JOURNAL OF COMPUTATIONAL PHYSICS 15, 117–133 (1974)

Solution of Two-Point Boundary-Value Problems with Jacobian Matrix Characterized by Large Positive Eigenvalues*

A. MIELE

Department of Mathematical Sciences, Rice University, Houston, Texas

A. K. AGGARWAL

Department of Mechanical and Aerospace Engineering and Materials Science, Rice University, Houston, Texas

AND

J. L. TIETZE†

Department of Electrical Engineering, Rice University, Houston, Texas

Received November 26, 1973

This paper treats the nonlinear, two-point boundary-value problem formulated by Troesch (Ref. 3) and studied by Roberts and Shipman (Ref. 4). Computationally speaking, this is a difficult problem, owing to the fact that the Jacobian matrix is characterized by large positive eigenvalues. The resulting numerical difficulties are reduced by treating the two-point boundary-value problem as a multipoint boundary-value problem. The modified quasilinearization algorithm of Refs. 5–6 is employed. This approach bypasses the integration of the nonlinear equations, which characterizes shooting methods. Computational results are also presented for another difficult nonlinear, two-point boundary-value problem, namely, the problem formulated by Holt (Ref. 7).

1. INTRODUCTION

This paper deals with differential equations of the form $\dot{x} - \varphi(x, t) = 0$, subject to p initial conditions and q final conditions, with $p + q = n$, where n is the dimension of the vector x. There are two main techniques for solving these

* This research was supported by the Office of Scientific Research, Office of Aerospace Research, United States Air Force, Grant No. AF-AFOSR-72-2185. This research is a condensation of the investigations described in Refs. 1-2.

† Present address: Technical Staff, TRW, Systems Group, Houston, Texas.

problems: (i) initial-value methods, also called shooting methods, and (ii) quasilinearization methods [8–13].

When technique (i) is employed, the initial conditions and the differential equations are satisfied at each stage of the process, while the final conditions are violated to some degree. A nominal solution is generated by choosing the initial vector $x(0)$ in a way consistent with the given initial conditions (this involves guessing the missing initial conditions) and by integrating forward the differential equations. A correction $\Delta x(0)$ to the initial vector $x(0)$ is generated so as to reduce the error in the final conditions at each iteration.

When technique (ii) is employed, the initial and final conditions are satisfied at each stage of the process, while the differential equations are violated to some degree. A nominal solution is generated by choosing the function $x(t)$ in a way consistent with the initial and final conditions. Then, a correction $\Delta x(t)$ to the vector function $x(t)$ is generated so as to reduce the error in the differential equations at each iteration.

Regardless of the technique employed, the correction $\Delta x(0)$ to the initial vector $x(0)$ or the correction $\Delta x(t)$ to the vector function $x(t)$ depends on the behavior of the Jacobian matrix $J = \varphi_x(x, t)$ associated with the differential system under consideration. Once a nominal solution is chosen, the Jacobian matrix depends on the time only and, hence, its eigenvalues are time dependent. If the eigenvalues have negative real parts, the two-point boundary-value approach of [8–13] leads to the solution in a reasonable number of iterations, providing the nominal solution is sufficiently close to the actual solution. On the other hand, if the eigenvalues have positive real parts,[1] the exponential growth of some of the components of the solution might lead to numerical difficulties, especially when the interval of integration is rather large.

To forestall the above difficulties, a multipoint approach to the two-point boundary-value problem was proposed in [14] and then developed in [15–16] in connection with shooting methods. The basic idea is to restrict the growth of some of the components of the solution by subdividing the interval of integration into m subintervals and then imposing continuity conditions at the interface between subintervals. The resulting approach exhibits characteristics which are intermediate between those of a shooting method and those of a finite-difference method.

More recently, a multipoint approach to the two-point boundary-value problem was developed in [5–6] within the frame of the modified quasilinearization method. The main idea is to bypass the integration of the nonlinear equations which characterizes the approaches of [14–16].

In this paper, the technique developed in [5–6] is employed to solve the difficult

[1] This is the case with the problems formulated by Troesch (Ref. 3) and Holt (Ref. 7).

two-point boundary-value problems formulated by Troesch in [3] and Holt in [7]. Section 2 summarizes the main results of [5–6]. Section 3 treats Troesch's problem, and Section 4 treats Holt's problem. Finally, the conclusions are given in Section 5.

2. Multipoint Approach

Consider the nonlinear differential equation[2]

$$\dot{x} - \varphi(x, t) = 0, \qquad 0 \leqslant t \leqslant 1, \tag{1}$$

where t is the time, x is the state, and φ is a continuous function of the arguments x and t. Here, t is a scalar, x an n-vector, and φ an n-vector.[3] At the initial time $t = 0$, p nonlinear conditions must be satisfied and, at the final time $t = 1$, q nonlinear conditions must be satisfied, with $p + q = n$. Therefore, the boundary conditions for the system (1) are expressed in the form

$$f[x(0)] = 0, \qquad g[x(1)] = 0, \tag{2}$$

where f is a p-vector and g is a q-vector.

In the two-point boundary-value approach (TPBVA), the interval of integration [0, 1] is a treated as a whole. The problem is to find the function $x(t)$ which solves Eq. (1) subject to the boundary conditions (2).

In the multipoint boundary-value approach (MPBVA), the interval of integration [0, 1] is divided into m subintervals by means of the $m - 1$ time stations $t_1, t_2, \ldots, t_{m-1}$, which are intermediate between the initial time $t_0 = 0$ and the final time $t_m = 1$. The problem is to find the function $x(t)$ which solves Eq. (1) subject to the boundary conditions (2) and the interface conditions or continuity conditions

$$x(t_i)_+ - x(t_i)_- = 0, \qquad i = 1, 2, \ldots, m - 1. \tag{3}$$

Performance Index

Since the system under consideration is nonlinear, approximate methods must be employed to obtain the solution numerically. Regardless of the method employed, any deviation from the solution can be measured in terms of a scalar performance index P, which is defined in such a way that $P = 0$ for the exact solution and $P > 0$ for any approximation to the solution. Specifically, P is given by[4]

$$P = \int_0^1 [\dot{x} - \varphi(x, t)]^T [\dot{x} - \varphi(x, t)]\, dt + f^T[x(0)]\, f[x(0)] + g^T[x(1)]\, g[x(1)] \tag{4}$$

[2] Without loss of generality, the interval of integration is normalized to be [0, 1].

[3] All vectors are column vectors.

[4] The superscript T denotes the transpose of a matrix.

and measures the cumulative error in the differential equations and the boundary conditions. Hence, one can use it as a guide during progression of a particular algorithm as well as to establish convergence. In Eq. (4), it is assumed implicitly that the function $x(t)$ satisfies the interface conditions (3) at every stage of the process.

Modified Quasilinearization Algorithm

The modified quasilinearization algorithm of [5–6] is designed to achieve a descent property on the performance index P. Let $x(t)$ denote the nominal function, let $\tilde{x}(t)$ denote the varied function, and let $\Delta x(t)$ denote the displacement leading from the nominal function to the varied function. These quantities satisfy the definition

$$\tilde{x}(t) = x(t) + \Delta x(t) = x(t) + \alpha A(t), \tag{5}$$

where α is the stepsize, $0 \leqslant \alpha \leqslant 1$, and where $A(t)$ denotes the displacement per unit stepsize.

The function $A(t)$ is determined by the following equations:

$$\dot{A} - \varphi_x^T(x, t)A + [\dot{x} - \varphi(x, t)] = 0, \tag{6}$$

$$f_x^T[x(0)]\, A(0) + f[x(0)] = 0, \qquad g_x^T[x(1)]\, A(1) + g[x(1)] = 0, \tag{7}$$

$$A(t_i)_+ - A(t_i)_- = 0, \qquad i = 1, 2, \ldots, m-1. \tag{8}$$

The method employed to solve the linear, nonhomogeneous system (6–8) is the method of particular solutions described in [5–6] and is not repeated here. For this problem, every iteration requires that $n + 1$ particular solutions be generated in each subinterval by forward integration. Therefore, the total number of particular solutions is $m(n + 1)$, where m is the number of subintervals. The solutions are combined linearly, and the $m(n + 1)$ constants are computed so as to satisfy Eqs. (6–8). We note that mn relations between the constants are supplied by Eqs. (7–8). The remaining m relations are supplied by the requirement that the sum of the constants pertaining to a particular subinterval be equal to one for that subinterval [5–6].

Once the function $A(t)$ is known, Eq. (5) becomes a one-parameter family, the parameter being the stepsize α. For this one-parameter family, the performance index P becomes a function of the form $\tilde{P}(\alpha)$. The stepsize α is computed by means of a bisection process, starting from $\alpha = 1$, until a reduction in the performance index is obtained.

After the stepsize α is known, the varied function $\tilde{x}(t)$ is computed with Eq. (5).

Then, the coefficients appearing in the differential system, are updated, and the next iteration of the modified quasilinearization algorithm is started. The process is terminated when the stopping condition $P \leqslant \epsilon$ is satisfied.

Multipoint Location

For a given nominal function $x(t)$, the Jacobian matrix $J = \varphi_x(x, t)$ is known along the interval of integration, and its eigenvalues $\lambda(t)$ can be computed. Even though φ_x and λ are time dependent, they can be regarded as constants over a time interval sufficiently small. Of course, this assumption is made only in order to gain a qualitative insight into the problem and not for computational purposes.

With the assumption $\lambda = \text{const}$ over a particular subinterval, the general solution of the homogeneous part of the differential Eq. (6) includes terms of the form $\exp[\lambda(t - t_i)]$, where λ is any eigenvalue and $t - t_i$ denotes the time elasped from the beginning of the subinterval. Let Λ denote the eigenvalue having the largest real part in the given subinterval. If the real part of Λ is positive, the term $\exp[\Lambda(t - t_i)]$ achieves the highest value at the end of the subinterval. Therefore, if $\Delta t = t_{i+1} - t_i$ denotes the length of the subinterval, this highest value is $\exp[\mathrm{Re}(\Lambda)\,\Delta t]$.

If the method of particular solutions is employed [5–6], the initial conditions for Eq. (6) are varied systematically at the beginning of each subinterval. A linear combination of these particular solutions is constructed, and the constants of the combination are determined so that all the equations and boundary conditions are satisfied. This yields a set of linear algebraic equations whose characteristic matrix has elements with order of magnitude ranging between 1 and $\exp[\mathrm{Re}(\Lambda)\,\Delta t]$. Since a given computer carries out numerical computations with finite precision, it is clear that the growth of the term $\exp(\Lambda\Delta t)$ must be contained, otherwise the solution of the linear system becomes meaningless.

The selection of the multipoints can now be placed in a better perspective. The spacing Δt must be determined so that the following inequality is satisfied:[5]

$$\exp[\mathrm{Re}(\Lambda)\,\Delta t] \ll 10^{\beta} \tag{9.1}$$

or

$$\mathrm{Re}(\Lambda)\,\Delta t \ll \beta \log_e 10 = 2.3\beta, \tag{9.2}$$

where β denotes the number of significant digits of a given computer. As an example, the IBM 370/155 computer is characterized by $\beta = 16$ in double-precision arithmetic. Therefore, for this machine, Ineq. (9.2) becomes

$$\mathrm{Re}(\Lambda)\Delta t \ll 36.8. \tag{10}$$

[5] We emphasize that the symbol Δt denotes the time interval between two consecutive multipoints and should not be confused with the integration stepsize h employed in each subinterval. Note that $h \ll \Delta t$; indeed, in the examples of Sections 3–4, $h = \Delta t/100$.

In summary, the selection of the multipoints is connected to the selection of the starting nominal solution. Even an approximate knowledge of the solution aids in the selection of the multipoints via the eigenvalues of the Jacobian matrix. As a general rule, the multipoints should be spaced more closely for those portions of the solution that have large positive eigenvalues and less closely for those portions that have small positive eigenvalues.

3. Problem of Troesch

In first-order form, the problem of Troesch [3] can be formulated as follows:[6]

$$\dot{x} = y, \qquad \dot{y} = k \sinh(kx), \tag{11}$$

$$x(0) = 0, \qquad x(1) = 1. \tag{12}$$

Its Jacobian matrix

$$J = \begin{bmatrix} 0 & k^2 \cosh(kx) \\ 1 & 0 \end{bmatrix} \tag{13}$$

is characterized by the following eigenvalues:

$$\lambda = \pm k \sqrt{[\cosh(kx)]}, \tag{14}$$

which, at the endpoints, become

$$\lambda(0) = \pm k, \qquad \lambda(1) = \pm k \sqrt{[\cosh(k)]}. \tag{15}$$

For relatively low values of k, the eigenvalues (15.2) are small, and problem (11–12) can be treated by employing two-point boundary-value techniques, such as those outlined in [8–13]. On the other hand, for relatively large values of k, the eigenvalues (15.2) are large, becoming $\lambda(1) = \pm 1049$ for $k = 10$. Thus, the use of multipoint boundary-value techniques, such as those presented in [5–6] and [14–16], becomes desirable.

The attention of these authors was attracted by the recent interesting work of Roberts and Shipman [4], who solved problem (11–12) by a combination of methods, namely, multipoint, continuation,[7] and perturbation in conjunction with shooting techniques. These authors wondered whether the multipoint approach of [5–6], employed in conjunction with the modified quasilinearization method, might yield a more direct, and yet more precise, solution than that given in [4]. Thus, the

[6] The symbols employed here denote scalar quantities.

[7] For a discussion of continuation techniques, see [17].

driving ideas were as follows: (i) to bypass the integration of the nonlinear equations which characterizes the approach of [4] and (ii) to bypass the solution of the sequence of multipoint boundary-value problems required by the continuation-perturbation technique of [4].

Experimental Conditions

Computations were performed using the IBM 370/155 computer at Rice University. The modified quasilinearization algorithm was programmed in FORTRAN IV and double-precision arithmetic. The linearized Eqs. (6–8) were integrated using Hamming's modified predictor-corrector method with a special Runge-Kutta procedure to start the integration routine [18]. The definite integral (4) was computed using a modified Simpson rule.

Solutions to problem (11–12) were computed for three values of the constant k, namely,

$$k = 5, \qquad k = 6, \qquad k = 10. \tag{16}$$

In all cases, the modified quasilinearization algorithm was employed iteratively until the following stopping condition was satisfied:

$$P \leqslant 10^{-12}. \tag{17}$$

In addition to the convergence condition (17), the following nonconvergence conditions were employed:

$$\text{(a)} \quad N \geqslant 50, \tag{18.1}$$

$$\text{(b)} \quad N_s \geqslant 10, \tag{18.2}$$

$$\text{(c)} \quad M \geqslant 10^{78}. \tag{18.3}$$

Here, N is the iteration number, N_s is the number of bisections of the stepsize required to ensure the decrease of the performance index at any iteration, and M is the modulus of any of the quantities employed in the algorithm. Satisfaction of Ineq. (18.1) indicates divergence or extreme slowness of convergence; satisfaction of Ineq. (18.2) indicates extreme smallness of the displacement $\Delta x(t)$; and satisfaction of Ineq. (18.3) indicates exponential overflow.

Multipoint Location

For $k = 5$ and $k = 6$, solutions were computed employing $m = 8$ subintervals and 50 integration steps per subinterval. For $k = 10$, solutions were computed employing $m = 14$ subintervals and 100 integration steps per subinterval.

Since the positive eigenvalue becomes quite large near the final point, the

multipoints were spaced so as to be more dense near the final point. For $m = 8$, the time stations were located at

$$t_0 = 0.00, \quad t_1 = 0.50, \quad t_2 = 0.75, \quad t_3 = 0.88, \quad t_4 = 0.94,$$
$$t_5 = 0.97, \quad t_6 = 0.98, \quad t_7 = 0.99, \quad t_8 = 1.00. \tag{19}$$

For $m = 14$, the time stations were located at

$$t_0 = 0.000, \quad t_1 = 0.500, \quad t_2 = 0.750, \quad t_3 = 0.850, \quad t_4 = 0.910,$$
$$t_5 = 0.940, \quad t_6 = 0.960, \quad t_7 = 0.970, \quad t_8 = 0.980, \quad t_9 = 0.990, \tag{20}$$
$$t_{10} = 0.992, \quad t_{11} = 0.994, \quad t_{12} = 0.996, \quad t_{13} = 0.998, \quad t_{14} = 1.000.$$

Nominal Functions

Since the differential system (11–12) is characterized by nonnegative value of $\ddot{x}$, it is natural to choose nominal functions such that $\ddot{x} \geqslant 0$. The following one-parameter family of nominal functions is consistent with (11.1), (12.1), and (12.2) for every value of the parameter γ:

$$x = t^\gamma, \qquad y = \gamma t^{\gamma-1}, \qquad \gamma \geqslant 1. \tag{21}$$

When these nominal functions are employed, Eq. (11.2) is violated, and the performance index (4) reduces to

$$P(k, \gamma) = \int_0^1 [\gamma(\gamma - 1)\, t^{\gamma-2} - k \sinh(kt^\gamma)]^2\, dt. \tag{22}$$

An obvious choice of the parameter γ is that which gives the minimum value to P for given k.

In this preliminary optimization, it is not essential that the optimum value of γ be computed exactly; the only important thing is that γ be in a proper range. With this in mind, the performance index (22) was computed for discrete values of the parameter γ [1]. Within the discrete set of values given to the parameter γ, the following values appear to be quasi-optimal:

$$\gamma = 12 \qquad \text{for} \quad k = 5, \tag{23.1}$$
$$\gamma = 20 \qquad \text{for} \quad k = 6, \tag{23.2}$$
$$\gamma = 150 \qquad \text{for} \quad k = 10. \tag{23.3}$$

This preliminary optimization required little computer time, since it involves straightforward quadratures. It proved to be useful computationally, since it enabled the multipoint boundary-value approach of [5–6] to converge rapidly to the solution. Indeed, from the computer time viewpoint, this preliminary optimization is approximately equivalent to one iteration of the modified quasilinearization algorithm of [5–6].

First Integral

The problem under consideration is characterized by the first integral

$$z = \cosh(kx) - y^2/2 = \text{const.} \tag{24}$$

This first integral is useful in checking the accuracy of the numerical procedure.

Computer Runs

Starting with the nominal functions (21) and (23), computer runs were made employing the multipoint boundary-value approach of [5–6]. Table I shows the

TABLE I

Performance index versus iteration number (Troesch's problem, MPBVA)

	$k = 5$	$k = 6$	$k = 10$
N	P	P	P
0	$0.46E + 03$	$0.28E + 04$	$0.26E + 07$
1	$0.61E + 01$	$0.10E + 03$	$0.16E + 07$
2	$0.62E - 02$	$0.53E + 00$	$0.13E + 06$
3	$0.88E - 08$	$0.29E - 04$	$0.62E + 04$
4	$0.20E - 19$	$0.11E - 12$	$0.86E + 02$
5			$0.50E - 01$
6			$0.32E - 07$
7			$0.20E - 19$

performance index P versus the iteration number N. The descent property on P was enforced by employing the stepsize $\alpha = 1$ at every iteration: there was no need for bisections. Convergence to the stopping condition (17) was achieved in $N = 4$ iterations for $k = 5$, $N = 4$ iterations for $k = 6$, and $N = 7$ iterations for $k = 10$. This rapid convergence is due to the excellent characteristics of the nominal functions (21) and (23).

Table II shows the converged initial values $x(0)$, $y(0)$, $z(0)$ and the converged final values $x(1)$, $y(1)$, $z(1)$ to seven significant figures. For the converged solutions, the tabulated functions $x(t)$, $y(t)$, $z(t)$ are given in [1] to four significant figures.

TABLE II

Terminal values at convergence (Troesch's problem, MPBVA)

	$k = 5$	$k = 6$	$k = 10$
$x(0)$	$0.0000000E + 00$	$0.0000000E + 00$	$0.0000000E + 00$
$y(0)$	$0.4575046E - 01$	$0.1795095E - 01$	$0.3583378E - 03$
$z(0)$	$0.9989534E + 00$	$0.9998389E + 00$	$0.9999999E + 00$
$x(1)$	$0.1000000E + 01$	$0.1000000E + 01$	$0.1000000E + 01$
$y(1)$	$0.1210050E + 02$	$0.2003576E + 02$	$0.1484064E + 03$
$z(1)$	$0.9989534E + 00$	$0.9998392E + 00$	$0.9999770E + 00$

According to the first integral (24), the quantity z should be constant along the interval of integration. The constancy of z is verified to seven significant figures for $k = 5$, to five significant figures for $k = 6$, and to four significant figures for $k = 10$.

As a further check of the accuracy of the solutions obtained, the differential system (11–12) was integrated forward employing the *converged initial conditions* given in Table II. The numerical results show that the computed value of $x(1)$ agrees with the prescribed final condition (12.2) to seven significant figures for $k = 5$, to seven significant figures for $k = 6$, and to six significant figures for $k = 10$.

Additional Computer Runs

Computer runs were also made employing the nominal functions (21) with

$$\gamma = 1. \tag{25}$$

The number of multipoints employed was $m = 8$ for $k = 5$ and $k = 6$ and $m = 14$ for $k = 14$. Convergence to the stopping condition (17) was achieved in $N = 6$ iterations for $k = 5$ and $N = 7$ iterations for $k = 6$. On the other hand, for $k = 10$, convergence was not achieved [nonconvergence (c), see Ineq. (18.3)]. Comparison of these data with those of Table I stresses the advantage accrued through the preliminary optimization of γ described by Eqs. (23).

4. Problem of Holt

In first-order form, the problem of Holt [7] can be formulated as follows:[8]

$$\dot{x} = \tau y, \quad \dot{y} = \tau z, \quad \dot{z} = \tau(-1.55xz + 0.10y^2 - u^2 + 0.20y + 1), \tag{26}$$

$$\dot{u} = \tau w, \qquad \dot{w} = \tau(-1.55xw + 1.10yu + 0.20u - 0.20), \tag{27}$$

$$x(0) = 0, \qquad y(0) = 0, \qquad u(0) = 0, \qquad y(1) = 0, \qquad u(1) = 1. \tag{28}$$

This difficult two-point boundary-value problem was treated by Roberts and Shipman in [16] and Jones in [19]. Roberts and Shipman employ a combination of multipoint and continuation techniques in conjunction with shooting methods. Jones employs an automatic continuation technique in conjunction with shooting methods. In order to further test the power of the method proposed in [5–6], these authors decided to reinvestigate problem (26–28).

Experimental Conditions

The experimental conditions employed for Holt's problem were identical with those employed for Troesch's problem. Solutions to problem (26–28) were computed for three values of the constant τ, namely,

$$\tau = 11.3, \qquad \tau = 13.3, \qquad \tau = 20.0. \tag{29}$$

Once more, the convergence condition for the modified quasilinearization algorithm was represented by Ineq. (17) and the nonconvergence conditions were given by Ineqs. (18).

Multipoint Location

For $\tau = 11.3$ and $\tau = 13.3$, solutions were computed employing $m = 4$ subintervals and 100 integration steps per subinterval. For $\tau = 20.0$, solutions were computed employing $m = 8$ subintervals and 100 integration steps per subinterval.

The magnitude of the eigenvalues at the endpoints is known *a priori* in Troesch's problem, while this is not the case in Holt's problem. Hence, for lack of better information, we spaced the multipoint uniformly. For $m = 4$, the time stations were located at

$$t_0 = 0.00, \qquad t_1 = 0.25, \qquad t_2 = 0.50, \qquad t_3 = 0.75, \qquad t_4 = 1.00. \tag{30}$$

[8] The symbols employed here denote scalar quantities.

For $m = 8$, the time stations were located at

$$t_0 = 0.000, \quad t_1 = 0.125, \quad t_2 = 0.250, \quad t_3 = 0.375, \quad t_4 = 0.500,$$
$$t_5 = 0.625, \quad t_6 = 0.750, \quad t_7 = 0.875, \quad t_8 = 1.000. \tag{31}$$

Nominal Functions

For all values of the parameter τ, the following nominal functions were employed:

$$x = 0, \quad y = 0, \quad z = 0, \quad u = t, \quad w = 0. \tag{32}$$

These nominal functions are consistent with the boundary conditions (28) but violate the differential constraints (26–27).

Computer Runs

Starting with the nominal functions (32), computer runs were made employing the multipoint boundary-value approach of [5–6]. Table III shows the stepsize α

TABLE III
Performance index versus iteration number (Holt's problem, MPBVA)

	$\tau = 11.3$		$\tau = 13.3$		$\tau = 20.0$	
N	α	P	α	P	α	P
0	—	$0.70E + 02$	—	$0.97E + 02$	—	$0.21E + 03$
1	1	$0.57E + 02$	1/2	$0.34E + 02$	1/2	$0.10E + 03$
2	1/2	$0.20E + 02$	1/2	$0.10E + 02$	1/4	$0.83E + 02$
3	1	$0.16E + 01$	1	$0.55E - 01$	1	$0.57E + 02$
4	1	$0.11E - 01$	1	$0.93E - 04$	1/8	$0.47E + 02$
5	1	$0.18E - 07$	1	$0.89E - 10$	1/2	$0.37E + 02$
6	1	$0.69E - 18$	1	$0.31E - 17$	1	$0.57E + 00$
7					1	$0.53E - 03$
8					1	$0.62E - 08$
9					1	$0.64E - 18$

and the performance index P versus the iteration number N. The descent property on P was enforced by bisecting the stepsize α when necessary. Convergence to the stopping condition (17) was achieved in $N = 6$ iterations for $\tau = 11.3$, $N = 6$ iterations for $\tau = 13.3$, and $N = 9$ iterations for $\tau = 20.0$.

Table IV shows the converged initial values $x(0)$, $y(0)$, $z(0)$, $u(0)$, $w(0)$ and the

TABLE IV

Terminal values at convergence (Holt's problem, MPBVA)

	$\tau = 11.3$	$\tau = 13.3$	$\tau = 20.0$
$x(0)$	$0.0000000000E + 00$	$0.0000000000E + 00$	$0.0000000000E + 00$
$y(0)$	$0.0000000000E + 00$	$0.0000000000E + 00$	$0.0000000000E + 00$
$z(0)$	$-0.9663117990E + 00$	$-0.9663117939E + 00$	$-0.9663118008E + 00$
$u(0)$	$0.0000000000E + 00$	$0.0000000000E + 00$	$0.0000000000E + 00$
$w(0)$	$0.6529095781E + 00$	$0.6529095785E + 00$	$0.6529095780E + 00$
$x(1)$	$-0.1092818831E + 01$	$-0.1186652039E + 01$	$-0.1189935452E + 01$
$y(1)$	$0.0000000000E + 00$	$0.0000000001E + 00$	$0.0000000000E + 00$
$z(1)$	$0.3652563642E - 01$	$0.8972612848E - 01$	$-0.8677066349E - 02$
$u(1)$	$0.1000000000E + 01$	$0.9999999999E + 00$	$0.9999999999E + 00$
$w(1)$	$0.1081711691E + 00$	$0.3887985425E - 02$	$0.8792789783E - 02$

converged final values $x(1)$, $y(1)$, $z(1)$, $u(1)$, $w(1)$ to ten significant figures. For the converged solutions, the tabulated functions $x(t)$, $y(t)$, $z(t)$, $u(t)$, $w(t)$ are given in [2] to four significant figures.

As a check of the accuracy of the solutions obtained, the differential system (26–28) was integrated forward employing the *converged initial conditions* given in Table IV. The numerical results show that the computed values of $y(1)$ and $u(1)$ do not agree with the prescribed final values (28.4) and (28.5). For $\tau = 11.3$ and $\tau = 13.3$, disagreement occurs even in the first significant figure. On the other hand, for $\tau = 20.0$, the forward integration was interrupted because of exponential overflow [nonconvergence (*c*), see Ineq. (18.3)].

In an attempt to analyze the above failure, the eigenvalues of the Jacobian matrix of the differential system (26–28) were computed at the multipoint locations. With reference to the real part of the eigenvalues, Table V shows the largest positive

TABLE V

Extreme values of the real part of the eigenvalues of the Jacobian matrix (Holt's problem, MPBVA)

	$\tau = 11.3$	$\tau = 13.3$	$\tau = 20.0$
Time station	$t = 0.50$	$t = 0.50$	$t = 0.25$
Largest positive eigenvalue (real part)	$\lambda = +31.29$	$\lambda = +33.94$	$\lambda = +56.65$
Time station	$t = 0.00$	$t = 0.00$	$t = 0.00$
Largest negative eigenvalue (real part)	$\lambda = -6.79$	$\lambda = -7.99$	$\lambda = -12.02$

eigenvalue as well as the largest negative eigenvalue. It indicates that the largest positive eigenvalue is 4–5 times greater than the largest negative eigenvalue.

Since integrating backward results in a change in the sign of the real part of the eigenvalues, it was felt that the integration difficulties might be lessened by reversing the sense of integration. With this in mind, the differential system (26–28) was integrated backward employing the *converged final conditions* given in Table IV. The numerical results show that the computed values of $x(0)$, $y(0)$, $u(0)$ agree with the prescribed initial values (28.1), (28.2), (28.3) to five decimal places for $\tau = 11.3$, to five decimal places for $\tau = 13.3$, and to three decimal places for $\tau = 20.0$.

Additional Computer Runs

The results of Table V, interpreted in the light of the discussion of Section 2, indicate that an easier treatment of Holt's problem is possible by integrating the system (26–28) in backward time rather than in forward time. Because of the reduced size of the positive eigenvalues,[9] the criterion (10) can be met even with $\Delta t = 1$, that is, solving Holt's problem with a two-point boundary value approach ($m = 1$).

With this in mind, solutions to Holt's problem (26–28) were computed with the two-point boundary-value approach of [12–13]. For $\tau = 11.3$ and $\tau = 13.3$, 400 integration steps were employed; and, for $\tau = 20.0$, 800 integration step were employed. For all values of τ, the nominal functions (32) were assumed.

For the two-point boundary-value approach in backward time, the behavior of

TABLE VI

Terminal values at convergence (Holt's problem, TPBVA)

	$\tau = 11.3$	$\tau = 13.3$	$\tau = 20.0$
$x(0)$	0.0000000000E + 00	0.0000000000E + 00	0.0000000000E + 00
$y(0)$	−0.9063118009E + 00	−0.9063117985E + 00	−0.9063118619E + 00
$z(0)$	−0.9063118009E + 00	−0.9063117985E + 00	−0.9063118619E + 00
$u(0)$	0.0000000000E + 00	0.0000000000E + 00	0.0000000000E + 00
$w(0)$	0.6529095769E + 00	0.6529095760E + 00	0.6529095773E + 00
$x(1)$	−0.1092818828E + 01	−0.1186652039E + 01	−0.1189935445E + 01
$y(1)$	0.0000000000E + 00	0.0000000000E + 00	0.0000000000E + 00
$z(1)$	0.3652567171E − 01	0.8972614674E − 01	−0.8676219384E − 02
$u(1)$	0.1000000000E + 01	0.1000000000E + 01	0.1000000000E + 01
$w(1)$	0.1081715368E + 00	0.3888089654E − 02	0.8793138212E − 02

[9] We emphasize that the sign of the eigenvalues changes when the integration sense is reversed.

the stepsize α and the performance index P is still represented by Table III, with the following exceptions: the converged values of the performance index are $P = 0.56E - 18$ for $\tau = 11.3$, $P = 0.43E - 22$ for $\tau = 13.3$, and $P = 0.54E - 18$ for $\tau = 20.0$.

Table VI shows the converged initial values $x(0)$, $y(0)$, $z(0)$, $u(0)$, $w(0)$ and the converged final values $x(1)$, $y(1)$, $z(1)$, $u(1)$, $w(1)$. For the converged solutions, the tabulated functions $x(t)$, $y(t)$, $z(t)$, $u(t)$, $w(t)$ are given in [2] to four significant figures.

As a check of the accuracy of the solutions obtained, the differential system (26–28) was integrated backward employing the *converged final conditions* given in Table VI. The numerical results show that the computed values of $x(0)$, $y(0)$, $u(0)$ agree with the prescribed values (28.1), (28.2), (28.3) to five decimal places for $\tau = 11.3$, to six decimal places for $\tau = 13.3$, and to five decimal places for $\tau = 20.0$. Therefore, the results of Table VI are more precise than those of Table IV.

5. Discussion and Conclusions

Two unusually difficult two-point boundary-value problems, those represented by Eqs. (11–12) and (26–28), were converted into multipoint boundary-value problems and solved by means of the modified quasilinearization approach of [5–6]. By properly adjusting the number of multipoints, the spacing between the multipoints, and the total number of integration steps, precise solutions were obtained.

The multipoint modified quasilinearization approach is especially useful in solving those nonlinear, two-point boundary-value problems where the Jacobian matrix is characterized by eigenvalues having large positive real parts at the solution or near the solution. This approach bypasses the integration of the nonlinear equations, which characterizes shooting methods. In addition, for the problems considered here, it eliminates the necessity for continuation and/or perturbation techniques.

Troesch's Problem

In this problem, the largest positive eigenvalue and the largest negative eigenvalue are identical in modulus. Therefore, the forward integration is characterized by the same degree of difficulty as the backward integration. This being the case, computational results were obtained only in forward integration.

A comparison between the present solutions and those obtained by Roberts and Shipman is given in [1]. The higher degree of precision characterizing the present solutions is due to several factors, namely: (i) the integrations were performed with Hamming's method rather than the Runge-Kutta method; (ii) the multipoints

were spaced with proper regard to the distribution of eigenvalues; and (iii) a larger number of integration steps was employed.

Holt's Problem

In this problem, the largest positive eigenvalue is 4–5 times greater than the largest negative eigenvalue. Therefore, the forward integration is characterized by a higher degree of difficulty than the backward integration.

In forward integration, Holt's problem could be solved with a MPBVA, and could not be solved with a TPBVA. This behavior is consistent with that found in [5–6] for several other boundary-value problems. However, the failure of the TPBVA seems to contradict the results of [12–13], where the TPBVA led to correct results for $\tau = 13.0$. Actually, this is not the case, and the explanation for the discrepancy is as follows: the present results were obtained on IBM 370/155, which is characterized by 16 significant figures in double-precision arithmetic; the results of [12–13] were obtained on Burroughs B-5500, which is characterized by 23 significant figures in double-precision arithmetic. Because of the larger number of significant figures, the B-5500 computer allows a larger spacing between the multipoints than the IBM 370/155 computer [see Ineq. (9.2)]. Indeed, for $\tau = 13.0$, the spacing allowed by the B-5500 computer covers the entire interval of integration!

In backward integration, Holt's problem could be solved with a TPBVA, thus eliminating the need for a MPBVA. This circumstance can be explained in the light of the discussion of Section 2: the positive eigenvalues associated with the backward integration are 4–5 times smaller than the positive eigenvalue associated with the forward integration.

Updating Technique

Within the context of the present paper, two methods for constructing the solution $A(t)$ of the linear, multipoint boundary-value problem (6–8) are possible. The *first method* requires saving the particular solutions at all the time stations considered in the integration process. In this case, the composite solutions $A(t)$ is obtained by combining linearly the particular solutions at the above time stations. The *second method* requires saving only the initial conditions employed to generate the particular solutions in each subinterval. In this case, the composite solution $A(t)$ is obtained by first computing the vectors $A(t_i)$, $i = 0, 1, \ldots, m$, at the beginning of each subinterval and then integrating the linear differential system (6) forward once more. Obviously, the second method requires less computer storage than the first method. Of course, this reduction in computer storage is obtained at the expense of one additional integration of the linear system (6) for each iteration.

REFERENCES

1. A. MIELE, A. K. AGGARWAL, AND J. L. TIETZE, Solution of a Two-Point Boundary-Value Problem with Jacobian Matrix Characterized by Extremely Large Eigenvalues, Rice University, Aero-Astronautics Report No. 107, 1972.
2. A. K. AGGARWAL, Some Numerical Results on Holt's Two-Point Boundary-Value Problem, Rice University, Aero-Astronautics Report No. 118, 1973.
3. B. A. TROESCH, Intrinsic Difficulties in the Numerical Solution of a Boundary-Value Problem, TRW, Redondo Beach, California, Internal Report No. NN-142, 1960.
4. S. M. ROBERTS AND J. S. SHIPMAN, Solution of Troesch's two-point boundary-value problem by a combination of techniques, *J. Comp. Phys.* **10** (1972), 232–241.
5. A. MIELE, K. H. WELL, AND J. L. TIETZE, Multipoint Approach to the Two-Point Boundary-Value Problem, Rice University, Aero-Astronautics Report No. 108, 1972.
6. A. MIELE, K. H. WELL, AND J. L. TIETZE, Multipoint approach to the two-point boundary-value problem, *J. Math. Anal. Applic.* **44** (1973), 625–642.
7. J. F. HOLT, Numerical solution of nonlinear two-point boundary-value problems by finite difference methods, *Commun. ACM* **7** (1964), 366–372.
8. R. E. BELLMAN AND R. E. KALABA, "Quasilinearization and Nonlinear Boundary-Value Problems," American Elsevier Publishing Company, New York, New York, 1965.
9. H. B. KELLER, "Numerical Methods for Two-Point Boundary-Value Problems," Blaisdell Publishing Company, Waltham, Massachusetts, 1968.
10. C. A. BAIRD, JR., Quasilinearization and the methods of finite difference and initial values, *J. Opt. Theory Applic.* **6** (1970), 320–330.
11. S. M. ROBERTS AND J. S. SHIPMAN, "Two-Point Boundary-Value Problems: Shooting Methods", American Elsevier Publishing Company, New York, New York, 1972.
12. A. MIELE AND R. R. IYER, General technique for solving nonlinear, two-point boundary-value problems via the method of particular solutions, *J. Optim. Theory Applic.* **5** (1970), 382–399.
13. A. MIELE AND R. R. IYER, Modified quasilinearization method for solving nonlinear, two-point boundary-value problems, *J. Math. Anal. Applic.* **36** (1971), 674–692.
14. D. D. MORRISON, J. D. RILEY, AND J. F. ZANCANARO, Multiple shooting methods for two-point boundary-value problems, *Commun. ACM* **5** (1962), 613–614.
15. M. R. OSBORNE, On shooting methods for boundary-value problems, *J. Math. Anal. Applic.* **27** (1969), 417–433.
16. S. M. ROBERTS AND J. S. SHIPMAN, Multipoint solution of two-point boundary-value problems, *J. Optim. Theory Applic.* **7** (1971), 301–318.
17. S. M. ROBERTS, J. S. SHIPMAN, AND C. V. ROTH, Continuation in quasilinearization, *J. Optim. Theory Applic.* **2** (1968), 164–178.
18. A. RALSTON, Numerical integration methods for the solution of ordinary differential equations, in "Mathematical Methods for Digital Computers" (A. Ralston and H. S. Wilf, Eds.), Vol. 1, John Wiley and Sons, New York, 1960.
19. D. J. JONES, Solution of Troesch's, and other, two-point boundary-value problems by shooting techniques, *J. Comp. Phys.* **12** (1973), 429–434.