ANALYSIS OF NUMERICAL SOLUTIONS FOR THE M.I.T. FLIGHT SIMULATOR

by

HIDEO MORI

B.S., University of Utah
(1950)

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
(1952)

Signature of Author Signature redacted
Department of Electrical Engineering, August 22, 1952

Certified by Signature redacted
Thesis Supervisor

Signature redacted
Chairman, Departmental Committee on Graduate Students

ANALYSIS OF NUMERICAL SOLUTIONS FOR THE M.I.T. FLIGHT SIMULATOR

by

HIDEO MORI

Submitted to the Department of Electrical Engineering on August 22, 1952 in partial fulfillment of the requirements for the degree of Master of Science.

ABSTRACT

The high level checking of large-scale analog computers can be accomplished by comparing a numerically calculated solution of the equations simulated with the solution obtained from the analog computer. A method of ascertaining the adequacy of the numerical solutions for the purpose of checking an analog computer is illustrated.

The errors committed at each interval of tabulation are examined and the propagated effect of these errors are found. The effects of the errors are assumed linear and superposition of the effects of the separate errors gives the total propagated error.

The propagated effect is found by solving a set of variational equations associated with the original differential equations. The solution of the variational equations are simplified by sectioning the solution such that the solution of a set of constant-coefficient linear differential equations is required for these sections.

The application of the method to the propagation of inaccuracies in analog computing elements is discussed as well as a comparison of alternate approaches to obtaining a numerical solution

Thesis Supervisor: Thomas Franklin Jones, Jr.

Title: Assistant Professor of Electrical Engineering

ACKNOWLEDGMENT

The author wishes to express his gratitude to Professor T. F. Jones, Jr., for supplying his time and knowledge during the supervision of this thesis, to the personnel of the Dynamic Analysis and Control Laboratory who made it possible to undertake this thesis, and to Miss Jeanne O'Brien for typing the manuscript.

TABLE OF CONTENTS

Chapter I

INTRODUCTION

The judicious use of large-scale analog computers requires some method of checking the accuracy of the solutions obtained from these computers. The M.I.T. Flight Simulator[1] uses the method of comparing a machine solution and a hand-calculated "check" solution. The effect of errors in the check solution, the comparison of various approaches to obtaining the check solution, and the comparison of two methods of error analysis are discussed in this thesis.

1.0. The Problem and the Objectives of this Thesis.

The M.I.T. Flight Simulator is an analog computer and as such it has many eccentricities. For example, the solution obtained from the computer may not be the solution of the desired problem because of errors in setup, gain errors in the amplifiers and other component errors, and calibration errors. Since no automatic indication that the solution is in error is possible, a method of ascertaining the accuracy of the computation of the desired problem is required before the solutions obtained from the computer can be accepted with any degree of certainty. The present method used to check the operation of the M.I.T. Flight Simulator consists of comparing the analog solution with a specific numerical solution of the desired problem. If an adequate check is achieved, then the computer is assumed to be set up properly and free of serious systematic errors. Since variations of the parameters of the problem normally do not seriously affect the accuracy of computation, solutions other than the

---

[1] Superscripts refer to references in Appendix C.

check solution are usually assumed to be correct. In other words, it is assumed that the computer can extrapolate solutions from the check solution without serious error. This fact has been verified by adjusting the computer to solve one check solution accurately and then changing the parameters and checking the results with other numerically calculated check solutions.

In addition to checking the setup of the computer, a check solution furnishes many of the data needed to use the flight simulator to its fullest capabilities. This is illustrated by the fact that the numerical solution is useful in choosing the proper time scale extension factors, scale factors, and gearbox ratios. These factors are determined largely by the maximum and minimum values of the variables and the rate of change of these variables.

Analyses of the accuracy of the M.I.T. Flight Simulator have been made many times, but no analysis of the check solutions has been made. Analysis of the check solution gives much information about the equations being solved, as well as the effect of errors omnipresent in the computer on the solution of the equations.

The objectives of this thesis are to determine the following: 1) a method of establishing the adequacy of the numerical check solutions for the flight simulator, 2) the adequacy of alternate approaches to obtaining the numerical check solutions, 3) the effects of machine errors on the solutions.

1.1. History of the Problem.

The growth of the M.I.T. Flight Simulator has resulted in the acceptance of larger, more complex problems than those handled previously by the

laboratory. The present hand computation method of obtaining numerical check solutions for these larger problems is approaching impracticability because of the excessive time required. A reduction of the time necessary for the computation can be made if accuracy can be sacrificed. The check solution need be accurate only to around 0.05 percent of the maximum deviation of a particular variable, since the best solutions obtained from the simulator are of that order of accuracy.

The process of obtaining a numerical solution is not an exact process. The numbers used in the computation are not exact since a roundoff error occurs when a number is confined to a certain number of significant figures, and since a truncation error results from the numerical approximations to the processes of integration and differentiation. These roundoff and truncation errors are not of serious magnitude on an instantaneous basis, but do have an accumulating and continuing effect which may render a computation useless. This continuing effect varies from point to point in the tabulation. That is, this propagated effect may be negligible at one point, but may be excessively large at another point. The method used for obtaining the numerical result also determines the effect of the error. For example, consider the question: What are the relative effects of an error when using extrapolation techniques as compared to those when using an iterative approach?

Other than the intuitive methods used by experienced numerical analysts, the author has encountered two methods of attacking this problem. Both methods require the solution of a system of variational equations. The method of solution of these variational equations is the only difference between the two approaches. One method gives the error as a function of time and the other gives the error at one point caused by errors

committed at previous times. Chapter II elaborates on the method used in this thesis, as advanced by Murray and Brock,[2] which solves the variational equations associated with the original differential equations by sectioning the solution and assuming constant coefficients for the variational equations. A sample problem using this method isgiven in chapter III to illustrate the final error.

The second scheme uses a weighting-function,[3] which is the solution of the adjoint equations associated with the variation equations. Chapter IV shows the correlation between the two methods by illustrating how to obtain the weighting function from the approach taken by this thesis. A discussion of various approaches to obtaining a check solution is given in chapter V along with sources of errors in numerical computation.

Chapter II

THE THEORY OF THE ERROR[2]

The M.I.T. Flight Simulator is designed to solve a set of simultaneous ordinary differential equations. The numerical solution for these equations is not the true solution due to roundoff and truncation errors present at each interval of computation. The error analysis technique presented in the following sections provides a means of compensating for the errors resulting from all variables by means of corrective time functions for each particular variable. The condition for the application of this theory is that errors from different sources such as roundoff and truncation are superposable, that is, the resultant correction is the sum of the effects of all the separate errors committed. The theory is not valid if the errors interact upon one another.

2.0. The Differential Equations and the Associated Variational Equations.

The system of differential equations solved by the computer can be arranged so that the first time derivative of each variable is equated to a function of the variables. A simple method of performing this arrangement is to note the inputs to each integrator in the setup diagram for the analog computer. The system of equations being solved by the simulator can be written thus,

$$Z_u = f_u(Z_1, Z_2, Z_3, \ldots Z_m, t) \qquad u = 1,2,2\ldots m \qquad 2.1$$

where $Z_u = \frac{d}{dt} Z_u$. The solution of this system of equations is

$$Z_u(t) = Z_u(0) + \int_0^t f_u(Z_1, Z_2, Z_3, \ldots Z_m, t)dt \qquad 2.2$$

Inasmuch as a numerical solution is to be obtained, the values of the variables will be available only at discrete time intervals. Consequently, the equations are modified as follows;

$$Z_{u,n} = Z_{u,n-1} + \int_{t_{n-1}}^{t_n} f_{u,n-1}(Z_1, Z_2, Z_3, \ldots . Z_m, t)dt. \qquad 2.3$$

The ideal solution above is not obtained, since errors are incurred at each numerical step. The solution obtained is the one in which the ubiquitous errors are also used in the computation and as a result Eq. 2.3 should be written as

$$\tilde{Z}_{u,n} = \tilde{Z}_{u,n-1} + \int_{t_{n-1}}^{t_n} f_{u,n-1}(\tilde{Z}_1, \tilde{Z}_2, \tilde{Z}_3, \ldots . \tilde{Z}_m, t)dt. \qquad 2.4$$

where the tilde indicates the value of the function with the errors included in them. If $CZ_{u,n}$ is the correction that must be applied to the computed solution to give the correct solution (the total error at this point is the negative of $CZ_{u,n}$), then by definition $\tilde{Z}_{u,n} + CZ_{u,n} = Z_{u,n}$. The substitution of $\tilde{Z}_{u,n} + CZ_{u,n}$ in Eq. 2.3 yields

$$\tilde{Z}_{u,n} + CZ_{u,n} = \tilde{Z}_{u,n-1} + CZ_{u,n-1} + \int_{t_{n-1}}^{t_n} f_{u,n-1}\Big[(\tilde{Z}_1 + CZ_1), (\tilde{Z}_2 + CZ_2), (\tilde{Z}_3 + CZ_3), \ldots . (\tilde{Z}_m + CZ_m), t\Big]\, dt. \qquad 2.5$$

The relationship for the correction can be obtained by taking the difference between Eqs. 2.4 and 2.5:

$$CZ_{u,n} = CZ_{u,n-1} + \int_{t_{n-1}}^{t_n} f_{u,n-1}\Big[(\tilde{Z}_1 + CZ_1), (\tilde{Z}_2 + CZ_2) \ldots . (\tilde{Z}_m + CZ_m), t\Big]\, dt - \int_{t_{n-1}}^{t_n} f_{u,n-1}\Big[Z_1, Z_2, Z_3 \ldots Z_m, t\Big]\, dt. \qquad 2.6$$

If the partial derivatives $\partial f_u/\partial Z_j$ exist and are continuous, then by expanding $f_u\left[(\tilde{Z}_1 + CZ_1),(\tilde{Z}_2 + CZ_2)..(\tilde{Z}_m + CZ_m),t\right]$ about $f_u(Z_1,Z_2,Z_3...Z_m,t)$ one obtains by neglecting all products of CZ_j

$$f_u\left[(\tilde{Z}_1 + CZ_1),(\tilde{Z}_2 + CZ_2),(\tilde{Z}_3 + CZ_3)...(\tilde{Z}_m + CZ_m),t\right] =$$

$$f_u\left[Z_1,Z_2,Z_3,...Z_m,t\right] + \sum_{j=1}^{m} \frac{\partial f_u}{\partial Z_j} CZ_j. \qquad 2.7$$

Substitution of Eq. 2.7 into Eq. 2.6 gives

$$CZ_{u,n} = CZ_{u,n-1} + \int_{t_{n-1}}^{t_n} \sum_{j=1}^{m} \frac{\partial f_u}{\partial Z_j} CZ_j \qquad 2.8$$

If both sides of Eq. 2.8 are differentiated with respect to time the result will be the system of the following variational equations.

$$\dot{CZ}_{u,n} = \sum_{j=1}^{m} \frac{\partial f_u}{\partial Z_j} CZ_j \qquad 2.9$$

2.1. The "H" Matrix and its Characteristic Values.

A method of systematization is desirable so that the equations can be handled efficiently. This is done by matrix notation wherever possible. Equation 2.9 can be written as

$$\left[\dot{CZ}\right] = \left[H\right]\left[CZ\right] \qquad 2.10$$

where H is an array of partial derivatives in the form

$$H = \begin{bmatrix} \frac{\partial f_1}{\partial Z_1} & \frac{\partial f_1}{\partial Z_2} & \cdots & \frac{\partial f_1}{\partial Z_m} \\ \frac{\partial f_2}{\partial Z_1} & \frac{\partial f_2}{\partial Z_2} & \cdots & \frac{\partial f_2}{\partial Z_m} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \frac{\partial f_m}{\partial Z_1} & \frac{\partial f_m}{\partial Z_2} & \cdots & \frac{\partial f_m}{\partial Z_m} \end{bmatrix} \qquad 2.11$$

If "H" is assumed constant, that is, the array of partial derivatives are constant, then the system of variation equations can be solved easily. If the equations are sectioned into time intervals during each of which the "H" matrix is essentially constant, a particularly simple solution to Eq. 2.10 can be found. The development which follows outlines this approach.

In transform notation the variational equations become

$$\lambda CZ] = [H] CZ] \qquad 2.12$$

or $[H - \lambda U] CZ] = 0$, where U is the unit matrix.

If $\lambda_1, \lambda_2, \lambda_3, \ldots \lambda_p$ are the characteristic values of the "H" matrix, then the solution for the variation equations can be assumed to be

$$CZ] = [V] [e^{\lambda_p t}] [V]^{-1} \gamma] \qquad 2.13$$

where $[e^{\lambda_p t}]$ is a diagonal matrix.

2.2. Characteristic and Dual Vectors.

Characteristic and dual vectors are used both to evaluate the constants by the inclusion of initial conditions and also to separate the

terms so that the effects of the errors present in each of the variables are noticeable.

The column matrix $[\gamma]$ is easily identified as the correction necessary at time t_α, since $[e^{\lambda_p t}]$ is equal to the unit matrix when $t = t_\alpha$.

$$[CZ(\alpha)] = [V][V]^{-1}[\gamma] = [\gamma] \tag{2.14}$$

Differentiation of Eq. 2.13 yields

$$[\dot{CZ}] = [V][\lambda_p e^{\lambda_p t}][V]^{-1}[\gamma] \tag{2.15}$$

but

$$[\lambda_p e^{\lambda_p t}] = [\lambda_p][e^{\lambda_p t}] \tag{2.16}$$

Substituting Eq. 2.13 into Eq. 2.9

$$[\dot{CZ}] = [H][V][e^{\lambda_p t}][V]^{-1}[\gamma] \tag{2.17}$$

then equating Eqs. 2.15 and 2.17

$$[H][V][e^{\lambda_p t}][V]^{-1}[\gamma] = [V][\lambda_p][e^{\lambda_p t}][V]^{-1}[\gamma] \tag{2.18}$$

reveals that

$$[H][V] = [V][\lambda_p] \tag{2.19}$$

Equation 2.19 is the condition for the solution of the V matrix, which is known as the characteristic vectors of the "H" matrix. The inverse matrix is known as the dual vectors. The time expression of Eq. 2.13 for the correction can be used only in the range where the approximation that the "H" matrix remains constant can be applied. The process of obtaining the "H" matrix and its associated characteristic values, characteristic vectors, and dual vectorsmust be repeated for each section. The amount of sectioning that is required will depend upon how rapidly the partial derivatives change.

The total correction in the solution will then be the sum of the corrections due to the separate errors. This idea of superposition requires that the errors from different sources do not interact upon one another. This method of approach has an error due to sectioning and an error due to linearizing. If the accuracy to which the error is sought is not great and if the sectioning is done judiciously, the combined error from these two sources can be neglected. The examples taken in the following chapters will show the validity of this statement.

Chapter III

A SAMPLE PROBLEM

The accuracy of the method of error analysis for a typical set of equations for the M.I.T. Flight Simulator was determined by the use of two numerically computed check solutions. One solution was computed with extreme vigilance using an iterative process with a small interval of tabulation and a large number of significant figures. The number of significant figures was reduced and the interval of tabulation was increased for the second solution. It was assumed that the first solution was the desired result and the correction to the second solution was obtained by noting the errors present in the second solution

3.0. Equations, "H" Matrix, and λ Roots.

The equations as obtained by noting the inputs to each integrator on the general setup diagram of the sample problem for the M.I.T. Flight Simulator in Fig. 3.1 is given below.

$$\begin{aligned} \dot{Y} &= A \sin\theta + B \sin kt \\ \dot{X} &= A \cos\theta + B \cos kt \\ \dot{\theta} &= \psi \\ \dot{\psi} &= f_{\psi}(X,Y,\theta,\psi) \end{aligned} \qquad 3.1$$

The inclusion of the variable ψ shows how higher order time derivatives can be handled by this method. The "H" matrix associated with this set of equations is shown in Eq. 3.2.

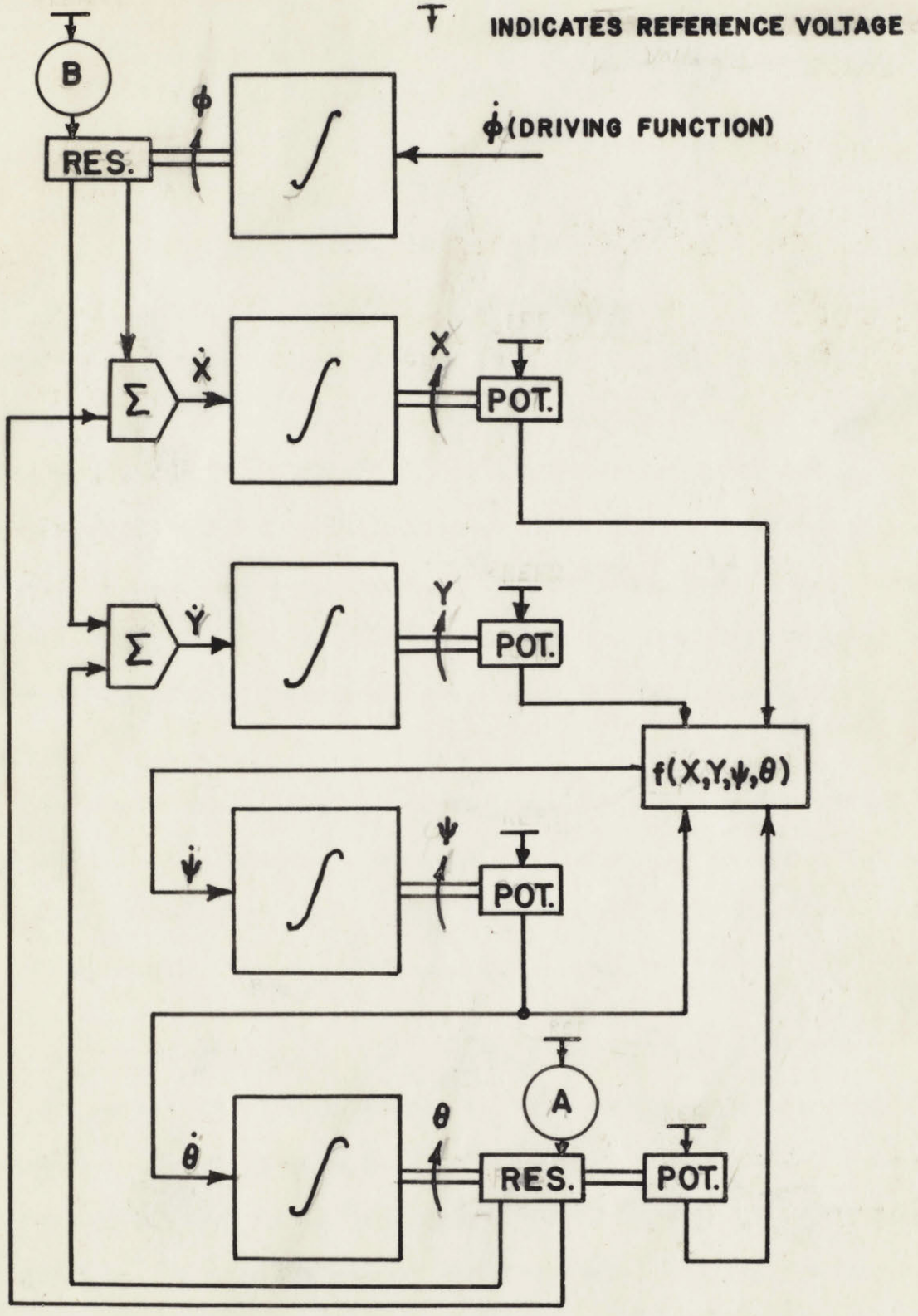


FIG. 3.1 SIMULATOR SETUP DIAGRAM FOR SAMPLE PROBLEM

$$H = \begin{bmatrix} 0 & 0 & A \sin\theta & 0 \\ 0 & 0 & -A\cos\theta & 0 \\ 0 & 0 & 0 & 1 \\ \frac{\partial f_\psi}{\partial Y} & \frac{\partial f_\psi}{\partial X} & \frac{\partial f_\psi}{\partial \theta} & \frac{\partial f_\psi}{\partial \psi} \end{bmatrix} \qquad 3.2$$

The first step in determining a satisfactory way of sectioning the equations consists of examining terms of the "H" matrix to determine which terms have large variations. The characteristic values of the "H" matrix were obtained at 0.1T (T is the total solution time) second intervals to assist in deciding the amount of sectioning required for the problem. One of the characteristic values was zero, another was a negative real number, and the final two were a pair of complex conjugate numbers. The loci of these roots are plotted on the complex plane in Fig. 3.2.

An examination of Fig. 3.2 showed that for the major portion of the solution the characteristic values did not change very rapidly. The assumption of a constant "H" matrix, which implied a set of constant characteristic values, was a valid one for sections of this problem.

3.1. Sectioning, Characteristic Vectors, and Dual Vectors.

On the basis of the assumption that a constant "H" matrix could be assumed for specific portions of the solution, the sample problem was sectioned into three parts, $0 \leq aT \leq 0.4T$ and $0.4T \leq aT \leq 0.8T$ and finally $0.8T \leq aT \leq T$. The solution will be worked for the first two sections using the values for the characteristic values, characteristic vectors, and dual vectors obtained by using the information obtained from the check solution at a = 0.2 and a = 0.6. The characteristic values are

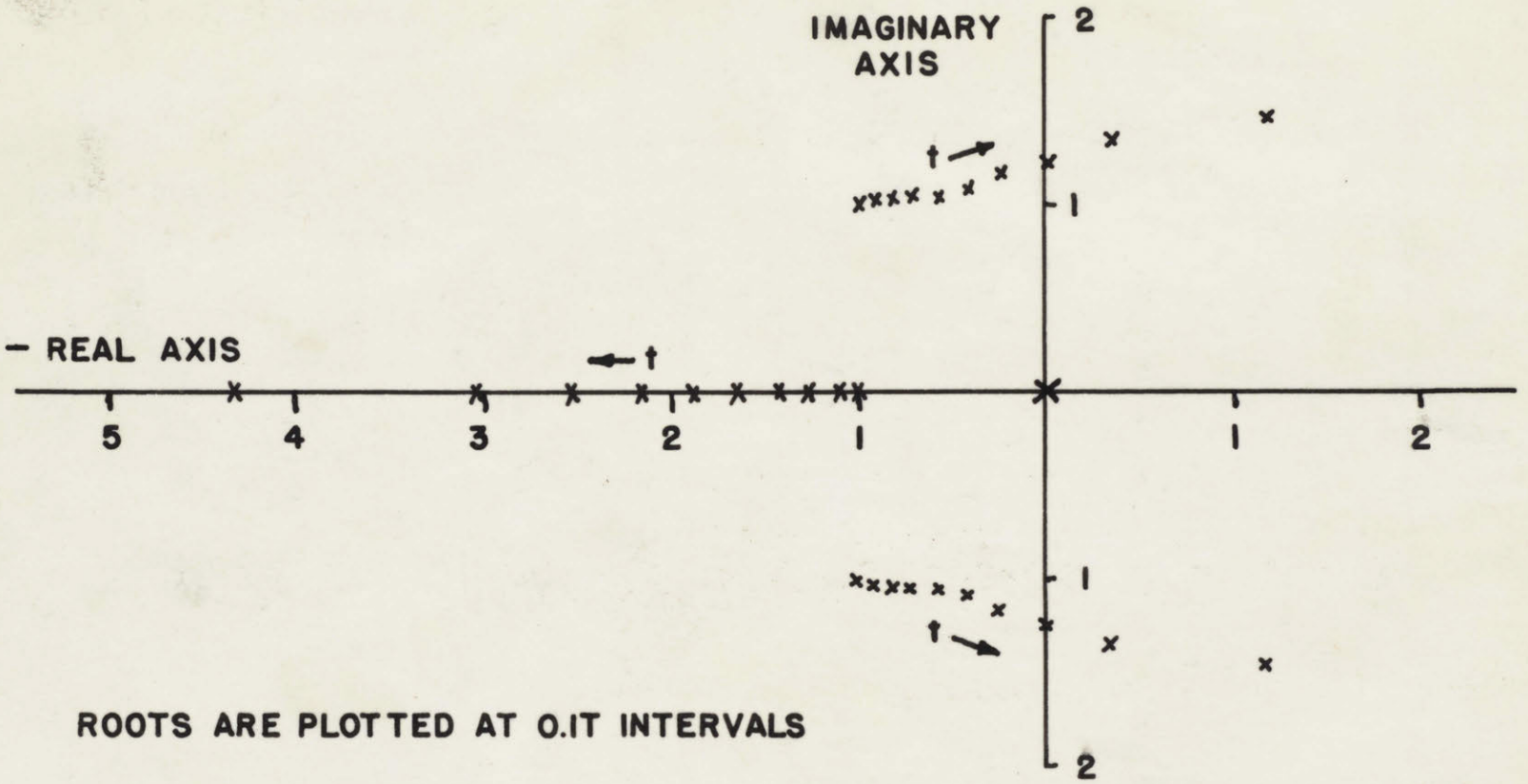


FIG. 3.2 LOCI OF λ ROOTS AS A FUNCTION OF TIME

0, -33.5, -16.6 ± j 55.2, and 0, -56.8, -4.9 ± j 62.2 for a = 0.2 and 0.6, respectively.

The characteristic vectors are

$$\begin{bmatrix} -111.1 & 1 & 1 & 1 \\ -0.225 & -274.7 & -274.7 & -274.7 \\ 0 & 1.54 & -3.8 + j\,2.51 & -3.8 + j\,2.51 \\ 0 & -0.46 & -0.228 - j\,0.758 & -0.228 + j\,0.758 \end{bmatrix} \quad 3.3$$

and

$$\begin{bmatrix} -222.0 & 1 & 1 & 1 \\ -3.51 & -13.05 & -13.05 & -13.05 \\ 0 & 0.211 & -0.252 + j\,0.040 & -0.252 - j\,0.040 \\ 0 & -0.037 & -0.0032 - j\,0.041 & -0.0032 + j\,0.041 \end{bmatrix} \quad 3.4$$

The dual vectors obtained by the inversion of the two matrices representing the characteristic vectors in Eqs. 3.3 and 3.4 are given below:

$$\begin{bmatrix} -.009 & -33 \times 10^{-6} & 0 & 0 \\ 7.3 \times 10^{-6} & -.0036 & 0.219 & 0.724 \\ 1.3 \times 10^{-8} + j\,2.2 \times 10^{-6} & -6.7 \times 10^{-6} - j\,1.1 \times 10^{-3} & -.109 + j\,.033 & -.362 + j\,.77 \\ 1.3 \times 10^{-8} - j\,2.2 \times 10^{-6} & -6.7 \times 10^{-6} + j\,1.1 \times 10^{-3} & -.109 - j\,.033 & -.362 - j\,.77 \end{bmatrix} \quad 3.5$$

and

$$\begin{bmatrix} -.0045 & 3.45 \times 10^{-4} & 0 & 0 \\ 7.18 \times 10^{-5} & -.0454 & 2.33 & 2.28 \\ 2.46 \times 10^{-4} + j\,3.47 \times 10^{-4} & -.0156 - j\,.022 & -1.16 + j\,.972 & -1.14 + j\,13.24 \\ 2.46 \times 10^{-4} - j\,3.47 \times 10^{-4} & -.0156 + j\,.022 & -1.16 + j\,.972 & -1.14 - j\,13.24 \end{bmatrix} \quad 3.6$$

Satisfactory assurance of the validity of the theory can be obtained by carrying out a few numerical examples. The correction needed in the variable X will be found in the following pages. Equation 2.13 shows that the correction that must be applied to X to obtain the desired solution is

$$CX(aT) = \sum_{\alpha=0}^{1} V_x \left[e^{\lambda_p bT} \right] [V]^{-1} \gamma(\alpha T) \qquad 3.7$$

where $b = (a-\alpha)$ with $a \geq \alpha$. This equation was valid only in the region $a_1 \leq \alpha \leq a_2$ where a_1 and a_2 were the upper and lower bounds of the region where the "H" matrix was assumed constant. For the region $0 \leq a \leq 0.4$ the correction, CX, given for the error present at time αT is $CX = CX_Y + CX_X + CX_\psi + CX_\theta$ where

$$CX_Y = \left[.002 - .002\, e^{-33.5bT} - 275\, e^{-16.6\,bT} (2.6 \times 10^{-8} \cos 55.2bT + 4.4 \times 10^{-6} \sin 55.2bT) \right] \gamma_y$$

$$CX_X = \left[7.39 \times 10^{-6} + e^{-33.5bT} + 275\, e^{-16.6\,bT} (1.34 \times 10^{-4} \cos 55.2bT + 2.2 \times 10^{-4} \sin 55.2\,bT) \right] \gamma_x$$

$$CX_\psi = \left[275 - .219 e^{-33.5bT} + e^{-16.6\,bT} (.219 \cos 5.52bT - .066 \sin 55.2bT \right] \gamma_\psi \qquad 3.8$$

$$CX_\theta = \left[275 - .724 e^{-33.5bT} + e^{-16.6\,bT} (.724 \cos 55.2bT - 1.54 \sin 55.2bT) \right] \gamma_\theta$$

Similarly for the region $0.4 \leq \alpha \leq 1$ the correction, CX, becomes

$$CX_Y = \left[.0158 - .0094\, e^{-56,8bT} - e^{-4.9bT} (.0064 \cos 62.2\,bT - .0091 \sin 62.2bT) \right] \gamma_Y$$

$$CX_X = \left[.0012 + .592\, e^{-56.8bT} - e^{-4.9bT} (.4064 \cos 62.2bT + .573 \sin 62.2bT) \right] \gamma_X$$

$$CX_{\psi} = -13.05\left[2.33\, e^{-56.8bT} + e^{-4.9bT}\,(-2.33 \cos 62.2bT + 1.94 \sin 62.2bT)\right]\gamma_{\psi}$$

$$CX_{\theta} = -13.05\left[2.28\, e^{-56.8bT} + e^{-4.9bT}\,(-2.28 \cos 62.2bT + 26.5 \sin 62.2bT)\right]\gamma_{\theta}$$

The constant term for the correction in X caused by γ_Y, γ_θ, and γ_X was neglected since the numerical value was less than 1/100th of the other quantities. This indicates that the lasting effect on X caused by an error in ψ, θ, or X is very small.

3.2. The Step Response.

The step responses (responses to a step of error), or sensitivity functions, are the separate terms of Eqs. 3.8 and 3.9. The relationship between the portions of CX caused by errors committed in X, Y, and θ are given in Figs. 3.3, 3.4, 3.5, and 3.6, respectively. The step response is the propagation of one error originating at one interval of tabulation in the check solution. The procurement of a function of the error present in each of the variables is necessary so that the convolution of the error functions and the step responses can be made to obtain the total correction.

3.3. Truncation Errors.

If the functions to be numerically handled are assumed analytic, then the two major errors are due to roundoff and truncation sources.[5] The effects of these errors are inherently affected by the integration formula used in the numerical process. An elaboration of this phase of the problem will be undertaken in chapter V. An iterative process utilizing the Newton-Gregory integration formula[6] with finite differences is used by the analysis section of the laboratory in obtaining the check solutions.

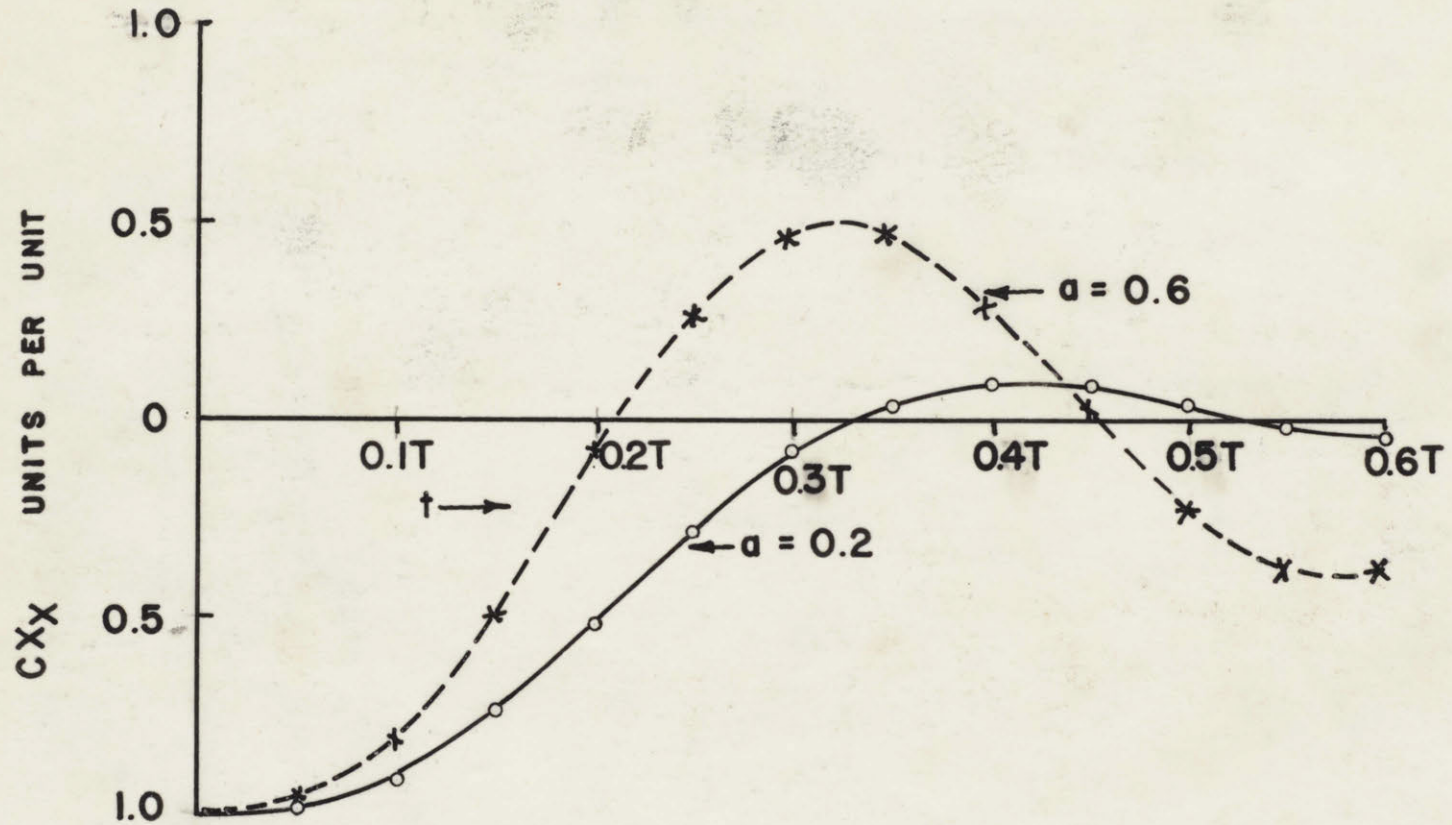


FIG. 3.3 CX CAUSED BY ERROR IN X

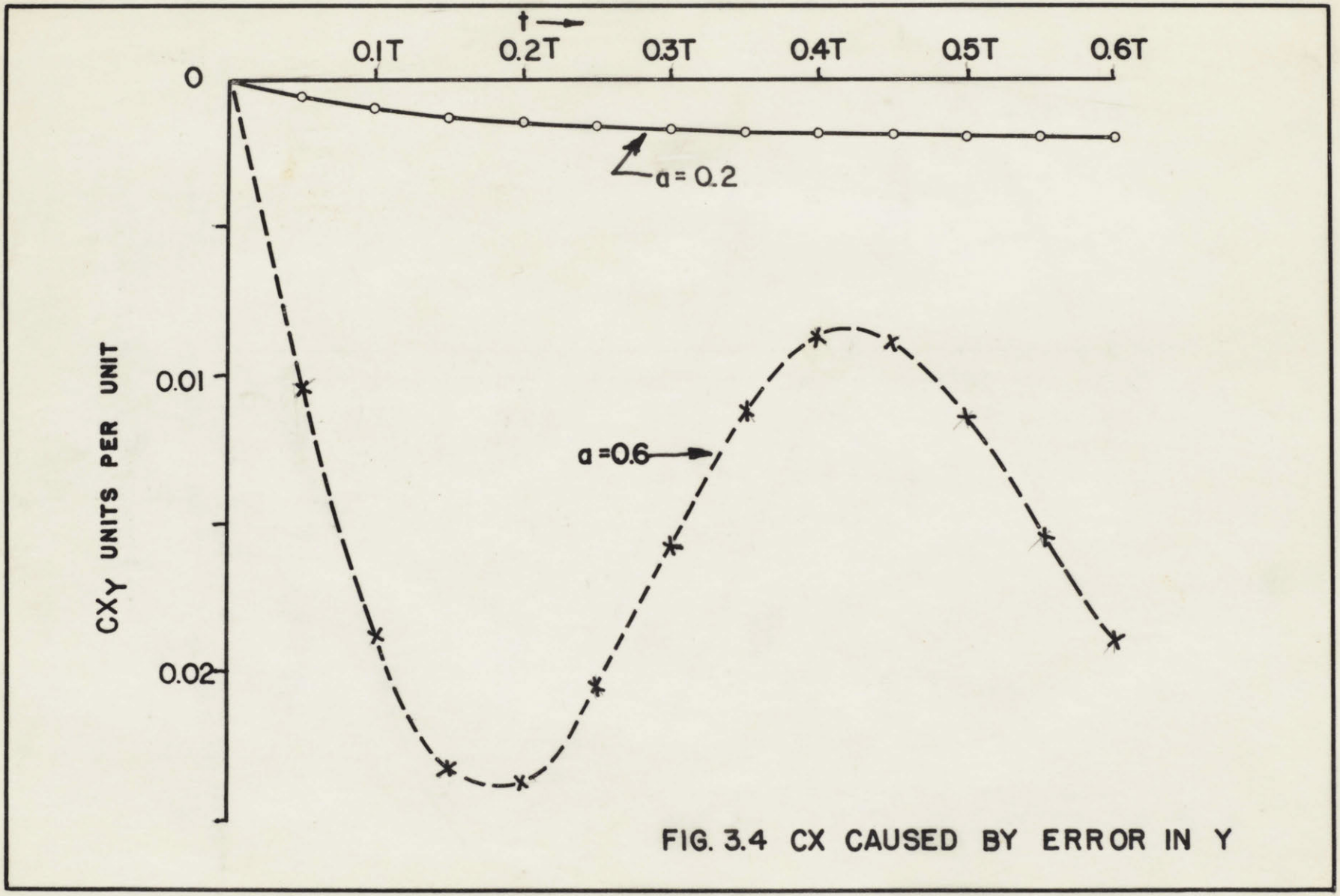


FIG. 3.4 CX CAUSED BY ERROR IN Y

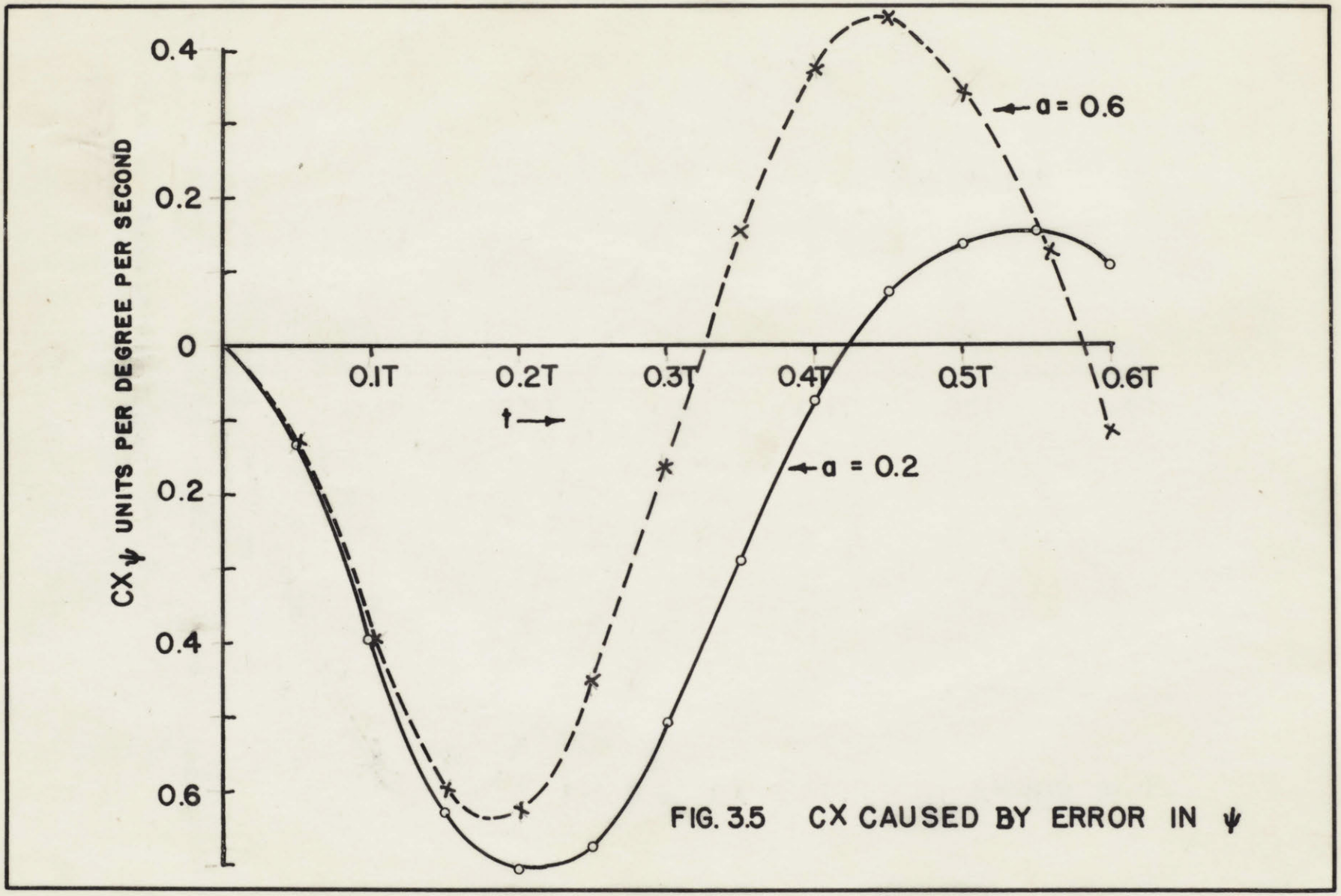


FIG. 3.5 CX CAUSED BY ERROR IN ψ

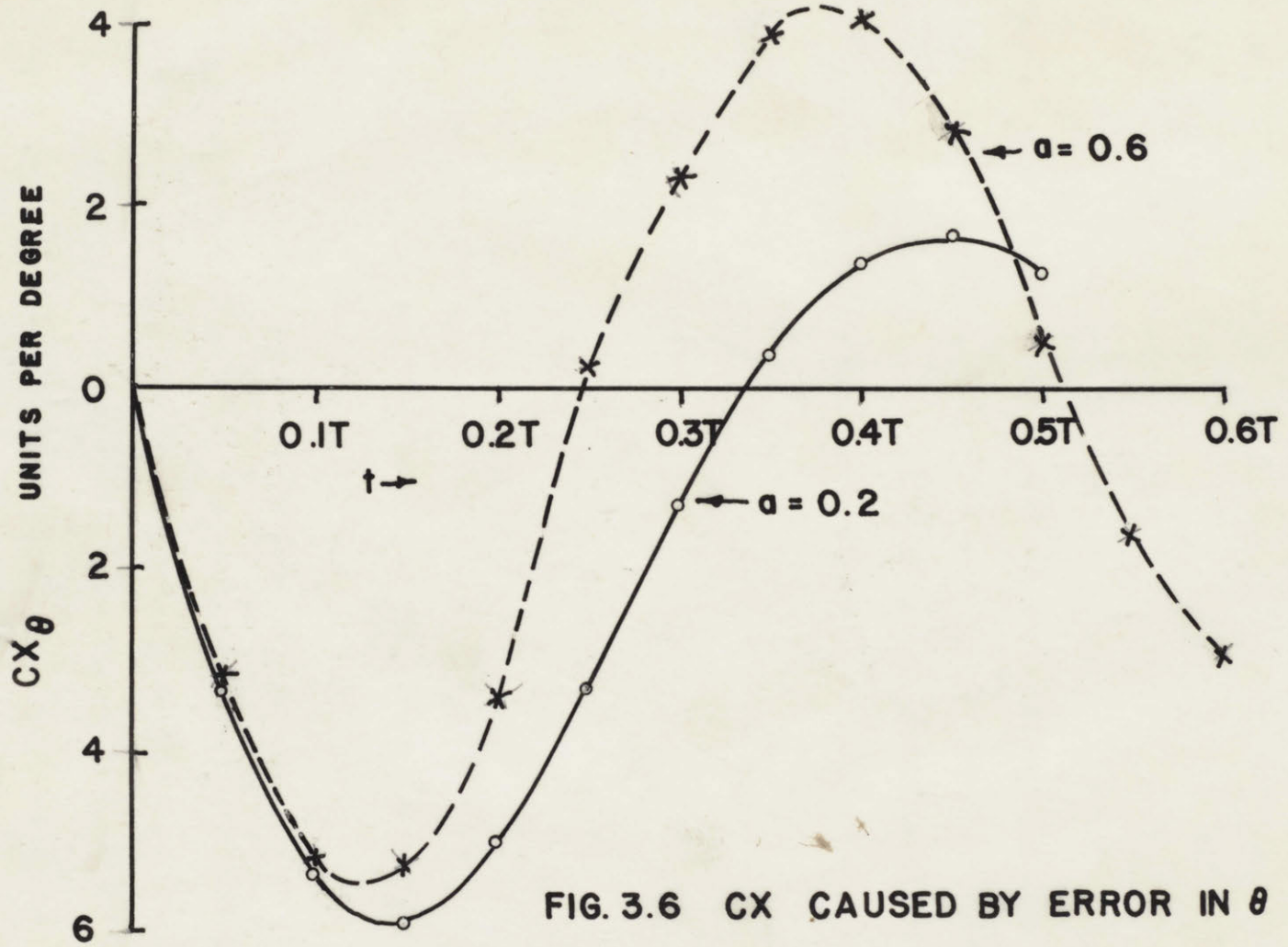


FIG. 3.6 CX CAUSED BY ERROR IN θ

The truncation error in the Newton-Gregory formula

$$f_n = f_{n-1} + h\left[f_n - \frac{1}{2}\Delta' - \frac{1}{12}\Delta'' - \frac{1}{24}\Delta''' - \frac{19}{720}\Delta^{iv} - \ldots\right] \qquad 3.10$$

is the first neglected term in the formula, therefore, the error committed by using n differences will be[8]

$$E_T = h\, C_{n+1}\Delta^{n+1} \qquad 3.11$$

In the check solution being studied the first differences were neglected in the integration of the variables X, Y, and θ. The second differences were neglected in ψ. Substitution of these neglected differences in Eq. 3.11 reveals that the truncation error for X, Y, and θ at each interval will be

$$E_T = -(.025T)\left(\frac{1}{2}\right)\Delta' = -\frac{T\Delta'}{80} \qquad 3.12$$

but for ψ will be

$$E_T = -(.025T)\left(\frac{1}{12}\right)\Delta'' = -\frac{T\Delta''}{480} \qquad 3.13$$

The plots of the first differences of X, Y, and θ multiplied by $\frac{T}{80}$ are given in Figs. 3.7, 3.8, and 3.9. The second differences of ψ multiplied by T/480 are given in Fig. 3.10.

Since a reduction in the amount of numerical work was desirable, the following scheme was used to reduce the work. The smoothed truncation error curves were sampled at 0.05T-second intervals. The size of the samples was increased by a factor of two since the number of error steps (intervals at which a truncation error occurs) was decreased by two.

An examination of the truncation error in ψ reveals that this error seems to be alternately positive and negative and as a result has very

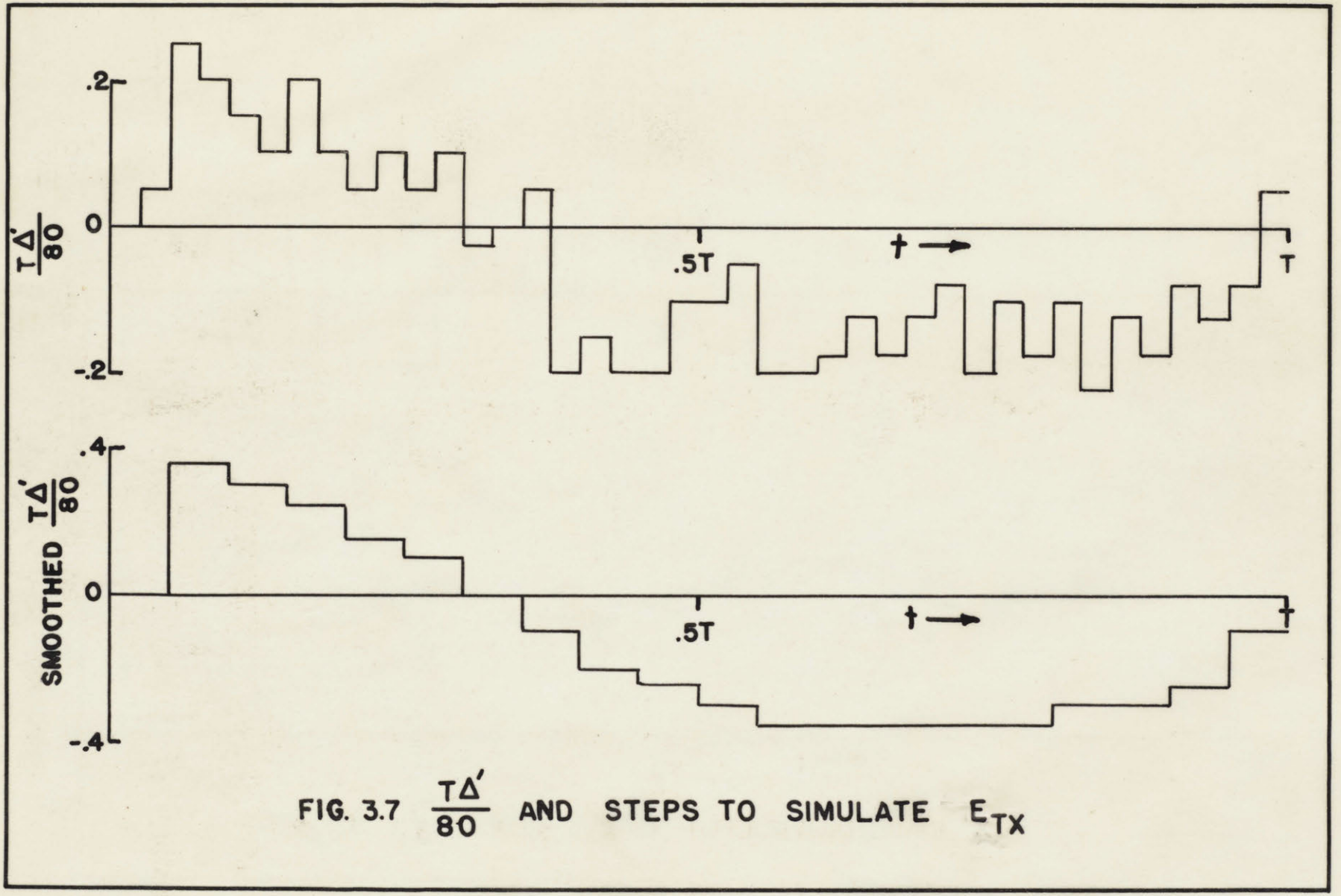


FIG. 3.7 $\frac{T\Delta'}{80}$ AND STEPS TO SIMULATE E_{TX}

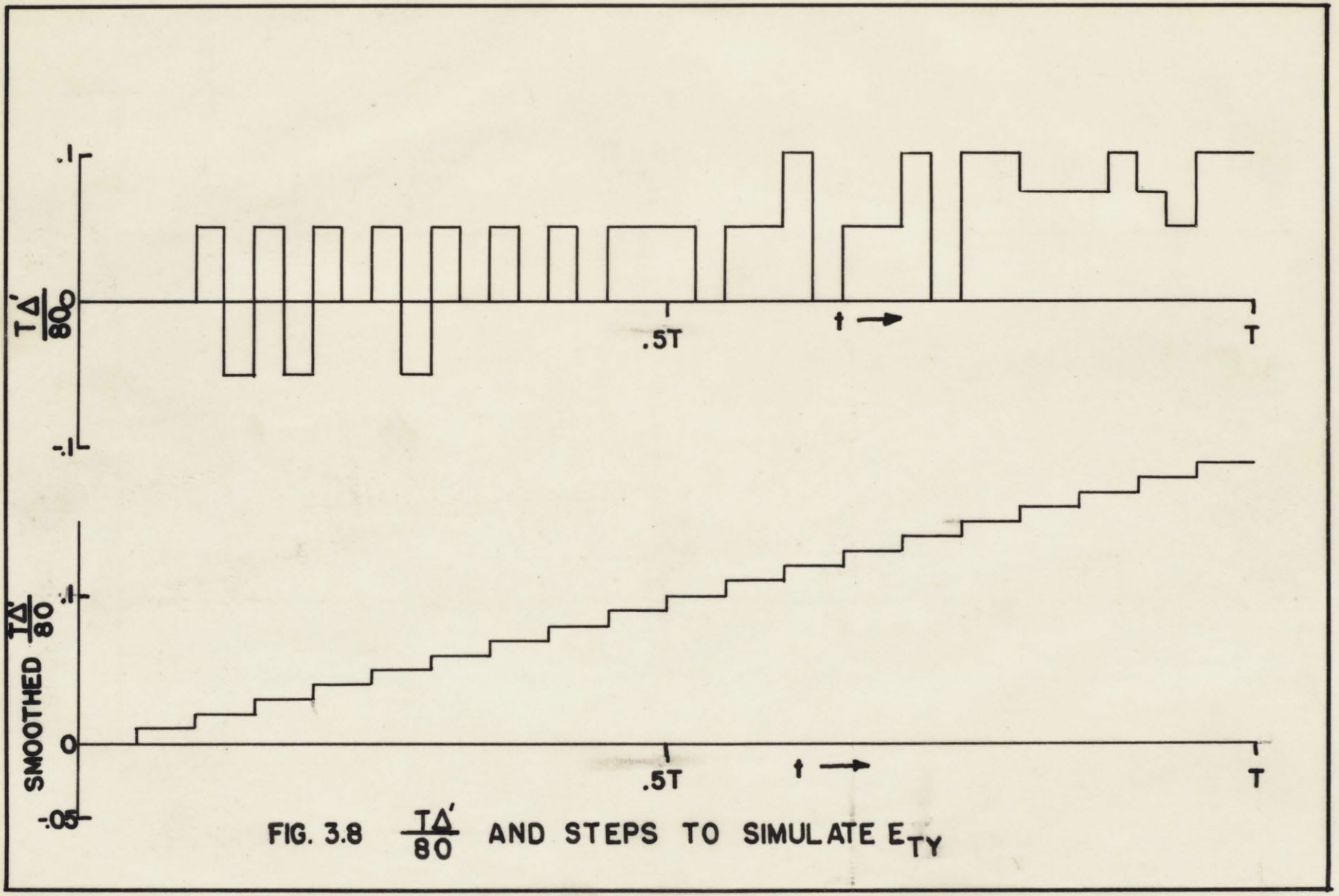


FIG. 3.8 $\frac{T\Delta'}{80}$ AND STEPS TO SIMULATE E_{TY}

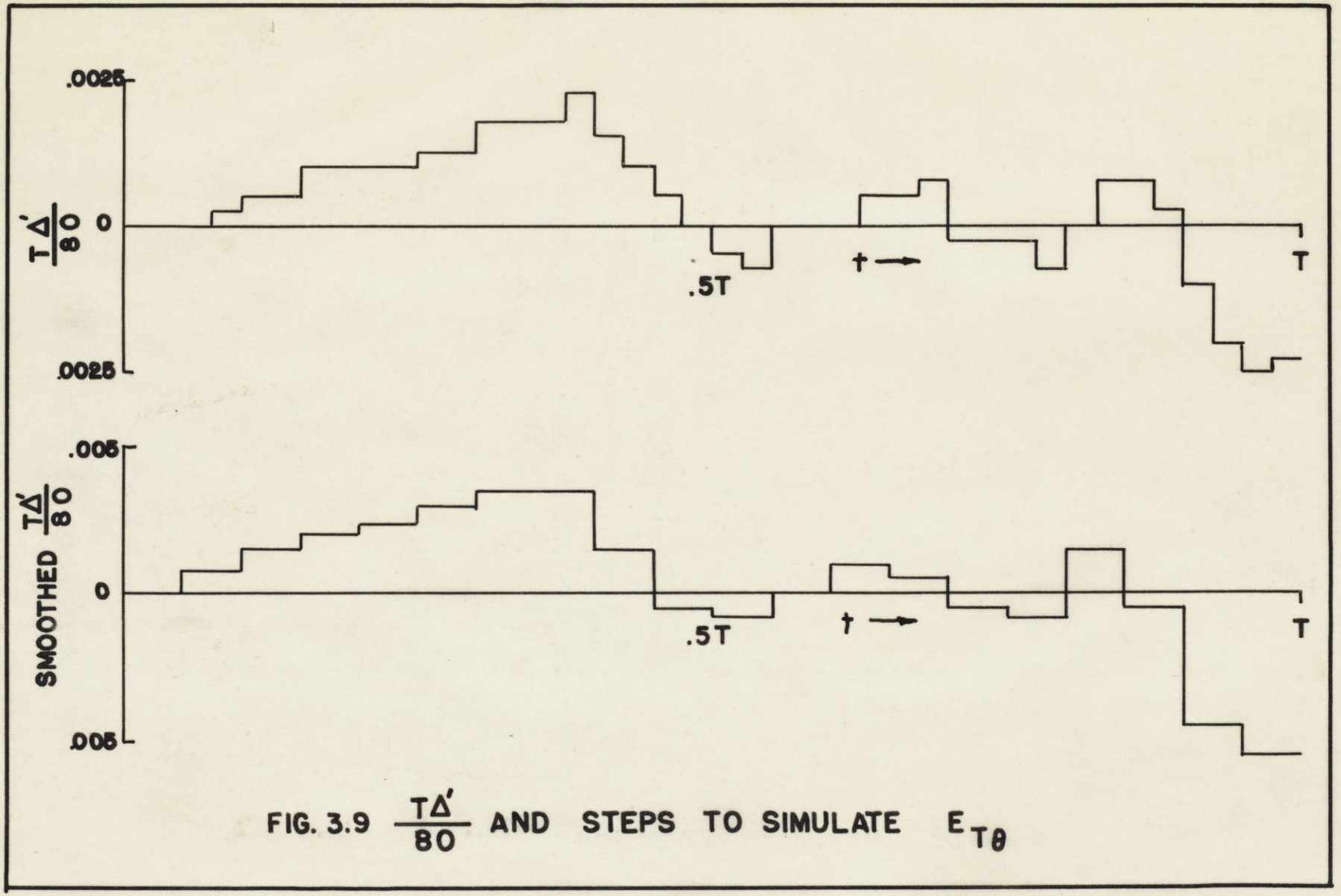


FIG. 3.9 $\frac{T\Delta'}{80}$ AND STEPS TO SIMULATE $E_{T\theta}$

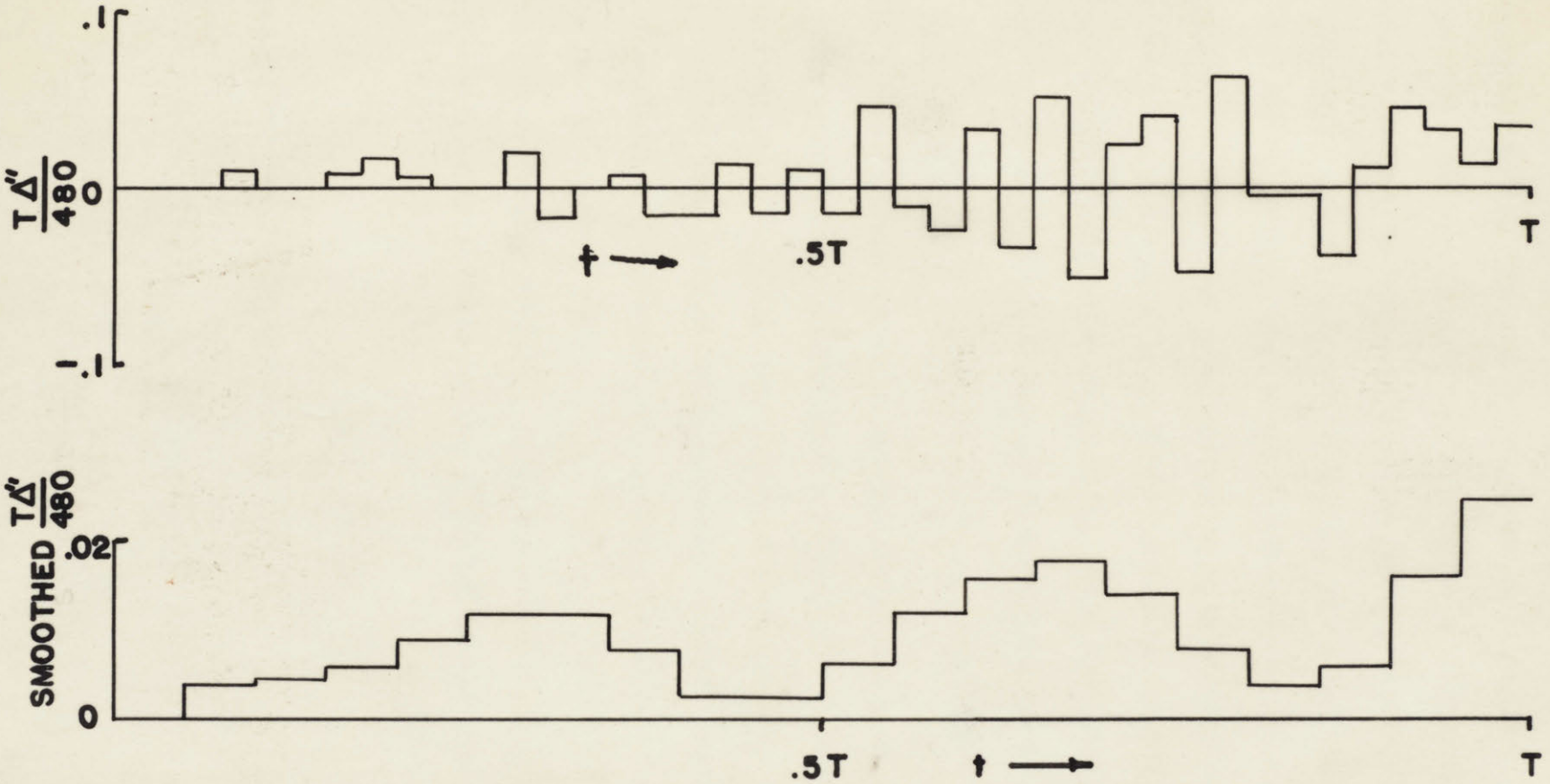


FIG. 3.10 $\frac{T\Delta''}{480}$ AND STEPS TO SIMULATE $E_{T\psi}$

little effect upon X. The truncation error in ψ is almost insignificant in comparison with the other errors when convolved with CX_{ψ}. This is evidenced by noting the scale used in the graphs relating these quantities.

Although a crude approximation for E_{TY} was made, the step response was such that the contribution to the total correction was negligible from this source of error. The major contribution to the correction arose from truncation errors in θ and X. An inspection of the magnitude of the step responses and the magnitude of $E_{T\theta}$ and E_{TX} shows that these two sources were the major contributors to the total correction curves.

3.4. Roundoff Errors.

The roundoff error cannot be handled as precisely as the truncation error since roundoff is a statistical variation and is not a function of the equations being handled. The probability of having a roundoff error can be handled. The maximum value caused by roundoff and a possible standard deviation of the error originating at each time interval will be attempted here.

The Newton-Gregory formula used by the computing section of the laboratory when written in terms of the ordinates rather than differences becomes

$$Z_n = Z_{n-1} + h\dot{Z}_n, \qquad 3.14$$

when neglecting the first differences, and

$$Z_n = Z_{n-1} + \frac{h}{2}\left[\dot{Z}_n + \dot{Z}_{n-1}\right] \qquad 3.15$$

$$Z_n = Z_{n-1} + \frac{h}{12}\left[5\dot{Z}_n + 8\,\dot{Z}_{n-1} - \dot{Z}_{n-2}\right] \qquad 3.16$$

when using first and second differences respectively. If the $\dot{Z}$'s are assumed correct and the maximum roundoff error in the $\dot{Z}$'s is ε, the maximum contribution to the Z's from the $\dot{Z}$'s will be[9]

$$E_{RM} = \varepsilon h \sum_j |c_j|. \qquad 3.17$$

The factor $\sum_j |c_j|$ would be equal to one when no differences or first differences are used, but would be equal to 14/12 when second differences are used. Equation 3.17 shows that the maximum error in Z caused by a roundoff of ε in $\dot{Z}$ will be reduced by the interval of tabulation, h. The standard deviation of the error can be found by using the material in Appendix A.

The preceding discussion indicates that the contribution of roundoff in $\dot{Z}$ to Z is not significant. The contribution of an error in the solution is attributed mainly to the roundoff in the variable. The maximum error in X would occur if a roundoff error of the same sign occurred in every variable for a period of time. The committed error function would be ε in this case, if maximum roundoff errors are assumed. If the probability of having a roundoff error of magnitude $0 \leq e \leq \varepsilon$ is uniform, then the probability of having an error e would be P(e)de and the standard deviation would be

$$\sigma = 2\sqrt{\int_0^{\varepsilon} e^2 P(e)de} = \frac{\sqrt{3}}{3}\varepsilon$$

The resultant error function would have a standard deviation equal to $\sqrt{3}\varepsilon/3$. Since the errors have an equal chance of being either positive or negative, the resultant error in X will be almost negligible, when the convolution with the step response is performed.

Roundoff errors can be assumed random when the change in the quantity at each interval is greater than the maximum roundoff error. The only quantity in the check solution that did not satisfy this criterion was the variable X. The change in X was less than the roundoff for a period of 8 intervals when X reached a maximum value somewhere between a = 0.5 and a = 0.7. With this exception the rest of the variables changed rapidly enough, so that the effects of roundoff could be neglected.

3.5. The Resultant Correction.

The resultant correction was obtained by summing the result of the convolutions of the approximation to the truncation error (roundoff errors are neglected) and the respective step responses. For the second section $0.4 \leq a \leq 0.8$ the correction was obtained by using the terminal conditions of the first section as well as the errors existing in the second section. The resultant correction is plotted in Fig. 3.11 as a dotted curve. The solid curve is the approximation to the actual difference between the two check solutions determined by the computing section.

The apparent displacement in the solid curve after aT = 0.45T seconds was caused by X reaching a maximum value and remaining at this maximum value for 0.15T seconds. The inclusion of this roundoff error in the calculated correction would give a closer, but not exact, agreement between the two curves.

The occurrence of this gross roundoff error cannot be predicted, but its effect is not serious since a cursory examination of Fig. 3.11 indicates that the cumulative error for the majority of the solution is in the order of magnitude of the roundoff error in X. Notice should be given also to

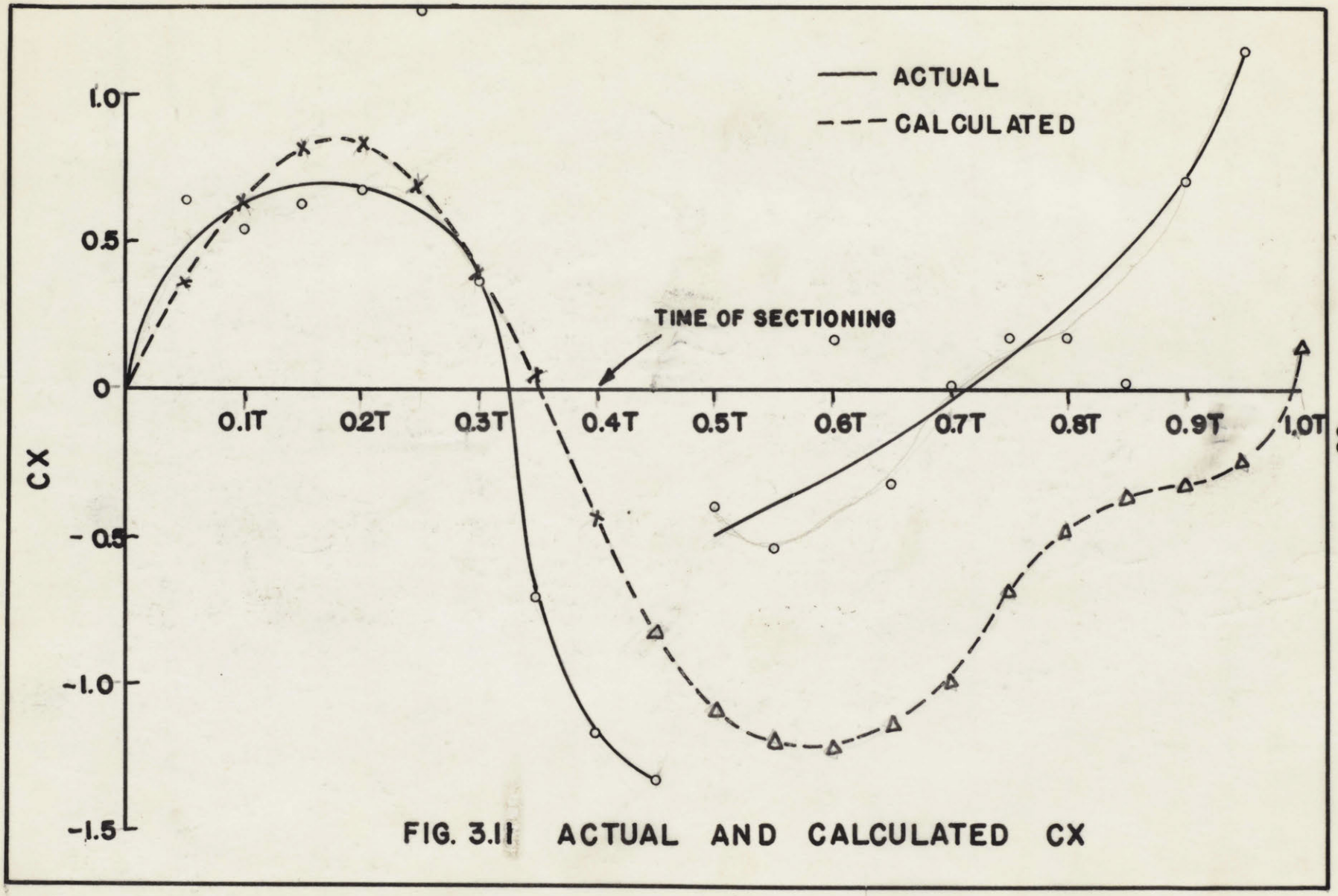


FIG. 3.11 ACTUAL AND CALCULATED CX

the smoothing effect of integration on errors in $\ddot{\theta}$. The further that the errors are removed from the point of interest, the less the effect of the error. The more integrations that have to be performed before the effect is felt, the less is the effect.

The more sections used, the more accurate the resultant error to the actual propagated error. If the amount of sectioning is carried to its extreme, the result is the solution of the variation equations without sectioning. A superficial examination of Fig. 3.11 indicates that the resultant correction gives the order of magnitude of the error, but not the exact form. These results indicate that a rough estimate of the error can be made from a very few sections.

The last section $.8 \leq a \leq 1$ was not performed since the object of this chapter was to verify the validity of the sectioning and linearizing assumptions for this error analysis technique. For the purposes of error study this complete survey need not be made since a set of step responses would give the desired information concerning the sensitivity of the various sections to errors. Determination of the step responses is the beginning to the error analysis study made by Rabow,[4] in which the step responses were obtained by superimposing a step (error) upon a solution and recording the variation that resulted from this step when compared to the original solution. The steps were inserted in the different variables and the results recorded at the various points of interest. The time at which the steps were initiated was also varied. The proper interpretation of these step responses gave the weighting function that related the error committed and the propagated error at a later time.

The "H" matrix technique gives a method of obtaining these step responses during the presetup stage. The responses obtained are only approximations to the experimentally obtained responses. Although the pulses of error used in this chapter were the errors present in using a numerical process, the errors could just as well have been errors in the computer functional elements. A more efficient use of error analysis can probably be made when the two techniques are combined; therefore, the following chapter will show how a step response may be obtained at various times during a solution by using the information obtained from the "H" matrix technique.

Chapter IV

The study of the propagation of errors in analog computers by Jones[3] used a weighting function which is the relationship between the error at various times during a solution and the error that has accumulated at a later time T in a solution. This weighting function can be obtained by properly analyzing a set of step responses for the particular system. This chapter illustrates the correlation between the step response obtained experimentally and the one calculated by the "H" matrix.

4.0. Adjoint Method.

The weighting function is the solution of the adjoint equations which are a set of equations in which the "H" matrix has been transposed with a negative sign.

$$\dot{a}] = [-H]_t a \quad 4.1$$

$$a] = [-H]_t^{-1} \dot{a}] \quad 4.2$$

$$\_a\_ = \_\dot{a}\_[-H]^{-1} \quad 4.3$$

If Eq. 4.3 is premultiplied with Eq. 2.11 the result is

$$\_a\_\dot{CZ}] = \_\dot{a}\_[-H]^{-1}[H]CZ] = -\_\dot{a}\_CZ] \quad 4.4$$

or

$$\_a\_\dot{CZ}] + \_\dot{a}\_CZ] = 0 \quad 4.5$$

but

$$\frac{d}{dt}\_a\_CZ] = \_a\_\dot{CZ}] + \_\dot{a}\_CZ]. \quad 4.6$$

Therefore

$$\frac{d}{dt}\left[\underset{\sim}{\alpha}\, CZ\right] = 0 \qquad 4.7$$

and by integrating from time t_α to T

$$\int_{t_\alpha}^{T} \frac{d}{dt}\left[\underset{\sim}{\alpha} CZ\right] dt = 0 \qquad 4.8$$

$$\alpha_1(T)CZ_1(T) + \alpha_2(T)CZ_2(T) \ldots .$$

$$- \left[\alpha_1(t_\alpha)CZ_1(t_\alpha) + \alpha_2(t_\alpha)CZ_2(t_\alpha) + \ldots .\right] = 0 \qquad 4.9$$

If the boundary conditions of $\alpha_n(T) = 1$, $\alpha_p(T) = 0$ when $p \neq n$, the equation reduces to

$$CZ_n(T) = \alpha_{n1}(t_\alpha)CZ_1(t_\alpha) + \alpha_{n2}(t_\alpha)CZ_2(t_\alpha) + \ldots .$$

An inspection of the final equations shows that the propagated error at time T is the weighted sum of the errors present at t_α in the variables. All errors must be changed to an equivalent error in one or more of the variables. The solutions of the adjoint equations may be rather difficult, but a system to obtain them experimentally was devised by Jones and verified by Rabow. The essence of this method was to purposely insert an error (a step) at various times in a solution and to compare the results with a solution that was run without the error (the step) inserted. The difference between the two solutions (with and without the error) was the step response to an error at the time when the step was initiated. The difference at a time T in the two solutions was plotted against the time at which the error was inserted as the base. The resultant curves gave the weighting function.

4.1. Pulse Response by Using "H" Matrix.

The step response at various points in a solution is obtained by the "H" matrix method. Since the problem is sectioned so that a constant "H" matrix can be assumed, the step response is the same regardless of when the error is inserted during a section of the problem. The responses are different for different sections. The time response to an error originating in a section is

$$\underline{CZ}_n = \underline{V}_n \left[e^{\lambda t} \right] \left[V \right]^{-1} \underline{\gamma}.$$

For the purpose of illustrating how to obtain these step responses, a hypothetical problem with two variables which has been sectioned into two parts will be used. The point of sectioning is at time t_s. The response in the two variables due to an error occurring at time t_α where $0 \leq t_\alpha \leq t_s$ or $t_s \leq t_\alpha \leq T$,

$$CZ_{11} = \left[K_1^1 e^{\lambda_1 (t-t_\alpha)} + K_1^2 e^{\lambda_2 (t-t_\alpha)} \right] \gamma_1(t_\alpha)$$

$$CZ_{21} = \left[K_2^1 e^{\lambda_1 (t-t_\alpha)} + K_2^2 e^{\lambda_2 (t-t_\alpha)} \right] \gamma_2(t_\alpha).$$

The propagated error in Z_1 caused by errors present in both Z_1 and Z_2 at time t_α is

$$CZ_{11} = \left[K_1^1 e^{\lambda_1 (t-t_\alpha)} + K_1^2 e^{\lambda_2 (t-t_\alpha)} \right] \gamma_1(t_\alpha)$$

$$CZ_{12} = \left[L_1^1 e^{\lambda_1 (t-t_\alpha)} + L_1^2 e^{\lambda_2 (t-t_\alpha)} \right] \gamma_1(t_\alpha).$$

Two types of step responses can be obtained, one when the step is inserted into the same variable that is being measured, and the other when the step is inserted into a different variable. If a step is inserted in

Z_1 during the period $t_s \leq t_\alpha \leq T$ and the response in Z_1 is desired, then CZ_{11}^2 is the desired step response. The superscripts refer to the section of the solution. A step initiated at a time $0 \leq t_\alpha \leq t_s$ in Z_1 requires that CZ_{11}^1 and CZ_{21}^1 for the first section be calculated until time $t_s - t_\alpha$. The step response during the first section is CZ_{11}^1. The remainder of the step response (during the second section) is obtained by calculating CZ_1^2 where the initiating steps are equal to the terminal values of CZ_{11}^1 and CZ_{21}^1 at time t_s. This procedure is carried out for as many times as there are sections in the solution. The above procedure is used to obtain the step response of an error in the same variable as the one being measured.

If the step is inserted in Z_2 and the response in Z_1 is desired, then for the interval $t_s \leq t_\alpha \leq T$ the desired response is CZ_{12}^2. If the step is initiated at $0 \leq t_\alpha \leq t_s$ then for the interval $t_\alpha < t_s$ the desired step response is CZ_{12}^1 during the first section. The value of CZ_{12}^1 and CZ_{22}^1 at time $t = t_s - t_\alpha$ is the magnitude of the step used at time t_s to obtain $CZ_{12}^2 + CZ_{11}^2$ for the second section.

4.2. Experimental Results.

The two step responses given in Fig. 4.1 as a solid line were obtained by Rabow by inserting a pulse into the input of integrators x and θ, respectively. This pulse in the integrand is equivalent to a step in the output. The responses are for the variable x. A general block diagram for the simulator setup used to obtain these pulse responses are given in Fig. 4.2, which simulates the equations given below.

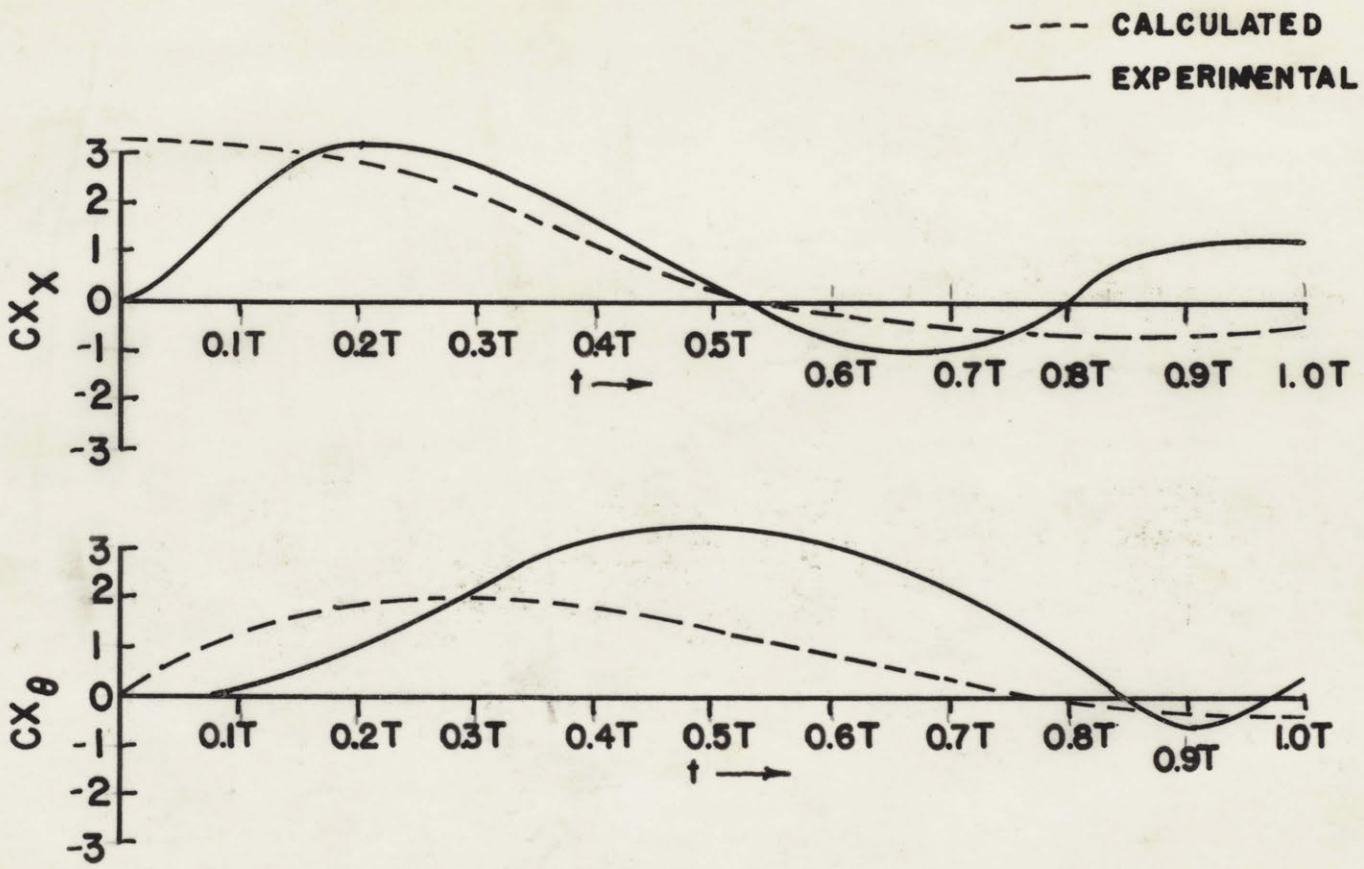


FIG. 4.1 EXPERIMENTAL AND CALCULATED STEP RESPONSES

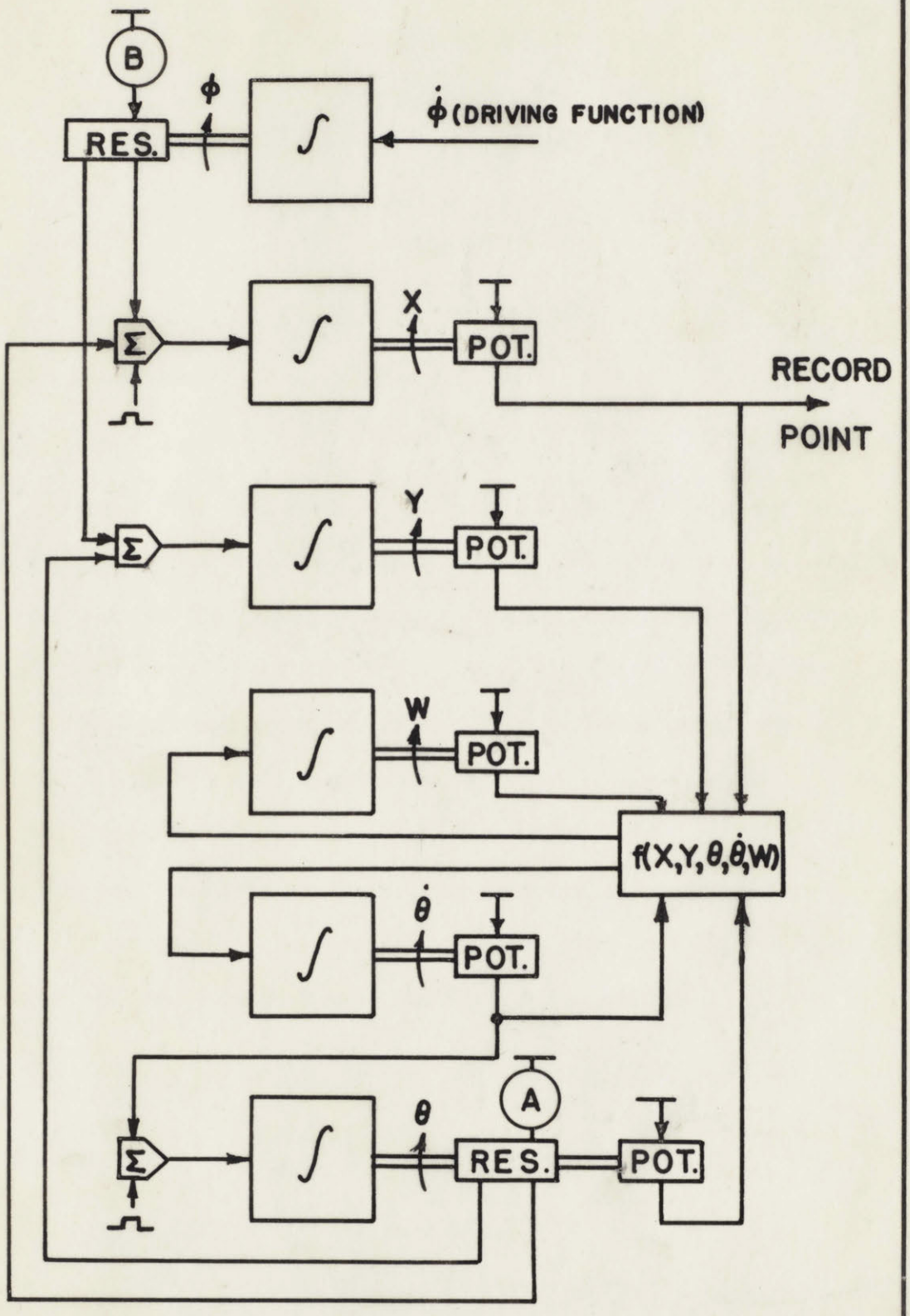


FIG.4.2 SIMULATOR SETUP DIAGRAM SHOWING RECORDING POINT AND PULSE INSERTION POINTS

$$\frac{dy}{dt} = A \cos \theta + B \cos \phi$$

$$\frac{dx}{dt} = A \sin \theta + B \sin \phi$$

$$\frac{d\dot{\theta}}{dt} = f(x, y, w, \dot{\theta}, \theta)$$

$$\frac{d\theta}{dt} = \dot{\theta}$$

$$\frac{dw}{dt} = f(\dot{\theta}, \theta, w)$$

The pulse widths used were 0.3 seconds and the pulse height was of such a magnitude that both the x and θ shafts were displaced by 0.00333 units. Inasmuch as the pulse is not an impulse, the shaft displacement is not an ideal step, but a step with a ramp front. As a result the CX_x in Fig. 4.2 did not start from 0.0033 but started from zero. The delay in the rise of CX_θ was also due to the fact that a pulse was used rather than a step.

The dotted line gives the same responses for an ideal step in using the initial conditions to the original differential equations in obtaining the "H" matrix. The equations for the step responses are

$$CX_x = 3.559 \times 10^{-5} e^{-525.4aT} - 3.394 \times 10^{-2} e^{-134.2aT} + e^{-10.2aT} \left[1.034 \cos 20.17aT + .8735 \sin 20.17aT\right]$$

$$CX_\theta = -.068 e^{-525.4aT} - 26.260 e^{-134.2aT} + e^{-10.2aT} \left[26.328 \cos 20.17aT - 1210 \sin 20.17aT\right]$$

The assumption of a constant "H" matrix cannot be made for an entire solution for the type of equations solved on the M.I.T. Flight Simulator.

A possible time of sectioning would be at a time of two thirds the solution time. This would give closer agreement at the end of the responses. This step response without sectioning gives an indication of what can be expected by using only the initial values of the variables in solving the "H" matrix.

The conclusions to be drawn here are that the "H" matrix method can give the step responses necessary for obtaining the weighting functions by using a very few sections. The experimentally obtained step response is an approximation to the step response in that a pulse rather than an impulse is used in the integrator inputs. The precision of the simulator dictates the lower limit for the size of the pulse to be used, and the upper limit is reached when the error cannot be assumed to be an increment additive to the solution. The weighting function obtained by using pulse responses obtained by the "H" matrix can be made to correspond closely to experimental results if the solution is sectioned into possibly three sections.

Chapter V

SOURCES OF ERROR

The truncation and roundoff errors present at each numerical tabulation are inherently associated with the integration formula used in the numerical process. The numerical technique used in the numerical computation (iterative or extrapolation) has a marked effect on these errors. The propagation of these errors has been explained in the previous sections of this thesis, but the actual source of these errors has not been discussed. This section will elaborate on these errors.

5.0. Roundoff Errors.

The process of numerical computation involves the use of present and past information to obtain present or future information. The process can be integration, differentiation, extrapolation or any other mathematical operation. The probability of having a roundoff in the result is dependent upon the coefficients multiplying the past information and the probability of roundoff in the past data. The maximum roundoff error will occur when the algebraic sum of the coefficients and the roundoff in the past ordinates or differences are a maximum. An indication of the magnitude of the maximum roundoff error that can occur is the sum of the absolute values of the coefficients multiplied by the roundoff. The probability of having a roundoff error in a sum of numbers is dependent upon the coefficients multiplying the numbers as explained in Appendix B. The roundoff error committed at any step can be assumed random if the variable changes more than the magnitude of the maximum roundoff between tabulation intervals. Although

the roundoff errors committed at each interval are comparable regardless of whether iteration or extrapolation techniques are used, the buildup of the error can become prohibitive in an extrapolation process. This buildup of the error will be explained in section 5.3.

5.1. Truncation Errors.

The magnitude of the truncation error is an indication of the accuracy of the numerical mathematical process of the variable to the true mathematical process. In the case of integration when a Newtonian integration formula is used the truncation error committed is approximately equal to the first neglected term in the formula, but if an infinite number of terms are used then the process is exactly equivalent to a true integration. The more rapidly the function to be integrated varies, the more terms of the formula must be used to reduce the truncation error. The errors committed in the various Newton formulas when the first, second, or third differences are neglected are discussed in Chapter 3. This time domain approximation to the error committed gives an indication of the error committed, but the signal frequency spectrum that the formula can handle is best expressed in the frequency domain. In the analog computation field, the use of transfer functions is prevalent and the numerical analyst may find it useful to look at the numerical processes in the frequency domain. The closeness of the various integration formulas to 1/s when expressed in the frequency domain is an indication of the adequacy of the formula to the functions being integrated.

The Newton Gregory formula with no differences reduces to the simple integration formula

$$\theta_n = \theta_{n-1} + T\dot{\theta}_n \qquad 5.1$$

which in the frequency domain becomes[7]

$$\frac{\theta(s)}{\dot{\theta}(s)} = \frac{T}{1 - e^{-sT}} . \qquad 5.2$$

When one difference is used the integration formula reduces to the trapezoidal rule

$$\theta_n = \theta_{n-1} + \frac{T}{2}(\dot{\theta}_n + \dot{\theta}_{n-1}) \qquad 5.3$$

and when two differences are used the integration formula becomes

$$\theta_n = \theta_{n-1} + \frac{T}{12}\left(5\dot{\theta}_n + 8\dot{\theta}_{n-1} - \dot{\theta}_{n-2}\right) \qquad 5.4$$

The corresponding transfer functions for these integration formulas become

$$\frac{\theta(s)}{\dot{\theta}_s} = \frac{T}{2}\left(\frac{1 + e^{-sT}}{1 - e^{-sT}}\right) \qquad 5.5$$

and

$$\frac{\theta(s)}{\dot{\theta}(s)} = \frac{T}{12}\left(\frac{5 + 8e^{-sT} - e^{-2sT}}{1 - e^{-sT}}\right), \qquad 5.6$$

respectively.

Each of the above integration formulas should reduce to 1/s in the range $s = -j\Omega/2$ and $s = +j\Omega/2$ where $\Omega = 2\pi/T$ if true integration is to be performed. By expanding e^{-sT} into a series, Eqs. 5.2, 5.5, and 5.6 can be reduced to the following when sT is made small enough so that the higher order terms can be neglected.

$$\frac{T}{1 - e^{-sT}} = \frac{T}{sT\left(1 - \frac{sT}{2!} + \frac{(sT)^2}{3!} - \ldots\right)} \qquad 5.7$$

$$\frac{T}{2}\,\frac{1 + e^{-sT}}{1 - e^{-sT}} = \frac{T}{2sT} \cdot \frac{2 - sT + \frac{(sT)^2}{2!} \ldots}{1 - \frac{sT}{2!} + \frac{(sT)^2}{3!} \ldots} \qquad 5.8$$

$$\frac{T}{12}\,\frac{5 + 8e^{-sT} - e^{-2sT}}{1 - e^{-sT}} = \frac{T}{12sT} \cdot \frac{12 - 6sT + \frac{4(sT)^2}{2!}}{1 - \frac{sT}{2!} + \frac{(sT)^2}{3!} - \ldots} \qquad 5.9$$

Equations 5.7, 5.8, and 5.9 will approximate 1/s as follows.

$$\frac{T}{1 - e^{-sT}} = \frac{1}{s}\,\frac{1}{1 - \frac{sT}{2!} + \frac{(sT)^2}{3!} - \frac{(sT)^3}{4!} + \ldots} \qquad 5.10$$

$$\frac{T}{2}\,\frac{1 + e^{-sT}}{1 - e^{-sT}} = \frac{1}{s}\,\frac{1 - \frac{sT}{2} + \frac{(sT)^2}{(2)(2!)} - \frac{(sT)^3}{(2)(3!)} + \ldots}{1 - \frac{sT}{2} + \frac{(sT)^2}{3!} - \frac{(sT)^3}{4!} + \ldots} \qquad 5.11$$

$$\frac{T}{12}\,\frac{s + 8e^{-sT} - e^{-2sT}}{1 - e^{-sT}} = \frac{1}{s}\,\frac{1 - \frac{sT}{2} + \frac{(sT)^2}{3!} + (sT)^4 - \ldots}{1 - \frac{sT}{2} + \frac{(sT)^2}{3!} - \frac{(sT)^3}{4!} + \ldots} \qquad 5.12$$

The results indicate that Eq. 5.9 gives the smallest error, Eq. 5.8 the next smallest and Eq. 5.7 the largest error. These results are consistent with the results from the time domain analysis. For a fixed maximum frequency signal Eq. 5.7 can have the same accuracy as Eq. 5.9

if the interval of tabulation T for Eq. 5.7 is decreased by a sufficient amount. When the numerical integration formulas are expressed in the frequency domain, the ability to handle a certain signal frequency spectrum with a specified accuracy becomes evident.

5.2. Extrapolation.

Numerical analysts use two basic methods of approach to obtain a numerical solution to the type of equations solved on the M.I.T. Flight Simulator. The two methods are the iteration and extrapolation methods. In the iteration technique, the equations are combined and arranged so that if a value for one variable is assumed, then all other variables can be obtained by numerical integration or other mathematical relationships. After the value for the first variable is chosen and the values for all the other variables have been computed, one iteration is completed when the newly obtained variables are replaced in the original equations to determine if a proper choice for the assumed quantity was made. If the two values (the assumed values and the values obtained by the numerical computations) do not agree within the limits of accuracy specified, then another iteration is performed. Not until an agreement within the tolerances specified is reached does the computer advance to the next interval. This process, which can be likened to a servo system, can use a crude integration formula, since the feedback loop (the original equations) is the most accurate measuring device that is available. The buildup of the errors cannot be great since each integrated value is substituted into the original equations for verification at each interval.

In the extrapolation method the values of the variable at previous intervals are weighted to obtain the value of the variable at the present interval. A relatively high accuracy formula which is not susceptible to roundoff errors must be used since no verification is performed at each interval. An error can be made at one interval and will not be corrected as in the iterative processes.

The buildup of errors has usually limited the number of consecutive steps that can be handled by extrapolation techniques. Wong[8] has shown that the choice of extrapolation formulas, so that the buildup of errors will at most be linear, must have the absolute magnitude of the largest coefficient multiplying the various past information around unity. As an illustration the following simultaneous linear differential equations were solved:

$$\begin{aligned} \dot{x} &= x - y - 1 + 2t \\ \dot{y} &= 2x - y + 3t + 1 \\ x(0) &= 1, \quad y(0) = 0 \end{aligned} \tag{5.13}$$

using the extrapolation formula below:

$$\begin{aligned} X_n &= X_{n-4} + X_{n-3} - X_{n-1} + \dot{TX}_{n-3} + \dot{TX}_{n-1} \\ Y_n &= Y_{n-4} + Y_{n-3} - Y_{n-1} + \dot{TY}_{n-3} + \dot{TY}_{n-1} \end{aligned} \tag{5.14}$$

After the solution of the equations had been calculated for a number of intervals, the same intervals were recalculated with an error inserted at one interval in x. The difference of the two solutions was the step response to an error. This response shown in Fig. 5.1 exhibits an unusual pattern, which was not expected from the analysis made by using the "H" matrix scheme. The "H" matrix for the set of equations in 5.13 becomes

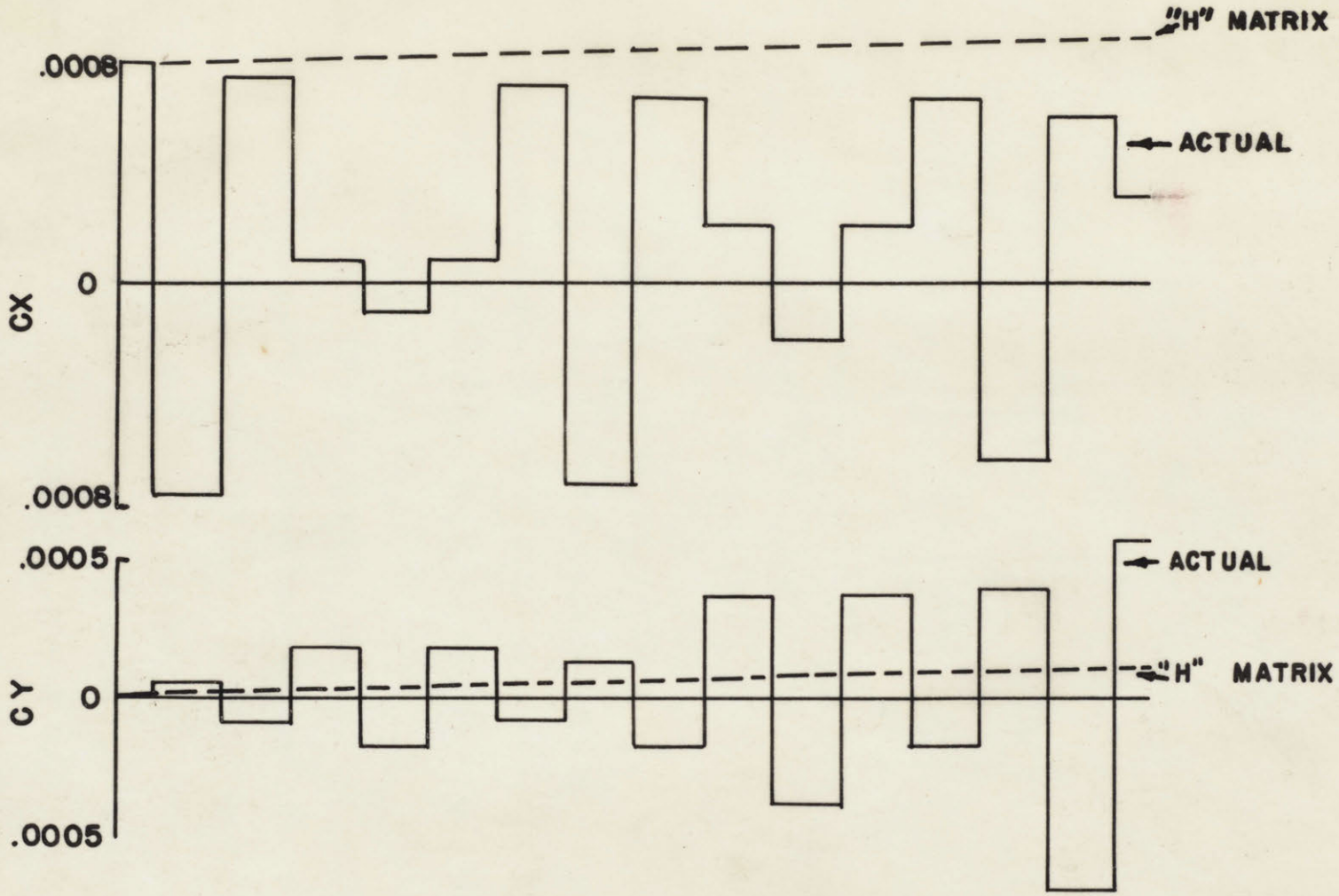


FIG. 5.1 STEP RESPONSE FOR AN EXTRAPOLATION PROCESS

$$H = \begin{bmatrix} 1 & -1 \\ 2 & -1 \end{bmatrix} \qquad 5.15$$

The characteristic roots of the "H" matrix are ±j1. The equations for the step response using characteristic and dual vectors as explained in chapter II become

$$\begin{bmatrix} CX \\ CY \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1-j & 1+j \end{bmatrix} \begin{bmatrix} e^{jt} & 0 \\ 0 & e^{-jt} \end{bmatrix} \begin{bmatrix} \frac{1+j}{2j} & -\frac{1}{2j} \\ \frac{-1-j}{2j} & \frac{1}{2j} \end{bmatrix} \begin{bmatrix} Y_x \\ Y_y \end{bmatrix} \qquad 5.16$$

or

$$CY_x = -\sin t$$
$$CX_x = 2 \sin (t + 45^\circ) . \qquad 5.17$$

The answer to the pattern was that an error in one variable was not inserted once, but four times by the extrapolation formula. An examination of the extrapolation formula in Eqs. 5.14 reveals that if an error was committed in one interval, then the pattern of the errors in succeeding intervals would be as follows: 1, -1, 1, 0, 0, 0, 1, -1, 1, 0, 0, 0, . . The pattern repeats itself after 6 intervals. The effect of the errors in the derivatives was neglected since these errors were diminished by the interval of tabulation (.03). The actual error in x would be the running sum of the errors committed at each step, therefore the extrapolation formula used has a linear buildup of errors. The smoothing effect due to decreasing the interval of tabulation which occurs in an iteration process was not present in the extrapolation technique. In fact the more steps required for a solution the more likely the accumulation of a large error when using an extrapolation technique.

The iteration process may require considerable time when a slow-speed computing machine is used and a large amount of iteration is required at each step. On the other hand the buildup of errors when using an extrapolation technique may render a computation useless if a large number of steps are required in the solution. A compromise between the two would be to use iteration, for example, at every tenth step and extrapolation at other times. A watch on the differences would indicate when a computation is becoming useless due to the inadequacy of the numerical process. The type of equations solved on the M.I.T. Flight Simulator is such that the errors cannot build up to an extremely large value.

Chapter VI

SUMMARY AND CONCLUSIONS

A correction can be found to compensate for the errors committed at each interval of tabulation. The errors present at each interval must be known to obtain the actual correction, but for error analysis purposes the obtaining of the step response is sufficient. These sensitivity functions indicate the susceptibility of the solution of the differential equations to an error in any of the variables. This scheme is not a catholicon for errors in a numerical solution, but is rather a method of assessing the adequacy of the solution for the purposes of checking the M.I.T. Flight Simulator.

6.0. Location of Characteristic Values on Complex Plane.

The behavior of the errors can be obtained by a cursory examination of the location of the roots of the characteristic equations in the complex plane. A root in the left half plane indicates that the errors will die out, whereas a root in the right half plane indicates that the errors will increase with time. The root locations change with time as well as with parametric changes of the differential equations. The movement of the root locations is towards the right half plane as the problem progresses for the type of equations simulated on the M.I.T. Flight Simulator. Although errors committed near the end of the solution do not die out, the buildup in the remaining time can not become excessive and the solution is still valid at the end.

6.1. Application to Study of Machine Errors and Approximation in Equations.

The scheme used in this thesis is not limited to a study of errors in numerical solutions, but can also be used to study the effects of computing machine errors in the solution. The effects of parametric changes as well as the effect of approximations used in the simulation of the equations can also be studied by this scheme.

6.2. Error Analysis Before and During Operation.

The weighting function method requires a considerable amount of analog computer time at a premium cost to perform an error analysis study. A reduction in the time required for a high order system can be accomplished by using the scheme outlined in this thesis to note the effects of errors in the various sections to the output quantities of the sections. For example, the effects of errors in the aerodynamic section on the output quantities of the aerodynamic section. The weighting function method could then be applied to note the effects of the output quantities on the final solution.

6.3. Extrapolation Versus Iteration.

The iterative method of numerical computation seems superior for the procurement of check solutions for the M.I.T. Flight Simulator. The extrapolation technique has a buildup of errors which cannot be tolerated for the hand computation of check solutions. The buildup of errors would necessitate carrying many more figures than those required for the final solution.

6.4. Use of Digital Computers.

The time required for a hand calculated solution is such that the maximum benefit is not derived from the check solution. The use of digital computers will be required to obtain the numerical solutions for the M.I.T. Flight Simulator if the magnitude of the problems becomes larger than that of the present time. The use of iteration processes is not as suited to use on digital computers as is the extrapolation process. The large number of tabulation intervals required for a check solution will require that some verification or iteration be performed at periodic intervals if an extrapolation technique is used. The need for a digital computer in the laboratory is exigent. The use of presently available digital computing facilities could be used, but the most desirable condition would be to have a digital computer in the computing section of the laboratory so that check solutions as well as error analysis could be performed in conjunction with the operation of the M.I.T. Flight Simulator.

APPENDIX A

STEP RESPONSE BY SOLVING LINEARIZED DIFFERENTIAL RESPONSE EQUATIONS ON AN ELECTRONIC ANALOG COMPUTER

The solution of the variation equations to obtain the step response entails a large amount of hand computation. The step response can be obtained by using a small amount of analog computing equipment. The autopilot servo simulator section[9] is ideally constructed to obtain these error sensitivity functions. This d-c equipment can be set up easily and rapidly. For example, the equations for the sample problem

$$\dot{C_Y} = A(C\theta)$$

$$\dot{C_X} = B(C\theta)$$

$$\dot{C\psi} = C(C_Y) + D(C_X) + E(C\psi) + F(C\theta)$$

$$\dot{C\theta} = C\psi$$

can be set up as shown in Fig. A.1. The necessary equipment consists of a few summing circuits, coefficients, and integrators. The number of integrators required is the same as the number of equations. The coefficients are the values of the elements of the "H" matrix, which are varied for the different sections of the problem.

The step response can be obtained by applying a pulse furnished by the sequence timers to the input of the integrators and recording the quantities at the points of interest. A set of these step responses is then obtained very rapidly and the coefficients are changed to correspond to the coefficients of the "H" matrix for a different section. In this manner the step response can be obtained to a problem before the setup

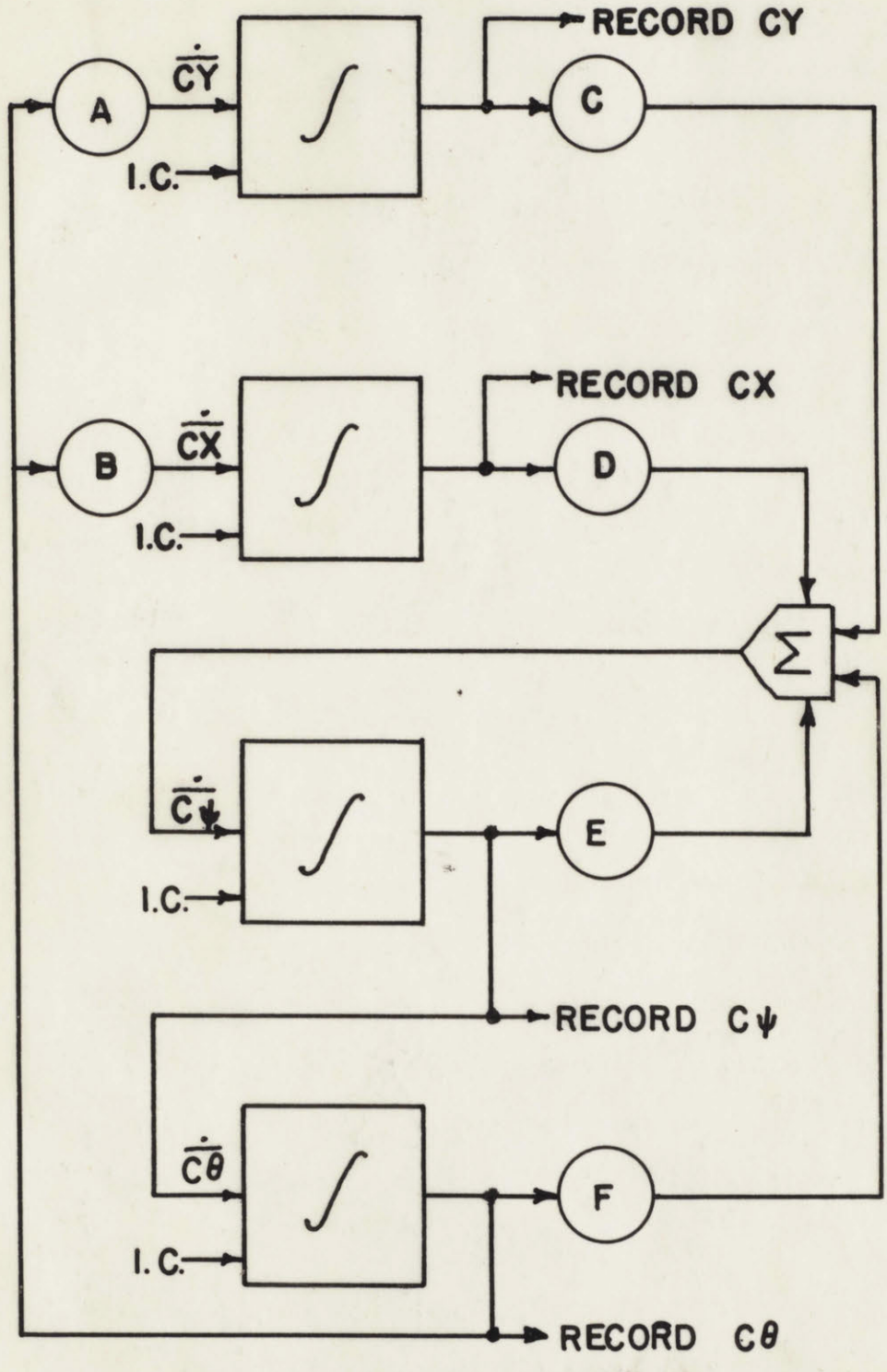


FIG. A.1 AUTOPILOT SERVO SIMULATOR SETUP TO OBTAIN STEP RESPONSES

stage of the problem has begun. The critical portions of the setup are found and attention is focused at these points during the simulator operation. A check solution must be available so that the coefficients of the d-c equipment can be set.

APPENDIX B

PROBABILITY OF A ROUNDOFF ERROR IN A SUM OF NUMBERS

What is the probability of introducing a roundoff error in a sum of numbers AX + BY + CY + ... when A,B,C,..... are exact numbers and X,Y,Z,..... are numbers which have roundoff errors in them? The probability of an error in the sum of two numbers AX + BY will be found and then extended to include the sum of several numbers. If the probability of having a roundoff a in X and b in Y is P(a) da and Q(b) db, respectively, then the probability of having a roundoff of s in the sum will be the probability of having a roundoff of a in X and a roundoff of s-a in Y occurring at the same time. A theory of compound probability states that if P_1 is the probability that event E_1 occurs, and P_2 is the probability that event E_2 occurs, then P_1P_2 is the probability that both events occur in either order or simultaneously if E_1 and E_2 are independent.[10]

The probability of a roundoff of s in the sum will be P(a) Q(s-a) for a chosen value of a, but for all possible values of a the integral of the product is taken.

$$R(s) = \int P(a)\, Q(s-a)\, da$$

The convolution of the two probability density curves of P(a) and Q(b) gives the probability density curve for the sum of two numbers.

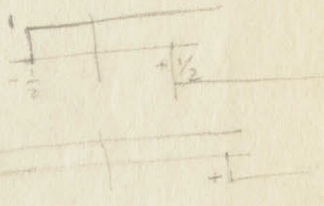
To obtain the probability density curve for the cases of more than the sum of two numbers, the probability density curve if first obtained for the sum of the two and then the resultant curve is convolved with the probability density curve of the following number. This process is

repeated for as many times as there are terms in the sum. The effect of the multiplying coefficients A, B, C, ... is to multiply the base of the probability density curve by the coefficient and to divide the ordinates of the curve.

As an example, the probability of having a roundoff error in the following sum will be found.

$$X + 2Y + 3Z$$

The numbers X, Y, and Z are rounded to the nearest whole number and the probability density curve is uniform between the limits of $\pm 1/2$. The probability density curve for the roundoff in X + 2Y becomes



$$\mathcal{L}^{-1}\left[\left(\frac{1 - e^{-sT}}{s}\right)\left(\frac{1 - e^{-2sT}}{2s}\right)\right],$$

since convolution in the time domain is multiplication in the frequency domain. Then the probability density curve for the roundoff in (X + 2Y) + 3Z becomes

$$\mathcal{L}^{-1}\left[\left(\frac{1 - e^{-sT}}{s}\right)\left(\frac{1 - e^{-2sT}}{2s}\right)\left(\frac{1 - e^{-3sT}}{3s}\right)\right].$$

APPENDIX C

REFERENCES

1. Hall, A. C., "A Generalized Analogue Computer for Flight Simulation", AIEE Preprint, 50-48, January, 1950.

2. Murray, F. J., and Brock, P., "Planning and Error Analysis for the Numerical Solution of a Test System of Differential Equations on the IBM Sequence Calculator", (Project Cyclone Report to Special Devices Center of Office of Naval Research, Contracts. Nos. N6ori-128 and N6140s-2890B).

3. Jones, T. F., Jr., The Propagation of Errors in Analog Computers, MIT Thesis for ScD in EE, 1952.

4. Rabow, G., Evaluation of Errors Occurring in the M.I.T. Flight Simulator, MIT Thesis for SM in EE, 1952.

5. Milne, W. E., Numerical Calculus, (Princeton University Press, Princeton, New Jersey, 1949).

6. Kopal, Z., Unpublished Notes for course 6.631.

7. Linvill, W. K., Unpublished Notes for course 6.54.

8. Wong, D. W., Digital Computer Solutions of Boundary Value Problems. MIT Thesis for SM in EE, 1952.

9. Johnson, E. C., An Electronic Simulator for Non-Linear Servomechanisms, MIT Thesis for SM in EE, 1949.

10. Reddick, H. W., and Miller, F. H., Advanced Mathematics for Engineers, (John Wiley and Sons, Inc., N. Y., 1948).

11. Schlesinger, F., "On the Errors in the Sum of a Number of Tabular Quantities", Astronomical Journal, Vol. XXX, pp. 183-190.