A numerical algorithm for some singularly perturbed boundary value problems

Basem S. Attili*

UAEU, Department of Mathematics and Computer Science, College of Science, P.O. Box 17551, Al-Ain, United Arab Emirates

Received 4 December 2003; received in revised form 29 August 2004

Abstract

A numerical algorithm is proposed to solve singularly perturbed linear two-point value problems. The method starts with a partial decoupling of the system to obtain two independent subsystems, fast and slow components. Each subsystem is then solved separately. A second-order finite difference scheme is used for this purpose. Numerical examples will be presented to show the efficiency of the method.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Singularly perturbed systems; Two point B.V.P’s; Finite differences

1. Introduction

Singularly perturbed boundary value problems are common in applied sciences. They often occur in, for example, optimal control, [24] and convection–diffusion equations. The presence of the perturbation parameter leads to difficulties when numerical techniques are used to solve such problems and convergence will not be uniform. This is since boundary layers appear in these problems, see [17,23–25]. One of the simplest examples of problems of such a type in a one-dimensional case is the problem of stationary diffusion process with a reacting substance that has the form

\[ \varepsilon^2 \frac{d^2}{dt^2} u(t) - c(t)u(t) = f(t), \quad t \in (0, 1) \]

Subject to \( u(t) = \phi(t), \quad t = 0, 1 \)  

(1.1)

*Tel.: +9713 7134286; fax: +973 767 1291.

E-mail address: b.attili@uaeu.ac.ae.
where $c(t) \geq c_0 > 0$ and $\varepsilon$ is a small parameter that characterizes the diffusion coefficient of the substance matter. When $\varepsilon$ tends to zero, boundary layers appear in a neighborhood of the boundary.

If we consider a special case of (1.1) where $c(t) = 1$ and $f(t) = -1$ with some boundary conditions; that is, (1.1) becomes

$$\varepsilon^2 \frac{d^2}{dt^2} u(t) - u(t) = -1, \quad t \in (0, 1)$$

Subject to $u(0) = u(1) = 0$. (1.2)

The exact solution to (2.1) has the form

$$u(t) = u(t; \varepsilon) = 1 - \frac{e^{-t/\varepsilon} + e^{-(1-t)/\varepsilon}}{1 + e^{-1/\varepsilon}}.$$ (1.3)

Note that $0 \leq u(t) \leq 1$ and $\lim_{\varepsilon \to 0} u(1/2) = 1$ and the maximum value of $u(t)$ occurs at $t = 1/2$.

Using the second-order central difference approximation to the second derivative of the form

$$\delta_{tt} z(t) = \frac{[(\delta_t - \delta_T) z(t)]}{h}, \quad \delta_t z(t) = \frac{(z(t + h) - z(t))}{h} \quad \text{and} \quad \delta_T z(t) = \frac{(z(t) - z(t - h))}{h}.$$ (1.4)

It is known, see [14,15], that the error in this formula (in the $l - \infty$ norm) depends on the parameter $\varepsilon$ and the step size $h$; that is

$$|u(t) - z(t)| \leq K(\varepsilon) h^2.$$ (1.5)

Moreover, for sufficiently small values of $\varepsilon$; that is, for $\varepsilon = \varepsilon(h) = 1/h$, the maximum error becomes larger than some positive constant $\bar{K}$, see [28,29]

$$\max_D |u(t) - z(t)| \geq \bar{K} > 0$$ (1.6)

as $h \to 0$. The estimate (1.5) shows that the error tends to zero and hence the scheme is given by

$$\varepsilon^2 \delta_{tt} z(t) - z(t) = -1; \quad z(0) = z(1) = 0,$$

where $\delta_{tt}$ is as given by (1.4) converges as $h \to 0$, while (1.6) shows that this difference scheme does not converge uniformly with respect to the small parameter $\varepsilon$. Even in the case when only the approximate solution is required, finite difference schemes and finite element methods produced unsatisfactory results, see [21,26,27]. It was shown in [29] that the results of using classical methods are also unsatisfactory even when a very fine grid is used. This suggests having numerical methods where the error in the approximate solution tends to zero as $h \to 0$ ($N \to \infty$) independently of the parameter $\varepsilon$; that is, uniform convergence is desired. Some special schemes can be found in for example Il’in [15] where...
special finite difference methods are used. Fitted schemes that allow the use of the meshes with an arbitrary distribution of nodes can be found in [1,2,6,11,15]. Problems that arise in viscoelastic flow were considered by Allen and Southwell [2], Attili [3], Davies et al. [5] and Karageorghis et al. [16]. Specially fitted schemes were used by Allen and Southwell [2] and some kind of regularization of the singularity was suggested by others before applying special schemes of boundary value solvers.

In this work, we will consider a more general problem rather than (1.1), which was used as an illustration. The approach we are going to follow starts with decoupling of the system (1.1) followed by applying a difference scheme of order 2. This will be done in Section 3. Some preliminaries will be given in Section 2. Also, in Section 3, we will describe the algorithm used to solve the resulting system. Finally, numerical examples in comparison with their exact solutions will be presented in Section 4.

2. Preliminaries

We will consider a general framework which is a system of first-order differential equations that has the form

\[
\frac{dx_1}{dt} = A_{11}(t)x_1 + A_{12}(t)x_2 + F_1(t), \\
\frac{\epsilon dx_2}{dt} = A_{21}(t)x_1 + A_{22}(t)x_2 + F_2(t),
\]

(2.1)

where \(0 < t < 1\), and subject to the linear boundary conditions of the form

\[L(x_1(0), x_1(1), x_2(0), x_2(1)) = 0.\]

Here \(x_i \in \mathbb{R}^{n_i}, i = 1, 2, \epsilon > 0\) is a small parameter, \(A_{ij}(t)\) are matrices with appropriate dimension and \(F_1(t), F_2(t)\) are given real vector functions of appropriate dimension. It is assumed that the functions \(A_{ij}(t), F_1(t)\) and \(F_2(t)\) together with \(L\) are such that the system given by (2.1) has a unique bounded solution. This system can be considered as a control stochastic linear system where \(F_1\) and \(F_2\) represent the control.

The analytical aspects of systems like (2.1) through asymptotic analysis are well established, see [17,23–25]. Studies conclude that the solution has two states; namely, one varies slowly and the other rapidly with respect to \(t\). Also, the presence of the small parameter \(\epsilon\) leads to what is called boundary layer phenomenon where the solution does not converge uniformly at end points. Similar behavior occurs when such nonuniformity occurs at an interior point called interior layer phenomenon, see [9,10,13,18,19].

As a result of such boundary layer behavior and the fast–slow states of the solutions, the two point boundary value problem given by (2.1) cannot be solved accurately using typical and usual numerical techniques. Different suggestions and approaches were used to overcome this problem. For example, Mattheij [22] considered partial decoupling to (2.1) before applying a difference scheme. Fitted methods on arbitrary distribution of nodes were considered by Il’in [15] while adaptive methods were considered in a way such that the nodes are distributed to achieve parameter uniform convergence.
3. Decoupling of the system

As mentioned before, the solution of the system given by (2.1) comprises of two states one varying slowly and the other rapidly with respect to the independent variable \( t \). Due to the nonuniform behavior and the slow–fast property of the solution, partial decoupling is needed. The purpose of such a decoupling is to transform the system in order to obtain two independent subsystems, one for the slow and the other for the fast state.

Before presenting the partial decoupling and for what follows, \( |v| \) denotes the supremum norm for vector \( v \), while for a matrix \( |G| \) denotes the vector induced supremum norm and \( \|G\| = \sup_{t \in [0,1]} |G(t)| \), where \( G(t) \) is a vector or a matrix function.

For our purpose and to obtain the decoupling, we will introduce the variable \( x_3 \) (new slow variable) given by
\[
x_3 = x_1 - \varepsilon N x_2.
\] (3.1)

If \( A_{22} \) is an invertible matrix, we may associate a system of lower dimension of the form
\[
x_3' = (A_{11} - N A_{21}) x_3 + (F_1 - N F_2),
\] (3.2)
where
\[
N(t) = N_0(t) + \varepsilon N_1(t) + O(\varepsilon^2)
\] (3.3)
with \( N_0 = A_{12} A_{22}^{-1} \) and \( N_1 = ((A_{11} - N_0 A_{21}) N_0 - \dot{N}_0) A_{22}^{-1} \) continuously differentiable matrix functions with the dot represents differentiation, see [9]. As a result, we have the following theorem.

**Theorem 1.** Assume that the matrix functions \( A_{ij}; i, j = 1, 2 \) are continuously differentiable such that \( \|A_{ij}(t)\| \leq b_{ij} \) and \( \|\dot{A}_{ij}(t)\| \leq c_{ij} \) where \( b_{ij} \) and \( c_{ij} \) are positive constants. If any eigenvalue \( \lambda(t) \) of \( A_{22}(t) \) satisfies the conditional stability conditions \( \text{Re}(\lambda(t)) \leq -c_3 \) or \( \text{Re}(\lambda(t)) \geq c_4 \) with \( c_3 \) and \( c_4 \) positive constants. Then there exists an \( \varepsilon_0 > 0 \) and conditionally differentiable matrix functions \( N(t) \), \( N_0(t) \) and \( N_1(t) \) given by (3.3) such that for \( 0 < \varepsilon < \varepsilon_0 \), the system (2.1) is transformed to
\[
x_3' = (A_{11} - N A_{21}) x_3 + (F_1 - N F_2),
\]
\[
\varepsilon x_2' = A_{21} x_3 + (\varepsilon A_{21} N + A_{22}) x_2 + F_2.
\] (3.4)

Similar results can be found in [30] and a more advanced and detailed one can be found in [9]. For the proof and details, the reader can refer to these references.

Now the system given by (3.4) is essential for the numerical treatment of the problem given by (2.1). The first part of (3.4) can be solved for \( x_3 \) using any two point boundary value problem solver provided that \( n_1 \) appropriate linearly independent boundary conditions are available. Once \( x_3 \) is computed, the second part of (3.4) can be solved for \( x_2 \). Note that this part can be written in the form
\[
\varepsilon x_2' = F(t) x_2 + g(t), \quad 0 < t < 1,
\] (3.5)
where \( F(t) = (\varepsilon A_{21} N + A_{22}) \) and \( g(t) = A_{21} x_3 + F_2 \). For the decoupling of the boundary conditions, note that (3.5) can be written in the form
\[
\varepsilon \frac{d(N^{-1} x_2)}{dt} = (D - \varepsilon N^{-1} \dot{N})(N^{-1} x_2) + N^{-1} g
\] (3.6)
with
\[ D = N^{-1}(N^{-1}(A_{21}N + A_{22})N - \varepsilon N), \]
where \( D = \text{diag}(D_1, D_2) \). With \( G = N^{-1}g = N^{-1}(A_{21}x_3 + F_2) \), we can split (3.6) into
\[ \varepsilon \dot{u}_1 = D_1 u_1 + G_1, \]
\[ \varepsilon \dot{u}_2 = D_2 u_2 + G_2, \]
(3.7)
where \( x_2 = (u_1, u_2) \). Leading to the decoupling of the boundary conditions into
\[ L(z(0), z(1), u_1(0), u_1(1), u_2(0), u_2(1)) = 0. \]

It should be noted that one can do without separating the boundary conditions. Instead of such separation, we can use the uniform asymptotic expansion of the system given in (2.1). Then numerically compute
the dominant part of the fundamental solution matrix, which in turn can be introduced in the boundary conditions too obtain the initial value of the solution of the given boundary value problem. Details of
such uniform expansions of systems of the form given by (2.1) can be found in [7,8].

We will treat (3.5) numerically using an appropriate difference scheme that converges uniformly to the solution \( x_2 \) assuming there are \( n_2 \) linearly independent boundary conditions. For that purpose, let \( h_j \) be a
grid defined on \([0, 1]\) with \( h = \max_j h_j \). Here \( t_0 = 0, t_N = 1 \) and \( t_k = kh; k = 0, 1, \ldots, N \). Let \( z_k = z(t_k) \) be the numerical approximation to the solution at the grid points and similarly
\( F_k = F(t_k), g_k = g(t_k) \) and \( x_{2k} = x_2(t_k) \). The error in the approximation will be given by \( e_k = z_k - x_{2k} \). With this notation, we
will have Euler’s forward implicit scheme given by
\[ \varepsilon \left( \frac{z_{k+1} - z_k}{h_k} \right) = F_{k+1}z_{k+1} + g_{k+1} \]
(3.8)
and Euler’s backward explicit scheme is given by
\[ \left( F_k + \frac{\varepsilon}{h_k} I \right) z_k = \frac{\varepsilon z_{k+1}}{h_k} - g_k, \]
(3.9)
where \( I \) is the identity matrix with appropriate order. These schemes given by (3.8) and (3.9) can be shown
to have a quadratic rate of convergence as stated in the following result.

**Theorem 2.** Assume \( F(t) \) and \( g(t) \) are twice continuously differentiable. Assume also that \( \| F(t) \|, \|
F^{-1}(t) \| \) and \( \| g(t) \| \) are bounded and the eigenvalues of \( F(t) \) are such that \( \text{Re}(\lambda(t)) \leq c_3 \) or \( \text{Re}(\lambda(t)) \geq c_4 \) with \( c_3 \) and \( c_4 \) positive constants. Then if \( h_k \) is chosen such that
\[ \min_k h_k > \max \left( \frac{\varepsilon}{\| F(t) \|}, \frac{2\varepsilon \| F(t) \|}{\| g(t) \|} \right), \]
(3.10)
then the schemes given by (3.8) and (3.9) solves (3.5) with \( z_k \) bounded and the error \( e_k \) is \( O(h^2) \) if
\[ h \| F(t) \| \gg 1. \]
(3.11)

The proof of the theorem can be obtained from a result given by Kreiss et al. [20] and so it is omitted. Note that there is an assumption on the minimum of the grid length given by (3.10) although in general one
requires $h < \varepsilon$ as a precondition on the grid to make any standard method accurate. This latter assumption will produce large round-off errors and that is why (3.10) is needed. Also that for the proof, condition (3.10) is a strong one and a weaker assumption can do the job; that is, instead of (3.10), the proof follows with the assumption
\[
\frac{2k}{h_k} < \|F_{k+1}^{-1}\| < \frac{h_k}{\varepsilon}.
\] (3.12)

**Remark 1.** If (3.11) is not satisfied, the error in the methods will be $O(h)$. In this case the differentiability conditions on $F(t)$ and $g(t)$ can be relaxed.

It is well known that when the eigenvalues of $F(t)$ are all negative (positive) then the forward (backward) integration is stable, see [4]. This suggests the use of some transformation on (3.5) in order for $F(t)$ to be block-diagonalized; that is, we require the system (3.5) to be transformed through the use of $N(t)$ (whose existence and uniqueness is guaranteed, see [12]) to
\[
\varepsilon \frac{d(N^{-1}x_2)}{dt} = (D - \varepsilon N^{-1} \dot{N})(N^{-1}x_2) + N^{-1}g,
\] (3.13)
where $D(t) = N^{-1}F(t)N(t) = \text{diag}(D_1, D_2)$ with $\text{Re}(\lambda(D_1(t))) < 0$ and $\text{Re}(\lambda(D_2)) > 0$.

Note that the stability of the system given by (3.13) will not change since the conditional stability of $F(t)$ implies that of $D$ and $(D - \varepsilon N^{-1} \dot{N})$, see [18, Chapter 5]. The same can be said about $A_{22}$ and $\varepsilon A_{21}N + A_{22}$. The system matrix $F(t)$ has a special block structure. It is important for it to remain so throughout the whole interval. If such a structure changes, one must find subintervals $(a_i, a_{i+1})$; $i = 0, 1, \ldots, n - 1$ on which the block structure does not alter. This means different equations of the form given by (3.13) needs to be solved on each subinterval and $a_i$ and $a_{i+1}$; $i = 0, 1, \ldots, k$ has to be among the grid points. These details will be combined in the following algorithm:

**Algorithm 3.**

**Step I: Compute**

$N_0(t) = A_{12}A_{22}^{-1}$,

$N_1(t) = ((A_{11} - N_0A_{21})N_0 - \dot{N}_0)A_{22}^{-1}$.

**Step II: Set**

$N(t) = N_0(t) + \varepsilon N_1(t)$.

**Step III: Form the transformed system (3.4).**

**Step IV: For the first part of (3.4)**

(i) Determine the partition points $a_0, a_1, \ldots, a_{n-1}$.

(ii) On each subinterval, use the QR-algorithm to compute the block-diagonalizing transformation $N_0, N_1, \ldots, N_{n-1}$.

**Step V: Compute**

(i) $D_i = N_i^{-1}((A_{22} + \varepsilon A_{21}N)N_i - \varepsilon \dot{N}_i)$, where $D_i = \text{diag}(D_{1i}, D_{2i})$ with $\text{Re}(\lambda(D_{1i}(t))) < 0$ and $\text{Re}(\lambda(D_{2i}(t))) > 0$.

(ii) $G_i = (G_{1i}^T, G_{2i}^T)^T = N_i^{-1}(A_{2i}^z + g_2)$.

(iii) $u_i = (u_{1i}^T, u_{2i}^T)^T = N_i^{-1}u$. 

Step VI: On each subinterval split the second part of (3.4) into
\[ \dot{u}_{1i} = D_{1i} u_{1i} + G_{1i}, \]
\[ \dot{u}_{2i} = D_{2i} u_{2i} + G_{2i}. \]

(3.14)

Step VII: Find the grid points on each subinterval by monitoring (3.11) and (3.12)
\[ \frac{2\varepsilon}{h_{ij}} < |D_i^{-1}(t_{i,j+1})| < \frac{h_{ij}}{\varepsilon} \quad \text{and} \quad h_{ij}|D_i(t_{i,j+1})| > 1, \]
where \( t_{i,j+1} \) refers to the \((j + 1)\)st grid point on the \(i\)th subinterval.

Step VIII: Combine the discretized form of both (3.13) and the first part of (3.4) and solve using
\[ \varepsilon \left( \frac{w_{i,j+1} - w_{i,j}}{h_i} \right) = K_1 K_2 [w_{i,j+1} - w_{i,j}] + K_1 [H_{i,j+1} - H_{i,j}], \]
where \( w = [z, y_1, y_2] \) the \(j\)th numerical approximation at the \(i\)th step and
\[ K_1 = \begin{bmatrix} -\frac{1}{\varepsilon} I_m & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I_{n_2} \end{bmatrix} \quad \text{with} \quad n_1 + n_2 = n, \]
\[ K_2 = \begin{bmatrix} A_{11} - NA_{21} & 0 & 0 \\ 0 & D_{1i} & 0 \\ 0 & 0 & D_{2i} \end{bmatrix} \]
subject to the boundary conditions
\[ L(z(0), z(1), y_{1i}(0), y_{1i}(1), y_{2i}(0), y_{2i}(1)) = 0. \]

4. Numerical results

In this section, we will illustrate the techniques developed in the previous sections by some numerical examples. Two examples from O’Mally [24] will be considered. One of the examples exhibits boundary layers at the end points and the other at an interior point.

Example 1. Consider the second-order linear equation
\[ \varepsilon \ddot{x} - 2t \dot{x} = 0, \quad -1 < t < 1, \]
\[ x(-1) = -1, \quad x(1) = 2. \]

(4.1)
The solution for \(-1 < t < 1\) is given by
\[ x(t) = 2 + \frac{1}{t} e^{\frac{t-1}{\varepsilon}} + O(1) \]
and graphically for \( \varepsilon = 0.01 \) by Fig. 1.

Transform (4.1) into a first-order system by letting \( \dot{x} = x_1 \) and \( \ddot{x} = \dot{x}_1 \) to obtain
\[ \dot{x} = x_1, \]
\[ \varepsilon \dot{x}_1 = 2tx_1. \]

(4.2)
subject to same boundary conditions. According to the notation of (2.1) we have $A_{11} = 0$, $A_{12} = 1$, $f_1(t) = 0$, $A_{21} = 0$, $A_{22} = 2t$, $f_2(t) = 0$. As a result the matrices used in (3.4) will be $N_0 = A_{12}A_{22}^{-1} = 1/t$, $N_1 = (A_{11} - N_0A_{21})N_0 - N_0A_{22}^{-1} = -1/4t^3$.

The choice of stretching points for Example 1 as well as Example 2 are given in Table 1.
The numerical solution for $\varepsilon = 0.01$ and $\varepsilon = 0.0001$ are given in Figs. 2 and 3, respectively. When violating condition 3.10 resulted in the solution given in Fig. 4.
Example 2. The second example has the form

\[ \varepsilon \ddot{x} + 2t \dot{x} - 2x = 0, \quad -1 < t < 1, \]
\[ x(-1) = -1, \quad x(1) = 2. \]  \hspace{1cm} (4.3)

Rewriting (4.3) as a first-order system leads to

\[ \dot{x} = x_1; \quad \varepsilon \dot{x}_1 = -2tx_1 + 2x. \]  \hspace{1cm} (4.4)

subject to same boundary conditions as in Example 1. Again the stretching points are given in Table 1 and the solution for \( \varepsilon = 0.01 \) and \( \varepsilon = 0.0001 \) are given in Figs. 5 and 6, respectively.

References


