JOURNAL OF COMPUTATIONAL PHYSICS 33, 70–85 (1979)

A Numerical Method for Linear Two-Point Boundary-Value Problems Using Compound Matrices

B. S. NG*

Department of Mathematical Sciences, Indiana University–Purdue University, Indianapolis, Indiana 46205

AND

W. H. REID

Department of Mathematics, University of Chicago, Chicago, Illinois 60637

Received August 11, 1978; revised December 4, 1978

An initial-value method, based on the use of certain compound matrices, is presented for the treatment of mathematically unstable two-point boundary-value problems. A distinctive feature of the method is that the solution is obtained from an auxiliary equation which is mathematically stable and it thereby avoids the well-known difficulties associated with, for example, the method of complementary functions. The method is described in detail for fourth-order equations but it is also shown how the method can easily be adapted to deal with second- and third-order equations.

1. INTRODUCTION

Boundary-value problems involving mathematically unstable ordinary differential equations frequently arise in many fields of application. Although such problems can sometimes be treated analytically by the methods of asymptotic analysis or singular perturbation theory, the numerical solution of problems of this type is generally a formidable task. In a recent paper [6], it was shown how certain compound matrices could be used for eigenvalue problems for linear ordinary differential equations and in this paper we wish to show how they can be used for mathematically unstable two-point boundary-value problems.

In Section 2, we give a description of the compound matrix method for fourth-order equations and, in Section 3, two examples are considered which, we believe, provide a severe test of the method. An attempt to extend these results to higher-order systems of equations has led to a number of further questions but, for reasons of simplicity, we have not touched on these matters. In Section 4, however, we indicate how the method can be adapted to deal with second- and third-order equations and

* Present address: Department of Mathematical and Computing Sciences, Old Dominion University, Norfolk, Virginia 23508.

0021-99991/79/100070–16$02.00/0

an example is considered for which the solution has boundary-layer behavior near *both* end points of the interval. Finally, in Section 5, we consider the relationship between the compound matrix method and the well-known Riccati method, and an alternative procedure is suggested for finding the solution which avoids the numerical difficulties associated with the use of the recovery transformation.

2. Description of the Method

For simplicity in the presentation, we shall discuss the method of compound matrices in terms of a boundary-value problem consisting of a single fourth-order differential equation subject to an equal number of separated boundary conditions at the end points. Consider then the equation

$$L_4(\phi) \equiv \phi^{iv} - a_1\phi''' - a_2\phi'' - a_3\phi' - a_4\phi = f, \tag{1}$$

where a_1, a_2, a_3, a_4 and f are functions of x and $0 \leqslant x \leqslant 1$. For our purposes it is convenient to rewrite (1) as a system of first-order equations. Thus, if we let $\boldsymbol{\phi} = [\phi, \phi', \phi'', \phi''']^T$ and $\mathbf{f} = [0, 0, 0, f]^T$, then (1) becomes

$$\boldsymbol{\phi}' = \mathbf{A}(x)\,\boldsymbol{\phi} + \mathbf{f}, \tag{2}$$

where

$$\mathbf{A}(x) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ a_4 & a_3 & a_2 & a_1 \end{bmatrix}. \tag{3}$$

The boundary conditions at $x = 0$ and 1 are then given by

$$\mathbf{P}\boldsymbol{\phi}(0) = \mathbf{p} \qquad \text{and} \qquad \mathbf{Q}\boldsymbol{\phi}(1) = \mathbf{q}, \tag{4a, b}$$

where

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \end{bmatrix} \qquad \text{and} \qquad \mathbf{Q} = \begin{bmatrix} q_{11} & q_{12} & q_{13} & q_{14} \\ q_{21} & q_{22} & q_{23} & q_{24} \end{bmatrix} \tag{5a,b}$$

are matrices of rank 2, and $\mathbf{p}$ and $\mathbf{q}$ are 2×1 column vectors. The solution to (2) subject to the boundary conditions (4) can now be written in the form

$$\boldsymbol{\phi} = \boldsymbol{\phi}_0 + \alpha\boldsymbol{\phi}_1 + \beta\boldsymbol{\phi}_2, \tag{6}$$

where $\boldsymbol{\phi}_0$ is any solution of (2) which satisfies the intial conditions (4a), while $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ are two linearly independent solutions of the homogeneous system

$$\boldsymbol{\phi}' = \mathbf{A}(x)\,\boldsymbol{\phi} \tag{7}$$

which satisfy the initial conditions

$$\mathbf{P}\boldsymbol{\phi}_i(0) = \mathbf{0} \qquad (i = 1, 2). \tag{8}$$

In the usual application of the method of complementary functions, ϕ_0, ϕ_1, and ϕ_2 must be computed separately. The boundary conditions at $x = 1$ then lead to a pair of linear equations from which, at least in principle, the constants α and β in (6) can be determined. Nevertheless, it is well-known [1] that the coefficient matrix of the linear equations for α and β can be highly ill-conditioned if the solutions of (7) exhibit inherent growth problems. In that case a further loss of accuracy will occur due to cancellation errors as ϕ must be obtained from (6) by superposition. In this section, therefore, we wish to describe a method, based on the use of certain compound matrices, which appears to overcome these numerical difficulties.

Consider then the 4×3 solution matrix of the inhomogeneous system (2)

$$\Phi_0 = \begin{bmatrix} \phi_0 & \phi_1 & \phi_2 \\ \phi_0' & \phi_1' & \phi_2' \\ \phi_0'' & \phi_1'' & \phi_2'' \\ \phi_0''' & \phi_1''' & \phi_2''' \end{bmatrix} \tag{9}$$

and the 4×2 solution matrix of the corresponding homogeneous system (7)

$$\Phi = \begin{bmatrix} \phi_1 & \phi_2 \\ \phi_1' & \phi_2' \\ \phi_1'' & \phi_2'' \\ \phi_1''' & \phi_1''' \end{bmatrix} \tag{10}$$

The 2×2 minors of Φ are

$$\begin{aligned} y_1 &= \phi_1\phi_2' - \phi_1'\phi_2, & y_4 &= \phi_1'\phi_2'' - \phi_1''\phi_2', \\ y_2 &= \phi_1\phi_2'' - \phi_1''\phi_2, & y_5 &= \phi_1'\phi_2''' - \phi_1'''\phi_2', \\ y_3 &= \phi_1\phi_2''' - \phi_1'''\phi_2, & y_6 &= \phi_1''\phi_2''' - \phi_1'''\phi_2'', \end{aligned} \tag{11}$$

and they satisfy the quadratic identity

$$y_1 y_6 - y_2 y_5 + y_3 y_4 = 0. \tag{12}$$

The 3×3 minors of Φ_0 are then given by

$$\begin{aligned} z_1 &= y_1\phi_0'' - y_2\phi_0' + y_4\phi_0, \\ z_2 &= y_1\phi_0''' - y_3\phi_0' + y_5\phi_0, \\ z_3 &= y_2\phi_0''' - y_3\phi_0'' + y_6\phi_0, \\ z_4 &= y_4\phi_0''' - y_5\phi_0'' + y_6\phi_0'. \end{aligned} \tag{13}$$

For later purposes, we also note the further identities

$$\begin{aligned} y_3z_1 - y_2z_2 + y_1z_3 &= 0, \\ y_5z_1 - y_4z_2 + y_1z_4 &= 0, \\ y_6z_1 - y_4z_3 + y_2z_4 &= 0, \\ y_6z_2 - y_5z_3 + y_3z_4 &= 0. \end{aligned} \tag{14}$$

These relations are not independent, however, since if any two of the identities (14) are given then the other two can be derived from them.

If we now let $\mathbf{y} = [y_1, \ldots, y_6]^T$ and $\mathbf{z} = [z_1, \ldots, z_4]^T$, then $\mathbf{y}$ is simply a second compound of Φ and $\mathbf{z}$ is a third compound of Φ_0. Moreover, by a direct calculation, it can easily be shown that the components of $\mathbf{y}$ must satisfy the equations

$$\begin{aligned} y_1' &= y_2, \\ y_2' &= y_3 + y_4, \\ y_3' &= a_3y_1 + a_2y_2 + a_1y_3 + y_5, \\ y_4' &= y_5, \\ y_5' &= -a_4y_1 + a_2y_4 + a_1y_5 + y_6, \\ y_6' &= -a_4y_2 - a_3y_4 + a_1y_6. \end{aligned} \tag{15}$$

Similarly, the components of $\mathbf{z}$ must satisfy the equations

$$\begin{aligned} z_1' &= z_2, \\ z_2' &= a_2z_1 + a_1z_2 + z_3 + y_1f, \\ z_3' &= -a_3z_1 + a_1z_3 + z_4 + y_2f, \\ z_4' &= a_4z_1 + a_1z_4 + y_4f. \end{aligned} \tag{16}$$

By using (11) and (13), we can immediately derive the initial conditions for $\mathbf{y}$ and $\mathbf{z}$ at $x = 0$. For example, if $\phi(0) = c$ and $\phi'(0) = d$, as these are the relevant boundary conditions at $x = 0$ for the numerical examples to be discussed in Section 3, then we can choose

$$\boldsymbol{\phi}_0(0) = [c, d, 0, 0]^T, \qquad \boldsymbol{\phi}_1(0) = [0, 0, 1, 0]^T, \qquad \text{and} \qquad \boldsymbol{\phi}_2(0) = [0, 0, 0, 1]^T. \tag{17}$$

The initial conditions for $\mathbf{y}$ and $\mathbf{z}$ are then given by

$$\mathbf{y}(0) = [0, 0, 0, 0, 0, 1]^T \qquad \text{and} \qquad \mathbf{z}(0) = [0, 0, c, d]^T. \tag{18a, b}$$

Once $\mathbf{y}$ and $\mathbf{z}$ have been obtained by integrating (15) and (16) from $x = 0$ to 1, subject to these or other appropriate initial conditions, the solution $\boldsymbol{\phi}$ can be determined in the following manner.

We first note that there must exist constants α and β such that

$$\begin{aligned} (\phi - \phi_0) &= \alpha\phi_1 + \beta\phi_2, \qquad (\phi - \phi_0)'' = \alpha\phi_1'' + \beta\phi_2'', \\ (\phi - \phi_0)' &= \alpha\phi_1' + \beta\phi_2', \qquad (\phi - \phi_0)''' = \alpha\phi_1''' + \beta\phi_2'''. \end{aligned} \tag{19}$$

We do not, of course, have any knowledge of ϕ_0, ϕ_1, and ϕ_2 separately. However, on eliminating α and β from (19) in four different ways and then using (13), it is easy to show that ϕ must satisfy

$$y_1\phi'' - y_2\phi' + y_4\phi = z_1, \tag{20}$$

$$y_1\phi''' - y_3\phi' + y_5\phi = z_2, \tag{21}$$

$$y_2\phi''' - y_3\phi'' + y_6\phi = z_3, \tag{22}$$

$$y_4\phi''' - y_5\phi'' + y_6\phi' = z_4. \tag{23}$$

Thus, once $\boldsymbol{\phi}(1)$ is known, these equations suggest that the solution ϕ to the boundary-value problem can be obtained by integrating any one of them backwards from $x = 1$ to 0. It should be mentioned, however, that $x = 0$ is often a regular singular point of Eqs. (20) to (23). For example, consider the case when $\mathbf{y}(0)$ is given by (18a). By considering the behavior of the solutions of Eqs. (20) to (23) as $x \to 0$, it can easily be shown [6] that $x = 0$ is a regular singular point of the equations and at that point they have exponents (2, 3), (−2, 2, 3), (−½, 2, 3), and (0, 2, 3), respectively. Thus, if one were to obtain ϕ by integrating Eqs. (21), (22), or (23) from $x = 1$ to 0, then some numerical difficulties would be expected due to the exponents −2, −½, and 0 of these equations respectively at $x = 0$. In this case, therefore, only Eq. (20) should be used to obtain ϕ.

To determine the initial condition for $\boldsymbol{\phi}$ at $x = 1$, we first rewrite Eqs. (20) to (23) as a system of linear equations for the unknown vector $\boldsymbol{\phi}(1)$ in the form

$$\mathbf{M}(1)\,\boldsymbol{\phi}(1) = \mathbf{z}(1), \tag{24}$$

where

$$\mathbf{M}(x) = \begin{bmatrix} y_4 & -y_2 & y_1 & 0 \\ y_5 & -y_3 & 0 & y_1 \\ y_6 & 0 & -y_3 & y_2 \\ 0 & y_6 & -y_5 & y_4 \end{bmatrix}. \tag{25}$$

It is easy to show that $\mathbf{M}(1)$ is of rank 2. For example, if we suppose that $y_4(1) \neq 0$ (say), then by using row reduction and the quadratic identity (12) we can show that

$$\mathbf{M}(1) \sim \begin{bmatrix} y_4(1) & -y_2(1) & y_1(1) & 0 \\ 0 & y_6(1) & -y_5(1) & y_4(1) \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{26}$$

Clearly then the boundary conditions (4b) together with Eq. (24) form a system of linear equations from which $\phi(1)$ can be uniquely determined provided that

$$\det \begin{bmatrix} q_{11} & q_{12} & q_{13} & q_{14} \\ q_{12} & q_{22} & q_{23} & q_{24} \\ y_4(1) & -y_2(1) & y(1) & 0 \\ 0 & y_6(1) & -y_5(1) & y_4(1) \end{bmatrix} \neq 0. \tag{27}$$

We note that this condition must be satisfied if the boundary-value problem is to have a unique solution. This follows from the fact, which can be verified by a direct calculation, that (27) is equivalent to the condition that $\det[\mathbf{Q}\Phi(1)] \neq 0$, where Φ is any solution matrix of the homogeneous system (7) which satisfies the homogeneous boundary conditions $\mathbf{P}\Phi(0) = 0$.

Suppose now that $\phi(1)$ has been determined in the manner described above. It then remains to be shown that the solution ϕ which satisfies these initial conditions and any one of Eqs. (20) to (23) is also a solution of Eq. (1). For this purpose let

$$\mathbf{w} = \mathbf{M}(x)\,\phi - \mathbf{z}, \tag{28}$$

where $\mathbf{w} = [w_1, \ldots, w_4]^T$. On differentiating (28) and rewriting in component form we have

$$w_1' = w_2, \tag{29}$$

$$w_2' = a_2 w_1 + a_1 w_2 + w_3 + y_1\{L_4(\phi) - f\}, \tag{30}$$

$$w_3' = -a_3 w_1 + a_1 w_3 + w_4 + y_2\{L_4(\phi) - f\}, \tag{31}$$

$$w_4' = a_4 w_1 + a_1 w_4 + y_4\{L_4(\phi) - f\}. \tag{32}$$

Since $\phi(1)$ is chosen so that (24) is satisfied, we must have $\mathbf{w}(1) = \mathbf{0}$. If we now suppose that ϕ has been obtained by integrating Eq. (20), then $w_1 \equiv 0$ and it follows from (29) that $w_2 \equiv 0$ also. Thus, Eqs. (30) to (33) become

$$0 = w_3 + y_1\{L_4(\phi) - f\}, \tag{33}$$

$$w_3' = a_1 w_3 + w_4 + y_2\{L_4(\phi) - f\}, \tag{34}$$

$$w_4' = a_1 w_4 + y_4\{L_4(\phi) - f\}. \tag{35}$$

On using (33) to eliminate $L_4\phi - f$ from (34) and (35), we obtain a pair of homogeneous first-order equations for w_3 and w_4. The only solution of these equations which satisfies the initial conditions $w_3(1) = w_4(1) = 0$ is the trivial one and hence it follows that $L_4\phi - f \equiv 0$. By a similar argument it can also be shown that if ϕ satisfies Eqs. (21), (22), or (23), then it is also a solution of Eq. (1).

Finally, it should be noted that the method described in this section can be simplified in the case of a boundary-value problem involving a homogeneous equation and homogeneous boundary conditions at $x = 0$ (say); for, if $f = 0$ and $\mathbf{P}\boldsymbol{\phi}(0) = \mathbf{0}$, then $\mathbf{z} \equiv 0$ for $0 \leqslant x \leqslant 1$. Thus, it is sufficient to compute $\mathbf{y}$ by integrating (15) from $x = 0$ to 1, and the solution ϕ can then be obtained by integrating Eq. (20) with $z_1 \equiv 0$ from $x = 1$ to 0.

3. Numerical Examples

3.1. *Conte's Problem*

To test the effectiveness of the compound matrix method on unstable boundary-value problems, we consider first an example discussed by Conte [1] in connection with the method of orthonormalization. Thus, consider the equation

$$\phi^{iv} - (1 + k^2)\,\phi'' + k^2\phi = -1 + \tfrac{1}{2}k^2x^2, \tag{36}$$

subject to the boundary conditions

$$\phi(0) = 1, \qquad \phi'(0) = 0 \tag{37}$$

and

$$\phi(1) = \tfrac{3}{2} + \sinh 1, \qquad \phi'(1) = 1 + \cosh 1. \tag{38}$$

This two-point boundary-value problem can, of course, be solved analytically and the exact solution

$$\phi(x) = 1 + \tfrac{1}{2}x^2 + \sinh x \tag{39}$$

has the distinctive feature of being independent of the parameter k. However, on rewriting Eq. (36) as a first-order system of the form (2), it can easily be seen that the eigenvalues of the coefficient matrix $\mathbf{A}$ are ± 1 and $\pm k$. Clearly then, when k is large, any attempt to determine ϕ by shooting or by the method of complementary functions will encounter severe difficulties because the inevitable presence of some multiples of the solutions $e^{\pm kx}$ will render the direct integration of Eq. (36) inherently unstable. In the compound matrix method, however, ϕ is obtained as the solution of a mathematically stable differential equation and the problem of destructive growth is thereby avoided.

Thus for this problem, with $a_1 = a_3 = 0$, $a_2 = 1 + k^2$, $a_4 = -k^2$, and $f = -1 + \tfrac{1}{2}k^2x^2$, we first integrate Eqs. (15) and (16) from $x = 0$ to 1 subject to the initial conditions

$$\mathbf{y}(0) = [0, 0, 0, 0, 0, 1]^T \qquad \text{and} \qquad \mathbf{z} = [0, 0, 1, 0]^T. \tag{40}$$

Although it is possible to solve for $\mathbf{y}(x)$ and $\mathbf{z}(x)$ analytically, it is sufficient for the

present purposes to note that for $x > 0$ and $kx \gg 1$ we have the asymptotic approximation

$$\mathbf{y}(x) \sim -\frac{e^{kx}\cosh x}{2(k^2-1)^2} \times \begin{bmatrix} k\tanh x - 2 + k^{-1}\tanh x \\ k^2\tanh x - k - \tanh x + k^{-1} \\ k^3\tanh x - k^2 - 1 + k^{-1}\tanh x \\ k^2 - 2k\tanh x + 1 \\ (k^2-1)(k - \tanh x) \\ k(k^2\tanh x - 2k + \tanh x) \end{bmatrix}. \tag{41}$$

The corresponding approximation to $\mathbf{z}(x)$ is somewhat complicated but it is not needed in the present discussion. Consider now the possibility of determining ϕ by integrating (20) from $x = 1$ to 0. To study the behavior of the solutions of (20) for large values of kx, we may replace y_1, y_2, and y_4 by their approximations (41). On then rewriting (20) in system form, it is found that the eigenvalues of the coefficient matrix associated with (20) are given by

$$\lambda_1(x) \sim k \qquad \text{and} \qquad \lambda_2(x) \sim \frac{k^2 - 2k\tanh x + 1}{k^2\tanh x - 2k + \tanh x} > 0. \tag{42}$$

Clearly this shows that Eq. (20) is stable with respect to backward integration from $x = 1$ to 0 but that it is unstable with repect to forward integration due to the large and positive eigenvalue k. A similar analysis leads to the same conclusion for Eqs. (22)

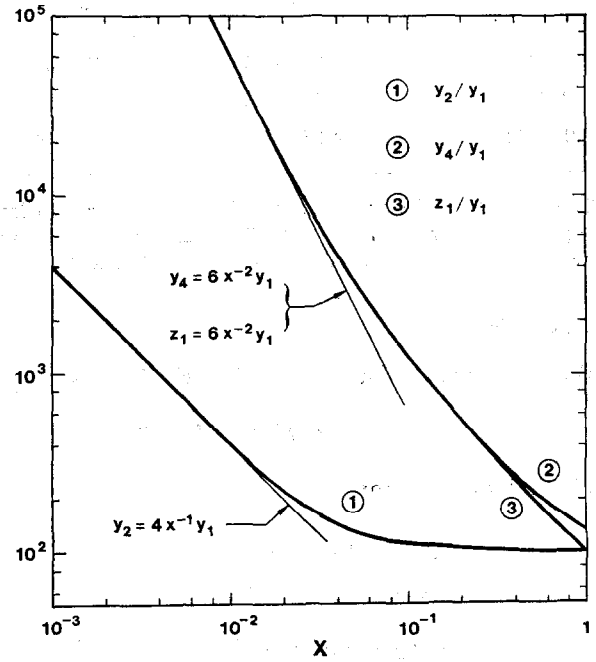

FIG. 1. The behavior of the ratios y_2/y_1, y_4/y_1 and z_1/y_1 for Conte's problem with $k = 10^2$.

and (23). Equation (21), however, is found to be unstable irrespective of the direction of integration since, in addition to k, one of the two remaining eigenvalues of its coefficient matrix is large and negative. These conclusions have also been confirmed by actual numerical experiments. Furthermore, by taking into account the possibly singular nature of the solutions of Eqs. (22) and (23) near $x = 0$ as discussed in Section 2, it is clear that only Eq. (20) should be used to compute the solution ϕ for this problem. In Figure 1 we show the behavior of the ratios y_2/y_1, y_4/y_1, and z_1/y_1 for $k = 10^2$ and we note that they all become unbounded as $x \to 0$. This shows that, even in the case of Eq. (20), backward integration from $x = 1$ to 0 cannot yield the final value of ϕ at $x = 0$, but such behavior does not appear to cause any other numerical difficulties.

We have also computed the solution ϕ for several values of k, ranging up to and including $k = 10^3$. For simplicity, the calculations were made by using a Runge–Kutta–Gill procedure with constant stepsize h and they were performed in single precision arithmetic on a CDC-6600 computer. In Table I, for example, we show the effect of stepsize on the maximum relative errors of ϕ and ϕ' among 50 equally spaced points at $x = 0.02(0.02)\,1$ for $k = 10^2$.

TABLE I[a]

	Maximum relative errors	
h	ϕ	ϕ'
0.001	$1.0E-06$	$5.6E-06$
0.0005	$6.5E-08$	$4.0E-07$
0.00025	$4.2E-09$	$2.6E-08$

[a] The effect of stepsize on the numerical solution of the boundary-value problem (36) to (38) with $k = 10^2$. The maximum relative errors are computed among 50 points located at $x = 0.02(0.02)1.0$.

3.2. *A Boundary-Layer Problem*

Consider next the homogeneous form of Eq. (36) which is given by

$$\phi^{iv} - (1 + k^2)\,\phi'' + k^2\phi = 0 \tag{43}$$

together with the boundary conditions

$$\phi(0) = 0, \qquad \phi'(0) = 0 \tag{44}$$

and

$$\phi(1) = 1, \qquad \phi'(1) = 0. \tag{45}$$

This problem was first studied by Flaherty and O'Malley [3]. It is of the singular perturbation type since ϕ' possesses a boundary layer of thickness $O(k^{-1})$ near each

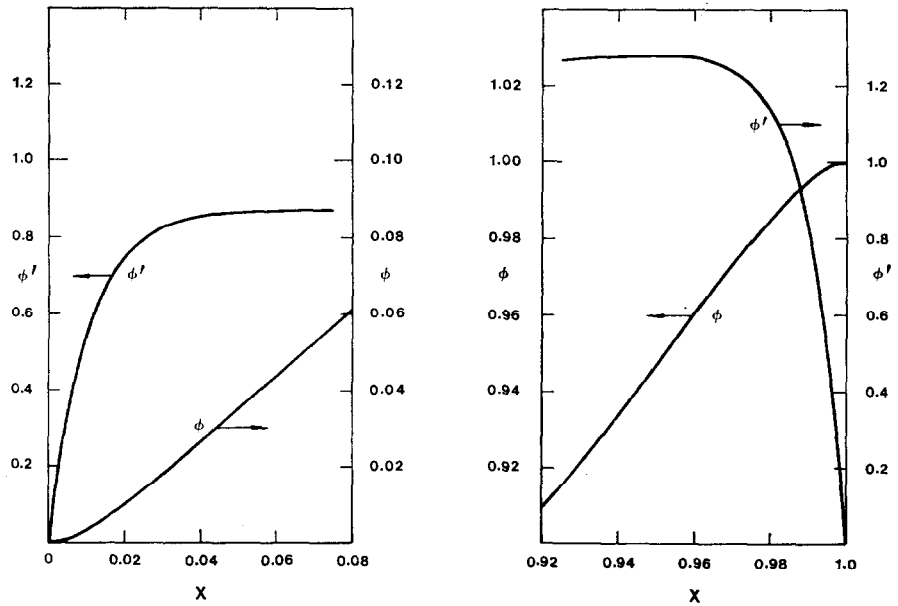

FIG. 2. The behavior of ϕ and ϕ' of the boundary-value problem (43) to (45) near the end points.

of the two end points. The behavior of ϕ and ϕ' in the vicinity of $x = 0$ and 1, as found by the present method, is shown in Fig. 2 for $k = 10^2$. Clearly, many of the same difficulties encountered in Conte's original problem must also arise in this example and the presence of the bondary layers in ϕ' provide a further test of the versatility of the compound matrix method.

Since Eq. (43) and the boundary conditions (44) at $x = 0$ are both homogeneous, we must have $\mathbf{z} \equiv 0$ and it is sufficient therefore to integrate Eqs. (15) from $x = 0$

TABLE II[a]

	Maximum relative errors					
	(i)		(ii)		(iii)	
h	ϕ	ϕ'	ϕ	ϕ'	ϕ	ϕ'
0.001	$3.6E-03$	$4.2E-03$	$1.6E-05$	$6.4E-06$	$3.7E-08$	$2.0E-06$
0.0005	$1.7E-04$	$2.9E-05$	$1.1E-06$	$4.4E-07$	$2.4E-09$	$1.3E-07$
0.00025	$9.4E-06$	$2.9E-06$	$7.0E-08$	$2.9E-08$	$1.5E-10$	$8.3E-09$

[a] The effect of stepsize on the numerical solution of the boundary-value problem (43) to (45) with $k = 10^2$. The maximum relative errors are computed among points located at (i) $x = 0.002(0.002) \cdot 0.02$, (ii) $x = 0.04(0.02)0.96$, and (iii) $x = 0.98(0.002)1.0$.

to 1 subject to the initial conditions (18a). The solution can then be obtained by backward integration of Eq. (20) with $z_1 \equiv 0$. We have computed ϕ and ϕ', using different but uniform stepsizes, for $k = 10^2$ and a comparison of these numerical results with the analytical solution is given in Table II.

It should be noted that results of comparable accuracy to those presented here and in Section 3.1 can very likely be obtained with a substantial reduction in the number of integration steps by using, for example, the Runge–Kutta–Fehlberg integration procedure. We have not exploited such a possibility here because our primary aim is is simply to show how reasonable accuracy can be achieved with a minimal amount of programming effort.

4. Second- and Third-Order Problems

Thus far we have restricted our discussion to boundary-value problems involving fourth-order equations. The basic ideas involved in the use of compound matrices can be generalized in a variety of ways, and in this section therefore we wish to show how the method can be applied to second- and third-order problems.

Consider first a single third-order equation

$$L_3(\phi) = \phi''' - b_1\phi'' - b_2\phi' - b_3\phi = f, \tag{46}$$

where b_1, b_2, b_3, and f are functions of x and $0 \leqslant x \leqslant 1$. We now let $\boldsymbol{\phi} = [\phi, \phi', \phi'']^T$; we shall also suppose that a single boundary condition is prescribed at $x = 0$ and is given by

$$\mathbf{P}\boldsymbol{\phi}(0) = p, \tag{47}$$

where, in this case, $\mathbf{P}$ is a 1×3 row vector and p is a constant. If, as before, we let $\boldsymbol{\phi}_0$ denote the solution of (46) which satisfies (47) and let $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ denote two linearly independent solutions of the corresponding homogeneous system, then the solution matrices $\boldsymbol{\Phi}_0$ and $\boldsymbol{\Phi}$ are given by

$$\boldsymbol{\Phi}_0 = \begin{bmatrix} \phi_0 & \phi_1 & \phi_2 \\ \phi_0' & \phi_1' & \phi_2' \\ \phi_0'' & \phi_1'' & \phi_2'' \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Phi} = \begin{bmatrix} \phi_1 & \phi_2 \\ \phi_1' & \phi_2' \\ \phi_1'' & \phi_2'' \end{bmatrix}. \tag{48}$$

The 2×2 minors of $\boldsymbol{\Phi}$ are now defined by

$$y_1 = \phi_1\phi_2' - \phi_1'\phi_2, \qquad y_2 = \phi_1\phi_2'' - \phi_1''\phi_2, \qquad y_3 = \phi_1'\phi_2'' - \phi_1''\phi_2', \tag{49}$$

and the only 3×3 minor of $\boldsymbol{\Phi}_0$ is simply its determinant which can be expressed in the form

$$z = y_1\phi_0'' - y_2\phi_0' + y_3\phi_0. \tag{50}$$

On differentiating (49) and (50) and eliminating the third derivatives by the use of (46), we then obtain

$$\begin{aligned} y_1' &= y_2, \\ y_2' &= y_3 + b_1 y_2 + b_2 y_1, \\ y_3' &= b_1 y_3 - b_3 y_1, \\ z' &= b_1 z + y_1 f, \end{aligned} \tag{51}$$

where the initial conditions for y_1, y_2, y_3, and z follow directly from the corresponding conditions on ϕ_0, ϕ_1, and ϕ_2. By using an argument similar to the one discussed in Section 2 for the fourth-order case, it follows immediately that the solution must satisfy the equation

$$y_1\phi'' - y_2\phi' + y_3\phi = z \tag{52}$$

together with the prescribed boundary conditions at $x = 1$.

The foregoing analysis can easily be adapted to deal with boundary-value problems involving a second-order equation of the form

$$L_2(\psi) = \psi'' - b_1\psi' - b_2\psi = f \tag{53}$$

together with separated boundary conditions at $x = 0$ and 1. If we now let $\psi = \phi'$, then Eq. (53) becomes

$$\phi''' - b_1\phi'' - b_2\phi' = f \tag{54}$$

and our discussion of third-order problems is then directly applicable on setting $b_3 \equiv 0$. Moreover, a further simplification is possible if we assume, as is often the case, that the boundary condition at $x = 0$ is imposed on either ψ or ψ'. In that case we have $y_3(0) = 0$ and, with $b_3 \equiv 0$, the third of Eqs. (51) shows that $y_3(x) \equiv 0$. Thus, it is sufficient to integrate the remaining three of Eqs. (51) from $x = 0$ to 1 to determine y_1, y_2, and z. The solution ψ can then be obtained by integrating the *first-order* equation

$$y_1\psi' - y_2\psi = z \tag{55}$$

from $x = 1$ to 0.

As an application of the procedure just described, we have used it to compute the solution of the equation

$$\psi'' + \psi' - k^2\psi = 0 \tag{56}$$

subject to the boundary conditions

$$\psi(0) = 1 \qquad \text{and} \qquad \psi(1) = \tfrac{1}{2}. \tag{57}$$

This problem is also of the singular perturbation type and it can easily be shown that ψ exhibits boundary-layer behavior in intervals of thickness $O(k^{-1})$ near each of the two end points. Moreover, the eigenvalues of the coefficient matrix associated with (56) are $\frac{1}{2}[-1 \pm (1 + 4k^2)^{1/2}]$. For large values of k, integration of (56) is therefore inherently unstable and it has been observed [3] that the solution of this boundary-value problem cannot be obtained by shooting for $k \gtrsim 30$. We encountered no difficulties, however, in computing ψ using the present method for values of k up to and

TABLE III

A Comparison of the Numerical and the Exact Solution of the Boundary-Value Problem (56) and (57) with $k = 10^2$ and $h = 0.0005$

x	Computed ϕ	Absolute error	Relative error
0.000	*a*	—	—
0.002	0.817913E + 00	0.27E − 05	3.4E − 06
0.004	0.668978E + 00	0.74E − 06	1.1E − 06
0.006	0.547164E + 00	0.38E − 06	6.9E − 07
0.008	0.447531E + 00	0.25E − 06	5.5E − 07
0.010	0.366040E + 00	0.18E − 06	4.9E − 07
0.020	0.139985E + 00	0.61E − 07	4.6E − 07
0.100	0.431804E − 04	0.36E − 10	8.4E − 07
0.200	0.186455E − 08	0.25E − 14	1.3E − 06
0.300	0.805118E − 13	0.15E − 18	1.8E − 06
0.400	0.347653E − 17	0.80E − 23	2.3E − 06
0.500	0.273869E − 21	0.75E − 27	2.7E − 06
0.600	0.259318E − 17	0.55E − 23	2.1E − 06
0.700	0.543397E − 13	0.86E − 19	1.6E − 06
0.800	0.113868E − 08	0.12E − 14	1.1E − 06
0.900	0.238608E − 04	0.13E − 10	5.3E − 07
0.980	0.683460E − 01	0.72E − 08	1.1E − 07
0.990	0.184859E + 00	0.98E − 08	5.3E − 08
0.992	0.225563E + 00	0.96E − 08	4.2E − 08
0.994	0.275228E + 00	0.87E − 08	3.2E − 08
0.996	0.335829E + 00	0.71E − 08	2.1E − 08
0.998	0.409774E + 00	0.43E − 08	1.1E − 08
1.000	0.500000E + 00	0.00	0.0

[a] Cannot be obtained numerically by the present method as discussed in the text.

including 10^3. A comparison of the numerical results with the analytical solution is given in Table III for $k = 10^2$. These results clearly show the rapid variation of the solution in the boundary layers near the two end points.

5. Relationship to the Riccati Method

Various initial-value methods have been proposed in the past to deal with mathematically unstable two-point boundary-value problems. Among these, the Riccati method [7] has attracted considerable attention recently. It is of some interest, therefore, to consider briefly certain relations between the Riccati method and the compound matrix method. In particular, a modification of the usual Riccati method for determining the solution will be suggested which appears to overcome the difficulties discussed by Nelson and Giles [5].

To fix ideas, we shall again consider an inhomogeneous fourth-order equation of the form (1). For simplicity, we shall also suppose that the boundary conditions at $x = 0$ and 1 are given by $\mathbf{u}(0) = \mathbf{p}$ and $\mathbf{v}(1) = \mathbf{q}$, where $\mathbf{u} = [\phi, \phi']^T$, $\mathbf{v} = [\phi'', \phi''']^T$, and $\mathbf{p}$ and $\mathbf{q}$ are constant 2-vectors. The first step in the application of the Riccati method to this boundary-value problem [4, 7] is to define a transformation of the form

$$\mathbf{u} = \mathbf{R}\mathbf{v} + \mathbf{g}. \tag{58}$$

It can then be shown [4, 7] that the 2×2 Riccati matrix $\mathbf{R}$ and the 2-vector $\mathbf{g}$ must satisfy the equations

$$\mathbf{R}' = \mathbf{A}_{11}\mathbf{R} - \mathbf{R}\mathbf{A}_{22} - \mathbf{R}\mathbf{A}_{21}\mathbf{R} + \mathbf{A}_{12} \tag{59}$$

and

$$\mathbf{g}' = (\mathbf{A}_{11} - \mathbf{R}\mathbf{A}_{21})\,\mathbf{g} - \mathbf{R}\mathbf{f}_2\,, \tag{60}$$

where $\mathbf{f}_2 = [0, f]^T$, and $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, $\mathbf{A}_{21}$, and $\mathbf{A}_{22}$ are 2×2 submatrices of the coefficient matrix in (2), i.e.

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}. \tag{61}$$

It can easily be seen from (58) that the natural initial conditions for $\mathbf{R}$ and $\mathbf{g}$ are simply $\mathbf{R}(0) = \mathbf{0}$ and $\mathbf{g}(0) = \mathbf{p}$.

The relationship between the elements of the Riccati matrix $\mathbf{R}$ and the components of the second compound $\mathbf{y}$ has been discussed by Davey [2] in connection with the removal of certain singularities from the Riccati method. Thus, following Davey, if we let

$$r_1 = y_3/y_6\,, \qquad r_2 = -y_2/y_6\,, \qquad r_3 = y_5/y_6\,, \qquad \text{and} \qquad r_4 = -y_4/y_6\,, \tag{62}$$

then the quadratic identity (12) becomes

$$y_1/y_6 = r_1 r_4 - r_2 r_3\,. \tag{63}$$

On differentiating (62), and then using (15) and (63), we can immediately show that the ratios (62), as the notation suggests, are, in fact, the elements of the Riccati matrix

$$\mathbf{R} = \begin{bmatrix} r_1 & r_2 \\ r_3 & r_4 \end{bmatrix}. \tag{64}$$

More generally, by rewriting Eqs. (15) in terms of ratios of the form y_i/y_j for some fixed j, the system of six linear equations can always be reduced to a system of four non-linear equations of Riccati type. It should be emphasized, however, that such a reduction is desirable only if a $y_j(x)$ can be chosen such that it does not vanish anywhere in the interval $0 \leqslant x \leqslant 1$.

The relationship between $\mathbf{g}$ and the third compound $\mathbf{z}$ can be dicussed in a similar way by considering ratios of the form z_i/y_j for some fixed j. Thus, if we let $s_i = z_i/y_6$ $(i = 1, 2, 3, 4)$, then a short calculation shows that Eqs. (16) can be rewritten as

$$s_1' = -(a_1 + a_3 r_4 + a_4 r_2)\, s_1 + s_2\,, \tag{65a}$$

$$s_2' = a_2 s_1 - (a_3 r_4 + a_4 r_2)\, s_2 + s_3 + (r_1 r_4 - r_2 r_3)\, f, \tag{65b}$$

$$s_3' = -a_3 s_1 - (a_3 r_4 + a_4 r_2)\, s_3 + s_4 - r_2 f \tag{66a}$$

and

$$s_4' = a_4 s_1 - (a_3 r_4 + a_4 r_2)\, s_4 - r_4 f. \tag{66b}$$

Furthermore, if we now let $\mathbf{g}_1 = [s_1\,, s_2]^T$ and $\mathbf{g}_2 = [s_3\,, s_4]^T$, then it can be shown that the identities (14) are equivalent to the single matrix equation

$$\mathbf{R}\mathbf{g}_1 + (\det \mathbf{R})\, \mathbf{g}_2 = 0. \tag{67}$$

By using this equation to eliminate s_3 and s_1 from Eqs. (65) and (66) respectively, we immediately obtain two uncoupled systems for $\mathbf{g}_1$ and $\mathbf{g}_2$. In particular, it is found that $\mathbf{g}_2$ satisfies Eq. (60) and hence $\mathbf{g}_2 \equiv \mathbf{g}$.

In the usual application of the Riccati method [7], the next step is to introduce a recovery transformation of the form

$$\mathbf{v}(0) = \mathbf{T}(x)\, \mathbf{v}(x) + \mathbf{h}(x), \tag{68}$$

where the 2×2 matrix $\mathbf{T}$ and the 2-vector $\mathbf{h}$ satisfy the equations

$$\mathbf{T}' = -\mathbf{T}(\mathbf{A}_{22} + \mathbf{A}_{21}\mathbf{R}), \tag{69}$$

$$\mathbf{h}' = -\mathbf{T}(\mathbf{A}_{21}\mathbf{h} + \mathbf{f}_2), \tag{70}$$

and the initial conditions $\mathbf{T}(0) = \mathbf{I}$ and $\mathbf{h}(0) = \mathbf{0}$. To compute the solution of the boundary-value problem, it is then necessary first to integrate Eqs. (59), (60), (69), and (70) subject to the appropriate initial conditions. Once $\mathbf{R}$, $\mathbf{g}$, $\mathbf{T}$, and $\mathbf{h}$ are known

on the interval $0 \leqslant x \leqslant 1$, the solution $\mathbf{u}$ and $\mathbf{v}$ can be obtained algebraically from (58) and (68) by first noting that

$$\mathbf{v}(0) = \mathbf{T}(1)\,\mathbf{q} + \mathbf{h}(1). \tag{71}$$

It has been observed by Nelson and Giles [5], however, that a direct application of this procedure may result in a severe loss of accuracy in the numerical solution due to possible cancellation errors in (68). These difficulties can partially be overcome by using the method of successive starts [5, 7]. Alternatively, Eqs. (20) to (23) can be rewritten in terms of the Riccati variables r_i and s_i $(i = 1, 2, 3, 4)$ in the form

$$(r_1 r_4 - r_2 r_3)\,\phi'' - r_2\phi' + r_4\phi = s_1\,, \tag{72}$$

$$(r_1 r_4 - r_2 r_3)\,\phi''' - r_1\phi' + r_3\phi = s_2\,, \tag{73}$$

$$-r_2\phi''' - r_1\phi'' + \phi = s_3\,, \tag{74}$$

$$-r_4\phi''' - r_3\phi'' + \phi' = s_4\,. \tag{75}$$

This suggests that once $\mathbf{R}$ and $\mathbf{g}_2$ have been obtained by integrating (59) and (60) from $x = 0$ and 1, $\mathbf{g}_1$ can be determined from (67), and the solution ϕ can then be obtained by integrating (72) backwards from $x = 1$ to 0. A further alternative would be to obtain $\mathbf{g}_1$ and $\mathbf{g}_2$ directly by integrating (65) and (66), thereby avoiding the use of (67). Thus, both of these procedures eliminate the need for integrating (69) and (70), and they also avoid the numerical difficulties associated with the use of the recovery transformation (68).

ACKNOWLEDGMENTS

The research reported in this paper has been supported in part by the Computing Services of Indiana University–Purdue University at Indianapolis (B.S.N.) and by the National Science Foundation under Grant MCS78-01249 with the University of Chicago (B.S.N. and W.H.R.).

REFERENCES

1. S. D. CONTE, *SIAM Rev.* **8** (1966), 309–321.
2. A. DAVEY, *J. Computational Phys.* **30** (1979), 137–144.
3. J. E. FLAHERTY AND R. E. O'MALLEY, JR., *Math. Comp.* **31** (1977), 66–93.
4. I. H. MUFTI, C. K. CHOW, AND F. T. STOCK, *SIAM Rev.* **11** (1969), 616–619.
5. P. NELSON, JR., AND C. A. GILES, *J. Computational Phys.* **10** (1972), 374–378.
6. B. S. NG AND W. H. REID, *J. Computational Phys.* **30** (1979), 125–136.
7. M. R. SCOTT, "Invariant Imbedding and its Applications to Ordinary Differential Equations," Addison–Wesley, Reading, Mass., 1973.