A Method for Integrating a Set of Ordinary Differential Equations Subject to a Type of Numerical Indeterminacy

1. Introduction. This paper proposes the numerical integration of a system of first order ordinary differential equations by means of a convergent sequence of equation systems. The schema suggested may prove useful whenever: (a) the equations of the set are coupled in an appropriate fashion; (b) ordinary point by point procedures fail due to subtraction which makes the analytic formula for the derivative of one of the variables numerically indeterminate.

The technique is adapted to machine use. On a C.P.C. it has solved a set of four simultaneous non-linear equations which describe a hypothetical one-dimensional free-radical flame [1]; on an IBM 701 it has integrated the seven simultaneous non-linear equations for a hydrogen-bromine flame.

2. Character of Equation Systems for Which the Method may be Useful. Certain physical problems such as free-radical flames are described by sets of simultaneous differential equations in which the calculation of one of the derivatives from its analytic formula involves serious loss of numerical accuracy due to subtraction. The conventional solution of such problems presumes the steady state hypothesis which replaces the differential equation by the algebraic approximation that the derivative vanishes identically. Unfortunately, questions such as the importance of free-radical diffusion in flames must be studied without the *ad hoc* assumption of the steady state. (When applied to free-radical flames, this approximation states that the derivative, dG_B/dt, of the fractional mass rate of flow of the free-radical B vanishes identically. Such an assumption must not be introduced but should be proved valid if we seek to establish the importance of diffusion.) Furthermore, it is generally important to determine the accuracy of such an approximation.

An outline of the qualitative features of one type of equation system which presents this numerical problem will clarify the essential character of the difficulty one meets when he applies traditional point-by-point procedures without assuming the steady state. Consider, therefore, a set of M simultaneous differential equations in an independent variable Z,

$$du_j(Z)/dZ = F_j,\ 1 \leq j \leq M, \tag{1}$$

where in general the F_j may be functions of any or all members of the set of dependent variables $\{u_l\}$ and, or, the independent variable Z. Suppose that the set of functions $\{F_j\}$ has four qualitative characteristics which will be enumerated presently. The first two of these four conditions state that the first function, F_1, is poorly determined numerically. As subsequent discussion explains, the combination of the four conditions describes a coupling which should cause traditonal point by point procedures of integration to fail and which suggests the particular successive approximation schema outlined in Section III.

The conditions are:

$$F_1 = \sum_k a_{1,k}, \qquad F_2 = \sum_k a_{2,k}. \tag{2}$$

(a) F_1 and F_2 are each given as a sum of terms.

(b) F_1 is smaller than one of the terms of equation (2), say, $a_{1,1}$, by a factor of 10^{-p}: $F_1/a_{1,1} = 10^{-p}$, p a positive integer.

(c) Calculation of an s digit value for one of the terms of F_2, say, $a_{2,1}$, requires that u_1 be known to s digits. Moreover, $a_{2,1}$ and F_2 are both of the same numerical magnitude.

(d) An s digit value for the term $a_{1,1}$ numerically determines an s digit value for u_2 (e.g., $a_{1,1}$ might be given analytically as $a_{1,1} = u_2 f$ where f is a function of other variables whose value can be computed accurately to s digits).

Since the conventional methods of computing the increments

$$\Delta_1 u_j(Z_n) = u_j(Z_n) - u_j(Z_{n-1}) \tag{3}$$

in the course of a point by point numerical integration require calculating the first derivatives, F_j, it follows from (a) and (b) that $F_1[\{u_l\}, Z_n]$ and therefore $\Delta_1 u_1(Z_n)$ can be computed to only $(s - p)$ digits whenever the $\{u_l(Z_n)\}$ are known to s digits. (This discussion assumes that an implicit method is being used in the integration. If an explicit method such as the Runge-Kutta were used, the sequence of subscripts on Z would be different, although the principles would be the same.) Furthermore, the error in $\Delta_1 u_1(Z_n)$ appears in $u_1(Z_n)$ and, therefore, by condition (c), in $F_2[\{u_j\}, Z_n]$. Unfortunately this introduces error first into $\Delta_1 u_2(Z_n)$ and thence into $u_2(Z_n)$. The completion of a vicious cycle through which this error reappears in the calculation of $F_1[\{u_l\}, Z_n]$ follows from conditions (b) and (d).

If the desired solution of the equation system actually does maintain conditions (b), (c), and (d) over an appreciable interval in Z, then these conventional methods might be expected to fail completely. The feedback of error might be expected to destroy quickly the approximate equality

$$F_1[\{u_l\}, Z] = 0. \tag{4}$$

In the case of the flame equations for one hypothetical free-radical system, the build-up of error was so rapid that in eight digit calculations the solution was lost in as few as six integration steps. (See Appendix A of reference [1] for an illustration of the failure of conventional techniques. A heuristic interpretation of this failure concludes the appendix.)

3. Proposed Iterative Solution.

A. *General basis of the method.* If the analytic formula for the second derivative of u_1

$$d^2u_1(Z)/dZ^2 = dF_1/dZ \tag{5}$$

did not present the same computational problem, the equation system could presumably be solved by the simple expedient of adopting F_1 as a variable and adding the differential equation for F_1 to the system. However, if the approximation $F_1 \backsimeq 0$ holds for any appreciable range, then the calculation of dF_1/dZ might

be expected to exhibit the same numerical indeterminacy as the calculation of F_1 itself. This is the case in several flame problems.

The very failure of an attempt to circumvent the difficulty by adding the equation for dF_1/dZ to the set suggests that perhaps the approximation $F_1 \cong 0$ is maintained over a considerable range in Z. In this case the four conditions (a)–(d) of Section 2 suggest that the solution to the complete set of equations might be approached by integrating the following sequence of equation systems. The kth approximating set of equations will make use of $(k-1)$st order approximations for u_1 and du_1/dt. For $k = 1$, the steady state approximation,

$$F_1[\{u_l\}, Z] = 0, \tag{6}$$

is made. According to conditions (b) and (d), equation (6) gives a numerically determinate implicit formula for u_2.

An approximation for $u_1^{(0)}$ can be obtained as follows. Recall that the hypothesis has been made that the analytic expression for dF_1/dZ exhibits the same numerical indeterminacy as F_1. Since, moreover, u_2 is supposed to make a numerically important contribution to $a_{1,1}$, which in turn has been supposed to be one of the most important terms of F_1, it is reasonable to hope that the term

$$(\partial a_{1,1}/\partial u_2)(du_2/dZ) = (\partial a_{1,1}/\partial u_2)F_2 \tag{7}$$

will make a numerically important contribution to dF_1/dZ and that a value for dF_1/dZ will determine without significant loss in accuracy a value for F_2. Now, according to condition (c), an s digit value for F_2 is supposed to determine an s digit value for u_1. Thus the approximation

$$dF_1/dZ = 0 \tag{8}$$

may provide a numerically determinate implicit equation for F_2 and therefore for $u_1^{(0)}$. These approximations are valid in some free-radical flame systems.

B. *Outline of the steps in the iteration scheme.*

1. *Step for* $k = 1$.

Solve the first set of approximate equations:

(9a) $F_1[\{u_l\}, Z] = 0$ (an implicit algebraic equation for $u_2^{(1)}$)

(9b) $(dF_1/dZ) = 0$ (an implicit algebraic equation for $u_1^{(0)}$)

(9c) $(du_j(Z)/dZ) = F_j, \qquad 3 \leq j \leq M.$

2. *Step for* $(k+1)$.

Assume that the $u_j^{(k)}$, $2 \leq j \leq M$ have been computed.

(a) Numerically differentiate $u_2^{(k)}$ to obtain $(du_2/dZ^{(k)}$.

(b) Since the $u_j^{(k)}$, $2 \leq j \leq M$ are known, calculate $u_1^{(k)}$ from the equation

$$(du_2/dZ)^{(k)} = F_2[\{u_l^{(k)}\}, Z]. \tag{10}$$

(Recall that condition (c) states that equation (10) is a numerically determinate formula for u_1.)

(c) Numerically differentiate $u_1^{(k)}$ to obtain $(du_1/dZ)^{(k)}$.

(d) Use the kth approximations which have just been computed for $u_1^{(k)}$

and $(du_1/dZ)^{(k)}$ to integrate numerically the $(k+1)$st set of approximate equations:

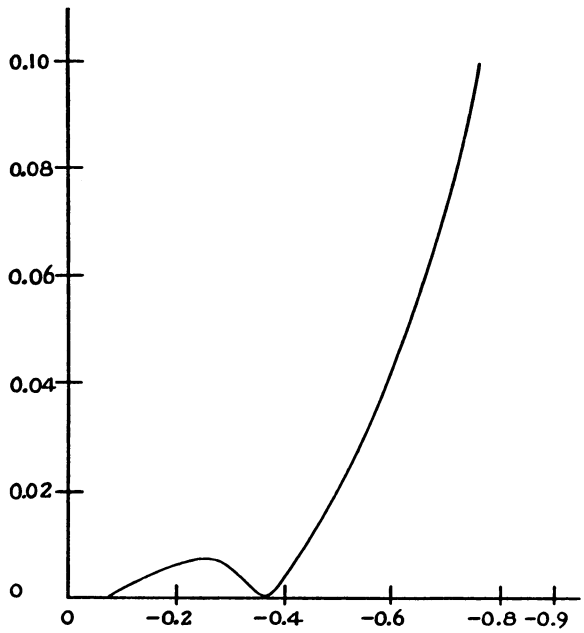
(11a) $$(du_1/dZ)^{(k)} = F_1[\{u_1^{(k)}, u_l^{(k+1)}\}, Z]$$

(an implicit algebraic equation for $u_2^{(k+1)}$)

(11b) $$(du_j/dZ)^{(k+1)} = F_j[\{u_1^{(k)}, u_l^{(k+1)}\}, Z]$$

$$(3 \leq j \leq M, \quad 2 \leq l \leq M).$$

A detailed illustration of the exact way in which one particular set of equations falls into the proposed successive approximation schema is given in Appendix C of reference [1]. The convergence is illustrated in the following graph.



Convergence of u_2: The ordinate is the fractional difference between the second and first approximations $[u_2^{(2)} - u_2^{(1)}]/u_2^{(2)}$ for the set of four simultaneous equations which describe one hypothetical free-radical flame. The maximum fractional error is always less than thirty percent and reaches 10 percent only after u_2 has dropped to one tenthousandth its initial value. The convergence of u_1 is equally good.

4. Conclusion. If the approximation schema converges to the solution (in the case of the two equation systems studied, the schema appears to converge to the correct answer), then it possesses two noteworthy advantages: (1) it succeeds when other point-by-point procedures fail because of numerical indeterminacy of one or more of the derivatives; (2) it can be readily programmed for machine calculation. Compared with the relaxation technique which also has been found to be stable towards this type of numerical indeterminacy, successive approximation requires more calculation. However, the relaxation method requires repeated use of judgment which can become difficult if the variables in the equations are badly cross-linked. Although by far the greater part of the calculational labor in a

relaxation treatment can be programmed for machine, that part which requires judgment would be difficult to code.

The disadvantage of successive approximation lies in increasing the length of the calculation. However, as the numerical indeterminacy grows more stringent, the initial approximation improves and the number of iterations required decreases. Moreover, this is a comparatively less serious disadvantage for a method suitable for machine use. Thus the number of times the calculation is to be repeated for a different set of input data, the character of the equations with respect to cross-linking, and the comparative labor of constructing the machine programs must all be considered when deciding which method will prove most convenient for any particular study.

EDWIN S. CAMPBELL

Department of Chemistry
New York University
New York, New York

1. EDWIN S. CAMPBELL, CM-847, University of Wisconsin Naval Research Laboratory, July 11, 1955.

A Smoothest Curve Approximation

A practical problem which often comes up in numerical work is the fitting of a curve to a finite set of known values in order to perform various operations such as integration. The usual method of approximation consists of fitting the points with one or more polynomials (independent of each other). By letting there be more points for each polynomial, with the polynomials being of comparable order, the error of approximation becomes asymptotic to a higher power of the interval length.

However, error analyses for such methods usually depend upon the boundedness of some derivative of a correspondingly high order [1]. But even if the function to be approximated is analytic, its correspondingly high order derivative may be of sufficient magnitude that for the given interval size, a simpler method would give better results. For instance, a fitting with an eighth order polynomial gives the following rule:

$$\int_0^8 f(t)dt = \frac{8}{28350}\,[989f(0) + 5888f(1) - 928f(2) + 10496f(3) - 4540f(4) + 10496f(5) - 928f(6) + 5888f(7) + 989f(8)].$$

An application of this formula to a positive function which was everywhere small over the range, apart from a sharp peak in the center, would lead to a negative result. Try to convince a prospector that there is a negative amount of mineral on his land because he finds a rich strike in the middle of it!

Frequently one contents oneself with a simpler rule which is repeated in blocks of so many intervals per block. However, this usually introduces discontinuities in the first derivative at the junctions of the blocks. If one were to integrate an in-