Planning and Error Considerations for the Numerical Solution of a System of Differential Equations on a Sequence Calculator

1. Introduction.—This paper deals with the solution of a specific system of fourteen ordinary differential equations,

$$(1) \qquad z_i' = f_i(z_1, \cdots, z_{14}, t), \qquad \text{where} \qquad i = 1, 2, \cdots, 14.$$

The system was solved for the United States Navy Special Devices Center on the IBM Selective Sequence Electronic Calculator. A discussion of basic theoretical planning and error evaluation will be presented herein. The complete solution process will be described in a joint report with PAUL BROCK of the Reeves Instrument Company, which it is hoped will be available from the Special Devices Center. The present theoretical discussion outlines reasons for the choice of method, the detailed step-by-step process decided upon, the considerations which led to choice of interval length, and those which are concerned with the choice of scale.

There were three types of error in the solution process, namely, the truncation error, the rounding error, and the error due to the presence of nonanalytic functions in the f's. It is possible to obtain the immediate error of each type and to develop a uniform method of evaluating the complete effect of each such error. In the case of the truncation error, there is a resonance effect, a build-up of error, which is different in nature from instability, as this system is stable in the usual sense. The rounding error is treated by probability methods. However, the present treatment of this error is somewhat simplified since the effect is small. The nonanalyticities require a special technique, which will be discussed later. The methods developed in this paper, with this specific example in mind, differ considerably from those of RADEMACHER.[1]

2. Equations.—The f_i's in this system can be calculated from a given set of values, $z_1, \cdots, z_{14}$, and t by means of the elementary operations of arithmetic and the functions: square root, absolute value, sine, cosine, and arcsine. A series expansion was used for the arcsine; Newton's Method was used for the square root; and a table was used for the sine and cosine evaluation. The nonanalyticities appeared because of expressions which occurred in the form, $|z_i - k|$, when certain of the dependent variables, z, pass through specific values. The initial values were in a region of analyticity, but nonanalyticities occurred eleven times during the solution process.

3. General Method.—Consider f_j as a function of time. Suppose f_j has been evaluated at $t - h$, $t - 2h$, $t - 3h$, and $t - 4h$. A cubic in time is passed through these values and the polynomial integrated between $t - h$ and t. The result is a value for Δz_j, the increment in z_j between $t - h$ and t. In the usual Milne method, the value, z_j, obtained at t is substituted in the f's and a new cubic formed which goes through the values of f at t, $t - h$, $t - 2h$, and $t - 3h$. This in turn is integrated between $t - h$ and t to yield a new estimate for Δz_j. The new value, z_j, can be used in the same way to obtain another value of Δz_j and the process continued until there is no

significant difference between successive values of z_j. A modification of this method is used. The first step in the above process is referred to as the "open" integration; the remaining steps are called "closed" integration.

The value of Δz_j obtained from the open integration is a simple linear combination of the values of f_j for $t - h$, $t - 2h$, $t - 3h$, and $t - 4h$, i.e.,

$$\Delta z_j = \frac{h}{24}[55f_j(t-h) - 59f_j(t-2h) + 37f_j(t-3h) - 9f_j(t-4h)]. \tag{2}$$

For a closed step

$$\Delta z_j = \frac{h}{24}[9f_j(t) + 19f_j(t-h) - 5f_j(t-2h) + f_j(t-3h)]. \tag{3}$$

4. Step-by-Step Error.—For simplicity, let $t = 0$ in the above equation. The usual remainder argument shows that, if P is a polynomial of the third degree which equals f_i at $t = -h, -2h, -3h$, and $-4h$, then for $-h \leqq t \leqq 0$

$$f_i - P = f_i^{(\mathrm{IV})}(x')\frac{(t+h)(t+2h)(t+3h)(t+4h)}{4!} \tag{4}$$

for some x' between $-4h$ and 0. Suppose $f_i^{(\mathrm{IV})}$ is a constant and integrate between $-h$ and 0; then, the error, $\epsilon_i^{(0)}$, in Δz_i of the open integration is

$$\epsilon_i^0 = .349 f_i^{(\mathrm{IV})} h^5. \tag{5}$$

If the correct value for f_i at $t = 0$ were known, a similar argument for the closed integration step would yield the error in the closed steps, i.e.,

$$\epsilon_i^c = -.0264 f_i^{(\mathrm{IV})} h^5. \tag{6}$$

However, in the first closed integration step for Δz_i, the following expression holds:

$$\Delta z_i^{c1} = \frac{h}{24}[9f_i^{(0)}(t) + 19f_i(t-h) - 5f_i(t-2h) + f_i(t-3h)], \tag{7}$$

where $f_i^{(0)}(t)$ is obtained by substituting the values of z involving the open integration error. Consequently, the first closed integration step involves an error

$$\epsilon_i^{c1} = \epsilon_i^c + \frac{9h}{24}\sum_{\alpha=1}^{14}\frac{\partial f_i}{\partial z_\alpha}\epsilon_\alpha^0. \tag{8}$$

Now, if h is sufficiently small, the second term is less important than the first. Consequently, if $f_i^{(\mathrm{IV})}$ is essentially constant, ϵ_i^0 and ϵ_i^{c1} differ in sign, and the values for the open integration and the first closed integration for Δz_i span the true value. The difference between these values forms an upper bound and a good estimate for the truncation error.

It may be noted that the same argument shows that the error, ϵ_i^{ck}, in the k^{th} step can be expressed in terms of the error, ϵ_i^{ck-1}, in the previous step, i.e.,

$$\epsilon_i^{ck} = \epsilon_i^c + \frac{9h}{24}\sum\frac{\partial f_i}{\partial z_\alpha}\epsilon_\alpha^{ck-1}. \tag{9}$$

This forms an easy basis for convergence considerations. The final convergence has an error which satisfies the equation,

$$\epsilon_i^{c\infty} - \frac{9h}{24} \sum \frac{\partial f_i}{\partial z_\alpha} \epsilon_\alpha^{c\infty} = \epsilon_i^{c}. \tag{10}$$

5. Actual Method Used.—The f's are complicated functions from the computational point of view, and their evaluation consumes much machine time. With this in mind, consider the relatively simple method given above. In view of equation (10) and the machine time required to evaluate the f's, there seems to be no point in carrying through the iterative process to the bitter end. In fact equations (8) and (10) suggest that considerations can be confined to an open and closed step. However, the price of an open and closed step in machine time is the same as the price of two open steps. It is possible to compare the error which follows from an open and closed step with interval, h, and the error of two open steps with interval, $\frac{h}{2}$. The latter can be obtained by dividing .349 in equation (5) by 16 and comparing the result, .022, with .0264 of equation (6). Thus, the open integration procedure, which is simplest from the coding point of view, is the best. (This argument is based on the use of cubic approximations.) Even if this situation were reversed, it would be better to choose the open integration method since a smaller step interval is used, and, hence, error estimates are more dependable. Similar objections may be raised to any other system which uses a larger h.

However, certain other considerations must be made. The choice of the interval, h, is necessarily based on the initial information given in the problem. Later in the run, h may be too large or too small. This can only be established by knowing the truncation error, which can be indicated by the difference between a closed and open step, as has been shown above. The procedure is to use an open integration at each step but, at every tenth step, to follow it by a closed integration. At each tenth step, the difference between the open and closed steps can be obtained from the printer in the sequence calculator, and one can infer from this whether the truncation error is tolerable.

If the truncation error is small enough, one is justified in increasing h by an integral multiple. This is readily done using the previously computed values of f. The coding for the problem was set up in such a way that one could start by feeding in cards with the properly spaced values of f for four points. To increase h then by an integral multiple, it was only necessary to take certain cards previously punched by the machine and feed them in. It was not necessary to decrease h, but if necessary, intermediate values of f could have been obtained by interpolation involving four differences.

6. Choice of h.—From physical reasoning and also from an actual consideration of the variational equation at the initial part of the run, it was determined that the system of equations is stable in the absolute sense. When an error is made, the effect of this error does not increase indefinitely, but, instead, dies down at a very slow rate.

Initially, the possibility of a run to $t = 100$ had to be planned for. Calculations based on the initial situation and these considerations indicated that

the interval, $h = .04$, is too large and that $h = .02$ would be about right. But the cost of the solution for $h = .02$ was prohibitive. Furthermore, it was felt that the assumption that all the errors would build up in the same direction was too conservative, and, hence, it was decided to use the interval, $h = .04$. Due, however, to the resonance effect which will be discussed later, the errors do effectively build up in the same direction so the stated reason is not sound. On the other hand, $f^{(IV)}$ does decrease as evidenced by the truncation error, and, while the total accuracy was not quite that planned, the result was adequate.

To begin the process, it is necessary to compute the values of f's at four points. These initial values were computed by finding the corresponding values of z by means of Taylor's series and computing the f's. Since it is relatively easy to increase h, an initial value of $h = .001$ was chosen, and the f's were computed for $t = .000$, $t = .001$, $t = .002$, and $t = .003$. Using $h = .001$, the computation process was carried to $t = .100$. This resulted in values at $t = .00$, $t = .02$, $t = .04$, and $t = .06$, which were used to start the run with $h = .02$. At $t = .08$ the values for z using the short interval, $h = .001$, and also an open step value for $h = .02$ and a closed step value were then available. Of course, the smaller h value yields a much more accurate value for z. It is interesting to observe that for every z, the accurate result lay between the open and closed values.

The $h = .02$ run was continued until about $t = 2.18$, and the $h = .04$ run began with $t = 2.10$. Again an overlap was permitted for checking purposes in order to be sure that the change of interval was correctly made. It was found that the shorter interval value again lay between the open and closed value for the longer interval.

The interval, $h = .04$, was continued until $t = 68$, where one of the variables, z, ran off scale, and the solution was no longer of interest. During an earlier false run, an attempt was made to increase the interval to $h = .12$, but this yielded large truncation errors. Therefore, the interval, $h = .04$, was used again. During the final true run, no such increase was attempted.

After the $h = .04$ run, the $h = .02$ run was repeated and continued in order to check the error of the $h = .04$ run. Unfortunately, only the run from $t = 2.10$ to $t = 2.90$ was duplicated at an interval, $h = .02$. However, this provided a valuable method of checking the error theory developed later.

7. Scaling Questions and Watches.—The registers in which a number can be entered contain 9 digit-positions. The scale for a quantity must be chosen so that the effect of truncation error can be clearly seen. On the other hand, some leeway against spill-over must be available. Consequently, the size of each quantity must be estimated in advance.

For this estimation, a certain amount of physical information is available. A set of assumptions, based on this information, were made concerning the sizes of the quantities involved, and values for all other quantities were estimated. However, the assumptions were not based on completely reliable information, and it was necessary to verify that, during the course of the computation, these assumptions remained valid. This was done by a system of "watches." Certain of the quantities were printed by the machine, and others could be obtained readily from printed results. During the complete calculation, up to the time the calculation was stopped, the assumptions held.

8. Hand Computations.—The Taylor's series for the z's had to be obtained by hand computation. This was done in triplicate, as it was believed that the error rate was approximately about 1 in a 100 calculation steps. Since the amount of work was of the order of magnitude of 10,000 steps, a calculation in duplicate would have too high a probability for a duplicate error.

In addition, it was also necessary to provide a hand computation of two steps of the machine computation for checking purposes. This was done for the $t = .004$ and $t = .005$ steps, but unfortunately many values appeared as zeros necessitating the provision of further material for checking purposes. In addition, a complete machine step consisting of an open and closed step at the later value, $t = 6.44$, was duplicated by hand and the values checked against the machine values. In these checking computations, the

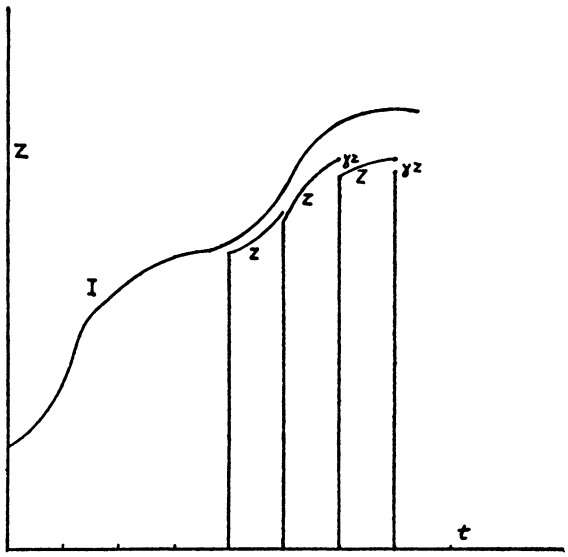

Illustration 1. Solid curve I is true solution. The broken curve is z as defined by equation (11). The jumps are γz.

same rounding procedures as those of the machine were used. Hence, a check demands precise digital agreement to every digit which appears.

Nevertheless, this numerical checking was inadequate to catch one error. It should be pointed out that a specific numerical calculation is logically of too low a type to constitute a complete coding check. A complete check should start with the coding and reconstruct the original mathematical relations. Thus, it is clear that a correct coding system is required.

9. Definition of the Error.—It has been mentioned that there are three types of errors inherent in the solution. Associated with every error there are two problems. It is necessary to estimate the error at the point where it is made. It is also necessary to evaluate the effect of the error on those portions of the solution which follow in time the point at which an error is made. Let the vector $z = (z_1, \cdots, z_{14})$ denote the computed solution. Let $(cz_1, \cdots, cz_{14})$

denote the correction vector which, when added to z, will yield the true solution. The computed solution exists only at the points at which a computation was made. Let z be defined between such points by the condition that it satisfies the equations

$$z_u' = f_u(z_1, \cdots, z_{14}, t) \tag{11}$$

and is continuous at the left-hand end point of each such interval. At the right-hand end point of any interval, z has a discontinuity corresponding to the error made at the right-hand point. (See Illustration 1.)

At each solution point the total error is made up of two parts. The first part, termed the rounding error, is the difference between the computed

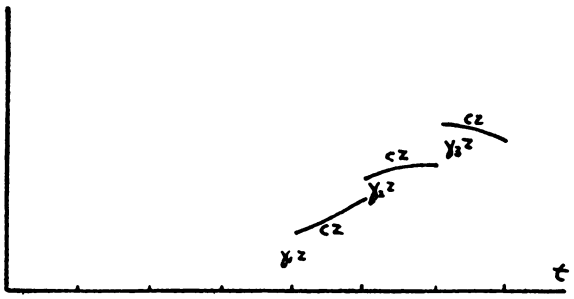

Illustration 2.

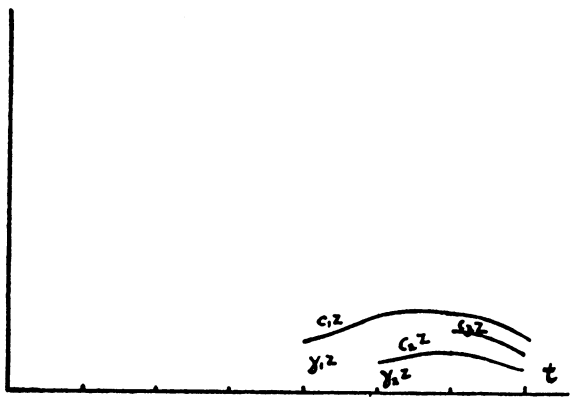

Illustration 3.

solution and the computationally perfect result of the same computation made with no rounding errors. The second part is the difference between the computationally perfect result and the result of a correct integration between the previous step and the present one. If the points involved in the integration process do not span a point of nonanalyticity for the f being integrated, the second part will be termed the truncation error. If, on the other hand, there is a point of nonanalyticity, for the f being integrated, present in the span covered by the integration process, this error will be called a nonanalyticity error. The distinction between truncation errors and non-

analyticity errors is set up purely with the situation of the present practical problem in mind and appears to be the most convenient in this case.

The total correction vector at the α computational point corresponding to the total error made at that point is denoted by $\gamma_\alpha z$; and the truncation correction, the rounding correction, and the nonanalyticity correction vectors for the corresponding errors are denoted by $\gamma_\alpha' z$, $\gamma_\alpha'' z$, and $\gamma_\alpha''' z$, respectively.

10. Differential Equation for the Error.—Since $z + cz$ is the desired solution of the system of equations, the following relationship holds, i.e.,

$$(z_u + cz_u)' = f_u(z_1 + cz_1, \cdots, z_{14} + cz_{14}, t). \tag{12}$$

For the intervals between points of computation, z also satisfies the equations (11), and, consequently,

$$(cz_u)' = f_u(z_1 + cz_1, \cdots, z_{14} + cz_{14}, t) - f_u(z_1, \cdots, z_{14}, t). \tag{13}$$

As z is now considered to be known, equation (13) is a differential equation on the vector, cz, which holds between points of computation. At points of computation, cz has a jump, γz, corresponding to the error made at this point. This jump condition and equation (13) determine precisely the correction, cz. (See Illustration 2.)

For practical computational reasons, equations (13) are made linear, i.e.,

$$(cz_u)' = \sum_v \frac{\partial f_u}{\partial z_v} cz_v. \tag{14}$$

The coefficients $\partial f_u/\partial z_v$ are regarded as constants over an interval which is large relative to the length of a step. Undoubtedly, this introduces an error into the error computation. However, consider the result of solving (14) as simply the zero order approximation to a solution of (13) by a PICARD iteration process. The use of (14) can be justified by estimating the corresponding solution of (13). By using higher terms in the expansion of the right-hand side of (13), one can readily justify the use of (14) for the purpose of attaining the real objective, the estimation of the size of cz_u.

One effect of the linearization of the differential equations is that the correction vector, cz, for the truncation error, rounding error, and nonanalyticity error can be separately computed, provided one error is assumed to have a negligible effect on the others. This is clear for an estimate of the truncation error and nonanalyticity error. This is also justified for the rounding errors since only maxima and variances for these errors are computed.

Thus, a truncation correction, $c'z$, can be defined as satisfying (14) between computation points and having a jump, $\gamma_\alpha' z$, at the α computation point. A similar definition holds for $c''z$ and $c'''z$, the rounding and nonanalyticity corrections. The sum $c'z + c''z + c'''z$ satisfies (14) between computation points and has the required jump, $\gamma_\alpha z$, at the α computation point.

A function, $c_\alpha' z$, is introduced which is zero before the α computation point, has a jump, $\gamma_\alpha' z$, at this point, and satisfies (14) after this point (see

Illustration 3). The definition of $c_\alpha''z$ and $c_\alpha'''z$ is similar, and one can readily show that

$$c'z = \sum_\alpha c_\alpha'z, \qquad c''z = \sum_\alpha c_\alpha''z, \qquad c'''z = \sum_\alpha c_\alpha'''z. \tag{15}$$

11. Error Effect Procedure.—In view of formula (15), the immediate problem is to compute $c_\alpha'z$. As was previously mentioned, intervals will be considered over which the equations (14) can be regarded as linear differential equations with constant coefficients. The following equation can then be written in vector form,

$$(cz)' = Hcz. \tag{16}$$

This equation is solved by finding the characteristic roots, $\lambda_1, \cdots, \lambda_{14}$, of H and the corresponding characteristic vectors. Although zero is a double characteristic root, there are 14 characteristic vectors, $x^{(1)}, \cdots, x^{(14)}$. The general solution is in the form,

$$y = \sum k_i e^{\lambda_i t} x^{(i)}. \tag{17}$$

Now $c_\alpha'z$ is the solution which at $t = t_\alpha$ has the value $\gamma_\alpha'z$. Let $X^{(1)}, \cdots, X^{(14)}$ be the set of dual vectors to the characteristic vectors $x^{(1)}, \cdots, x^{(14)}$, i.e.,

$$x^{(i)} \cdot X^{(j)} = \delta_{ij}. \tag{18}$$

Then

$$c_\alpha'z = \sum_i \gamma_\alpha'z \cdot X^{(i)} e^{\lambda_i (t - t_\alpha)} x^{(i)}. \tag{19}$$

Of course, the λ_i's, $x^{(i)}$'s, $X^{(i)}$'s assume complex values, but $c_\alpha'z$ will be real since $\gamma_\alpha'z$ is real. A similar formula holds for $c_\alpha''z$ and $c_\alpha'''z$.

In computing the λ_i's, a further assumption was made which permitted one to factor the determinant $|H - \lambda|$ into two quintics, a quadratic, and λ^2. The quintics were solved by finding a real root by NEWTON'S Method, and the remaining quartics were solved by a method of LIN.[2]

The process by which the characteristic vectors and their duals were obtained was somewhat complicated, since it was desirable to avoid fourteenth-order matrix operations. However, this was specific to the problem and will not be discussed here.

12. Truncation Error.—In view of the above mechanism for evaluating the effect, $c_\alpha'z$, of an error, $\gamma_\alpha'z$, one has the problem of obtaining $\gamma_\alpha'z$ in such a form that the summation process of (15) can be effectively carried out. For $\gamma_\alpha'z$ the difference is found between the open and closed integration values at every tenth step. This was expressed in analytic form. However, there is a more effective way of obtaining $\gamma_\alpha'z$.

The f_i for each i is given. Furthermore, by inspection it is seen that each f_i can be expressed approximately in the form

$$f_i = f_i^0(t) + a_1 e^{-\alpha_1 t} \cos(\beta_1 t + \gamma_1) + a_2 e^{-\alpha_2 t} \cos(\beta_2 t + \gamma_2) + a_3 e^{-\lambda_0 t}, \tag{20}$$

where $f_i^0(t)$ is a slowly changing function, while $\alpha_1 + i\beta_1$, $\alpha_2 + i\beta_2$, and λ_0 are roots of the characteristic equations for H. The truncation correction

which is needed when an open integration is performed on the function, f, is

$$(21)\quad \int_t^{t+h} f\,dt - \frac{1}{24}h[55f(t) - 59f(t-h) + 37f(t-2h) - 9f(t-3h)].$$

This expression for the correction is linear, and if f is expressed as a linear combination of terms in the form $Ae^{-\lambda t}$ (where λ is complex), then for each such term the error is

$$(22)\quad hAe^{-\lambda t}\left[\frac{i - e^{-\lambda h}}{\lambda h} - \frac{1}{24}(55 - 59e^{\lambda h} + 37e^{2\lambda h} - 9e^{3\lambda h})\right] = Ae^{-\lambda t}h\Gamma_0(\lambda h).$$

Thus, in general, if

$$f = \sum_{i=1}^{N} a_i e^{-\lambda_i t},$$

the truncation correction is

$$(23)\quad \gamma z = \sum_{i=1}^{N} h a_i e^{-\lambda_i t}\Gamma_0(\lambda_i h).$$

From the point of view of frequency analysis, $\Gamma_0(\lambda h)$ is exceedingly interesting. For, if $\Gamma_0(u)$ is expressed as a power series in u, we obtain

$$(24)\quad \Gamma_0(u) = .349u^4 + .716u^5 + .789u^6 + .613u^7 + .375u^8 + .191u^9 + .084u^{10} + \cdots \text{ (see footnote 3).}$$

This reflects the fact that the method of integration used was set up to integrate cubic polynomials perfectly. From the frequency standpoint, this means that there is a fourth-order contact at $\lambda = 0$. On the other hand, this suggests an investigation of other integration formulae which would be precise at certain other frequencies particularly chosen for the problem at hand. Such an investigation has been initiated with some interesting results. For instance, an arbitrarily good approximation cannot be found over a connected λh interval of length exceeding 2π no matter how many constants are available in the integration formula; but for a smaller interval such an approximation can be found.

The above formula indicates that the contribution of the low frequency terms to $\gamma' z_i$ is negligible, and indeed only one term makes an effective contribution to $\gamma' z_i$. Consequently, one can obtain $c_\alpha' z$ by (23) and (19). Furthermore, by a summation, based on a geometrical progression, it is possible to find $c' z_j$ using (15).

Since the same frequency appears in both (23) and (19), it is noted that when the summation (15) is carried out, a term in the expansion of cz_α is obtained in the form,

$$(k_1 t + k_2)e^{-\alpha_1 t}\cos(\beta_1 t + \gamma),$$

while the other terms are exponentials without t in the coefficients. As far as this term is concerned, the total error builds up until the rather small exponential, α_1, eventually cuts it down. This was the important term in the error analysis.

Thus even though the errors, $\gamma_\alpha' z$, oscillate in sign, their total effect builds up. The result is analogous to the solution of a differential equation,

$$L(y) = g(x),$$

where $L(y)$ is a linear differential operator with constant coefficients, whose indicial equation has roots with negative real parts and in which $g(x)$ contains a term corresponding to one of these roots. This phenomenon produces a resonance effect.

This resonance phenomenon will occur whenever the original system of equations (1) and the variational equation (16) are close enough so that terms with frequencies associated with the variational equations appear in the f's of (1) regarded as functions of the time. If practical, the method of integration should discriminate against these frequencies.

As was mentioned previously, a stretch from $t = 2.10$ to $t = 2.90$ was available over which there is a solution corresponding to an interval of $h = .04$ and one corresponding to $h = .02$. It is readily seen from the above that the error for the smaller interval should be approximately $\frac{1}{16}$ that for the larger interval. Now the theoretically computed estimate of the error can be compared with the difference between these solutions. The agreement between the two curves was very good in both size and shape.

Since the error is a solution of a system of linear equations, the improvement in $\gamma_\alpha' z$ obtained by halving the interval will be reflected in $c_\alpha z$. Thus, halving the interval is an excellent method of estimating the error when h is small enough to justify the linearization upon which (14) is based. Unfortunately, the cost in machine time is tripled by this procedure.

13. Rounding Error.—To find the correction vector, $\gamma'' z$, for the rounding error, it is necessary to compare the correct calculation with the actual calculation. Naturally, this involves a study of the computation process used for each f. The correction, $\gamma_\alpha'' z_j$, for the j^{th} component is made up of a sum of a number of terms, a few of which are periodic, but most of which can be regarded as chance variables. Since the total effect of rounding error in this problem is not great compared with the other errors, it was felt that certain simplifications were justified. Hence, the $\gamma_\alpha'' z$ was considered to be in the form

$$\gamma_\alpha'' z = (a_1 \epsilon_1^\alpha, a_2 \epsilon_2^\alpha, \cdots, a_{14} \epsilon_{14}^\alpha), \tag{25}$$

where the a_j is a maximum value for the component involved and the ϵ_j^α's are independent evenly-distributed chance variables taking on values between -1 and 1. However, in the analysis, $a_j \epsilon_j^\alpha$ could be replaced by a more precise expression for this component if it were necessary.

It follows from (25), (9), and (15), that the total effect of the rounding error is in the form

$$c'' z = \sum_{\alpha, j} E_{\alpha, j}(t) \epsilon_j^\alpha. \tag{26}$$

The maximum can be estimated by proper summation methods; and it is also possible using SCHWARZ's inequality and these same summation methods to obtain a bound for the variance of $c'' z$.

14. Points of Nonanalyticity.—In those cases in which an f_i was nonanalytic, it could be written in the form

$$f_i = F_i + H_i \operatorname{sign}(z_j - A), \tag{27}$$

where F_i and H_i are analytic and A is a constant. In general, $i \neq j$, and it was a relatively simple matter to compute both the correct integral for the term, $H_i \operatorname{sign}(z_j - A)$ and the value obtained from the method of integration used. There is a discrepancy due to the fact that this method of integration assumes that the integrand is a polynomial while the term $H_i \operatorname{sign}(z_j - A)$ is in general discontinuous. However, this difficulty persists over four points, and the successive errors obtained in this way tend to cancel. Therefore, except for a temporary disturbance, the effect of an individual nonanalyticity is slight. Since there are relatively few of these, the total effect of the nonanalyticities is readily calculated by means of (19) and (15).

15. General Planning Considerations.—Experience suggests that the preliminary analysis of a problem of this sort be based on the matrix, $H = ((\partial f_i/\partial z_\alpha))$, which, of course, can be evaluated initially. The characteristic roots of H indicate the frequencies present. The total error permitted can be divided into three not necessarily equal parts, one for truncation, one for rounding, and one for miscellaneous. In view of the resonance error, the total truncation error should be divided by the expected length of the computation to obtain the error permitted per unit interval, ϵ_0, provided the situation is stable. If the open integration formula is used and if one λ, λ', predominates among the characteristic roots of H, then h, the interval width to be used initially, can be obtained from

$$\epsilon_0 = \Gamma_0(\lambda' h). \tag{28}$$

If more than one λ is important in H from this point of view, ϵ_0 will have to be further subdivided and a portion allocated to each effective root. If some λ has positive real parts, i.e., if the situation is unstable, it is usually permissible to make a change in scale and a similar discussion applies.

After having left the region around the initial value, h can be determined by comparing the result of an open and closed integration at specified intervals. In this connection, it should be pointed out that the quantity, $\Gamma_c(\lambda h)$, which determines the error for a closed integration in the same way as Γ_0 did for an open integration, is given by

$$-(.0264u^4 + .0341u^5 + .0246u^6 + .0129u^7 + .0055u^8 + .0020u^9 + .0006u^{10} + \cdots). \tag{29}$$

The change in sign again indicates that the results of an open and closed integration span the true value, and it is apparent that the difference is an excellent estimate of the truncation error.

16. Conclusion.—It seems clear from the above that a plan of solution should be based on considerations of over-all error rather than upon consideration of error per step. One such plan is given. This plan indicates a method for choosing h initially and also during the course of the solution. The ideas used can be supplemented by algebraic methods to estimate the total error when the solution process has been completed.

Our experience involved a stable system of differential equations, but the use of frequency analysis is justified in most cases, including unstable cases in which a scale change occurs.

F. J. M.

[1] Harvard University, Computation Laboratory, *Annals*, v. 16, p. 176–187. [*MTAC*, v. 3, p. 437.]

[2] SHIH-NGE LIN, "Numerical solution of complex roots of quartic equations," *Jn. Math. Phys.*, v. 26, 1947, p. 279–283.

[3] This formula assumes that the computed and hence available f is the function to be integrated. If one wishes to make an allowance for the distinction between this f and the correct f associated with the true solution, a difference equation must be solved. In case λ is a characteristic root of H, the corrected value Γ_0 is obtained by dividing the above formula (24) by

$$1 + .5000u + .1667u^2 + .0417u^3 - .0972u^4 - .1424u^5 \cdots.$$

This correction is significant.

RECENT MATHEMATICAL TABLES

761[A].—R. COUSTAL, "Calcul de $\sqrt{2}$, et réflexion sur une espérance mathématique," Acad. Sci. Paris, *Comptes Rendus*, v. 230, 1950, p. 431–432.

The first four terms of the binomial expansion of

$$\sqrt{2} = a(1 - 2x)^{-\frac{1}{2}}$$

where

$$a^2 = 2 - 4x$$

and a is an approximation to $\sqrt{2}$, good to 333D, were used to obtain $\sqrt{2}$ to 1032D. Besides this value the author gives the distribution of digits in the 1033S values of $\sqrt{2}$ and $1/\sqrt{2}$. In the first 1000D in the $\sqrt{2}$, the digits 0–9 have the following frequencies

108, 98, 109, 82, 100, 104, 90, 104, 113, 92.

Such a distribution has a chi-square of 8.38. The probability of such a value from a normal distribution is almost exactly 1/2. For $1/\sqrt{2}$ the probability is merely .05. [Compare MTAC, v. 4, p. 109–111].

The author "reflects" on the paradox that if one takes the product of the first 1033 digits of the decimal expansion of a real number x in the interval $0 < x < 1$, the expected value of the product is $(9/2)^{1033} > 10^{674}$, whereas the probability that it is exactly zero is $1 - (9/10)^{1033} > 1 - 10^{-47}$.

D. H. L.

762[C].—H. S. UHLER, "A mathematician's tribute to the state of Israel," *Scripta Mathematica*, v. 14, 1949, p. 281–283.

The author gives ln 173 and ln 5709 to 290D.

763[D, H, L].—C. N. DAVIES, "The sedimentation and diffusion of small particles," R. Soc. London, *Proc.*, v. 200, 1949, p. 100–113.

This paper contains a table of the first 16 positive roots of the equation

$$2\alpha x + \tan x = 0$$