GENERAL FRAMEWORK, STABILITY AND ERROR ANALYSIS FOR NUMERICAL STIFF BOUNDARY VALUE METHODS

U.M. Ascher* and R.M.M. Mattheij**

Technical Report 87-28

July 1987

Abstract

This paper provides a general framework, called "theoretical multiple shooting", within which various numerical methods for stiff boundary value ordinary differential problems can be analyzed. A global stability and error analysis is given, allowing (as much as possible) the specificities of an actual numerical method to come in only locally. We demonstrate the use of our results for both one-sided and symmetric difference schemes. The class of problems treated includes some with internal (e.g. "turning point") layers.

Subject classification: AMS(MOS): 65L10.

Keywords: Stiff, boundary value problems, theoretical multiple shooting, global error analysis, stability, one-sided schemes, symmetric schemes.

* Department of Computer Science, University of British Columbia, Vancouver, B.C., Canada V6T1W5.

** Faculteit Wiskunde en Informatica, Technische Universiteit Eindhoven, 5600 MB Eindhoven, The Netherlands.

1. Introduction

Stiff ordinary differential equations (ODEs) often have solutions which exhibit narrow regions of very fast variation (so-called boundary or interior layers) which connect wider regions where the solution varies more slowly. When solving such a problem numerically, one wants to avoid using a uniformly dense discretization mesh as would be dictated by the fast modes of the stiff ODE. Thus, for a linear ODE system

$$\mathbf{y}' = A(x)\mathbf{y} + \mathbf{q}(x) \qquad a < x < b \tag{1.1}$$

($\mathbf{y}' \equiv \frac{d\mathbf{y}}{dx}$, $A(x) \in \mathbf{R}^{n\times n}$), using a mesh

$$\pi : a = x_1 < x_2 < \cdots < x_N < x_{N+1} = b, \tag{1.2a}$$

$$h_i := x_{i+1} - x_i, \quad h := \max_{1\le i\le N} h_i, \tag{1.2b}$$

one must reckon with $h||A|| \gg 1$. (The usual non-stiff numerical analysis framework assumes that $h||A|| \ll 1$, so then essentially a fundamental solution of (1.1) is followed closely pointwise everywhere on the interval $[a,b]$ by any consistent 0-stable discretization.)

The purpose of this paper is to provide a general framework within which various numerical methods for stiff boundary value problems (BVPs) can be analyzed. We further give a global stability and error analysis, allowing (as much as possible) the specificities of an actual numerical method to come in only locally. We demonstrate the use of our results for some one-sided and symmetric schemes.

Thus, consider the ODE (1.1), subject to the well-scaled boundary conditions (BC) (cf. [dHMa2])

$$\mathbf{By} \equiv B_a\mathbf{y}(a) + B_b\mathbf{y}(b) = \beta. \tag{1.3}$$

(Here B_a, $B_b \in \mathbf{R}^{n\times n}$, and we assume for later convenience that $[B_a \mid B_b]$ has orthonormal

rows.) It is often convenient (but not necessary for our results to hold) to assume as an expression of the stiffness that $A(x)$ and $\mathbf{q}(x)$ depend on a small parameter ϵ and may become unbounded as $\epsilon \to 0$. But the BVP is assumed to be well-conditioned independently of ϵ, i.e. the constant κ is of moderate size, where

$$\kappa := \max(\kappa_1, \kappa_2), \tag{1.4a}$$

$$\kappa_1 := ||\Phi||_{[a,b]}, \tag{1.4b}$$

$$\kappa_2 := ||G||_{[a,b]}, \tag{1.4c}$$

$\Phi(x)$ is the fundamental solution of (1.1) satisfying

$$\mathbf{B}\Phi = I, \tag{1.5}$$

and $G(x,s)$ is Green's function for (1.1), (1.3). Here and in the sequel we use the l_2-norm for vectors and matrices (with $|\cdot|$ for vectors and $||\cdot||$ for matrices) and the notation

$$||\mathbf{z}||_{[c,d]} := \max_{c \le x \le d} |\mathbf{z}(x)|. \tag{1.6}$$

(In (1.4c) the L_1-norm can be used. This is often advantageous for singular perturbation problems, cf. [dHMa1], [dHMa2].)

Barring a rapidly oscillatory case, we generally expect a solution profile which has boundary layers and/or interior layers connecting longer subintervals where the solution varies slowly (we will say that it is *smooth* there), which may be described by a *segmentation*,

$$a = t_1 < t_2 < \ldots < t_M < t_{M+1} = b \tag{1.7}$$

such that M is fixed (independent of ϵ), and on each subinterval $[t_j, t_{j+1}]$, precisely one of the following occurs:

(i) The solution has a boundary layer. Then $j=1$ (for a layer near a) or $j=M$ (for a layer near b) and $t_{j+1}-t_j \to 0$ as $\epsilon \to 0$.

(ii) The solution has an interior layer. Here $1<j<M$ and $t_{j+1}-t_j \to 0$ as $\epsilon \to 0$.

(iii) The solution is smooth on the subinterval, i.e. for some positive integer p

$$||\mathbf{y}^{(\nu)}||_{[t_j,t_{j+1}]} \leq const \qquad \nu = 0,1, \ldots, p \tag{1.8}$$

(where *const* is independent of ϵ).

A number of authors have considered such a segmentation in the numerical context, e.g. [KNB], [AsWe1], [We]. In [KNB] conditions are given for determining a "long" segment, i.e. where (1.8) holds. Basically, three types of modes can then be identified, based on the sign and size of the eigenvalues λ of $A(x)$: Fast decreasing modes corresponding to $\mathrm{Re}(\lambda)<0$, $|Re(\lambda)| \gg h$, fast increasing modes corresponding to $\mathrm{Re}(\lambda)>0$, $|Re(\lambda)| \gg h$, and slow modes for which $|\lambda| \ll h$. The fast modes must contribute only very little to the solution in segments where it is smooth, so they do not necessarily have to be approximated pointwise well there.

The analytic approach used to handle each of the short subintervals of types (i) and (ii) is to apply a stretching transformation to the independent variable, as implied numerically by placing a dense mesh there. (The choice of such a transformation is not always simple, but we will leave this out of our treatment.) The general effect of this stretching is to yield bounded coefficients for the ODE on the segment, even though the segment length in the new variable may become infinite as $\epsilon \to 0$.

Remark

Our distinction between fast and slow modes with respect to a discretization mesh implies two "time scales". In singular perturbation terminology we may actually have more than two time scales, e.g. a multideck system like

$$diag\ (\epsilon_1,\epsilon_2,\ldots,\epsilon_n)\mathbf{y}' = A\mathbf{y}$$

where $\epsilon_1,\epsilon_2,\ldots,\epsilon_n$ are positive scalars, some of which being small; but we still only need distinguish between fast and slow modes.

□

A bounded transformation may be envisioned which splits the modes on each segment three ways, into fast decreasing, fast increasing and slow. However, this splitting is not global: A mode may change in different segments from fast to slow or from slow to fast, altering the dimensions of the spaces of the three different mode types. In particular, such changes may give rise to internal layers which may be associated with so-called *turning points*.

On the other hand we know [Ma], [dHMa2] that since the BVP is well-conditioned the ODE (1.1) must have a dichotomy [Co]: There exists a projection P of rank p and a constant $\hat{K}$ such that

$$||\Phi(x)P\Phi^{-1}(t)|| \le \hat{K}, \quad x>t \tag{1.9a}$$

$$||\Phi(x)(I-P)\Phi^{-1}(t)|| \le \hat{K}, \quad x\le t, \tag{1.9b}$$

i.e. p modes never increase rapidly, and $n-p$ modes never decrease rapidly. (The popular view of a turning point for a 2^{nd} order scalar ODE as a location where a fast increasing mode switches direction into a fast decreasing one is incorrect: The fast increasing mode switches into a slow one, while a slow mode switches into a fast decreasing one. Thus, if there were only fast modes, a turning point would not be possible in a well-conditioned problem.) In our global analysis we will rely on the latter concept, unlike e.g. [KNB].

The radically different solution behaviour that may occur in different regions of the interval on which the BVP is defined suggests that different numerical methods and/or analyses should be applied on regions with very different solution characteristics. Hence the importance of reduction of global considerations to local (segment-wise) ones. Our main tool to achieve this

reduction is a *theoretical multiple shooting* framework, gradually developed in §§ 2,3 and 4. In this we follow and extend [dHMa1].

There are two general approaches to numerically solve stiff problems which contain different types of modes mixed together. One approach is to apply first a transformation in order to separate modes of different types. Upon decoupling increasing and decreasing fast modes (cf. [Ma]), a one-sided ("upwinded") scheme may then be applied in the appropriate directions, resulting in good, rapidly decaying approximations to rapidly decaying modes (in any direction of decay). Riccati and orthonormalization methods have been proposed ([DiRu], [Da], [LoMa], [Me]). A further separation of fast and slow modes (which must be done, if at all, in a segmented way) has been carried out to advantage in [KNB], [BrLo]. We give a global stability and error analysis for such methods in §5.

The one-sided approach is particularly useful when the ODE is in a decoupled form to begin with. But for the general case the practical transformation algorithms may be cumbersome and slow, even for linear problems. The other approach is to use numerical schemes which are capable, at least to some acceptable degree, of simultaneously handling the various types, thus eliminating the need for an explicit decoupling transformation. For initial value problems, for instance BDF schemes internally separate fast decreasing and slow modes (there are no fast increasing ones) usually successfully, and provide decaying approximations to fast decaying modes as well. For BVPs one is led to symmetric one-leg difference (or collocation) schemes. These adequately separate fast and slow modes (cf. [As1]) and preserve the dichotomy of the underlying ODE. However, the preservation of dichotomy is only done in a weak sense, since fast modes are approximated by slow ones. Consequently, various kinds of local errors do not get damped and so their effect spreads globally (cf. [We], [AsWe2], [As1], [AsJa]). One outcome of this is that dense grids must be used in layer regions, another is that errors in approximating

the direction of various modes may build up unfavourably in some pathological cases (cf. [As2]). Thus, the practicality of these schemes is somewhat marred by certain theoretical deficiencies. (Still, almost all practical problems of this sort to date have been solved by symmetric schemes.) In §6 we give a global stability and error analysis for numerical methods of the second approach, culminating in Theorem 6.9. This extends previous analyses to BVPs which may contain interior layers as well.

2. Theoretical Multiple Shooting

We begin by describing the theoretical multiple shooting framework [dHMa1]. Consider the segmentation (1.7) described in §1, where on each segment $[t_j,t_{j+1}]$ the solution of the BVP (1.1), (1.3) is of one type, be it a boundary layer, an interior layer, or a region of smooth variation. For each segment we define a BVP

$$\mathbf{y}' = A(x)\mathbf{y} + \mathbf{q}(x), \quad t_j \le x \le t_{j+1}, \tag{2.1a}$$

$$\mathbf{B}_j\mathbf{y} \equiv B_{1j}\mathbf{y}(t_j^+) + B_{2j}\mathbf{y}(t_{j+1}^-) = \mathbf{s}_j, \tag{2.1b}$$

where B_{1j}, $B_{2j} \in \mathbf{R}^{n\times n}$ and the vectors $\mathbf{s}_j \in \mathbf{R}^n$ are to be specified later. For notational convenience we require, as for B_a, B_b, that

Assumption 2.2

The matrix $[B_{1j} \mid B_{2j}]$ has orthonormal rows, $1 \le j \le M$.

□

Let $\Phi_j(x)$ be the fundamental solution of (2.1a) satisfying

$$\mathbf{B}_j\Phi_j = I \tag{2.3a}$$

and let $\mathbf{v}_j(x)$ be the particular solution of (2.1a) satisfying

$$\mathbf{B}_j\mathbf{v}_j = 0. \tag{2.3b}$$

Then the solution $\mathbf{y}(x)$ of (2.1) can be written as

$$\mathbf{y}(x) = \Phi_j(x)\mathbf{s}_j + \mathbf{v}_j(x), \quad t_j \le x \le t_{j+1}, \quad 1 \le j \le M. \tag{2.4}$$

By requiring that $\mathbf{y}(x)$ also be a solution of (1.1), (1.3), we can patch together the pieces in (2.4) via

$$\mathbf{y}(t_j^-) = \mathbf{y}(t_j^+), \quad 2 \le j \le M. \tag{2.5}$$

This gives

$$\Phi_j(t_{j+1})\mathbf{s}_j - \Phi_{j+1}(t_{j+1})\mathbf{s}_{j+1} = \beta_j := \mathbf{v}_{j+1}(t_{j+1}) - \mathbf{v}_j(t_{j+1}), \quad 1 \le j \le M-1 \tag{2.6a}$$

which, together with the BC (1.3) written as

$$B_a\Phi_1(t_1)\mathbf{s}_1 + B_b\Phi_M(t_{M+1})\mathbf{s}_M = \beta_M := \beta - B_a\mathbf{v}_1(t_1) - B_b\mathbf{v}_M(t_{M+1}), \tag{2.6b}$$

yields a system of nM linear equations for $\mathbf{s}^T = (\mathbf{s}_1^T, \ldots, \mathbf{s}_M^T)$ which we write as

$$\mathbf{A}\mathbf{s} = \mathbf{b}, \qquad \mathbf{b}^T = (\beta_1^T, \ldots, \beta_M^T). \tag{2.6c}$$

The name 'theoretical multiple shooting' can now be explained in that (2.6) resembles in form the well-known standard multiple shooting method. But here we *do not* require the BC (2.1b) to be initial conditions, hence no integration of possibly ill-conditioned initial value problems is specified.

Two basic questions arise with regard to the above formulation:

(i) Given that the BVP (1.1), (1.3) is well-conditioned, what is needed to ensure that the BVPs (2.1) be well-conditioned?

(ii) Given that we have approximations for $\Phi_j(x)$, $\mathbf{v}_j(x)$ and $\mathbf{s}_j$, say $\hat{\Phi}_j(x)$, $\hat{\mathbf{v}}_j(x)$ and $\hat{\mathbf{s}}_j$ respectively, leading to an approximate solution $\hat{\mathbf{y}}(x)$ via

$$\hat{\mathbf{y}}(x) = \hat{\Phi}_j(x)\hat{\mathbf{s}}_j + \hat{\mathbf{v}}_j(x), \quad t_j \le x \le t_{j+1}, \quad 1 \le j \le M, \tag{2.7}$$

what can be said about this $\hat{\mathbf{y}}(x)$ as an approximation to $\mathbf{y}(x)$?

Of course, the "$\hat{\ }$" notation is a simplified way to denote a numerical approximation, which may in fact be defined only pointwise on a mesh (1.2).

3. Conditioning and stability of theoretical multiple shooting

To answer question (i), we must connect the global fundamental solution and Green's function introduced in §1 (see (1.4), (1.5)) to the local ones. We have

$$\Phi_j(x) = \Phi(x)\ (\ \mathbf{B}_j\Phi\ )^{-1} \qquad 1 \le j \le M, \tag{3.1}$$

$$\mathbf{s}_j = (\mathbf{B}_j\Phi)\{\sum_{k=1}^{M-1} \Phi^{-1}(t_j)\, G(t_j,t_k)\beta_k + \beta_M\} = \mathbf{B}_j\{\sum_{k=1}^{M-1} G(\cdot,t_k)\beta_k + \Phi(\cdot)\beta_M\} \tag{3.2}$$

(cf. [dHMa1]). Moreover,

$$\mathbf{v}_j(x) = \int_{t_j}^{t_{j+1}} G_j(x,s)\mathbf{q}(s)\,ds \tag{3.3}$$

where for each j, $1 \le j \le M$, $G_j(x,s)$ is the Green's function of the BVP (2.1). It can be easily verified that the local and global Green's functions are related as

$$G_j(x,s) = G(x,s) - \Phi_j(x)\ \mathbf{B}_j G(\cdot,s). \tag{3.4}$$

Upon defining the local conditioning constants

$$\kappa_{1j} := ||\Phi_j||_{[t_j,t_{j+1}]}, \tag{3.5a}$$

$$\kappa_{2j} := ||G_j||_{[t_j,t_{j+1}]}, \tag{3.5b}$$

we find immediately that these two constants can be bounded in terms of each other:

Lemma 3.6

For $1 \le j \le M$,

$$\kappa_{2j} \le \kappa_2(1 + 2\kappa_{1j}), \tag{3.6a}$$

$$\kappa_{1j} \le 2\kappa_{2j}. \tag{3.6b}$$

Proof: The inequality (3.6a) follows from (3.4) and assumption 2.2. For (3.6b) note that

$$||\Phi_j(x)|| = ||\Phi_j(x)[B_{1j}|B_{2j}]|| \le ||\Phi_j(x)B_{1j}|| + ||\Phi_j(x)B_{2j}|| = ||G_j(x,t_j)|| + ||G_j(x,t_{j+1})||.$$

□

Lemma 3.6 implies that if the original BVP (1.1), (1.3) is well-conditioned, and if the BC (2.1b) are properly chosen for each j so that κ_{1j} is of moderate size, then the BVP (2.1) is well-conditioned. Thus we obtain a "local" dichotomy with a moderate dichotomy constant on the segment (t_j,t_{j+1}). Assuming that κ_{1j} in (3.5a) are bounded, we obtain stability of the theoretical multiple shooting approach.

Theorem 3.7

Suppose that there is a moderate constant K_1 such that for each j, $1 \le j \le M$,

$$\kappa_{1j} \le K_1. \tag{3.7a}$$

Further suppose that the BVP (1.1),(1.3) is well-conditioned (κ in (1.4) is of moderate size). Then the following hold:

(i) The local BVPs (2.1) are *well-conditioned*, with $\kappa_{2j} \le \kappa(1+2K_1)$.

(ii) The theoretical multiple shooting method is stable: there is a moderate constant $K_2=2\kappa K_1$ such that

$$cond(\mathbf{A}) \le K_2 M. \tag{3.7b}$$

(iii) The vector **s** is bounded in terms of the original data by

$$|\mathbf{s}| \le \kappa(|\beta| + 2\sum_{j=1}^{M} \kappa_{2j} ||\mathbf{q}||_{1,[t_j,t_{j+1}]}) \le \kappa(|\beta| + 2\kappa(1+2K_1) ||\mathbf{q}||_1). \tag{3.7c}$$

Above we have used the notation $||\cdot||_{1,[c,d]}$ for the L_1 norm on an interval $[c,d]$. The interval is omitted if it is $[a,b]$.

□

The proof of this theorem is straightforward, noting (3.2).

Note that if a family of singularly perturbed problems depending on a small parameter ϵ is considered, then the results of Theorem 3.7 hold uniformly in ϵ, provided that the bound (3.7a) holds uniformly, too.

4. Dichotomy and the choice of local boundary conditions

The utility of Theorem 3.7 still depends on finding suitable local BC for (2.1) so that (3.7a) hold with a constant K_1 of moderate size. In order to construct such local BC we use the dichotomic structure (1.9) of the problem (assumed to hold uniformly in ϵ, $0<\epsilon\le\epsilon_0$, say). Since this concept is global, we are able to generalize earlier more restricted efforts like [AsWe1], [We], to problems including turning points.

Without loss of generality we may take in (1.9)

$$P = \begin{pmatrix} 0 & 0 \\ 0 & I_p \end{pmatrix}, \tag{4.1}$$

so we may partition the fundamental solution $\Phi(x)$ as

$$\Phi(x) = (\Phi^1(x) \mid \Phi^2(x)) \tag{4.2}$$

where $\Phi^1(x) \in \mathbf{R}^{n\times(n-p)}$ and $\Phi^2(x) \in \mathbf{R}^{n\times p}$ denote the nondecreasing and the nonincreasing parts, respectively. This induces a natural choice for the local BC as follows: Let $Q_{1j} \in \mathbf{R}^{n\times p}$ and $Q_{2j} \in \mathbf{R}^{n\times(n-p)}$ be two matrices with orthonormal columns such that

$$Q_{1j}^T\Phi^1(t_j) = 0, \quad Q_{2j}^T\Phi^2(t_{j+1}) = 0. \tag{4.3}$$

Then define

$$B_{1j} := \begin{bmatrix} 0 \\ Q_{1j}^T \end{bmatrix}, \quad B_{2j} := \begin{bmatrix} Q_{2j}^T \\ 0 \end{bmatrix}. \tag{4.4}$$

It follows that $\mathbf{B}_j\Phi$ is a block diagonal matrix with full rank blocks, whence $\mathbf{B}_j\Phi$ is nonsingular and by (3.1) $\Phi_j(x)$ is well-defined. We may now bound the local conditioning constants in terms of the global dichotomy constant:

Theorem 4.5

If (1.9) holds and the local BC are chosen as in (4.4) then

$$\kappa_{1j} \le 2\hat{K}, \tag{4.5a}$$

$$\kappa_{2j} \le \hat{K}. \tag{4.5b}$$

Proof: The bound (4.5a) follows immediately form (4.5b) and (3.6b). To see (4.5b), write for

the Green's function

$$G_j(x,s) = \begin{cases} \Phi_j(x) B_{1_j} \Phi_j(t_j) \Phi_j^{-1}(s), & x > s \\ -\Phi_j(x) B_{2_j} \Phi_j(t_{j+1}) \Phi_j^{-1}(s) & x \le s \end{cases}$$

and substitute (3.1) to obtain

$$G_j(x,s) = \begin{cases} \Phi(x) P \Phi^{-1}(s), & x > s \\ -\Phi(x)(I-P)\Phi^{-1}(s) & x \le s \end{cases}$$

and the result follows from (1.9).

□

We can now substitute the bounds (4.5) in Theorem 3.7 to obtain

Corollary 4.6

With the dichotomy (1.9) holding, and choosing the local BC as in (4.4), Theorem 3.7 holds with $K_1 = 2\hat{K}$, $K_2 = 4\kappa\hat{K}$, and

$$|s| \le \kappa(|\beta| + 2\hat{K}||\mathbf{q}||_1). \tag{4.6a}$$

□

We have answered the first question posed at the end of §2. In the following two sections we investigate the second question posed there, namely, when are given approximations $\hat{\Phi}_j(x)$, $\hat{v}_j(x)$ and $\hat{s}_j$ appropriate for having a globally meaningful result? We shall focus on the resulting multiple shooting matrix $\hat{\mathbf{A}}$, which is formed similarly to $\mathbf{A}$ (cf. (2.6)) from the approximate quantities.

5. Stability and global error analysis I

If the approximate theoretical multiple shooting matrix $\hat{\mathbf{A}}$ is a sufficiently accurate approximation of $\mathbf{A}$ (element-wise) then the stability of the numerical method leading to this $\mathbf{A}$ is almost directly related to the conditioning of the original BVP. We also obtain directly a localized pointwise error estimate:

Theorem 5.1

Suppose that the assumptions of Theorem 3.7 hold. In addition suppose that there are constants δ_1, δ_2 such that, for $1 \le j \le M$,

$$\|\hat{\Phi}_j(t_j) - \Phi_j(t_j)\|, \ \|\hat{\Phi}_j(t_{j+1}) - \Phi_j(t_{j+1})\| \le \delta_1, \tag{5.1a}$$

$$2\kappa M\delta_1 =: \gamma < 1, \tag{5.1b}$$

$$|\hat{\mathbf{v}}_j(t_j) - \mathbf{v}_j(t_j)|, \ |\hat{\mathbf{v}}_j(t_{j+1}) - \mathbf{v}_j(t_{j+1})| \le \delta_2. \tag{5.1c}$$

Then the approximate theoretical multiple shooting method is well defined and stable. Furthermore,

$$\|\hat{A}^{-1}\| \le \frac{\kappa M}{1-\gamma}, \tag{5.1d}$$

$$|\hat{\mathbf{s}} - \mathbf{s}| \le \frac{1}{1-\gamma}(\hat{\gamma} + \gamma|\mathbf{s}|) \tag{5.1e}$$

where $\hat{\gamma} := 2\kappa M\delta_2$, and

$$|\hat{\mathbf{y}}(x) - \mathbf{y}(x)| \le \frac{K_1}{1-\gamma}(\hat{\gamma}+\gamma|\mathbf{s}|) + |(\hat{\Phi}_j(x)-\Phi_j(x))\hat{\mathbf{s}}_j| + |\hat{\mathbf{v}}_j(x) - \mathbf{v}_j(x)|, \tag{5.1f}$$

$$t_j \le x \le t_{j+1}, \ 1 \le j \le M.$$

Proof: From (5.1a),

$$\|\hat{\mathbf{A}} - \mathbf{A}\| \le 2\delta_1.$$

Then (5.1b) guarantees the estimate (5.1d) by a standard perturbation argument. The result (5.1e) also follows using standard algebraic manipulations. Finally, to obtain (5.1f) we write

$$\hat{y}(x) - y(x) = \hat{\Phi}_j(x)\hat{s}_j - \Phi_j(x)s_j + \hat{v}_j(x) - v_j(x) =$$
$$= \Phi_j(x)(\hat{s}_j - s_j) + (\hat{\Phi}_j(x) - \Phi_j(x))\hat{s}_j + \hat{v}_j(x) - v_j(x),$$

and take norms.

□

5.1 Applications

In order to appreciate the importance of Theorem 5.1, we first note that s is bounded via Corollary 4.6, whence $\hat{s}$ and its distance from s are bounded via (5.1e). By our choice of local BC we have well-conditioned local BVPs if the original, given BVP is well-conditioned. This leads to the following conclusion:

If

(i) $\Phi_j(x)$ and $v_j(x)$ are approximated sufficiently accurately *at the segments ends* ($x=t_j,t_{j+1}$);

(ii) the smooth solution components are similarly approximated sufficiently accurately *throughout* each segment (t_j,t_{j+1});

then by (5.1f) we can expect controllably small errors $|\hat{y}(x) - y(x)|$. For (i) and (ii) we again remark that "slow" and "fast" is a local, segment-wise notion.

In short, this theorem allows us to concentrate on well-conditioned BVPs each defined on a segment of one type only, for instance with the solution smooth throughout the segment. The error is then localized. Moreover, on a smooth solution segment $[t_j,t_{j+1}]$, $||\hat{\Phi}_j(x) - \Phi_j(x)||$ need not be small for a small error to be obtained in $y(x)$, only $|(\hat{\Phi}_j(x) - \Phi_j(x))\hat{s}_j|$ and $|\hat{v}_j(x) - v_j(x)|$ matter. These correspond to the smooth components of the solution only.

Suppose that we use a finite difference or a marching discretization method for a stiff BVP (1.1), (1.3), based on a mesh (1.2). To fit this into the framework presented here, we can take a subset of the points of π for the theoretical multiple shooting mesh (1.7). (The mesh (1.2) has to be such that a choice of a subset suitable for the application of Theorem 5.1 is possible.) The requirements above are then satisfied, in principle, for one-sided schemes.

Example 1

Consider applying the backward Euler scheme with a uniform step size h to the initial value problem

$$\epsilon y' = -y + q(x) \quad 0<x<1$$
$$y(0) = q(0) + 1,$$

with $q(x)$ a smooth function, $||q||=1$, and $0<\epsilon \ll h \ll 1$. This scheme yields

$$y_{i+1} = \frac{\epsilon h^{-1}}{1+\epsilon h^{-1}} y_i + \frac{1}{1+\epsilon h^{-1}} q(x_{i+1}), \quad i=1,\ldots,h^{-1}.$$

The exact solution has a boundary layer of the form $e^{-x/\epsilon}$ at $x=0$ and is smooth away from 0. Using $h \gg \epsilon$, this layer is then skipped over by the mesh.

To apply the theoretical multiple shooting framework, consider the segments defined by $t_1=0$, $t_2=h$, $t_3=1$, i.e. $M=2$, and define approximate quantities in between mesh points using linear interpolation. We make the stable choice $B_{11} = B_{12} = 1$, $B_{21} = B_{22} = 0$. Clearly, in the segment $[t_1,t_2]$ we do not have a pointwise accurate approximation to $\Phi_1(x) = e^{-x/\epsilon}$, but at $x=t_2$ we do have

$$\hat{\Phi}_1(t_2) = \frac{\epsilon h^{-1}}{1+\epsilon h^{-1}} \approx 0 \approx e^{-h/\epsilon} = \Phi_1(t_2),$$

and similarly $\hat{v}_1(t_2) \approx v_1(t_2)$. Therefore δ_1 and δ_2 in (5.1) are very small (in fact they shrink to 0 as $\epsilon \to 0$). According to (5.1f), then, the approximation of the smooth solution on $[h,1]$ depends only on accuracy considerations for this segment, and is essentially not affected by the poor pointwise approximation in the layer segment $[0,h)$.

□

If we want a uniformly accurate approximate solution on $[a,b]$ then we must have a fine (dense) mesh in layer regions. This is the approach taken in [KNB], [BrLo] and [DOR], for example. (It assumes, incidentally, that in a way the location of layers is known, or can be found out directly from the problem coefficients - a nontrivial assumption for nonlinear problems.) Picking the points t_j to be the mesh points where the mesh changes from fine to coarse (or just outside layers), the conditions of Theorem 5.1 are satisfied, because in smooth regions where large steps h_i are taken, one-sided schemes for the decoupled ODE are used which damp the fast modes. If the explicit decoupling transformations which these methods perform prior to discretization are well-conditioned (see the above cited references for this) then by (5.1d) the methods are stable, with a stability constant directly related to the conditioning constant of the given BVP.

Theorem 5.1 also allows for use of different discretization schemes in different segments of the interval $[a,b]$. For instance, symmetric schemes can be more economical in layer regions, provided that a dense mesh is used there (see, e.g., [KNB]).

6. Stability and global error analysis II

In this section we investigate cases where the conditions of Theorem 5.1 do not hold. In particular, if $\Phi_j(x)$ is not approximated well at mesh points of a segment $[t_j,t_{j+1}]$ with a smooth

solution, then it may not be approximated well at the segment's endpoints, so in general (5.1a,b) does not hold. (This will occur, with a reasonable numerical scheme, only if the mesh is coarse, whence the concentration on a smooth solution segment.) In such a case $||\hat{\mathbf{A}} - \mathbf{A}||$ is not necessarily small, so we cannot conclude (5.1d), and must resort to other considerations to obtain stability. Yet it is often possible to give useful results without (5.1), as we shall now show.

What we will insist on is that the dichotomy structure of the ODE solution space be preserved by the numerical method. This still allows the approximant to a mode to be "slow" even when the actual exact one is "fast". Recall that for accuracy reasons alone there is no need to approximate the fast modes well in smooth solution regions. (Some undesirable side effects of such a poor fast mode approximation may result, though, as mentioned in §1. We will return to this later.) Symmetric difference schemes preserve the dichotomy in this way, but one-sided schemes like BDF, without prior decoupling and upwinding, do not.

From this description it transpires that for establishing boundedness of $||\hat{\mathbf{A}}||$ (i.e. stability) we cannot rely on closeness to $\mathbf{A}$ and must consider $\hat{\mathbf{A}}$ directly. To this end, define the block diagonal matrix

$$\mathbf{Q} = diag(Q_1,\ Q_2,\ \ldots,\ Q_{M-1},\ I) \tag{6.1a}$$

where

$$Q_j := \begin{pmatrix} Q_{2j}^T \\ Q_{1,j+1}^T \end{pmatrix} \tag{6.1b}$$

(see (4.3) for Q_{2j}, $Q_{1,j+1}$). The j^{th} row blocks of $\mathbf{A}$ and $\hat{\mathbf{A}}$ are

$$\Psi_j := (\Phi_j(t_{j+1}) \mid -\Phi_{j+1}(t_{j+1})), \tag{6.2a}$$

and

$$\hat{\Psi}_j := (\hat{\Phi}_j(t_{j+1}) \mid -\hat{\Phi}_{j+1}(t_{j+1})), \tag{6.2b}$$

respectively. By definition we have the structure

$$Q_j\Psi_j = \begin{pmatrix} I & 0 & | & U_{j2} & V_{j2} \\ V_{j1} & U_{j1} & | & 0 & -I \end{pmatrix}, \tag{6.3a}$$

where $U_{j1} \in \mathbb{R}^{p\times p}$ and $U_{j2} \in \mathbb{R}^{(n-p)\times(n-p)}$ are nonsingular blocks.

The blocks V_{j1}, V_{j2} can be further shown to be *zero* as follows: By (3.1),

$$\Phi_j(x) = \Phi_{j+1}(x)[(\mathbf{B}_{j+1}\Phi)(\mathbf{B}_j\Phi)^{-1}]. \tag{6.4}$$

Since $\mathbf{B}_j\Phi$ is block diagonal with the same block structure for each j, the matrix $(\mathbf{B}_{j+1}\Phi)(\mathbf{B}_j\Phi)^{-1}$ also has this block structure, say

$$(\mathbf{B}_{j+1}\Phi)(\mathbf{B}_j\Phi)^{-1} = \begin{pmatrix} R_j^1 & 0 \\ 0 & R_j^2 \end{pmatrix},$$

where $R_j^1 \in \mathbb{R}^{(n-p)\times(n-p)}$, $R_j^2 \in \mathbb{R}^{p\times p}$ are nonsingular. Hence from (6.4),

$$\Phi_j^1(x) = \Phi_{j+1}^1(x)R_j^1\ ,\ \Phi_j^2(x) = \Phi_{j+1}^2(x)R_j^2\ .$$

Substituting this into (6.3a) and observing (6.2) and (4.3), the above claim follows and we obtain

$$Q_j\Psi_j = \begin{pmatrix} I & 0 & | & U_{j2} & 0 \\ 0 & U_{j1} & | & 0 & -I \end{pmatrix}. \tag{6.3b}$$

For $Q_j\hat{\Psi}_j$ we can write similarly to (6.3a),

$$Q_j\hat{\Psi}_j = \begin{pmatrix} I & 0 & | & \hat{U}_{j2} & \hat{V}_{j2} \\ \hat{V}_{j1} & \hat{U}_{j1} & | & 0 & -I \end{pmatrix}, \tag{6.5}$$

however we do not generally have an additional zero structure like in (6.3b).

Yet, an additional structure can be observed in (6.5), assuming that the smooth solution components are approximated sufficiently accurately. Firstly, let us denote by p_j and k_j the dimensions of the spaces of *fast decreasing* modes and *fast increasing* modes, respectively, on the j^{th} segment. Next, note that it is not restrictive to assume that the block matrices

$$Q_{1,j+1}^T \hat{\Phi}_j^2(t_{j+1}),\ Q_{2j}^T \hat{\Phi}_{j+1}^1(t_{j+1}),$$

are upper triangular, for otherwise a Gram-Schmidt algorithm may be invoked, with the orthogonal part incorporated into the matrices $Q_{1,j+1}^T$ and Q_{2j}^T. Hence we conclude that it is not restrictive to assume that the matrices $\hat{U}_{j1}$, $\hat{U}_{j2}$ are upper triangular. By a similar argument we see that, *alternatively*, we may assume that the matrices $\hat{V}_{j1}$, $\hat{V}_{j2}$ have a zero upper-left triangle (for this Gram-Schmidt is invoked from right to left). Finally, it is not restrictive to assume that layer segments alternate with smooth solution regions in (1.7). We combine all this into

Assumption 6.6

(i) Let M be odd ($M \le N$), and set $\hat{M} := \frac{1}{2}(M-1)$. Let the solution on each segment $[t_{2l}, t_{2l+1}]$ be smooth, $l=1,\ldots,\hat{M}$, the other segments being layer regions. Assume further that there is some (small) $\delta>0$ such that on each layer segment $[t_{2l-1}, t_{2l}]$, $l=1,2,\ldots,\hat{M}+1$,

$$||\Phi_j(x) - \hat{\Phi}_j(x)||,\ |v_j(x) - \hat{v}_j(x)| \le \delta, \quad j=2l-1,\ t_j \le x \le t_{j+1}. \tag{6.6a}$$

For the smooth segments assume that $\hat{\Phi}_{2l}(x)$ exists, and that the $p-p_{2l}$ smooth nonincreasing modes and the $n-p-k_{2l}$ smooth nondecreasing modes are approximated (in *sup* norm) up to δ, $l=1,2,\ldots,\hat{M}$.

(ii) Let $Q_{1,j+1}^T$ be chosen such that

$\hat{V}_{j1}$ has a zero upper-left triangle, j even,

$\hat{U}_{j1}$ is upper triangular, j odd.

(iii) Let Q_{2j}^T be chosen such that

$\hat{V}_{j2}$ has a zero upper-left triangle, j odd,

$\hat{U}_{j2}$ is upper triangular, j even.

(iv) For j odd, let the contribution of the fast modes in Φ_j which decay towards the "layer's end" $t_{j\pm1}$ be $O(\delta)$ at $t_{j\pm1}$.

□

This yields

Lemma 6.7

For j odd, the last p_j rows of $\hat{U}_{j1}$ and the first $p-p_j$ columns of $\hat{V}_{j2}$ are $O(\delta)$. Hence

$$|| \hat{V}_{j2} \hat{U}_{j1} || = O(\delta), \qquad \text{j odd.} \tag{6.7a}$$

For j even, the last k_j rows of $\hat{U}_{j2}$ and the first $n-p-k_j$ columns of $\hat{V}_{j1}$ are $O(\delta)$. Hence

$$|| \hat{V}_{j1} \hat{U}_{j2} || = O(\delta), \qquad \text{j even.} \tag{6.7b}$$

Proof: Let j be odd. By Assumption 6.6 (iv) and (6.6a), the contribution of fast components to $\hat{U}_{j1}$ is only $O(\delta)$. If this contribution was zero then the Gram-Schmidt algorithm would have yielded zero in the last p_j rows of $\hat{U}_{j1}$, so a perturbation argument gives the claim for this block. As for $\hat{V}_{j2}$, this is a residual arising from a smooth region. Since V_{j2}=0 and the smooth solution components are approximated well, it follows that $\hat{V}_{j2}$ is an $O(\delta)$ perturbation of a matrix with rank at most $p-p_j$. Orthogonalization from right to left then gives the claimed structure. The estimate (6.7a) follows.

The proof for j even is similar.

For illustration we display the case for M=5, i.e. we have *three* layer segments sandwiched with *two* smooth regions as in Figure 6.1.

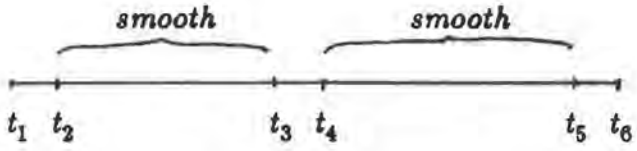

Figure 6.1 - layers and smooth solution regions

Hence we have, e.g. a turning point at $[t_3,t_4]$. By neglecting $O(\delta)$ differences between exact and approximate on layer regions (noting (6.3b)), the matrix $\mathbf{Q}\hat{\mathbf{A}}$ has the structure

(6.8)

$$
\begin{array}{|cc|cc|cc|cc|}
\hline
I & 0 & \hat{U}_{12} & \hat{V}_{12} & & & & \\
0 & U_{11} & 0 & -I & & & & \\
\hline
 & & I & 0 & U_{22} & 0 & & \\
 & & \hat{V}_{21} & \hat{U}_{21} & 0 & -I & & \\
\hline
 & & & & I & 0 & \hat{U}_{32} & \hat{V}_{32} \\
 & & & & 0 & U_{31} & 0 & -I \\
\hline
 & & & & & & I & 0 & U_{42} & 0 \\
 & & & & & & \hat{V}_{41} & \hat{U}_{41} & 0 & -I \\
\hline
\multicolumn{2}{|c|}{B_a\Phi_1(a)} & & & & & & & \multicolumn{2}{c|}{B_b\Phi_5(b)} \\
\hline
\end{array}
$$

The main result of this section can now be stated and proved:

Theorem 6.9

Suppose that the BVP (1.1), (1.3), (1.4) is well-conditioned and consider a discretization method in the theoretical multiple shooting framework, with Assumption 6.6 holding (for δ small enough).

Then there is a constant c of moderate size such that the following stability estimate holds:

$$\|[\hat{\mathbf{A}}]^{-1}\| \leq c\, M \max\{\kappa,\ \max_{1\leq l\leq \hat{M}} \{\|\hat{\Phi}_{2l}(t_{2l})\|,\ \|\hat{\Phi}_{2l}(t_{2l+1})\|\}\}. \tag{6.9a}$$

Proof: We consider $\mathbf{Q}\hat{\mathbf{A}}$, with $\mathbf{Q}$ given by (6.1), first for separated BC. Since (1.1), (1.3) is well-conditioned, we may consider without loss of generality only the BC $B_a = B_{11}$, $B_b = B_{M2}$. Then

$$B_a\Phi_1(a) = \begin{pmatrix} 0 & 0 \\ 0 & I_p \end{pmatrix},\quad B_b\Phi_M(b) = \begin{pmatrix} I_{n-p} & 0 \\ 0 & 0 \end{pmatrix}.$$

Permuting row blocks of $\mathbf{Q}\hat{\mathbf{A}}$ so that the identity blocks form the main diagonal (cf. (6.8)), we may consider a Gauss-Jordan elimination procedure with identity blocks as pivots, which yields an explicit expression for $[\mathbf{Q}\hat{\mathbf{A}}]^{-1}$. The bound (6.9a) is then obtained upon estimating the blocks encountered in this process. Upon considering the elimination, attention is reduced to super-blocks of the form

$$\begin{bmatrix} I & \hat{U}_{2l,2} & 0 & 0 \\ 0 & I & 0 & \hat{V}_{2l+1,2} \\ \hat{V}_{2l,1} & 0 & -I & 0 \\ 0 & 0 & \hat{U}_{2l+1,1} & -I \end{bmatrix}.$$

By Lemma 6.7, we may consider this block as an $O(\delta)$ perturbation of one where

$$\| \hat{V}_{2l+1,2} \hat{U}_{2l+1,1} \| = 0.$$

For the latter, we may diagonalize the second row using the fourth, then the first using the second, then the third using the first, and lastly the fourth using the third. Only elements covered by the proposed bound in (6.9a) appear in the thus-formed inverse. A perturbation argument for δ small enough, using (6.7a), concludes the proof for separated BC.

For general BC, consider the partition notation

$$B_a\Phi_1(a) =: \overline{B_a} \equiv \begin{pmatrix} \overline{B_a^{I1}} & \overline{B_a^{I2}} \\ \overline{B_a^{Z1}} & \overline{B_a^{Z2}} \end{pmatrix}, \quad B_b\Phi_M(b) =: \overline{B_b} \equiv \begin{pmatrix} \overline{B_b^{I1}} & \overline{B_b^{I2}} \\ \overline{B_b^{Z1}} & \overline{B_b^{Z2}} \end{pmatrix}.$$

Our dichotomy assumption (1.9) implies that $\overline{B_a}$ must control the p nonincreasing modes; consequently, it is not restrictive to assume that the $p \times p$ block $\overline{B_a^{Z2}}$ is nonsingular and that

$$\overline{B_a^{Z1}} = 0, \quad \overline{B_a^{I2}} = 0.$$

Furthermore, $\overline{B_b}$ has the same block structure as $\overline{B_a}$, i.e. we may assume that its off-diagonal blocks are zero, too. This follows because $\overline{B_a} = B_a\Phi(a)(\mathbf{B}_1\Phi)^{-1}$ and

$$\overline{B_b} = B_b\Phi(b)(\mathbf{B}_M\Phi)^{-1} = (I - B_a\Phi(a))(\mathbf{B}_M\Phi)^{-1} = (I - \overline{B_a}(\mathbf{B}_1\Phi))(\mathbf{B}_M\Phi)^{-1},$$

with $(\mathbf{B}_1\Phi)(\mathbf{B}_M\Phi)^{-1}$ and $(\mathbf{B}_M\Phi)^{-1}$ being block-diagonal matrices.

Before considering $\mathbf{Q}\hat{\mathbf{A}}$ for this case, let us consider the exact matrix $\mathbf{QA}$ for which Theorem 3.7 holds. We will use row-block elimination only, and attempt to obtain separated BC. Thus, we proceed to eliminate $\overline{B_a^{I1}}$. We use the $(2j-1)^{st}$ (block-) rows in succession, $j=1,\ldots,M-1$, adding each time an appropriate multiple to the $(2M-1)^{st}$ row. This eliminates $\overline{B_a^{I1}}$ and cleans up the BC rows, but replaces $\overline{B_b^{I1}}$ by

$$\tilde{B}_b^{11} := \overline{B}_b^{I1} + \overline{B}_a^{I1} U_{12} U_{22} \cdots U_{M-1,2}. \tag{6.10a}$$

Next we proceed to eliminate $\overline{B}_b^{I2}$. We use in succession (block-) rows $2j$, $j=M-1,\ldots,1$, yielding separated BC with $\overline{B}_a^{I2}$ replaced by

$$\tilde{B}_a^{22} := \overline{B}_a^{I2} + \overline{B}_b^{I2} U_{M-1,1} \cdots U_{21} U_{11}. \tag{6.10b}$$

Now, from the well-conditioning we must have that the matrices defined in (6.10) are nonsingular with moderate condition numbers. Furthermore, note that the addition to the original block in (6.10a) involves nondecreasing modes evaluated at the left end of *each* segment j and, similarly, the addition to the original block in (6.10b) involves nonincreasing modes evaluated at the right end of *each* segment j, $j=1,\ldots,M$. Thus, these additions involve (significant) propagation of slow modes only.

After this we are ready to attack the approximate case. We neglect $O(\delta)$ effects assuming that everything is exact in layer regions. In particular, the BC rows are "exact". Then we proceed as for **A** above to eliminate $\overline{B}_a^{I1}$. But before using row $2j-1$ for j odd, we use row $2j$ to eliminate $\hat{V}_{j2}$. The fill-in that this causes is only $O(\delta)$ by (6.7a). Following this "preprocessing", elimination is done as before, and we obtain an approximation of $\tilde{B}_b^{11}$,

$$\overline{B}_b^{I1} + \overline{B}_a^{I1} \hat{U}_{12} U_{22} \hat{U}_{32} \cdots U_{M-1,2} + O(\delta) = \tilde{B}_b^{11} + O(\delta).$$

The latter estimate holds because fast modes are damped in layer regions and slow modes are assumed to be well-approximated everywhere on the interval $[a,b]$. Similarly, we next eliminate $\overline{B}_b^{I2}$, preprocessing row $2j$ by row $2j-1$ for j even. The fill-in is $O(\delta)$ by (6.7b), and a similar argument yields that here, too, we have an $O(\delta)$ approximation to $\tilde{B}_a^{22}$ of (6.10b).

A perturbation argument for small δ shows that we have reduced the problem to one with separated BC, for which the claimed bound has been proved earlier. This completes the proof for the nonseparated BC case as well.

□

Having shown stability, error estimation can be done in the usual way. Thus, if the stability conditions of Theorem 6.9 hold with the right hand side of (6.9a) being of moderate size, then we may substitute (local truncation and boundary) errors in place of solutions in stability bounds. The stability and error analysis may be done on each segment separately. The so obtained segment error estimates are then magnified in the global error estimate by at most the bound in (6.9a).

6.1 Applications

We now consider symmetric one-step finite difference schemes, for which the analysis in this section applies and the one in §5 does not. As in §5.1, we envision a mesh (1.2) such that a subset of π forms an adequate theoretical multiple shooting mesh (1.7).

First we observe the implications of Theorem 6.9, given that its conditions are satisfied. Again we note the crucial point that on the smooth solution segments $[t_{2l},t_{2l+1}]$, where the mesh is coarse, the error is driven by the local truncation error, which depends on the smooth components of the solution alone (and not on $||\hat{\Phi}_{2l}(x) - \Phi_{2l}(x)||$). This explains on a theoretical basis the often satisfactory performance, for stiff BVPs with internal layers, of a code like COLSYS [ACR], which attempts to adaptively refine the mesh (1.2) based essentially only on the solution profile. Earlier analysis efforts, e.g. [AsWe1], [AsWe2], [As1], assume a much more restricted problem setting which in particular does not include internal layers. (We remark, though, that these works do give specific information on layer discretization, while here we do not.)

Observe that there is one difference between this segment analysis and the usual non-stiff analysis for one-step schemes. In the latter, one automatically assumes that the BC are satisfied exactly by the approximate solution, while here we must admit boundary errors at segment ends. Assumption 6.6 (iv) requires that this error be small. This means not only that the layer mesh should be sufficiently fine, but also that it should cover "all" of the layer.

Let us now consider Part (i) of Assumption 6.6 (the other parts do not cause difficulty, as already discussed above). It is important to stress that this assumption *does not* follow from using a "reasonable" discretization mesh for a well-conditioned BVP (reasonable, that is, based on approximation considerations for the solution profile). Clearly, a dense mesh in layer regions is a must for obtaining good accuracy anywhere with symmetric schemes; but in case that a boundary layer is missing in the exact solution (say because a reduced solution of a singularly perturbed problem happens to satisfy the BC), then it is less clear at a first glance that the mesh structure specified in the assumption is required. Indeed, if a dense layer mesh is missing in such a case, it is well-known that the resulting scheme may or may not be stable (but it is accurate if stable), and examples for both instances can be easily constructed.

Moreover, less obvious is the question whether $\hat{\Phi}_{2l}$ exists (and is moderately bounded; for otherwise the bound in (6.9a) is not very meaningful). Indeed, since fast decaying modes are approximated with symmetric schemes by slow ones, errors in the mode directions are not damped, and these may in principle build up in an unfortunate way. In restricted singular perturbation contexts, this has been linked to the need to require well-conditioning of an auxiliary matrix [We], [AsWe2], or BVP [Kr], [As1], [As2]. The auxiliary BVP involves a special ratio of the mesh step h and the BVP's small parameter ϵ, and so its possible singularity is more a theoretical curiosity than a practical problem. (An example is given in [As2, Example 1].) Indeed, in the highly unlikely event that it occurs at all for a given BVP, if a code with an

adaptive mesh selection algorithm is used then it would usually respond to obtaining the poorer approximation by changing the mesh, thus inadvertantly eliminating the instability.

Of slightly more practical interest is the case where we let $\epsilon \to 0$ keeping h fixed (or just looking at $\epsilon \ll h^2$). The stability condition in [We], [AsWe2] translates to the algebraic condition that the matrix

$$\begin{pmatrix} Q_{1j}^T \\ Q_{2j}^T \end{pmatrix}$$

(cf. (4.3), (4.4)) have a bounded inverse.

Example 2

This is an adaptation of an example in [We]. Consider the ODE

$$\epsilon \mathbf{y}' = A(x)\mathbf{y} + \mathbf{q}(x) \qquad a < x < b$$

with

$$A(x) = T(x;\nu) \begin{pmatrix} -1 & 0 \\ 0 & 2 \end{pmatrix} T^{-1}(x;\nu), \quad \mathbf{q}(x) = T(x;\nu)\,[\epsilon \mathbf{w}_p'(x) - \hat{A}\mathbf{w}_p(x)],$$

where T is the reflection matrix

$$T(x;\nu) = T^{-1}(x;\nu) = \begin{pmatrix} \sin \nu x & \cos \nu x \\ \cos \nu x & -\sin \nu x \end{pmatrix},$$

$$\hat{A} = \begin{pmatrix} -1 & \nu\epsilon \\ -\nu\epsilon & 2 \end{pmatrix}, \quad \mathbf{w}_p(x) := \begin{pmatrix} \sin 5x + \cos 10x \\ \cos 7x \end{pmatrix},$$

and consider the BC for $a=0$, $b=1$,

$$B_a = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \quad B_b = \begin{pmatrix} \cos\nu & -\sin\nu \\ 0 & 0 \end{pmatrix}, \quad \beta = 0.$$

These BC correspond to (4.4). Using the transformation $\mathbf{w} := T^{-1}\mathbf{y}$ we have

$$\epsilon \mathbf{w}' = \hat{A}\mathbf{w} + T^{-1}(x;\nu)\mathbf{q}(x),$$

for which $\mathbf{w}_p(x)$ is a particular solution. The fundamental solution for the ODE in $\mathbf{w}$ can be easily found, so $\mathbf{y}(x)$ can be found as well, and the BVP can be verified to be well-conditioned for $0<\epsilon\ll\nu$. However, the matrix B_a+B_b is singular for $\nu = \pi(l+1/2)$, l=0,1,..., so we expect stability problems with symmetric difference (or collocation) schemes for these values of ν.

In [AMR, Tables 10.4,10.5] we list numerical results for this example, using collocation at Gaussian points for two values of ν: $\nu=\pi/4$, where no stability trouble occurs and the theoretical results of [AsWe2] are clearly demonstrated, and the troublesome case $\nu=\pi/2$, where the results are generally poor.

Note that the instability demonstrated in this example depends on the coarse mesh being coarse throughout the smooth solution segment. A simple way to get rid of this instability, proposed in [dHMa1], is to add a few mesh points $O(\epsilon)$ apart in the middle of the interval. This creates a "layer" region there which separates the original long segment into two segments with coarse meshes. We may then consider the ODE on each of these segments, with BC as in (4.4). With the changed segment lengths, $\hat{\Phi}_2$ and $\hat{\Phi}_4$ are both nicely bounded. Theorem 6.9 may now be invoked to prove that the resulting scheme is stable, justifying the suggestion of [dHMa1]. Computational results comparable to those for $\nu=\pi/4$ are obtained.

□

References

[As1] U. Ascher, "On some difference schemes for singular singularly-perturbed boundary value problems", Numer. Math. 46 (1985), 1-30.

[As2] U. Ascher, "Two families of symmetric difference schemes for singular perturbation problems", in *Numerical Boundary Value ODEs,* U. Ascher and R.D. Russell (eds.), Progress in Scientific Computing Vol. 5, Birkhauser, Boston, 1985.

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

[AsJa] U. Ascher and S. Jacobs, "On collocation implementation for singularly perturbed two-point problems", U. British Columbia Computer Science Tech. Rep. 86-19, 1986.

[AMR] U. Ascher, R.M.M Mattheij and R.D. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations,* Prentice-Hall (1987).

[AsWe1] U. Ascher and R. Weiss, "Collocation for singular perturbation problems I: first order systems with constant coefficients", SIAM J. Numer. Anal. 20 (1983), 537-557.

[AsWe2] U. Ascher and R. Weiss, "Collocation for singular perturbation problems II : linear first order systems without turning points", Math. Comp. 43 (1984), 157-187.

[BrLo] D. Brown and J. Lorenz, "A high order method for stiff BVPs with turning points", SIAM J. Sci. Stat. Comp. (1987).

[Co] W.A. Coppel, *Dichotomies in Stability Theory,* Lecture Notes in Mathematics vol. 629, Springer-Verlag, Berlin, 1978.

[Da] A. Davey, "An automatic orthonormalization method for solving stiff BVPs", J. Comp. Phys. 51 (1983), 343-356.

[DOR] L. Dieci, M.R. Osborne and R.D. Russell, "A Riccati transformation method for solving boundary value problems, I: theoretical aspects", 1986.

[DiRu] L. Dieci and R.D. Russell, "Riccati and other methods for singularly perturbed BVP", in Proceedings of BAIL IV, J. Miller (ed.), 1986.

[dHMa1] F. de Hoog and R.M.M. Mattheij, "Conditioning, dichotomy, and scaling for two-point BVPs", in *Numerical Boundary Value ODEs,* U. Ascher and R.D. Russell (eds.), Progress in Scientific Computing Vol. 5, Birkhauser, Boston, 1985.

[dHMa2] F. de Hoog and R.M.M. Mattheij, "On dichotomy and well-conditioning in BVP", SIAM J. Numer. Anal. 24 (1987), 89-105.

[Kr] H.-O. Kreiss, "Centered difference approximation to singular systems of ODEs", Symposia Mathematica X, Inst. Nazionale di Alta Math., 1972.

[KNB] H.-O. Kreiss, N.K. Nichols, and D.L. Brown, "Numerical methods for stiff two-point boundary value problems", SIAM J. Num. Anal. 23 (1986), 325-368.

[LoMa] P.M. van Loon and R.M.M. Mattheij, "Stable continuous orthonormalisation techniques for linear boundary value problems", to appear in J. Austr. Math. Soc. Series B (1987).

[Ma] R.M.M. Mattheij, "Decoupling and Stability of Algorithms for Boundary Value Problems", SIAM Review (1985), 1-44.

[Me] G.H. Meyer, "Continuous orthonormalization for boundary value problems", J. Comp. Phys. 62 (1986), 248-262.

[We] R. Weiss, "An analysis of the box and trapezoidal schemes for linear singularly perturbed boundary value problems", Math Comp. 42 (1984), 41-67.