ON THE NUMERICAL INTEGRATION OF TWO-POINT BOUNDARY VALUE PROBLEMS

JUN 80 J E FLAHERTY, R E O'MALLEY

N00014-76-C-0326
Abstract

We consider a class of two-point boundary value problems for systems of the form \( \frac{dx}{dt} = f(x,y,t,\varepsilon) \), \( \varepsilon \frac{dy}{dt} = g(x,t,\varepsilon) \), where the matrix \( g(x,t,0) \) has a hyperbolic splitting with a fixed number of stable and unstable eigenvalues. Solutions to such boundary value problems can then be expected to have boundary layer behavior near both endpoints in the limit \( \varepsilon \to 0 \). We obtain uniform asymptotic representations of solutions. Our analysis shows, in particular, how to determine the reduced order boundary value problem satisfied by the limiting interior solution. Orthogonal matrix methods are used to determine this reduced problem and appropriate boundary layer corrections in a computationally effective manner. Numerical experiments with model problems illustrate the possibility of multiple solutions and show how our asymptotic results can be used in combination with the COLSYS code for solving two-point problems via collocation.

Initial value problems for stiff systems of ordinary differential equations are now considered to be relatively tractable numerically (cf. Enright et al. (1975)), though still the subject of current research and difficulty (cf., e.g., Miranker (1979) and Kreiss (1978)). Codes appropriate for two-point problems whose solutions involve rapid changes near the endpoints are not readily available, though these problems occur in many applications. We must study them as a prelude to analyzing more difficult problems with interior layers of rapid transition at unknown locations.

*This work was supported in part by the Air Force Office of Scientific Research under Grant AFOSR-80-0192, and by the Office of Naval Research under Contract N00014-76-C-0326.
Specifically, let us consider singular perturbation problems of the form

\begin{align*}
&1. \quad \dot{x} = f(x,y,t,\varepsilon), \\
&2. \quad \varepsilon \dot{y} = g(x,y,t,\varepsilon) = g_1(x,t,\varepsilon) + g_2(x,t,\varepsilon)y, \\
&3. \quad A(x(0),y(0),\varepsilon) = a_1(x(0),\varepsilon) + a_2(x(0),\varepsilon)y(0) = 0, \\
&4. \quad B(x(1),y(1),\varepsilon) = b_1(x(1),\varepsilon) + b_2(x(1),\varepsilon)y(1) = 0,
\end{align*}

where $x$ and $y$ are vectors of dimensions $m$ and $n$, respectively, and where the $n \times n$ matrix $g_2(x,t,0)$ has $k > 0$ (strictly) stable and $n-k > 0$ (strictly) unstable eigenvalues for all $x$ and for $0 \leq t \leq 1$. We shall assume that there are $q \geq k$ initial conditions (3) and $r = m + n - q \geq n-k$ terminal conditions (4), and that the coefficients in (1)-(4) have asymptotic expansions in $\varepsilon$ with all terms being smooth functions of their remaining arguments. Limiting solutions of (1)-(4) will be sought as the small positive parameter $\varepsilon$ tends toward zero.

O'Malley (1980) discusses the behavior of solutions to more nonlinear problems. We've simplified that treatment by supposing that the (potentially) stiff differential system (2) and the separated boundary conditions (3)-(4) are linear in the "fast" variable $y$. Vasil'eva and Butuzov (1978) and O'Malley and Flaherty (1980) discuss problems where $g_2(x,t,0)$ has a nullspace, while Miranker and Wahba (1976) and Kreiss (1979) discuss problems where $g_2(x,t,0)$ has purely imaginary eigenvalues. The numerical treatment of certain interior and shock layer problems is considered in Kreiss (1975), Hemker (1977), Osher (1980), and elsewhere.

For the special problem (1)-(4), we can expect bounded solutions to have the limiting form

\begin{align*}
&\quad x(t,\varepsilon) = X_0(t) + 0(\varepsilon), \\
&\quad y(t,\varepsilon) = Y_0(t) + \mu_0(t/\varepsilon) + \nu_0((1-t)/\varepsilon) + 0(\varepsilon),
\end{align*}

although unbounded solutions to such problems are also possible (cf. O'Malley (1970) and Ferguson (1975)). Indeed, impulsive endpoint behavior is important in applications to control and systems theory (cf. O'Malley (1978)). Here, we'll ask that $\mu_0(t)$ and $\nu_0(0)$ decay to zero as $t$ and $0$, respectively, tend to infinity. Then, $\mu_0$ will provide a rapid transition from $y(0,0)$ to $Y_0(0)$ near $t = 0$, and $\nu_0$ will allow analogous nonuniform convergence as $\varepsilon \to 0$ at the terminal endpoint. The limiting solution, $(X_0,Y_0)$, within $(0,1)$ will necessarily satisfy the limiting system

\begin{align*}
&\quad \dot{X}_0 = f(X_0,Y_0,t,0), \\
&\quad g(X_0,Y_0,t,0) = 0,
\end{align*}

so

\begin{align*}
&\quad Y_0(t) = G(X_0,t) \equiv -q_2^{-1}(X_0,t,0)g_1(X_0,t,0),
\end{align*}
and $X_0$ must satisfy the $m$th order nonlinear (and non-stiff) system

$$X_0 = F(X_0, t) \equiv f(X_0, G(X_0, t), t, 0).$$

We need to specify an appropriate set of $m$ boundary conditions for our limiting solution $X_0(t) = x(t, 0)$ within $(0, 1)$. Thus, we seek a "cancellation law" (cf. Wasow (1944) and Harris (1973)) which selects a combination of $q-k$ of the limiting initial conditions (3) and of $r-n+k$ of the terminal conditions (4) to be satisfied by $X_0$. The limiting solution $(X_0, Y_0)$ will not generally satisfy the $m$ limiting boundary conditions, i.e. $A_0(X_0(0)) \equiv A(X_0(0), Y_0(0), 0)$ and $B_1(X_0(1)) \equiv B(X_0(1), Y_0(1), 0)$ will not vanish. We will, however, find $m$ boundary conditions for $X_0$ of the form

$$
\begin{align*}
\Phi(X_0(0)) &\equiv Z_{0+}(X_0(0))A_0(X_0(0)) = 0, \\
\Psi(X_0(1)) &\equiv Z_{1-}(X_0(1))B_1(X_0(1)) = 0,
\end{align*}
$$

for appropriate matrices $Z_{0+}$ and $Z_{1-}$ with ranks $q-k$ and $r-n+k$, respectively. The nonlinear "reduced" two-point problem (7)-(9) can be expected to have no solution, a unique solution, or many solutions.

To begin our numerical discussion, recall that the Schur decomposition (cf., e.g., Stewart (1973)) guarantees the existence of an orthogonal matrix $E$ such that

$$
\begin{bmatrix}
T_-(x, t) & U(x, t) \\
0 & T_+(x, t)
\end{bmatrix} =
$$

where $T_-$ is $k \times k$ and upper triangular with the stable eigenvalues of $g_2$ and where $T_+$ is upper triangular with $n-k$ unstable eigenvalues. The matrix $E$ can be determined numerically through the QR algorithm, i.e. through a sequence of Householder transformations. We actually need much less than this, and would prefer a procedure which only block triangularized $g_2$. Partitioning

$$
E = [E_\pm E_\mp],
$$

after its $k$th column, note that $E_\pm (E_\mp$ spans the stable (unstable) eigenspace of $g_2$, so $P = E_\pm E_\mp$ and $Q = E_\mp E_\pm$ are complementary projections onto these eigenspaces.

In order to have a limiting solution of the form (4), the initial layer correction term $\mu_0(t)$ must be a decaying solution of

$$
\frac{d\mu_0}{dt} = g_2(X_0(0), 0, 0)\mu_0 \equiv G_0\mu_0,
$$
\[
\frac{d\mu_0}{dt} = \lim_{c \to 0} c \left( \frac{dy}{dt} - \frac{dy_0}{dt} \right) \text{ and } t = cT. \text{ Thus}
\]

(13) \[
\mu_0(T) = \exp(G_0T)\mu_0(0),
\]

will decay provided \( \mu_0(0) \) lies in the stable eigenspace of \( G_0 \), i.e., \( \mu_0(0) = E_0-c_0 \) where \( E_0 = E_0(X_0(0),0) \).

Since we are seeking solutions of the form (5), our representations (6) and (13) imply that the \( q \) limiting initial conditions (3) take the form \( A(x(0,0),y(0,0),0) = A_0(X_0(0),Y_0(0)+E_0-c_0,0) = A_0(X_0(0)) + a_2(X_0(0),0)E_0-c_0 = 0 \), so \( c_0 = -(a_2E_0-)^+A_0 \) where the dagger represents a matrix pseudoinverse. Assuming that \( a_2E_0- \) has its maximal rank \( k \), there will be an orthogonal matrix \( Z_0 = (Z_0^- Z_0^+) \) which reduces it to its row echelon form, i.e., \( Z_0 a_2E_0- = (V_0^0) \) where \( V_0 \) is \( k \times k \) and nonsingular. After multiplying by \( Z_0 \), the first \( k \) rows of the product provide the unknown (but, due to the flexibility in selecting \( Z_0 \), not necessarily unique) vector

(14) \[
\mu_0(0) = -E_0V_0^{-1}Z_0^0A_0,
\]

needed to specify \( \mu_0(t) \). The remaining \( q-k \) equations of the product provide the initial conditions (8) needed to define the reduced problem for \( X_0(t) \).

In analogous fashion, if we suppose that \( b_2+1_+ = b_2(X_0(1),0) \cdot E_+(X_0(1),1) \) has its maximal rank \( n-k \), and let \( Z_1 = (Z_1^- Z_1^+) \) be such that \( Z_1 b_2+1_+ = (V_1^0) \) is row-reduced, the decaying terminal boundary layer correction term

(15) \[
V_0(0) = -\exp(g_2(X_0(1),1,0) \sigma)E_{+1} V_1^{-1}Z_{1+}B_0,
\]

becomes (usually nonuniquely) specified as does the set of terminal conditions (9) for \( X_0(t) \).

Because the orthogonal matrices \( E_0, Z_0, E_1 \), and \( Z_1 \) depend on the corresponding \( m \) vectors \( x_0(0) = x(0,0) \) and \( x_0(1) = x(1,0) \), we shall usually have to use Newton's method to actually determine them whenever \( g_2(x,t,0), a_2(x,0), \) or \( b_2(x,0) \) actually depend on the \( x \) variable at \( t = 0 \) or \( 1 \). [Alternatively, unless the boundary conditions (3) and (4) directly prescribe or determine \( x(0,c) \) or \( x(1,c) \), a shooting technique (cf. Keller (1968)), which involves guessing the endvalue \( x_0(0) \) or \( x_0(1) \) of the slow \( x(t,0) \) vector and integrating \( X_0 \) to the opposite endpoint, could be attempted. Such a procedure would be far better than shooting for the full vector \( (x,y) \), an approach which does not work well for problems with boundary layer behavior at both}
Newton's method requires a satisfactory initial guess \( X_0(t) \) for \( X_0(t) \). We would proceed to determine successive iterates \( X^V(t) \) until we achieved numerical convergence. At the \( V \)th step, we would calculate an approximation \( X^V \) to \( E(X_0(p),p,0) \) for \( p = 0 \) and \( 1 \) by applying the QR procedure to \( q_2(X(p),p,0) \). One QR step with the matrix \( Z^V \) will reduce each of \( a_2(X(0),0)E_-(X(0),0) \) and \( b_2(X^V(1),1)E_+(X^V(1),1) \) to row echelon form. A linear problem for \( \delta X^V(t) = X^{V+1}(t) - X^V(t) \) is thereby determined, presuming we can neglect the influence of derivatives of \( Z \) and \( Z_i+1 \) on the selection of the boundary conditions.

To illustrate such problems numerically, we have experimented with the simple harmonic oscillator system

\[
\begin{align*}
\dot{x} &= 1 - x, \\
\dot{y}_1 &= y_2, \\
\dot{y}_2 &= \alpha^2(x)y_1 + \beta(x),
\end{align*}
\]

made nonlinear since the coefficients \( \alpha(x) = 1 + 2x \) and \( \beta(x) = 8x(1-x) \) depend on the slow vector \( x \). The three linear boundary conditions are

\[
\begin{align*}
x(0) + y_1(0) &= 0, \\
-bx(0) + y_2(0) &= 0, \\
\text{and } x(1) + y_1(1) &= 0.
\end{align*}
\]

Since \( g_2 \) has one positive and one negative eigenvalue whenever \( \alpha \) is nonzero, we can expect the limiting solution within \( (0,1) \) to be determined by a cancellation law which retains only one initial condition. Thus, the reduced problem will consist of the limiting system

\[
\begin{align*}
\dot{x}_0 &= 1 - x_0, \\
\dot{y}_{10} &= 0, \\
\alpha^2(x_0)\dot{y}_{10} + \beta(x_0) &= 0,
\end{align*}
\]

and some combination of \( x_0(0) + y_{10}(0) \) and \(-bx_0(0) \) set to zero.

The \( E \) matrix for this example is explicitly determined as

\[
E(x,t) = \frac{1}{\sqrt{1 + \alpha^2(x)}} \begin{bmatrix} 1 & |\alpha(x)| \\ -|\alpha(x)| & 1 \end{bmatrix},
\]

and we can take \( Z_0 \) to be any multiple of \( E'(X_0(0),0) \). Thus, the initial condition (8) appropriate for the reduced system (19) is

\[
|\alpha(x_0(0))|(x_0(0) + y_{10}(0)) - bx_0(0) = 0.
\]

Rewriting this as a function of \( x_0 = x_0(0) \) alone, \( x_0 \) must satisfy a cubic equation with the three roots \( x_0 = 0 \) and \( \frac{1}{4}(b \text{ sgn } \alpha_0 - 6 \pm ((b \text{ sgn } \alpha_0 - 4)^2 + 48)^{1/2} \) selected so that \( \text{ sgn } \alpha_0 = \pm 1 \) according to the sign of \( 1 + 2x_0 \). (For \( b = 2 \), the corresponding initial values are \( x_0 = 0, 0.803, \) and \(-4.29 \).) Values of \( x_0 \) must be avoided for which the corresponding \( \alpha(X_0) = 3 + 2(x_0 - 1)e^{-t} \) has a zero within \( 0 \leq t \leq 1 \).
Since \( X_0(0) + Y_{10}(0) = -bX_0(0)/(1 + 2X_0(0)) \) neither of the separate limiting initial conditions \( A_0(X_0(0)) = 0 \) will be satisfied unless \( bX_0(0) = 0 \). Only then, will the initial boundary layer term \( V_0(t) \) be trivial. Unless \( Y_{10}(1) = -Y_0(1) = -1 - (X_0 - 1)e^{-1} \), the terminal boundary layer term \( V_0(0) \) is also nontrivial.

We are guaranteed that the three "outer" solutions \((X_0,Y_0)\) so obtained provide limiting solutions within \((0,1)\) for corresponding actual solutions in the limit \( \varepsilon \to 0 \) (cf. Hoppensteadt (1971)). Moreover, adding the corresponding boundary layer corrections \( V_0(t/\varepsilon) \) and \( V_0((1-t)/\varepsilon) \) (cf. (13) and (16)) will improve the outer approximation \( Y_0(t) \) in the endpoint boundary layers.

One would expect our approximation (5) to be a good first guess to provide a two-point boundary value solver. This is not necessarily true, however, because of stiffness as \( \varepsilon \to 0 \). Our experience, instead, suggests using our (then crude) asymptotic solution as a first guess for problems where \( \varepsilon \) is only moderately small and then gradually reducing \( \varepsilon \). This scheme allows mesh points to be gradually redistributed in boundary layer regions of rapid transition, and provides good numerical results.

We used the COLSYS code of Ascher et al. (1979), but note that more specialized methods for singular perturbation problems might also be tried (cf. Flaherty and Mathon (1980)). We calculated three distinct solutions to our boundary value problem for \( b = 0, 1, 2, \) and \(-2\) following the procedure: The asymptotic solution and a uniform mesh was used as an initial guess for \( \varepsilon = 10^{-1} \). The \( \varepsilon \)-sequence \( 10^{-1}, 10^{-2}, 10^{-4}, \) and \( 10^{-6} \) was used, with the previously computed solution and mesh used as the initial approximation. If the solution failed to converge or used more than 222 subintervals, \( \varepsilon = 10^{-3} \) and/or \( \varepsilon = 10^{-5} \) values were inserted into the sequence. Calculations used both cubic and quintic splines, and error tolerances of \( 10^{-6} \) for \( x \) and \( Y_1 \) and either \( 10^{-3} \) or \( 10^{-6} \) for \( y_2 \). In general, good accuracy was obtained with both cubic and quintic splines, though calculation time was generally less with the lower order polynomials. Alternative solutions with slightly different initial guesses took much more computing time. For \( b = -2 \) and \( X_0(0) = -2.80 \), we have a turning point since \( a(X_0(0.930)) = 0 \). Actual solutions become unbounded there as \( \varepsilon \to 0 \), though for moderate values of \( \varepsilon \) (like \( 10^{-2} \)) COLSYS could provide a corresponding solution with difficulty. For \( b = 0 \), and \( X_0(0) = -3.5 \), \( a \) has a zero above \( t = 1 \). This forces the boundary layer jump \( \|Y_{10}(0) - Y(1,0)\| \) to be large (89.8 compared to 0.4 and 0.6 for the other roots \( X_0(0) = 0 \) and 0.5), yet we get good results for \( \varepsilon = 10^{-6} \). As a typical example, we display our results (Figures 1, 2, and 3) for
\( b = 2 \) and tabulate our results (Table 1) for cubic splines. The column headings for Table 1 are: (1) \( r \), (2) the number of subintervals used by COLSYS to solve the problem with the given tolerances, (3) and (4) the number of subintervals in the right (left) boundary layer, defined as \( t \leq 10c \) (\( t \geq 1 - 10c \)), (5) the percentage of subintervals in the boundary layers, (6) time to perform the calculation, excluding input/output, and (7) accumulated CPU time on an IBM 3033.

References


18. S. Osher (1980), "Nonlinear singular perturbation problems and one-sided difference schemes," preprint, Mathematics Department, University of California, Los Angeles.


### TABLE

<table>
<thead>
<tr>
<th>$x_0(0) = 0$:</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon$</td>
<td>no.</td>
<td>no. sub.</td>
<td>no. sub.</td>
<td>% sub.</td>
<td>CP time</td>
<td>acc. CP time</td>
</tr>
<tr>
<td></td>
<td>subints.</td>
<td>$t \leq 10\varepsilon$</td>
<td>$t &gt; 1-10\varepsilon$</td>
<td>in BL</td>
<td>(sec.)</td>
<td></td>
</tr>
<tr>
<td>$10^{-1}$</td>
<td>80</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.589</td>
<td>.589†</td>
</tr>
<tr>
<td>$10^{-2}$</td>
<td>80</td>
<td>23</td>
<td>25</td>
<td>60</td>
<td>.650</td>
<td>1.239</td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td>116</td>
<td>14</td>
<td>50</td>
<td>55</td>
<td>1.816</td>
<td>3.055</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>134</td>
<td>15</td>
<td>71</td>
<td>64</td>
<td>1.668</td>
<td>4.723</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$x_0(0) = 0.803$:</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon$</td>
<td>no.</td>
<td>no. sub.</td>
<td>no. sub.</td>
<td>% sub.</td>
<td>CP time</td>
<td>acc. CP time</td>
</tr>
<tr>
<td></td>
<td>subints.</td>
<td>$t \leq 10\varepsilon$</td>
<td>$t &gt; 1-10\varepsilon$</td>
<td>in BL</td>
<td>(sec.)</td>
<td></td>
</tr>
<tr>
<td>$10^{-1}$</td>
<td>80</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.551</td>
<td>.551</td>
</tr>
<tr>
<td>$10^{-2}$</td>
<td>160</td>
<td>78</td>
<td>62</td>
<td>88</td>
<td>1.216</td>
<td>1.767</td>
</tr>
<tr>
<td>$10^{-3}$</td>
<td>128</td>
<td>59</td>
<td>51</td>
<td>86</td>
<td>1.165</td>
<td>2.932*</td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td>128</td>
<td>53</td>
<td>49</td>
<td>80</td>
<td>1.280</td>
<td>4.212</td>
</tr>
<tr>
<td>$10^{-5}$</td>
<td>182</td>
<td>75</td>
<td>67</td>
<td>78</td>
<td>2.152</td>
<td>.6.364*</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>196</td>
<td>83</td>
<td>75</td>
<td>81</td>
<td>2.728</td>
<td>9.092</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$x_0(0) = -4.29$:</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon$</td>
<td>no.</td>
<td>no. sub.</td>
<td>no. sub.</td>
<td>% sub.</td>
<td>CP time</td>
<td>acc. CP time</td>
</tr>
<tr>
<td></td>
<td>subints.</td>
<td>$t \leq 10\varepsilon$</td>
<td>$t &gt; 1-10\varepsilon$</td>
<td>in BL</td>
<td>(sec.)</td>
<td></td>
</tr>
<tr>
<td>$10^{-1}$</td>
<td>160</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.344</td>
<td>1.344</td>
</tr>
<tr>
<td>$10^{-2}$</td>
<td>142</td>
<td>52</td>
<td>47</td>
<td>70</td>
<td>1.932</td>
<td>3.276</td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td>196</td>
<td>69</td>
<td>57</td>
<td>64</td>
<td>4.331</td>
<td>7.607</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>222</td>
<td>87</td>
<td>77</td>
<td>74</td>
<td>8.599</td>
<td>16.206</td>
</tr>
</tbody>
</table>

† When the tolerance on $y_2$ was changed from $10^{-3}$ to $10^{-6}$ the calculation with $\varepsilon = 10^{-1}$ used 222 subintervals and took 2.483 sec.

*Solution jumped to $x_0(0)$ solution without these values of $\varepsilon$. 

9
Figures 1, 2, and 3:

Solution $y_1(t)$ for the third order problem (16)-(17) with $b = 2$ and $x_0(0) = 0$, 0.803, and -4.29, respectively.
Two-point boundary value problems for systems \( \dot{x} = f(x,y,t,\varepsilon) + g_2(x,t,\varepsilon)y \) where the matrix \( g_2(x,t,0) \) has a hyperbolic splitting into a fixed number of stable and unstable eigenvalues. Solutions can then have layer behavior near both endpoints as \( \varepsilon \to 0^+ \). We obtain uniform asymptotic representations of solutions. Our analysis shows how to determine the reduced problem for the limiting interior solution. Orthogonal matrix methods are used to find this reduced problem and appropriate boundary layer correctors in a computationally effective manner. Numerical experiments illustrate the utility of multiple solutions and show how our asymptotic results can be used in practice.