COLLOCATION FOR SINGULAR PERTURBATION PROBLEMS II: LINEAR FIRST ORDER SYSTEMS WITHOUT TURNING POINTS

by

U. Ascher* and R. Weiss**

Technical Report 82-4

May 1982

Abstract

We consider singularly perturbed linear boundary value problems for ODES, with variable coefficients, but without turning points. Convergence results are obtained for collocation schemes based on Gauss and Lobatto points, showing that highly accurate numerical solutions for these problems can be obtained at a very reasonable cost using such schemes, provided that appropriate meshes are used. The implementation of the numerical schemes and the practical construction of corresponding meshes are discussed.

These results extend those of a previous paper which deals with systems with constant coefficients.

Subject classification: 65L10

---

*Department of Computer Science, The University of British Columbia, 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

In part I of this work [2], (hereinafter referred to as "Part I") we have considered the numerical solution of singularly perturbed boundary value ordinary differential equations with constant coefficients. Our attention was focused on symmetric collocation schemes, which include the midpoint (or box) and the trapezoidal difference schemes as special cases. We have shown that such schemes can be used to compute highly accurate numerical solutions at a very reasonable cost, provided that appropriate meshes are used. Such a mesh consists, in general, of three parts: Two fine grids near the boundaries, to cover the possible two layer regions, and a coarser grid in between.

Similar results for the variable coefficient case are obtained in Weiss [9] for the trapezoidal and midpoint schemes. The eigenvalues of the "fast component" part of the differential equations are assumed to stay away from the imaginary axis for all values of the independent variable. Thus, turning points are excluded from the discussion. In the passage from constant to variable coefficients, the analysis had to be extended significantly.

In this paper we extend the results of the two papers mentioned above to include convergence results for the collocation schemes based on Gauss and Lobatto points for linear two-point boundary value problems which have a uniformly bounded inverse and which are restricted as in [9].
In addition, we describe an implementation of these schemes, discuss practical mesh construction and demonstrate our results numerically.

The general problem considered in this paper is of order n+m, with n equations singularly perturbed,

$$\varepsilon \underset{\sim}{y}' = A_{11}(t,\varepsilon)\underset{\sim}{y} + A_{12}(t,\varepsilon)\underset{\sim}{z} + \underset{\sim}{f}_1(t) \qquad 0 \le t \le 1 \tag{1.1}$$

$$\underset{\sim}{z}' = A_{21}(t,\varepsilon)\underset{\sim}{y} + A_{22}(t,\varepsilon)\underset{\sim}{z} + \underset{\sim}{f}_2(t) \tag{1.2}$$

plus the boundary conditions for $\underset{\sim}{x}(t) = \begin{pmatrix} \underset{\sim}{y}(t) \\ \underset{\sim}{z}(t) \end{pmatrix}$

$$B_0 \underset{\sim}{x}(0) + B_1 \underset{\sim}{x}(1) = \underset{\sim}{\beta} . \tag{1.3}$$

The assumption (2.3) below on the eigenvalues of A_{11} plus the other regularity assumptions lead to the conclusion that the solution of (1.1) - (1.3) consists of a smooth curve away from the boundaries, possibly connected at each end to the boundary by a thin transition layer. As was pointed out in Part I, with Gauss or Lobatto schemes these boundary layer solutions must be approximated accurately, because otherwise layer errors would propagate throughout the entire interval of integration. The meshes used for collocation thus consist of three parts: Two fine grids near each boundary, with maximum mesh spacing $h_L \le K\varepsilon$ for a suitable constant K, connected to a much sparser mesh away from the boundaries with minimum mesh spacing $\underline{h} >> \varepsilon$. The determination of the sparse mesh is based on the accuracy needed in the approximation of the reduced solution. The total number of mesh points required to meet a given error tolerance can be made to be independent of ε.

Of course, the mesh described above becomes highly nonuniform for very small ε. However, higher order collocation methods can handle such nonuniformity, see Part I and Ascher, Pruess and Russell [1]. Thus they are preferable to convergence acceleration methods in this context.

Following a short section where some results on the analytic solution of (1.1) - (1.3) are gathered for later use is Section 3, where the numerical schemes, their implementation and properties and the convergence results are presented. In §3.1 we describe a careful implementation of the collocation schemes which uses local-unknowns elimination, resulting with a well-conditioned system of linear equations (3.59) of a familiar sparse

structure, independent of the order of the scheme. This implementation is used both for the analysis and for the numerical calculations in following sections.

In §3.2 we consider a transformation of the dependent variables, needed for the analysis. Whereas in Part I this transformation commutes with the collocation operator, here it does not, and the resulting residue is shown to be sufficiently small in norm so that it can be considered as a small perturbation in regions where the mesh is dense, i.e. in boundary layer regions.

In §3.3 the mesh is described, together with the general collocation solution decomposition on each of its three parts. Then, in §3.4, our convergence results are stated. Theorem 3.1 summarizes the results for the layer regions near the boundaries while theorem 3.2 describes our results in the region away from the boundaries. Theorem 3.3 then states the combined results of the previous two theorems on the entire interval. This theorem is essentially the same as theorem 5.3 of Weiss [9]. An outline of its proof is followed by some remarks on the practical calculation of solutions by these schemes. Finally, we discuss the construction of the mesh near a boundary when there is no boundary layer because of a particularly lucky choice of boundary values. Such "luck" occurs frequently in practical problems. We find the argument in Kreiss and Kreiss [4] in this regard incomplete.

Sections 4 and 5 are devoted almost entirely to the proofs of theorems 3.1 and 3.2, respectively. In §4 we also discuss the layer mesh construction and show that the number of mesh points needed to achieve overall accuracy δ for any ε, $0 < \varepsilon \leq \varepsilon_1$, is $O(\delta^{-\frac{1}{p}})$, where p is the order of superconvergence of the method, defined in (3.42). This, provided that the mesh

defined in (3.46), (3.47) is used. If a uniform layer mesh is used instead then the number of mesh points needed is $O(-\delta^{-1/P} \ln \delta)$. But the actual advantage of (3.46), (3.47) over a uniform layer mesh is more significant than these bounds would indicate; see table 4.2 of Part I.

It is interesting to note that, perhaps contrary to one's first intuition, the analysis for the "long" interval away from the boundaries, where the solution varies slowly, is much more gruelling than the analysis for the layer intervals, where the solution varies very rapidly. In fact, the solution in the layer is dominated by a rapidly decreasing exponential and so its form is very smooth and simple to approximate, provided that we have a layer mesh with step sizes proportional to ε, affecting a stretching transformation. Indeed, it is the simple, exponential form of the layer solution which enables us to come up with the a-priori error equidistributing mesh (3.46), (3.47), whereas in general such meshes can be constructed only adaptively. Markowich and Ringhofer [6] had a similar success with problems on infinite intervals.

In §6, we seal this paper with a numerical example demonstrating our theoretical results.

2. Analytic preliminaries

In this section we mention some analytic results needed in the sequel and develop some notation. Since this section covers the same ground as §2 of Weiss [9], we allow ourselves to omit some details here.

Consider the linear problem (1.1), (1.2) where $A_{ij} = A_{ij}(t,\varepsilon)$ and $\underset{\sim}{f}_i = \underset{\sim}{f}_i(t,\varepsilon)$ are assumed, for simplicity, to be in $C^\infty([0,1] \times [0,\varepsilon_0])$ for some $\varepsilon_0 > 0$, $1 \le i, j \le 2$. Further, assume that

$$A_{11}(t,0) = E(t)\,\Lambda(t)\,E^{-1}(t) \tag{2.1}$$

with $E \in C^\infty[0,1]$,

$$\Lambda(t) = \operatorname{diag}\{\lambda_1(t),\ldots,\lambda_n(t)\} \tag{2.2}$$

and

$$\operatorname{re}(\lambda_i(t)) \begin{cases} < 0, & i = 1,\ldots,n_- \\ > 0, & i = n_-+1,\ldots,n \end{cases}, \qquad t \in [0,1] \tag{2.3}$$

Let $n_+ := n-n_-$, and denote $\Lambda_-(t) = \operatorname{diag}\{\lambda_1(t),\ldots,\lambda_{n_-}(t)\}$, $\Lambda_+(t) = \operatorname{diag}\{\lambda_{n_-+1}(t),\ldots,\lambda_n(t)\}$.

We wish to decouple the slow components $\underset{\sim}{z}$ from the fast ones and to diagonalize the remaining system for $\underset{\sim}{y}$. This is possible for a system with constant coefficients, see Part I; here, we can only almost get it. With $L(t)$ a smooth solution to

$$\varepsilon L' = -LA_{11} + \varepsilon(A_{22}L - LA_{12}L) + A_{21} \tag{2.4}$$

define the transformation

$$\begin{pmatrix} \underset{\sim}{u} \\ \underset{\sim}{v} \end{pmatrix} = \begin{pmatrix} E^{-1} & 0 \\ -\varepsilon L & I \end{pmatrix} \begin{pmatrix} \underset{\sim}{y} \\ \underset{\sim}{z} \end{pmatrix} \tag{2.5}$$

(See Weiss [9] for justification). The system (2.1) - (2.2) is then transformed into

$$\varepsilon \underset{\sim}{u}' = (\Lambda + \varepsilon B_{11})\underset{\sim}{u} + B_{22}\underset{\sim}{v} + \underset{\sim}{g}_1 \tag{2.6}$$

(2.7) $\quad \underset{\sim}{v}' = B_{22}\underset{\sim}{v} + \underset{\sim}{g}_2$

where B_{11}, B_{12}, $\underset{\sim}{g}_1$, $\underset{\sim}{g}_2$ are smooth functions of t and ε.

For the transformed system (2.6) - (2.7), a desirable representation of the solution is obtained [5]: Writing it compactly as

(2.8) $\quad H\underset{\sim}{w} = \underset{\sim}{g}$

with $\underset{\sim}{w}(t) = \begin{pmatrix} \underset{\sim}{u}(t) \\ \underset{\sim}{v}(t) \end{pmatrix}$, $\underset{\sim}{g}(t) = \begin{pmatrix} \underset{\sim}{g}_1(t) \\ \underset{\sim}{g}_2(t) \end{pmatrix}$, and introducing the maps $P_- \in \mathbb{R}^{n_- \times n}$ an $P_+ \in \mathbb{R}^{n_+ \times n}$ defined by

(2.9) $\quad P_-\underset{\sim}{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_{n_-} \end{pmatrix}$, $P_+\underset{\sim}{x} = \begin{pmatrix} x_{n_-+1} \\ \vdots \\ x_n \end{pmatrix}$, $\underset{\sim}{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$

we have

Theorem 2.1. The system (2.8) subject to boundary conditions

(2.10) $\quad P_-\,\underset{\sim}{u}(0) = \underset{\sim}{\eta}_- \in \mathbb{R}^{n_-}$, $\quad P_+\,\underset{\sim}{u}(1) = \underset{\sim}{\eta}_+ \in \mathbb{R}^{n_+}$, $\quad \underset{\sim}{v}(0) = \underset{\sim}{\eta}_0 \in \mathbb{R}^m$

has a unique solution which satisfies

(2.11) $\quad ||\underset{\sim}{w}|| \le \text{const}\,(||\underset{\sim}{g}|| + ||\underset{\sim}{\eta}_-|| + ||\underset{\sim}{\eta}_+|| + ||\underset{\sim}{\eta}_0||)$,

provided that ε is sufficiently small, $0 < \varepsilon \le \varepsilon_1$. Also, for any $q \ge 0$ there is a particular solution $\underset{\sim}{w}_p(t) = \begin{pmatrix} \underset{\sim}{u}_p(t) \\ \underset{\sim}{v}_p(t) \end{pmatrix}$ of (2.8) which satisfies

(2.12) $\quad \sum_{j=0}^{q} ||\frac{d^j \underset{\sim}{w}_p}{dt^j}|| \le \text{const}, \quad 0 < \varepsilon \le \varepsilon_1$.

Now, define matrix solutions W_-, W_+ and W_0 to the homogeneous problem (2.8) with $\underset{\sim}{g} = 0$ as follows:

(i) $W_- = \begin{pmatrix} U_- \\ 0 \end{pmatrix}$, $U_- \in \mathbb{R}^{n \times n_-}$, where U_- satisfies

$$\varepsilon U_-' = (\Lambda + \varepsilon B_{11})U_-, \quad P_-U_-(0) = I, \quad P_+U_-(1) = 0$$

(ii) $W_+ = \begin{pmatrix} U_+ \\ 0 \end{pmatrix}$, $U_+ \in \mathbb{R}^{n \times n_+}$, where U_+ satisfies

$$\varepsilon U_+' = (\Lambda + \varepsilon B_{11})U_+; \quad P_-U_+(0) = 0, \quad P_+U_+(1) = I$$

(iii) $W_0 = \begin{pmatrix} U_0 \\ V_0 \end{pmatrix}$, $U_0 \in \mathbb{R}^{n \times m}$, $V_0 \in \mathbb{R}^{m \times m}$, where

$$HW_0 = 0;\ V_0(0) = I,\ P_- U_0(0) = S_-(\varepsilon),\ P_+ U_0(1) = S_+(\varepsilon)$$

and $S_- \in \mathbb{R}^{n_- \times m}$, $S_+ \in \mathbb{R}^{n_+ \times m}$ can be chosen by theorem 2.1 such that

$$\sum_{j=0}^{q} \left\| \frac{d^j W_0}{dt^j} \right\| \leq \text{const.} \tag{2.13}$$

Then we obtain the desired representation of the general solution to (2.6) - (2.7):

Theorem 2.3 Any solution $\underset{\sim}{w}$ of (2.8) can be written as

$$\underset{\sim}{w} = W_- \underset{\sim}{\zeta}_- + W_+ \underset{\sim}{\zeta}_+ + W_0 \underset{\sim}{\zeta}_0 + \underset{\sim}{w}_p \tag{2.14}$$

with $\underset{\sim}{\zeta}_- \in \mathbb{R}^{n_-}$, $\underset{\sim}{\zeta}_+ \in \mathbb{R}^{n_+}$ and $\underset{\sim}{\zeta}_0 \in \mathbb{R}^m$. The (smooth) particular solution satisfies (2.12), and the matrices W_-, W_+ and W_0, defined above, have the asymptotic expansions

$$\begin{aligned} W_0(t) &= \sum_{j=0}^{q} W_{0j}(t)\, \varepsilon^j + O(\varepsilon^{q+1}) \\ U_-(t) &= \sum_{j=0}^{q} U_{-j}(t/\varepsilon)\, \varepsilon^j + O(\varepsilon^{q+1}) \\ U_+(t) &= \sum_{j=0}^{q} U_{+j}((t-1)/\varepsilon)\, \varepsilon^j + O(\varepsilon^{q+1}) \end{aligned} \tag{2.15}$$

From the expansions (2.15) it is clear that

$$\begin{aligned} U_{-0}(t/\varepsilon) &= \exp\left(\Lambda_-(0)\, t/\varepsilon\right) \\ U_{+0}\left(\frac{t-1}{\varepsilon}\right) &= \exp\left(\Lambda_+(1)\, (t-1)/\varepsilon\right). \end{aligned} \tag{2.16}$$

Consider now the linear boundary value problem (1.1) - (1.3). The boundary conditions are transformed by (2.5) into a similar form for $\underset{\sim}{w}(0)$ and $\underset{\sim}{w}(1)$ and substituting the representation (2.14) into these boundary conditions we get

$$M(\varepsilon)\underset{\sim}{\zeta} = \hat{\underset{\sim}{\beta}}(\varepsilon), \tag{2.17}$$

where $\underset{\sim}{\zeta} = (\underset{\sim}{\zeta}_-, \underset{\sim}{\zeta}_+, \underset{\sim}{\zeta}_0)^T$ and the matrix M has an expansion

$$M(\varepsilon) = \sum_{j=0}^{q} M_j \varepsilon^j + O(\varepsilon^{q+1}). \tag{2.18}$$

We assume that M_0 is nonsingular. This is equivalent to assuming that the boundary value problem (1.1) - (1.3) is well posed, i.e. for ε small enough

$$(2.19) \qquad \left|\left|\begin{matrix}\underset{\sim}{v}\\ \underset{\sim}{z}\end{matrix}\right|\right| \le \text{const } \left(\left|\left|\begin{matrix}\underset{\sim}{f}_1\\ \underset{\sim}{f}_2\end{matrix}\right|\right| + ||\underset{\sim}{\beta}||\right)$$

with the constant independent of ε.

It is clear that the preceding representation of the general solution of (2.8) can be made analogously on any interval $[\underline{t},\bar{t}] \subset [0,1]$ with the solution matrices appropriately defined. In particular, in (2.15), (2.16), $\underline{t}$ would replace 0 and $\bar{t}$ would replace 1. Denoting by $(\upsilon_-)_\ell$ and $(\upsilon_+)_\ell$ the ℓ-th columns of υ_- and υ_+ we get

$$(2.20) \qquad \left|\left|\frac{d^j(\upsilon_-)_\ell}{dt^j}\right|\right| \le \text{const } \varepsilon^{-j}[\exp\{\text{re}(\lambda_\ell(\underline{t}))(t-\underline{t})/\varepsilon\} + O(\varepsilon)],$$

$$\underline{t} \le t \le \bar{t},\ \ell=1,\ldots,n_-\quad j = 0,1,\ldots,q$$

$$(2.21) \qquad \left|\left|\frac{d^j(\upsilon_+)_\ell}{dt^j}\right|\right| \le \text{const } \varepsilon^{-j}[\exp\{\text{re}(\lambda_\ell(\bar{t}))(t-\bar{t})/\varepsilon\} + O(\varepsilon)],$$

$$\underline{t} \le t \le \bar{t},\ \ell=n_-+1,\ldots,n,\ j = 0,1,\ldots,q.$$

3. Numerical solutions and their convergence

3.1. The numerical schemes and their implementation

In section 3 of Part I we have presented some classes of collocation methods and discussed their equivalent Runge-Kutta formulation and some of their properties. Here we mention only some of these details again and rely on familiarity with Part I for the rest.

A collocation procedure under consideration is completely determined in terms of k points ($k \geq 1$),

$$0 \leq \rho_1 < \ldots < \rho_k \leq 1, \tag{3.1}$$

which we take to be either Gauss or Lobatto points, and a mesh

$$\begin{aligned} &\Delta: 0 = t_1 < t_2 < \ldots < t_N < t_{N+1} = 1 \\ &h_i := t_{i+1} - t_i, \quad 1 \leq i \leq N, \quad h := \max\{h_i, \ 1 \leq i \leq N\} \end{aligned} \tag{3.2}$$

On a given mesh Δ, the collocation solution $\underset{\sim}{x}_\Delta(t) = \begin{pmatrix} \underset{\sim}{y}_\Delta(t) \\ \underset{\sim}{z}_\Delta(t) \end{pmatrix}$ to (1.1) - (1.3) is a continuous piecewise polynomial vector function of degree at most k satisfying the boundary conditions (1.3) and the differential equations (1.1), (1.2) at the collocation points

$$t_{ij} := t_i + h_i\rho_j \qquad i = 1,\ldots,N, \ j = 1,\ldots,k. \tag{3.3}$$

Inside each subinterval $[t_i, t_{i+1}]$, the polynomials $\underset{\sim}{y}_\Delta(t)$ and $\underset{\sim}{z}_\Delta(t)$ can be represented in terms of the values

$$\begin{aligned} &\underset{\sim}{y}_i := \underset{\sim}{y}_\Delta(t_i), \quad \underset{\sim}{z}_i := \underset{\sim}{z}_\Delta(t_i), \qquad 1 \leq i \leq N+1 \\ &\underset{\sim}{y}_{ij} := y_\Delta(t_{ij}), \quad \underset{\sim}{z}_{ij} := \underset{\sim}{z}_\Delta(t_{ij}) \qquad 1 \leq i \leq N, \ 1 \leq j \leq k \end{aligned} \tag{3.4}$$

(strictly speaking, for Lobatto points some additional derivative values are required as well), which satisfy the difference equations

$$\frac{\varepsilon}{h_i}(\underset{\sim}{y}_{ij} - \underset{\sim}{y}_i) = \sum_{\ell=1}^{k} \hat{a}_{j\ell}(A_{11}(t_{i\ell},\varepsilon)\underset{\sim}{y}_{i\ell} + A_{12}(t_{i\ell},\varepsilon)\underset{\sim}{z}_{i\ell} + \underset{\sim}{f}_1(t_{i\ell})) \qquad 1 \leq j \leq k \tag{3.5}$$

$$(3.6)\qquad \frac{1}{h_i}(z_{ij} - z_i) = \sum_{\ell=1}^{k} \hat{a}_{j\ell}(A_{21}(t_{i\ell},\varepsilon)y_{i\ell} + A_{22}(t_{i\ell},\varepsilon)z_{i\ell} + f_2(t_{i\ell}))$$

For Lobatto points, $\rho_k = 1$ and $\rho_1 = 0$. Thus $y_{i+1} = y_{ik}$, $z_{i+1} = z_{ik}$ and equations (3.5), (3.6) are trivial for j=1. For Gauss points, $\rho_k < 1$, $\rho_1 > 0$ and we extend the range of j in (3.5), (3.6) to include j = k+1 as well, with $y_{i+1} = y_{i,k+1}$, $z_{i+1} = z_{i,k+1}$ and $\hat{a}_{k+1,\ell} = \hat{b}_\ell$, $\ell=1,\ldots,k$; see §3 of Part I for the definitions of the constants $\hat{a}_{j\ell}$, $\hat{b}_\ell$, as well as the matrices $\hat{A}$ and $\bar{A}$ used later.

In the sequel, we shall adhere to the following notational convention, used already above. The collocation approximation to a function $\psi(t)$ is denoted by $\psi_\Delta(t)$. Its values at mesh points are ψ_i, $1 \le i \le N+1$, and those at collocation points are ψ_{ij}, $1 \le i \le N$, $1 \le j \le k$. Also, ψ^c will denote the vector formed by the restriction of $\psi(t)$ to $\Delta \cup \{t_{ij};\ 1 \le i \le N,\ 1 \le j \le k\}$. As well, c, K and c_j, j = 0,1,2... will denote constants independent of ε and Δ.

Next, we describe a particular, careful implementation of the collocation schemes which is used both for the numerical calculations reported in §6 and for the analysis in §5. The differential equations (1.1), (1.2) are written as one system

$$(3.7)\qquad x' = A(t)x + f(t)$$

for which the numerical method is written in Runge-Kutta form

$$(3.8)\qquad x_{i+1} = x_i + h_i \sum_{j=1}^{k} \hat{b}_j F_{ij} \qquad 1 \le i \le N$$

$$(3.9)\qquad F_{ij} = x'_\Delta(t_{ij}) = A(t_{ij})x_{ij} + f(t_{ij}) = A(t_{ij})(x_i + h_i \sum_{\ell=1}^{k} \hat{a}_{j\ell} F_{i\ell}) + f(t_{ij}).$$

The unknowns F_{ij} (or x_{ij}) for each interval $[t_i,t_{i+1}]$ are local and can be eliminated locally. (This is sometimes referred to as "condensation of parameters" - see Ascher, Pruess and Russell [1]). We choose to locally

eliminate the $\underset{\sim}{F}_{ij}$ in case that $\rho_k < 1$, and the $\underset{\sim}{x}_{ij}$ in case that $\rho_k = 1$. These choices avoid unnecessary loss of digits due to cancellation error, as can be readily verified for the example $y' = y/\varepsilon + 1/\varepsilon$ with $0 < \varepsilon << 1$.

Consider Gauss points first. Equations (3.9) can be written as $(n+m)k$ linear equations

$$\underset{\sim}{J}_i \underset{\sim}{F}_i = \underset{\sim}{R}_i \tag{3.10}$$

where

$$\underset{\sim}{F}_i = (\underset{\sim}{F}_{i1},\ldots,\underset{\sim}{F}_{ik})^T , \quad \underset{\sim}{R}_i = C_A \underset{\sim}{x}_i + \underset{\sim}{f}_i$$

$$C_A = \begin{bmatrix} A(t_{i1}) \\ \vdots \\ A(t_{ik}) \end{bmatrix} , \quad \underset{\sim}{f}_i = \begin{pmatrix} \underset{\sim}{f}(t_{i1}) \\ \vdots \\ \underset{\sim}{f}(t_{ik}) \end{pmatrix} \tag{3.11}$$

$$J_i = I - h_i \begin{bmatrix} \hat{a}_{11}A(t_{i1}) & \hat{a}_{12}A(t_{i1}) & \ldots. & \hat{a}_{1k}A(t_{i1}) \\ \hat{a}_{21}A(t_{i2}) & \hat{a}_{22}A(t_{i2}) & \ldots. & \hat{a}_{2k}A(t_{i2}) \\ \vdots & & & \\ \hat{a}_{k1}A(t_{ik}) & \ldots\ldots\ldots\ldots & & \hat{a}_{kk}A(t_{ik}) \end{bmatrix} = I - h_i D_A(\hat{A}\otimes I) \tag{3.12}$$

in which I stands for an identity matrix of the appropriate dimension ($n+m$ or $(n+m)k$) and $D_A = \text{diag}\,\{A(t_{i1}),\ldots,A(t_{ik})\}$. (The dependence on i is suppressed in C_A and D_A). Introducing for notational purposes the $(n+m)\times(n+m)k$ matrix

$$\hat{B} = [\hat{b}_1 I,\ldots,\hat{b}_k I] \tag{3.13}$$

we can write (3.8) as

$$\underset{\sim}{X}_{i+1} = \Gamma_i \underset{\sim}{x}_i + \underset{\sim}{g}_i \qquad 1 \le i \le N \tag{3.14}$$

where

$$\Gamma_i = I + h_i \hat{B} J_i^{-1} C_A , \quad \underset{\sim}{g}_i = h_i \hat{B} J_i^{-1} \underset{\sim}{f}_i \tag{3.15}$$

The difference equations (3.14) together with the boundary equations corresponding to (1.3)

$$B_0 \underset{\sim}{X}_1 + B_1 \underset{\sim}{X}_{N+1} = \underset{\sim}{\beta} \tag{3.16}$$

form a set of (N+1)(n+m) linear equations whose size and structure are independent of k for the solution values at the mesh points.

For Lobatto points we perform a similar elimination of local parameters, but now our parameters are $\underset{\sim}{x}_i = \underset{\sim}{x}_{i1}, \underset{\sim}{x}_{i2}, \ldots, \underset{\sim}{x}_{i,k-1}, \underset{\sim}{x}_{ik} = \underset{\sim}{x}_{i+1}$. Instead of (3.8), (3.9) we write, as in (3.5), (3.6)

$$1/h_i(\underset{\sim}{x}_{ij} - \underset{\sim}{x}_i) = \sum_{\ell=1}^{k} \hat{a}_{j\ell}(A(t_{i\ell})\underset{\sim}{x}_{i\ell} + \underset{\sim}{f}(t_{i\ell})) \qquad 2 \le j \le k \tag{3.17}$$

and this can be written as (n+m)(k-1) linear equations

$$\bar{J}_i \bar{\underset{\sim}{x}}_i = \bar{\underset{\sim}{R}}_i \tag{3.18}$$

where

$$\bar{\underset{\sim}{x}}_i = (\underset{\sim}{x}_{i2}, \ldots, \underset{\sim}{x}_{ik})^T, \qquad \bar{\underset{\sim}{R}}_i = (\bar{\underset{\sim}{R}}_{i2}, \ldots, \bar{\underset{\sim}{R}}_{ik})^T$$

$$\bar{\underset{\sim}{R}}_{ij} = [I + h_i \hat{a}_{j1} A(t_i)]\underset{\sim}{x}_i + h_i \sum_{\ell=1}^{k} \hat{a}_{j\ell} \underset{\sim}{f}(t_{i\ell}) \tag{3.19}$$

$$\bar{J}_i = I - h_i \begin{bmatrix} \hat{a}_{22}A(t_{i2}) & \cdots & \hat{a}_{2k}A(t_{ik}) \\ \vdots & & \\ \hat{a}_{k2}A(t_{i2}) & \cdots & \hat{a}_{kk}A(t_{ik}) \end{bmatrix} = I - h_i(\bar{A}\otimes I)\bar{D}_A \tag{3.20}$$

with $\bar{D}_A = \text{diag}\{A(t_{i2}), \ldots, A(t_{ik})\}$ and $\bar{A}$ is a nonsingular matrix, as in (3.14) of Part I.

Since $\rho_k = 1$, $\underset{\sim}{x}_{i+1}$ is obtained as the last n+m rows of $\bar{J}_i^{-1}\bar{\underset{\sim}{R}}_i$. Partitioning $\bar{J}_i^{-1}$ into blocks of size (n+m)x(n+m), $\bar{J}_i^{-1} = ((\bar{J}_i^{-1})_{j\ell})_{j,\ell=2}^{k}$, we get difference equations of the form (3.14) where now, instead of (3.15),

$$\Gamma_i = \sum_{\ell=2}^{k} (\bar{J}_i^{-1})_{k\ell}[I + h_i \hat{a}_{\ell 1} A(t_i)], \quad \underset{\sim}{g}_i = h_i \sum_{\ell=2}^{k} (\bar{J}_i^{-1})_{k\ell} \sum_{s=1}^{k} \hat{a}_{\ell s} \underset{\sim}{f}(t_{is}). \tag{3.21}$$

An advantage of the difference equations (3.14), (3.16), obtained both for Gauss and for Lobatto points, is that even when some rows of A(t) of (3.7) depend on $1/\varepsilon$ and $\varepsilon \ll h_i$, the components of Γ_i and $\underset{\sim}{g}_i$ remain bounded and are constructed accurately.

3.2 Transformation of variables

Consider the linear problem (1.1) - (1.3) and the transformed system (2.6), (2.7). Since the latter has a structure more amenable to analysis, we will rely on it in parts of our treatment. However, we stress that the actual numerical procedure is applied to (1.1), (1.2) and not to (2.6), (2.7).

In the constant coefficient case, the operators of collocation and the transformation (2.5) commute. Here they do not, in general. Thus, if we define vector functions $\underset{\sim}{u}_\Delta(t)$, $\underset{\sim}{v}_\Delta(t)$ by

$$\begin{pmatrix} \underset{\sim}{u}_\Delta \\ \underset{\sim}{v}_\Delta \end{pmatrix} = \begin{pmatrix} E^{-1} & 0 \\ -\varepsilon L & I \end{pmatrix} \begin{pmatrix} \underset{\sim}{y}_\Delta \\ \underset{\sim}{z}_\Delta \end{pmatrix} \tag{3.22}$$

then $\underset{\sim}{u}_\Delta$, $\underset{\sim}{v}_\Delta$ collocate the transformed equations, but are not necessarily piecewise polynomials of degree at most k. Correspondingly, applying the transformation (3.22) to the difference equations (3.5), (3.6) we obtain

$$\frac{\varepsilon}{h_i}(\underset{\sim}{u}_{ij} - \underset{\sim}{u}_i) = \sum_{\ell=1}^{k} \hat{a}_{j\ell}\{[\Lambda(t_{i\ell}) + \varepsilon B_{11}(t_{i\ell})]\underset{\sim}{u}_{i\ell} + B_{12}(t_{i\ell})\underset{\sim}{v}_{i\ell} + \underset{\sim}{g}_1(t_{i\ell})\} + \frac{\varepsilon}{h_i}\underset{\sim}{R}_{ij} \qquad 1 \le i \le N \tag{3.23}$$

$$\frac{1}{h_i}(\underset{\sim}{v}_{ij} - \underset{\sim}{v}_i) = \sum_{\ell=1}^{k} \hat{a}_{j\ell}\{B_{22}(t_{i\ell})\underset{\sim}{v}_{i\ell} + \underset{\sim}{g}_2(t_{i\ell})\} + \frac{1}{h_i}\underset{\sim}{S}_{ij} \tag{3.24}$$

where $\varepsilon/h_i\ \underset{\sim}{R}_{ij}$ and $1/h_i\ \underset{\sim}{S}_{ij}$ consist of linear operators acting on $\underset{\sim}{u}_{i\ell}$, $\underset{\sim}{v}_{i\ell}$, $\ell=1,\ldots,k$ for Lobatto points and $\ell=1,\ldots,k+1$ for Gauss points, and inhomogeneities. We now show that their norms are $O(h_i)$ and so they can be dealt with as small perturbations when h_i is small.

Lemma 3.1: For each i, $1 \le i \le N$,

$$\begin{pmatrix} \varepsilon/h_i\ \underset{\sim}{R}_{ij} \\ 1/h_i\ \underset{\sim}{S}_{ij} \end{pmatrix} = h_i\{\underset{\sim}{\phi}_{ij}[\underset{\sim}{u}_{i1},\ldots,\underset{\sim}{u}_{iq};\ \underset{\sim}{v}_{i1},\ldots,\underset{\sim}{v}_{iq}] + \underset{\sim}{\psi}_{ij}[\underset{\sim}{g}_1(t_{i1}),\ldots,\underset{\sim}{g}_1(t_{iq}),\ \underset{\sim}{g}_2(t_{i1}),\ldots,\underset{\sim}{g}_2(t_{iq})]\} \tag{3.25}$$

where $\underset{\sim}{\phi}_{ij}$, $\underset{\sim}{\psi}_{ij}$ are bounded linear operators,

$$||\underset{\sim}{\phi}_{ij}||,\ ||\underset{\sim}{\psi}_{ij}|| \le c, \qquad 1 \le i \le N,\ 1 \le j \le q$$

with q = k for Lobatto points, q = k+1 for Gauss points. □

Those readers who wish to skip the proof of this lemma can do so without loss of continuity.

Proof: Writing $\underset{\sim}{u}'_\Delta(t)$ and $\underset{\sim}{v}'_\Delta(t)$ in terms of their polynomial interplants of order k on $[t_i, t_{i+1}]$ we get

$$\underset{\sim}{u}'_\Delta(t) = \sum_{j=1}^{k} \underset{\sim}{u}'_\Delta(t_{ij}) L_j\left(\frac{t-t_i}{h_i}\right) + \frac{1}{k!}\, \underset{\sim}{u}_\Delta^{(k+1)}(\zeta_t) \prod_{j=1}^{k} (t-t_{ij}) \qquad t_i \le \zeta_t \le t_{i+1}$$

where L_j are the Lagrange polynomials. Integrating,

$$\underset{\sim}{u}_{ij} - \underset{\sim}{u}_i = h_i \sum_{\ell=1}^{k} \underset{\sim}{u}'_\Delta(t_{i\ell})\hat{a}_{j\ell} + \frac{1}{k!}\int_{t_i}^{t_{ij}} \underset{\sim}{u}_\Delta^{(k+1)}(\zeta_t) \prod_{\ell=1}^{k} (t-t_{i\ell})dt$$

and so, by (3.23),

$$\underset{\sim}{R}_{ij} = \frac{1}{k!}\int_{t_i}^{t_{ij}} \underset{\sim}{u}_\Delta^{(k+1)}(\zeta_t) \prod_{\ell=1}^{k} (t-t_{i\ell})dt \tag{3.26}$$

with a similar expression for $\underset{\sim}{S}_{ij}$, $\underset{\sim}{v}_\Delta$ replacing $\underset{\sim}{u}_\Delta$.

Next, since $\underset{\sim}{y}_\Delta$ and $\underset{\sim}{z}_\Delta$ are polynomials of degree at most k on $[t_i, t_{i+1}]$, by the transformation (3.22),

$$\underset{\sim}{u}_\Delta^{(k+1)}(\tau) = \sum_{\nu=1}^{k} \binom{k+1}{\nu} (E^{-1}(\tau))^{(\nu)} \underset{\sim}{y}_\Delta^{(k+1-\nu)}(\tau) \qquad t_i \le \tau \le t_{i+1} \tag{3.27}$$

$$\underset{\sim}{v}_\Delta^{(k+1)}(\tau) = -\varepsilon \sum_{\nu=1}^{k} \binom{k+1}{\nu} L^{(\nu)}(\tau) \underset{\sim}{y}_\Delta^{(k+1-\nu)}(\tau) \tag{3.28}$$

and

$$\varepsilon \underset{\sim}{y}_\Delta^{(\ell)}(\tau) = \varepsilon \frac{d^{\ell-1}}{d\tau^{\ell-1}}(\underset{\sim}{y}'_\Delta(\tau)) = \sum_{j=1}^{k} \varepsilon \underset{\sim}{y}'_\Delta(t_{ij}) \frac{d^{\ell-1}}{d\tau^{\ell-1}} L_j\left(\frac{\tau-t_i}{h_i}\right). \tag{3.29}$$

Replacing the vectors $\varepsilon \underset{\sim}{y}'_\Delta(t_{ij})$ through the collocation equations (3.9) and

the transformation (3.22), and substituting into (3.27), (3.28), we obtain

$$(3.30) \qquad \|\underset{\sim}{u}^{(k+1)}\| \le h_i\phi/\varepsilon, \ \|\underset{\sim}{v}_{\Delta}^{(k+1)}\| \le h_i\phi$$

$$(3.31) \qquad \phi = ch_i^{-k} \max\{\|\underset{\sim}{u}_{ij}\|, \|\underset{\sim}{v}_{ij}\|, \|\underset{\sim}{g}_1(t_{ij})\|, \|\underset{\sim}{g}_2(t_{ij})\|, \ 1 \le j \le q\}$$

Finally, substituting (3.30), (3.31) into (3.26) and the corresponding expression for $\underset{\sim}{S}_{ij}$, the desired result (3.25) is obtained.

QED.

3.3 The mesh and the decomposition of numerical solutions

The meshes considered in this paper have the following structure. Near the boundaries, the step sizes h_i are comparable to ε. Specifically, there are given numbers $0 < N_0, N_1 < N$ and constants K_0, K_1, such that

$$(3.32) \qquad \begin{aligned} h_i &\le K_0\varepsilon, \quad i = 1,\ldots,N_0 \\ h_i &\le K_1\varepsilon, \quad i = N-N_1+1,\ldots,N. \end{aligned}$$

In between, much larger step sizes may be used, i.e. $h_i >> \varepsilon$, $i = N_0+1,\ldots,N-N_1$. We will assume for convenience of notation that h, the largest step size, occurs away from the boundaries. Such a mesh is depicted in figure 1 below.

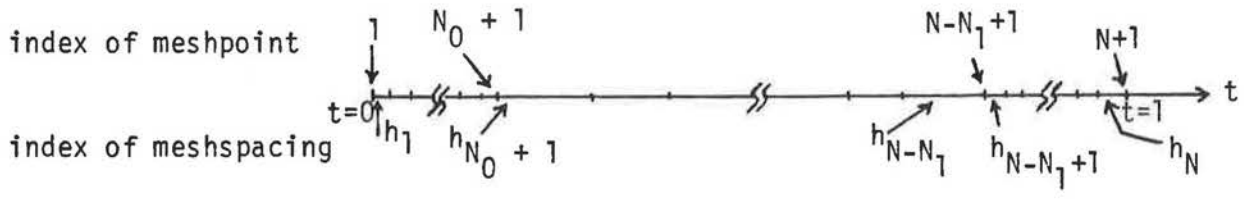

Figure 1: The mesh

For brevity of notation we set $\underline{i} = N_0 + 1$, $\bar{i} = N-N_1 + 1$ and write

(3.33) $\quad t_{\underline{i}} = T_0\varepsilon\ , \quad t_{\bar{i}} = 1 - T_1\varepsilon$

Our next step is to write down a decomposition representation to the discrete solution of (3.23), (3.24), similar to the representation (2.14) for the analytic solution. Moreover, we write down such a representation for each of the three parts of the mesh.

We write the system (3.23), (3.24), in analogy to (2.8) as

(3.34a) $\quad (H\underset{\sim}{w}_\Delta)(t_{ij}) = \underset{\sim}{g}(t_{ij}) \qquad j = 1,\ldots,q,\ i = 1,\ldots,N$

or in shorthand as

(3.34b) $\quad H_\Delta \underset{\sim}{w}_\Delta = \underset{\sim}{g}_\Delta,$

where $\underset{\sim}{w}_\Delta(t) = \begin{pmatrix} \underset{\sim}{u}_\Delta(t) \\ \underset{\sim}{v}_\Delta(t) \end{pmatrix}$ is in the class S_Δ of functions defined by (3.22) with $\underset{\sim}{y}_\Delta(t)$, $\underset{\sim}{z}_\Delta(t)$ continuous piecewise polynomial vector functions of degree at most k. Let

(3.35) $\quad W^M_{\Delta s}(t) = \begin{pmatrix} U^M_{\Delta s}(t) \\ V^M_{\Delta s}(t) \end{pmatrix} \in \mathbb{R}^{(n+m)\times n_s}\ , \quad t\in[t_{M_0},\ t_{M_1}]$

be matrix valued functions with columns in class S_Δ. Here s stands for -, + or 0, $n_0 := m$, M stands for I, II or III, to denote the three mesh regions considered, and so $I_0 = 1$, $I_1 = \underline{i} = II_0$, $II_1 = \bar{i} = III_0$, $III_1 = N+1$.

<u>On the interval $[0,\ T_0\varepsilon]$</u>: Define $W^I_{\Delta s}$, $\underset{\sim}{w}^I_{p\Delta} \in S_\Delta$ as follows (omitting the superscript I):

(3.36a) $\quad H_\Delta W_{\Delta -} = 0,\ P_- U_{\Delta -}(0) = I,\ P_+ U_{\Delta -}(T_0\varepsilon) = 0,\ V_{\Delta -}(0) = 0;$

(3.36b) $\quad H_\Delta W_{\Delta +} = 0,\ P_- U_{\Delta +}(0) = 0,\ P_+ U_{\Delta +}(T_0\varepsilon) = I,\ V_{\Delta +}(0) = 0;$

(3.36c) $\quad H_\Delta W_{\Delta 0} = 0,\ P_- U_{\Delta 0}(0) = S_-(\varepsilon),\ P_+ U_{\Delta 0}(T_0\varepsilon) = P_+ U_0(T_0\varepsilon),\ V_{\Delta 0}(0) = I;$

(3.36d) $\quad H_\Delta \underset{\sim}{w}_{P\Delta} = \underset{\sim}{g}_\Delta,\ P_- \underset{\sim}{u}_{P\Delta}(0) = P_- \underset{\sim}{u}_P(0),\ P_+ \underset{\sim}{u}_{P\Delta}(T_0\varepsilon) = P_+ \underset{\sim}{u}_P(T_0\varepsilon),$

$$\underset{\sim}{v}_{P\Delta}(0) = \underset{\sim}{v}_P(0)$$

The general solution of (3.34) on the left layer mesh is written formally as:

(3.37) $\underset{\sim}{w}^I_\Delta = W^I_{\Delta-}\,\underset{\sim}{\zeta}^I_- + W^I_{\Delta+}\,\underset{\sim}{\zeta}^I_+ + W^I_0\,\underset{\sim}{\zeta}^I_0 + \underset{\sim}{w}^I_{P\Delta}$

with $\underset{\sim}{\zeta}^I_- \in \mathbb{R}^{n-}$, $\underset{\sim}{\zeta}^I_+ \in \mathbb{R}^{n+}$, $\underset{\sim}{\zeta}_0 \in \mathbb{R}^m$.

On the interval $[1 - T_1\varepsilon, 1]$: In precise analogy to the above, define $W^{III}_{\Delta s}$ and $\underset{\sim}{w}^{III}_{P\Delta}$, again omitting superscripts:

(3.38a) $H_\Delta W_{\Delta-} = 0,\ P_- U_{\Delta-}(1-T_1\varepsilon) = I,\ P_+ U_{\Delta-}(1) = 0,\ V_{\Delta-}(1-T_1\varepsilon) = 0;$

(3.38b) $H_\Delta W_{\Delta+} = 0,\ P_- U_{\Delta+}(1-T_1\varepsilon) = 0,\ P_+ U_{\Delta+}(1) = I,\ V_{\Delta+}(1-T_1\varepsilon) = 0;$

(3.38c) $H_\Delta W_{\Delta 0} = 0,\ P_- U_{\Delta 0}(1-T_1\varepsilon) = P_- U_0(1-T_1\varepsilon),\ P_+ U_{\Delta 0}(1) = S_+(\varepsilon),$
$V_{\Delta 0}(1-T_1\varepsilon) = 0$

(3.38d) $H_\Delta \underset{\sim}{w}_{P\Delta} = \underset{\sim}{g}_\Delta,\ P_- \underset{\sim}{u}_{P\Delta}(1-T_1\varepsilon) = P_- \underset{\sim}{u}_P(1-T_1\varepsilon),\ P_+ \underset{\sim}{u}_{P\Delta}(1) = P_+ \underset{\sim}{u}_P(1),$
$\underset{\sim}{v}_{P\Delta}(1-T_1\varepsilon) = \underset{\sim}{v}_P(1-T_1\varepsilon)$

The general solution of (3.34) on the right layer mesh is written formally as

(3.39) $\underset{\sim}{w}^{III}_\Delta = W^{III}_{\Delta-}\,\underset{\sim}{\zeta}^{III} + W^{III}_{\Delta+}\,\underset{\sim}{\zeta}^{III}_+ + W^{III}_0\,\underset{\sim}{\zeta}^{III}_0 + \underset{\sim}{w}^{III}_{P\Delta}$

with $\underset{\sim}{\zeta}^{III}_- \in \mathbb{R}^{n-}$, $\underset{\sim}{\zeta}^{III}_+ \in \mathbb{R}^{n+}$, $\underset{\sim}{\zeta}_0 \in \mathbb{R}^m$.

On the interval $[T_0\varepsilon, 1-T_1\varepsilon]$: Define W^{II}_Δ, $W^{II}_{\Delta 0}$ and $\underset{\sim}{w}^{II}_{P\Delta}$ as follows: Let $Y_0(t)$, $Z_0(t)$ be obtained from $U_0(t)$, $V_0(t)$, via the (inverse) transformation (2.5). Then W^{II}_Δ and $W^{II}_{\Delta 0}$ are obtained via the transformation (3.22) from $X_\Delta = \begin{pmatrix} Y_\Delta \\ Z_\Delta \end{pmatrix}$ and $X_{\Delta 0} = \begin{pmatrix} Y_{\Delta 0} \\ Z_{\Delta 0} \end{pmatrix}$, respectively, which are defined as follows:

(3.40a) X_Δ and $X_{\Delta 0}$ satisfy the homogeneous equations (3.8), (3.9) with $\underset{\sim}{f} = \underset{\sim}{0}$.

(3.40b) $Y_\Delta(T_0\varepsilon) = I,\ Z_\Delta(T_0\varepsilon) = 0;$

(3.40c) $Y_{\Delta 0}(T_0\varepsilon) = Y_0(T_0\varepsilon),\ Z_{\Delta 0}(T_0\varepsilon) = Z_0(T_0\varepsilon).$

The particular solution $\underset{\sim}{w}^{II}_{P\Delta}$ is defined, e.g., by

(3.40d) $H_\Delta \underset{\sim}{w}_{P\Delta} = \underset{\sim}{g}_\Delta,\ w_{P\Delta}(T_0\varepsilon) = \underset{\sim}{w}_P(T_0\varepsilon)$

The general solution of (3.34) on the long interval away from the layers is written formally as

(3.41) $$\underset{\sim}{w}_{\Delta}^{II} = W_{\Delta}^{II}\underset{\sim}{\zeta}^{II} + W_{\Delta 0}^{II}\underset{\sim}{\zeta}_{0}^{II} + \underset{\sim}{w}_{P\Delta}^{II}$$

with $\underset{\sim}{\zeta}^{II} \in \mathbb{R}^n$, $\underset{\sim}{\zeta}_0 \in \mathbb{R}^m$.

3.4 Convergence results

Below we state the various results regarding the convergence of the numerical methods, culminating in theorem 3.3 The proofs for those results which have not been proven elsewhere are contained in the next two sections. Denote by p the "regular" superconvergence order of the schemes under consideration, i.e.

(3.42) $p = 2k$ for a k-stage Gauss scheme

$= 2(k-1)$ for a k-stage Lobatto scheme.

Also, define the seminorms on collocation solutions,

(3.43a) $$||\psi_{\Delta}||_{\Delta} := \max\{||\psi_i||;\quad 1 \le i \le N+1\}$$

(3.43b) $$||\psi_{\Delta}||_{c} := \max\{||\psi_{ij}||;\quad 1 \le i \le N,\ 1 \le j \le k\}$$

where the vector norms used are maximum norms. Thus $||\underset{\sim}{\psi}_{\Delta}^{c}|| = \max\{||\psi_{\Delta}||_{\Delta}, ||\psi_{\Delta}||_{c}\}$. Also $||\psi_{\Delta}||_{\Delta}^{M}$, $||\psi_{\Delta}||_{c}^{M}$ will denote the seminorm where the range of i in (3.43) is restricted to $M_0 \le i \le M_1$, M = I, II or III. For a matrix whose columns are collocation solutions, a maximum on the column norms is taken.

For the "short" intervals $[0, t_{\underline{i}}]$ and $[t_{\bar{i}}, 1]$ we have

Theorem 3.1

(a) The solution representations (3.37) and (3.39) are valid (i.e. their components can be computed in a stable way)

(b) With h_L the maximum step size in the layers ($h_L \le \varepsilon \max\{K_0, K_1\}$ by (3.32)), the "smooth" components satisfy

(3.44) $$||W_{\Delta 0} - W_0||^I_\Delta,\ ||W_{\Delta 0} - W_0||^{III}_\Delta,\ ||\underset{\sim}{w}_{P\Delta} - \underset{\sim}{w}_P||^I_\Delta,$$
$$||\underset{\sim}{w}_{P\Delta} - \underset{\sim}{w}_P||^{III}_\Delta \le ch_L^P$$

(c) The auxiliary solution components in the layers (for which there are no counterparts in the exact solution decomposition) satisfy

(3.45) $$W^I_{\Delta+}(T_0\varepsilon) = \begin{pmatrix} 0_{n_-xn_+} \\ I_{n_+xn_+} \\ 0 \end{pmatrix} + O(\varepsilon),\quad W^{III}_{\Delta-}(1 - T_1\varepsilon) = \begin{pmatrix} I_{n_-xn_-} \\ 0 \end{pmatrix} + O(\varepsilon)$$

(d) For a given accuracy tolerance δ, $\delta \ge c\varepsilon$, the layer meshes can be constructed as follows: With

$\mu := \max\{|\lambda_j(0)|,\ j=1,\dots,n_-\}$, $\nu := \min\{-re(\lambda_j(0)),\ j=1,\dots,n_-\} > 0$, define

(3.46) $$h_1 := \frac{\varepsilon}{\mu}\left[\frac{\nu}{\mu|c_\gamma|}\right]^{1/P}\delta^{1/P}$$

(3.47) $$h_i := h_{i-1}\exp\{\tfrac{1}{p}\ \tfrac{\nu}{\varepsilon}\ h_{i-1}\} \quad \text{until } t_{i+1} \ge T_0\varepsilon$$

where c_γ is a known constant (cf. Part I) and

(3.48) $$T_0 := \nu^{-1}|\ln\delta|.$$

The right end layer is constructed in an analogous way, depending on $\lambda_j(1)$, $j = n-n_-+1,\dots,n$. Then

(3.49a) $$W^I_{\Delta-}(t_i) = \begin{pmatrix} \exp(\Lambda_-(0)t_i/\varepsilon) \\ 0 \end{pmatrix} + O(\delta) \qquad 1 \le i \le \underline{i}$$

(3.49b) $$W^{III}_{\Delta+}(t_i) = \begin{pmatrix} 0 \\ \exp(\Lambda_+(1)(t_i-1)/\varepsilon) \end{pmatrix} + O(\delta) \qquad \bar{i} \le i \le N+1$$

The proof of this theorem is given in §4. We note that the assumption $\delta \ge c\varepsilon$ is not essential, see §6.

For the "long" interval $[t_{\underline{i}}, t_{\bar{i}}]$ we have

Theorem 3.2 Let

(3.50) $$\kappa := \varepsilon\underline{h}^{-1}, \qquad \underline{h} := \min\{h_i\ ; \qquad \underline{i} \le i \le \bar{i}\ \}$$

(a) The solution representation (3.41) is valid for ε sufficiently small such that $\kappa < c$ for Lobatto points or $\varepsilon \sum_{i=\underline{i}}^{\bar{i}} h_i^{-1} < c$ for Gauss points, where c is a constant of order 1.

(b) The first n fundamental solution components satisfy

$$(3.51a) \qquad W_\Delta^{II}(t_i) = (-1)^{k(i-\underline{i})} \begin{pmatrix} E^{-1}(t_i) \\ 0 \end{pmatrix} + O(\kappa) \qquad \underline{i} \le i \le \bar{i}$$

for Gauss points and

$$(3.51b) \qquad W_\Delta^{II}(t_i) = (-1)^{(k+1)(i-\underline{i})} \begin{pmatrix} \Lambda^{-1}(t_i)E^{-1}(t_i)A_{11}(t_{\underline{i}}) \\ 0 \end{pmatrix} + O(\kappa) \qquad \underline{i} \le i \le \bar{i}$$

for Lobatto points.

(c) Define the error e as follows:

For a k-stage Gauss scheme, $e := h^k$ and, if k is odd and the mesh is locally almost uniform, i.e.,

$$(3.52) \qquad h_{i+1} = h_i(1 + O(h_i)) \text{ for all i odd } \underline{\text{or}} \text{ all i even,}$$

then $e = h^{k+1}$.

For a k-stage Lobatto scheme $e := Kh^P + \varepsilon h^{k-1}$, and, if k is even and the mesh is locally almost uniform then $e = Kh^P + \varepsilon h^k$. Also, if the slow components $\underset{\sim}{z}$ are absent from (1.1), (1.3) then $K = 0$.

Then we have

$$(3.53) \qquad ||\underset{\sim}{w}_{P\Delta}^{II} - \underset{\sim}{w}_P||_\Delta^{II},\ ||W_{\Delta 0}^{II} - W_0||_\Delta^{II} \le ce.$$

The proof of this theorem is given in §5.

The condition on ε and the mesh for Gauss points in (a) is slightly annoying. However, it can be argued that for $O(\kappa)$ perturbations to be small, say $O(h)$, this condition has to be satisfied anyway.

The central theorem, summarizing our efforts for linear problems with variable coefficients, follows:

Theorem 3.3: Assume that the boundary value problem (1.1) - (1.3) is well posed, uniformly for $0 < \varepsilon \le \varepsilon_0$, and denote the solution by $\underset{\sim}{x}(t) = \begin{pmatrix} \underset{\sim}{y}(t) \\ \underset{\sim}{z}(t) \end{pmatrix}$. Assume further that

$$\det \begin{pmatrix} P_- E^{-1}(0) \\ P_+ E^{-1}(1) \end{pmatrix} \neq 0 . \tag{3.54}$$

Then for $0 < \varepsilon \le \varepsilon_1$ there are positive constants δ_0, h_0 and κ_0 such that for any δ, $0 < c_0\varepsilon \le \delta \le \delta_0$ and any mesh $\Delta = \Delta(\varepsilon)$ satisfying (3.46), (3.47), similar conditions at the right end layer, $\kappa \le \kappa_0$ and $h \le h_0$, the k-stage collocation scheme based on Gauss or Lobatto points has a unique solution $\underset{\sim}{x}_\Delta(t)$ which satisfies

$$\| \underset{\sim}{x}_\Delta - \underset{\sim}{x} \|_\Delta \le c(e + \delta) \tag{3.55}$$

where e is defined in part (c) of theorem 3.2.

The proof of theorem 3.3 is essentially identical to that of theorem 5.3 in Weiss [9], so we only give an outline here.

Proof outline for theorem 3.3

The basic task is to patch up the solution representations (3.37), (3.41) and (3.39) on the various segments of the mesh. The problem in transformed variables (3.23), (3.24), (or (3.34) for short) is considered, first under the boundary conditions

$$P_- \underset{\sim}{u}_1 = \underset{\sim}{\eta}_- \in \mathbb{R}^{n-}, \quad P_+ \underset{\sim}{u}_{N+1} = \underset{\sim}{\eta}_+ \in \mathbb{R}^{n+}, \quad \underset{\sim}{v}_1 = \underset{\sim}{\eta}_0 \in \mathbb{R}^m . \tag{3.56}$$

This corresponds to the differential problem (2.8), (2.10) which theorem 2.1 guarantees to be well-behaved for any given parameter vectors $\underset{\sim}{\eta}_-$, $\underset{\sim}{\eta}_+$ and $\underset{\sim}{\eta}_0$.

Thus, the 3(n+m) components of the parametric representations (3.37), (3.41) and (3.39), i.e. of $\underset{\sim}{\zeta}^\Delta := (\underset{\sim}{\zeta}^I_-, \underset{\sim}{\zeta}^I_+, \underset{\sim}{\zeta}^I_0, \underset{\sim}{\zeta}^{II}, \underset{\sim}{\zeta}^{II}_0, \underset{\sim}{\zeta}^{III}_-, \underset{\sim}{\zeta}^{III}_+, \underset{\sim}{\zeta}^{III}_0)$,

are fixed by the 3(n+m) linear equations consisting of (3.56) plus the matching conditions

$$(3.57) \qquad \underset{\sim}{w}_{\Delta}^{I}(t_{\underline{i}}) = \underset{\sim}{w}_{\Delta}^{II}(t_{\underline{i}}) \ , \ \underset{\sim}{w}_{\Delta}^{II}(t_{\bar{i}}) = \underset{\sim}{w}_{\Delta}^{III}(t_{\bar{i}}).$$

In analogy to (2.11), the resulting 3(n+m) x 3(n+m) constraint matrix should have a uniformly bounded inverse for δ, h and κ sufficiently small. Theorems 3.1 and 3.2 furnish us with information on the structure of key blocks of this matrix in (3.45), (3.49) and (3.51). Examining the resulting structure, it becomes apparent that the principal part of the constraint matrix is nonsingular iff the matrix

$$\begin{pmatrix} P_{-}E^{-1}(t_{\underline{i}}) \\ P_{+}E^{-1}(t_{\bar{i}}) \end{pmatrix}$$

is nonsingular. The condition for the latter to hold uniformly in ε is (3.54).

Now, theorem 2.3 guarantees that the exact solution $\underset{\sim}{w}(t)$ has a similar decomposition and by (2.14) its parameter vector corresponding to $\underset{\sim}{\zeta}^{\Delta}$, which is determined so as to satisfy (2.10), can be written as

$$\underset{\sim}{\zeta} = (\underset{\sim}{\zeta}_{-}, \underset{\sim}{0}, \underset{\sim}{\zeta}_{0}, \underset{\sim}{0}, \underset{\sim}{\zeta}_{0}, \underset{\sim}{0}, \underset{\sim}{\zeta}_{+}, \underset{\sim}{\zeta}_{0}).$$

The stability of the constraint matrix plus the convergence results of theorems 3.1 and 3.2 imply that

$$(3.58) \qquad ||\underset{\sim}{\zeta} - \underset{\sim}{\zeta}^{\Delta}|| \leq c(e + \delta)$$

Finally, the well-posedness of the problem guarantees, as described in §2, the safe transformation back to the $\underset{\sim}{x}(t)$ variables with the choice of bounded $\underset{\sim}{\eta}_{+}$, $\underset{\sim}{\eta}_{-}$, $\underset{\sim}{\eta}_{0}$ to satisfy the original boundary conditions. This completes the proof outline of theorem 3.3.

In practice, the transformation (3.22) or (2.5) is not needed and the difference equations are solved in the original variables. Thus, the

matrices and vectors Γ_i, $\underset{\sim}{g}_i$, of (3.14) are assembled according to (3.15) or (3.21) and the linear system of (n+m)(N+1) equations

$$(3.59)\qquad \begin{bmatrix} -\Gamma_1 & I & & & & \\ & -\Gamma_2 & I & & & \\ & & \ddots & \ddots & & \\ & & & & -\Gamma_N & I \\ B_0 & & & & & B_1 \end{bmatrix} \begin{bmatrix} \underset{\sim}{x}_1 \\ \underset{\sim}{x}_2 \\ \vdots \\ \underset{\sim}{x}_N \\ \underset{\sim}{x}_{N+1} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{g}_1 \\ \underset{\sim}{g}_2 \\ \vdots \\ \underset{\sim}{g}_N \\ \underset{\sim}{\beta} \end{bmatrix}$$

is solved for the superconvergent mesh values. The structure of the matrix in (3.59) is the familiar one for the trapezoidal or midpoint (box) scheme and is independent of k. The condition number of the matrix is a modest O(N) and in particular is independent of ε (cf. theorem 6.2 of Part I).

Indeed, it is a good practice in actual computation to roughly estimate the condition number of the above matrix for two values of ε, say. If that condition number seems to get large as ε decreases then something is "wrong": The mesh may be inadequate, or (3.54) does not hold or, perhaps most commonly, the differential problem is not well posed uniformly in ε. How to deal with the latter two cases will be discussed in a subsequent paper.

The meshes under consideration, which are described in fig. 1 and (3.32), (3.33), make sense from the simple point of view of approximating the solution profile independently of the differential equations. Thus, small subintervals are used where the solution varies fast and much larger subintervals are used elsewhere. However, the simple approximation point of view of the mesh may be somewhat misleading in the case where there is

no boundary layer, say at t = 0, even though $n_- > 0$, because the reduced solution happens to hit the boundary at the prescribed value(s) of $\underset{\sim}{\beta}$. In this case caution should be exercised, as the following discussion indicates.

Suppose for simplicity that the reduced solution, obtained by setting $\varepsilon = 0$ in (1.1) - (1.3), satisfies (1.3), and that a uniform mesh with step size $h >> \varepsilon$, adequate to approximate the smooth solution well, is used in (3.5), (3.6). In §5 we show that setting $\varepsilon = 0$ we get expressions (5.22) for Gauss points and (5.51) for Lobatto points. Here $\underline{i} = 1$, $\bar{i} = N+1$, and we obtain

$$\underset{\sim}{\bar{y}}_{N+1} = (-1)^{kN}\, \underset{\sim}{\bar{y}}_1 + \underset{\sim}{\phi} \tag{3.60a}$$

for Gauss points and

$$\underset{\sim}{\bar{y}}_{N+1} = (-1)^{(k+1)N}\, A_{11}^{-1}(1)A_{11}(0)\underset{\sim}{\bar{y}}_1 + \underset{\sim}{\phi} \tag{3.60b}$$

for Lobatto points, where $\underset{\sim}{\phi}$ is some inhomogeneity. Splitting the boundary condition matrices as

$$B_0 = [B_0^y;\ B_0^z]\ , \quad B_1 = [B_1^y;\ B_1^z] \tag{3.61}$$

and substituting into the boundary conditions (3.16) gives the n+m equations

$$B_y\, \underset{\sim}{\bar{y}}_1 + B_z\, \underset{\sim}{\bar{z}}_1 = \underset{\sim}{\hat{\beta}} \tag{3.62}$$

where B_z and $\underset{\sim}{\hat{\beta}}$ are appropriate quantities of no interest and

$$B_y = \begin{cases} B_0^y + (-1)^{kN} B_1^y & \text{for Gauss points} \\ B_0^y + (-1)^{(k+1)N} B_1^y A_{11}^{-1}(1)A_{11}(0) & \text{for Lobatto points} \end{cases} \tag{3.63}$$

Now, clearly a necessary condition for (3.62) to have a unique solution is

$$\text{rank}\ (B_y) = n \tag{3.64}$$

This turns out not to be the case for some very simple examples, e.g., the

example in §6 of Part I.

If (3.64) holds then the uniform mesh is adequate for this problem with a smooth solution. The condition number of the matrix in (3.59) is then bounded independently of ε. However, if (3.64) does not hold then the reduced problem is singular. The condition number of the matrix in (3.59) is then, at best, $O(h/\varepsilon)$, and roundoff errors of this size are introduced in the course of computation.

There is a very simple remedy to this situation: Add to the mesh <u>one point</u> $O(\varepsilon)$ away from each boundary where there is no layer (despite the existence of eigenvalues of the corresponding sign in A_{11}). This one point layer mesh is sufficient for theorem 3.3 to hold: The more elaborate layer mesh of (3.46), (3.47) is constructed for accuracy reasons and therefore is not needed here.

4. Boundary layer regions

In this section we consider the linear problem (1.1), (1.2) on the subinterval $[0, T_0\varepsilon]$, where a fine mesh satisfying (3.32) is assumed. Analogous results hold for the subinterval $[1 - T_1\varepsilon, 1]$. Let

(4.1) $\quad h_L := \max\{h_i,\ 1 \le i \le N_0\}.$

Following Weiss [9] we consider the transformed system (2.6) - (2.7) for an easier analysis.

4.1 Stability and results for smooth and auxiliary solution components

First, consider one equation

(4.2) $\quad \varepsilon y' = \lambda y + f \qquad 0 \le t \le T_0\varepsilon$

with $\mathrm{re}(\lambda(t)) \le -\bar{\lambda} < 0$ and $\lambda(t)$ is piecewise constant: $\lambda(t) = \lambda(t_i)$, $t_i \le t < t_{i+1}$. From (3.5) we get for Gauss points

(4.3) $\quad y_{ij} = y_i + \frac{h_i}{\varepsilon} \sum_{\ell=1}^{k} \hat{a}_{j\ell}[\lambda(t_i)y_{i\ell} + f(t_{i\ell})] \qquad 1 \le j \le k$

So, eliminating the local unknowns y_{ij} and substituting into the corresponding expression for y_{i+1}, we get (see Part 1, §4)

(4.4) $\quad y_{i+1} = \gamma(\frac{\lambda(t_i)h_i}{\varepsilon})y_i + \frac{h_i}{\varepsilon}\,\hat{b}^T[(\frac{\varepsilon}{\lambda(t_i)h_i} - \hat{A})^{-1}\hat{A} + I]f_i$

where $f_i = (f(t_{i1}),\ldots,f(t_{ik}))^T$, $1 = (1,\ldots,1)^T \in \mathbb{R}^k$, and

(4.5) $\quad \gamma(\zeta) = 1 + \hat{b}^T(\zeta^{-1}I - \hat{A})^{-1}1$

is the growth factor. The matrix $\zeta^{-1}I - \hat{A}$ is nonsingular for all ζ satisfying $\mathrm{re}(\zeta) < 0$.

Solving the recurrence relation (4.4) we get

(4.6) $\quad y_{i+1} = [\prod_{\ell=1}^{i}\gamma(\frac{\lambda(t_\ell)h_\ell}{\varepsilon})]y_1 + \frac{1}{\varepsilon}\sum_{j=0}^{i-1}[\prod_{\ell=i-j+1}^{i}\gamma(\frac{\lambda(t_\ell)h_\ell}{\varepsilon})]h_{i-j}\xi_{i-j}^T f_{i-j}$

where $\xi_{i-j}^T := \hat{b}^T[\frac{\varepsilon}{\lambda(t_i)h_i} - \hat{A})^{-1}\hat{A} + I]$ is a bounded vector by (3.32). Now,

since the method is A-stable we have

(4.7a) $\quad |\gamma(\zeta)| \le 1 \qquad re(\zeta) < 0.$

Furthermore, since $\gamma'(0) = 1$ it follows that for any set S of the form

$$S \equiv S(\alpha_1,\alpha_2,\beta) = \{\zeta \mid 0 < |\zeta| \le \beta, \ \frac{\pi}{2} + \alpha_1 \le \arg\zeta \le \frac{3\pi}{2} - \alpha_2\}$$

with $\alpha_1,\alpha_2 > 0$, $\alpha_1 + \alpha_2 \le \pi$, $\beta < \infty$, there is a positive constant $\mu = \mu(\alpha_1, \alpha_2, \beta)$ such that

(4.7b) $\quad |\gamma(\zeta)| \le e^{\mu re(\zeta)}, \qquad \zeta \in S$

By (3.32),

$$|\frac{\lambda(t_\ell)h_\ell}{\varepsilon}| \le |\lambda(t_\ell)|K_0 \le \beta$$

for some well defined constant β of moderate size. Using (4.7b) we get

$$|\prod_{\ell=i-j+1}^{i} \gamma(\frac{\lambda(t_\ell)h_\ell}{\varepsilon})| \le \prod_{\ell=i-j+1}^{i} e^{-\mu\bar{\lambda}h_\ell/\varepsilon} = e^{-\mu\bar{\lambda}(t_{i+1} - t_{i+1-j})/\varepsilon}$$

and

$$\left|\frac{1}{\varepsilon}\sum_{j=0}^{i-1}\left[\prod_{\ell=i-j+1}^{i}\gamma(\frac{\lambda(t_\ell)h_\ell}{\varepsilon})\right]h_{i-j}\right| \le \frac{1}{\varepsilon}\sum_{j=0}^{i-1} e^{-\mu\bar{\lambda}(t_{i+1}-t_{i+1-j})/\varepsilon} h_{i-j}$$

$$\le \frac{1}{\varepsilon}\int_{t_1}^{t_{i+1}} e^{-\mu\bar{\lambda}(t_{i+1}-s)/\varepsilon} ds \le \frac{1}{\mu\bar{\lambda}}$$

So, substituting in (4.6) we get

(4.8) $\quad |y_{i+1}| \le |y_1| + c||f^c||$

where

(4.9) $\quad c = (\bar{\lambda}\mu)^{-1} \max_{1\le j\le i} ||\xi_j||$

It is straightforward to show that a similar result is obtained also for the Lobatto points.

This is the desired stability result for one equation. Next, consider the differential system (2.6) - (2.7) and its corresponding collocation discretization (3.23) - (3.24).

Theorem 4.1 The difference equations (3.23), (3.24) subject to the boundary conditions

(4.10) $\quad P_- \underset{\sim}{u}_1 = \underset{\sim}{\eta}_- \in \mathbb{R}^{n_-}, \quad P_+ \underset{\sim}{u}_{\tilde{1}} = \underset{\sim}{\eta}_+ \in \mathbb{R}^{n_+}, \quad \underset{\sim}{v}_1 = \underset{\sim}{\eta}_0 \in \mathbb{R}^m$

have a unique solution $\underset{\sim}{u}^c_\Delta$, $\underset{\sim}{v}^c_\Delta$ which satisfies

(4.11) $\quad ||\underset{\sim}{u}^c_\Delta||, \ ||\underset{\sim}{v}^c_\Delta|| \le K\{||\underset{\sim}{\eta}_-|| + ||\underset{\sim}{\eta}_+|| + ||\underset{\sim}{\eta}_0|| + ||\underset{\sim}{g}^c||\}$

provided that ε is small enough, $0 < \varepsilon \le \varepsilon_0$.

ε_0 is sufficiently small to enable a contraction argument below and depends on the bounds in lemma 3.1 and on $\max_{0\le t\le T_0\varepsilon} ||\Lambda'(t)||$. (To recall, by $\underset{\sim}{u}^c$ we mean the restriction of $\underset{\sim}{u}_\Delta(t)$ to the mesh points plus the collocation points).

Proof: We consider the case for Gauss points; the case for Lobatto points is treated similarly. Our strategy is to consider first the simplified difference equations

(4.12) $\quad \frac{\varepsilon}{h_i}(\underset{\sim}{u}_{ij} - \underset{\sim}{u}_i) = \sum_{\ell=1}^{k} \hat{a}_{j\ell}\{\Lambda(t_i)\underset{\sim}{u}_{i\ell} + \underset{\sim}{f}_1(t_{i\ell})\}$

$1 \le j \le k+1$

(4.13) $\quad \frac{1}{h_i}(\underset{\sim}{v}_{ij} - \underset{\sim}{v}_i) = \sum_{\ell=1}^{k} \hat{a}_{j\ell}\{B_{22}(t_{i\ell})\underset{\sim}{v}_{i\ell} + \underset{\sim}{g}_2(t_{i\ell})\}$

where $\underset{\sim}{f}_1(t_{i\ell}) := \underset{\sim}{g}_1(t_{i\ell}) + B_{12}(t_{i\ell})\underset{\sim}{v}_{i\ell}$, and to treat the difference between (4.12) - (4.13) and (3.23) - (3.24) as a perturbation term of order h_L.

The components $\{\underset{\sim}{v}_{ij}\}$ in (4.13), (4.10) are now completely separated from the components $\{\underset{\sim}{u}_{ij}\}$. For $\underset{\sim}{v}_\Delta(t)$ the usual theory applies. This is a Runge-Kutta scheme for a non-stiff initial value problem, and certainly for ε small and h_L satisfying (3.32), $\underset{\sim}{v}_\Delta(t)$ exists and satisfies

(4.14) $\quad ||\underset{\sim}{v}^c|| \le c\{||\underset{\sim}{\eta}_0|| + ||\underset{\sim}{g}_2||\}$

Now, for (4.12) note that since $\Lambda(t)$ is diagonal, the vector system decouples into n scalar components. For each of the first n_- components we

can apply the estimate (4.8) directly, since $re(\lambda_j(t_i)) < 0$, $1 \le j \le n_-$. For the last n_+ components, $re(\lambda_j(t_i)) > 0$, and we have to reverse the direction of integration, from right to left. Thus, for such a component, (4.8) is changed to read

$$|y_i| \le |y_{\bar{i}}| + c||f^c|| \tag{4.15}$$

which is compatible with the end conditions (4.10). We obtain that the difference equations (4.12) subject to (4.10) possess a solution $\underset{\sim}{u}_\Delta$ satisfying

$$||\underset{\sim}{u}_\Delta||_\Delta \le ||\underset{\sim}{\eta}_-|| + ||\eta_+|| + c||\underset{\sim}{f}_1|| \le ||\underset{\sim}{\eta}_-|| + ||\underset{\sim}{\eta}_+|| + c_1\{||g_\Delta|| + ||\underset{\sim}{\eta}_0||\} \tag{4.16}$$

It is now easy to show a similar result for $\underset{\sim}{u}_{ij}$ by expressing them in terms of $\underset{\sim}{u}_i$ using (4.12).

This completes the treatment of the major part of the difference operator. Now, the equations (4.12) - (4.13) differ from (3.23) - (3.24) by terms of order h_i (or ε) only, and a standard perturbation argument completes the proof

Q.E.D.

Now, with the stability result (4.11) and the linearity of the problem, part (a) of theorem 3.1 easily follows. Next, consider the "smooth" components $W_{\Delta 0}(t)$ and $\underset{\sim}{w}_{P\Delta}(t)$. These correspond to the components in the exact solution decomposition which vary slowly across the boundary layer region. Substitution of $W_{\Delta 0} - W_0$ and $\underset{\sim}{w}_{P\Delta} - \underset{\sim}{w}_P$ into (4.11) immediately yields that

$$||W^c_{\Delta 0} - W^c_0||,\ ||\underset{\sim}{w}^c_{P\Delta} - \underset{\sim}{w}^c_P|| \le ch_L = 0(\varepsilon) \tag{4.17}$$

and this is really all we need. However, more can be obtained by applying the standard collocation analysis (Russell [7], Weiss [8]) to the original variables (i.e., analyzing the error in (3.5), (3.6)). After transforming

back to $\underset{\sim}{w}$, part (b) of theorem 3.1 is obtained.

Consider part (c) of theorem 3.1. We write

$$W^I_{\Delta+} = F + G, \qquad F = \begin{pmatrix} 0 \\ F_+ \\ 0 \end{pmatrix}, \qquad G = \begin{pmatrix} U^I_+ \\ G_+ \\ V^I_+ \end{pmatrix} \tag{4.18}$$

with F satisfying the homogeneous equations (4.12), (4.13), subject to

$$F(t_{\underset{\sim}{1}}) = \begin{pmatrix} 0 \\ I \\ 0 \end{pmatrix} \tag{4.19}$$

and G is the rest. The difference equations for F are again decoupled and so A-stability immediately implies that F is bounded. We now have to show that G is small. But comparing (3.36b) with what F satisfies, it is apparent that G satisfies the difference equations (3.23), (3.24), with inhomogeneous terms of size $O(\varepsilon + h_2) = O(\varepsilon)$ and under homogeneous boundary conditions as in (3.36b). Using stability, part (c) of theorem 3.1 is proven.

4.2. Mesh selection in the layer regions

In §4.1 we have shown parts (a) - (c) of theorem 3.1. Here we treat the dominant components of the solution decomposition, $W^I_{\Delta-}(t)$. Analogous results for $W^{III}_{\Delta+}(t)$ will be omitted.

First, consider one homogeneous equation with constant coefficients,

$$\varepsilon y' = \lambda y, \qquad y(0) = 1, \tag{4.20}$$

with the solution $y(t) = \exp(\frac{\lambda}{\varepsilon}t)$, and denote $\hat{\lambda} := -\mathrm{re}(\lambda) > 0$. In Part I it was shown that, given a tolerance $\delta < 1$, the following mesh generates an approximation accurate to within this tolerance,

$$h_1 := \frac{\varepsilon}{|\lambda|}\, c_p \delta^{1/p}, \quad c_p := \left[\frac{\hat{\lambda}}{|\lambda||c_\gamma|}\right]^{1/p} \tag{4.21}$$

$$h_i := h_{i-1} \exp\{\frac{1}{p}\frac{\hat{\lambda}}{\varepsilon} h_{i-1}\} = h_1 \exp\{\frac{1}{p}\frac{\lambda}{\varepsilon} t_i\}, \qquad i = 2,\ldots,N_0. \tag{4.22}$$

Here c_γ is a known constant depending only on p, and N_0 is determined so that $t_{N_0+1} \geq T_0\varepsilon > t_{N_0}$. Since we would like the contribution of the fast decaying solution to be below δ on the "long" interval $[t_{\underline{i}}, t_{\bar{i}}]$, it is natural to set T_0 so that

$$\exp(\frac{-\hat{\lambda}}{\varepsilon} T_0\varepsilon) = \delta$$

i.e.

$$T_0 = \frac{1}{\hat{\lambda}}(-\ln\delta) \tag{4.23}$$

Note that the mesh defined by (4.21), (4.22) satisfies (3.32) because its steps are monotonically increasing and $h_{\underline{i}} = c_{p/|\lambda|}\varepsilon$. Also, beyond $t_{\underline{i}}$ the mesh becomes much sparser, depending only upon the accuracy needs for the reduced solution. Thus, the magnitude of $|y(t_{\underline{i}})|$ is propagated essentially undamped by the numerical scheme outside the layer region.

Next we let λ in (4.20) vary as in (4.2), i.e., $\lambda(t) = \lambda(t_i)$, $t_i \leq t < t_{i+1}$. Then (cf. Part I, §4.2)

$$y_{i+1} = \prod_{j=1}^{i} \gamma(\frac{\lambda(t_j)h_j}{\varepsilon}) = \prod_{j=1}^{i} \gamma(\frac{\lambda(0)h_j}{\varepsilon}) + R_i \tag{4.24}$$

where

$$R_i = \prod_{j=1}^{i} \gamma(\frac{\lambda(0)h_j}{\varepsilon})[\sum_{j=1}^{i} (1 + O(\frac{t_j h_j}{\varepsilon})) - 1] \tag{4.25}$$

Lemma 4.1 The residue R_i satisfies

$$|R_i| \leq c\varepsilon \tag{4.26}$$

provided that $\varepsilon(\ln\delta)^2$ is bounded by a constant. Thus, the mesh (4.21), (4.22) with $\lambda(0)$ replacing λ guarantees an approximation error of $O(\delta) + O(\varepsilon)$.

Proof: Let $t_{i+1} \leq -c_1\varepsilon \ln\delta$ with $\delta = O(\varepsilon)$. As for (4.6) and the following arguments, we obtain

$$|\sum_{j=1}^{i} \gamma(\frac{\lambda(0)h_j}{\varepsilon})| \leq \exp(-\mu t_{i+1}/\varepsilon).$$

Writing

$$\exp(-\kappa t_j h_j/\varepsilon) \le 1 + O(t_j h_j/\varepsilon) \le \exp(\kappa t_j h_j/\varepsilon)$$

we have

$$\exp(-\kappa t_{i+1}^2/2\varepsilon) \le \prod_{j=1}^{i} (1 + O(t_j h_j/\varepsilon) \le \exp(\kappa t_{i+1}^2/2\varepsilon)$$

So, by (4.25)

$$|R_i| \le \exp(-\mu t_{i+1}/\varepsilon)[\exp(\kappa t_{i+1}^2/2\varepsilon) - 1]$$

Now, for $t_{i+1} = -c_1\varepsilon \ln\delta$, $\kappa t_{i+1}^2/2\varepsilon = O(\varepsilon(\ln\delta)^2)$ and so

$$\exp(ct_{i+1}^2/2\varepsilon) - 1 \le c_2 t_{i+1}^2/2\varepsilon$$

Also

$$\exp(-\mu t_{i+1}/\varepsilon) = c_3\delta$$

So

$$|R_i| \le C_4\varepsilon\delta(\ln\delta)^2 \le c\varepsilon$$

QED

Turning to the differential system (2.6), (2.7), we once again consider the difference equations (3.23), (3.24) as an $O(\varepsilon)$ perturbation of (4.12), (4.13). The homogeneity and boundary conditions of (3.36a) plus the decoupling of (4.13) from (4.12) clearly imply that $V_{\Delta^-}(t) \equiv 0$ and $P_{+\cup\Delta^-}(t) \equiv 0$. Also, for each of the first n_- components, the previous results for one equation apply provided that the mesh in (4.21), (4.22) is chosen according to it. Taking the most stringent of these choices will produce $O(\delta)$ accuracy for all components. This is clearly achieved by the choice (3.46), (3.47). Part (d) of theorem 3.1 is then proven and hence, the proof of theorem 3.1 is complete.

The practical importance of using the mesh (3.46), (3.47) instead of, say, a uniform mesh has been demonstrated in table 4.2 of Part I. We now

supplement this by some a priori estimates of N_0, the number of mesh points needed in the layer.

Theorem 4.2 Asymptotically, for ε and δ small,

(4.27) $\quad N_0 = O(\delta^{-1/p})$.

Note that in (4.27) N_0 is independent of ε. Also, a uniform mesh with T_0 given by (4.23) would yield $N_0 = O(\delta^{-1/p}(-\ln\delta))$.

Proof: It is sufficient to consider (4.21), (4.22). Then

$$N_0 = \sum_{i=1}^{N_0} h_i/h_i = \frac{1}{h_1} \sum_{i=1}^{N_0} h_i \exp\{-\frac{1}{p}\frac{\hat{\lambda}}{\varepsilon} t_i\} \approx \frac{1}{h_1} \int_0^{T_0\varepsilon} \exp\{-\frac{1}{p}\frac{\hat{\lambda}}{\varepsilon} t\}dt =$$

$$- \frac{p\varepsilon}{\hat{\lambda}h_1} [1 - \exp\{- \frac{\hat{\lambda}T_0}{p}\}]$$

Substituting (4.23) for T_0 and (4.21) for h_1,

(4.28) $\quad N_0 \approx \frac{p}{c_p} \frac{|\lambda|}{\hat{\lambda}} (\delta^{-1/p} - 1) \approx \frac{|\lambda|}{\hat{\lambda}} pc_p^{-1}\delta^{-1/p}$

This proves our claim.

QED

Further, the constants c_p of (4.21) can be shown to increase as p is increased (see Part I for $|c_\gamma|$). Thus, the estimate (4.28) also indicates that for a given accuracy δ, N_0 decreases as p (or k) is increased. Note that C_p also reflects a relative efficiency of higher order methods for problems where the eigenvalues have significant imaginary parts.

The mesh (3.46), (3.47) may be more demanding than necessary in case that eigenvalues of different magnitude are present in $\Lambda_-(t)$. At a given t, $0 \le t \le T_0\varepsilon$, the eigenvalue which imposes the smallest step size is the one for which the magnitude of the p-th derivative of the solution, $(\frac{|\lambda|}{\varepsilon})^p \exp\{-\frac{\hat{\lambda}}{\varepsilon} t\}$, is largest. Thus, if for instance,

$|\lambda_1| = \max\{|\lambda_j|,\ j = 1,\ldots,n_-\}$, then we can use (4.21) with $\lambda := \lambda_1$ in place of (3.46) and then construct the mesh using (4.22) (with $\lambda := \lambda_1$) until $t_{i+1} \geq \hat{t}_1$, where

$$\text{(4.29)} \qquad \hat{t}_1 := \min\{\hat{t}_{1j};\ \hat{t}_{1j} > 0\},\quad \hat{t}_{ij} := \varepsilon p \frac{re(\lambda_j) - re(\lambda_1)}{\ln |\lambda_{1/\lambda_j}|}$$

Then, in case that $\hat{t}_1 < T_0\varepsilon$, switch to $\lambda := \lambda_\ell$ where ℓ gives the minimum in (4.29) and continue with (4.22), etc. However, the overhead involved in constructing such a mesh is worth it only in special cases, as described above.

5. The interval away from boundaries

On the "long" interval $[t_{\underline{i}}, t_{\bar{i}}]$ we use the original problem variables and do not apply the transformation (2.5), because we can deal with the system (1.1), (1.2) directly in a simpler fashion. Thus our difference equations are (3.5), (3.6).

Consider first the "reduced" equations, where (3.5) is replaced by

$$(5.1) \qquad 0 = \sum_{\ell=1}^{k} \hat{a}_{j\ell}\{A_{11}(t_{i\ell})\underset{\sim}{y}_{i\ell} + A_{12}(t_{i\ell})\underset{\sim}{z}_{i\ell} + \underset{\sim}{f}_1(t_{i\ell})\} \qquad 1 \le j \le k$$

(For simplicity, assume that A_{rs} and $\underset{\sim}{f}_r$ are independent of ε, $1 \le r, s \le 2$). We will show that the solutions of (5.1), (3.6), denoted by $\bar{\underset{\sim}{y}}_{\Delta}(t)$, $\bar{\underset{\sim}{z}}_{\Delta}(t)$, are bounded at the collocation points in terms of their data and approximate their analytic counterparts well at these points.

Let $\underset{\sim}{x}(t) = \begin{pmatrix} \underset{\sim}{y}(t) \\ \underset{\sim}{z}(t) \end{pmatrix}$ be a smooth solution to the problem (1.1) - (1.2) on $[t_{\underline{i}}, t_{\bar{i}}]$. Recall that we can write

$$(5.2) \qquad \underset{\sim}{y}(t) = \bar{\underset{\sim}{y}}(t) + \underset{\sim}{\eta}(t), \qquad \underset{\sim}{z}(t) = \bar{\underset{\sim}{z}}(t) + \underset{\sim}{\zeta}(t)$$

where $\bar{\underset{\sim}{y}}(t)$, $\bar{\underset{\sim}{z}}(t)$ solve the "reduced" equations

$$(5.3) \qquad \underset{\sim}{0} = A_{11}\bar{\underset{\sim}{y}} + A_{12}\bar{\underset{\sim}{z}} + \underset{\sim}{f}_1 \qquad t_{\underline{i}} \le t \le t_{\bar{i}}$$

$$(5.4) \qquad \bar{\underset{\sim}{z}}' = A_{21}\bar{\underset{\sim}{y}} + A_{22}\bar{\underset{\sim}{z}} + \underset{\sim}{f}_2$$

$$(5.5) \qquad \bar{\underset{\sim}{z}}(t_{\underline{i}}) = \underset{\sim}{z}(t_{\underline{i}})$$

and, for any integer $q \ge 0$,

$$(5.6) \qquad \sum_{j=0}^{q} \left|\left|\frac{d^j \underset{\sim}{\eta}}{dt^j}\right|\right| = O(\varepsilon), \quad \sum_{j=0}^{q} \left|\left|\frac{d^j \underset{\sim}{\zeta}}{dt^j}\right|\right| = O(\varepsilon) \qquad t_{\underline{i}} \le t \le t_{\bar{i}}$$

Now, for each j, $1 \le j \le k$, we can write by (5.1)

$$(5.7) \qquad \bar{\underset{\sim}{y}}_{ij} = -A_{11}(t_{ij})[A_{12}(t_{ij})\bar{\underset{\sim}{z}}_{ij} + \underset{\sim}{f}_1(t_{ij})].$$

Substituting into (3.6), this gives

$$(5.8) \qquad \frac{1}{h_i}(\bar{\underset{\sim}{z}}_{ij} - \bar{\underset{\sim}{z}}_{i}) = \sum_{\ell=1}^{k} \hat{a}_{j\ell}\,[A_{22}(t_{i\ell}) - A_{21}(t_{i\ell})A_{11}^{-1}(t_{i\ell})A_{12}(t_{i\ell})]\bar{\underset{\sim}{z}}_{i\ell} + \\ + \underset{\sim}{f}_2(t_{i\ell}) - A_{21}(t_{i\ell})A_{11}^{-1}(t_{i\ell})\,\underset{\sim}{f}_1(t_{i\ell})$$

The solution $\bar{z}_\Delta(t)$ is then the collocation solution of the non-stiff differential equations

$$(5.9) \qquad z' = [A_{22} - A_{21} A_{11}^{-1} A_{12}]z + [f_2 - A_{21} A_{11}^{-1} f_1] \qquad t_i \le t \le t_{\bar{i}}$$

which result from substituting (5.3) into (5.4), and the usual collocation theory implies not only that

$$(5.10) \qquad ||\bar{z}^c|| \le c_1(||\bar{z}_i|| + ||f_1^c|| + ||f_2^c||)$$

but also that

$$(5.11) \qquad ||\bar{z}_\Delta - \bar{z}||_\Delta = O(h^p) , \quad ||\bar{z}_\Delta - \bar{z}||_c = O(h^{k+1}) + O(h^p)$$

if we take $\bar{z}_i := z(t_i)$. Here p is defined in (3.42) and the seminorms are defined in (3.43). In the seminorms above and throughout this section, the superscript II is omitted.

Substituting into (5.7) we obtain similar results to (5.10), (5.11) for the fast components <u>at the collocation points</u>:

$$(5.12a) \qquad ||\bar{y}_\Delta||_c \le c_2(||\bar{z}_i|| + ||f_1^c|| + ||f_2^c||)$$

$$(5.12b) \qquad ||\bar{y}_\Delta - \bar{y}||_c = O(h^{k+1})$$

In fact, if m = 0, i.e. there are no slow components z in (1.1), then from (5.7), (5.3),

$$(5.12c) \qquad \bar{y}_{ij} = \bar{y}(t_{ij}) \qquad 1 \le i \le N, \ 1 \le j \le k.$$

Now, for <u>Lobatto points</u>, (or any other set of collocation points with $\rho_k = 1$), since the mesh points are also collocation points we get from (5.11), (5.3), (5.4),

$$(5.12d) \qquad ||\bar{y}_\Delta - \bar{y}||_\Delta = O(h^p)$$

$$(5.12e) \qquad ||\bar{y}_\Delta - \bar{y}||_\Delta = 0 \qquad \text{if } m = 0$$

This leads to the essential difference between Gauss and Lobatto points in part (c) of theorem 3.2

The above results, however, do not yield even general stability for

the fast components, without being supplemented. Thus we take a closer look at equations (3.5), (3.6) and their implementation when

(5.13) $\quad \varepsilon << \underline{h} := \min\{h_i\ ;\ \underline{i} \le i \le \bar{i}\}$

5.1 The case for Gauss points

Writing

(5.14) $\quad \hat{f}_{ij} := A_{12}(t_{ij})z_{ij} + f_1(t_{ij})\quad \hat{f}_i = (\hat{f}_{i1},\ldots,\hat{f}_{ik})^T$

we can consider (3.5) separately from (3.6), with $\hat{f}_{i\ell}$ the inhomogeneous terms. As in (3.10) - (3.15) we obtain

(5.15) $\quad y_{i+1} = \Gamma_i y_i + g_i$

where Γ_i and g_i are given by (3.15), (3.11) - (3.13) with n replacing n+m, $\varepsilon^{-1}A_{11}$ replacing A and $\varepsilon^{-1}\hat{f}_i$ replacing f_i. We have

(5.16) $\quad J_i^{-1} = (I - h_i\varepsilon^{-1}D_{A_{11}}(\hat{A}\otimes I))^{-1} = \varepsilon h_i^{-1}(\varepsilon h_i^{-1}I - \hat{G}_i)^{-1},\ \hat{G}_i = D_{A_{11}}(\hat{A}\otimes I)$

As in equation (4.15) of Part I we write for the nonsingular matrix $\hat{G}_i$

(5.17) $\quad (\varepsilon h_i^{-1}I - \hat{G}_i)^{-1} = (\varepsilon h_i^{-1}\hat{C}_i - I)\hat{G}_i^{-1}$

provided that $\varepsilon < h_i||\hat{G}_i^{-1}||^{-1}$. Now,

(5.18a) $\quad \hat{B}\hat{G}_i^{-1}C_{A_{11}} = \hat{B}(\hat{A}\otimes I)^{-1}D_{A_{11}}^{-1}C_{A_{11}} = (\hat{b}^T\hat{A}^{-1}\underset{\sim}{1})I,$

where $\underset{\sim}{1} = (1,\ldots,1)^T \in \mathbb{R}^k$. From (4.17) of Part I and (5.18a),

(5.18b) $\quad I - \hat{B}\hat{G}_i^{-1}C_{A_{11}} = (1 - \hat{b}^T\hat{A}^{-1}\underset{\sim}{1})I = (-1)^k I$

and this is the leading term of Γ_i. We get in (5.15)

(5.19) $\quad y_{i+1} = (-1)^k y_i + \varepsilon h_i^{-1}\hat{H}_i y_i + \hat{B}(\varepsilon h_i^{-1}\hat{C}_i - I)\hat{G}_i^{-1}\hat{f}_i$

where the matrix $\hat{H}_i$ is bounded and depends on $\hat{b}$, $\hat{A}$ and $A_{11}(t_{ij})$, $j = 1,\ldots,k$. Clearly, both $\hat{H}_i$ and the matrix multiplying $\hat{f}_i$ in (5.19) vary smoothly with i. Further, since

(5.20) $\quad y_{ij} = A_{11}^{-1}(t_{ij})(-\hat{f}(t_{ij}) + \varepsilon F_{ij})$

(cf. (3.9)) we get for

(5.21a) $\hat{\underset{\sim}{y}}_i := (\underset{\sim}{y}_{i1},\ldots,\underset{\sim}{y}_{ik})^T$

the equations

(5.21b) $\hat{\underset{\sim}{y}}_i = -D_{A_{11}}^{-1}\hat{\underset{\sim}{f}}_i - \varepsilon h_i^{-1}D_{A_{11}}^{-1}(\varepsilon h_i^{-1}\hat{C}_i - I)\hat{G}_i^{-1}(\hat{\underset{\sim}{f}}_i + C_{A_{11}}\underset{\sim}{y}_i)$.

Now, again set $\varepsilon = 0$. Then we get for the "reduced" solution by (5.19)

(5.22) $\bar{\underset{\sim}{y}}_i = (-1)^{k(i-\underline{i})}\bar{\underset{\sim}{y}}_i - \sum_{j=\underline{i}}^{i-1}(-1)^{k(i-1-j)}\hat{B}\hat{G}_j^{-1}\hat{\underset{\sim}{f}}_j$

This general form of the "reduced" solution yields the following stability result.

Lemma 5.1. The difference equations (5.1), (3.6), using Gauss points for $\bar{\underset{\sim}{y}}_\Delta$, $\bar{\underset{\sim}{z}}_\Delta$ with $\bar{\underset{\sim}{y}}_{\underline{i}}$ and $\bar{\underset{\sim}{z}}_{\underline{i}}$ specified, have a unique solution provided that h is small enough. This solution satisfies (5.10) - (5.12) and

(5.23) $$||\bar{\underset{\sim}{y}}_\Delta||_\Delta \le ||\bar{\underset{\sim}{y}}_{\underline{i}}|| + c\{||\underset{\sim}{z}_{\underline{i}}|| + ||\underset{\sim}{f}_1^c|| + ||\underset{\sim}{f}_2^c|| + \sum_{i=0}^{\hat{i}}\sum_{j=1}^{k}||\underset{\sim}{f}_1(t_{\underline{i}+2i,j}) - \underset{\sim}{f}_1(t_{\underline{i}+2i+1,j})||\}$$

where $\hat{i}$ is the integral part of $\frac{1}{2}(\bar{i} - \underline{i})$.

Proof: Firstly, note that by (5.14) and the stability result (5.10) for the slow components,

(5.24) $||\hat{\underset{\sim}{f}}_j|| \le c(||\underset{\sim}{z}_{\underline{i}}|| + ||\underset{\sim}{f}_1^c|| + ||\underset{\sim}{f}_2^c||$

Next, we distinguish between two cases.

I. k is odd. We have

(5.25) $\hat{B}\sum_{j=\underline{i}}^{i-1}(-1)^{i-1-j}\hat{G}_j^{-1}\hat{\underset{\sim}{f}}_j = \hat{B}\sum_{j=\underline{i}}^{i-1}(\hat{G}_j^{-1}\hat{\underset{\sim}{f}}_j - \hat{G}_{j-1}^{-1}\hat{\underset{\sim}{f}}_{j-1})$

where j takes on only even or only odd values, with a possible additional end term. Each term in the right hand sum of (5.25) is written as

(5.26) $\hat{G}_j^{-1}\hat{\underset{\sim}{f}}_j - \hat{G}_{j-1}^{-1}\hat{\underset{\sim}{f}}_{j-1} = \hat{G}_j^{-1}(\hat{\underset{\sim}{f}}_j - \hat{\underset{\sim}{f}}_{j-1}) + (\hat{G}_j^{-1} - \hat{G}_{j-1}^{-1})\hat{\underset{\sim}{f}}_{j-1}$

The second term in (5.26) is bounded by $O(h_j)||\hat{f}_{j-1}||$ and so, even after we sum these quantities up in (5.25), the bound of (5.24) holds. For the first term on the right hand side of (5.26) we have to examine expressions of the form

$$(5.27) \qquad [A(t_{j\ell})\bar{z}_{j\ell} - A(t_{j-1,\ell})\bar{z}_{j-1,\ell}] + [f_1(t_{j\ell}) - f_1(t_{j-1,\ell})]$$

The term in the first square brackets is again bounded in norm by $O(h_j)||\bar{z}^c_\Delta||$, in view of (5.11). The remaining term gives rise to the rightmost term in the bound (5.23).

II. k is even. Here we have to examine the sum $\hat{B}\sum_{j=i}^{i-1}\hat{G}_j^{-1}\hat{f}_j$ and make sure that it does not grow like h^{-1}, despite the fact that its terms do not alternate in sign. But by (5.16), (5.18b),

$$(5.28) \qquad \hat{B}\hat{G}_j^{-1} = \hat{B}(\hat{A}\otimes I)^{-1}D_{A_{11}}^{-1}, \quad \hat{b}^T\hat{A}^{-1}1 = 0.$$

Thus

$$(5.29) \qquad \hat{B}\hat{G}_j^{-1}\hat{f}_j = \hat{B}(\hat{A}\otimes I)^{-1}\begin{pmatrix} A_{11}^{-1}(t_{j1})\hat{f}_{j1} \\ \vdots \\ A_{11}^{-1}(t_{jk})\hat{f}_{jk}\end{pmatrix} = \hat{B}(\hat{A}\otimes I)^{-1}\left[\begin{pmatrix} A_{11}^{-1}(t_{j1})\hat{f}_{j1} \\ \vdots \\ A_{11}^{-1}(t_{jk})\hat{f}_{jk}\end{pmatrix} - \begin{pmatrix} A_{11}^{-1}(t_{j1})\hat{f}_{j1} \\ \\ A_{11}^{-1}(t_{j1})\hat{f}_{j1}\end{pmatrix}\right]$$

This brings us to examine again terms like (5.27) and the remainder of the proof is, therefore, as for the case when k is odd

QED

Consider now the approximation error $\bar{e}(t) = \bar{y}_\Delta(t) - \bar{y}(t)$, where $\bar{y}(t)$ is defined in (5.2) - (5.5). This error satisfies the difference equations (5.1), (3.6) (with $\bar{z}_\Delta - \bar{z}$ replacing $\bar{z}_\Delta$) with the local truncation error, appropriately represented, as the inhomogeneous term. Choosing

$\bar{\underset{\sim}{y}}_\Delta(t_{\underline{i}}) := \bar{\underset{\sim}{y}}(t_{\underline{i}})$ we obtain for the error $\bar{\underset{\sim}{e}}_i$ at mesh points, by (5.22),

$$\bar{\underset{\sim}{e}}_i = \sum_{j=\underline{i}}^{i-1} (-1)^{k(i-1-j)} \underset{\sim}{K}_j h_j^{k+1} \tag{5.30}$$

with $\underset{\sim}{K}_j$ varying smoothly with j, i.e.

$$\underset{\sim}{K}_j - \underset{\sim}{K}_{j-1} = O(h_j) \tag{5.31}$$

(This is well-known to hold for the principal term of the local truncation error). Thus, in general

$$||\bar{\underset{\sim}{y}}_\Delta - \bar{\underset{\sim}{y}}||_\Delta = O(h^k) \tag{5.32}$$

and a sharper estimate is obtained when k is odd and (3.52) holds. For then we can arrange the sum in (5.30) in pairs of terms of different sign and equal value up to O(h). Thus we get in this case

$$||\bar{\underset{\sim}{y}}_\Delta - \bar{\underset{\sim}{y}}||_\Delta = O(h^{k+1}) \tag{5.33}$$

Note that at the collocation points we always have this better estimate.

Next, consider the "full" scheme (3.5), (3.6). We write the collocation solution for (1.1), (1.2) as

$$\underset{\sim}{y}_\Delta(t) = \bar{\underset{\sim}{y}}_\Delta(t) + \underset{\sim}{\eta}_\Delta(t), \quad \underset{\sim}{z}_\Delta(t) = \bar{\underset{\sim}{z}}_\Delta(t) + \underset{\sim}{\zeta}_\Delta(t) \tag{5.34}$$

choosing $\bar{\underset{\sim}{y}}_{\underline{i}} = \bar{\underset{\sim}{y}}(t_{\underline{i}})$, $\bar{\underset{\sim}{z}}_i = \bar{\underset{\sim}{z}}(t_{\underline{i}})$. Thus, $\underset{\sim}{\eta}_{\underline{i}} = \underset{\sim}{\eta}(t_{\underline{i}}) = O(\varepsilon)$, $\underset{\sim}{\zeta}_{\underline{i}} = \underset{\sim}{\zeta}(t_{\underline{i}}) = \underset{\sim}{0}$. Substituting in (3.5), (3.6) we get for $\underset{\sim}{\eta}_\Delta$ and $\underset{\sim}{\zeta}_\Delta$ (note that $\bar{\underset{\sim}{y}}(t)$ is smooth and (5.12b), (5.32) hold),

$$\sum_{\ell=1}^{k} \hat{a}_{j\ell}\{A_{11}(t_{i\ell})\underset{\sim}{\eta}_{i\ell} + A_{12}(t_{i\ell})\underset{\sim}{\zeta}_{i\ell}\} = \varepsilon h_i^{-1}(\underset{\sim}{\eta}_{ij} - \underset{\sim}{\eta}_i + \bar{\underset{\sim}{y}}_{ij} - \bar{\underset{\sim}{y}}_i) = \varepsilon h_i^{-1}(\underset{\sim}{\eta}_{ij} - \underset{\sim}{\eta}_i) + O(\varepsilon) \tag{5.35}$$

$$1 \le j \le k$$

$$\sum_{\ell=1}^{k} \hat{a}_{j\ell}\{A_{21}(t_{i\ell})\underset{\sim}{\eta}_{i\ell} + A_{22}(t_{i\ell})\underset{\sim}{\zeta}_{i\ell}\} = h_i^{-1}(\underset{\sim}{\zeta}_{ij} - \underset{\sim}{\zeta}_i) \tag{5.36}$$

$$\underset{\sim}{\eta}_{\underline{i}} = O(\varepsilon), \quad \underset{\sim}{\zeta}_{\underline{i}} = \underset{\sim}{0}. \tag{5.37}$$

Now, the right hand side terms of (5.35) are considered as inhomogeneities (to conform to (5.1), (3.6) we multiply them as a vector by $\hat{A}\,\hat{A}^{-1}$; this

does not change anything essential) and lemma 5.1 is applied. By (5.23), (5.10), (5.21),

$$(5.38)\qquad ||\underset{\sim}{\eta}^c_\Delta|| \le c\varepsilon\{\sum_{i=\underline{i}}^{\bar{i}} h_i^{-1}||\eta^c_\Delta|| + \underline{h}^{-1}||\bar{\underset{\sim}{y}}^c_\Delta||\}$$

$$(5.39)\qquad ||\underset{\sim}{\zeta}^c_\Delta|| \le c\varepsilon h^{-1}(||\underset{\sim}{\eta}^c_\Delta|| + ||\bar{\underset{\sim}{y}}^c_\Delta||)$$

Hence, if $\bar{c} := c\varepsilon \sum_{i=\underline{i}}^{\bar{i}} h_i^{-1} < 1$ then, with $\kappa := \varepsilon\underline{h}^{-1}$,

$$(5.40)\qquad ||\underset{\sim}{\eta}^c_\Delta|| \le \frac{c\kappa}{1-\bar{c}}||\bar{\underset{\sim}{y}}^c||\ ,\quad ||\underset{\sim}{\zeta}^c_\Delta|| \le \frac{c\kappa}{1-\bar{c}}||\bar{\underset{\sim}{y}}^c_\Delta||.$$

A combination of (5.40) with (5.23), (5.10) and (5.12a) now gives the desired stability result for the collocation solution $\underset{\sim}{y}_\Delta(t)$, $\underset{\sim}{z}_\Delta(t)$. Part (a) of theorem 3.2 then follows for the Gaussian points.

Next, consider the first n fundamental solution components. The collocation approximations $Y_\Delta(t)$, $Z_\Delta(t)$ are defined by the homogeneous difference schemes (3.5), (3.6) and the initial conditions (3.40b). Thus, for the "reduced" problem with $\varepsilon = 0$ we get by (5.8), (5.19),

$$(5.41)\qquad \bar{Z}_\Delta(t) \equiv 0\ ,\quad \bar{Y}_\Delta(t_i) = (-1)^{k(i-\underline{i})}I \qquad \underline{i} < i \le \bar{i}.$$

Furthermore, by repeating the argument leading to (5.40) it becomes clear that $Y_\Delta(t_i)$, $Z_\Delta(t_i)$ are only $O(\kappa)$ away from $\bar{Y}_\Delta(t_i)$ and $\bar{Z}_\Delta(t_i)$. Thus, applying the transformation (3.22), (3.51a) is obtained.

Finally, for a smooth solution $\underset{\sim}{x}(t) = \begin{pmatrix}\underset{\sim}{y}(t)\\ \underset{\sim}{z}(t)\end{pmatrix}$ which may be a transformation of $\underset{\sim}{w}_p(t)$ or of a column of $W_0(t)$, consider the approximation error. We write

$$(5.42)\qquad \underset{\sim}{y}_\Delta(t) - \underset{\sim}{y}(t) = (\bar{\underset{\sim}{y}}_\Delta(t) - \bar{\underset{\sim}{y}}(t)) + (\underset{\sim}{\eta}_\Delta(t) - \underset{\sim}{\eta}(t)),\ \underset{\sim}{z}_\Delta(t) - \underset{\sim}{z}(t) = (\bar{\underset{\sim}{z}}_\Delta(t) - \bar{\underset{\sim}{z}}(t)) + (\bar{\underset{\sim}{\zeta}}_\Delta(t) - \bar{\underset{\sim}{\zeta}}(t))$$

(see (5.2), (5.34)) and estimate the quantities on the right hand sides. We already have (5.11), (5.12b), (5.32) and (5.33). For the remaining

terms we write the difference equations as in (5.35) - (5.37) with $\underset{\sim}{\eta}_{i\ell} - \underset{\sim}{\eta}(t_{i\ell})$ replacing $\underset{\sim}{\eta}_{i\ell}$ etc. Additional local error terms of size $O(\varepsilon h_i^k)$ are easily incorporated into the right hand sides and the initial conditions corresponding to (5.37) are homogeneous. Thus, a result similar to (5.40) is obtained for the errors. This completes the proof of theorem 3.2 for Gauss points.

5.2 The case for Lobatto points

For Lobatto points we proceed in a similar way as above, establishing stability first. With $\hat{\underset{\sim}{f}}_{ij}$ as in (5.14) we obtain (5.15) where Γ_i and $\underset{\sim}{g}_i$ are given by (3.21), (3.20), with n replacing n+m, $\varepsilon^{-1}A_{11}$ replacing A and $\varepsilon^{-1}\hat{\underset{\sim}{f}}_i$ replacing $\underset{\sim}{f}_i$. Then,

$$\bar{J}_i^{-1} = \varepsilon h_i^{-1}(\varepsilon h_i^{-1} I - \bar{G}_i)^{-1}, \quad \bar{G}_i = (\bar{A} \otimes I)\bar{D}_{A_{11}} \tag{5.43}$$

and we write, as in (5.17), provided that $\varepsilon < h_i ||\bar{G}_i^{-1}||^{-1}$,

$$(\varepsilon h_i^{-1} I - \bar{G}_i)^{-1} = (\varepsilon h_i^{-1}\bar{C}_i - I)\bar{G}_i . \tag{5.44}$$

The last rows of $\bar{G}_i^{-1}$ are

$$A_{11}^{-1}(t_{i+1})(\hat{\bar{a}}_{k2} I, \ldots, \hat{\bar{a}}_{kk} I) \tag{5.45}$$

where $(\hat{\bar{a}}_{k2}, \ldots, \hat{\bar{a}}_{kk})$ is the last row of $\bar{A}^{-1}$ and hence $-\sum_{\ell=2}^{k} \hat{\bar{a}}_{k\ell}\hat{a}_{\ell 1} = \gamma(-\infty) = (-1)^{k+1}$ (cf. equations (4.19), (4.20) of Part I). Substitution in (3.21) yields

$$\underset{\sim}{y}_{i+1} = (-1)^{k+1} A_{11}^{-1}(t_{i+1}) A_{11}(t_i)\underset{\sim}{y}_i + \varepsilon h_i^{-1}\bar{H}_i \underset{\sim}{y}_i + (\varepsilon h_i^{-1}\bar{C}_i - I)\bar{Q}_i \hat{\underset{\sim}{f}}_i \tag{5.46}$$

where $\bar{H}_i$ and $\bar{Q}_i$ are bounded matrices independently of ε, which vary smoothly with i. For the other collocation points we obtain in precisely the same way (cf. (3.18))

$$\underset{\sim}{y}_{ij} = c_{ij} A_{11}^{-1}(t_{ij}) A_{11}(t_i)\underset{\sim}{y}_i + \varepsilon h_i^{-1}\bar{H}_{ij}\underset{\sim}{y}_i + (\varepsilon h_i^{-1}\bar{C}_i - I)\bar{Q}_{ij}\hat{\underset{\sim}{f}}_i \tag{5.47}$$

$$2 \le j \le k - 1$$

for appropriate constants c_{ij} and bounded matrices $\bar{H}_{ij}$, $\bar{Q}_{ij}$.

Now, the equivalent of lemma 5.1 is covered already in (5.12a), and in a preferable way: Here for the "reduced" solution no $\bar{y}_{\underline{i}}$ values are specified (like for the reduced solution of the continuous problem) and no factor like the sum on $f_{\underline{l}}$ values in (5.23) appears. Then, writing (5.34) for the "full" solution, followed by (5.35), (5.36) and

$$\zeta_{\underline{i}} = 0, \tag{5.48}$$

we obtain

$$\begin{aligned} ||\eta^c_{\Delta}|| &\le c\,\varepsilon\,\underline{h}^{-1}(\,||\eta^c_{\Delta}|| + ||\bar{y}^c_{\Delta}||) \\ ||\zeta^c_{\Delta}|| &\le c\,\varepsilon\,h^{-1}(||\eta^c_{\Delta}|| + ||\bar{y}^c_{\Delta}||) \end{aligned} \tag{5.49}$$

Hence if $c\kappa < 1$ then (5.40) is obtained. The distinction from the case for Gauss points is, however, that the condition relating ε and the mesh in order to enable the contraction argument is more pleasant here.

Next, consider the collocation approximation to $Y_{\Delta}(t)$, $Z_{\Delta}(t)$. The solution is an $O(\kappa)$ perturbation of the collocation solution to the "reduced" problem, where we consider the homogeneous equations (5.1), (3.6) subject to the initial conditions (3.40b). By (5.8) and (5.46),

$$\bar{Z}_{\Delta}(t) \equiv 0, \quad \bar{Y}_{\Delta}(t_i) = (-1)^{(k+1)(i-\underline{i})} A_{11}^{-1}(t_i) A_{11}(t_{\underline{i}}) \tag{5.50}$$

Applying the transformation (3.22) and noting that $E^{-1}A_{11}^{-1} = \Lambda^{-1}E^{-1}$, we obtain (3.51b).

Finally, consider the approximation error for a smooth solution $\underset{\sim}{x}(t)$. Writing the error as in (5.42), we need to consider $\underset{\sim}{\nu}_{i\ell} := \underset{\sim}{\eta}_{i\ell} - \underset{\sim}{\eta}(t_i)$. Again writing the difference equations as in (5.35), (5.36), (5.48) for the errors in $\underset{\sim}{\eta}_{i\ell}$ and $\underset{\sim}{\zeta}_{i\ell}$, we obtain the "reduced" form of the difference equations with inhomogeneous terms given by the local truncation error terms of size $O(\varepsilon h_i^k)$ with a smoothly varying principal error function.

This is precisely as for the Gauss points, except that there the error in $\underset{\sim}{\eta}$ is always dominated by the error in $\bar{\underset{\sim}{y}}$ (given that $\kappa << 1$), while here the error in $\underset{\sim}{\eta}$ may sometimes dominate, see (5.12d), (5.12e).

The error $\underset{\sim}{\nu}_i = \underset{\sim}{\eta}_i - \underset{\sim}{\eta}(t_i)$ is obtained analogously to $\bar{\underset{\sim}{e}}_i$ for Gauss points. We write the general solution to (5.46) for $\varepsilon = 0$ as

$$(5.51) \qquad \bar{\underset{\sim}{y}}_i = (-1)^{(i-\underline{i})(k+1)} A_{11}^{-1}(t_i) A_{11}(t_{\underline{i}}) \bar{\underset{\sim}{y}}_{\underline{i}} - \sum_{j=\underline{i}}^{i-1} (-1)^{(i-j-1)(k+1)} A_{11}^{-1}(t_{j+1}) \bar{Q}_j \hat{\underset{\sim}{f}}_j$$

and obtain for $\underset{\sim}{\nu}_i$ upon substitution

$$(5.52) \qquad \underset{\sim}{\nu}_i = \sum_{j=\underline{i}}^{i-1} (-1)^{(k+1)(i-j-1)} \underset{\sim}{K}_j \varepsilon h_j^k$$

with $\underset{\sim}{K}_j$ varying smoothly with j as in (5.31). Thus

$$(5.53) \qquad \| \underset{\sim}{\eta}_\Delta - \underset{\sim}{\eta} \|_\Delta \le c\varepsilon h^{k-1}$$

and, in case that k is even and the mesh is locally almost uniform,

$$(5.54) \qquad \| \underset{\sim}{\eta}_\Delta - \underset{\sim}{\eta} \|_\Delta \le c\varepsilon h^{k}.$$

The remaining error in the slow components clearly satisfies similar bounds. This completes the proof of theorem 3.2.

6. Numerical examples

The following numerical results were computed on an Amdahl 470-V/8 computer with a 14-hexadecimal-digits mantissa. The notation $a\text{-}b \equiv a\times 10^{-b}$ is used throughout.

Example (Hemker [3]). Consider

$$\varepsilon u'' + (2 + \cos \pi t)u' - u = f(t) \qquad 0 \le t \le 1 \tag{6.1a}$$

where

$$f(t) = -(1 + \varepsilon\pi^2)\cos \pi t - \pi(2 + \cos \pi t)\sin \pi t + (1 - \alpha + \frac{3}{2\varepsilon}\pi^2 t^2)e^{-3t/\varepsilon} \tag{6.1b}$$

subject to

$$u(0) = \alpha , \quad u(1) = -1. \tag{6.1c}$$

The solution is

$$u(t) = \cos \pi t + (\alpha - 1)e^{-3t/\varepsilon} + O(\varepsilon^2) \tag{6.2}$$

Thus, when $\alpha \neq 1$ we have a boundary layer at $t = 0$ only.

When converting to a first order system note that if we use the usual variables u, u', then the problem does not have a bounded inverse (since $u'(0) \sim '/\varepsilon$). Instead we integrate once, as in Kreiss and Kreiss [4], Kreiss and Nichols [5], obtaining with $y := u$ the system

$$\varepsilon y' = -(2 + \cos \pi t)y + z \tag{6.3a}$$

$$z' = (1 - \pi \sin \pi t)y + f(t) \tag{6.3b}$$

$$y(0) = \alpha, \; y(1) = -1. \tag{6.3c}$$

So our matrix A_{11} is a negative scalar function of t here.

First we choose $\alpha = 1$ and use uniform meshes (clearly (3.64) holds). The results are listed in table 1, where under "E" we list the maximum error at mesh points and under "rate" the measured convergence rate in h. The results for Gauss and Lobatto points confirm part (c) of theorem 3.2. In addition, we list for comparison numerical results using collocation at

Table 1: Example 1 with a smooth solution throughout, $\varepsilon = 10^{-10}$

	Gauss points				Radau points				Lobatto points				
k	N	cond	E	rate	k	N	cond	E	rate	k	N	cond	E	rate
1	10	.23+3	.64-1		1	10	.62+2	.20		2	10	.40+3	.65-1	
	20	.45+3	.16-1	2.0		20	.12+3	.10	0.9		20	.76+3	.17-1	2.0
	40	.88+3	.40-2	2.0		40	.25+3	.53-1	1.0		40	.15+4	.43-2	2.0
2	10	.87+2	.47-2		2	10	.66+2	.33-3		3	10	.16+3	.30-4	
	20	.16+3	.12-2	2.0		20	.13+3	.40-4	3.0		20	.29+3	.19-5	4.0
	40	.31+3	.29-3	2.0		40	.25+3	.49-5	3.0		40	.55+3	.12-6	4.0
3	10	.23+3	.16-3		3	10	.66+2	.18-5		4	10	.40+3	.41-6	
	20	.45+3	.98-5	4.0		20	.13+3	.54-7	5.0		20	.76+3	.68-8	5.9
	40	.88+3	.61-6	4.0		40	.25+3	.17-8	5.0		40	.15+4	.11-9	6.0
4	10	.88+2	.88-5		4	10	.66+2	.21-8		5	10	.16+3	.70-10	
	20	.16+3	.55-6	4.0		20	.13+3	.17-10	7.0		20	.29+3	.28-12	8.0
	40	.31+3	.34-7	4.0		40	.25+3	.13-12	7.0		40	.55+3	.12-13	*

* rate polluted by roundoff errors

the unsymmetric Radau points (see Part I). The usage of the latter schemes is possible here because all the eigenvalues of A_{11} have the same sign in their real part. For the examples discussed in Weiss [9] or in §6 of Part I, for instance, the Radau schemes are unstable unless the transformation 2.5 is explicitly applied (and this time not just for analysis) and the schemes are upwinded. Therefore, we stay with the symmetric schemes.

Next we set $\alpha = 0$, obtaining a steep boundary layer near $t = 0$. Results are listed in table 2. Here the meshes are constructed by taking the corresponding meshes of table 1 and adding a layer mesh according to (4.21), (4.22) with $\hat{\lambda} = -\lambda = 3$. The accuracy tolerance δ is chosen to be just below the smooth solution error for the finest mesh in table 1, for each scheme. Here we list under "E" the maximum error on all mesh points with "rate" the convergence rate in the maximum mesh width h. Note the relatively small number of mesh points needed to achieve high accuracy with the higher order schemes, particularly of Lobatto type.

Also listed in table 2 are some results when $\delta << \varepsilon << 1$. This is not covered by our analysis (see part (d) of theorem 3.1), because we are primarily concerned in this paper with what happens when $\varepsilon \to 0$. However, as indicated by the numerical results, the analysis can be extended to cover this case as well. Indeed, when $\delta << \varepsilon$ then a denser mesh is constructed in the layer regions and this only makes the situation more regular.

Other examples have been tried as well. In particular, numerical solutions for the example in Weiss [9], which for some parameter values violates condition (3.54), have been computed. Their behaviour is similar to that reported in [9] for the midpoint and trapezoidal schemes and their discussion is therefore omitted.

Table 2: Example 1 with a boundary layer, $\alpha = 0$

	Gauss points					Lobatto points				
ε	k	δ	N	E	rate	k	δ	N	E	rate
10^{-10}	1	1.-3	32	.21-1		2	1.-3	32	.13-1	
			42	.54-2	2.0			42	.32-2	2.0
			62	.15-2	1.8			62	.80-3	2.0
	2	1.-4	20	.63-2		3	1.-7	57	.22-4	
			30	.16-2	2.0			67	.13-5	4.0
			50	.39-3	2.0			87	.82-7	4.0
	3	1.-7	26	.10-3		4	1.-10	54	.75-7	
			36	.62-5	4.1			64	.11-8	6.0
			56	.39-6	4.0			84	.10-9	3.5
	4	1.-8	22	.12-4		5	1.-10	30	.11-9	
			32	.73-6	4.0			40	.70-10	
			52	.45-7	4.0			60	.70-10	
10^{-4}	3	1.-7	25	.10-3		3	1.-7	56	.20-4	
			35	.62-5	4.1			66	.11-5	4.2
			55	.38-6	4.0			86	.86-7	3.7
	4	1.-8	21	.12-4		4	1.-10	53	.61-7	
			31	.66-6	4.1			63	.11-8	5.7
			51	.26-7	4.6			83	.94-10	3.6

REFERENCES

1. U. Ascher, S. Pruess and R.D. Russell, "On spline basis selection for solving differential equations", (1981) To appear in SIAM J. Numer. Anal.

2. U. Ascher and R. Weiss, "Collocation for singular perturbation problems I: First order systems with constant coefficients". (1981) To appear in SIAM J. Numer. Anal.

3. P.W. Hemker, "A numerical study of stiff two-point boundary value problems", Math. Centrum, Amsterdam (1977).

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

5. H.O. Kreiss and N. Nichols, "Numerical methods for singular perturbation problems", Uppsala University, Dept. of Computer Science Report #57 (1975).

6. P.A. Markowich and C.A. Ringhofer, "Collocation methods for boundary value problems on 'long' intervals", (1981) To appear in Math. Comp.

7. R.D. Russell, "Collocation for systems of boundary value problems", Numer. Math. 23 (1974), 119-133.

8. R. Weiss, "The application of implicit Runge-Kutta and collocation methods to boundary-value problems", Math. Comp. 28 (1974), 449-464.

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