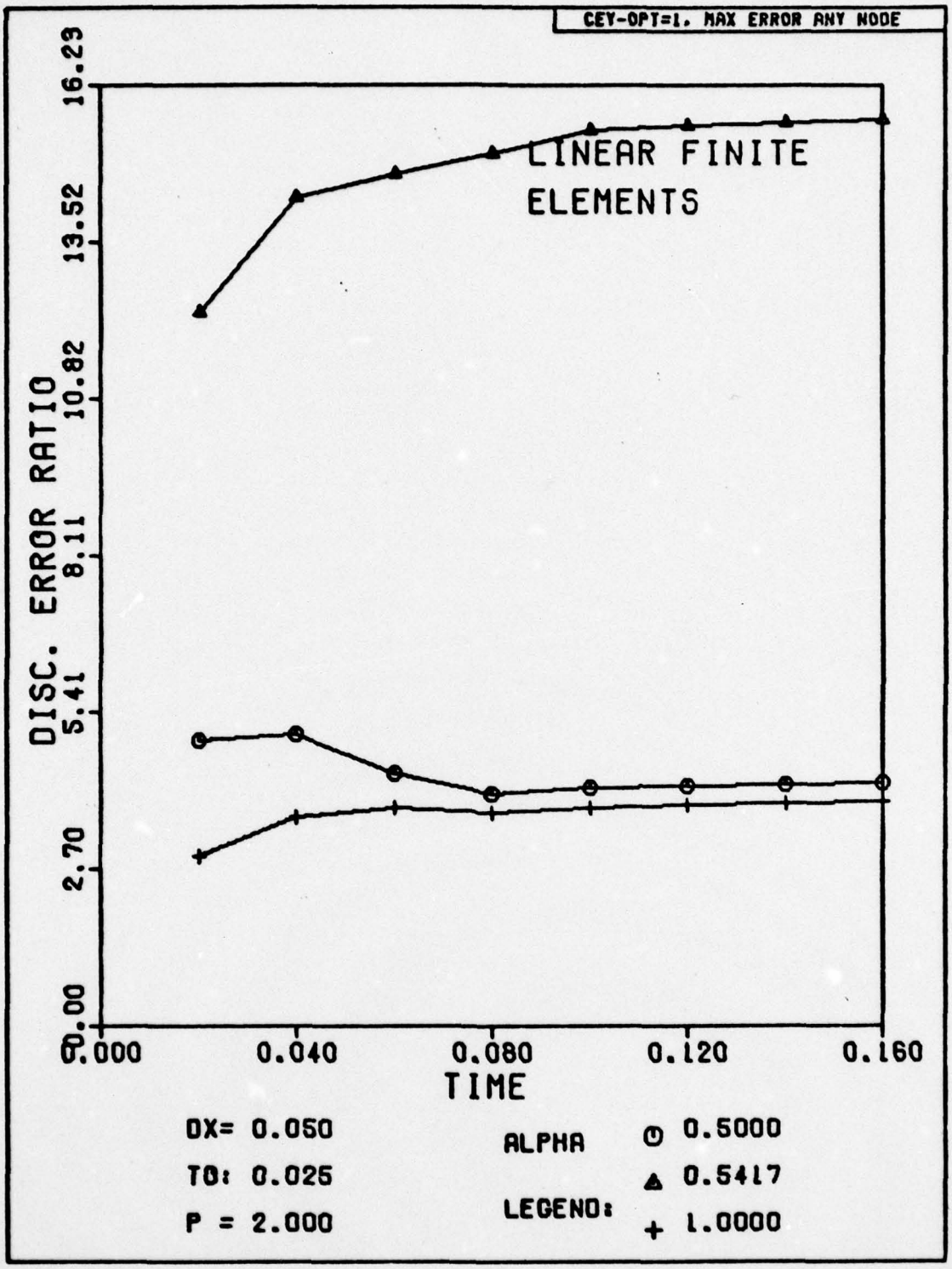


Fig. H-133. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

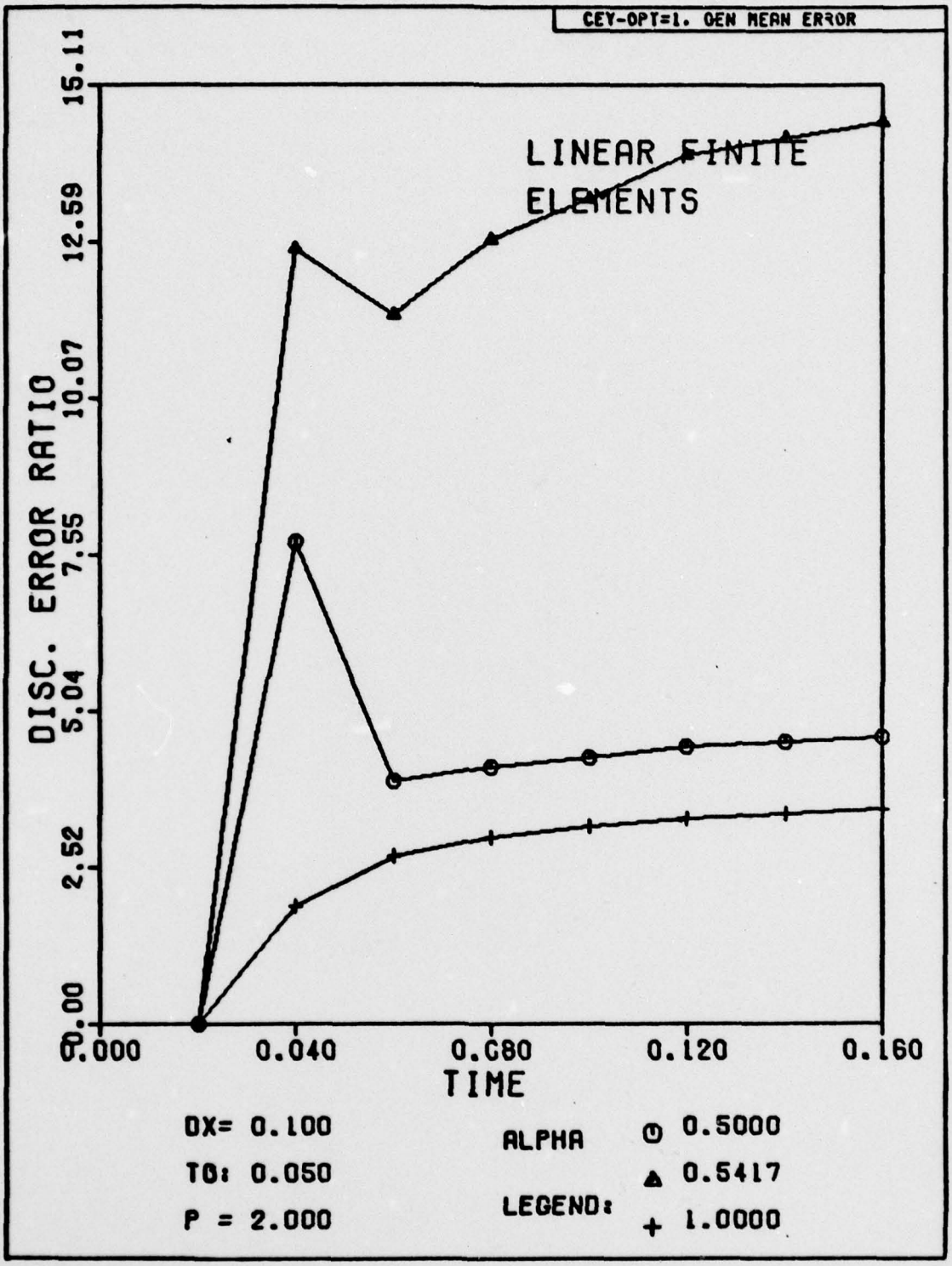


Fig. H-134. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

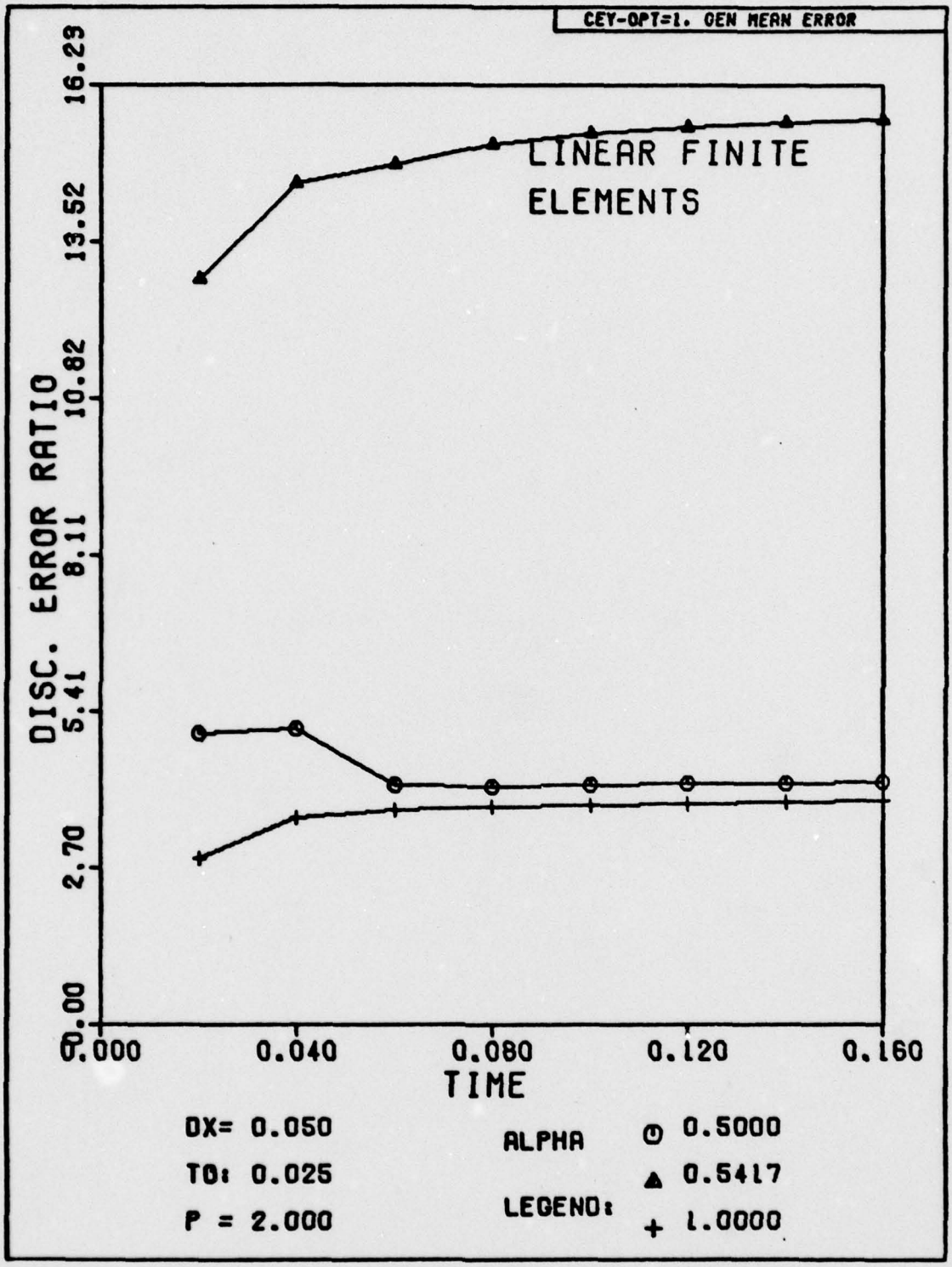


Fig. H-135. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

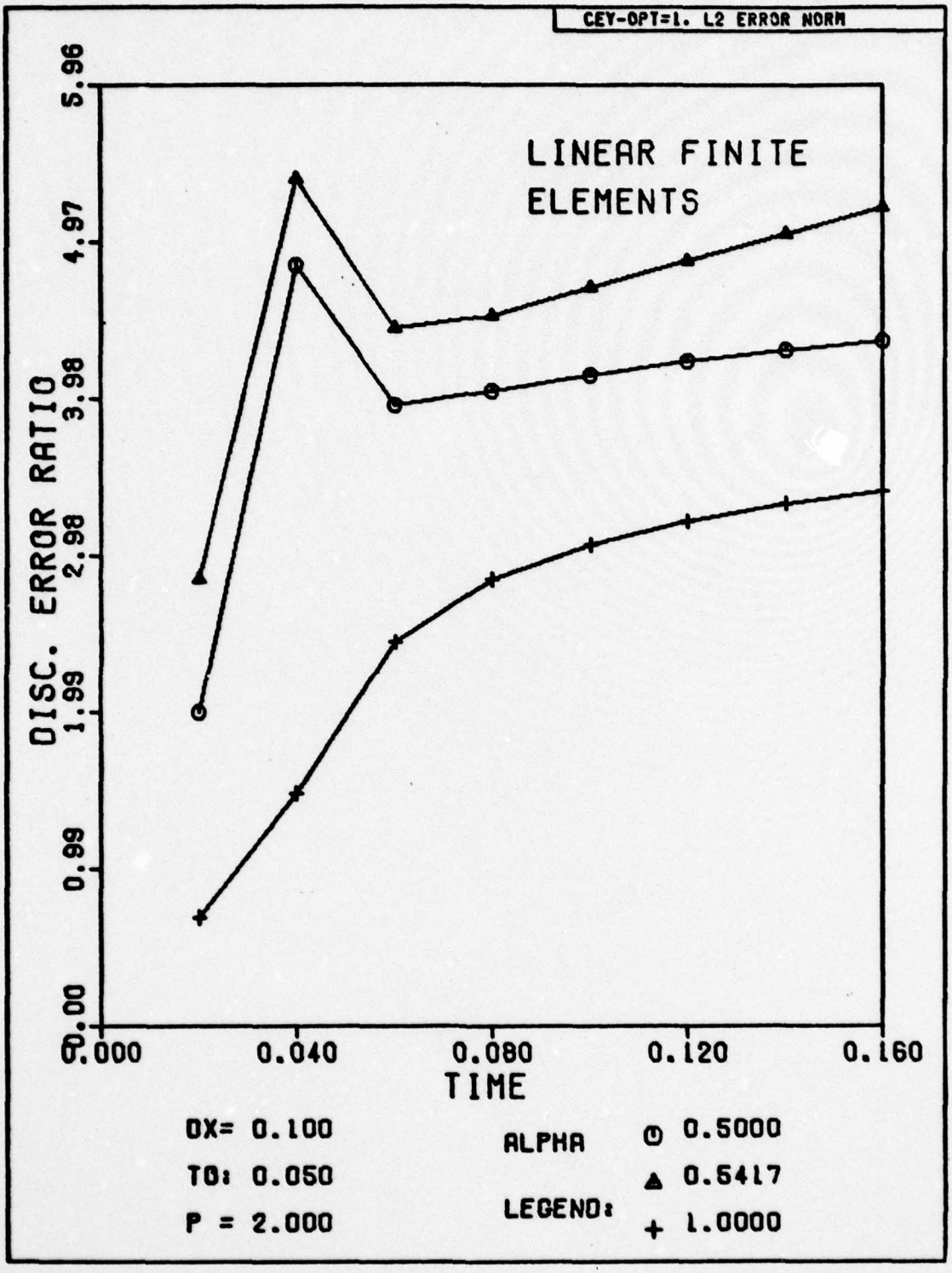


Fig. H-136. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

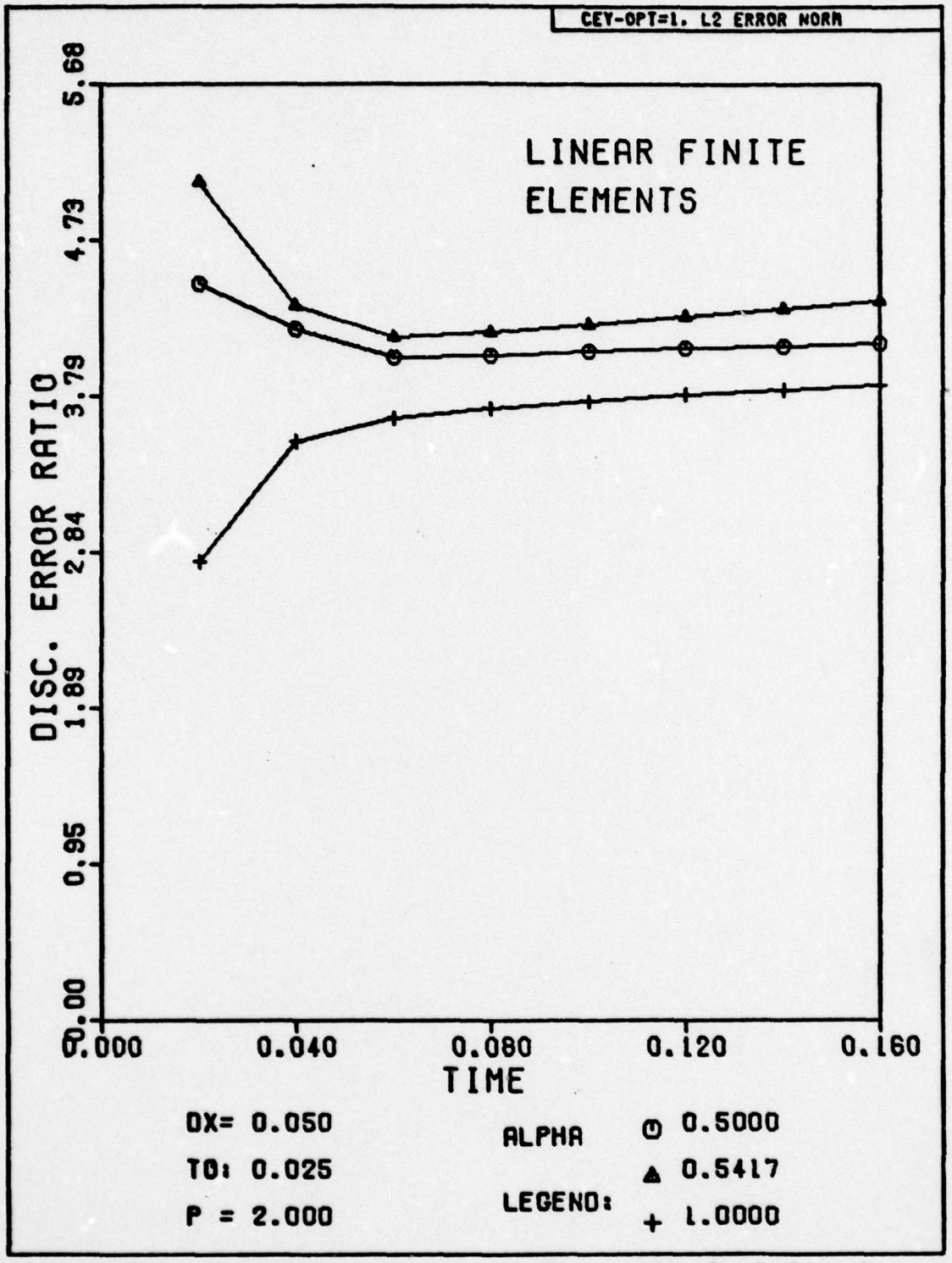


Fig. H-137. Discretization Error Ratio Versus Time for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

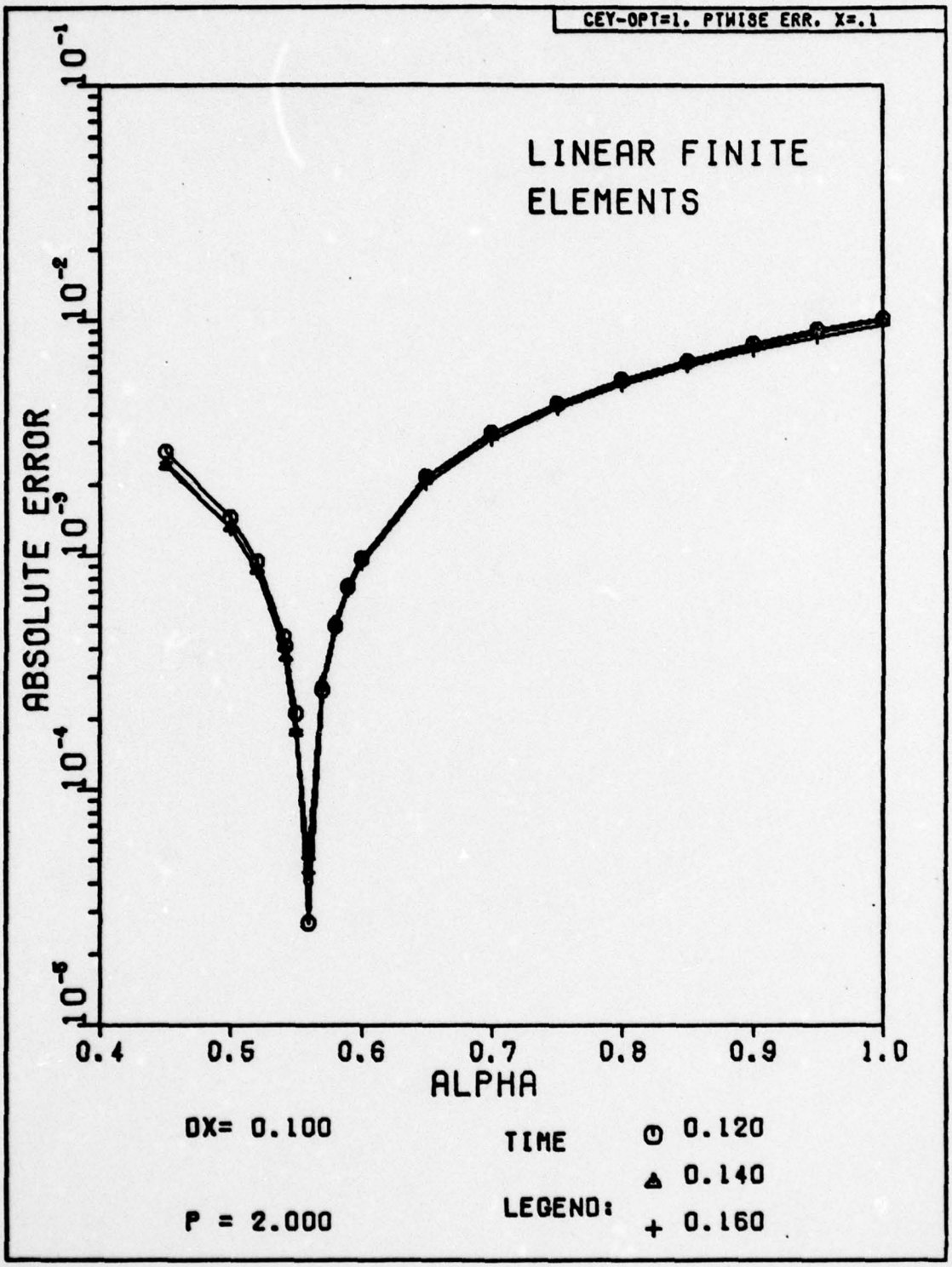


Fig. H-138. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

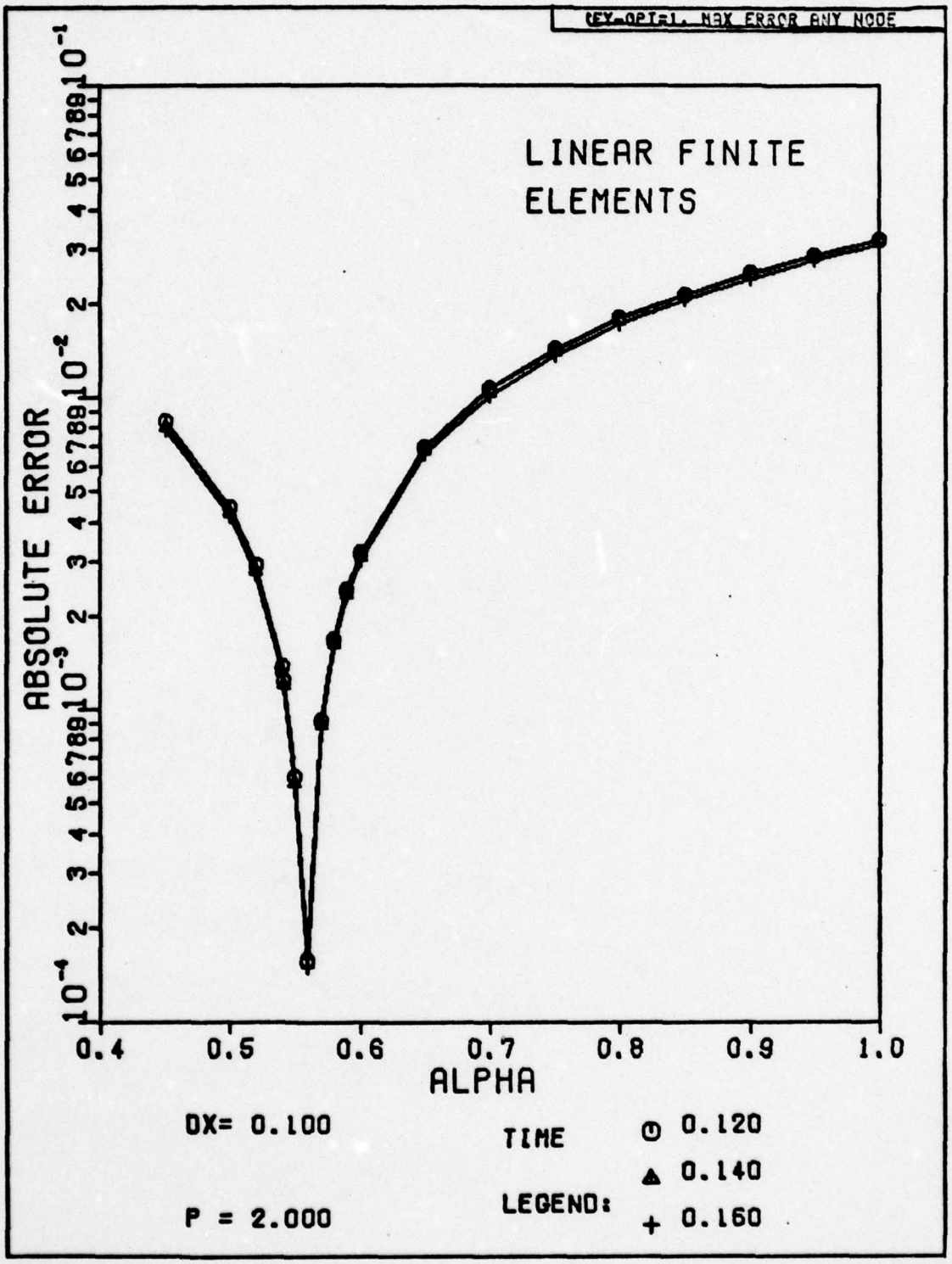


Fig. H-139. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

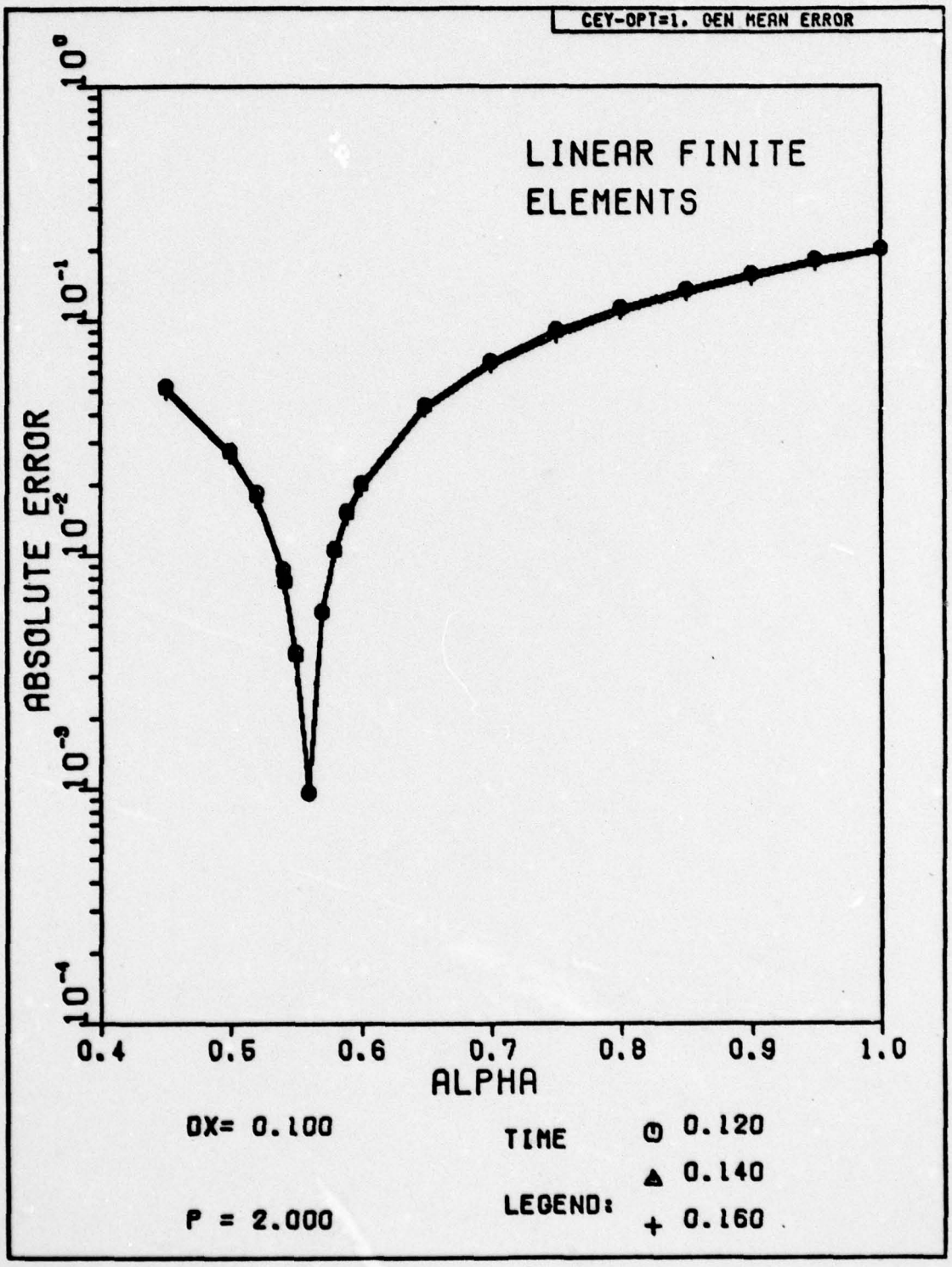


Fig. H-140. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

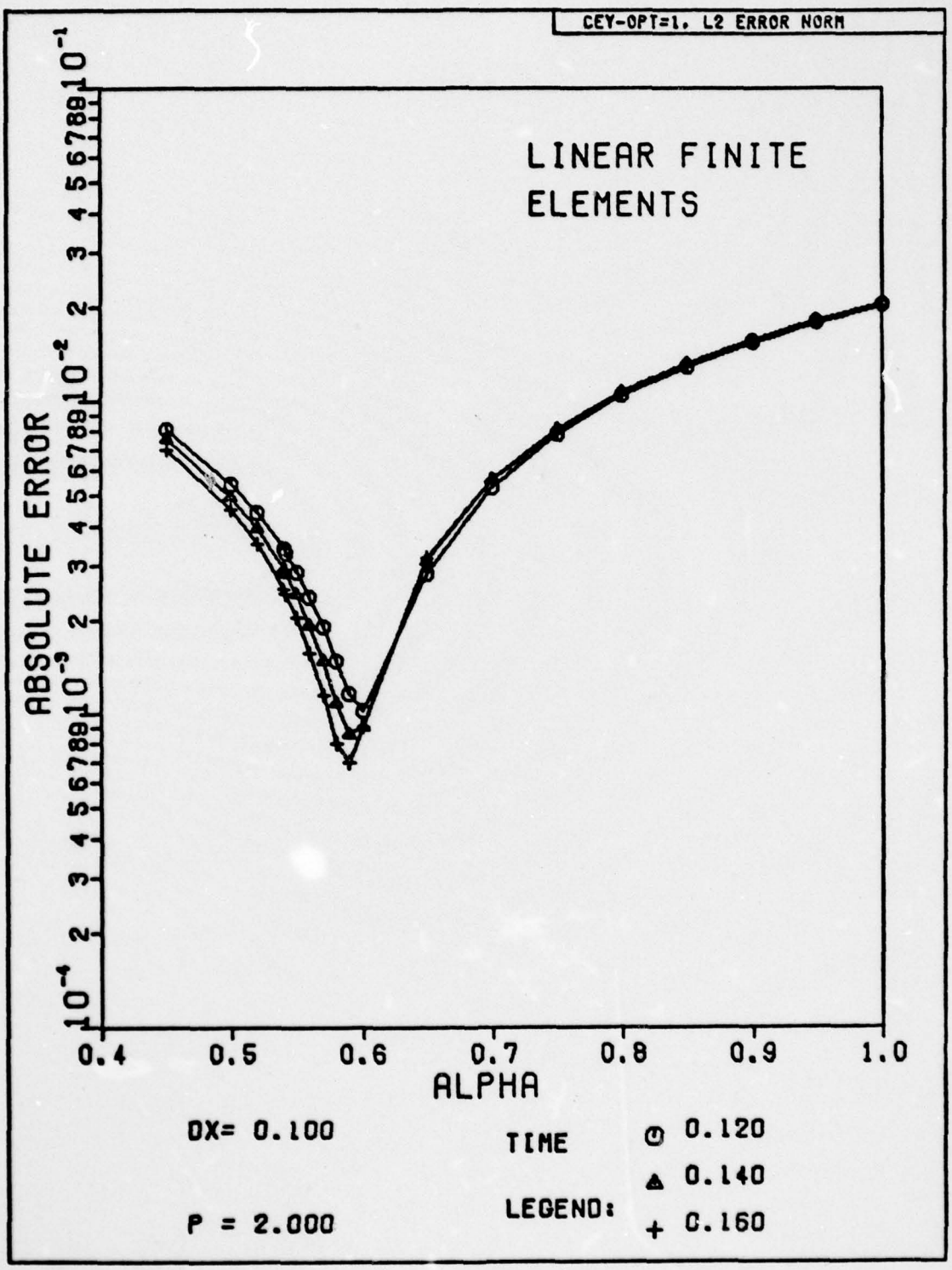
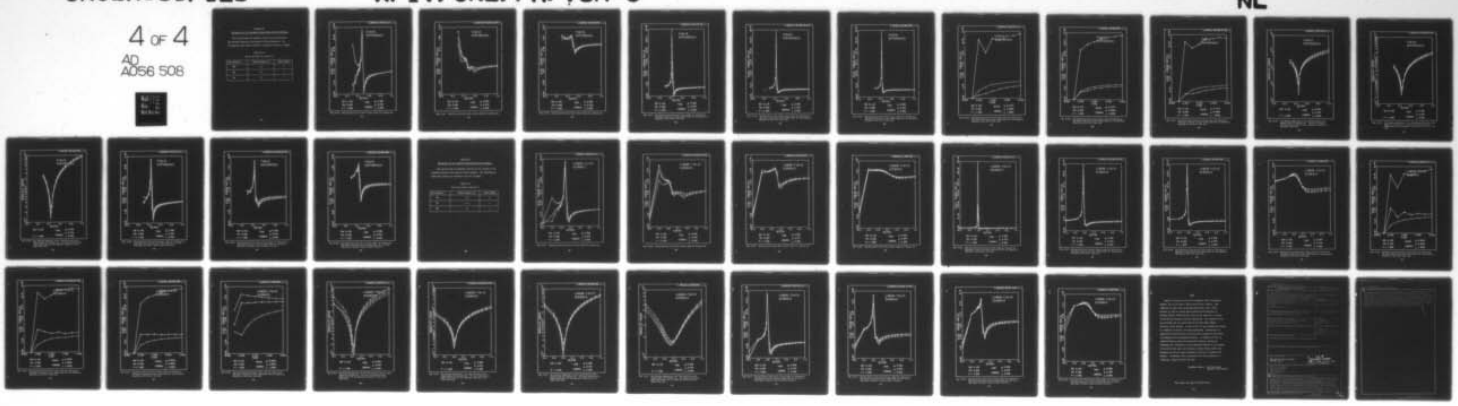


Fig. H-141. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem One. The exact solution has been substituted for the numerical solution at the first time step.

AD-A056 508 AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OHIO SCH--ETC F/G 20/13
AN INVESTIGATION OF THE NUMERICAL METHODS OF FINITE DIFFERENCES--ETC(U)
MAR 78 C R MARTIN

UNCLASSIFIED AFIT/GNE/PH/78M-6 NL

4 OF 4
AD
A056 508



END
DATE
FILMED
8-78
DDC

Section III

The Results for the Secondary Problem Using Finite-Differences

This section shows the graphical results for the solution of the secondary problem by the method of finite-differences. The following key shows which options are included in this set of graphs.

Table H-III

Key to the Plots in Section III

Run Identifier	Fourier Modulus (p)	Option Number
CDE	1.0	0
CDF	1.0	1
CDG	1.0	2

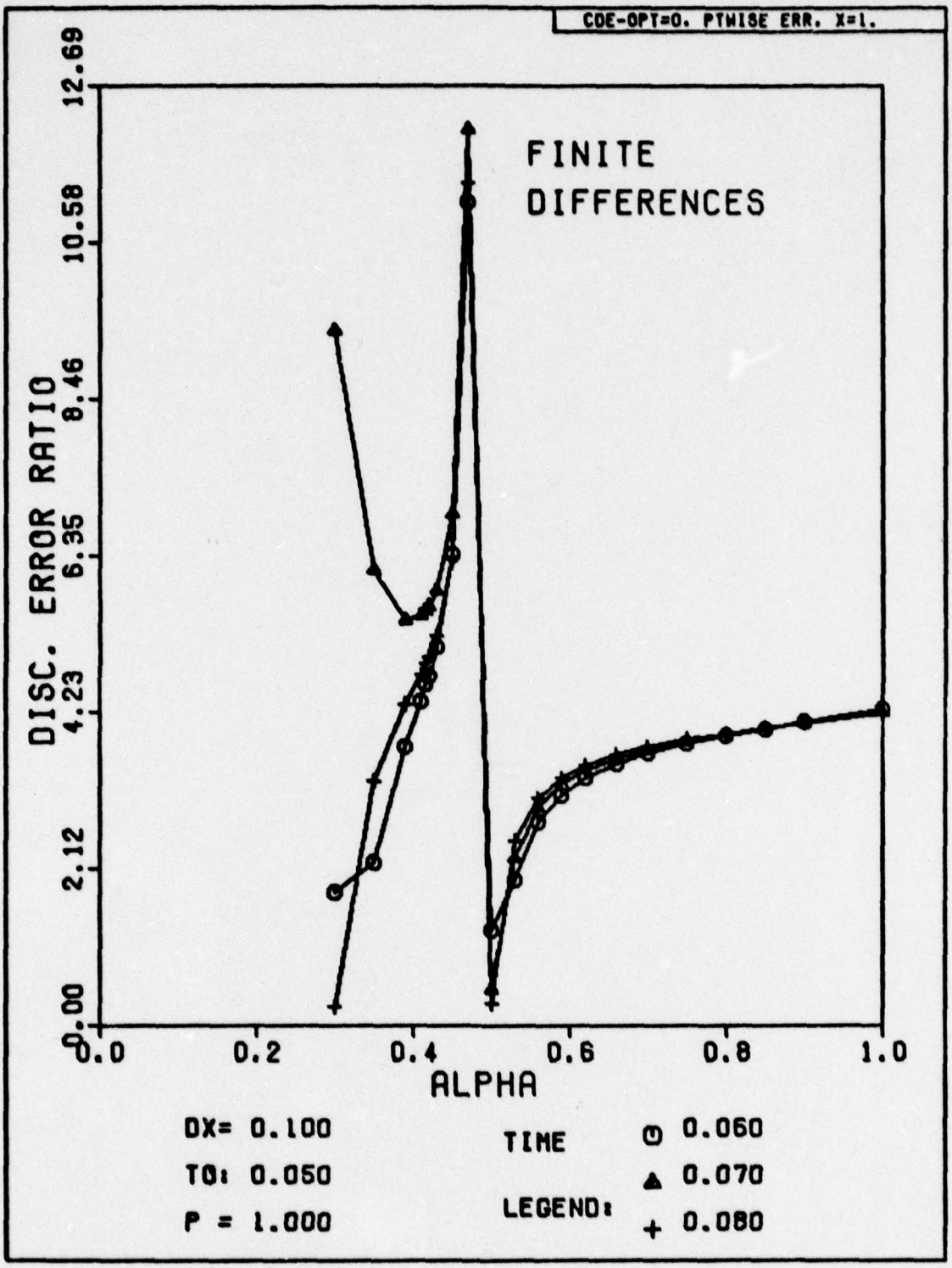


Fig. H-142. Discretization Error Ratio Versus Alpha for Problem Two.

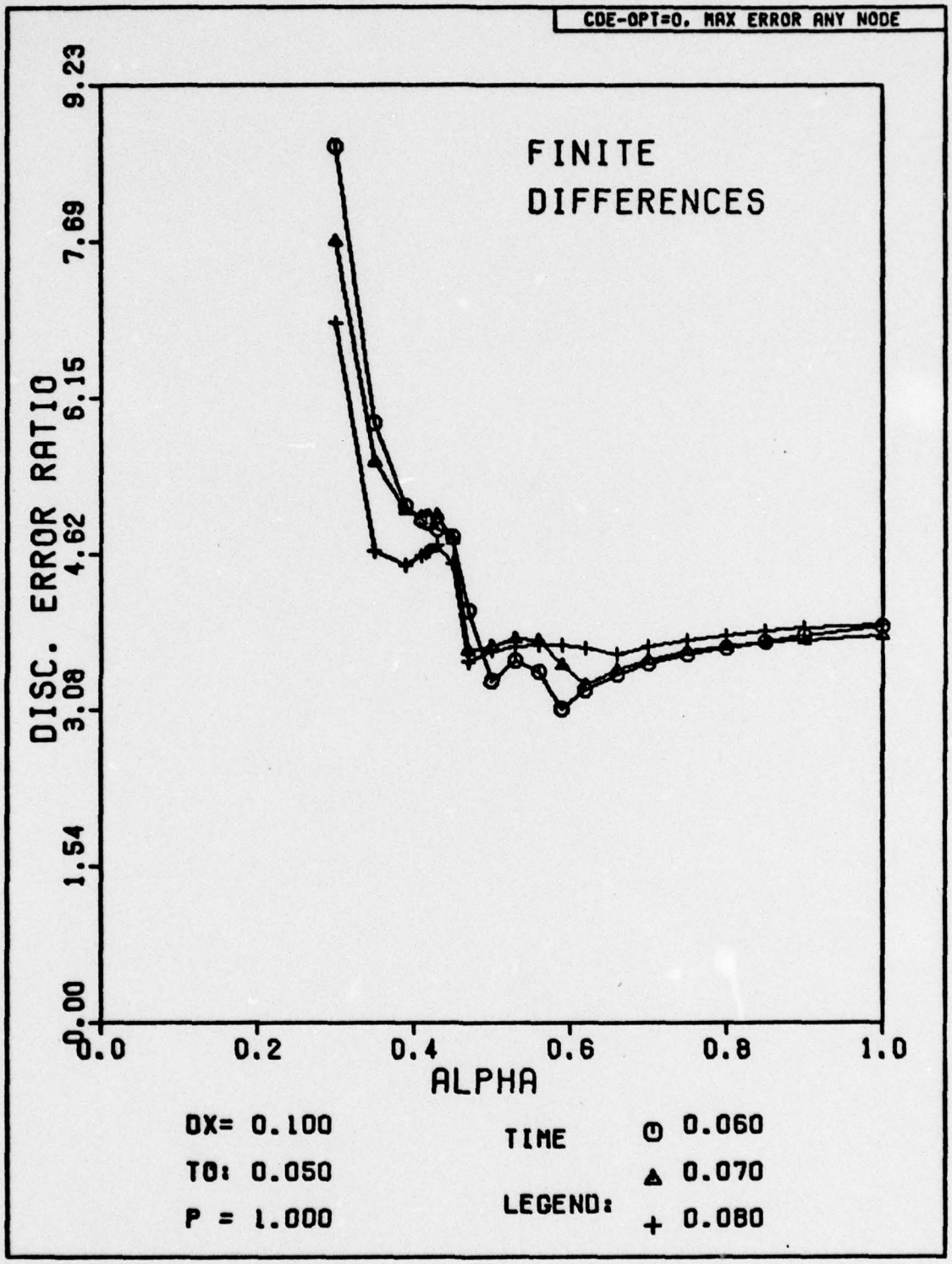


Fig. H-143. Discretization Error Ratio Versus Alpha for Problem Two.

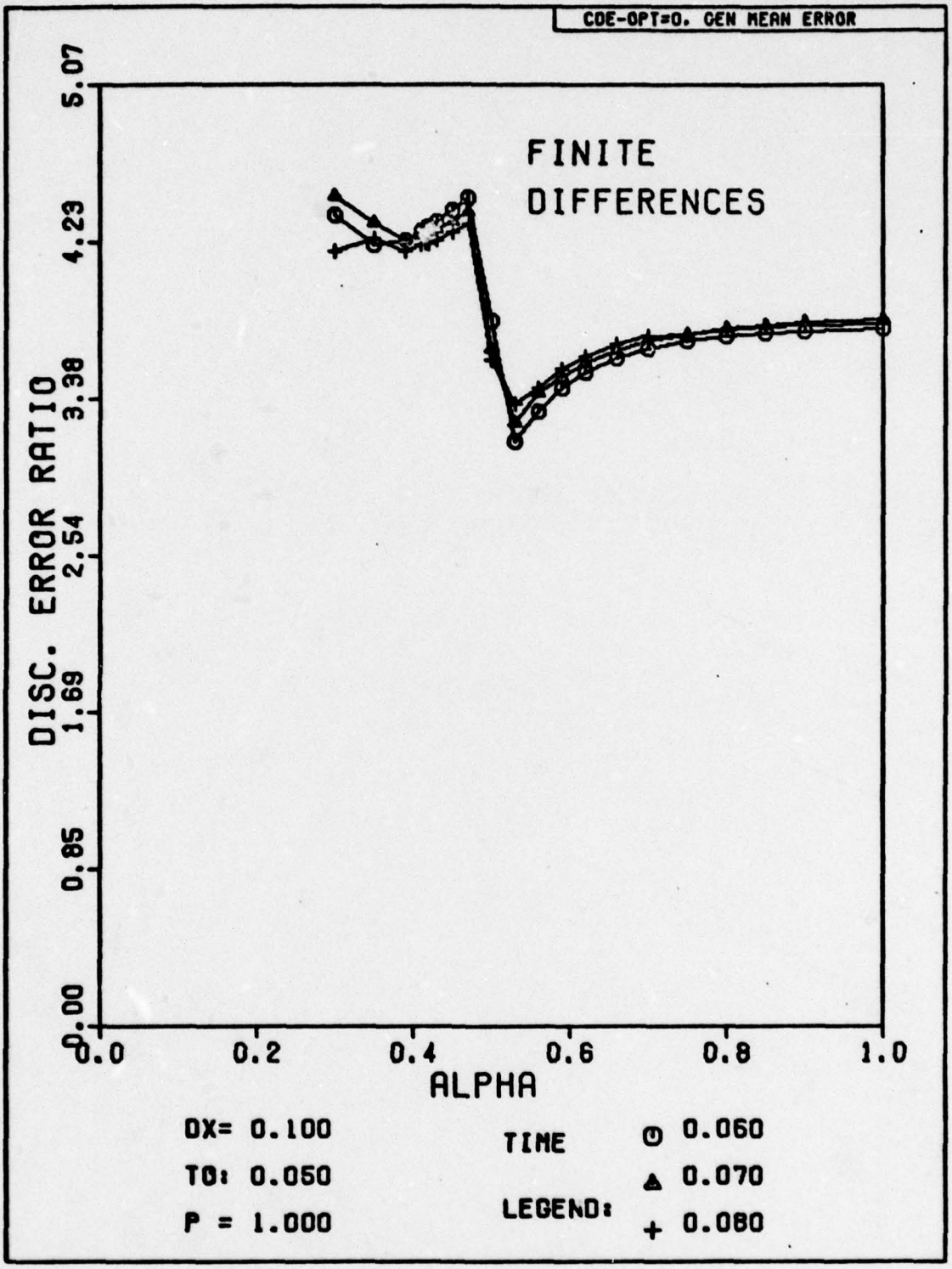


Fig. H-144. Discretization Error Ratio Versus Alpha for Problem Two.

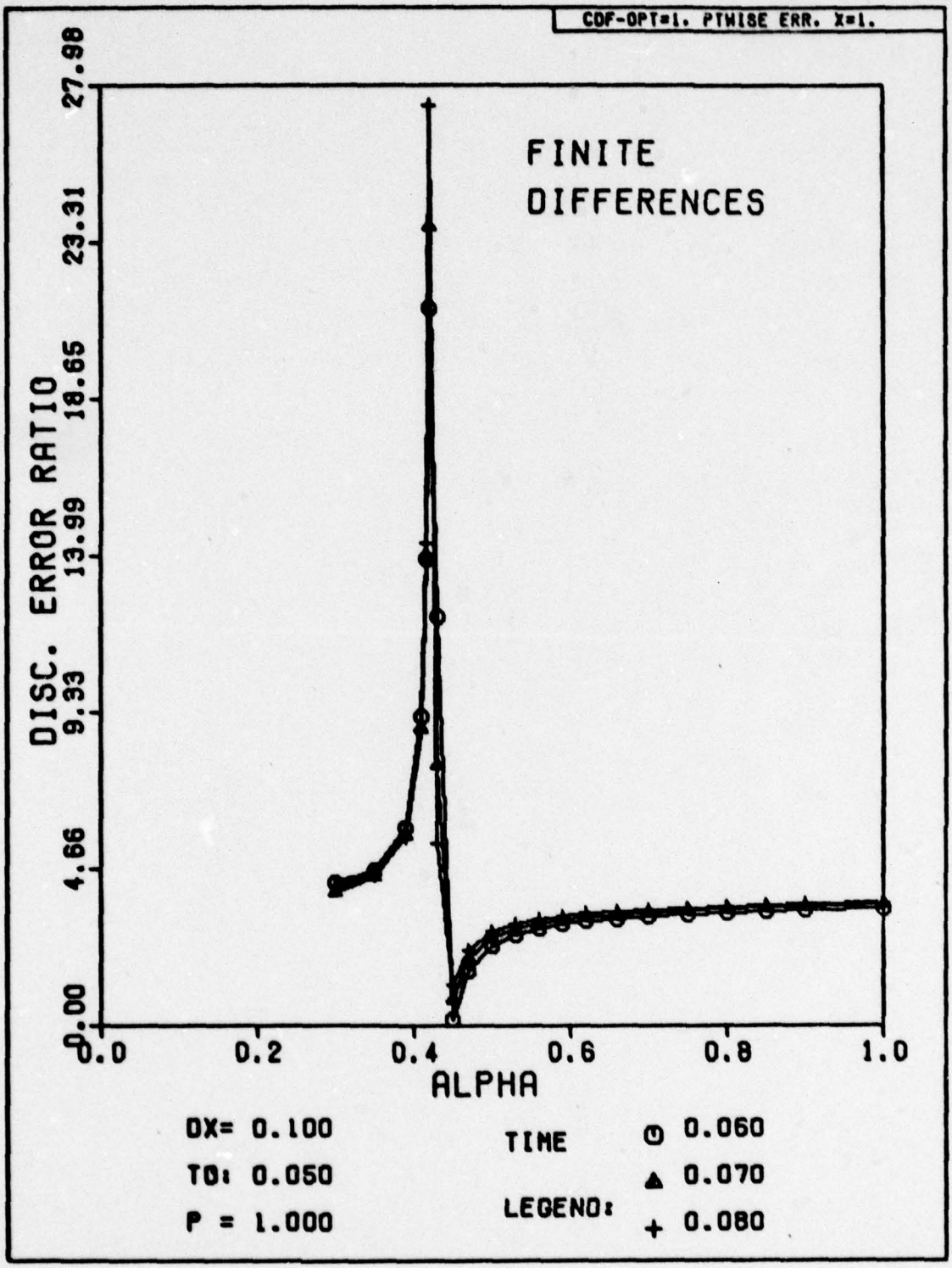


Fig. H-145. Discretization Error Ratio Versus Alpha for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

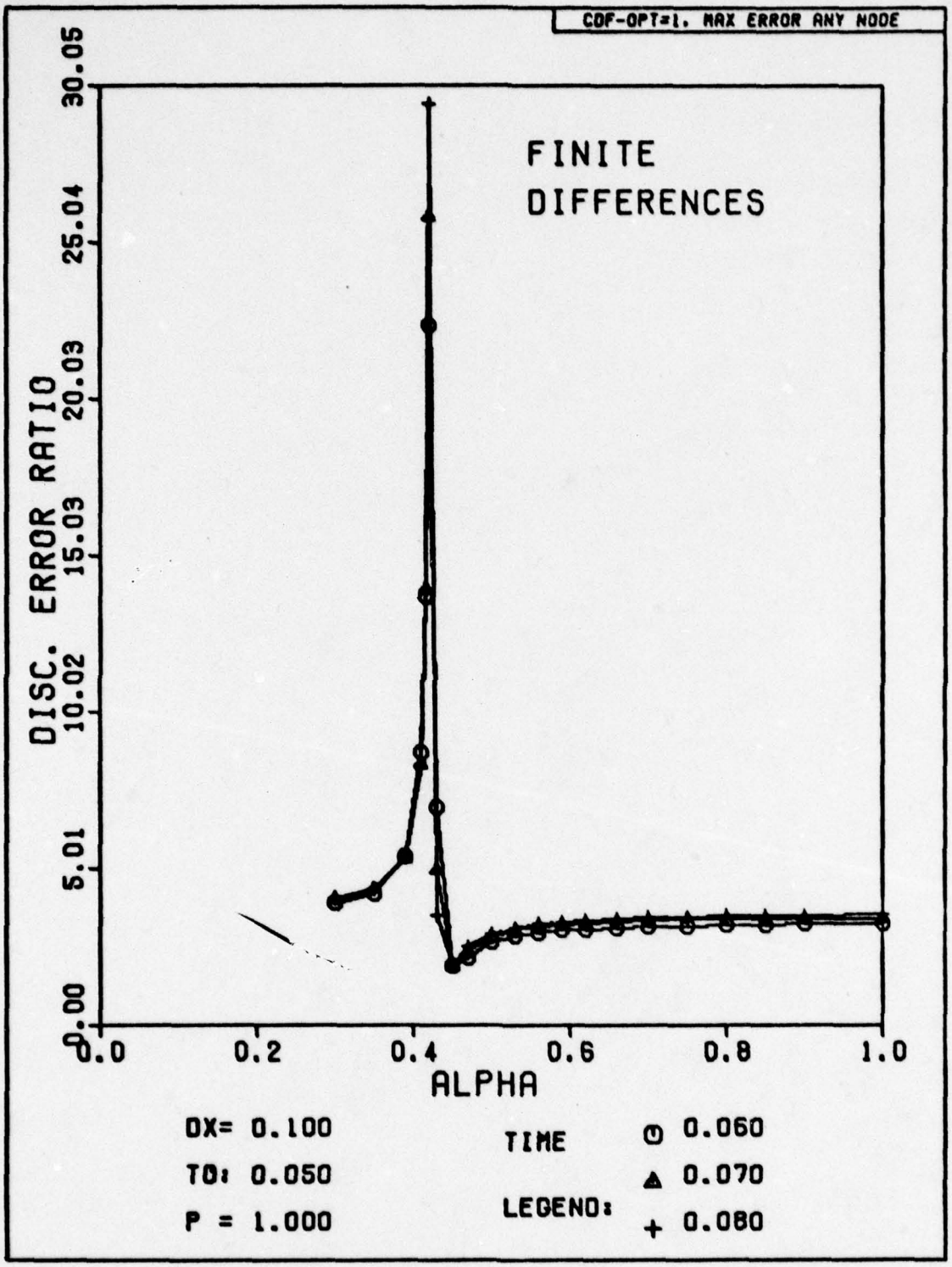


Fig. H-146. Discretization Error Ratio Versus Alpha for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

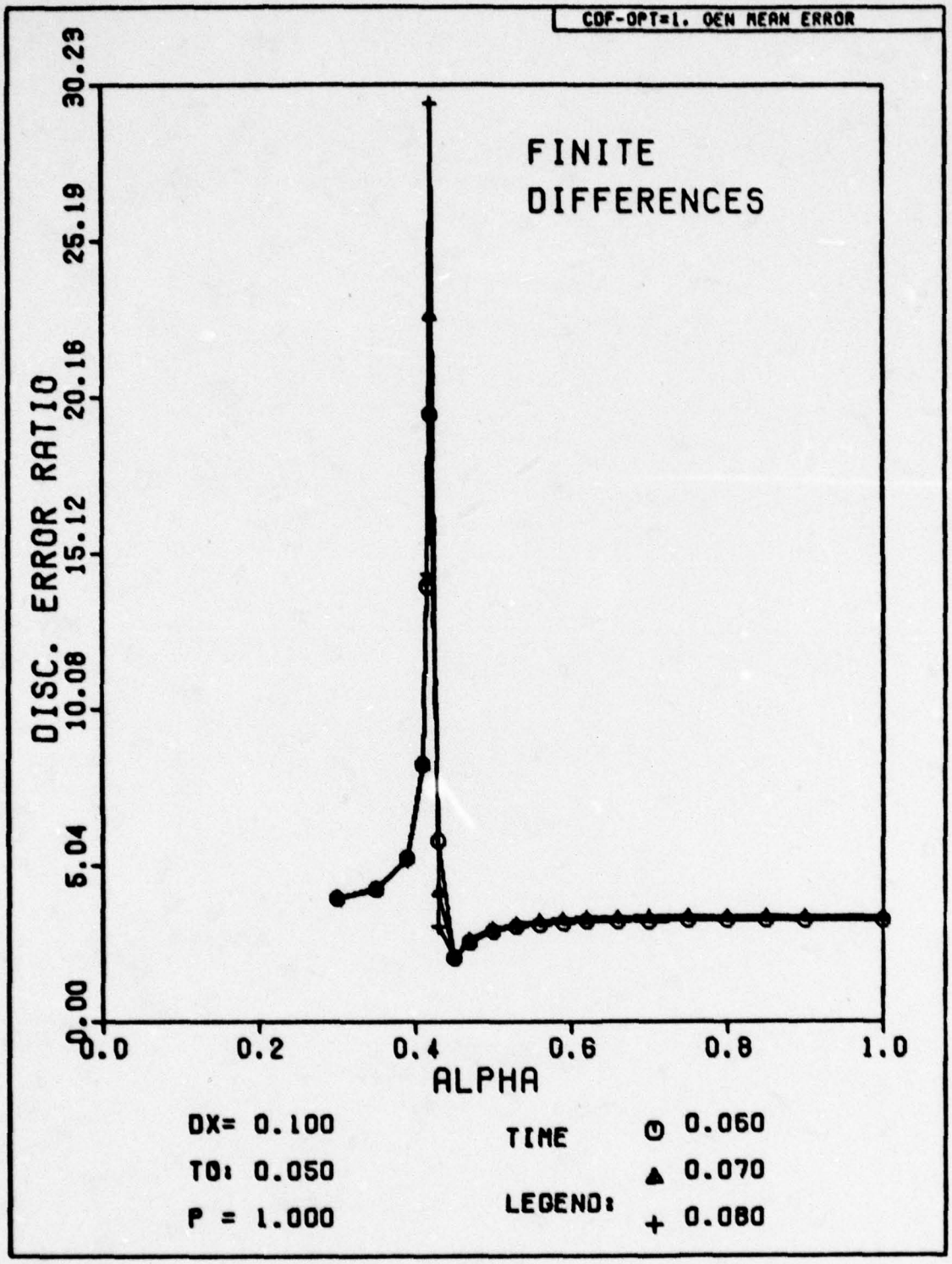


Fig. H-147. Discretization Error Ratio Versus Alpha for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

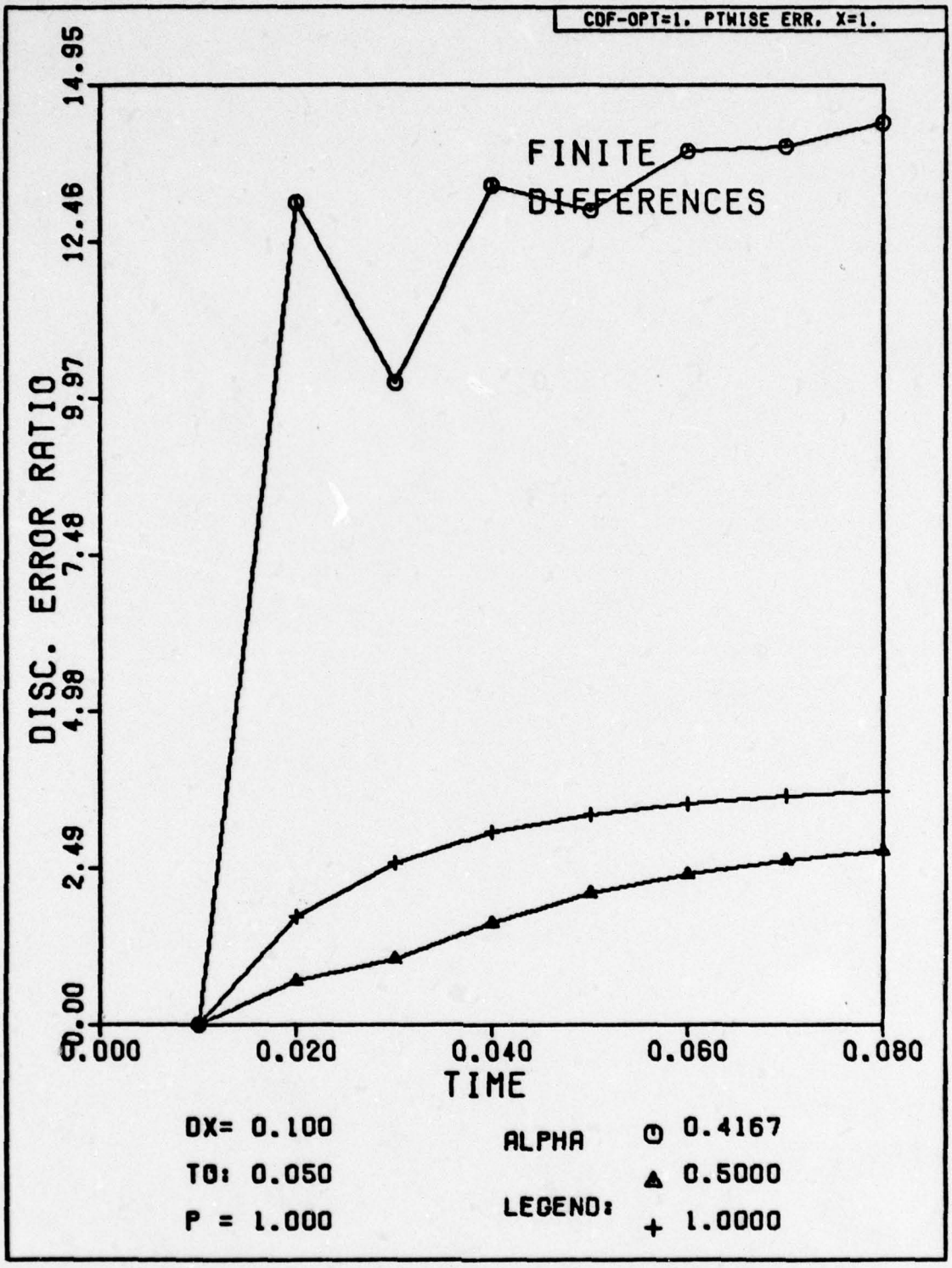


Fig. H-148. Discretization Error Ratio Versus Time for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

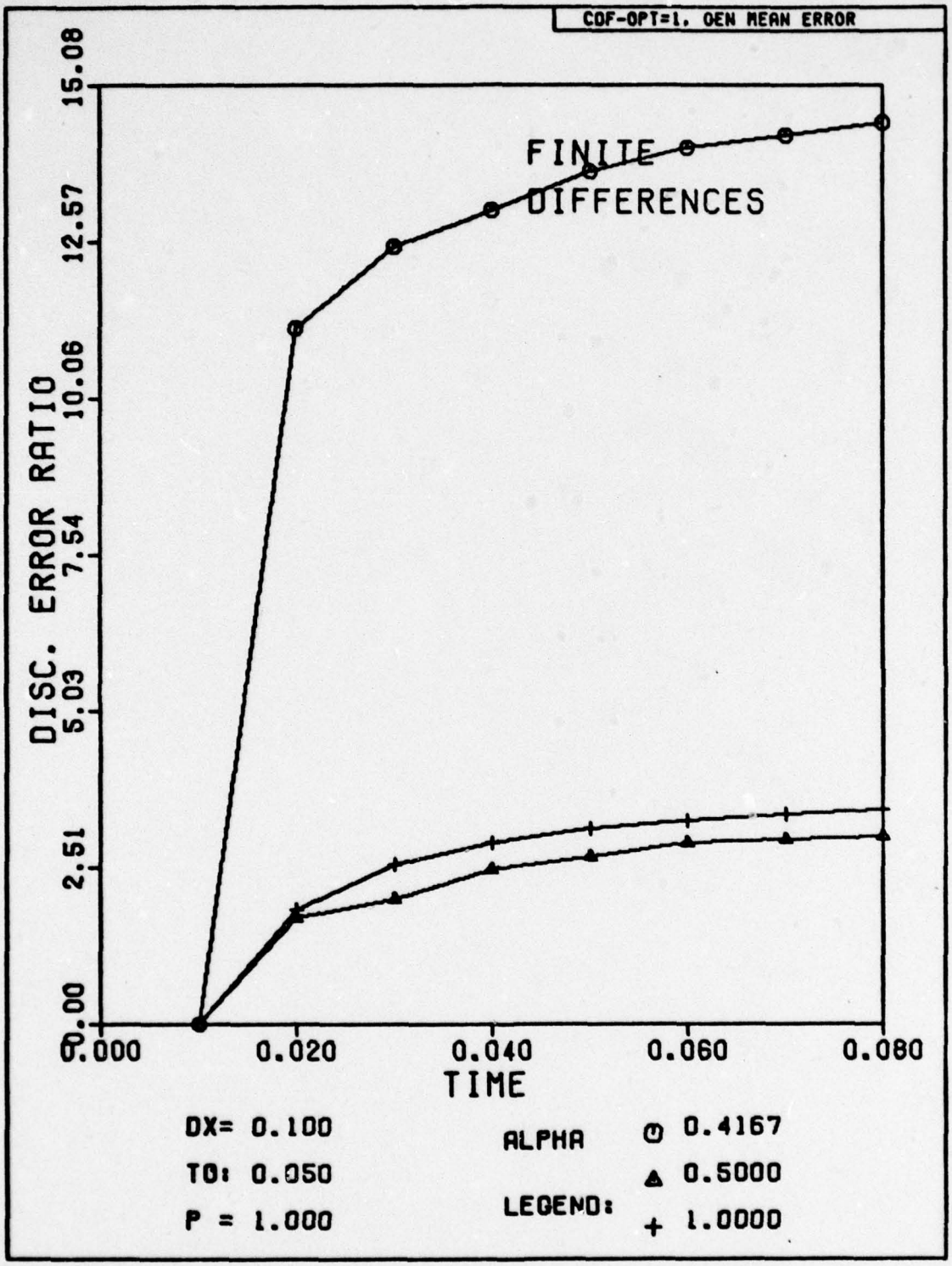


Fig. H-149. Discretization Error Ratio Versus Time for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

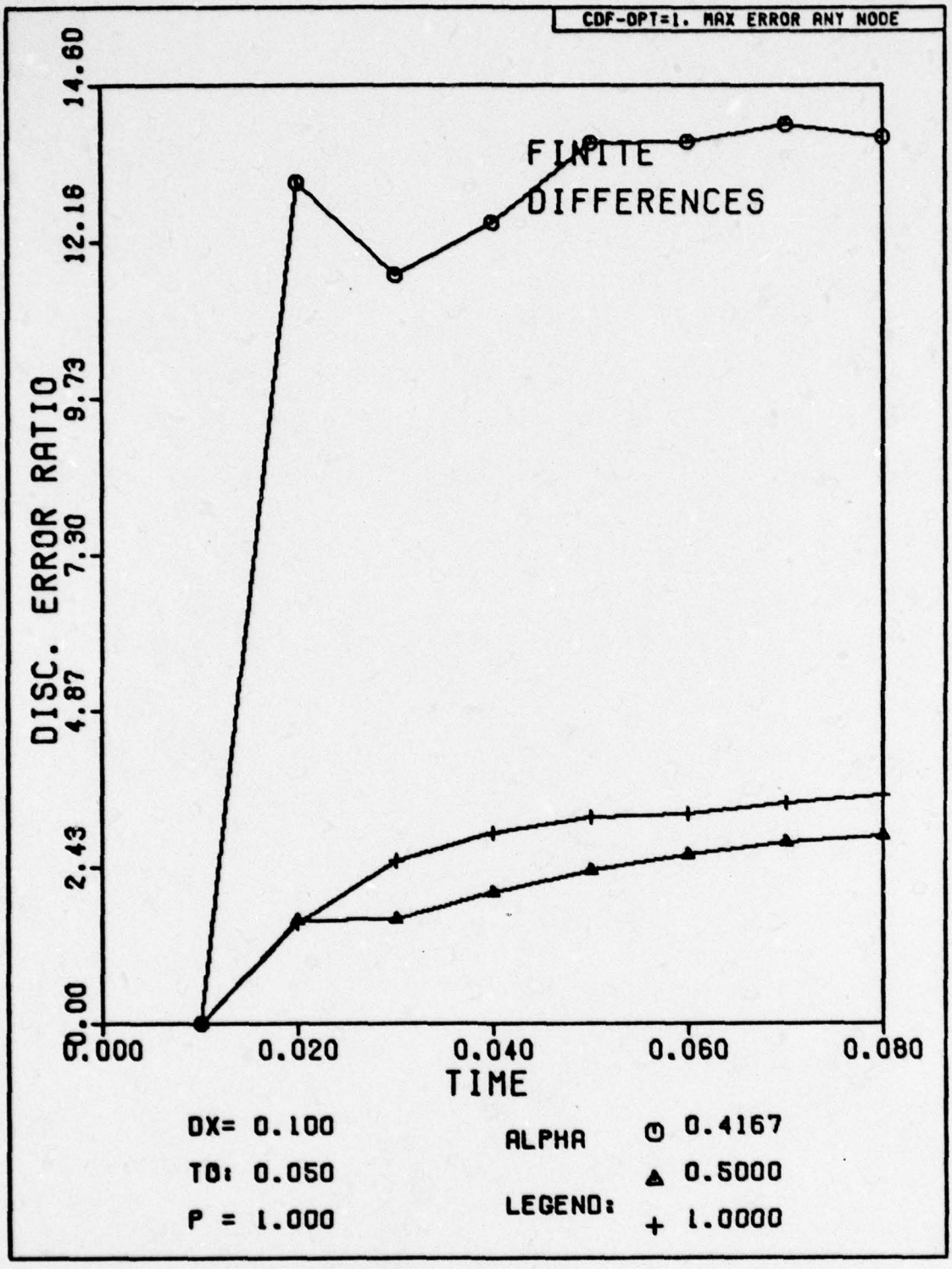


Fig. H-150. Discretization Error Ratio Versus Time for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

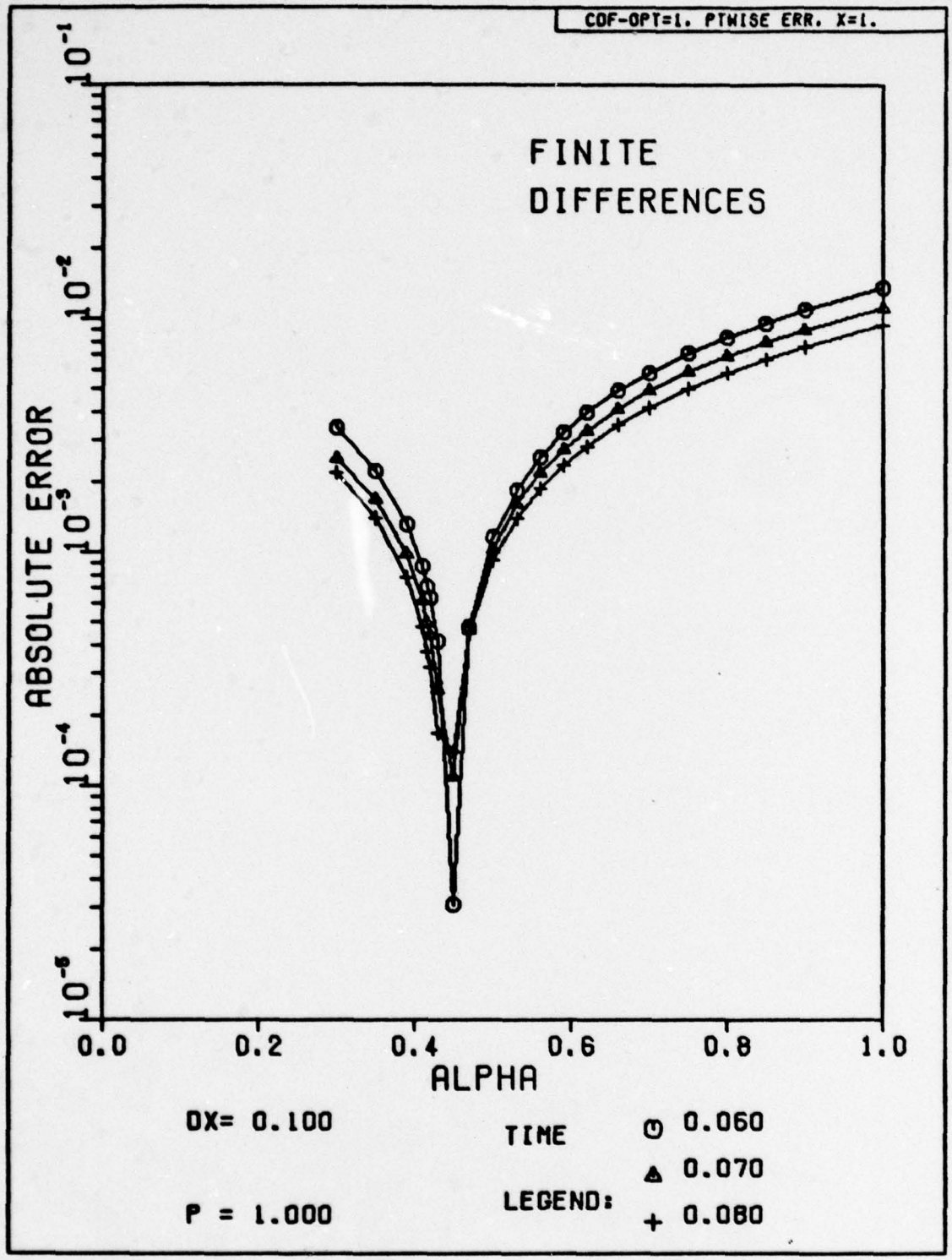


Fig. H-151. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

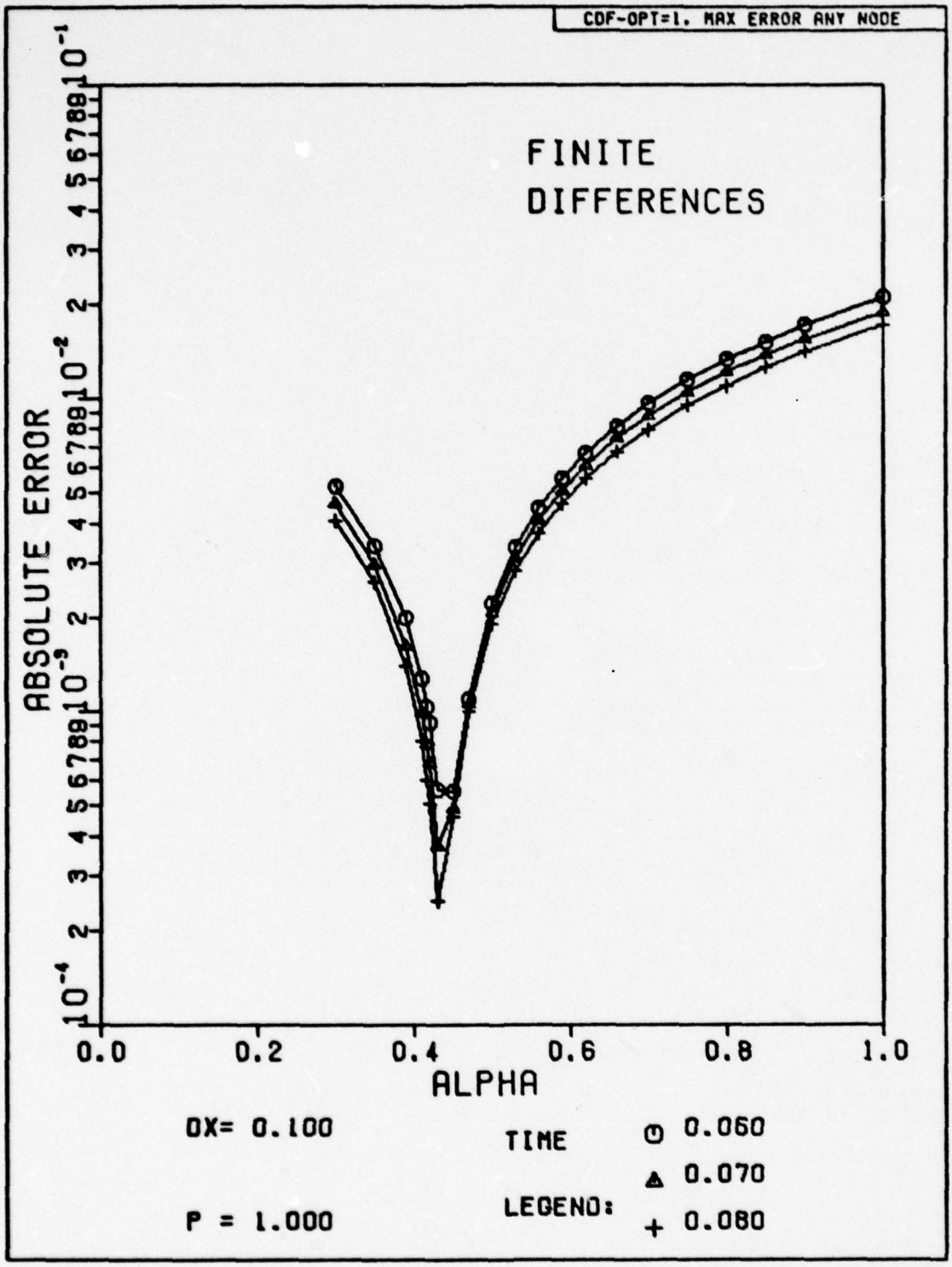


Fig. H-152. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

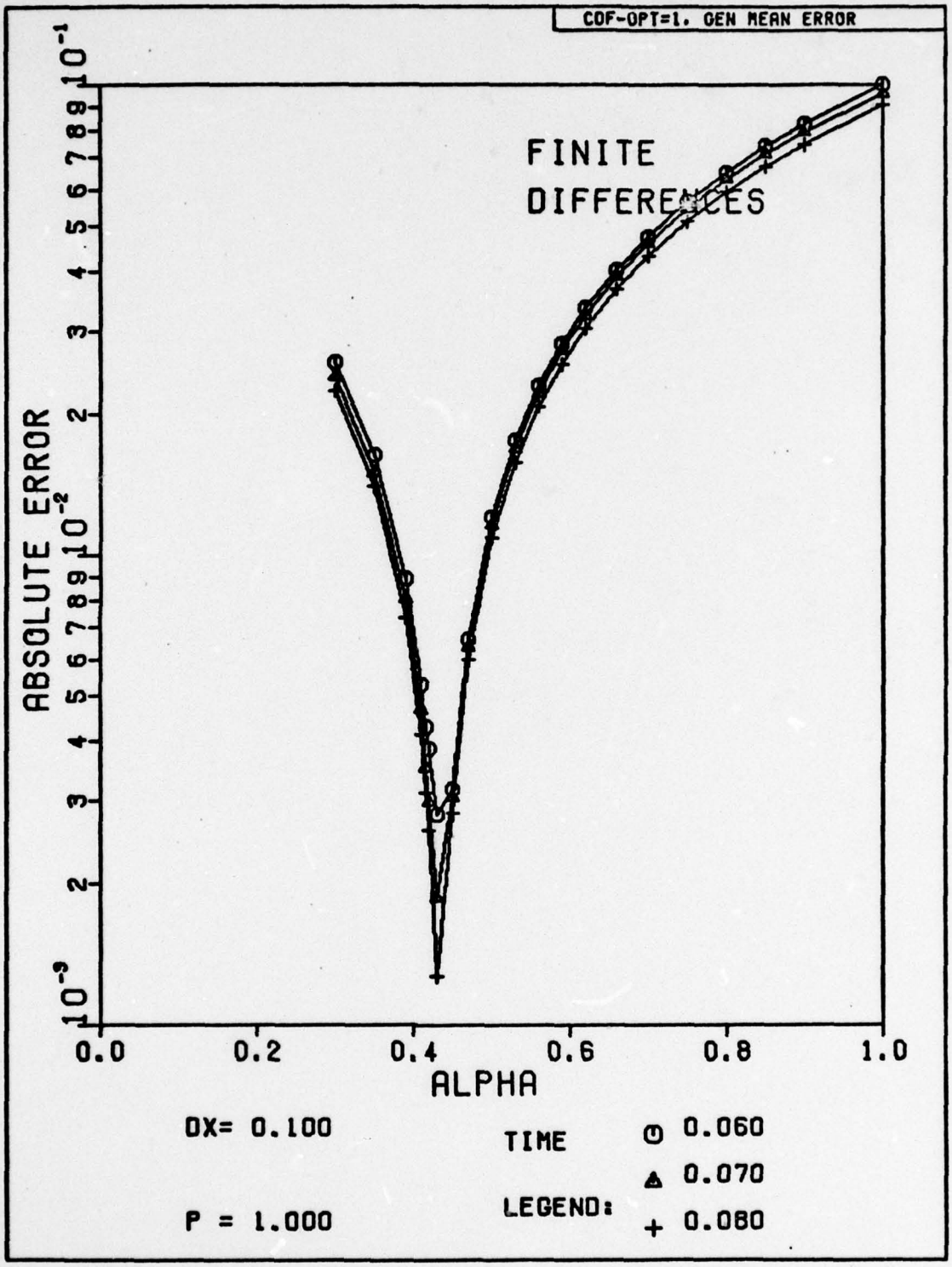


Fig. H-153. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

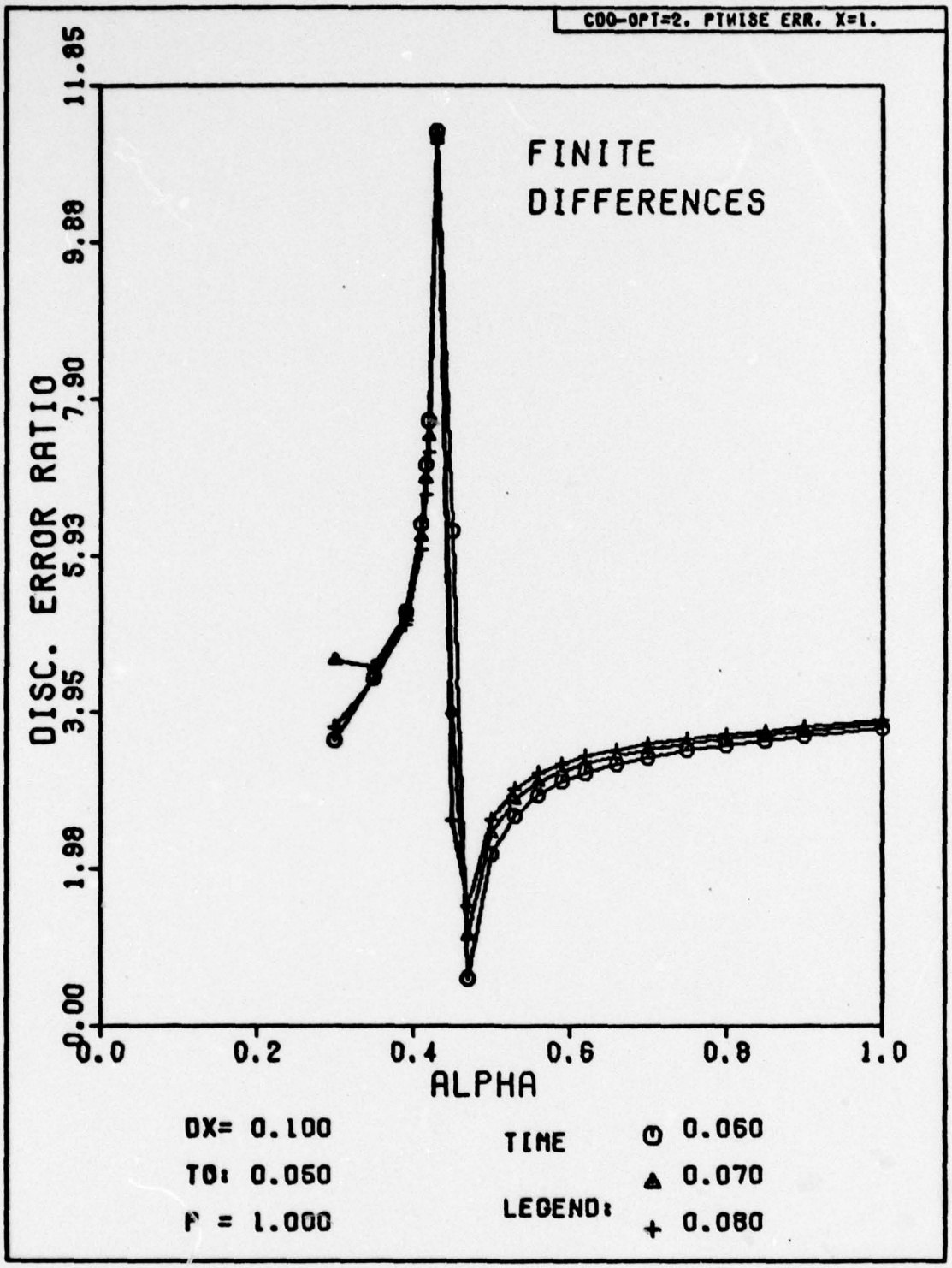


Fig. H-154. Discretization Error Ratio Versus Alpha for Problem Two. The space-time grid has been subdivided for the first time step to obtain a more accurate solution.

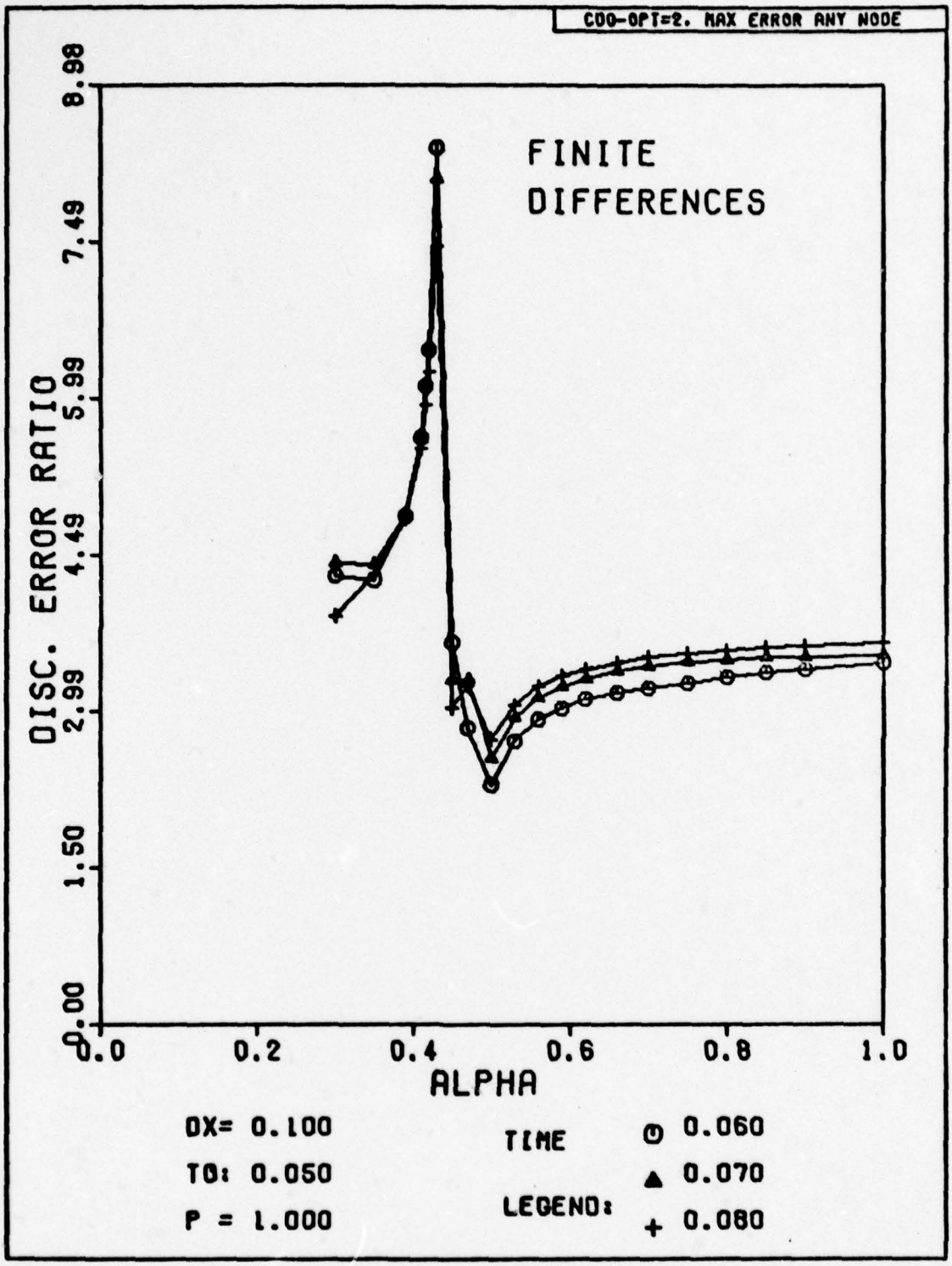


Fig. H-155. Discretization Error Ratio Versus Alpha for Problem Two. The space-time grid has been subdivided for the first time step to obtain a more accurate solution.

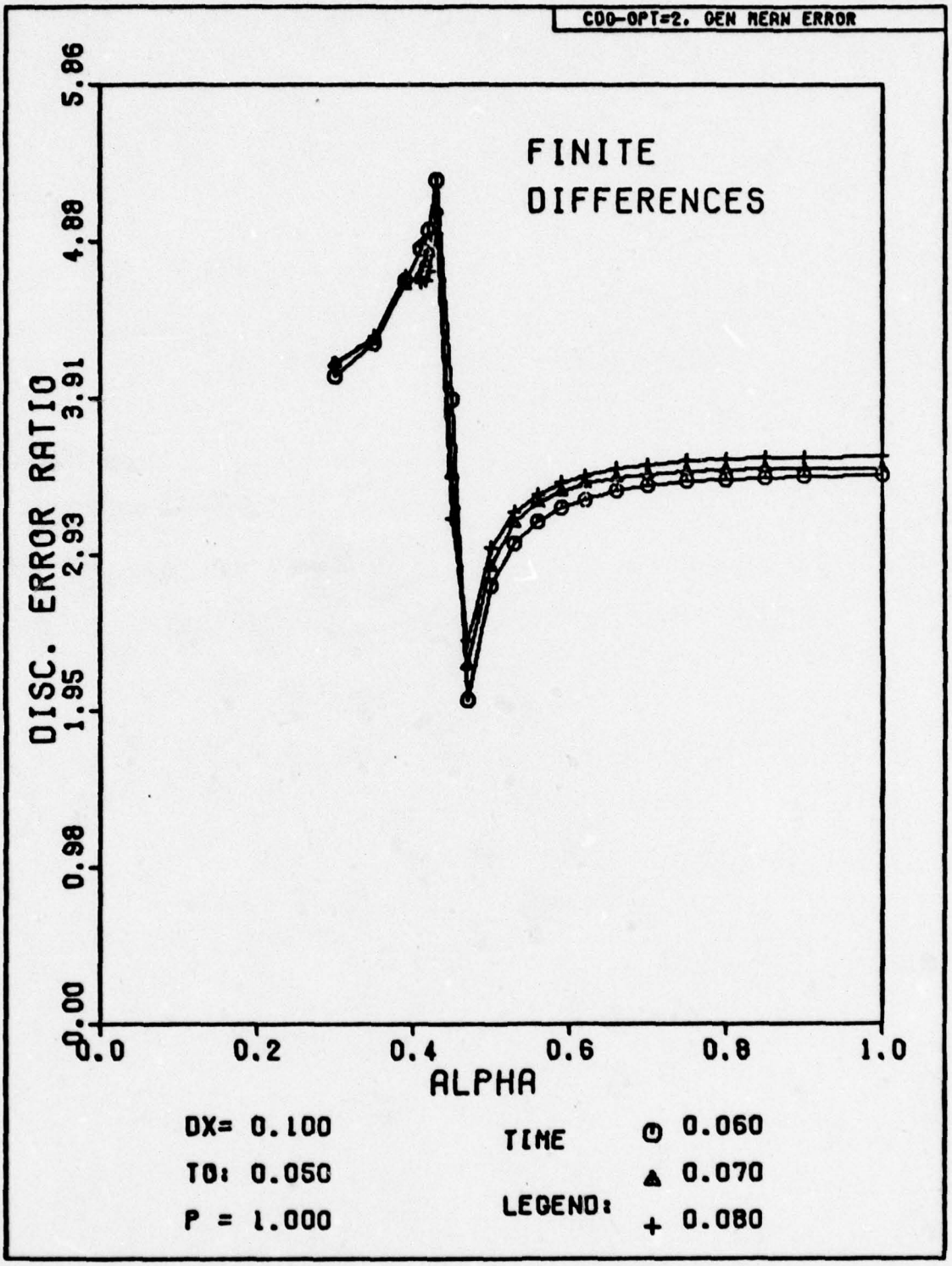


Fig. H-156. Discretization Error Ratio Versus Alpha for Problem Two. The space-time grid has been subdivided for the first time step to obtain a more accurate solution.

Section IV

The Results for the Secondary Problem Using Finite-Elements

This section shows the graphical results for the solution of the secondary problem by the method of finite-elements. The following key shows which options are included in this set of graphs.

Table H-IV

Key to the Plots in Section IV

Run Identifier	Fourier Modulus (p)	Option Number
CES	1.0	0
CET	1.0	1
CEU	1.0	2

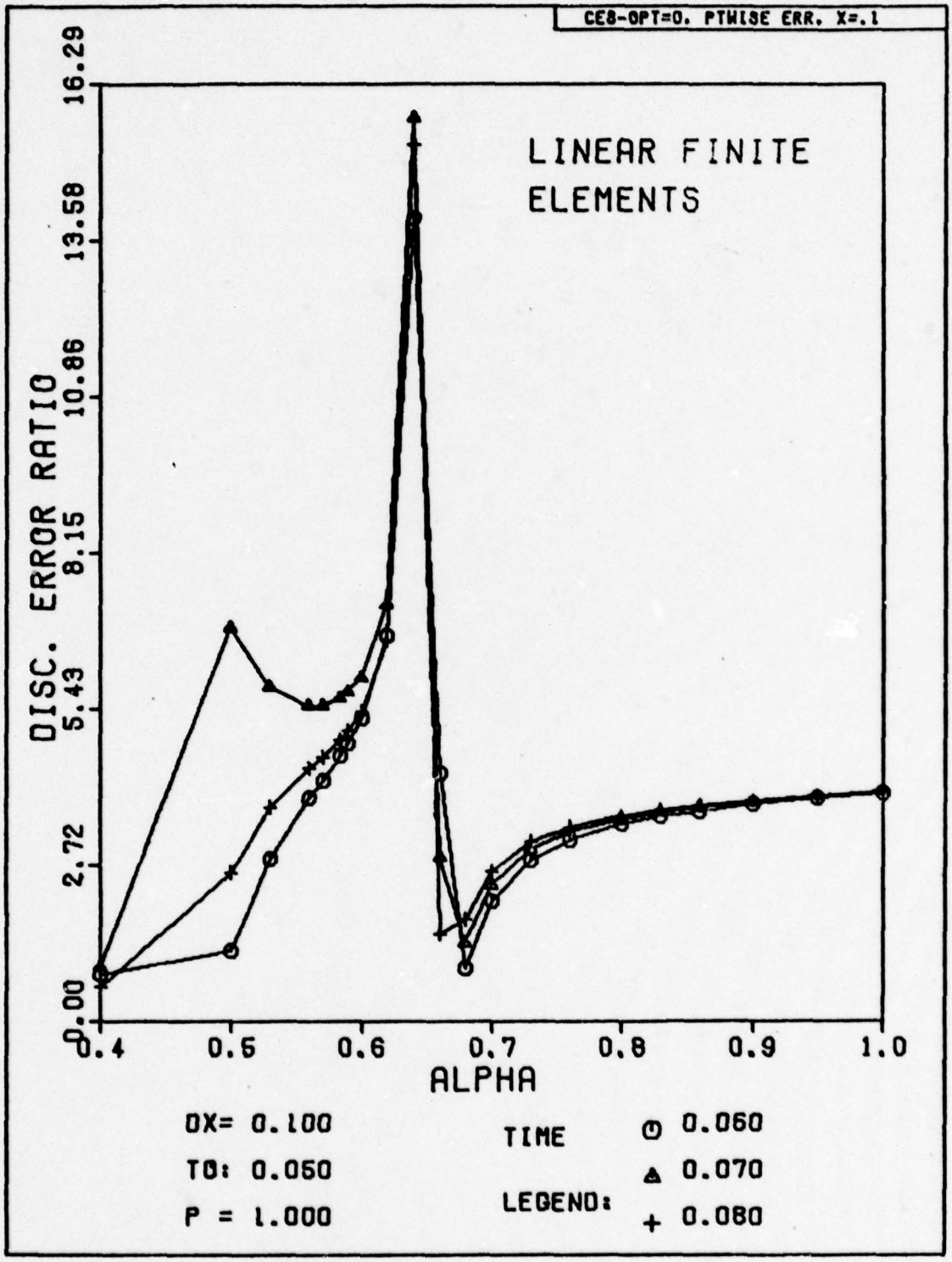


Fig. H-157. Discretization Error Ratio Versus Alpha for Problem Two.

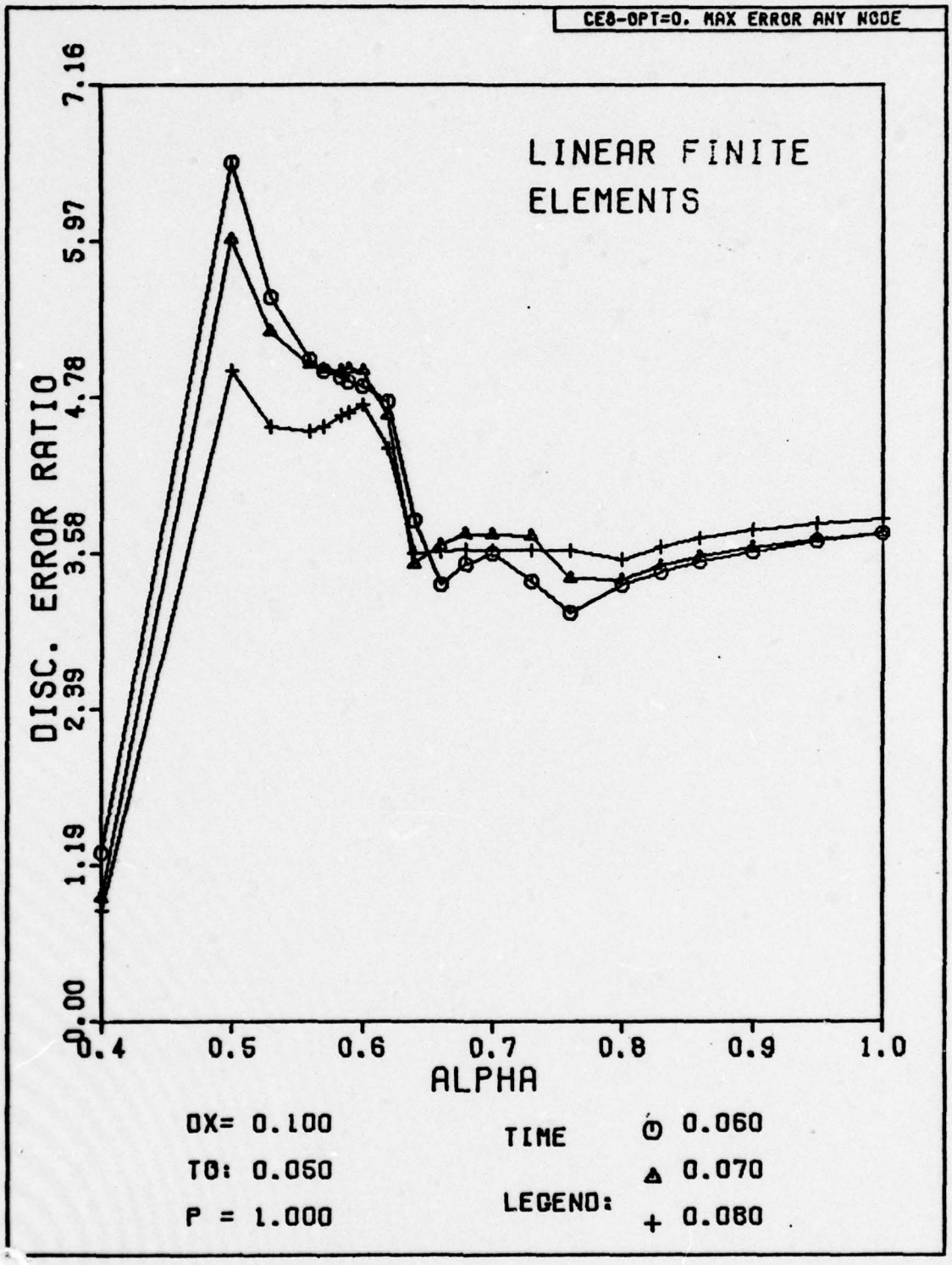


Fig. H-158. Discretization Error Ratio Versus Alpha for Problem Two.

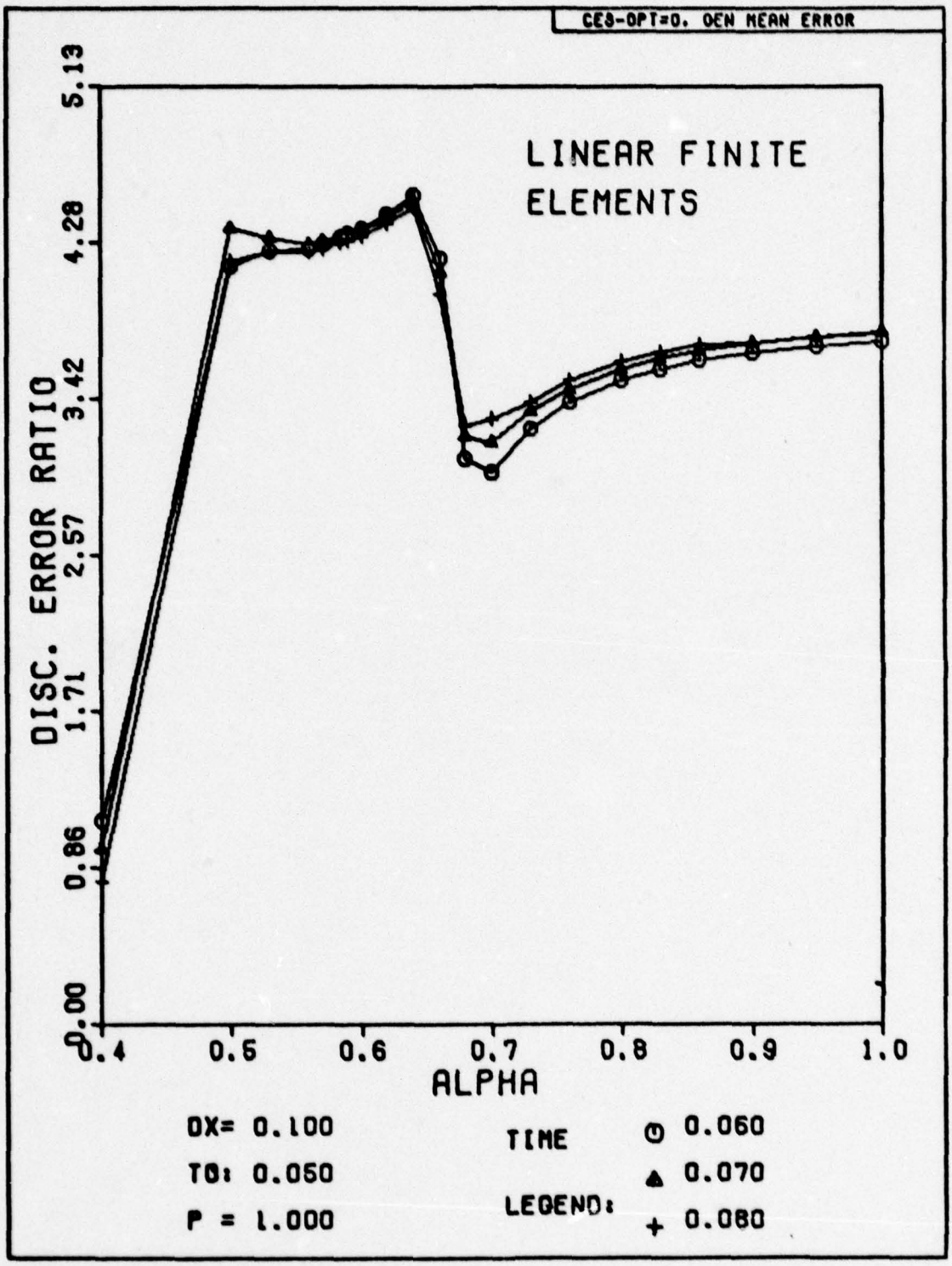


Fig. H-159. Discretization Error Ratio Versus Alpha for Problem Two.

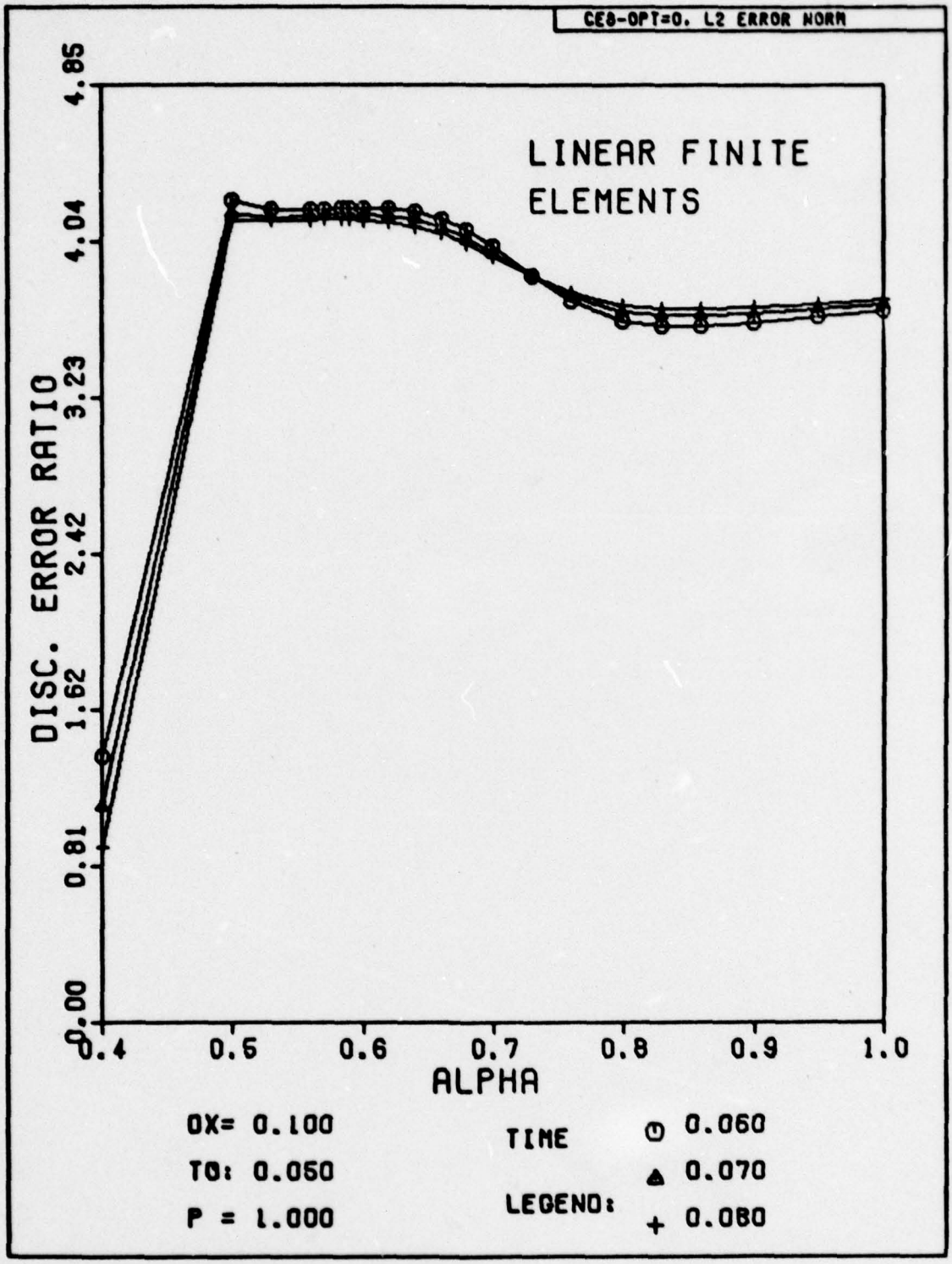


Fig. H-160. Discretization Error Ratio Versus Alpha for Problem Two.

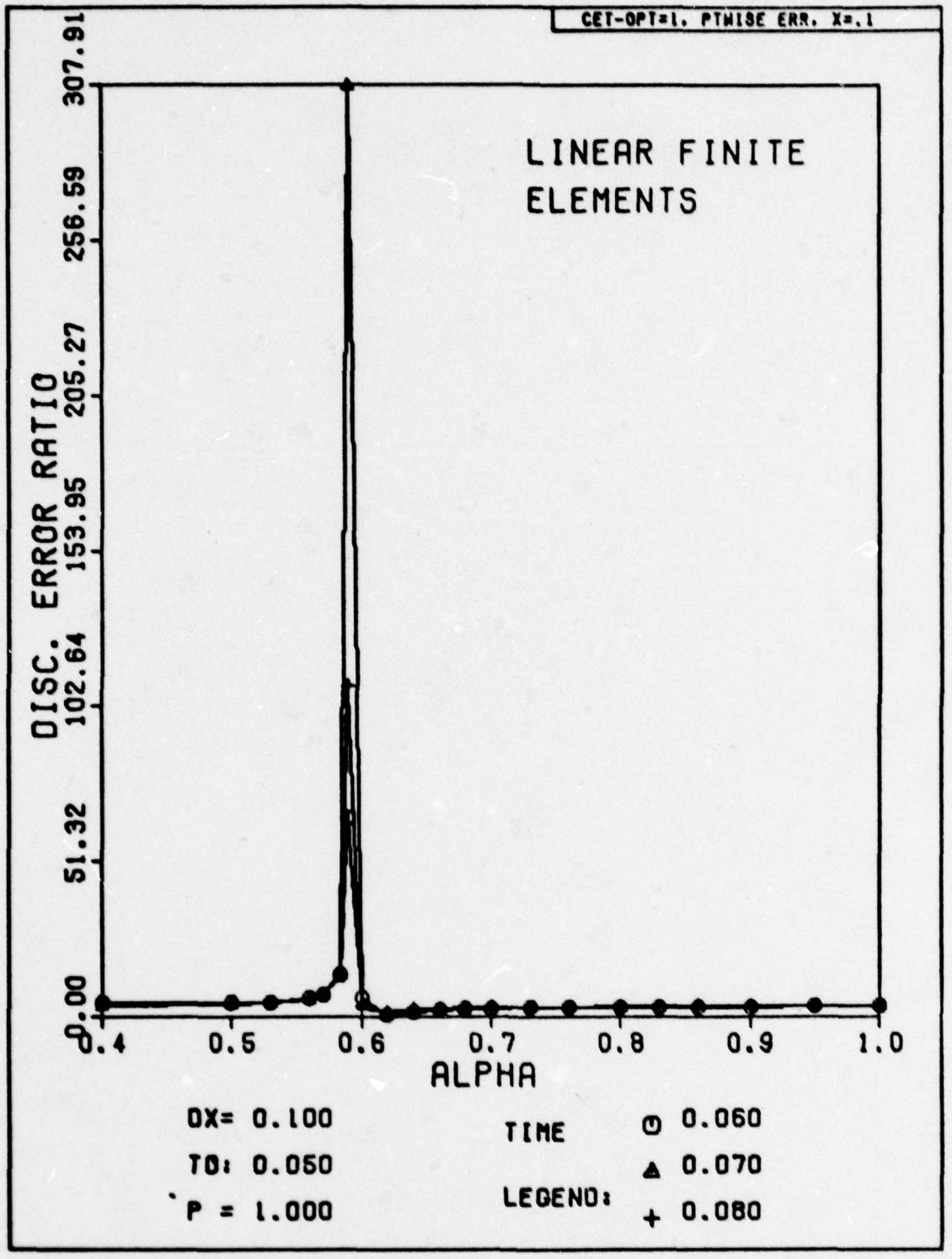


Fig. H-161. Discretization Error Ratio Versus Alpha for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

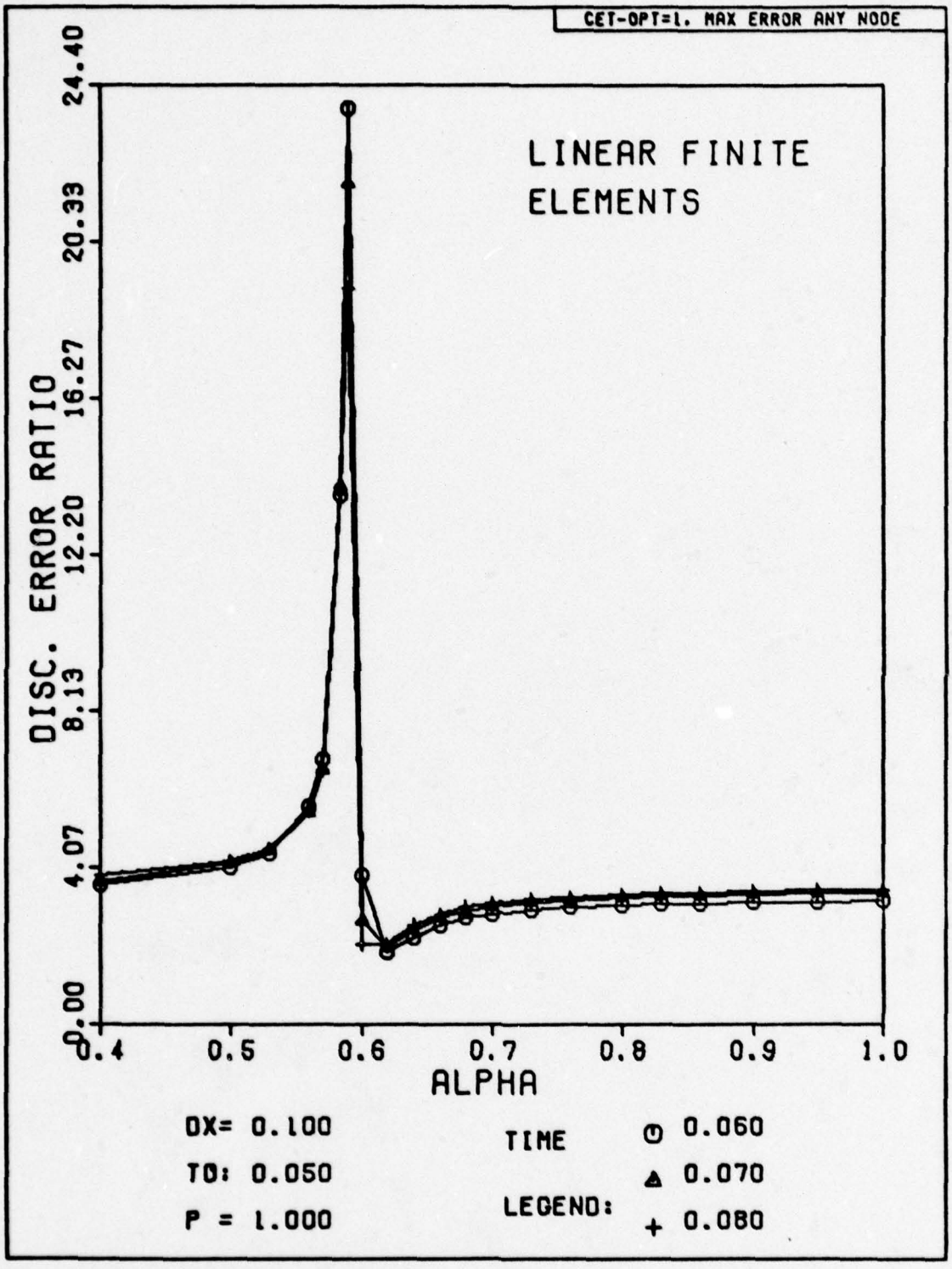


Fig. H-162. Discretization Error Ratio Versus Alpha for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

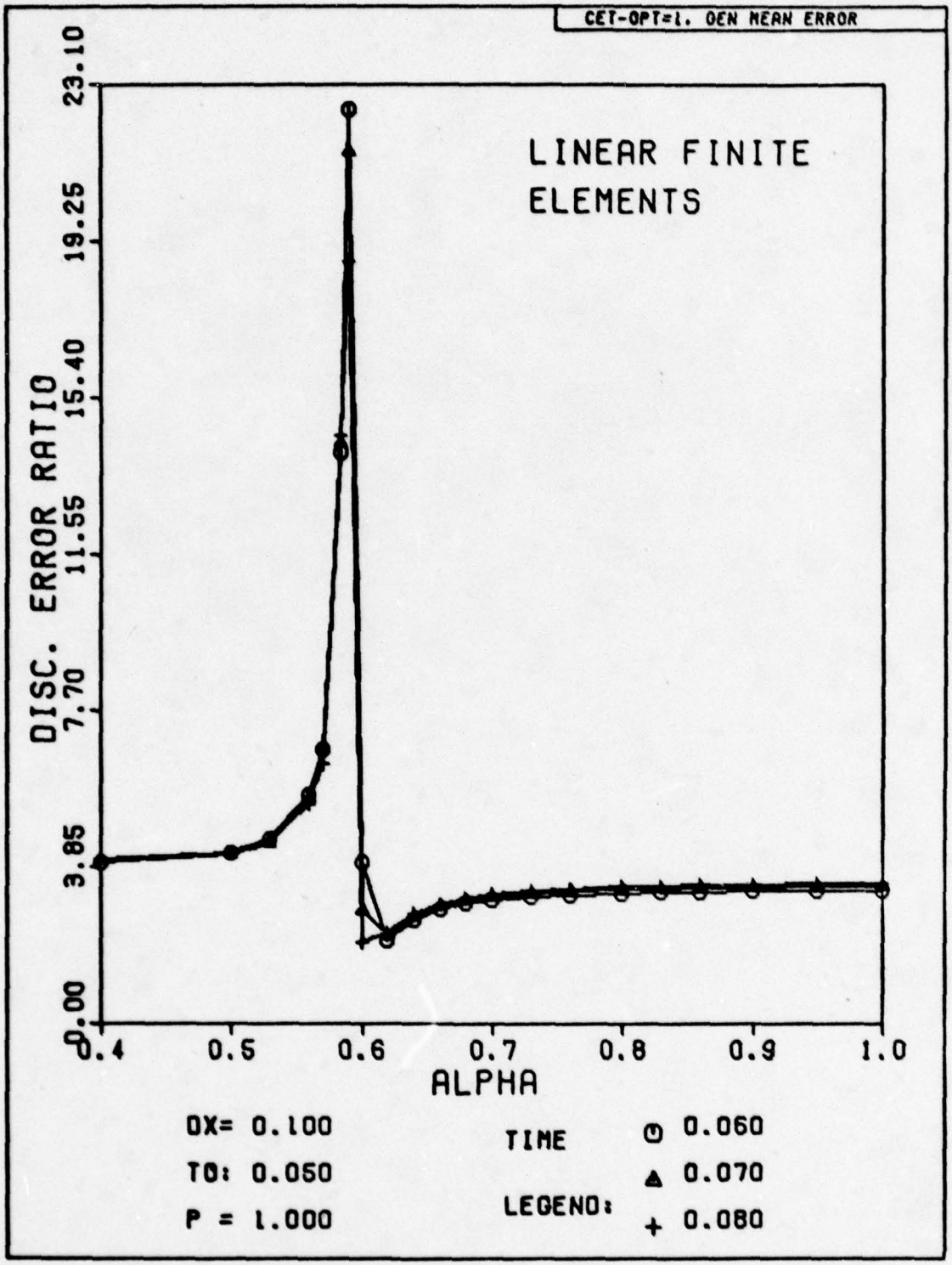


Fig. H-163. Discretization Error Ratio Versus Alpha for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

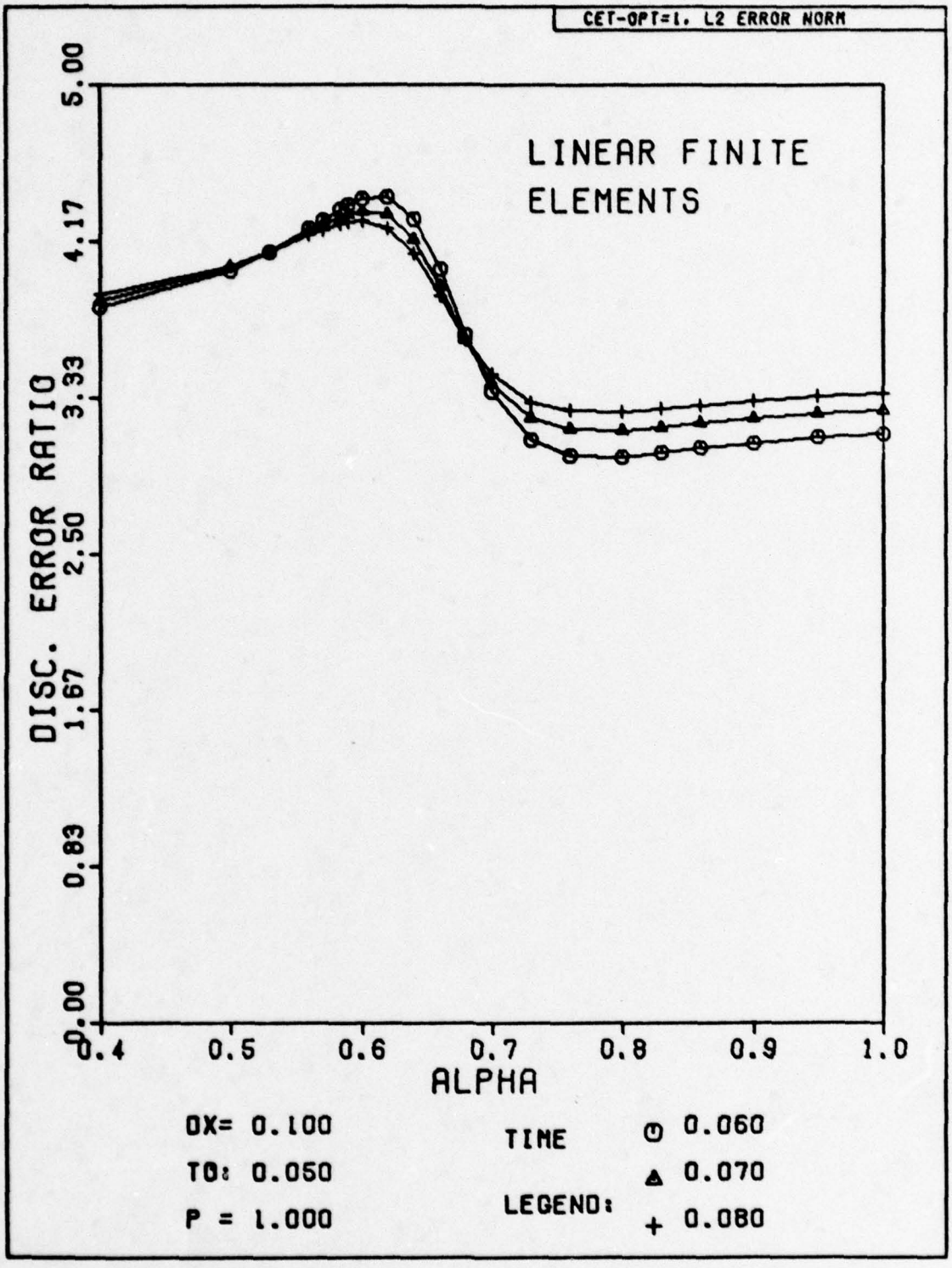


Fig. H-164. Discretization Error Ratio Versus Alpha for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

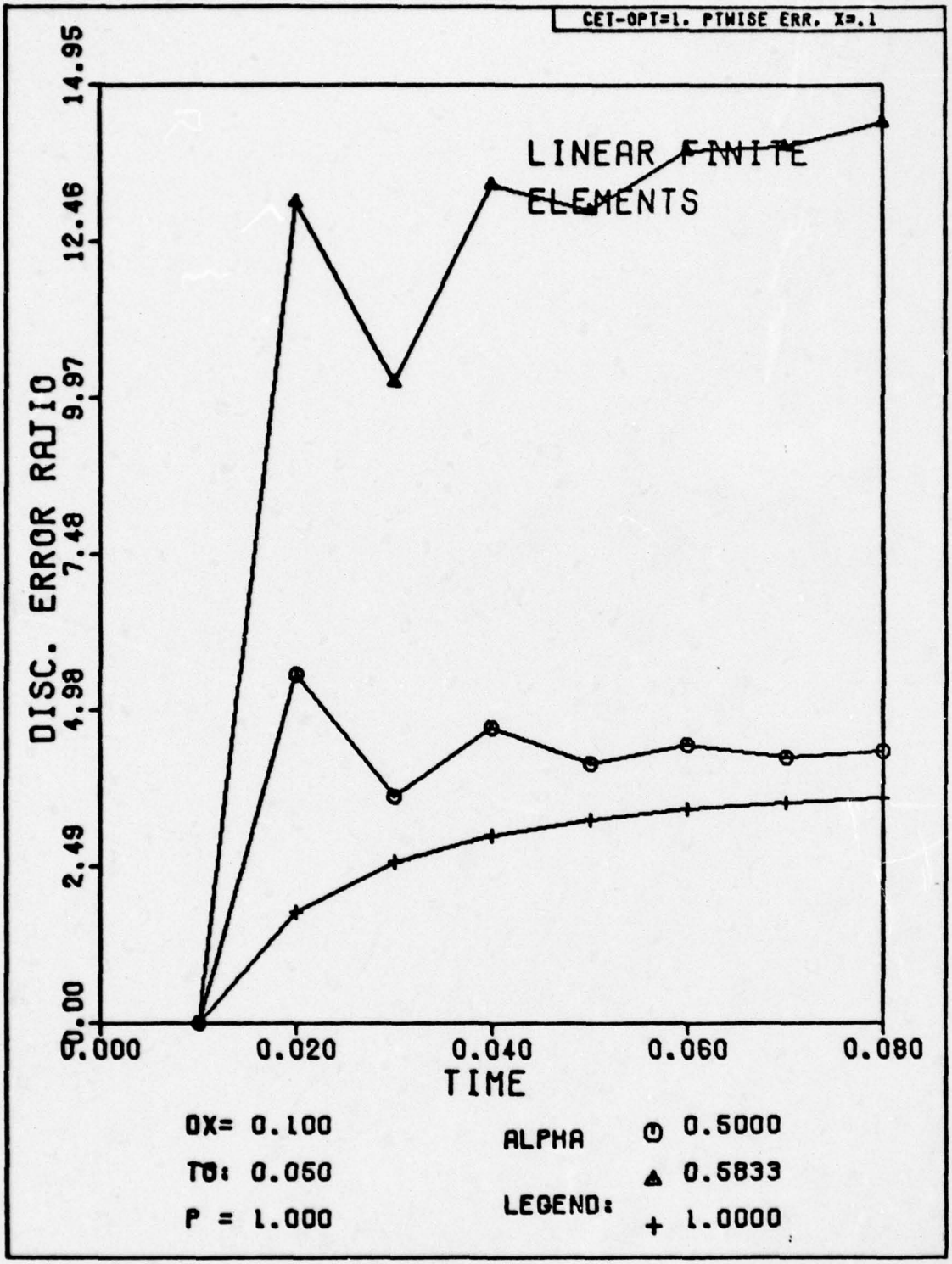


Fig. H-165. Discretization Error Ratio Versus Time for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

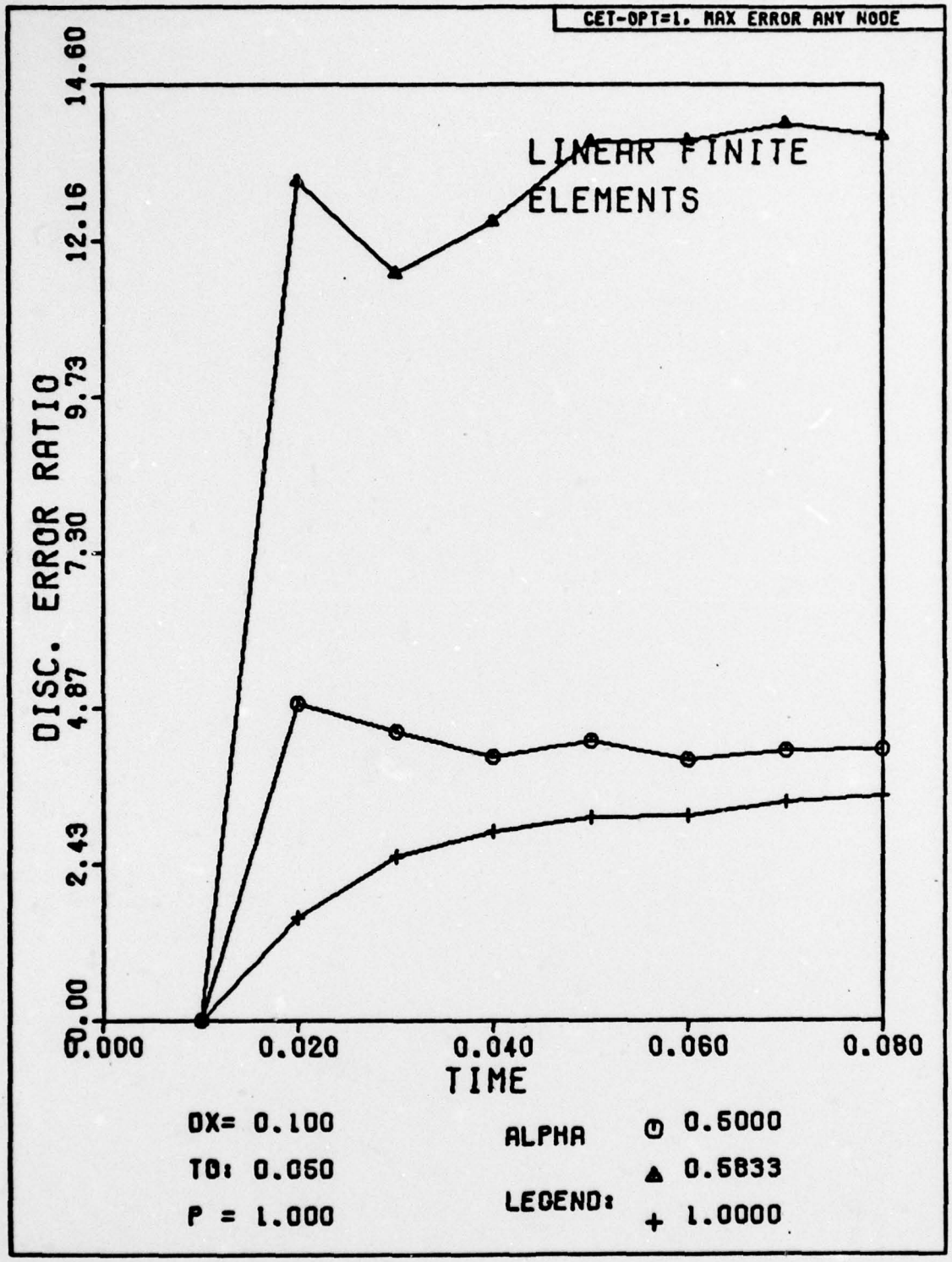


Fig. H-166. Discretization Error Ratio Versus Time for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

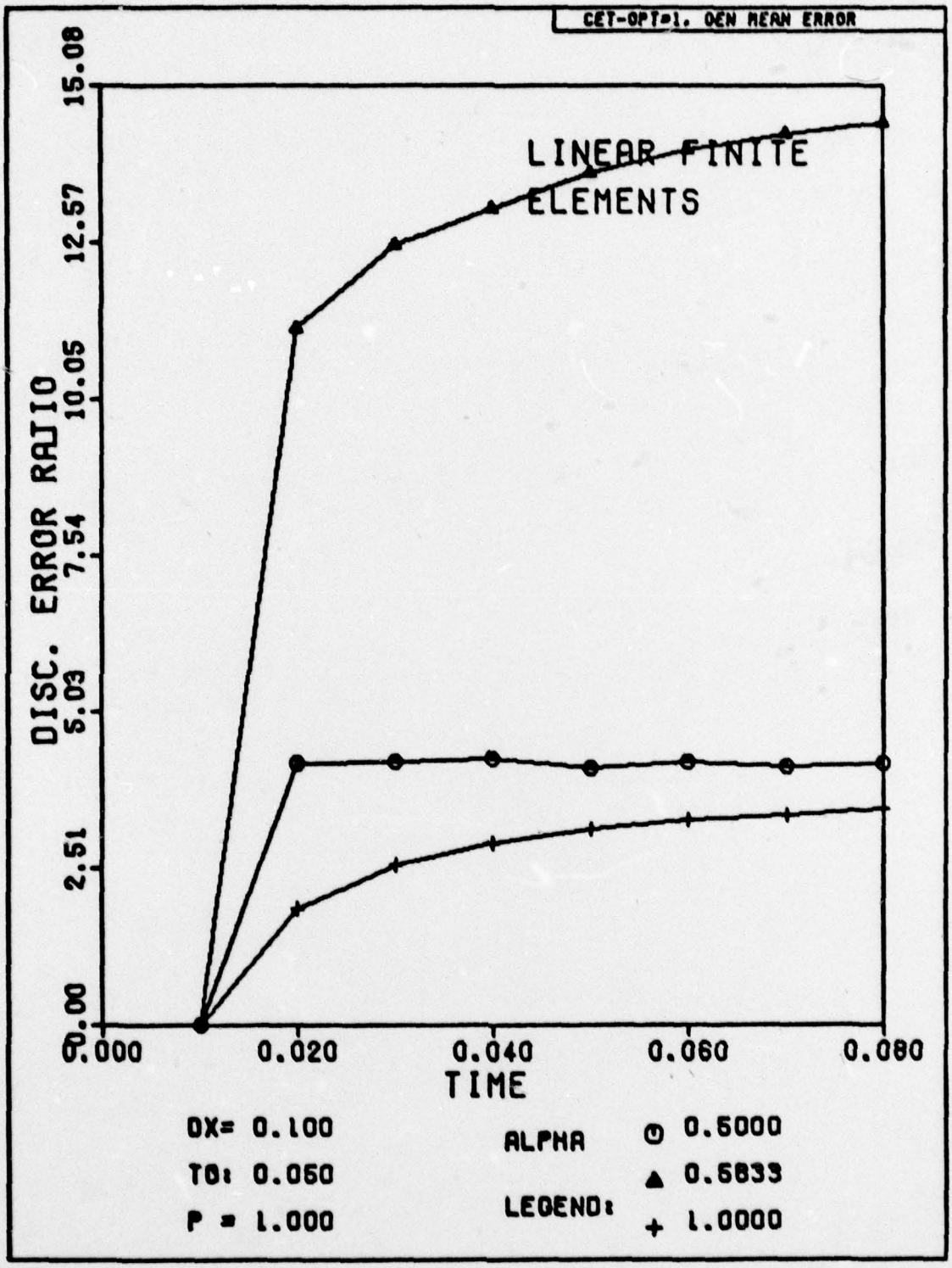


Fig. H-167. Discretization Error Ratio Versus Time for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

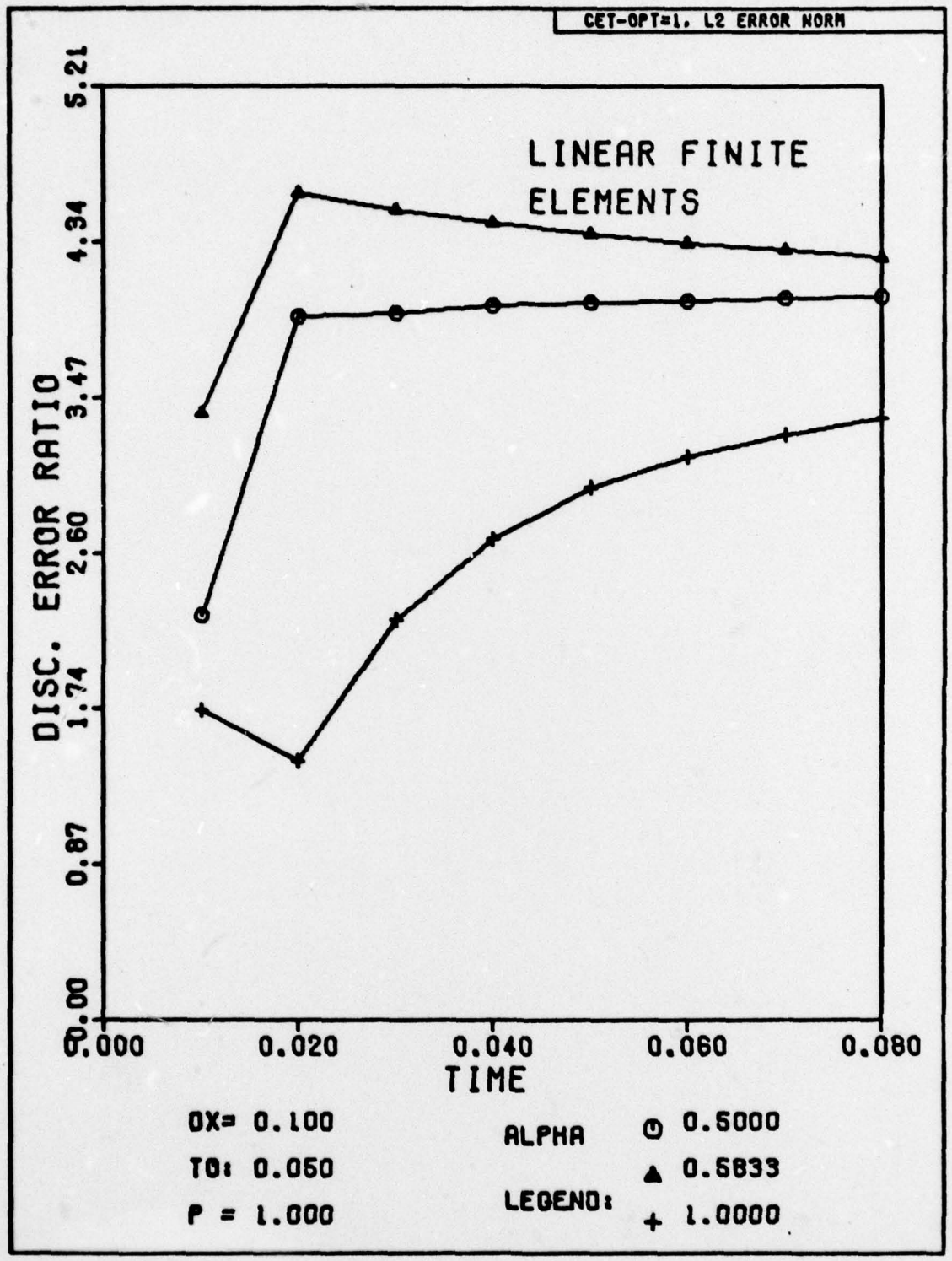


Fig. H-168. Discretization Error Ratio Versus Time for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

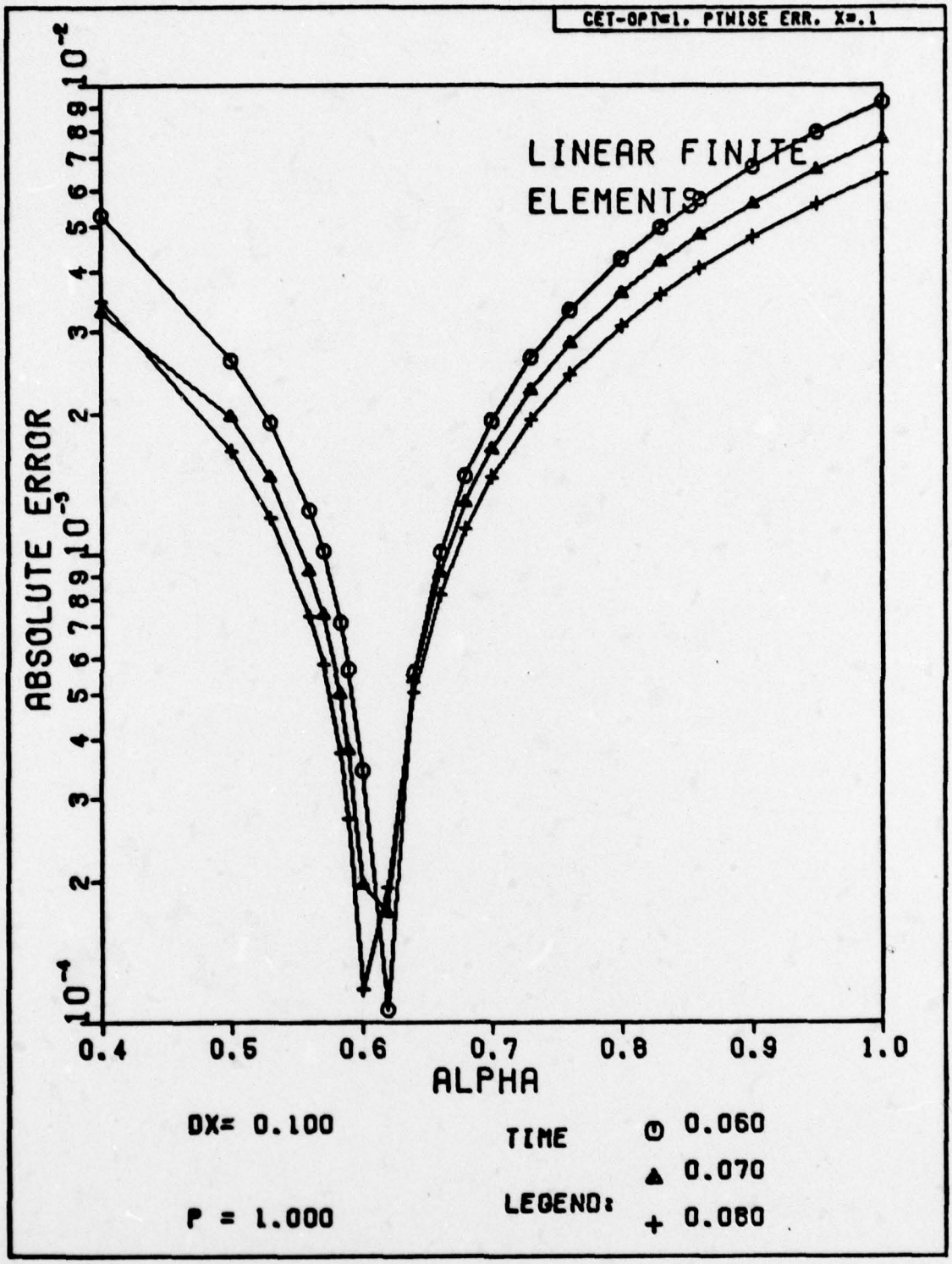


Fig. H-169. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

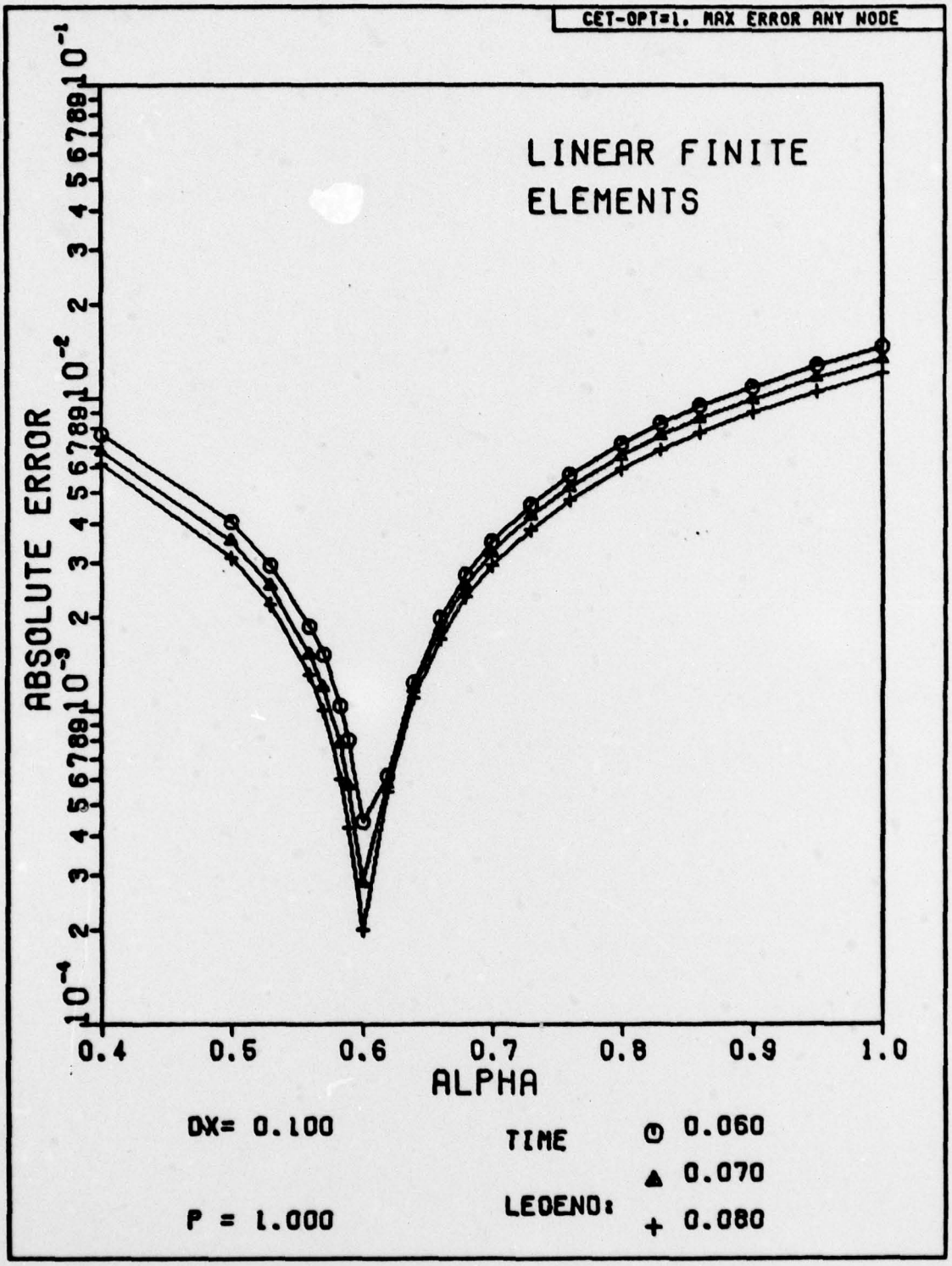


Fig. H-170. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

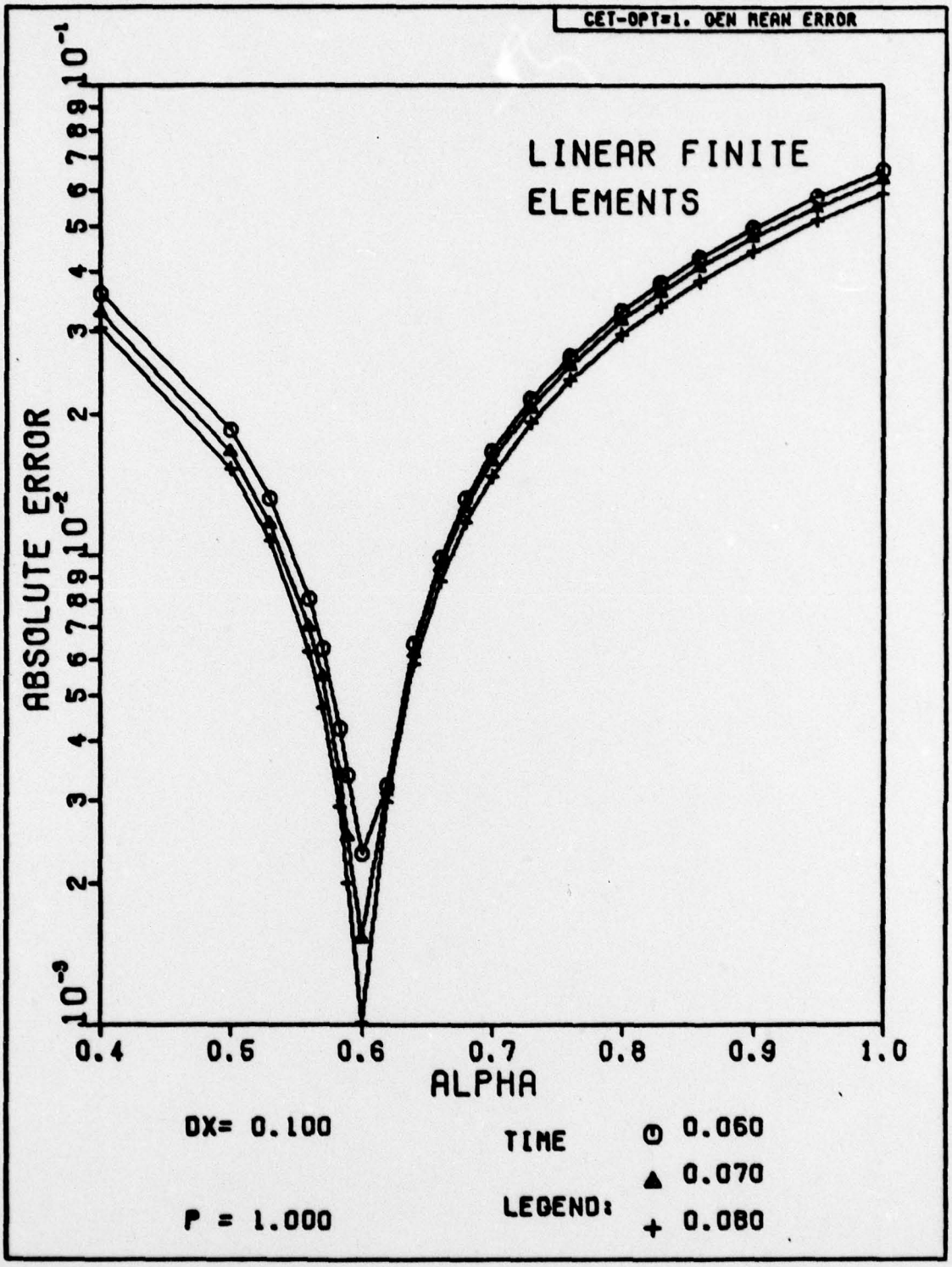


Fig. H-171. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

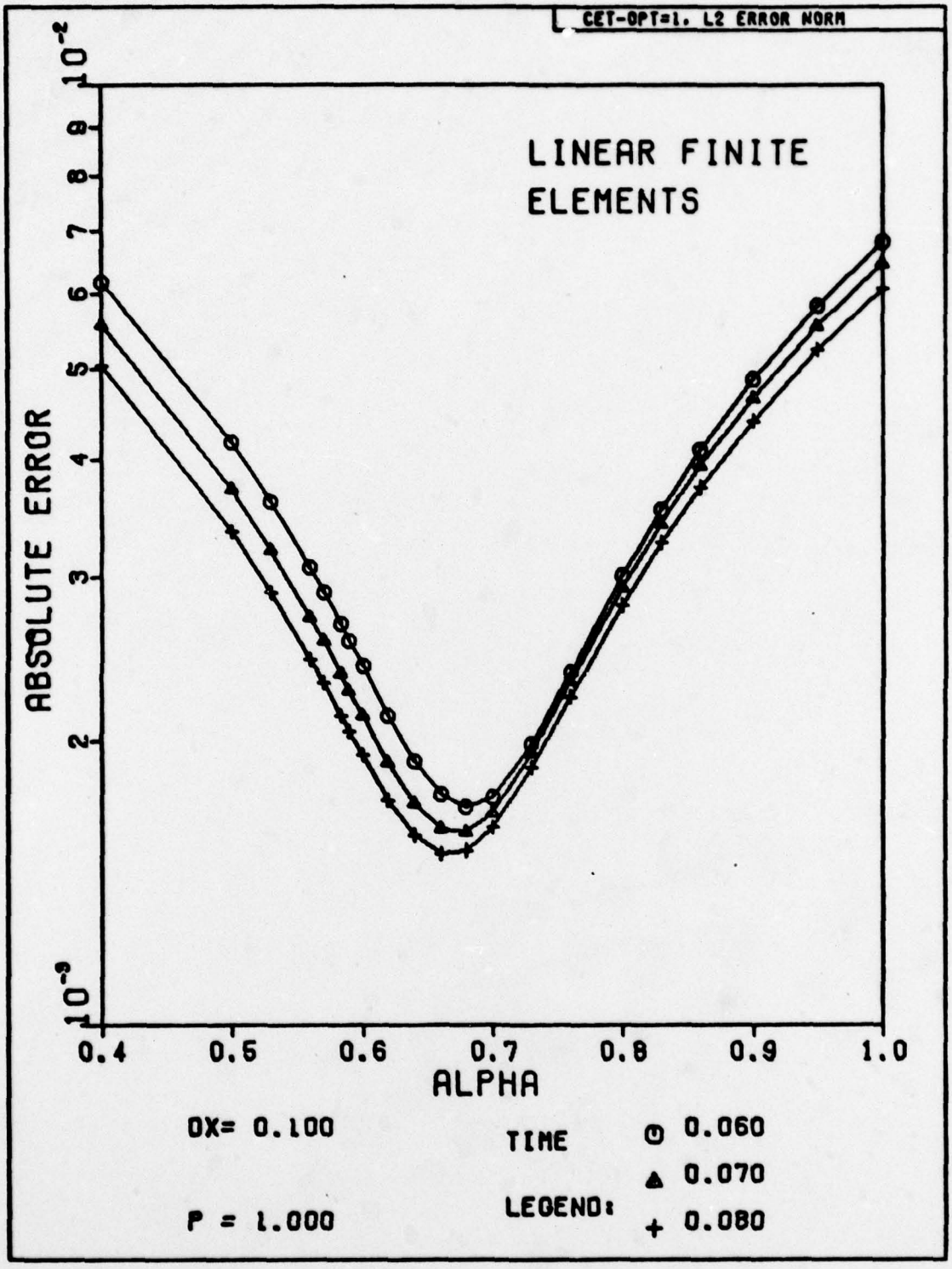


Fig. H-172. The Absolute Magnitude of the Discretization Error Versus Alpha for Problem Two. The exact solution has been substituted for the numerical solution at the first time step.

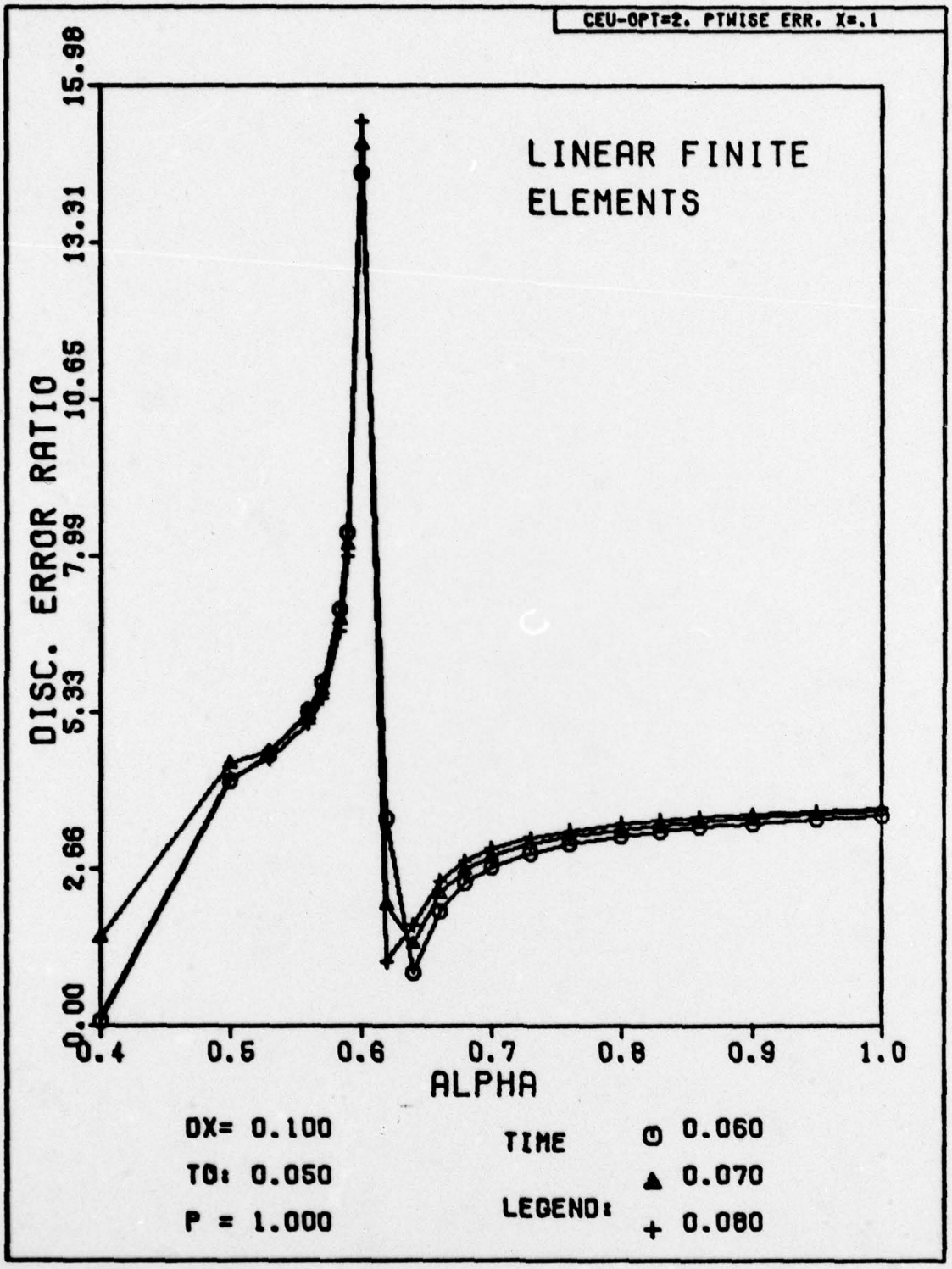


Fig. H-173. Discretization Error Ratio Versus Alpha for Problem Two. The space-time grid has been subdivided for the first time step to obtain a more accurate solution.

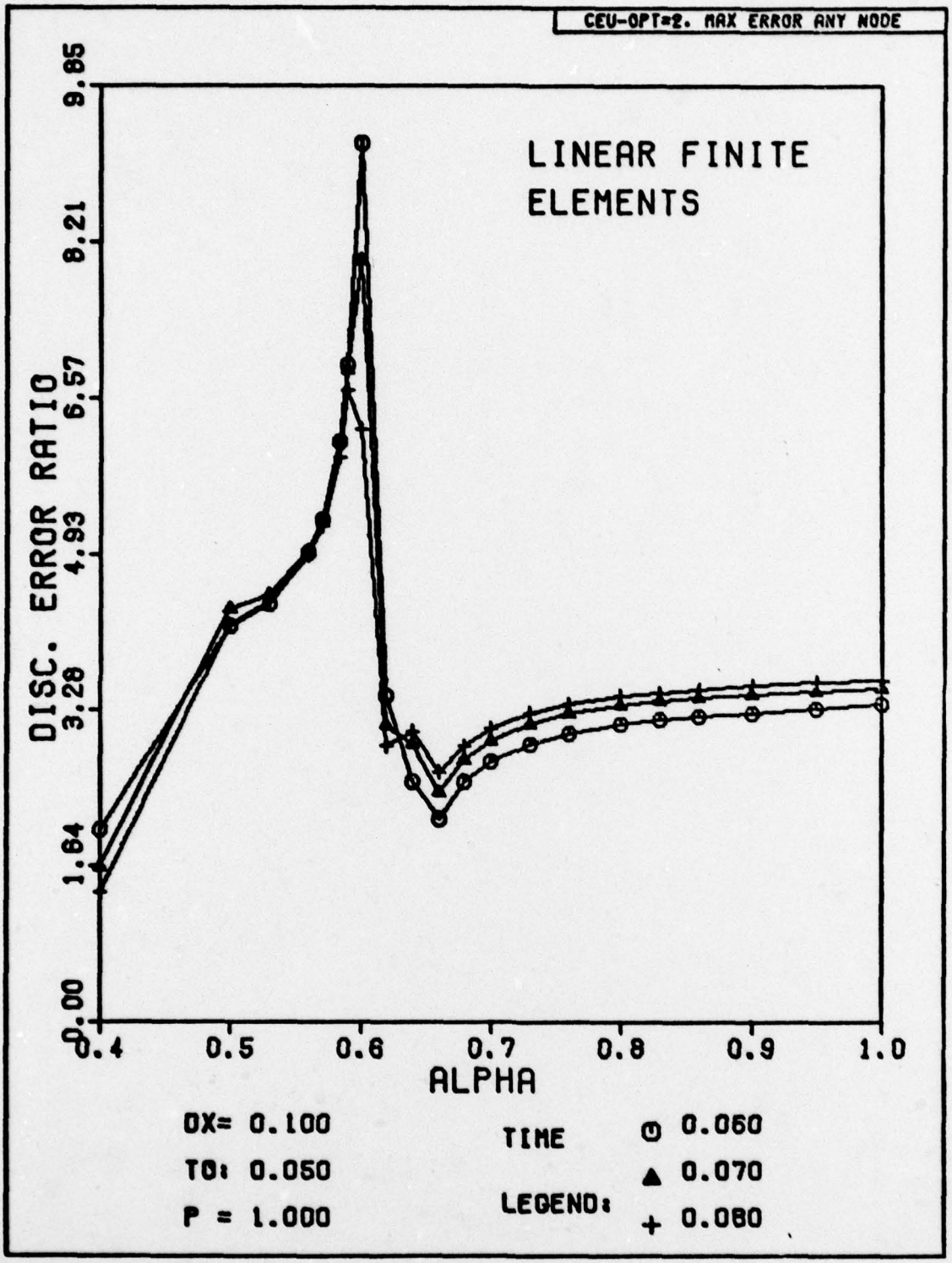


Fig. H-174. Discretization Error Ratio Versus Alpha for Problem Two. The space-time grid has been subdivided for the first time step to obtain a more accurate solution.

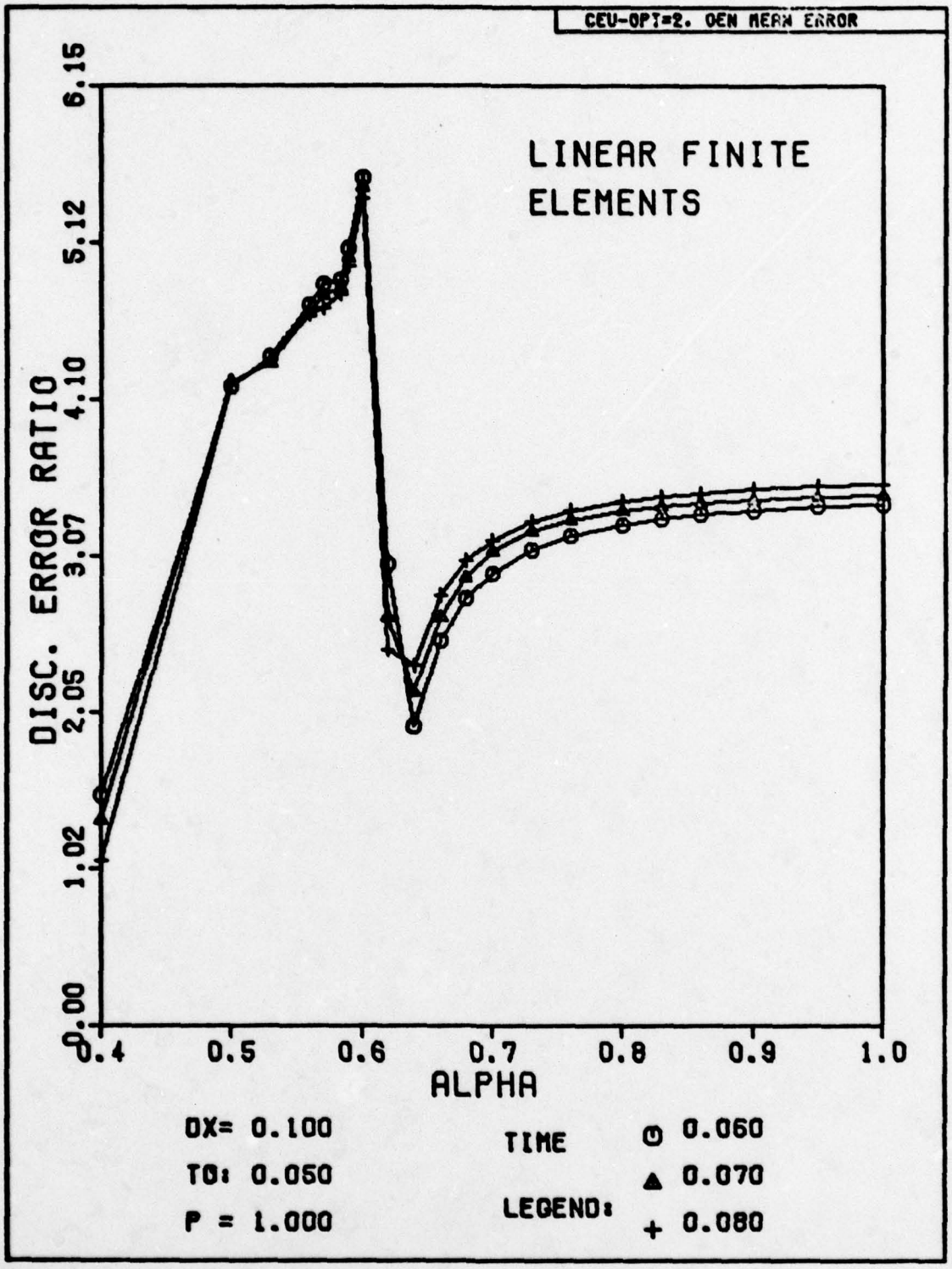


Fig. H-175. Discretization Error Ratio Versus Alpha for Problem Two. The space-time grid has been subdivided for the first time step to obtain a more accurate solution.

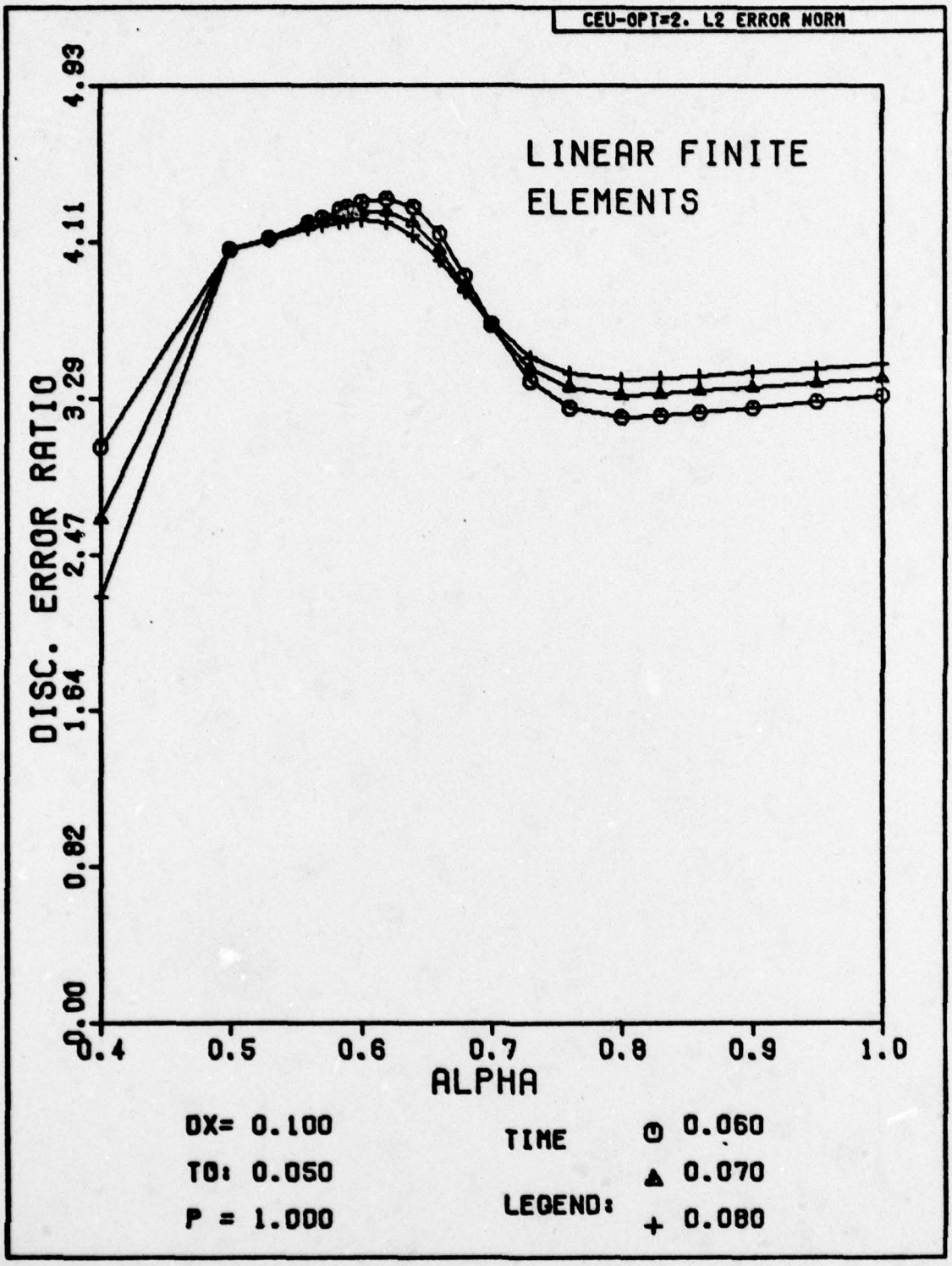


Fig. H-176. Discretization Error Ratio Versus Alpha for Problem Two. The space-time grid has been subdivided for the first time step to obtain a more accurate solution.

Vita

Charles R. Martin was born on 25 September, 1950 in Wiesbaden, Germany, the son of Clyde J. Martin and Lillian R. Martin. Upon completion of high school at Watauga High School, Boone, North Carolina in 1968, he entered North Carolina State University at Raleigh, Raleigh, North Carolina, where he was enrolled in a cooperative education program in Nuclear Engineering. His cooperative work was performed with the Nuclear Division of Duke Power Company, Charlotte, North Carolina. In May of 1973, he was graduated with Honors as a Bachelor of Science in Nuclear Engineering. Concurrently, he completed the Reserve Officers Training Corps program and was awarded the designation of Distinguished Graduate. In November of 1973, he completed Missile Combat Crew Operational Readiness Training at Vandenburg AFB, California as a Distinguished Graduate of that program. For the next three years, he served as a Deputy Missile Combat Crew Commander and Missile Launch Procedures Instructor at Malmstrom AFB, Montana. In September 1976 he entered the Air Force Institute of Technology, Wright-Patterson AFB, Ohio.

Permanent Address: 5257 Vann Street
Raleigh, North Carolina

This thesis was typed by Sharon Flores.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: AFIT/GNE/PH/78M-6 ✓	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): 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	5. TYPE OF REPORT & PERIOD COVERED: MS THESIS 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): MARTIN, CHARLES R.	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Air Force Institute of Technology (AFIT-EN) Wright-Patterson AFB, Ohio 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: 17 March 1978 13. NUMBER OF PAGES: 317
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)

18. SUPPLEMENTARY NOTES

Approved for public release;
IAW AFR 190-17.

Jerral F. Guess

JERRAL F. GUESS, Captain, USAF
Director of Information

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 transient heat conduction equation, with Dirichlet and Neumann boundary conditions, is solved by the methods of finite-differences and finite-elements, and the numerical solutions are investigated with respect to accuracy and stability. A general six point finite-difference expression is used for which there exists a high order accurate modification. The finite-element method used is based on a stationary variational principle. Several methods for treating accuracy and convergence problems which result from a discontinuity in the initial condition are investigated. The Crank-Nicolson method

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

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Next Page

is a special case of both the finite-difference and finite-element methods. The finite-difference version of the Crank-Nicolson method is shown to be more accurate than the finite-element version, especially when a discontinuity exists between the initial condition and the boundary conditions. The high order accurate schemes for both finite-differences and finite-elements are shown to be equivalent for the case of linear elements. Some of the results suggest the possibility of finding a finite-element scheme which is highly accurate in a mean square sense over the entire solution domain.