JOURNAL OF COMPUTATIONAL PHYSICS 10, 374–378 (1972)

Note

A Useful Device for Certain Boundary-Value Problems

As originally conceived, the method of invariant imbedding was a technique for determining the "missing" initial conditions at one endpoint of the interval underlying a linear two-point boundary-value problem. The solution of the boundary-value problem, if desired, was to be obtained by solving the resulting initial-value problem. A number of recent works [1–6] have appeared which have in common the basic idea of somehow solving the boundary-value problem entirely within the framework of invariant imbedding. The purpose of this note is to show that a straightforward application of the method of Scott [3, 6] on a digital computer can, under certain circumstances, lead to large cancellation errors, and to describe a modification of this method which seems to aid in overcoming these difficulties.

We consider the problem

$$u'(z) = a(z)\,u(z) + b(z)\,v(z) + s^{+}(z), \tag{1}$$

$$-v'(z) = c(z)\,u(z) + d(z)\,v(z) + s^{-}(z), \tag{2}$$

$$u(0) = 0, \qquad v(x) = \alpha. \tag{3}$$

The solution is given by

$$v(z) = \frac{1}{T(z)}\{T(x)\,\alpha + [e_l(x) - e_l(z)]\}\} \tag{4}$$

$$u(z) = R(z)\,v(z) + e_r(z), \tag{5}$$

where R, T, e_r and e_l are determined by the initial-value problem

$$R'(z) = b(z) + [a(z) + d(z)]\,R(z) + c(z)\,R^2(z), \tag{6}$$

$$T'(z) = [d(z) + c(z)\,R(z)]\,T(z), \tag{7}$$

$$e_r'(z) = [a(z) + c(z)\,R(z)]\,e_r(z) + R(z)\,s^{-}(z) + s^{+}(z), \tag{8}$$

$$e_l'(z) = [c(z)\,e_r(z) + s^{-}(z)]\,T(z), \tag{9}$$

$$R(0) = e_r(0) = e_l(0) = 0, \tag{10}$$

$$T(0) = 1. \tag{11}$$

Remark. That (4)–(11) define a solution of (1)–(3) can be directly verified. Background, motivation, an extension to systems, and an extension to more general boundary conditions are given in [6].

It is generally believed that an advantage accrues to invariant imbedding, relative to a straightforward application of the method of superposition, because initial-value problems for the system (6)–(9) can be solved sumerically with more accuracy than can such problems for the original system (1)–(2). However, granting this point, there remains the possibility, in any digital implementation, of large errors due to cancellations in the additions occurring in (4) and (5).

In order to illustrate this difficulty, Table I displays numerical results for the problem

$$u' = v + 1, \tag{12}$$

$$v' = u, \tag{13}$$

$$u(0) = 0, \qquad v(20) = 1. \tag{14}$$

The unmodified method of (Scott's version of) invariant imbedding obviously produces a rather poor approximation to the solution, but even so it is considerably better than the large negative solution produced by the method of superposition when it was applied to this unstable system. The major source of error in the unmodified invariant imbedding algorithm is roundoff error occurring in the subtraction $e_i(x) - e_i(z)$ in (4). For example, the exact value of $e_i(z)$ is $(1/\cosh z) - 1$ thus $e_i(x) = e_i(20) = -1 + 1.0306(-10)$, $e_i(18) = -1 + 7.61499(-9)$. Thus an integration scheme would need to have eight places of accuracy, which is a great deal to ask, in order to give even one significant figure in the direct computation

TABLE I

$u(z)$ for (12)–(14) As Given By Various Techniques

	$u(z)$		
z	Exact	Invariant imbedding (unmodified)	Successive starts ($z_i = 2i$)
5	6.1(−7)	4.9(−4)	4.3(−5)
10	0.00009	0.0359	0.00014
14	0.0050	0.6441	0.0050
16	0.0366	1.0183	0.0367
18	0.2707	1.1353	0.2706
19	0.7358	1.3677	0.7354
20	2.0000	2.0000	2.0000

of $e_l(20) - e_l(18)$. In fact, even the exact solution of the differential equations would yield a nonsensical final answer if the subtraction $e_l(x) - e_l(z)$ were carried out in fewer than eight digits of precision. (The small number $e_l(x) - e_l(z)$ gives a significant contribution to the final answer because it is divided by $T(z) = 1/\cosh z$, which is small of the same order.)

In the usual procedure for solving (1)–(3) via (4)–(11), the initial-value problem (6)–(11) is numerically integrated once from $z = 0$ to $z = x$. During this integration there are stored the values of R, T, e_r and e_l corresponding to those z at which the values of u and v are desired. (If this storage is undesirable, then it can be obviated at the cost of a second numerical integration of (6)–(11) from $z = 0$ to $z = x$.) In an attempt to minimize the effect of the cancellation error associated with the calculation of $e_l(x) - e_l(z)$ in (4), we recommend the following modification of this procedure.

Select values $0 = z_0 < z_1 < \cdots < z_n = x$ such that $\max(z_i - z_{i-1})$ is on the order of a characteristic length for the system (1)–(2). During the numerical integration substitute the initial-value problems

$$y_i'(z) = [c(z)\, e_r(z) + s^-(z)]\, T(z), \qquad z_{i-1} \leqslant z \leqslant z_i\,, \tag{15}$$

$$y_i(z_{i-1}) = 0, \tag{16}$$

for Eq. (9), but integrate the remaining Eqs. of (6)–(11) as previously, with $y_i(z)$ being stored rather than $e_l(z)$. After the integration compute $e_l(x) - e_l(z)$ at the desired values of z by the formula

$$e_l(x) - e_l(z) = [y_i(z_i) - y_i(z)] + y_{i+1}(z_{i+1}) + \cdots + y_n(z_n), \qquad z_{i-1} < z \leqslant z_i\,. \tag{17}$$

For ease of reference we shall denote this procedure as the method of *successive starts*. The results in Table I show that it may yield greatly increased accuracy, and it seems in no case to decrease the accuracy.

In conclusion, we would like to make the following remarks.

Remark 1. All numerical integrations for the example were performed using a Runge–Kutta integration scheme with a fixed step size of 0.2. This step size was selected as near optimal after some experimentation. For this example the use of a more accurate integration scheme would probably increase the accuracy of the final answer only marginally; however, there are other problems for which a more accurate integration scheme might be imperative. The computations were done in single precision on an IBM System\360.

Remark 2. The most important source of roundoff errors in (3)–(4) seems to be in the subtraction $e_l(x) - e_l(z)$ in (4). Although errors are certainly possible in forming $T(x)\alpha + [e_l(x) - e_l(z)]$, these will be important only at those z such

that the effect of the boundary conditions, represented by $T(x)\,\alpha/T(z)$, is equal in magnitude, but opposite in direction, to the effect of the inhomogeneous terms, represented by $[e_i(x) - e_i(z)]/T(z)$. At such z the value of $v(z)$, both actual and as computed from (4), will probably be small relative to its maximum value over all z, and consequently this cancellation error will not significantly affect accuracy relative to this maximum value. A similar argument holds for the addition in (5). (Incidentally, roundoff error in this operation is primarily responsible for the relative inaccuracy at $z = 5$.) We conclude that even with the successive starts modifications, this version of invariant imbedding can only be expected to yield results which are accurate *relative to the maximum value, over z, of the solution.* This criterion of accuracy will be acceptable for most problems.

Remark 3. From (9) it appears that, for example, the inaccuracy in forming $e_i(x) - e_i(z)$ is likely to be particularly troublesome when z is near x and $T(z)$ is a generally decreasing function having a value several orders of magnitude below unity at $z = x$. This type of behavior for $T(z)$ is not too uncommon, since $T(z)$ will vary like $e^{-z/L}$ in many interesting situations, where L is a characteristic length for the problem. [For example, L may be taken as the reciprocal of the maximum over $0 \leqslant z \leqslant x$ of some relevant norm for the given matrix defined by the right side of (1)–(2).] In such a situation this cancellation error is likely to be particularly severe for "long" problems in which the problem length x is several times larger than L, and it will then be imperative that some device, such as successive starts, be adopted to avoid this source of error.

Remark 4. The difficulty associated with computation of $e_i(x) - e_i(z)$ does not arise for homogeneous problems, as e_r and e_i are both identically zero for such problems. It seems likely that a corresponding difficulty will appear in some of the methods described in Refs. [1–5], but that this was not observed because either the development or the computational examples were limited to homogeneous problems.

ACKNOWLEDGMENT

The authors appreciate receiving comments by Professor R. C. Allen, of the University of New Mexico, on an earlier version of this work.

REFERENCES

1. R. E. BELLMAN, H. H. KAGIWADA, AND R. E. KALABA, *Comm. ACM* **10** (1967), 100.
2. I. H. MUFTI, C. K. CHOW, AND F. T. STOCK, *SIAM Rev.* **11** (1969), 616.
3. M. R. SCOTT, *J. Math. Anal. Appl.* **28** (1969), 112.

4. R. C. ALLEN, JR., AND G. M. WING, *J. Math. Anal. Appl.* **29** (1970), 141.
5. E. D. DENMAN, "Coupled Modes in Plasmas, Elastic Media, and Parametric Amplifiers," American Elsevier, New York, 1970.
6. P. NELSON, JR., AND M. R. SCOTT, *J. Math. Anal. Appl.* **34** (1971), 628.

PAUL NELSON, JR. AND C. A. GILES
*Oak Ridge National Laboratory,**
Oak Ridge, Tennessee 37830

* Operated by Union Carbide Corporation for the U. S. Atomic Energy Commission.

Printed in Belgium