ON SOME QUESTIONS ARISING IN THE APPROXIMATE SOLUTION OF NONLINEAR DIFFERENTIAL EQUATIONS*

By

R. BELLMAN, *RAND Corporation, Santa Monica, California*

AND

J. M. RICHARDSON, *Hughes Research Laboratories, Malibu, California*

Abstract. A new approach to the approximate solution of nonlinear differential equations is explored. The basic idea is to rewrite the nonlinear equations in the form of an infinite sequence of coupled linear equations by application of the Carleman linearization process. The sequence is truncated at a finite stage by a linear closure approximation involving the minimization of the mean square error. Attention is given to the stability of the truncated sequence of linear equations with respect to propagation of error due to closure back to the earlier members of the sequence. The use of suitably defined orthogonal polynomials to simplify closure approximations is considered. The generalization of the general method to the multidimensional case is treated. Consideration is given to the concept of self-consistent closure methods in which the averaging of the squared closure error depends upon the approximate linear equations derived thereby. A specific example of the last is treated analytically in closed form and a numerical comparison is made with the exact solution.

1. Introduction. The standard approaches to the approximate solution of nonlinear differential equations, analytic or computational, involve a linearization of one type or another. The finite difference techniques involve linearization over a small interval, while the methods of successive approximation such as quasilinearization [1] use linearization over the entire interval. In this report we wish to explore a new approach which raises a number of interesting analytic problems. Some we can answer, but others appear to be of great complexity. For some applications of these techniques, see two previous papers [2, 3].

2. Carleman linearization. It appears to have been pointed out first by Carleman [4] that any nonlinear differential equation could be converted, in numerous ways, into a linear differential equation of infinite order. Consider, for example, the scalar equation

$$\frac{du}{dt} = -u + u^2, \qquad u(0) = c, \tag{1}$$

and ignore the fact that we can solve this equation explicitly. Introduce the denumerable set of variables $u_k(t)$, $k = 1, 2, \cdots$, by means of the relation

$$u_k(t) = u^k. \tag{2}$$

Then, using (1), we obtain the differential equation

$$\frac{du_k}{dt} = ku^{k-1}\frac{du}{dt} = ku^{k-1}(-u + u^2) = -ku_k + ku_{k+1}\,. \tag{3}$$

*Received May 17, 1962.

If we affix the appropriate initial values $u_k(0) = c^k$ and establish a suitable uniqueness theorem, we can assert the equivalence of the original nonlinear differential equation and the infinite linear system of (3), where k assumes the values 1, 2, $\cdots$. As we shall discuss below, there are many other ways of accomplishing this end.

It is natural to attempt to approximate to the solution of the infinite system by cutting it off at a finite stage and using the set of linear equations

$$\frac{du_k}{dt} = -ku_k + ku_{k+1}, \qquad k = 1, 2, \cdots, N, \tag{4}$$

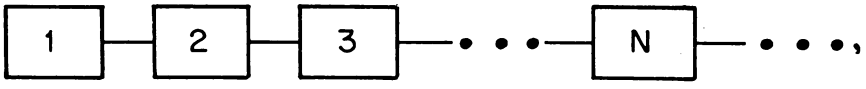
$u_k(0) = c^k$. As we see clearly, the first N equations involve $N + 1$ functions. Hence, we must approximate to u_{N+1} by some function of the previous u_i, $i = 1, 2, \cdots, N$, in order to obtain a closed set of equations.

In order to preserve the utility of this approximation scheme, we want u_{N+1} to be a linear function of the u_i,

$$u_{N+1} = \sum_{i=1}^{N} a_{iN}u_i . \tag{5}$$

How then should the coefficients a_{iN} be chosen, and what are the merits of this method? These are topics we shall discuss below.

3. Stability. Consider a linear series of coupled systems



where the state of the kth box at time t is denoted by $u_k(t)$. The equations of (3) assert that the rate of change of the kth box at time t is determined by the state of the kth system and that of its immediate neighbor on the right.

We now wish to ascertain how a slight change in the state of the Nth box affects the behavior over time of the first box. Will this slight change be propagated to the left with an increase or decrease in intensity, and will it build up over time? These are stability problems, and they involve the usual stability criteria for nonlinear differential equations [5].

4. Discussion of a particular case. Perhaps the simplest form of closure is that where $a_{iN} = 0$, $i = 1, 2, \cdots, N$, which means that the term u_{N+1} is simply omitted. We wish then to compare the solution of (4) with the solution of the modified system

$$\frac{dv_k}{dt} = -kv_k + kv_{k+1}, \qquad k = 1, 2, \cdots, N-1, \tag{6}$$

$$\frac{dv_N}{dt} = -Nv_N ,$$

with $v_k(0) = u_k(0) = c^k$.

Let us consider only the case where $| c | = | u(0) |$ is chosen sufficiently small that $| u(t) | \leq 2 | c |$ for $t \geq 0$. That this can be done is an immediate consequence of the Poincaré-Lyapunov stability theorem [5]. Hence, $| u_k(t) | \leq (2 | c |)^k$ for $t \geq 0$.

Using the Nth equations of each system, eqs. (6) and (4), we see that we may write

$$u_N = v_N + N \int_0^t e^{-N(t-s)} u_{N+1}(s)\, ds. \tag{7}$$

Hence

$$| u_N - v_N | \leq N \int_0^t e^{-N(t-s)}(2 | c |)^{N+1} \, dt \leq (2 | c |)^{N+1}. \tag{8}$$

Consider now the $(N - 1)$st equations

$$\frac{du_{N-1}}{dt} = -(N - 1)u_{N-1} + (N - 1)u_N , \tag{9}$$

$$\frac{dv_{N-1}}{dt} = -(N - 1)v_{N-1} + (N - 1)v_N .$$

Subtracting, we see that

$$u_{N-1} - v_{N-1} = (N - 1) \int_0^t e^{-(N-1)(t-s)}(u_N - v_N) \, ds, \tag{10}$$

from which

$$| u_{N-1} - v_{N-1} | \leq \left((N - 1) \int_0^t e^{-(N-1)(t-s)} \, ds\right)(2 | c |)^{N+1} \leq (2 | c |)^{N+1}. \tag{11}$$

It is easy to see inductively that

$$| u_k - v_k | \leq (2 | c |)^{N+1}, \tag{12}$$

for $k = N - 1, N - 2, \cdots, 1$.

We see then that in this case, provided that $2 | c | \leq 1$, the larger the value of N, the better the approximation. Furthermore, we see that the degree of approximation depends critically upon the sign of the coefficient of u, which is to say on the asymptotic behavior of the solutions of the linear approximation.

5. Alternative linearization. We have taken advantage of the particular analytic nature of the right side of the equation

$$\frac{du}{dt} = g(u), \qquad u(0) = c. \tag{13}$$

As in Carleman's original presentation, we can treat any continuous, or even suitably discontinuous right-hand side in the following fashion.

Let $\{p_n(u)\}$ be a sequence of orthonormal functions over a u-interval $[a, b]$ and expand $g(u)$ in an orthogonal series,

$$g(u) \sim \sum_{m=1}^{\infty} a_m p_m(u). \tag{14}$$

Then

$$\frac{dp_n(u)}{dt} = p_n'(u)g(u) \sim \sum_{m=1}^{\infty} a_{mn} p_m(u), \tag{15}$$

for $n = 1, 2 \cdots$, furnishes, together with (14), the desired infinite linear system.

To apply this idea to the original closure problem, observe that an expansion in Legendre polynomials (assuming that u varies only over $-1 \leq u \leq 1$) has the advantage that $P_n(u)$ is itself the best linear approximation of the form $\sum_{k=0}^{n-1} a_{kn} P_k(u)$ in the mean square sense over $[-1, 1]$.

Hence if our infinite linear system had the form

$$\frac{dP_1(u)}{dt} = a_{11}P_1 + a_{12}P_2 , \tag{16}$$

$$\frac{dP_2(u)}{dt} = a_{21}P_1 + a_{22}P_2 + a_{23}P_3 ,$$

$$\vdots$$

$$\frac{dP_n(u)}{dt} = a_{n1}P_1 + a_{n2}P_2 + \cdots + a_{nn}P_n + a_{n,n+1}P_{n+1} ,$$

$$\vdots$$

we could neglect P_{n+1} and obtain a closed finite system, secure in the knowledge that we had used the best mean-square fit P_{n+1} as a linear combination of the preceding P_k .

Similarly, if we used Tschebyscheff polynomials, we would have the best fit by neglecting the $(n + 1)$st polynomial in the nth equation in the sense of the norm $\max_{-1\le u\le 1} | \cdots |$.

The advantages of these expansions lie in the fact that the best Tschebyscheff fit to u^n for $-1 \le u \le 1$ is of the order of magnitude of 2^{-n}, i.e., we can find coefficients a_{kn} such that

$$\max_{-1\le u\le 1} \left| u^n - \sum_{k=0}^{n-1} a_{kn}u^k \right| = 2^{-n}, \tag{17}$$

and the minimizing polynomial is precisely

$$u^n - \sum_{k=0}^{n-1} a_{kn}u^k = \frac{\cos n(\cos^{-1} u)}{2^n}. \tag{18}$$

Similarly, the mean-square fit using Legendre polynomials is of the same order of magnitude.

This very much more precise approximation will be particularly useful when the linear approximation to a set of differential equations possesses oscillatory solutions, i.e., when the matrix has zero characteristic roots.

6. Multidimensional case. Consider now the set of nonlinear differential equations

$$\frac{dx_i}{dt} = \sum_{j=1}^{N} a_{ij}x_j + \sum_{j,k=1}^{N} a_{ijk}x_jx_k , \qquad x_i(0) = c. \tag{19}$$

Taking the x_jx_k as new variables $y_{jk} = x_jx_k$, we see that

$$\frac{dy_{jk}}{dt} = x_j\left(\sum_{l=1}^{N} a_{kl}x_l + \sum_{l,m} a_{klm}x_lx_m\right) + x_k\left(\sum_{l=1}^{N} a_{jl}x_l + \sum a_{jlm}x_lx_m\right)$$

$$= \sum_{l=1}^{N} a_{kl}y_{jl} + \sum_{l=1}^{N} a_{jl}y_{kl} + \cdots . \tag{20}$$

Introducing vector-matrix notation we can write

$$\frac{dx}{dt} = Ax + By, \tag{21}$$

$$\frac{dy}{dt} = A_2y + \phi(x).$$

Here x is an n-dimensional vector, y and $\phi(x)$ are n^2-dimensional vectors, A is an $n \times n$ matrix, A_2 is an $n^2 \times n^2$ matrix, and B is a $2n \times n$ matrix. Similarly, we can form the $n^3 \times n^3$ system satisfied by the functions $x_i x_j x_k$, etc. Fortunately, the matrices A_2, A_3, $\cdots$ which we obtain in this way have a simple structure. They are the iterated Kronecker sums of the matrix [6] A,

$$A_2 = A \oplus A, \qquad A_3 = A \oplus A_2 = A \oplus A \oplus A. \tag{22}$$

The characteristic roots of A_2 are $\lambda_i + \lambda_j$, where the characteristic roots of A are λ_i, $i = 1, 2, \cdots, N$.

Hence, under the assumption that A is a stability matrix, namely that all of its characteristic roots have negative real parts, we can establish a result analogous to that derived in Section 4 for the first-order differential equation.

7. Self-consistent methods. Questions of greater difficulty are posed by the following variant of the foregoing technique. Returning to the equation

$$\frac{du}{dt} = -u + u^2, \qquad u(0) = c, \tag{23}$$

for simplicity of exposition, suppose that we approximate to u^2 by a linear combination

$$u^2 \simeq a_1 u + a_2 \tag{24}$$

over the interval $[0, T]$, using a mean-square norm as a measure of approximation. We wish to choose a_1 and a_2 to minimize the quantity

$$f(a_1, a_2) = \int_0^T (u^2 - a_1 u - a_2)^2 \, dt. \tag{25}$$

The a_1 and a_2 are readily determined by the conditions

$$\begin{aligned} \int_0^T u^2 \, dt &= a_1 \int_0^T u \, dt + a_2 T \\ \int_0^T u^3 \, dt &= a_1 \int_0^T u^2 \, dt + a_2 \int_0^T u \, dt. \end{aligned} \tag{26}$$

The first difficulty we face is that the coefficients a_1 and a_2 depend upon the ultimate solution, which we do not know. The best we can do is to use the solution of the approximating linear equation as an approximation to u in (26). For finite T, an insoluble transcendental equation is obtained. However, in the limit $T \to \infty$ (26) reduces to a pair of soluble equations, and the method can be carried through with the final result in closed form.

Substitution of (24) into (23) yields the expression

$$\frac{du}{dt} = -(1 - a_1)u + a_2, \tag{27}$$

whose solution is

$$\begin{aligned} u &= (c - b)e^{-(1-a_1)t} + b \\ b &= a_2/(1 - a_1) \end{aligned} \tag{28}$$

where the initial condition $u(0) = c$ has been used.

Assuming that $a_1 < 1$, the first of eqs. (26) with u given by (28) reduces to

$$\left(\frac{a_2}{1-a_1}\right)^2 - a_1\left(\frac{a_2}{1-a_1}\right) + a_2 = 0 \tag{29}$$

in the limit $T \to \infty$. Eq. (29) has two solutions for a_2 in terms of a_1, namely

$$a_2 = 0, \qquad -(1-a_1)(1-2a_1). \tag{30}$$

The root $a_2 = -(1-a_1)(1-2a_1)$ is of no interest since an investigation of this possibility reveals that it is an attempt of the method to approximate the solution of the original nonlinear equation (23) in the neighborhood of the *unstable* equilibrium solution $u = 1$. On the other hand, the root $a_2 = 0$ corresponds to the approximation of the solution in the neighborhood of the *stable* equilibrium solution $u = 0$. Therefore, the treatment will be henceforth limited to the latter case ($a_2 = 0$).

Setting $a_2 = 0$ (i.e., $b = 0$) in (28), giving the solution to the approximate linear equation, substituting the result in the second of eqs. (24), and letting $T \to \infty$, we obtain

$$\frac{c^3}{3(1-a_1)} = a_1 \frac{c^2}{2(1-a_1)} \tag{31}$$

or

$$a_1 = 2/3c. \tag{32}$$

Thus the self-consistent version of the solution (28) is

$$u = ce^{-(1-2/3c)t}. \tag{33}$$

This result can be compared with the exact result

$$u = [(1-c)c^{-1}e^t + 1]^{-1}. \tag{34}$$

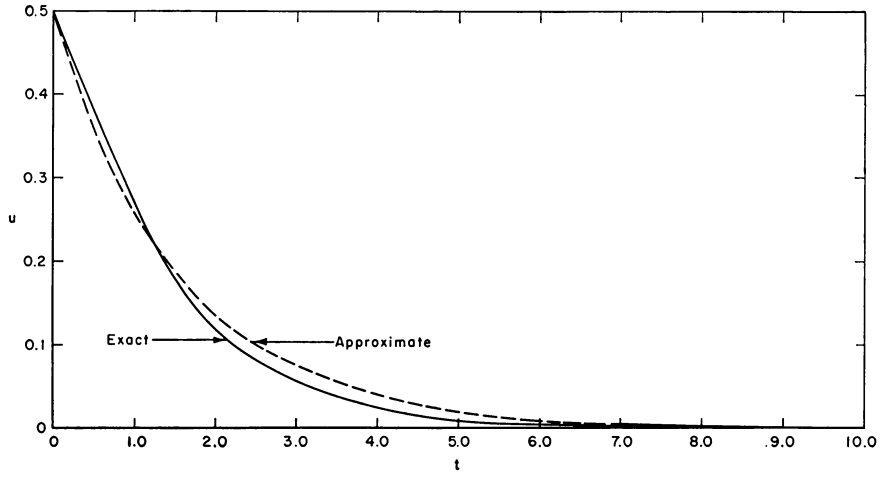
In Table I the approximate and exact solutions are compared for selected values of t

Comparison of Approximate and Exact Solutions ($c = 1/2$)

t	u_{exact} (Eq. 34)	u_{approx} (Eq. 33)	$u_{\text{approx}} - u$
0	0.500	0.500	0.000
0.1	0.475	0.468	-0.007
0.2	0.450	0.438	-0.012
0.5	0.378	0.358	-0.020
1.0	0.269	0.257	-0.012
2.0	0.119	0.132	0.013
5.0	0.00669	0.0178	0.0111
10.0	0.000045	0.00063	0.00058

for the case $c = \frac{1}{2}$. The exact and approximate solutions are compared graphically in Fig. 1. It is to be noted that the maximum absolute error is only about 0.02. The approximate solution is surprisingly good considering the crudity of the approximation method.

In the case where $T < \infty$ we can use successive approximations and proceed as follows. Let u_0 be an initial approximation to the solution of (23) and compute $a_1^{(0)}$, $a_2^{(0)}$



from (26) using u_0 in place of u. Let u_1 be determined as the solution of

$$\frac{du_1}{dt} = -u_1 + a_1^{(0)} + a_2^{(0)}u_1 , \qquad u_1(0) = c, \tag{35}$$

and then $a_1^{(1)}$, $a_2^{(1)}$ obtained from (26) with u replaced by u_1 and so on.

In this way we converge to the self-consistent approximate solution. It seems fairly difficult to obtain good estimates of the degree of approximation.

References

1. R. Kalaba, J. Math. and Mech. **8**, 519 (1959)
2. R. Bellman and J. M. Richardson, Proc. Natl. Acad. Sci. U. S. **47**, 1191 (1961)
3. J. M. Richardson and R. Bellman, *Perturbation techniques*, Symposium on Nonlinear Oscillations, Kiev, U. S. S. R., to appear
4. T. Carleman, Ark. Mat. Astron. Fys. **22***B*, 1 (1932)
5. R. Bellman, *Stability Theory of Differential Equations* (McGraw-Hill Book Company, Inc., New York, 1954)
6. R. Bellman, *Introduction of Matrix Analysis* (McGraw-Hill Book Company, Inc., New York, 1960)