COLLOCATION FOR SINGULAR PERTURBATION PROBLEMS III: NONLINEAR PROBLEMS WITHOUT TURNING POINTS

by

U. Ascher* and R. Weiss**

Technical Report 82-9

October 1982

Abstract

A class of nonlinear singularly perturbed boundary value problems is considered, with restrictions which allow only well-posed problems with possible boundary layers, but no turning points. For the numerical solution of these problems, a close look is taken at a class of general purpose, symmetric finite difference schemes arising from collocation.

It is shown that if locally refined meshes, whose practical construction is discussed, are employed, then high order uniform convergence of the numerical solution is obtained. Nontrivial examples are used to demonstrate that highly accurate solutions to problems with extremely thin boundary layers can be obtained in this way at a very reasonable cost.

Keywords: Collocation, singular perturbation, nonlinear problems, symmetric difference schemes

* Department of Computer Science, Univ. of B.C. Vancouver, Canada V6T 1W5. The research of this author was supported in part under NSERC grant A4306.

** Institut fur Angewandte und Numerische Mathematik, Technische Universitat Wien, 1040 Wien, Gusshausstrasse 27-29, Austria.

1. Introduction

The numerical solution of nonlinear singular perturbation problems presents some major computational difficulties. At the same time, such problems are abundant in applications, e.g. in semiconductor theory, diffusion-convection processes with a dominant convection term, fluid dynamic problems with large Reynolds numbers, etc.

The basic computational difficulty arising can be roughly described as follows. Suppose that a grid with maximum spacing h is used to discretize the differential problem. While normally h can be assumed to be small compared to the differential problem parameters, in a singular perturbation problem the parameter ε, which multiplies some of the highest derivatives appearing in the problem formulation, is so small that for practical reasons we must consider the case

$$h \geq \varepsilon, \text{ or even } h >> \varepsilon.$$

Furthermore, from the solution approximation point of view, one has to deal with transition layers, i.e. regions where the solution profile varies rapidly, with gradients proportional to some negative power of ε.

Consider boundary value problems of this type for ordinary differential equations. To recall, there are two classes of general purpose methods for boundary value ODEs. (See, e.g., Keller [13]). The first is that of initial value techniques like multiple shooting. Such techniques fail to perform adequately for singular perturbation problems, requiring $h = O(\varepsilon)$, essentially for the same reason that causes grief when simple shooting is applied to moderately "stiff" problems.

The other class of general purpose methods is that of centered difference (or collocation) schemes. Such schemes offer more hope for singular perturbation problems, requiring small mesh spacing (of size comparable to ε) only in layer regions, not everywhere. However, if the mesh is not fine enough in a layer region then the numerical solution is polluted everywhere by oscillatory error components.

In recent years, a large number of special purpose methods have been proposed which do not require an accurate representation by the mesh of layer regions, while still producing accurate numerical solutions outside the layers. The upwinded Euler scheme is a popular example of such methods, where the error generated inside the layer regions is quickly damped outside.

The approach highlighted in this paper is that of centered, general purpose difference schemes with local mesh refinement in layer regions. Since this is currently less popular than special, (explicitly or implicitly) upwinded schemes, we now discuss the relative merits of these approaches.

Firstly, as has been noted by others as well, while centered schemes tend to produce wiggly numerical noise if the mesh is inadequate, special one-sided schemes tend to have too much "artificial viscosity", i.e. to be inaccurate (typically only first order in h) and to smear a layer information over a number of neighboring mesh elements. A smooth solution curve can actually be considered worse than one containing numerical ripples if it is wrong, because its form is more deceptive, cf. Gresho and Lee [11].

A natural idea here is to obtain a first, relatively inaccurate, solution by a special purpose method and then to switch to a centered, more accurate, general purpose method with mesh points distributed according to the obtained first solution profile. However, the general implementation of such a switch is far from being trivial, and a simpler and probably more robust technique is to do continuation in ε, using the centered method all along (i.e., solve a sequence of problems with decreasing ε, the first with ε=h, say, and the last with the desired value of ε, gradually upgrading the mesh).

It should be noted that higher order special purpose methods of Runge-Kutta type exist as well (Ringhofer [18], Ascher and Weiss [3]). These methods do share the drawbacks of special purpose schemes mentioned below.

One-sided special purpose schemes generally require upwinding (explicit or implicit). Thus, unless the problem is already in a form where growing and decaying fundamental matrix components are separated, a costly transformation is required to bring it to such form, see Kreiss and Kreiss [15].

As a third argument, we feel that with the above special purpose schemes there is more opportunity for the phenomenon where the difference problem has spurious solutions to crop up. These difference solutions do not correspond to any solution of the differential problem (see, e.g., Beyn and Doedel [5]). This is so because the difference operator is not exactly modeled after the differential operator any more. Of course, if a centered scheme is used with an inadequate mesh (say, uniform) then spurious solutions may easily result as well; see, e.g., Kellog, Shubin and Stephens [14]. This, however, is less important for practical purposes, because such a numerical procedure is nonsensical anyway.

Finally, note that a nonlinear singular perturbation problem where the location of a transition layer is not precisely known can be far more difficult to solve numerically than one with only known layer locations. At least in the latter case we can always flood the transition layer region with sufficiently many mesh points so as to remove the singular perturbation effect and obtain a solution, even if not in the most cost-effective way. Now, in case that the layer location depends on the solution and is not exactly known, the upwinding of a special scheme may be done in the wrong direction. A backward Euler scheme then becomes a Forward Euler scheme, which is not even

A-stable. This may result in difficulties in the convergence of the nonlinear iteration. A centered scheme as discussed in this paper is A-stable in both directions of integration and hence less prone to disastrous results in such circumstances. Continuation in ε can again be used in principle with centered schemes to methodically refine the mesh appropriately. For an example, see Wan and Ascher [21]. Still, the success of such a process may well depend on a (perhaps vague) a-priori knowledge of a desirable initial solution profile.

The purpose of this paper is to consider the computational implementation and performance of a class of symmetric, or centered, collocation schemes which include the most familiar finite difference schemes as special cases. We consider the application of these schemes to a general, but restricted, class of nonlinear singularly perturbed problems which are well-posed and allow for boundary layers only. This class of problemsis relatively well understood analytically. The analytical knowledge allows us to take a close look at the numerical schemes, and we believe that this is an essential step towards understanding the performance of these schemes on wider classes of problems. Computational experience with such schemes crudely applied to a number of various problems has already been reported (e.g. Hemker et al. [12], Ascher [1], Wan and Ascher [21]).

In part I [3] and part II [4] of this work (hereinafter referred to as "Part I" and "Part II", respectively) we have considered symmetric collocation schemes for the numerical solution of linear singularly perturbed problems. In particular, Lobatto and Gauss collocation points have been considered. The simplest instances of these methods are the well-known trapezoidal and midpoint difference schemes. The ideas in these papers have been put into use in Spudich and Ascher [20].

We have shown that these symmetric schemes produce highly accurate numerical solutions at a very reasonable cost, provided that appropriately fine meshes are used near the boundaries, where the analytic solution may have steep boundary layers.

Here we extend these results to nonlinear problems, where Newton's method of quasilinearization is used and the resulting linearized boundary value problems are solved using the collocation implementation discussed in Part II. We demonstrate the potential of these schemes on three examples which appear in the literature.

It turns out that the convergence results as well as the mesh construction in Part II extend, with slight modifications, to the nonlinear case. However, the extension is not trivial. The differences between the linear and nonlinear problems are highlighted in the next section, which prepares the analytic preliminaries.

In section 3 we describe the numerical schemes used, state the convergence results and outline their proofs. Practical mesh construction is discussed in section 4.

In section 5 we discuss in detail our numerical experience with three examples. The numerical schemes are shown to produce highly accurate solutions to problems with extremely thin boundary layers, at a very reasonable cost.

2. Analytic preliminaries

Consider the problem of order n+m for $x(t,\varepsilon) = (y(t,\varepsilon), z(t,\varepsilon))$,

$$\begin{aligned} &(2.1) \qquad \varepsilon y' = f(t,y,z,\varepsilon) \\ &(2.2) \qquad z' = g(t,y,z,\varepsilon) \end{aligned} \qquad 0 \le t \le 1$$

$$(2.3) \qquad b(x(0); x(1); \varepsilon) = 0.$$

Here $\varepsilon > 0$ is a small parameter, y and f have n components, z and g have m, and b is a boundary vector function of size n+m. The nonlinear functions f, g and b have asymptotic expansions in ε, with the coefficients being smooth functions of the other variables.

We assume that the Jacobian matrix $f_y(t,y,z,o)$ of the "fast" solution components has a regular splitting with $n_- \geq 0$ (strictly) stable and $n_+ := n-n_- \geq 0$ (strictly) unstable eigenvalues, for $0 \le t \le 1$ and (y,z) in an appropriate domain. It is natural then to look for a solution $x^*(t,\varepsilon) = (y^*(t,\varepsilon), z^*(t,\varepsilon))$ which has the representation

$$\begin{aligned} &(2.4) \qquad y^*(t,\varepsilon) = \bar{y}(t) + \mu(\tau) + \nu(\sigma) + 0(\varepsilon) \\ &(2.5) \qquad z^*(t,\varepsilon) = \bar{z}(t) + 0(\varepsilon). \end{aligned} \qquad 0 \le t \le 1$$

Here

$$(2.6) \qquad \tau = t/\varepsilon \quad , \qquad \sigma = (t-1)/\varepsilon,$$

$\bar{y}(t)$ and $\bar{z}(t)$ are solutions of the reduced equations

$$\begin{aligned} &(2.7) \qquad 0 = f(t,\bar{y},\bar{z},0) \\ &(2.8) \qquad \bar{z}' = g(t,\bar{y},\bar{z},0) \end{aligned} \qquad 0 \le t \le 1$$

subject to m appropriate boundary conditions, and μ and ν are left end and right end layer correction functions. They satisfy

$$(2.9) \qquad \frac{d}{d\tau}\mu = f(0,y(0) + \mu(\tau), z(0), 0) \qquad 0 \le \tau < \infty$$

$$(2.10) \qquad \frac{d}{d\sigma}\nu = f(1,y(1) + \nu(\sigma), z(1), 0) \qquad -\infty < \sigma \le 0$$

and μ and ν decay exponentially to 0 as $\tau \to \infty$, $\sigma \to -\infty$, respectively. Equations

(2.7) - (2.10) arise from the representation (2.4), (2.5) by equating O(1) terms with respect to ε in (2.1), (2.2).

Now, to construct the solution as in (2.4), (2.5), we substitute into the boundary conditions to obtain

$$b((\bar{y}(0) + \mu(0), \bar{z}(0)); (\bar{y}(1) + \nu(0), \bar{z}(1)); 0) = 0. \tag{2.11}$$

The requirement that μ and ν decay exponentially implies that $\mu(0)$ and $\nu(0)$ must be on the stable manifolds of their corresponding equations and we write these equations as

$$\begin{aligned} &\phi_- (\bar{y}(0), \bar{z}(0), \mu(0)) = 0 && (n_+ \text{ eqns}) \\ &\phi_+ (\bar{y}(1), \bar{z}(1), \nu(0)) = 0 && (n_- \text{ eqns}) . \end{aligned} \tag{2.12}$$

Thus in (2.11), (2.12) we have 2n+m constraints for the 4n+2m unknowns $\bar{x}(0)$, $\bar{x}(1)$, $\mu(0)$, $\nu(0)$.

Eliminating $\mu(0)$ and $\nu(0)$ from (2.11), (2.12) (in principle) leaves a set of m equations to be satisfied by $\bar{y}(0)$, $\bar{z}(0)$, $\bar{y}(1)$ and $\bar{z}(1)$ alone, and these are the boundary conditions for the reduced equations (2.7), (2.8) (cf. Episova [8], O'Malley [17]). Flaherty and O'Malley [9] construct the reduced boundary conditions numerically in case that (2.1) and (2.3) are linear in y, which implies that (2.11) and (2.12) are linear in y, μ and ν.

Note that everything is much simpler when f is linear in y. Not only can the manifolds (2.12) be explicitly found but also, and more importantly, (2.9) and (2.10) imply that μ and ν are simply decaying exponential functions. In the more general case, (2.9) and (2.10) are just general systems of ODEs.

The reduced differential equations (2.7), (2.8) plus the reduced boundary conditions form the reduced problem whose solution(s) $\bar{x}(t) = (\bar{y}(t), \bar{z}(t))$ is referred to as a reduced solution. Different reduced solutions yield different solutions to our problem (2.1) - (2.3) provided they can serve in the ansatz

(2.4), (2.5). To enable this we assume that the Jacobian matrix f_y at a reduced solution has a hyperbolic splitting for all $0 \le t \le 1$.

For simplicity we also assume that at the reduced solution, $f_y(t, \bar{y}(t), \bar{z}(t), 0)$ is nondefective. Thus, there is a nonsingular (smooth) matrix function $E(t)$ s.t.

$$E^{-1}(t)\, f_y(t, \bar{y}(t), \bar{z}(t), 0)E(t) = \Lambda(t) \equiv \mathrm{diag}\{\lambda_1(t),\dots,\lambda_n(t)\} \tag{2.13}$$

and

$$\mathrm{re}(\lambda_j(t)) \begin{cases} < 0 & j=1,\dots,n_- \\ > 0 & j=n_-+1,\dots,n \end{cases} \qquad 0 \le t \le 1 \tag{2.14}$$

Consider now a linearization of our problem (2.1) - (2.3) about an appropriate function $\hat{x} = (\hat{y}, \hat{z})$, which we write in operator form as

$$L[\hat{x}]x = s[\hat{x}] \tag{2.15}$$

In detail, (2.15) is written as

$$\varepsilon y' - A_{11}y - A_{12}z = s_1 \tag{2.16}$$

$$z' - A_{21}y - A_{22}z = s_2 \qquad 0 \le t \le 1 \tag{2.17}$$

$$B_0x(0) + B_1x(1) = \beta \tag{2.18}$$

where

$$\begin{aligned} &A_{11}(t, \varepsilon) = f_y(t, \hat{y}, \hat{z}, \varepsilon),\ A_{12}(t, \varepsilon) = f_z(t, \hat{y}, \hat{z}, \varepsilon) \\ &A_{21}(t, \varepsilon) = g_y(t, \hat{y}, \hat{z}, \varepsilon),\ A_{22}(t, \varepsilon) = g_z(t, \hat{y}, \hat{z}, \varepsilon) \\ &B_0 = \frac{\partial b(x_1, x_2)}{\partial x_1},\ \ B_1 = \frac{\partial b(x_1, x_2)}{\partial x_2}\ \text{ at } x_1 = \hat{x}(0),\ \ x_2 = \hat{x}(1). \end{aligned} \tag{2.19}$$

We assume that the linear operator $L[x^*]$ has a bounded inverse, independent of ε, for $0 < \varepsilon \le \varepsilon_0$. Lipschitz continuity of L (which follows, e.g., if f, g, b are twice differentiable with respect to the dependent variables) then implies a similar bound for $L[\hat{x}]$ at points $\hat{x}$ near x^* and, in particular, at the constructed solution of (2.4), (2.5) for ε small enough.

Our problem (2.1) - (2.3) can be written in the form (2.16) - (2.19) with $\hat{x} = x^*$, simply by defining $s[x^*]$ appropriately, e.g.

$$s_1(t, \varepsilon) = f(t, y^*, z^*, \varepsilon) - A_{11}y^* - A_{12}z^*$$

etc. The problem then looks like the one considered previously in Part II, but there is a difference: Here, the matrices A_{ij} and inhomogeneous terms s_i, $1 \le i, j \le 2$, are not slowly varying near the interval ends, since they contain the boundary layer effects of y^*. Thus, while on the "long" interval $O(\varepsilon) < t < 1 - O(\varepsilon)$ away from the layers not much change is expected, in the boundary layer regions a richer solution behaviour is now allowed, compared to the usual linear, variable coefficient case, as described above in connection to (2.9) and (2.10).

In Part I and Part II a layer mesh was constructed which took advantage of the known exponential decay of $\mu(\tau)$. In the nonlinear case, then, this mesh construction is less clear. Fortunately, by the exponential decay of μ and ν, the mesh construction can be applied for values of t which correspond to sufficiently large values of τ and $-\sigma$, if we know enough about the reduced solution $\bar{x}(0)$, $\bar{x}(1)$. The practical construction of a mesh in the layer regions near the boundaries is discussed in sections 4 and 5.

3. Numerical schemes and their convergence

To solve the problem (2.1) - (2.3) numerically we use k-stage, c^0-collocation as described in §3 of Part I and in §3 of Part II. The same notation is adhered to here. Thus, on a given mesh

$$\Delta: 0 = t_1 < t_2 < \ldots < t_N < t_{N+1} = 1 \tag{3.1}$$
$$h_i := t_{i+1} - t_i, \quad 1 \le i \le N, \quad h := \max_{1 \le i \le N} h_i$$

the solution $x_\Delta(t) = (y_\Delta(t), z_\Delta(t))$ is a continuous piecewise polynomial vector function which satisfies the boundary conditions (2.3) and the differential equations (2.1), (2.2) at the collocation points

$$t_{ij} := t_i + h_i \rho_j \qquad 1 \le i \le N, \ i \le j \le k. \tag{3.2}$$

The points $0 \le \rho_1 < \ldots < \rho_k \le 1$ are chosen to be the Gauss or Lobatto points. This gives the difference scheme

$$\varepsilon h_i^{-1}(y_{ij} - y_i) = \sum_{\ell=1}^{k} \hat{a}_{j\ell} f(t_{i\ell}, y_{i\ell}, z_{i\ell}, \varepsilon) \tag{3.3}$$

$$1 \le i \le N, \ 2-r \le j \le k+r$$

$$h_i^{-1}(z_{ij} - z_i) = \sum_{\ell=1}^{k} \hat{a}_{j\ell}\, g(t_{i\ell}, y_{i\ell}, z_{i\ell}, \varepsilon) \tag{3.4}$$

$$b(x_1; x_{N+1}; \varepsilon) = 0 \tag{3.5}$$

for $x_i = x_\Delta(t_i)$, $x_{ij} = x_\Delta(t_{ij})$, where $\hat{a}_{j\ell}$ are known constants and $x_{i,k+1} = x_{i+1}$. For Gauss points, $r=1$, $\rho_1 > 0$, $\rho_k < 1$ (so mesh points are not collocation points) and $\hat{a}_{k+1,\ell} = \hat{b}_\ell$, $\ell=1,\ldots,k$. The simplest of these schemes, with $k=1$, is the midpoint rule

$$\varepsilon h_i^{-1}(y_{i+1} - y_i) = f(t_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}, z_{i+\frac{1}{2}}, \varepsilon) \tag{3.6}$$
$$h_i^{-1}(z_{i+1} - z_i) = g(t_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}, z_{i+\frac{1}{2}}, \varepsilon)$$

where $t_{i+\frac{1}{2}} = t_i + \frac{1}{2}h_i$, $y_{i+\frac{1}{2}} = \frac{1}{2}(y_i + y_{i+1})$ and $z_{i+\frac{1}{2}} = \frac{1}{2}(z_i + z_{i+1})$.

For Lobatto points, $r=0$, $\rho_1=0$ and $\rho_k=1$. Thus the mesh points t_i are collocation points. The simplest of these schemes, with $k=2$, is the trapezoidal rule

$$\varepsilon h_i^{-1}(y_{i+1} - y_i) = \tfrac{1}{2}(f(t_i, y_i, z_i, \varepsilon) + f(t_{i+1}, y_{i+1}, z_{i+1}, \varepsilon))$$

(3.7)

$$h_i^{-1}(z_{i+1} - z_i) = \tfrac{1}{2}(g(t_i, y_i, z_i, \varepsilon) + g(t_{i+1}, y_{i+1}, z_{i+1}, \varepsilon)).$$

Denote by ψ^c the restriction of a function $\psi(t)$ to $\Delta \cup \{t_{ij}; 1 \le i \le N, 1 \le j \le k\}$. Equations (3.3) - (3.5) form a nonlinear algebraic system for x_Δ^c, which we attempt to solve by Newton's method of quasilinearization. Equivalently, and more naturally for implementation, the quasilinearization can be done before discretization. Thus, given an initial guess $x_\Delta^0(t)$, a sequence of iterates $x_\Delta^0(t), x_\Delta^1(t), \ldots, x_\Delta^j(t), \ldots$ is generated as follows: With $x_\Delta^j(t)$ known, define

(3.8) $$x_\Delta^{j+1}(t) := x_\Delta^j(t) + \xi_\Delta(t)$$

where $\xi_\Delta(t)$ is the collocation solution of the linear problem

(3.9) $$L[x_\Delta^j]\xi = s[x_\Delta^j]$$

with

(3.10) $$s_1 = f(t, y_\Delta^j, z_\Delta^j, \varepsilon) - \varepsilon(y_\Delta^j)'$$

(3.11) $$s_2 = g(t, y_\Delta^j, z_\Delta^j, \varepsilon) - (z_\Delta^j)'$$

(3.12) $$\beta = b(x_\Delta^j(0); x_\Delta^j(1); \varepsilon)$$

(cf. (2.16) - (2.19)).

The formulation of (3.9) as a difference scheme is similar to (3.3) - (3.5) except that the resulting algebraic equations for ξ_Δ^c are linear and can be written as

(3.13) $$L_\Delta[x_\Delta^j]\xi_\Delta^c = s^c[x_\Delta^j]$$

with L_Δ a possibly large, sparse matrix (see §3 of Part II).

Key questions regarding the use of our schemes are the definition of suitable meshes Δ, the existence of solutions to (3.3) - (3.5), their approximation properties with respect to the exact solution of the boundary value problem and the convergence of iterative methods, in particular of Newton's method. The question of convergence of Newton's method is, of course, closely related to that of the stability of the linearized difference operator. Our results regarding these questions are summarized in the following theorem.

THEOREM. Assume the following:

(a) The boundary value problem satisfies the assumptions of §2. Denote by $S_r(x) = \{u\varepsilon C[0,1]; ||u-x|| \leq r\}$ a sphere with radius $r > 0$ around $x\varepsilon C[0,1]$.

(b) The matrix condition

$$\det \begin{vmatrix} P_-E^{-1}(0) \\ P_+E^{-1}(1) \end{vmatrix} \neq 0 \tag{3.14}$$

holds, where $E(t)$ has been defined in (2.13) and

$$P_- = [I \quad 0]\varepsilon R^{n_-xn} \;, \quad P_+ = [0 \quad I]\varepsilon R^{n_+xn} \;, \tag{3.15}$$

I being appropriate identity matrices.

(c) The following mesh construction is used: For a given tolerance δ, $0 < c\varepsilon \leq \delta < 1$, near $t = 0$,

$$h_i := \begin{cases} \varepsilon \; c_u\delta^{1/p} & t_{i+1} \leq \gamma\varepsilon \leq T_0\varepsilon \\ h_{i-1} \exp\{\frac{1}{p}\frac{\nu}{\varepsilon} h_{i-1}\} & \gamma\varepsilon < t_{i+1} \leq T_0\varepsilon \end{cases} \tag{3.16}$$

where

$$p = \begin{cases} 2k & \text{k-stage Gauss scheme} \\ 2(k-1) & \text{k-stage Lobatto scheme} \end{cases} \tag{3.17}$$

$$T_0 = \nu^{-1} \ln \delta^{-1} \tag{3.18}$$

$$\nu = \min\{-\mathrm{re}(\lambda_j(0)), \; j=1,\ldots,n_-\} > 0 \tag{3.19}$$

and c, c_u and γ are positive constants, independent of ε and δ. A similar formula is used near $t = 1$ to construct $(1>)t_N > \ldots > t_{N-N_1+1}(\geq 1-T_1\varepsilon)$ based on the eigenvalues $\lambda_{n_-+1}(1),\ldots,\lambda_n(1)$. In between, a much sparser mesh is used, with

(3.20) $\qquad \varepsilon\underline{h}^{-1}(\bar{i}-\underline{i}) \leq \text{const.}$

where

(3.21) $\qquad \underline{h} := \min\{h_i,\ \underline{i} \leq i < \bar{i}\},\quad \underline{i} := N_0,\quad \bar{i} = N-N_1+1$

<u>THEN</u>, for each scheme of the class considered there are positive constants $\varepsilon_0, \delta_0, h_0, r_0, K_0$, c and $K \geq 0$, independent of ε and Δ, such that there is a unique isolated solution $x_\Delta \in S_{r_0}(x^*)$ provided that $0 < \varepsilon \leq \varepsilon_0$, $\delta \leq \delta_0$, $h \leq h_0$ and $\varepsilon\underline{h}^{-1} \leq K_0$ for a Lobatto scheme, $\varepsilon \sum_{i=\underline{i}}^{\bar{i}-1} h_i^{-1} \leq K_0$ for a Gauss scheme. Further, Newton's method converges quadratically to this solution provided that the starting iterate x_Δ^0 satisfies $x_\Delta^0 \in S_{r_1}(x^*)$, with r_1 sufficiently small. Finally, the numerical solution satisfies

(3.22) $\qquad ||x_\Delta(t_i) - x^*(t_i)|| \leq c(e+\delta) \qquad\qquad 1 \leq i \leq N+1$

where e stands for the following:

(3.23) $$e = \begin{cases} h^{k+q} & \text{k-stage Gauss scheme} \\ Kh^p + h^{k-1+q} & \text{k-stage Lobatto scheme} \end{cases}$$

In (3.23), $K=0$ if $m=0$, p is defined in (3.17) and $q=1$ if the mesh is locally almost uniform, i.e.,

(3.24) $\qquad h_{i+1} = h_i(1 + O(h_i))$ for all i odd <u>or</u> all i even, $\underline{i} \leq i < \bar{i}$

and k is odd for Gauss, even for Lobatto schemes: otherwise, $q=0$. For z the usual superconvergence results hold, i.e. the error at mesh points is $O(h^p)$.

Before giving an outline of the proof to the THEOREM, we wish to remark on some of its details, so that a reader with a primary interest in the algorithm can skip the proof.

The condition (3.14) is a restriction on the differential problem which has nothing to do with its well-posedness. It is a limitation on the applicability of our numerical schemes with full success and is needed to guarantee the stability of the discretization process on the interval $[t_{\underline{i}}, t_{\bar{i}}]$. There, $h_i >> \varepsilon$ and thus the difference operator does not closely approximate the differential operator any more. Computational difficulties can arise when (3.14) is violated, as discussed in Weiss [22] and in Part II.

The principle underlying the layer mesh definition (3.16) is that of keeping the error at the mesh points below the tolerance δ by approximating functions of the type $\exp\{-\lambda_j(0)t/\varepsilon\}$, $j=1,\dots,n_-$, which determine the decay in the boundary layer, see Part II. Of course, to be really constructive one needs to pin down the constants in (3.16) and to provide a working estimate of ν of (3.19). This is done in the next section.

Finally, it is clear that the size of the constant r_1 determining how close x_Δ^0 has to be to x_Δ for Newton's method to converge is very important practically. Unfortunately, our result in this respect is somewhat incomplete: While we can show that for Lobatto schemes r_1 can be chosen independently of ε and Δ, for Gauss schemes our analysis leads to the condition that r_1 shrinks like $(\bar{i} - \underline{i})^{-1}$, as the mesh on $[t_{\underline{i}}, t_{\bar{i}}]$ becomes dense. However, in practice we have never experienced a difference in the domain of attraction of the Newton iteration for the two types of schemes.

Since the proof of the THEOREM is loaded with technicalities, we proceed to give only an outline of it, in an attempt to keep the paper readable.

Outline of the proof of the THEOREM

We consider a k-stage Lobatto scheme and remark about Gauss schemes at the end of this section. Let c denote a generic constant. Also, all norms appearing in the sequel are appropriately restricted maximum norms. As a first step, consider the application of the collocation scheme to the linearized problem at the exact solution x^* of (2.1) - (2.3),

$$(3.25) \qquad L[x^*]x = s[x^*]$$

(recall (2.16) - (2.19)), where the right hand side functions are

$$(3.26) \qquad \begin{aligned} s_1(t) &= f(t,y^*(t),z^*(t),\varepsilon) - A_{11}(t)y^*(t) - A_{12}(t)z^*(t) \\ s_2(t) &= g(t,y^*(t),z^*(t),\varepsilon) - A_{21}(t)y^*(t) - A_{22}(t)z^*(t) \end{aligned}$$

with the ε dependence of the functions involved omitted for brevity. The solution of (3.25), (3.26) is of course x^* as well.

The collocation scheme for (3.25), (3.26) is

$$(3.27a) \qquad \varepsilon h_i^{-1}(y_{ij}-y_i) - \sum_{\ell=1}^{k} \hat{a}_{j\ell}(A_{11}(t_{i\ell})y_{i\ell} + A_{12}(t_{i\ell})z_{i\ell}) = \sum_{\ell=1}^{k} \hat{a}_{j\ell}\, s_1(t_{i\ell}) \equiv \tilde{s}_{ij}$$

$$1 \le i \le N,\ 2 \le j \le k$$

$$(3.27b) \qquad h_i^{-1}(z_{ij}-z_i) - \sum_{\ell=1}^{k} \hat{a}_{j\ell}(A_{21}(t_{i\ell})y_{i\ell} + A_{22}(t_{i\ell})z_{i\ell}) = \sum_{\ell=1}^{k} \hat{a}_{j\ell}\, s_2(t_{i\ell}) \equiv \tilde{r}_{ij}$$

$$(3.27c) \qquad B_0x_1 + B_1x_{N+1} = \beta$$

We now derive stability and convergence results for (3.27). Systems of this form have been investigated in Part II, with only one essential difference: There the matrices A_{ij} were assumed to be smooth functions of the slow variable t for $0 \le t \le 1$, which is not the case here. Still, the key idea of the treatment of Part II can be imported here: To establish unique solvability of (3.27) we consider the discrete system separately at first on the three intervals $[0, t_{\underline{i}}]$, $[t_{\underline{i}}, t_{\bar{i}}]$ and $[t_{\bar{i}}, 1]$, subject to the following special boundary conditions.

I: On $[0, t_{\underline{i}}]$

$$C_-y_1 = \alpha_I,\ z_1 = \beta_I,\ P_+E^{-1}(t_{\underline{i}})y_{\underline{i}} = \gamma_I$$

II: On $[t_{\underline{i}}, t_{\bar{i}}]$

$$P_-E^{-1}(t_{\underline{i}})y_{\underline{i}} = \alpha_{II},\ z_{\underline{i}} = \beta_{II},\ P_+E^{-1}(t_{\bar{i}})y_{\bar{i}} = \gamma_{II}$$

III: On $[t_{\bar{i}}, 1]$

$$P_-E^{-1}(t_{\bar{i}})y_{\bar{i}} = \alpha_{III},\ z_{\bar{i}} = \beta_{III},\ C_+y_{N+1} = \gamma_{III}.$$

Here, $\alpha_{I,II,III} \varepsilon R^{n_-}$, $\gamma_{I,II,III} \varepsilon R^{n_+}$ and $\beta_{I,II,III} \varepsilon R^m$ are arbitrary parameters. The matrices C_- and C_+ are chosen so that the problems

$$\frac{d}{d\tau}\xi = f_y(0, \bar{y}(0) + \mu(\tau), \bar{z}(0), 0)\xi \qquad 0 \leq \tau < \infty$$

$$C_-\xi(0) = 0, \quad \lim_{\tau\to\infty} \xi(\tau) = 0$$

and

$$\frac{d}{d\sigma}\zeta = f_y(1, \bar{y}(1) + \nu(\sigma), \bar{z}(1), 0)\zeta \qquad -\infty < \sigma \leq 0$$

$$C_+\zeta(0) = 0, \quad \lim_{\sigma\to-\infty} \zeta(\sigma) = 0$$

which result from linearizing the layer equations, have only the trivial solutions.

On the "long" interval $[t_{\underline{i}}, t_{\bar{i}}]$ the procedure of Part II is immediately applicable and yields, with (3.14), unique solvability of the problem for all parameters α_{II}, β_{II}, γ_{II} and $\tilde{s}_\Delta$, $\tilde{r}_\Delta$. (The latter two are arbitrary right hand sides in (3.27a) and (3.27b), respectively). Also, it yields the explicit dependence of the solution on the parameters and, since (3.20) holds, the bounds

$$(3.28) \qquad ||x_\Delta^c||_{II} \leq c(||\alpha_{II}|| + ||\beta_{II}|| + ||\gamma_{II}|| + (\bar{i}-\underline{i})||\tilde{s}_\Delta^c||_{II} + ||\tilde{r}_\Delta^c||_{II}).$$

$$(3.29) \qquad ||x_\Delta^c||_{II} \leq c(||\alpha_{II}|| + ||\beta_{II}|| + ||\gamma_{II}|| + ||s_{1\Delta}^c||_{II} + ||\tilde{r}_\Delta^c||_{II}).$$

On the layer interval $[0, t_{\underline{i}}]$, where the $A_{ij}(t)$ vary like $\mu(t/\varepsilon)$, we can employ the results of Markowich and Ringhofer [16] who treat the layer equations in the variable $\tau = t/\varepsilon$ with fast components only. Their mesh construction is as in (3.16) (with an insignificant modification in (3.19)), and a contraction argument, based on the fact that $t_{\underline{i}} << 1$, allows the inclusion of slow variables as well. Thus we obtain the unique solvability of the problem, a representation of the solution in terms of the parameters α_I, β_I and γ_I and the bound

$$(3.30) \qquad ||x_\Delta^c||_I \leq c(||\alpha_I|| + ||\beta_I|| + ||\gamma_I|| + ||\tilde{s}_\Delta^c||_I + ||\tilde{r}_\Delta^c||_I)$$

An analogous situation occurs for the boundary layer at the other end.

The next step is to patch the solutions obtained above together, to obtain a solution of (3.27) by requiring that the representations are identical at $t_{\underline{i}}$ and $t_{\bar{i}}$ and that the boundary conditions (3.27c) be satisfied. This results in a linear system of equations for the parameters, of dimension 3(n+m). Due to assumption (a) of the theorem, the matrix involved has a bounded inverse; the details follow closely those of Weiss [22]. Hence we finally obtain for (3.27)

$$(3.31) \qquad ||x_\Delta^c|| \leq c(||\beta|| + ||\tilde{s}_\Delta^c||_I + (\bar{i}-\underline{i})||\tilde{s}_\Delta^c||_{II} + ||\tilde{s}_\Delta^c||_{III} + ||\tilde{r}_\Delta^c||)$$

$$(3.32) \qquad ||x_\Delta^c|| \leq c(||\beta|| + ||s_{1\Delta}^c|| + ||\tilde{r}_\Delta^c||).$$

Having thus established stability for (3.27), we turn to convergence. The solutions of each of the three discrete problems are approximations to the (general) solutions of the continuous problem (3.25) on the three intervals, subject to the special boundary conditions. The relevant convergence results are described in Markowich and Ringhofer [16] for the layer intervals and in Part II for the "long" interval in between. Using these error estimates in the patching procedure, we obtain the following error estimates relating (3.25) - (3.26) and (3.27). At mesh points,

$$(3.33) \qquad ||x_i - x^*(t_i)|| \leq c(\delta+e) \quad , \qquad i=1,\ldots,N+1,$$

while at collocation points other than mesh points

(3.34) $$||x_{ij} - x^*(t_{ij})|| \leq c(\delta^{k/p} + Kh^k + \varepsilon h^{k-1}) \qquad i=1,\dots,N,\ j=2,\dots,k-1.$$

This completes the description for the linearized problem. After this preparation we turn to the analysis of the nonlinear scheme (3.3) - (3.5). We employ the contraction mapping principle, the application of which to a nonlinear problem

(3.35) $$u = N(u)$$

proceeds in two main steps:

(i) Defining an approximate solution $\hat{u}$ of the problem which leads to a small residual $\hat{u} - N(\hat{u})$, and

(ii) Obtaining a sufficiently small bound on the Lipschitz constant of N in a vicinity of $\hat{u}$.

To put our discrete nonlinear system in the form (3.25), we write it as

(3.36a) $$\varepsilon h_i^{-1}(y_{ij}-y_i) - \sum_{\ell=1}^{k} \hat{a}_{j\ell}(A_{11}(t_{i\ell})y_{i\ell} + A_{12}(t_{i\ell})z_{i\ell}) = \\ = \sum_{\ell=1}^{k} \hat{a}_{j\ell}(f(t_{i\ell},y_{i\ell},z_{i\ell},\varepsilon) - A_{11}(t_{i\ell})y_{i\ell} - A_{12}(t_{i\ell})z_{i\ell}) \\ 1 \leq i \leq N,\ 2 \leq j \leq k$$

(3.36b) $$h_i^{-1}(z_{ij} - z_i) - \sum_{\ell=1}^{k} \hat{a}_{j\ell}(A_{21}(t_{i\ell})y_{i\ell} + A_{22}(t_{i\ell})z_{i\ell}) = \\ = \sum_{\ell=1}^{k} \hat{a}_{j\ell}(g(t_{i\ell},y_{i\ell},z_{i\ell},\varepsilon) - A_{21}(t_{i\ell})y_{i\ell} - A_{22}(t_{i\ell})z_{i\ell})$$

(3.37) $$B_0x_1 + B_1x_{N+1} = B_0x_1 + B_1x_{N+1} - b(x_1; x_{N+1}; \varepsilon),$$

where the matrices A_{rs}, B_r are as above. In concise form, (3.36), (3.37) are written as

(3.38) $$L_\Delta[x^*]x_\Delta = F(x_\Delta).$$

By (3.31) we know that $L_\Delta^{-1}[x^*]$ exists, whence we write (3.38) as

(3.39) $$x_\Delta = L_\Delta[x^*]^{-1}F(x_\Delta) \equiv N(x_\Delta)$$

which is of the type (3.35). Next we establish a contraction argument as follows.

As an approximate solution $\hat{x}_\Delta$ of (3.36) - (3.37) we choose the solution of (3.27). When substituting $\hat{x}_\Delta$ into (3.36) a residual $\tilde{d}_{ij}$ of the form

$$\tilde{d}_{ij} = \sum_{\ell=1}^{k} \hat{a}_{j\ell} d_{ij} \qquad 1 \le i \le N, \quad 2 \le j \le k$$

is obtained, where the vector d_Δ^c formed from the d_{ij} values is bounded in norm by the right hand term of (3.34). Hence, by (3.32)

$$(3.40) \qquad ||\hat{x}_\Delta - N(\hat{x}_\Delta)|| \le c(\delta^{k/p} + Kh^k + \varepsilon h^{k-1})$$

So, $\hat{x}_\Delta$ produces a small residual of (3.39), as desired.

Further, it is clear that due to the smoothness of f, g and b as functions of x and due to (3.32), the Lipschitz constant of N_Δ in a suitably restricted sphere about $\hat{x}_\Delta$ can be made sufficiently small. The contraction mapping principle then yields the existence and uniqueness of a solution x_Δ of (3.39) in this sphere, and the bound

$$(3.41) \qquad ||x_\Delta - \hat{x}_\Delta|| \le c||\hat{x}_\Delta - N(\hat{x}_\Delta)||.$$

Combining (3.41), (3.40) and (3.33), (3.34) we finally obtain the convergence result at collocation points for (3.3) - (3.5),

$$(3.42) \qquad ||x_{ij} - x^*(t_{ij})|| \le c(\delta^{k/p} + Kh^k + \varepsilon h^{k-1}) \qquad 1 \le i \le N, 1 \le j \le k.$$

It is now easy to see that Newton's method can be applied to (3.39) with quadratic convergence. Further, Newton's method is clearly invariant under the transformation that carries (3.38) to (3.39), so the quadratic convergence result applies to the scheme actually employed in practice.

The result (3.42) corresponds to the global convergence estimate for Lobatto schemes in the usual (not singularly perturbed) case. To obtain the corresponding superconvergence results at mesh points, note that

$$(3.43) \qquad ||x^{*c} - x_\Delta^c|| \le ||x^{*c} - \hat{x}_\Delta^c|| + ||\hat{x}_\Delta^c - x_\Delta||.$$

To bound the second term in this inequality, we compare (3.36) to (3.27), apply a Taylor expansion to the right hand side of (3.36) and utilize (3.32) once again to yield

$$(3.44) \qquad ||\hat{x}^c_\Delta - x_\Delta|| \leq c||x^{*c} - x_\Delta||^2.$$

Hence from (3.33), (3.42), (3.43) and (3.44) we finally obtain the desired result (3.22). This completes the proof for Lobatto-type schemes.

The analysis for Gauss-type schemes is significantly less pleasant. The basic reason is that special favourable things occur at collocation points which, in the case of a Gauss scheme, do not include the mesh points. This property, which is actually welcomed in some applications (because it allows for a slick implementation for problems with discontinuous coefficients or problems with artificial singularities) causes here weaker convergence and stability properties than those enjoyed by Lobatto schemes, and a harder analysis to prove them. In particular, weaker convergence properties are already evident in the desired estimates (3.22), (3.23) (which are sharp), a factor of $\bar{i} - \underline{i}$ creeps into the stability estimate (3.29) (hence (3.32)) and the patching procedure at $t_{\underline{i}}$ and $t_{\bar{i}}$ is harder to justify because these are not collocation points any more.

Using a more elaborate analysis we were able to show existence of a discrete solution, unique in a sphere about the exact solution, and the convergence estimates (3.22), (3.23). The convergence of Newton's method, however, is guaranteed in our analysis only when the starting approximation is already in a sphere about x_Δ whose radius shrinks like $(\bar{i} - \underline{i})^{-1}$.

4. Mesh construction

In this section we discuss the practical mesh construction in the layer region $[0, T_0\varepsilon]$, where T_0 is given by (3.18), (3.19). An analogous construction holds, of course, for the right end layer region $[1-T_1\varepsilon, 1]$, while in between a sparse mesh, fine enough only to approximate the reduced solution, is used.

The purpose of the mesh selection is to obey a uniform error tolerance δ (which is considered as an estimate, not a bound), and the strategy is to equidistribute the error with respect to $\mu(\tau)$, which is the dominant solution component in this region, see (2.4), (2.5). This is already the strategy behind the definition (3.16) and we wish here to somewhat refine and precisely specify this selection.

The proposed mesh construction is as follows:

$$(4.1) \qquad h_1 := \varepsilon/\lambda\left[\frac{\nu}{\lambda\,|c_\gamma|}\right]^{1/p}\delta^{1/p} \quad , \qquad \lambda := \max\{|\lambda_j(0)|,\ j=1,\dots,n_-\}$$

$$(4.2) \qquad h_i := h_{i-1}\exp\{1/p\ \frac{\nu}{\varepsilon}\, h_{i-1}\} \qquad i=2,\dots,N_0 \text{ until } t_{N_0} \le T_0\varepsilon \le t_{N_0+1}$$

Here p is defined in (3.17), c_γ is a known constant depending on p and defined in Part I and ν is a slight modification of (3.19), to be discussed below.

The mesh selection strategy (4.1), (4.2) can be easily seen to be at least as conservative as (3.16) for suitable constants c_u and γ. The number of mesh points N_0 obtained by this construction is independent of ε and is proportional to $\delta^{1/p}$, see theorem 4.2 of Part II. Comparing (4.1), (4.2) to the mesh (3.46), (3.47) of Part II, we see that they are essentially the same. The difference is that ν here cannot be determined based on the eigenvalues of $A_{11}(0,0)$.

Since we do not really wish to always compute the reduced solution, for the practical evaluation of ν (at "$\tau \to \infty$", which is only $O(\varepsilon|\ln \delta|)$ away from (3.19)) we can calculate the eigenvalues of $A_{11} = f_y$ at $t = T_0\varepsilon$, say. These eigenvalues will, of course, depend on the currently available approximation to the reduced solution. Thus, a strategy blending the nonlinear Newton iterations with mesh refinement suggests itself. Luckily, however, the solution is not very sensitive to the exact location of the layer mesh points, so re-evaluation of the eigenvalues (and the corresponding redefinition of the layer mesh) is usually not needed more than once.

Consider the function $\mu(\tau)$, with respect to which we are choosing the mesh. In (4.2), we are simply capitalizing on its known exponentially decaying behaviour for τ large enough. If f of (2.1) is linear in y then (2.9) implies that $\mu(\tau)$ has that known exponential behaviour right from $\tau = 0$. Thus, in Part II we have used the "exponential mesh" throughout the layer region. In the more general nonlinear case, what one needs, strictly speaking, is a general nonstiff ODE error control for (2.9), of which the first constant steps in (3.16) are a primitive instance. However, a sophisticated error control can hardly be justified here and, in fact, the mesh (4.1), (4.2) can be used as well to obtain an error proportional to δ, with a moderate constant of proportionality.

5. Numerical examples

In order to test the theoretical results numerically and to demonstrate the power of the obtained schemes, a computer program was written. Newton's method of quasilinearization is implemented and the linearized problems are solved by collocation using local parameter elimination, as described in §3 of Part II. The mesh in possible layer regions is automatically constructed using (4.1) - (4.2), as discussed in the previous section, with the initial solution profile, provided by the user, being used to calculate the eigenvalues at the two boundary points.

That initial solution is an approximation to the reduced solution and is expected to be smooth near the boundaries. Note that the knowledge and use of the reduced solution as an initial guess for Newton's method does not guarantee its convergence. However, the constructed mesh is right and so Newton's method usually converges in practice.

The input tolerance δ is used to control both the layer meshes construction and the convergence of the nonlinear iteration. Optionally, the condition numbers of the matrices L_Δ of (3.13) are calculated. The emphasis in the implementation was on flexibility, rather than efficiency; the efficient implementation of these schemes will be discussed elsewhere.

For the calculations reported below, a floating point system with 14-hexadecimal-digit mantissa was used.

Example 1 (Carrier [6], Chin and Krasny [7]) Consider the problem

$$(5.1) \qquad \varepsilon^2 u'' = 1-2b(1-t^2)u - u^2, \qquad -1 \le t \le 1$$

$$(5.2) \qquad u(-1) = u(1) = 0,$$

where $b \le 0$ is a parameter.

The reduced solution about which the representation (2.4), (2.5) makes sense is

$$(5.3) \qquad \bar{u}(t) = -b(1-t^2) - \sqrt{b^2(1-t^2)^2 +1}$$

To convert to a 1st order system, set

$$(5.4) \qquad y_1 = u, \quad y_2 = \varepsilon u'$$

Using symmetry we then obtain the problem

$$(5.5) \qquad \varepsilon y_1' = y_2$$

$$0 \le t \le 1$$

$$(5.6) \qquad \varepsilon y_2' = 1-2b(1-t^2)y_1 - y_1^2$$

$$(5.7) \qquad y_2(0) = 0, \qquad y_1(1) = 0$$

Thus we have only fast components and the eigenvalues of f_y are

$$(5.8) \qquad \lambda_2(t) = \sqrt{-2(y_1(t) + b(1-t^2))} \;, \qquad \lambda_1(t) = -\lambda_2(t)$$

So, at the reduced solution (5.3), $\lambda_1(t)$ and $\lambda_2(t)$ are real and stay away from 0. Also, $\nu = \sqrt{2}$ in what corresponds to (3.19) for the boundary layer at the right end. (Clearly there is a boundary layer only near $t = 1$). Note that the boundary conditions (5.2) imply that, evaluated at the exact solution, the eigenvalues of the Jacobian matrix vanish at $t = 1$. This, however, does not cause analytic or computational difficulties for our procedures.

Numerical solutions were calculated for $b = 0,1$, and $\varepsilon = 10^{-2},10^{-3}, 10^{-6},10^{-10}$. Some typical values are listed in table 1. The results in [6], [7] were verified. The number of mesh points, the condition numbers of L_Δ and the number of nonlinear iterations needed, were all found to be essentially independent of ε and of b for the above range of parameters. With $\delta = 10^{-6}$ and 10 uniform subintervals away from the layers, we tried a number of initial guesses x_Δ^0: With the reduced solution as x_Δ^0, 3 iterations were needed for convergence on a mesh with $N = 28$, automatically constructed for a 4-stage Lobatto scheme.

Table 1: Selected solution values for example 1
with $b = 1$

ε	$u(0)$	$\varepsilon u'(1)$
10^{-2}	-2.414093	1.174918
10^{-3}	-2.414212	1.156703
10^{-6}	-2.414214	1.154703
10^{-10}	-2.414214	1.154701

With $x_\Delta^0 \equiv \begin{pmatrix}-2\\0\end{pmatrix}$, which gives an $O(1)$ perturbation to the eigenvalues λ_1, λ_2 used to construct the mesh, a mesh with a similar structure was constructed and nonlinear convergence took 4 iterations. Results in both cases were indeed accurate to within δ. On the other hand, with $x_\Delta^0 \equiv 0$ the mesh construction produced an inadequate uniform mesh of 10 subintervals, because $\lambda_2 \approx 0$ near $t = 1$. Thus, while the process is not very sensitive to inaccuracies in the profile and end values of the reduced solution, of course not every initial guess automatically produces a suitable mesh and some care is needed in the design of the initial solution profile. This observation is even more pronounced in the next example.

<u>Example 2</u> Flaherty and O'Malley [9]

This example demonstrates that finding the reduced solutions of a problem may help in more ways than one. Here, multiple solutions are detected. Consider the problem

$$(5.9) \qquad \varepsilon y_1' = \qquad y_2$$

(5.10) $\varepsilon y_2' = \alpha^2(z)y_1 \qquad + \beta(z) \qquad\qquad 0 \le t \le 1$

(5.11) $z' = \qquad\qquad - z + 1$

(5.12) $z(0) + y_1(0) = 0, \; -bz(0) + y_2(0) = 0, \; z(1) + y_1(1) = 0,$

where

(5.13) $\alpha(z) = 1 + 2z \;, \quad \beta(z) = 8z(1-z)$

and b is a parameter. Note that the nonlinearities appear as functions of the slow component alone.

Clearly, the eigenvalues at the reduced solution are

(5.14) $\lambda_1 = \alpha(\bar{z}(t)) \;, \quad \lambda_2 = -\lambda_1$

Also, the reduced solution is given by

(5.15) $\bar{y}_1(t) = -\dfrac{8\bar{z}(t)(1 - \bar{z}(t))}{(1 + 2\bar{z}(t))^2} \;, \quad \bar{y}_2(t) = 0$

$\bar{z}(t) = 1 + e^{-t}[\bar{z}(0) - 1]$

It is much less easy to see what values $\bar{z}(0)$ may take. One could experimentally use (5.15) with a variety of values for $\bar{z}(0)$ as initial approximations for (5.9) - (5.12). However, using the technique described in [9], Flaherty and O'Malley obtained that z(0) may have precisely the following three values,

(5.16) $z(0) = 0, \; \frac{1}{4}[bs - 6 \pm ((bs - 4)^2 + 48)^{\frac{1}{2}}] \;, \; s = \mathrm{sign}(\alpha(\bar{z}(0)))$

The entire construction holds, by (5.14), (5.15), only if

(5.17) $\lambda_1(t) = 3 + 2(\bar{z}(0) - 1)e^{-t} \neq 0 \;, \qquad\qquad 0 \le t \le 1$

Following [9] we have calculated 3 solutions for each of the parameter values b = 2,0,-2. In [9], the general purpose code COLSYS [2], which implements collocation at Gaussian points, was used with the reduced solution as the initial guess. The authors had some difficulties in carrying out the calculations for small ε, and continuation in ε was needed (see [9], [10]).

The reason for these difficulties (as noted by the authors themselves in private communication) is that a uniform mesh was initially used, before allowing COLSYS to adapt it for a given problem. Thus, the approximate solution on the initial mesh had large oscillations throughout the interval [0,1] and was not close in norm to either the exact differential solution or the initial guess.

Here, using the a-priori graded meshes described above, we have encountered no difficulty at all for all cases where (5.17) holds, even with very small ε. No continuation in ε was needed. Using $\delta = 10^{-6}$, the mesh construction of §4 with 10 uniform subintervals away from the boundaries and the reduced solution for the initial guess, solutions were calculated with 3 Gauss, 5 Gauss and 4 Lobatto points per subinterval. The first two choices of collocation points were used by Flaherty and O'Malley [9]. The Lobatto scheme with $k = 4$ has, by (4.1), (4.2), the same mesh construction and computational cost as the Gauss scheme with $k = 3$, while by the THEOREM, its accuracy in h away from the layers is 6, the same as of the Gauss scheme with $k = 5$. It is therefore interesting to compare the actual performance of these 3 schemes.

Computing solutions for $\varepsilon = 10^{-3}, 10^{-6}, 10^{-12}$, we have found that, as in the previous example, the number of mesh points ($N = 28$ for $k = 3$, $N = 18$ for $k = 5$ Gauss points), the condition numbers of L_Δ and the number of nonlinear iterations needed (usually one) were essentially independent of ε and b, as long as (5.17) holds. The Lobatto scheme was particularly accurate for some cases, notably of the negative values of $\bar{z}(0)$ for $b = 2$ and $b = 0$. To understand why, consider the error term e in (3.23). For the Lobatto scheme $e = Kh^6 + \varepsilon h^4$ and the constant K arises from the approximation of the slow components z. Here z is very smooth and is approximated very well. Thus K is very small and, when ε is very small, the error e of (3.23) is very small.

In order to measure computational errors approximately, we have used the reduced solutions for very small ε away from the layers, and additional calculations with denser meshes, to obtain "exact solutions". We have subsequently verified for the above calculations that the layer error tolerance δ has been met to an order of magnitude.

The negative values of $\bar{z}(0)$ provide the more challenging cases. It can be verified that $\lambda_1(t)$ has a zero at

$$(5.18) \qquad \bar{t} = \ln \tfrac{1}{6}(b + 10 + \sqrt{(b + 4)^2 + 48})$$

and so we have a turning point at $\bar{t}$ if $0 < \bar{t} < 1$, and the theory then breaks down. This is the case for $b = -2$ and we note in passing that, while solutions now become unbounded as $\varepsilon \to 0$, for values of ε which are not extremely small solutions can still be calculated using COLSYS, as pointed out by Flaherty and O'Malley [9]. For $b = 0$, $\bar{t} = \ln 3 > 1$, but $\bar{t}$ is close to 1. There results a large boundary layer jump (cf. [9]), however, the condition number of the problem and of L_Δ does not blow up as $\varepsilon \to 0$. Accurate solutions were obtained for this case as well, using the Lobatto scheme with N = 28. Some sample values are given in table 2.

Table 2: Sample solution values for example 2 with $b = 0$, $\bar{z}(0) = -3.5$

ε	$y_1(1)$	$y_2(1)$
10^{-3}	.6555561	-26.70139
10^{-6}	.6554576	-27.71479
10^{-12}	.6554575	-27.71592

Example 3 Flaherty and O'Malley [10]

This problem arises when considering a nonlinear elastic beam which rests on a foundation with nonlinear resistance to deflection. One is led to the system

$$(5.19) \qquad \varepsilon y_1' = -y_2$$

$$(5.20) \qquad \varepsilon y_2' = \phi(z_1)\cos z_2 - y_1(\sec z_2 + \varepsilon y_2 \tan z_2)$$

$$0 \le t \le 1$$

$$(5.21) \qquad z_1' = \sin z_2$$

$$(5.22) \qquad z_2' = y_1$$

where for ϕ we took $\phi(z_1) = z_1 - 1$. See [10] for the development and analysis of this problem.

Now, assuming that y_2 is bounded, we get the eigenvalues

$$(5.23) \qquad \lambda_1(t) = \sqrt{\sec \bar{z}_2(t)} \ , \qquad \lambda_2(t) = -\lambda_1(t)$$

and the reduced solution system

$$(5.24) \qquad \bar{z}_1' = \sin \bar{z}_2 \ , \qquad \bar{y}_1 = \phi(\bar{z}_1)\cos^2 \bar{z}_2$$

$$(5.25) \qquad \bar{z}_2' = \bar{y}_1 \ , \qquad \bar{y}_2 = 0$$

which is referred to as the hanging cable system [10]. The latter system can be integrated if two "reduced" boundary conditions are provided. This can be easily done in the case where the beam is simply supported,

(5.26) $y_1(0) = y_1(1) = 0$

(5.27) $z_1(0) = z_1(1) = 0$

For then, (5.26) is dropped and (5.27) is retained for the reduced solution.

Applying our numerical schemes to the problem (5.19) - (5.22), (5.26), (5.27), we have once again encountered no difficulty. Rather than integrating the hanging cable system, we simply used the following initial approximation.

(5.28) $y_1 = t(1-t),\ y_2 = 0,\ z_1 = \sin \pi t,\ z_2 = 1/2t^2 - 1/3t^3$

With tolerances and mesh sizes as in the previous examples, 3 iterations were needed for convergence.

Table 3: Selected solution values for example 3 with simple support boundary conditions

ε	$y_2(0)$	$z_2(0)$	$y_1(0.5)$	$z_1(0.5)$
10^{-2}	.867460	.426679	-.891701	.108247
10^{-4}	.863935	.434442	-.891686	.108314
10^{-6}	.863899	.434519	-.891686	.108314
10^{-12}	.863899	.434520	-.891686	.108314

Other types of boundary conditions are discussed in [10]. One case is of clamped supports at both endpoints, where (5.27) is retained but (5.26) is replaced by

(5.29) $z_2(0) = z_2(1) = 0$

As argued by Flaherty and O'Malley [10], this set of boundary conditions leads to a problem with unbounded inverse, as $\varepsilon \to 0$ which is therefore not covered by our theory. The analysis in Schmeiser [19] shows that the reduced problem is still given by (5.24), (5.25) and that there are boundary layers of magnitude $O(1)$ in z_2 and boundary layers of magnitude $O(1/\varepsilon)$ in y_1 and y_2. Hence we expect, with a slight modification of the mesh selection procedure, to be able to solve with the same techniques for the clamped supports as well, with almost the same success as before. (Note, however, that we cannot avoid condition numbers of order $O(1/\varepsilon)$ in L_Δ; but that alone would only bother us with unrealistically small values of ε). An extension of our analysis and computations for singular singular-perturbation problems, which covers the above case, will be reported in the near future.

References

1. U. Ascher, "Solving boundary-value problems with a spline-collocation code", J. Comp. Phys. 34 (1980), 401-413.

2. U. Ascher, J. Christiansen and R.D. Russell, "Collocation software for boundary-value ODEs", Trans. Math. Software 7 (1981), 209-222.

3. U. Ascher and R. Weiss, "Collocation for singular perturbation problems I: First order systems with constant coefficients", to appear in SIAM J. Numer. Anal.

4. U. Ascher and R. Weiss, "Collocation for singular perturbation problems II: Linear first order systems without turning points", Tech. Rep. 82-4 (1982), Dept. Comp. Sci., Univ. of B.C.

5. W.-J. Beyn and E. Doedel, "Stability and multiplicity of solutions to discretizations of nonlinear differential equations", SIAM J. Sci. Stat. Comput. 2 (1981), 404-415.

6. G.F. Carrier, "Singular perturbations and geophysics", SIAM Rev. 12 (1970), 175-193.

7. R.C.Y. Chin and R. Krasny, "A hybrid asymptotic-finite element method for stiff two-point boundary value problems", (1981) manuscript.

8. V.A. Episova, "Asymptotic properties of general boundary value problems for singularly perturbed conditionally stable systems of ordinary differential equations", Differential Equs. 11 (1975), 1457-1465.

9. J.E. Flaherty and R.E. O'Malley, Jr., "On the numerical integration of two-point boundary value problems for stiff systems of ordinary differential equations", Proc. BAIL I conf. (1980), J. Miller (ed.), Dublin.

10. J.E. Flaherty and R.E. O'Malley, Jr., "Singularly perturbed boundary value problems for nonlinear systems, including a challenging problem for a nonlinear beam", Proc. Oberwolfach Conf., summer (1981).

11. P.M. Gresho and R.L. Lee, "Don't suppress the wiggles - they're telling you something", in AMD 34 (1979) - ASME, T. Hughes (ed.).

12. P.W. Hemker, H. Schippers and P.M. de Zeeuw, "Comparing some aspects of two codes for two-point boundary-value problems", NW 98/80 Math. Centrum, Amsterdam (1980).

13. H.B. Keller, "Numerical solution of two point boundary value problems", SIAM monograph 24 (1976).

14. R.B. Kellog, G.R. Shubin and A.B. Stephens, "Uniqueness and the cell Reynolds number", SIAM J. Numer. Anal. 17 (1980), 733-739.

15. B. Kreiss and H.O. Kreiss, "Numerical methods for singular perturbation problems", SIAM J. Numer. Anal. 18 (1981), 262-276.

16. P.A. Markowich and C.A. Ringhofer, "Collocation methods for boundary value problems on 'long' intervals", to appear in Math. Comp.

17. R.E. O'Malley, Jr., "On multiple solutions of singularly perturbed systems in the conditionally stable case", in Singular perturbations and asymptotics, R. Meyer and S. Parter (eds.), (1980).

18. C. Ringhofer, "On collocation schemes for quasilinear singularly perturbed boundary value problems", Manuscript (1982).

19. C. Schmeiser, "Behandlung eines nichtlinearen balkenmodelles mit methoden der singulären störungsrechnung", Ms. thesis (1981), Tech. Univ. Wien.

20. P. Spudich and U. Ascher, "Calculation of complete theoretical seismograms in vertically varying media using collocation methods", Manuscript (1982).

21. F.Y.M. Wan and U. Ascher, "Horizontal and flat points in shallow cap dimpling" Tech. Rep. 80-5, Inst. Appl. Math. and Stat., University of British Columbia, Vancouver, Canada (1980).

22. R. Weiss, "An analysis of the box and trapezoidal schemes for linear singularly perturbed boundary value problems", Manuscript (1982).