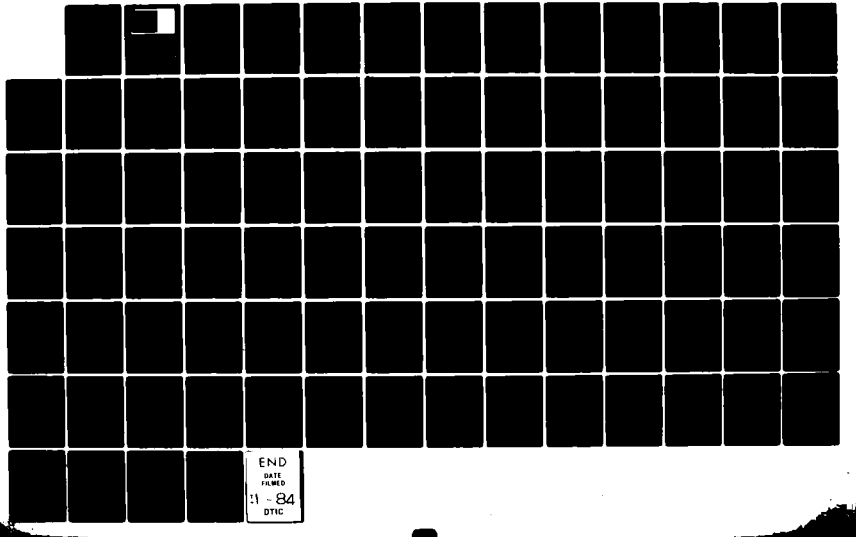
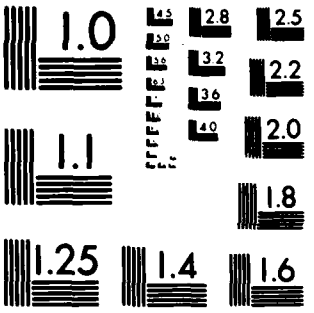
AD-A136 427 NUMERICAL METHODS FOR STIFF TWO-POINT BOUNDARY VALUE PROBLEMS(U) WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER H KREISS ET AL. NOV 83 MRC-TSR-2599 1/1

UNCLASSIFIED N00014-83-K-0422 F/G 12/1 NL



END
DATE FILMED
11-84
DTIC



MRC Technical Summary Report #2599

NUMERICAL METHODS FOR STIFF TWO-POINT BOUNDARY VALUE PROBLEMS

Heinz-Otto Kreiss, N. K. Nichols,
and David L. Brown

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53705

November 1983

(Received September 27, 1983)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

Office of Naval
Research
800 North Quincy St.
Arlington, VA 22217

Air Force Office of
Scientific Research
Building 410, Bolling AFB
Washington, DC 20332

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

NUMERICAL METHODS FOR STIFF TWO-POINT BOUNDARY VALUE PROBLEMS†

Heinz-Otto Kreiss*, N. K. Nichols**, and David L. Brown*

Technical Summary Report #2599
November 1983

ABSTRACT

The authors

We consider the two-point boundary value problem for stiff systems of ordinary differential equations. For systems that can be transformed to essentially diagonally dominant form with appropriate smoothness conditions, a priori estimates are obtained. Problems with turning points can be treated with this theory, and we discuss this in detail. They give robust difference approximations and present error estimates for these schemes. In particular they give a detailed description of how to transform a general system to essentially diagonally dominant form and then stretch the independent variable so that the system will satisfy the correct smoothness conditions. Numerical examples are presented for both linear and nonlinear problems.

AMS (MOS) Subject Classification: 65L10

Key Words: Stiff, ODE, Boundary value problem, Turning points, Difference equations, Essentially diagonally dominant form, Robust, Upwinding

Work Unit Number 3 (Numerical Analysis and Scientific Computing)

---

†All numerical computations were made on the Caltech Applied Mathematics Department Fluid Dynamics VAX.

*Department of Applied Mathematics, California Institute of Technology 217-50, Pasadena, CA 91125.

**Department of Mathematics, University of Reading, England RG6-2AH.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.
Research partially supported by:
Office of Naval Research Contract No. N00014-83-K-0422; and Air Force Office of Scientific Research Contract No. AFOSR-82-0321.

NUMERICAL METHODS FOR STIFF TWO-POINT BOUNDARY VALUE PROBLEMS†

Heinz-Otto Kreiss*, N. K. Nichols**, and David L. Brown*

1. Introduction

In this paper we consider the two-point boundary value problem for a linear system of n ordinary differential equations (ODEs)

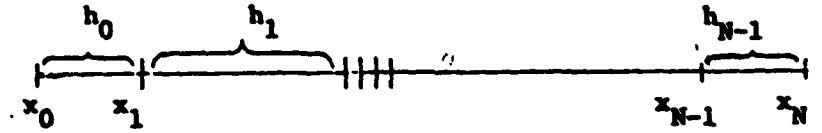
$$\frac{dy}{dx} = A(x)y + F(x), \qquad 0 \le x \le c, \tag{1.1}$$

subject to n linearly independent boundary conditions

$$B_0 y(0) + B_1 y(c) = g. \tag{1.2}$$

Here $y^T = (y^{(1)}, y^{(2)}, \cdots, y^{(n)})$ [1] is a vector function with n components and B_0, B_1 and $A(x) \in C^p$ [2] are $n \times n$ matrices. We assume furthermore that the vector $F(x) \in C^p$.

We are particularly interested in the case when the problem (1.1)-(1.2) can be characterized as being *stiff* but not highly oscillatory. The "stiffness" of a system of ordinary differential equations is defined relative to a computational grid on which the system is to be solved by means of e.g. a difference approximation:



We divide the x-axis into subintervals of variable length h_j with gridpoints

---

†All numerical computations were made on the Caltech Applied Mathematics Department Fluid Dynamics VAX.

*Department of Applied Mathematics, California Institute of Technology 217-50, Pasadena, CA 91125.

**Department of Mathematics, University of Reading, England RG6-2AH.

[1] If y is a vector then y^T denotes its transpose and y^* its adjoint. The vector norm is defined by $|y| = \max |y^{(i)}|$. Similar notations hold for matrices, for example $|A| = \sup_y |Ay| / |y|$. Furthermore for vector functions $\|y(x)\|_{0,c} = \max_{0 \le x \le c} |y(x)|$ denotes the maximum norm.

[2] $A(x) \in C^j$ if the elements of A are j times continuously differentiable.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.
Research partially supported by:
Office of Naval Research Contract No. N00014-83-K-0422; and Air Force Office of Scientific Research Contract No. AFOSR-82-0321.

$x_0 = 0$, $x_\nu = \sum_{j=0}^{\nu-1} h_j$, $\nu = 1, 2, \ldots, N$; $x_N = c$ and denote by $u_\nu = u(x_\nu)$ vector functions defined on the grid. Defining $h = \max_j h_j$, we say that the system (1.1) is *stiff* if $h|A| \gg 1$. The case when it is possible to choose h such that $h|A| \ll 1$ everywhere has been treated many times before and it is not our aim to discuss that situation here.

There are essentially two features of stiff boundary value problems that makes their solution by numerical methods difficult. One is that the matrix A has large eigenvalues of both signs. The second is that there may be *turning points* in the problem. The concept of a turning point is not particularly well defined in the literature; however for our purposes, we take it to mean a subinterval of $0 \le x \le c$ over which an eigenvalue of A changes its order of magnitude: In general, if the eigenvalues of A vary over several orders of magnitude as a function of x, there are difficulties.

It is typical that the solutions of stiff boundary value problems vary over several different scales. Therefore, it is intuitively clear that in order to obtain an accurate numerical solution to such problems the computational grid must be constructed in such a way that the solution of the problem is everywhere smooth with respect to the grid. Alternately, we can think of changing the original problem by introducing a "stretched" variable $\tilde{x} = \tilde{x}(x)$ such that the solutions $y(\tilde{x})$ of the transformed problem

$$\tilde{x}'(x)\,\frac{dy}{d\tilde{x}} = Ay + F$$

varies slowly with respect to a *uniform* grid $h_\nu \equiv h$. Since $y(x)$ can change by several orders of magnitude we have to be careful about what we mean by "varying slowly". We must first scale y correctly. We assume that this is achieved by a positive scaling function $\varphi \in C^p$. The smoothness constant of this scaling

function is defined by

$$\hat{K} := \hat{K}_{[a,b],p} := \sup_{a \le x \le b} \max_{0<\nu \le p} \left| (d^\nu \varphi / dx^\nu) / \varphi \right|.$$

Thus φ can grow like $e^{\hat{K}x}$ or decay like $e^{-\hat{K}x}$. If, for example, $\hat{K} = 10$, then we can obtain a change of scale over a short interval.

Definition 1.1: A scaling function varies slowly for $a \le x \le b$ with respect to a uniform gridlength h if $hK \ll 1$.

We now consider y and assume that we have chosen a slowly varying scaling function φ such that $y = \varphi \tilde{y}$. We can now think of $\tilde{y}$ as being of order $O(1)$ and therefore we define its smoothness constant $\tilde{K}$ by

$$\tilde{K} := \tilde{K}_{[a,b],p} := \sup_{a \le x \le b} \max_{0<\nu \le p} \left| d^\nu \tilde{y} / dx^\nu \right|.$$

This leads to

Definition 1.2: y varies slowly for $a \le x \le b$ with respect to a scaling function φ and a uniform mesh with meshsize h if

$$Kh \ll 1, \quad K = \max(\tilde{K}, \hat{K})$$

The definition above is useful for numerical purposes only if we can determine the smoothness constant K without detailed knowledge of $y(x)$. In the next section we begin by considering scalar equations and show that K can be determined in terms of the properties of the coefficients of the differential equation. In sections 3 and 4 these results are then generalized to systems of ODEs which can be smoothly transformed to essentially diagonally dominant form.

For the problem (1.1),(1.2), the constant K can be determined in every subinterval $b_1 \le x \le b_2$ of $0 \le x \le c$. If K does not change by orders of magnitude from subinterval to subinterval, then we can use a uniform meshsize h with

$K\bar{h} \ll 1$ everywhere on the whole interval. However, as we pointed out above, it is typical in stiff problems that there are several different scales on which the solution varies; for example there can be boundary and internal layers. These variations manifest themselves in values of K which vary over several orders of magnitude between subintervals. In sections 9 and 10 we discuss how to construct the stretched variable $\tilde{x}(x)$ mentioned above so that the new smoothness constant is of the same order everywhere. This stretched variable leads to a nonuniform mesh in the old variable x.

In the remainder of the paper we discuss difference approximations for stiff systems and give numerical examples. There are two basic classes of difference approximations that can be used; these are centered schemes and onesided schemes. Onesided schemes have the apparent disadvantage that the differential equation must be transformed to an appropriate form before they can be applied. However, centered schemes are not as reliable as onesided schemes, and as we show in section 5, they cannot be expected to give acceptable results in general unless the system has been transformed to the proper blocked form. Even then, the combination of onesided and centered schemes which we advocate in that section will perform better than a strictly centered scheme. This is true basically because the fundamental estimates for the former more closely mimic those of the differential equation than do the estimates for the centered schemes.

In sections 6 through 8 we discuss difference approximations in more detail. Section 6 is concerned with difference approximations for scalar equations, and sections 7 and 8 cover diagonally and essentially diagonally dominant systems. In section 9 we describe how to transform a general system to essentially diagonally form. Finally, in sections 10 and 11 we give a detailed description of our numerical procedures and present some numerical results.

2. Analytic properties of scalar equations.

In this section we consider scalar complex valued equations

$$\begin{aligned} &\frac{dy}{dx} = a(x)y + f(x), \\ &y(0) = y_0, \quad 0 \le x \le c \qquad \mathrm{Re}\, a < 0. \end{aligned} \tag{2.1}$$

We are interested in the case that $a(x)$ is large but we are not interested in the case that $y(x)$ is highly oscillatory. Therefore we assume that there is a constant $\rho \sim O(1)$ such that

$$|a_I(x)| \le \rho |a_R| \qquad a_I = \mathrm{Im} a, \qquad a_R = \mathrm{Re} a < 0. \tag{2.2}$$

Later on we want to solve (2.1) by difference approximation on a grid with essentially uniform gridsize h. Because our difference approximation will only depend on point values of $a(x)$ and $f(x)$ it is important that the behavior of these point values represents well the behavior of the continuous functions. An appropriate assumption is that there exist a natural number $p > 0$ and a constant $K_1 > 1$ with $K_1 h \ll 1$ such that

$$\begin{aligned} &\| (|a| + 1)^{-1} \frac{d^\nu a}{dx^\nu} \|_{0,c} \le K_1, \quad \| (|f| + 1)^{-1} \frac{d^\nu f}{dx^\nu} \|_{0,c} \le K_1, \\ &\qquad \nu = 1,2,\ldots,p. \end{aligned} \tag{2.3}$$

The number p, which will depend on the method we use, will be chosen later.

We can use (2.3) to estimate the local variation of $a(x)$ and $f(x)$. Letting $\varphi = (f + sgn f)^{-1} df / dx$ and $\psi = \varphi sgn f$ we can write

$$\frac{df}{dx} = \varphi(x) f(x) + \psi(x)$$

i.e.

$$f(x) = \int_0^x e^{\int_\eta^x \varphi(t)dt} \psi(\eta) d\eta + e^{\int_0^x \varphi(t)dt} f(0)$$

or

$$f(x) = e^{xg_1(x)} f(0) + xg_2(x)$$

where by (2.3)

$$|g(x)| \le K_1, \quad |g_2(x)| \le K_1 e^{K_1 x}.$$

Therefore for $|K_1 x| \le 1$

$$|f(x) - f(0)| \le |e^{xg_1} - 1|\,|f(0)| + |x|\,|g_2| \le |K_1 x|(1 + O(|K_1 x|))(|f(0)| + 1).$$

A similar estimate holds for $a(x)$. Thus we have

Lemma 2.1: For all x_1 with $K_1|x_1 - x_0| \le 1$

$$|a(x_1) - a(x_0)| \le |K_1(x_1 - x_0)|(1 + O(|K_1(x_1 - x_0)|))|a(x_0)|,$$
$$|f(x_1) - f(x_0)| \le |K_1(x_1 - x_0)|(1 + O(|K_1(x_1 - x_0)|))\,(|f(x_0)| + 1).$$

If the expressions (2.3) are not bounded then we cannot expect that the solution of (2.1) varies slowly. As an example, consider the differential equation

$$\frac{dy}{dx} = -\frac{x + \varepsilon}{\varepsilon} y + \varepsilon^{-\frac{1}{2}}, \quad y(0) = 0, \quad 0 \le x \le 1.$$

Here $\varepsilon > 0$ is a small constant (for example $\varepsilon = 10^{-6}$). Now

$$(|a(x)| + 1)^{-1} |da(x)/dx| = |(x + 2\varepsilon)^{-1}|$$

becomes arbitrarily large for $\varepsilon \to 0$. For $x\sqrt{\varepsilon} \gg 1$ the solution is to first approximation given by

$$y = \frac{\varepsilon^{\frac{1}{2}}}{x + \varepsilon}$$

and varies slowly. However, for $0 \le x \le const.\sqrt{\varepsilon}$ it changes rapidly. To see this we introduce a new variable $\tilde{x} = x/\sqrt{\varepsilon}$ and obtain

$$dy/d\tilde{x} = -(\tilde{x} + \varepsilon^{\frac{1}{2}})y + 1, \quad y(0) = 0$$

i.e. $y(x)$ varies slowly as a function of $\tilde{x}$ but not as a function of x. This can also be seen from the graph of $y(x)$.

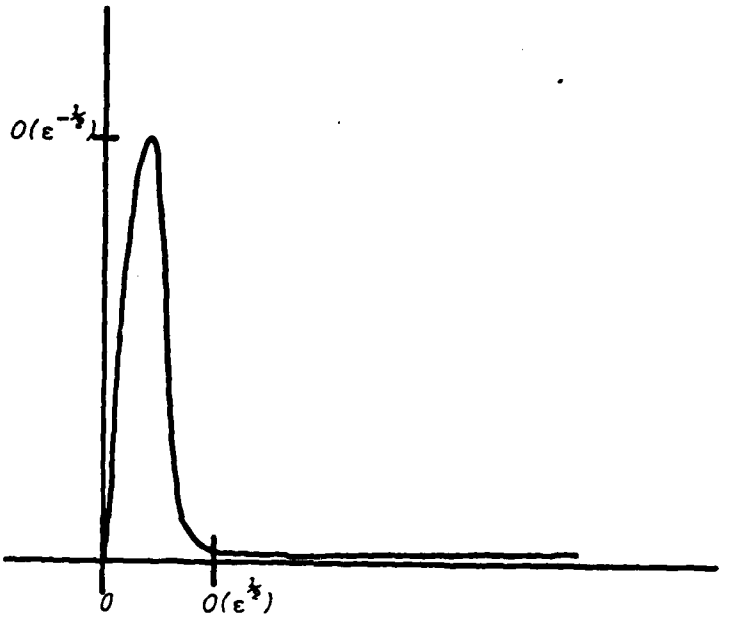


Fig. 2.1

The conditions (2.2), (2.3) still do not guarantee that the solutions of the differential equation change slowly. Consider, for example,

$$dy/dx = -\varepsilon^{-1}y, \quad y(0) = 1, \quad 0 < \varepsilon \ll 1.$$

All the above conditions are satisfied. However the solution is given by

$$y(x) = e^{-x/\varepsilon}$$

which forms a boundary layer and decays rapidly from 1 to zero. This possibility can be avoided if we assume also that $a(0)$ is not too large, i.e. there is a constant $K_2 > 1$ with $K_2 h \ll 1$ such that

$$|a_R(0)| \le K_2, \quad \text{i.e.} \quad |a(0)| \le (\rho + 1)K_2. \tag{2.4}$$

Observe that the conditions (2.3), (2.4) do not prevent $a(x)$ from becoming large because (2.3) allows a rather rapid exponential growth of $a(x)$ away from the boundary. Numerically this corresponds to the requirement that we use an exponentially stretched mesh to represent boundary layers in the solution.

The above conditions are not sufficient to guarantee that the solution $y(x)$ stays bounded. Therefore we assume explicitly that the equation has been properly scaled such that

$$\|y(x)\|_{0,c} \le 1. \tag{2.5}$$

The following lemma says that (2.5) implies the condition $|f|/(|a| + 1) \le$ constant.

Lemma 2.2: Assume that $c \ge 2/K_1$. Then there is a constant C_3 which depends only on K_1 such that

$$\left\| \frac{f(x)}{|a(x)| + 1} \right\|_{0,c} \le C_3 \|y\|_{0,c}. \tag{2.5a}$$

Proof. Without restriction we can assume that $\|y\|_{0,c} = 1$. Let x_0 be a point with

$$\left| \frac{f(x_0)}{|a(x_0)| + 1} \right| \ge 1.$$

If $|c - x_0| \ge 1/K_1$, we consider (2.1) for $x \ge x_0$ and introduce new variables

$$\tilde{x} = \frac{(x - x_0)}{|a(x_0)| + 1}, \quad \tilde{y} = \frac{|a(x_0)| + 1}{|f(x_0)|} y.$$

Then

$$d\tilde{y} / d\tilde{x} = (1 + |a(x_0)|)^{-1} a(x)\tilde{y} + f(x)/|f(x_0)|.$$

By lemma 2.1

$$\frac{|a(x)|}{|a(x_0)| + 1} \leq \left| \frac{a(x_0)}{|a(x_0)| + 1} \right| + \left| \frac{a(x) - a(x_0)}{|a(x_0)| + 1} \right|$$

$$\leq 1 + O(|K_1(x - x_0)|).$$

and $\frac{|f(x)|}{|f(x_0)|} \geq 1 - O(|K_1(x - x_0)|)$. Therefore

$$\tilde{y}(x_1) = \int_{x_0}^{x_1} \frac{f(x)}{f(x_0)} d\tilde{x} + \tilde{y}(x_0) + \int_{x_0}^{x_1} \frac{a(x)}{1 + |a(x_0)|} \tilde{y}\, d\tilde{x}$$

implies

$$\|\tilde{y}\|_{x_0, x_1} \geq |\tilde{y}(x_1)| \geq (x_1 - x_0)(1 - O(|K_1(x_1 - x_0)|) -$$

$$(x_1 - x_0)(1 + O(|K_1(x_1 - x_0)|)\|\tilde{y}\|_{x_0, x_1}.$$

Thus for $|x_1 - x_0| = \delta / K_1$, $\delta > 0$ sufficiently small,

$$\|\tilde{y}\|_{x_0, x_1} = \frac{|a(x_0)| + 1}{|f(x_0)|} \|y\|_{x_0, x_1} \geq \frac{\delta}{2K_1}.$$

Since the last inequality holds for any x_0, the inequality of the lemma holds in this case with $C_3 = 2K_1/\delta$. If $|c - x_0| < 1/K_1$ then by considering (2.1) for $x \leq x_0$ instead, we obtain the desired inequality in a similar way. This proves the lemma.

We shall now prove that under the conditions of (2.2)-(2.5) there is a constant C such that the solution $y(x)$ of (2.1) satisfies

$$\|d^\nu y / dx^\nu\|_{0,c} \leq C. \qquad \nu = 1,2,\ldots,p. \tag{2.6}$$

This is easiest to show when $|a(x)|$ is small for all x: We have

Lemma 2.3: Assume that (2.5a) holds, that conditions (2.2)-(2.5) are satisfied and that

$$\|a_R(x)\|_{0,c} \le K_2.$$

Then there is a constant C which depends only on K_1, K_2, ρ such that (2.6) holds.

Proof: (2.2) implies

$$\|a(x)\|_{0,c} \le (\rho + 1)K_2.$$

Therefore by (2.3)

$$\|d^\nu a / dx^\nu\|_{0,c} \le K_1((\rho + 1)K_2+1) \quad \nu = 1,2,\ldots,p.$$

By (2.5a) and (2.3)

$$\|f(x)\|_{0,c} \le C_3(\|a\|_{0,c} + 1) \quad i.e. \quad \|d^\nu f / dx^\nu\|_{0,c} \le \text{const.}, \quad \nu = 1,2,\ldots,p.$$

Using the differential equation, (2.6) follows.

We assume now that

$$a_R(x) \le -1, \tag{2.7}$$

i.e. we allow $|a_R|$ to become arbitrarily large. We will use

Lemma 2.4: The solution of (2.1) with $a_R(x) = \mathrm{Re}\, a < 0$ satisfies the estimates

$$|y(x)| \le x \max_{0<\eta<x} |f(\eta)| + e^{\int_0^x a_R(\xi)d\xi} |y(0)|, \tag{2.8a}$$

$$|y(x)| \le \max_{0<\eta<x} |f(\eta)/a_R(x)| + e^{\int_0^x a_R(\xi)d\xi} |y(0)|. \tag{2.8b}$$

Proof: The solution of (2.1) can be written down explicitly:

$$y(x) = \int_0^x e^{\int_\eta^x a(\xi)d\xi} f(\eta)d\eta + e^{\int_0^x a(\xi)d\xi} y(0). \qquad \text{—}$$

The first estimate follows from the inequality

$$|e^{\int_\eta^x a(\xi)d\xi}| \le e^{\int_\eta^x a_R(\xi)d\xi} \le 1.$$

Furthermore

$$|y(x)| \le \int_0^x e^{\int_\eta^x a_R(\xi)d\xi} |f(\eta)|d\eta + |e^{\int_0^x a_R(\xi)d\xi}||y(0)| \le$$

$$\left| \int_0^x a_R(\eta) e^{\int_\eta^x a_R(\xi)d\xi} |f(\eta)/a_R(\eta)|d\eta \right| + e^{\int_0^x a_R(\xi)d\xi} |y(0)| \le$$

$$\left| \int_0^x \frac{d}{d\eta} e^{\int_\eta^x a_R(\xi)d\xi} d\eta \right| \cdot \max_{0\le\eta\le x} |f(\eta)/a_R(\eta)| + e^{\int_0^x a_R(\xi)d\xi} |y(0)|$$

which gives us (2.8b).

Using lemma 2.4 we can obtain

Lemma 2.5: Consider (2.1) and assume that the condition (2.2) is satisfied. Then there is a constant C_1 which depends only on ρ such that

$$|d^\nu y(x)/dx^\nu| \le \qquad (2.9a)$$

$$C_1 \left\{ \sum_{j=0}^{\nu-1} \| a^{-1} \frac{d^{j+1}a}{dx^{j+1}} \|_{0,x} \| \frac{d^j y}{dx^j} \|_{0,x} + \| a^{-1} \frac{d^\nu f}{dx^\nu} \|_{0,x} + | \frac{d^\nu y(0)}{dx^\nu} | \right\} \quad ,\nu = 0,1,2,\ldots,p.$$

If furthermore (2.3), (2.5a), and (2.7) hold, then there is a constant C_2 which depends only on K_1 and ρ such that

$$|d^{\nu}y(x)/dx^{\nu}| \le C_2\left\{\|y\|_{0,s} + 1 + \sum_{j=1}^{\nu}\left|\frac{d^j y(0)}{dx^j}\right|\right\} \tag{2.9b}$$

Proof: For $\nu = 0$ the first two estimates follow from (2.2) and lemma 2.4 since

$$|f/a_R| = |f/a||a/a_R| \le (1+\rho)\,|f/a|.$$

Now let $u = dy/dx$. Differentiating (2.1) gives us

$$\frac{du}{dx} = au + \left(\frac{da}{dx}\right)y + \frac{df}{dx}. \tag{2.10}$$

Thus by lemma 2.4

$$\|du/dx\|_{0,s} \le \|a_R^{-1}da/dx\|_{0,s}\|y\|_{0,s} + \|a_R^{-1}df/dx\|_{0,s} + |dy/dx|_{x=0}.$$

Now

$$\|a_R^{-1}da/dx\|_{0,s} \le (1+\rho)\|a^{-1}da/dx\|_{0,s}$$
$$\|a_R^{-1}df/dx\|_{0,s} \le (1+\rho)\|a^{-1}df/dx\|_{0,s} \le (1+\rho)\left\|\frac{|f|+1}{|a|}\right\|_{0,s}\left\|\frac{df/dx}{|f|+1}\right\|_{0,s}$$

and if $|a| \ge 1$ we have also that

$$\|a^{-1}da/dx\|_{0,s} \le 2\|(|a|+1)^{-1}da/dx\|_{0,s}$$

and

$$\left\|\frac{(|f|+1)}{a}\right\|_{0,s} \le 2\left\|\frac{f}{|a|+1}\right\|_{0,s} + 1$$

and so (2.9a,b) follow for $\nu = 1$ from (2.2), (2.3), (2.5a) and (2.7).

The estimates for higher derivatives are obtained by repeated differentiation of (2.1). This proves the lemma.

We can now prove the main result of this section

Theorem 2.1: Assume that $C \geq 2/K_1$ and that the conditions (2.2)-(2.5) hold. Then there is a constant C which depends only on K_1, K_2, ρ such that (2.6) is valid.

Proof: Lemma 2.2 tells us that (2.5a) holds. Now divide the interval $0 \leq x \leq c$ into as few subintervals $c_\tau \leq x \leq c_{\tau+1}$ as possible such that at least one of the two conditions

$$|a_R(x)| \leq K_2, \quad a_R(x) \leq -1$$

holds in the whole subinterval. If $|a_R| \leq K_2$ is valid then we can estimate the derivatives by lemma 2.3. If $a_R \leq -1$ then we can obtain the estimate using lemmas 2.2 and 2.5 if we have a bound for $\sum_{j=0}^{\nu} |d^j y(c_\tau)/dx^j|$. The interval $0 = c_0 \leq x \leq c_1$ is included in the case $|a_R| \leq K_2$, so we are only concerned with the remaining intermediate points c_τ, $\tau = 1,2,\ldots$. For these points we obtain this bound from the estimate for the previous subinterval $c_{\tau-1} \leq x \leq c_\tau$. This proves the theorem.

Finally, we can eliminate the condition (2.5), i.e. the assumption that we have scaled the solution beforehand. If $\|y\|_{0,c} > 1$ then $\tilde{y} = y/\|y\|_{0,c}$ satisfies

$$d\tilde{y}/dx = a(x)\tilde{y} + \tilde{f}, \quad \tilde{f} = f/\|y\|_{0,c}. \tag{2.10}$$

Now

$$\|(|\tilde{f}| + 1)^{-1} d^\nu \tilde{f}/dx^\nu\|_{0,c} \leq \|(|f| + 1)^{-1} d^\nu f/dx^\nu\|_{0,c}$$

implies that (2.10) satisfies all our conditions. Thus

$$\begin{aligned} &\|d^\nu \tilde{y}/dx^\nu\|_{0,c} \leq C, \\ &\text{i.e.} \quad \|d^\nu y/dx^\nu\|_{0,c} \leq C\|y\|_{0,c} \quad \text{for} \quad \|y\|_{0,c} > 1. \end{aligned} \tag{2.11}$$

Combining (2.6) and (2.11) gives us

Theorem 2.2: If the conditions (2.2)-(2.4) are satisfied, then we obtain the estimate

$$\| d^{\nu} y / dx^{\nu} \|_{0,c} \leq C(\| y \|_{0,c} + 1), \quad \nu = 1,2,\ldots,p. \tag{2.12}$$

Remark: If $hC \ll 1$ then $y(x)$ is slowly varying with respect to the scaling function $\varphi = \| y \|_{0,c}$. We can obtain a much more precise result by using $\varphi = \sqrt{(|f|^2 + 1)/(|a|^2 + 1)}$ as a scaling function. To explain this we assume that $a \gg 1$, $f \gg 1$ and use $\varphi = |f/a|$. Then

$$\left| \frac{d\varphi / dx}{\varphi} \right| \leq |f'/f| + |a'/a| \sim 2K$$

and $\tilde{y} = y/\varphi$ satisfies

$$\tilde{y}' = (a - \varphi'/\varphi)\tilde{y} + a, \quad i.e. \quad |\tilde{y}| \approx |a/(a-\varphi)| \sim 1.$$

3. Diagonally dominant systems.

We consider now systems of differential equations. A reasonable assumption for such systems is that the coefficients change slowly, i.e. that they satisfy conditions of the same type as we described for scalar equations. Unfortunately, this assumption is not sufficient to guarantee that the solutions of the systems also vary slowly. Consider, for example, the system

$$\frac{d}{dx}\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 0 & x^2 \\ \varepsilon^{-1} & 0 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} - \begin{bmatrix} 0 \\ \varepsilon^{-1/2}x \end{bmatrix}, \quad -1 \le x \le 1, \quad 0 < \varepsilon \ll 1,$$

$$v(-1) = \alpha, \quad v(1) = \beta \tag{3.1}$$

In a neighborhood of $x = 0$ the functions x^2, ε^{-1}, $\varepsilon^{1/2}x$ satisfy the conditions (2.3) with $K_1 = 1$. However, the solution is not smooth. In component form (3.1) is given by

$$u' = x^2 v, \quad \varepsilon v' = u - \varepsilon^{1/2}x$$

i.e.

$$\varepsilon v'' = x^2 v - \varepsilon^{1/2}.$$

A graph of the solution is given in fig. 3.1.

There is however a class of problems that behave like scalar equations and we shall describe this class now.

Definition 3.1: The matrix function

$$A(x) = \begin{bmatrix} a_{11}(x) & a_{12}(x) & \cdots & a_{1m}(x) \\ a_{21}(x) & \cdots & \cdots & a_{2m}(x) \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1}(x) & \cdots & \cdots & a_{mm}(x) \end{bmatrix}$$

is diagonally dominated if

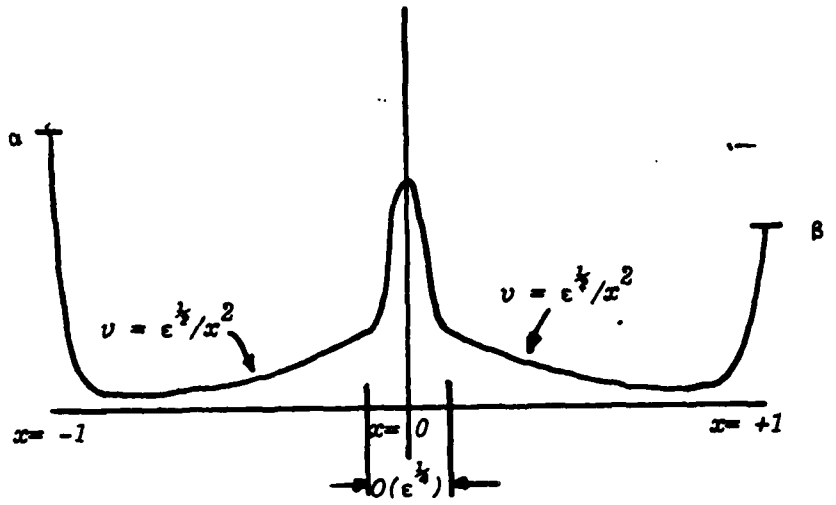


Fig. 3.1

α)

$$\begin{aligned} &\mathrm{Re}\,a_{ii} < 0 \quad i = 1,2,\ldots,r, \\ &\mathrm{Re}\,a_{ii} > 0 \quad i = r+1,\ldots,m. \end{aligned}$$

β) There is a constant $\delta > 0$ independent of x such that

$$\sum_{j \neq i}^{m} |a_{ij}| \leq (1-\delta)|\mathrm{Re}\,a_{ii}|, \quad i = 1,2,\ldots,m$$

γ) There is a constant $\rho \sim O(1)$ such that

$$|\mathrm{Im}\,a_{ii}| \leq \rho |\mathrm{Re}\,a_{ii}|, \quad i = 1,2,\ldots,m.$$

If A is diagonally dominated then it is appropriate to write the system (1.1) in the form

$$\frac{dy}{dx} = \Lambda(I + B)y + F, \quad 0 \leq x \leq c \tag{3.2}$$

where

$$\Lambda = \begin{pmatrix} \Lambda^I & 0 \\ 0 & \Lambda^{II} \end{pmatrix}, \quad B = \begin{pmatrix} 0 & b_{12} & b_{13} & \cdots & b_{1m} \\ b_{21} & 0 & b_{23} & \cdots & b_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ b_{m1} & \cdots & \cdots & \cdots\, b_{m,m-1} & 0 \end{pmatrix}$$

with

$$\Lambda^I = diag\,(a_{11},a_{22},\cdots,a_{rr}),\quad \Lambda^{II} = diag\,(a_{r+1,r+1},a_{r+2,r+2},\cdots,a_{mm}).$$

Such a system is said to be in ***diagonally dominant form***. Corresponding to (2.3) and (2.4) we make

Assumption 3.1: There is a constant K_1 such that

$$\|\Lambda^{-1}\frac{d^\nu \Lambda}{dx^\nu}\|_{0,c} \le K_1,\qquad \|\frac{d^\nu B}{dx^\nu}\|_{0,c} \le K_1,\qquad \nu = 1,2,\ldots,p\ .$$

Assumption 3.2

$$\|\Lambda^{-1}\frac{d^\nu F}{dx^\nu}\|_{0,c} \le \tilde{K}_1,\qquad \nu = 1,2,\ldots,p.$$

Remark: Corresponding to the previous section we can replace assumption 3.2 with

$$\|(\left|F^{(i)}\right| + 1)^{-1}\frac{d^\nu F^{(i)}}{dx^\nu}\|_{0,c} \le K_1.$$

We shall always assume that $|a_{ii}| \gg 1$ and therefore assumption 3.2 means that we have already scaled the equation properly.

Assumption 3.3: There is a constant K_2 such that

$$\left|\Lambda^I(0)\right| \le K_2,\qquad \left|\Lambda^{II}(c)\right| \le K_2.$$

We shall show that under these conditions the solutions of the system (3.2) change slowly. We start again with a couple of lemmata:

Lemma 3.1: Assume that $A(x)$ is diagonally dominated. Then the solutions of (3.1) satisfy the estimate

$$|y(x)| \le \frac{2}{\omega}\left\{\|(\Lambda + \Lambda^*)^{-1}F\|_{0,c} + s(x)\right\} \tag{3.3}$$

where

$$s(x) = e^{-\frac{1}{2}\int_0^x a^I(\xi)d\xi}|y^I(x)| + e^{\frac{1}{2}\int_x^c a^{II}(\xi)d\xi}|y^{II}(c)|. \tag{3.4}$$

Here we have used the notation

$$y^I(x) = (y^{(1)}, \cdots, y^{(r)})^T, \quad y^{II}(x) = (y^{(r+1)}, \cdots, y^{(m)})^T$$
$$a^I(x) = \min_{1\le\mu\le r}|\mathrm{Re}\,a_\mu(x)|, \quad a^{II}(x) = \min_{r+1\le\mu\le m}|\mathrm{Re}\,a_\mu(x)|.$$

Proof: We consider first the case $y^I(0) = y^{II}(c) = 0$. Let $\mathbf{M}$ denote the space of all continuous functons g with $g^I(0) = g^{II}(c) = 0$. The differential equation

$$L_0 y = \frac{dy}{dx} - \Lambda y = F$$

has a unique solution in $\mathbf{M}$ given by

$$\begin{pmatrix} y^I \\ y^{II} \end{pmatrix} = L_0^{-1}F = \begin{pmatrix} \int_0^x e^{\int_\eta^x \Lambda^I(\xi)d\xi} F^I(\eta)d\eta \\ \int_c^x e^{\int_\eta^x \Lambda^{II}(\xi)d\xi} F^{II}(\eta)d\eta \end{pmatrix}.$$

Let L_1 denote the operator defined by

$$L_1 y = (A - \Lambda)y, \quad y \in \mathbf{M}.$$

Then we can write the differential equation (3.2) in the form

$$(I - L_0^{-1}L_1)y = L_0^{-1}F.$$

In the same way as in lemma 2.2 it follows that

$$\| L_0^{-1} f(x) \|_{0,c} \le 2 \| (\Lambda + \Lambda^*)^{-1} F(x) \|_{0,c}.$$

Furthermore

$$\| L_0^{-1} L_1 y \|_{0,c} \le 2 \| (\Lambda + \Lambda^*)^{-1} L_1 y \|_{0,c} \le (1 - \delta) \| y \|_{0,c}$$

and the estimate (3.3) follows for the case that $y^I(0) = y^{II}(c) = 0$.

Now we consider the case where $f \equiv 0$ and $y^I(0)$, $y^{II}(c)$ are arbitrary. We let y be the solution of (3.2) and write

$$y = v + D y_0, \quad y_0 = \begin{bmatrix} y^I(0) \\ y^{II}(c) \end{bmatrix}, \quad D = \begin{bmatrix} e^{\int_0^x \Lambda^I(\xi) d\xi} & 0 \\ 0 & e^{\int_0^x \Lambda^{II}(\xi) d\xi} \end{bmatrix}.$$

Then v satisfies

$$v' = Av + (A - \Lambda) D y_0, \quad v^I(0) = v^{II}(c) = 0.$$

By (3.3) we obtain

$$\| v(x) \|_{0,c} \le \delta^{-1} \| 2(\Lambda + \Lambda^*)^{-1} (A - \Lambda) \|_{0,c} \| D \|_{0,c} \cdot | y_0 | \le$$
$$\le \delta^{-1} \cdot \max_i \left\{ \sum_{j \ne i} \frac{| a_{ij} |}{| \mathrm{Re} a_{ii} |} \right\} | y_0 | \le \delta^{-1} (1 - \delta) | y_0 |.$$

Thus

$$| y(x) | \le \delta^{-1} (1 - \delta) | y_0 | + | y_0 | = \delta^{-1} | y_0 |. \tag{3.5}$$

To show that in addition to being bounded, $y(x)$ decreases exponentially away from the boundary, we consider the solution of (3.2) with

$$f \equiv 0, \quad y^I(0) = y_0^I, \quad y^{II}(c) = 0.$$

Let $y = e^{p(x)} z$, $p(x) = -\frac{\delta}{2} \int_0^x a^I(\xi) d\xi$. Then $z(x)$ satisfies the equation

$$z'(x) = (A(x) + \frac{\delta}{2}a^I(x)I)z(x).$$

Using (3.5) we obtain

$$|z(x)| \leq \frac{2}{\delta}|y_0^I|.$$

Therefore

$$|y(x)| \leq e^{p(x)}|z(x)| \leq 2\delta^{-1}e^{p(x)}|y_0^I|.$$

The corresponding result holds for $y^I(0) = 0$, $y^{II}(c) = y_1^{II}$. Thus we have proved the complete estimate (3.3).

In the same way as for the scalar equation we can now use the last lemma to discuss the smoothness of the solutions of (3.2).

Lemma 3.2: Consider the system (3.2). Assume that $A(x)$ is diagonally dominated and that the conditions of assumption 3.1 are satisfied. Then there is a constant C_1 which depends only on K_1, ρ, and δ such that

$$\|\frac{d^\nu y}{dx^\nu}\|_{0,c} \leq C_1\left\{\sum_{j=0}^{\nu}\|\Lambda^{-1}d\frac{d^j F}{dx^j}\|_{0,c}\right\} + \sum_{j=0}^{\nu}\left\{\left|\frac{d^j y^I}{dx^j}\right|_{x=0} + \left|\frac{d^j y^{II}}{dx^j}\right|_{x=c}\right\}.$$

Proof: The proof resembles the proof of lemma 2.5 closely. For $\nu = 0$ the estimate is given by lemma 3.1. Let $u = dy/dx$ and differentiate (3.2). Then

$$\frac{du}{dx} = A(x)u + \tilde{F}, \quad \tilde{F} = \frac{dF}{dx} + \left\{\frac{dA}{dx}\right\}y. \tag{3.6}$$

Using the estimate for y we obtain an estimate for $\tilde{F}$, and lemma 3.1 gives us the estimate for u. This process can be continued and the lemma is proved.

We can now prove the main result of this section.

Theorem 3.1: Consider the diagonally dominant system (3.2) and assume that assumptions 3.1-3.3 hold. Then there is a constant C which depends only on K_1, $\tilde{K}_1$, ρ, and δ such that

$$\| \frac{d^\nu y}{dx^\nu} \|_{0,c} \le C\left[\| y \|_{0,c} + 1 \right], \quad \nu = 1,2,\ldots,p. \tag{3.7}$$

Proof: By lemma 3.2 and assumption 3.2

$$\| \frac{d^\nu y}{dx^\nu} \|_{0,c} \le C_1 \left[\nu \tilde{K}_1 + \sum_{j=0}^{\nu} \left[\left| \frac{d^j y^I}{dx^j} \right|_{x=0} + \left| \frac{d^j y^{II}}{dx^j} \right|_{x=c} \right] \right].$$

By assumptions 3.2 and 3.3 and since A is diagonally dominated,

$$\sum_{j=1}^{m} |a_{ij}| \le 2K_2, \quad |F^{(i)}| \le K_2 \tilde{K}_1, \quad i = 1,2,\ldots,r$$

holds at $x = 0$. Therefore the differential equations gives us

$$\left| \frac{dy^I(0)}{dx} \right| \le 2K_2 |y(0)| + K_2 \tilde{K}_1.$$

Correspondingly we get for $x = c$

$$\left| \frac{dy^{II}(c)}{dx} \right| \le 2K_2 |y(c)| + K_2 \tilde{K}_1.$$

Thus by lemma 3.2 we can estimate $\| dy/dx \|_{0,c}$. The estimates for higer derivatives are obtained as before by differentiation.

If $\Lambda^I(0)$, $\Lambda^{II}(c)$ are not $O(1)$ then we introduce an exponential stretching such that they are bounded in the stretched variables. Let

$$a_1 = \max_{1 \le i \le r} |a_{ii}(0)| > 1$$

and introduce new variables by

$$x = a_1^{-1}\int_0^{\tilde{x}} \varphi(\xi)d\xi, \quad \varphi(\xi) = e^{\int_0^{\xi} g(\eta)d\eta}$$

where g will be defined below. Then the system (1.1) becomes

$$\frac{dy}{d\tilde{x}} = a_1^{-1}\varphi(\tilde{x})Ay + a^{-1}\varphi(\tilde{x})F(x). \tag{3.8}$$

Let $\tilde{x}_1 = K_1^{-1}\log a_1$ and choose

$$g(\eta) = \begin{cases} K_1 & \text{for } 0 \le \eta \le \tilde{x}_1 \\ 0 & \text{for } \eta \ge \tilde{x}_1 \end{cases} \qquad \text{i.e.} \quad \varphi(\tilde{x}) = \begin{cases} e^{K_1 x} & \text{for } 0 \le \tilde{x} \le \tilde{x}_1 \\ a_1 & \text{for } \tilde{x} > \tilde{x}_1 \end{cases}$$

Then

$$x = \begin{cases} (e^{K_1 x} - 1)/a_1 K_1 & \text{for } 0 \le \tilde{x} \le \tilde{x}_1 \\ \tilde{x} + x_1 - \tilde{x}_1 & \text{for } \tilde{x} > \tilde{x}_1 \end{cases}, \quad x_1 = (1 - a_1^{-1})/K_1.$$

Now treat the neighborhood of $x = c$ correspondingly. Then $\Lambda^I(0)$, $\Lambda^{II}(c)$ are bounded and assumptions 3.1 and 3.2 hold for $p = 1$ with K_1, $\tilde{K}_1$ replaced by $2K_1$, $2\tilde{K}_1$. Thus the estimate of theorem 3.1 is valid for $p = 1$. In particular $\| \frac{dy}{dx} \|_{x_1, x_2}$ (where $x_2 = c \quad |1 \quad a_2^{-1}| / K_1$, and $a_2 = \max_{r<i\le m} |a_{ii}|$) is already bounded in the unstretched variable because for $x_1 \le x \le x_2$ no stretching occured.

To obtain estimates for higher derivatives we could replace $g(\eta)$ by a smoother function. However, this is not necessary. Apply the stretching to the differential equation (3.6) for $u = dy/dx$. The we get a bound for $du/d\tilde{x}$. On any of the subintervals $0 \le \tilde{x} \le x_1$, $x_1 \le x \le x_2$, $x_2 \le x \le c$ differentiation commutes with stretching and therefore we can estimate the derivatives on every subinterval in terms of $\|y\|_{0,c} + 1$. In particular we have for the subinterval away from the boundaries:

Theorem 3.2: Assume that assumptions 3.1 and 3.2 hold and that $c \geq 2/K_1$. Then there is a constant C which depends only on K_1, $\tilde{K}_1$, K_2, ρ and δ such that

$$\left\| \frac{d^\nu y}{dx^\nu} \right\|_{1/K_1, c-1/K_1} \leq C\left(\|y\|_{0,c} + 1 \right)$$

4. Essentially diagonally dominant systems.

The class of diagonally dominant systems is not broad enough to include many interesting problems. Therefore, in this section we generalize our results to systems which are essentially diagonally dominant or which can be transformed smoothly to systems of that type.

Definition 4.1 A matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot\cdot & \cdot\cdot & \cdots & \cdots \\ a_{n1} & \cdot\cdot & \cdots & a_{nn} \end{pmatrix}$$

is called *essentially diagonally dominated* if there is a constant K_0 with $K_0 h \ll 1$ such that A can be partitioned into the form

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \qquad A_{11} = \Lambda(I + B), \quad \Lambda = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & 0 \\ \cdots & \cdots & \cdots & . \\ 0 & \cdots & \cdots & a_{mm} \end{pmatrix},$$

where A_{11} is diagonally dominated and

$$\| \Lambda^{-1} A_{12} \|_{0,c} \le K_0, \qquad \| A_{2j} \|_{0,c} \le K_0, \quad j = 1,2.$$

We consider now systems (1.1)

$$\frac{d}{dx} \begin{pmatrix} y \\ w \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} y \\ w \end{pmatrix} + \begin{pmatrix} G \\ H \end{pmatrix} \tag{4.1}$$

where A is essentially diagonally dominated. Such a system is said to be in *essentially diagonally dominant form*. We are interested in the case that $h|A_{11}| \gg 1$. We make

Assumption 4.1:

$$\|\Lambda^{-1}d^{\nu}\Lambda/\,dx^{\nu}\|_{0,c} \le K_1, \quad \|\Lambda^{-1}d^{\nu}A_{12}/\,dx^{\nu}\|_{0,c} \le K_1,$$

$$\|d^{\nu}B/\,dx^{\nu}\|_{0,c} \le K_1, \quad \|d^{\nu}A_{2j}/\,dx^{\nu}\|_{0,c} \le K_1, \quad j = 1,2; \quad \nu = 1,2,\ldots,p\ .$$

Assumption 4.2:

$$\|\Lambda^{-1}d^{\nu}G/\,dx^{\nu}\|_{0,c} \le \tilde{K}_1, \qquad \|d^{\nu}H/\,dx^{\nu}\|_{0,c} \le \tilde{K}_1, \quad \nu = 0,1,2,\ldots p.$$

Assumption 4.3:

$$|\Lambda^{I}(0)| \le K_2, \quad |\Lambda^{II}(c)| \le K_2.$$

We want to show that the estimate of theorem 3.1 is still valid.

Theorem 4.1: There is a constant $\tilde{C}$ which depends only on K_1, $\tilde{K}_1$, K_2, ρ and δ such that

$$\|d^{\nu}\tilde{y}/\,dx^{\nu}\|_{0,c} \le \tilde{C}(\|\tilde{y}\|_{0,c} + 1), \qquad \tilde{y} = \begin{Bmatrix} y \\ w \end{Bmatrix}. \tag{4.2}$$

Proof: We have by (4.1) and assumptions (4.1) and (4.2) that

$$\|dw/\,dx\|_{0,c} \le K_1\|\tilde{y}\|_{0,c} + \tilde{K}_1.$$

Write the equations for y in the form

$$dy/\,dx = A_{11}y + G_1 + G$$

where $G_1 = A_{12}w$. Now,

$$\|\Lambda^{-1}dG_1/\,dx\|_{0,c} \le K_0\|dw/\,dx\|_{0,c} + K_1\|w\|_{0,c}\ .$$

so in the same way as the proof of theorem 3.1 we obtain

$$\|dy/\,dx\|_{0,c} \le \text{const.}\ (\|\tilde{y}\|_{0,c} + 1)\ .$$

This proves the theorem for $p = 1$. For higher derivatives it follows by differentiation of the differential equations.

If the assumption 4.3 is not satisfied then we obtain from theorem 3.2 that the derivatives are still bounded in the interior of $0 \le x \le c$. Thus we have

Theorem 4.2: Assume that in the neighborhood $|x_0 - x| \le \beta$, $\beta \ge 2/K_1$ of every point x_0 with $1/K_1 \le x_0 \le c - 1/K_1$ we can write the system (1.1) in the form (4.1) such that assumptions 4.1 and 4.2 hold. Then we obtain

$$\| d^\nu \tilde{y} / dx^\nu \|_{1/K_1, c-1/K_1} \le C(\| \tilde{y} \|_{0,c} + 1). \tag{4.3}$$

Remarks: It is important to note that the dimension, m, of the large block need not be a constant on the entire interval $0 \le x \le c$. As the assumptions of the theorem state, it is only necessary that we be able to block the system into the form (4.1) in the neighborhood of every point in the interior of the interval. Thus the theorem applies to problems with "turning points", since by this we mean a problem in which one of the eigenvalues of the large block A_{11} locally changes size by an order of magnitude.

Observe also that the estimate (4.3) is invariant with respect to smooth transformations, i.e. we can replace y locally by $\tilde{y} = S(x)y$ where

$$\begin{aligned} &\| S \|_{x_0-\beta, x_0+\beta} + \| S^{-1} \|_{x_0-\beta, x_0+\beta} \le K_0, \\ &\| d^\nu S / dx^\nu \|_{x_0-\beta, x_0+\beta} \le K_1, \quad \nu = 1,2,\ldots p, \end{aligned} \tag{4.4}$$

and an estimate of the type (4.3) is still valid. Therefore we can relax the conditions of theorem 4.2. Instead of assuming that A has the form (4.1) we need only assume that in the neighborhood of every point there is a transformation S, satisfying (4.4), such that SAS^{-1} is of the form (4.1) and satisfies the conditions of theorem 4.2.

We shall now derive estimates for essentially diagonally dominant systems. We consider A_{11} to be the large part and A_{12}, A_{21}, A_{22} the $O(1)$ part of the system (4.1).

It is well known that the solutions of the system

$$\frac{dw}{dx} = A_{22}w + \tilde{F}, \quad 0 \le x \le c$$

satisfy an estimate

$$\| w \|_{0,c} \le K_3(|w(0)| + c \| \tilde{F} \|_{0,c}), \quad K_3 \le \exp[\| A_{22} \|_{0,c}\, c]. \tag{4.5}$$

Note that since we do not assume that the eigenvalues κ of A_{22} satisfy $\mathrm{Re}\kappa < 0$, this bound for K_3 is realistic and so the estimate (4.5) is useful only if c is sufficiently small.

We can now estimate the solutions of (4.1) in terms of G, H and $y^I(0)$, $y^{II}(0)$, $w(0)$. Lemma 3.1 and (4.5) give us

$$\| y \|_{0,c} \le 2\delta^{-1} \| (\Lambda + \Lambda^*)^{-1} A_{12} \|_{0,c} \;\; \| w \|_{0,c} + D_1 ,$$

$$\| w \|_{0,c} \le K_3 c \, \| A_{21} \|_{0,c} \;\; \| y \|_{0,c} + D_2 ,$$

where $D_1 = 2\delta^{-1}(\| (\Lambda + \Lambda^*)^{-1} G \|_{0,c} + |y^I(0)| + |y^{II}(c)|)$, and $D_2 = K_1(|w(0)| + c \| H \|_{0,c})$. Therefore we have

Lemma 4.1 Assume that

$$2\delta^{-1} c K_3 \| (\Lambda + \Lambda^*)^{-1} A_{12} \|_{0,c} \;\; \| A_{21} \|_{0,c} \le 1 - \delta^*, \quad \delta^* > 0,$$

then

$$\begin{aligned} \delta^* \| y \|_{0,c} &\le D_1 + 2\delta^{-1} \| (\Lambda + \Lambda^*)^{-1} A_{12} \|_{0,c} D_2, \\ \delta^* \| w \|_{0,c} &\le D_2 + K_3 c \, \| A_{21} \|_{0,c} D_1. \end{aligned} \tag{4.6}$$

If c is not small then we can derive global estimates in the following way. We divide the interval $0 \le x \le c$ into subintervals $c_i \le x \le c_{i+1}$, $i = 0,1,2,\ldots,q-1$, $c_0 = 0$, $c_q = c$, $c_{i+1}-c_i$ sufficiently small. On every subinterval we write $y = y_P + y_H$, $w = w_P + w_H$ where y_P, w_P denote the solution of the differential equation with boundary conditions

$$y_P^I(c_i) = y_P^{II}(c_{i+1}) = w_P(c_i) = 0. \tag{4.7}$$

and y_H, w_H is the solution of the homogeneous differential equation with

$$y_H^I(c_i) = y^I(c_i), \quad y_H^{II}(c_{i+1}) = y^{II}(c_{i+1}), \quad w_H(c_i) = w(c_i). \tag{4.8}$$

By lemma 4.1 $y_P^{II}(c_i)$, $y_P^I(c_{i+1})$, and $w_P(c_{i+1})$ are bounded and are uniquely determined by (4.7). Also, remembering that the differential equations are linear, we can write

$$y_H^{II}(c_i) = P^{II}, \quad y_H^I(c_{i+1}) = P^I, \quad w_H(c_{i+1}) = \tilde{P}.$$

Here P^{II}, P^I, and $\tilde{P}$ are linear relations with bounded coefficients in $y^I(c_i)$, $y^{II}(c_{i+1})$ and $w(c_i)$. Thus in every subinterval we obtain n linear relations

$$y^{II}(c_i) = P^{II} + y_P^{II}(c_i), \quad y^I(c_{i+1}) = P^I + y_P^I(c_{i+1}), \quad w(c_{i+1}) = \tilde{P} + w_P(c_{i+1})$$

for the variables $y(c_i)$, $y(c_{i+1})$, $w(c_i)$, and $w(c_{i+1})$. There are $n(q+1)$ unknowns $y(c_j)$, $w(c_j)$ and q subintervals. The missing relations are obtained from the boundary conditions for the original problem. Thus the $y(c_j)$, $w(c_j)$ can be obtained as the solution of a linear system of equations. Whether we obtain reasonable bounds for the original problem (1.1), (1.2) depends on the condition number of that linear system.

5. The choice of difference methods

In this section we shall discuss our choice of difference methods. Perhaps the simplest stiff problem is given by

$$\varepsilon \frac{dy}{dx} = -y + f(x), \quad y(0) = y_0 \tag{5.1}$$

where $0 < \varepsilon \ll 1$ is a very small positive constant and $f(x)$ is a smooth function with derivatives which are $O(1)$. The solution $y = y_S + y_B$ consists of a smooth part

$$y_S(x) = f(x) + O(\varepsilon), \tag{5.2}$$

which can be obtained by an asymptotic expansion, and a boundary layer part

$$y_B = e^{-x/\varepsilon}(y_0 - y_S(0)), \tag{5.3}$$

which is a solution of the homogeneous equation

$$\varepsilon \frac{dy}{dx} = -y. \tag{5.4}$$

Thus the solution is smooth except in a boundary layer near $x = 0$ where it changes rapidly.

Now consider a uniform grid $x_\nu = \nu h$, $\nu = 0,1,2,\ldots$; $0 < h \ll 1$. There are two standard types of difference approximations. One type is centered schemes of which the trapezoidal rule is an example:

$$\varepsilon \frac{u_{\nu+1} - u_\nu}{h} = -\frac{u_{\nu+1} + u_\nu}{2} + \frac{f_{\nu+1} + f_\nu}{2}, \quad u_0 = y_0. \tag{5.5}$$

The other type is one-sided schemes, such as the implicit Euler method:

$$\varepsilon \frac{v_{\nu+1} - v_\nu}{h} = -v_{\nu+1} + f_{\nu+1}, \quad v_0 = y_0. \tag{5.6}$$

The solutions $u_\nu = u_{S\nu} + u_{B\nu}$, $v_\nu = v_{S\nu} + v_{B\nu}$ of (5.5) and (5.6) respectively, consist again of a smooth part

$$u_{S_\nu} = f_\nu + O(\varepsilon), \quad v_\nu = v_{S\nu} + 0(\varepsilon)$$

and a boundary layer part

$$u_{B\nu} = \kappa^\nu (y_0 - u_{S_0}), \quad \kappa = \frac{1 - \frac{1}{2}(h/\varepsilon)}{1 + \frac{1}{2}(h/\varepsilon)}; \tag{5.7}$$

$$v_{B\nu} = \tau^\nu (y_0 - v_{S_0}), \quad \tau = \frac{1}{1 + h/\varepsilon}$$

where κ^ν, τ^ν are solutions of the corresponding homogeneous difference equations. If $h \ll \varepsilon$ then $\kappa \sim \tau \sim e^{-h/\varepsilon}$ and therefore $\kappa^\nu \sim \tau^\nu \sim e^{-x_\nu/\varepsilon}$ i.e. the solutions of the difference schemes approximate the solution of the differential equation well. However, we are interested in the case that $\varepsilon \ll h$. In this case

$$\kappa \sim -1, \qquad |\tau| \sim 0. \tag{5.7a}$$

Thus u_ν is in general highly oscillatory everywhere and does not approximate $y(x)$ well at all. In contrast, $|v_\nu - y(x_\nu)|$ is small away from the boundary layer. The advantage of one-sided methods in this situation is clear.

Onesided schemes have a major drawback, however, when they are to be used for solving systems of equations. For the equation

$$\varepsilon \frac{dy}{dx} = y + f, \quad y(c) = y_0, \quad x \le c,$$

the appropriate onesided scheme is the *explicit* Euler method,

$$\varepsilon \frac{v_{\nu+1} - v_\nu}{h} = v_\nu + f_\nu, \quad i.e. \quad v_\nu = \tau v_{\nu+1} + (1 + \frac{\varepsilon}{h})^{-1} f_\nu, \quad v_N = y_0, \tag{5.8}$$

because we start the integration at $x = c$ and calculate the solution for

decreasing values of ν. This construction of onesided schemes can be generalized to systems of the form

$$\frac{dy}{dx} = \begin{bmatrix} A_{-1} & 0 & 0 \\ 0 & A_0 & 0 \\ 0 & 0 & A_{+1} \end{bmatrix} y + By + F, \quad y = \begin{bmatrix} y^I \\ y^{II} \\ y^{III} \end{bmatrix}, \tag{5.9}$$

where

$$|hA_0| \ll 1, \quad |hB| \ll 1, \quad |hA_{\pm 1}| \gg 1,$$

and the eigenvalues $\kappa(A_{-1})$, $\kappa(A_{+1})$ of A_{-1} , A_{+1} respectively satisfy the inequalities

$$-h\,\mathrm{Re}\,\kappa(A_{-1}) \gg 1, \quad h\,\mathrm{Re}\,\kappa(A_{+1}) \gg 1.$$

An approximation to (5.9) on a nonuniform mesh is

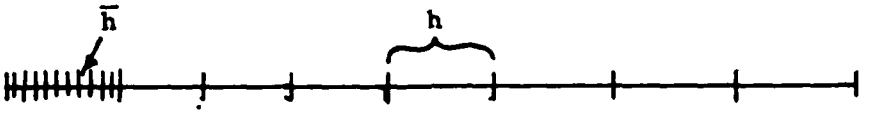
$$\begin{aligned} \frac{u^I_{\nu+1} - u^I_\nu}{h_\nu} &= A_{-1}(x_{\nu+1})u^I_{\nu+1} + G^I_{\nu+1}, \\ \frac{u^{II}_{\nu+1} - u^{II}_\nu}{h_\nu} &= \tfrac{1}{2}(A_0(x_\nu)u^{II}_\nu + A_0(x_{\nu+1})u^{II}_{\nu+1}) + \tfrac{1}{2}(G^{II}_\nu + G^{II}_{\nu+1}), \\ \frac{u^{III}_{\nu+1} - u^{III}_\nu}{h_\nu} &= A_{+1}(x_\nu)u^{III}_\nu + G^{III}_\nu, \end{aligned} \tag{5.10}$$

where

$$G_\nu = B(x_\nu)u_\nu + F_\nu,$$

i.e. we use implicit or explicit Euler for the variables corresponding to "large" eigenvalues of negative or positive sign, respectively, and we use the trapezoidal rule for the variables corresponding to "small" eigenvalues. If the system (1.1) is not already of the form (5.9), however, we must transform it to that form before we can tell which combination of these methods to use. Since this

transformation can be expected to be somewhat difficult to implement numerically, the question arises as to whether centered schemes can still be successfully employed for stiff problems, since they do not require a priori knowledge of such a transformation before they can be written down. The answer is that in many cases, centered schemes can be used together with appropriately chosen ***nonuniform meshes***. Consider again the trapezoidal rule (5.5). Instead of using a uniform mesh we now use a nonuniform mesh made up of two uniform meshes with meshwidths $\bar{h} \ll \varepsilon$ and h respectively.



In the boundary layer $0 \le x \le \bar{x}$, $\bar{x} = 0(\varepsilon|\log\varepsilon|)$ we use $\bar{h}$ and return to h for $x > \bar{x}$. In this case the boundary layer part of the solution is given by

$$u_{B\nu} = \begin{cases} e^{-x_\nu/\varepsilon}(y_0 - u_{S0}) & x_\nu \le \bar{x} \\ \pm(-1)^\nu e^{-x_\nu/\varepsilon}(y_0 - u_{S0}) & x_\nu > \bar{x} \end{cases}$$

It is clear that by choosing $\bar{x}$ sufficiently large we can make $|u_{B\nu} - y_B(x_\nu)|$ small everywhere. For systems we proceed correspondingly. We use a fine grid in the neighborhood of $x = 0,c$ and a coarse grid in the interior. On this coupled grid we approximate (1.1) by

$$\frac{u_{\nu+1} - u_\nu}{h_\nu} = \tfrac{1}{2}(A(x_{\nu+1})u_{\nu+1} + A(x_\nu)u_\nu) + \tfrac{1}{2}(F_\nu + F_{\nu+1}). \qquad (5.11)$$

Weiss and Ascher have considered the use of methods of this type in [2],[6]. The

collocation methods they discuss can be considered as generalizations of (5.11) and of the Box scheme, given by

$$\frac{u_{\nu+1} - u_\nu}{h_\nu} = \tfrac{1}{2}A\left(\frac{x_\nu + x_{\nu+1}}{2}\right)(u_\nu + u_{\nu+1}) + f\left(\frac{x_\nu + x_{\nu+1}}{2}\right). \tag{5.12}$$

A code based on those methods is discussed in [1]. For general systems (1.1) where the matrix $A(x)$ is Hermetian or the equation is already in "almost" block form,

$$y' = \begin{bmatrix} \frac{1}{\varepsilon}A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} y + F$$

where

$$h(|A_{11}^{-1}| + |A_{11}| + |A_{12}| + |A_{21}| + |A_{22}|) \ll 1,$$

centered schemes can be expected to behave properly provided that the boundary layers are properly resolved as discussed above. However, if the system has not been blocked beforehand, or if there are turning points present in the problem, then in general we cannot expect good results from centered schemes. The oscillatory nature of these schemes (see (5.7a)) in regions of the mesh where $h|A| \gg 1$ makes reasonable error estimates difficult to obtain for general problems. This is perhaps best illustrated with the following examples.

Consider the system

$$\frac{d}{dx}\begin{bmatrix} y \\ w \\ v \end{bmatrix} + \begin{bmatrix} -\frac{x}{\varepsilon} & 0 & \frac{1}{\varepsilon} \\ \frac{1}{\varepsilon} & \frac{x^2}{\varepsilon} & \frac{1}{\varepsilon} \\ -\frac{1}{2} & 0 & 0 \end{bmatrix}\begin{bmatrix} y \\ w \\ v \end{bmatrix} = 0, \quad -1 \le x \le 1 \tag{5.13}$$

with boundary conditions

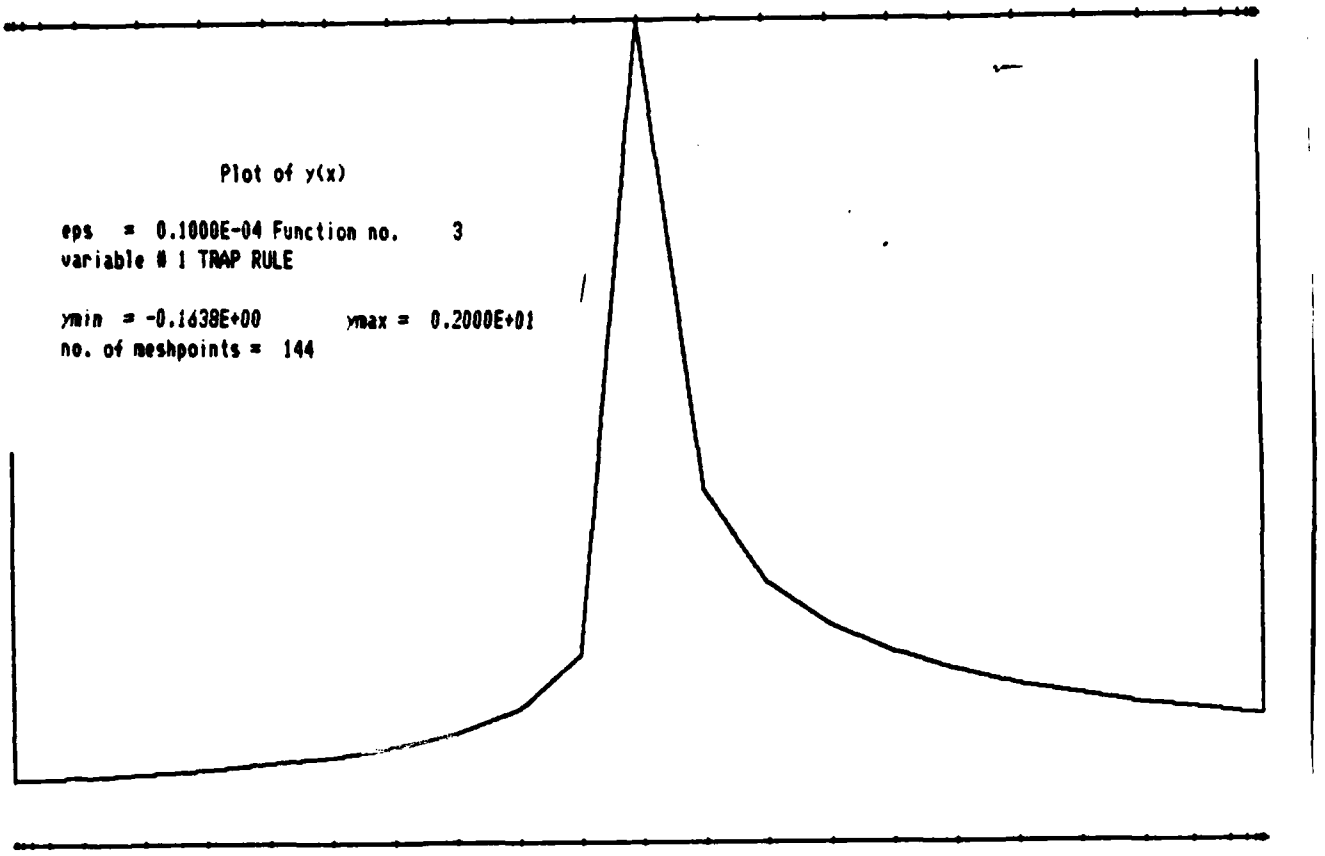


Figure 5.1

$$y(-1) = 1, \quad y(1) = 2, \quad w(-1) = 0 .$$

For $\varepsilon \ll 1$ the solution to this problem consists of boundary layers in the variables y and w near $x = \pm 1$ which connect to the constant states $y \sim 0$, $w \sim 0$ in the region away from the boundaries. It is easy to verify, in fact that

$$y(x) = e^{-(x+1)/\varepsilon} + 2e^{(x+1)/\varepsilon} + O(\varepsilon)$$

is the leading order representation for y. Since there is no non-smooth behavior in the interior of the interval, it might seem reasonable that this solution could

be computed numerically using a centered scheme and a nonuniform mesh to resolve the boundary layers. Figure 5.1, which shows the results of approximating (5.13) with $\varepsilon = 10^{-6}$ using the Trapezoidal rule, dramatically demonstrates that this conclusion is incorrect. In this figure, only the approximation to $y(x)$ is shown. The horizontal lines at the top and bottom of the plot indicate the locations of the mesh points used in the computation. For scale purposes the values of "ymin" and "ymax" in the legend accompanying each plot indicate the locations of these lines. The behavior of the computed solution near the center of the interval is clearly unacceptable.

One might suspect that the behavior near $x = 0$ is due to the potential turning point behavior of this problem in that region. The equations for y and v lead to the equation

$$\varepsilon y'' - xy' - \tfrac{1}{2}y = 0 \tag{5.14}$$

for $y(x)$. It is easy to verify that there is a potential for nonsmooth behavior in a neighborhood of size $O(\sqrt{\varepsilon})$ near $x = 0$. In figure 5.2, the mesh has been refined accordingly near $x = 0$. In this figure both y and w are shown. Note that while y now appears smooth, w still exhibits an unacceptable error near $x = 0$. Figure 5.3 shows a plot of the approximation to w computed using the Box scheme (5.12). The behavior near $x = 0$ is different but still unacceptable in this case.

It is possible to eliminate erroneous behavior of the type exhibited in figures 5.2 and 5.3 by using adaptive refinement of the computational mesh. However, the solution will only become smooth in the neighborhood of $x = 0$ once the meshwidths there satisfy $h = O(\varepsilon)$. We consider this to be an unaccept-

able restriction since in general turning point behavior occurs on larger scales than $O(\varepsilon)$ and hence would require *less* refinement.

Finally in figure 5.4 we show the results of a computation using the combination of onesided and centered schemes that we advocate and discuss in the following sections. One-sided schemes have the advantage over centered schemes that in regions where $h|A| \gg 1$, they mimic the damping behavior of the differential equations. This means that local errors are damped out quickly by one-sided schemes. In contrast, when using centered schemes, errors tend to be nonlocal due to the oscillatory behavior of the methods. As can be seen from the examples above, this can result in significant errors when solving systems of equations.

Because of the difficulties of this type that can arise when using centered schemes, we have chosen instead to use schemes of the type (5.10). Although it might seem that this involves more work that if we were to employ centered schemes, this extra work is in fact necessary if we wish to guarantee that our methods be robust. As we have shown, the "cheaper" centered methods can fail on problems of any generality.

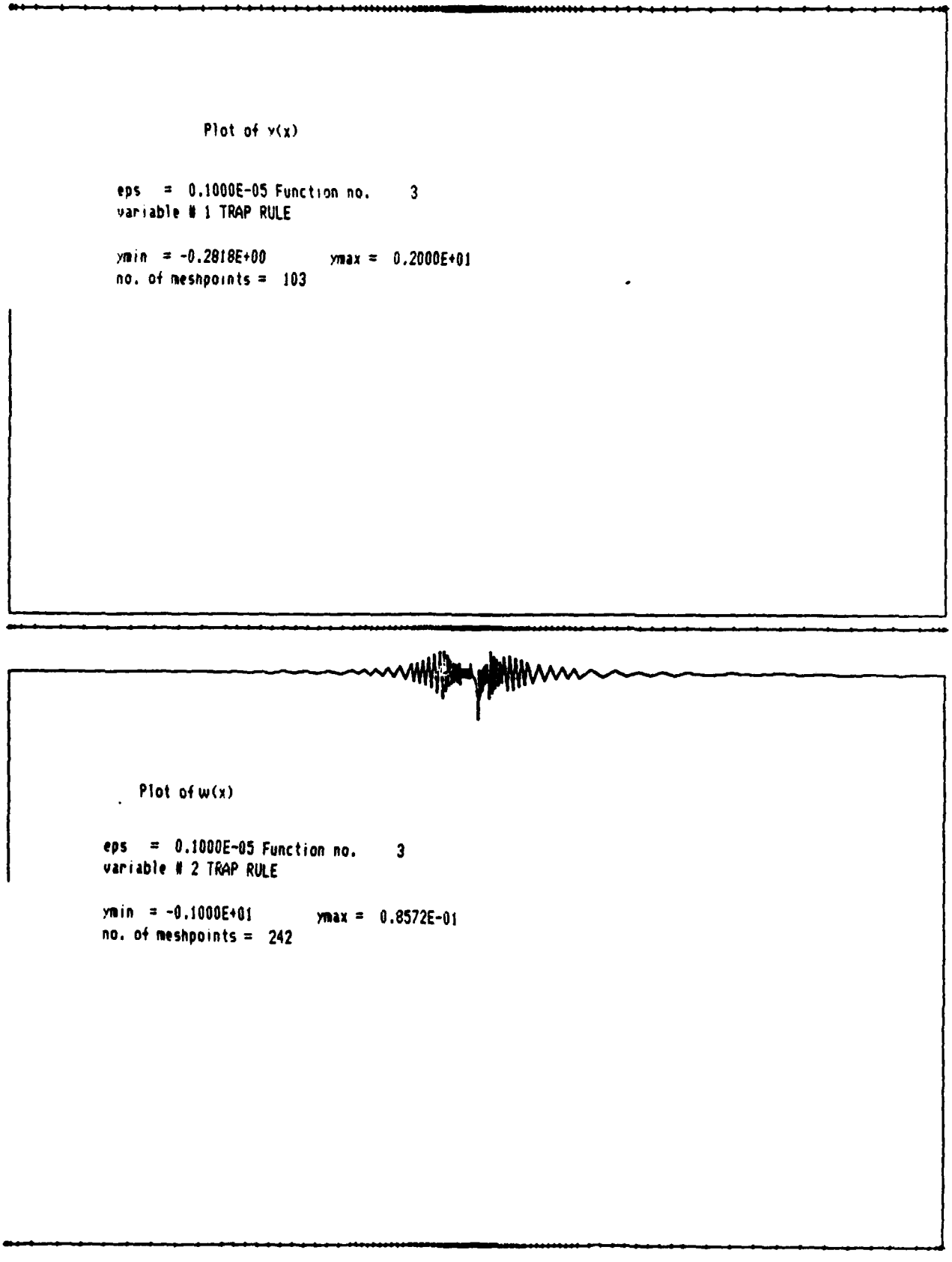
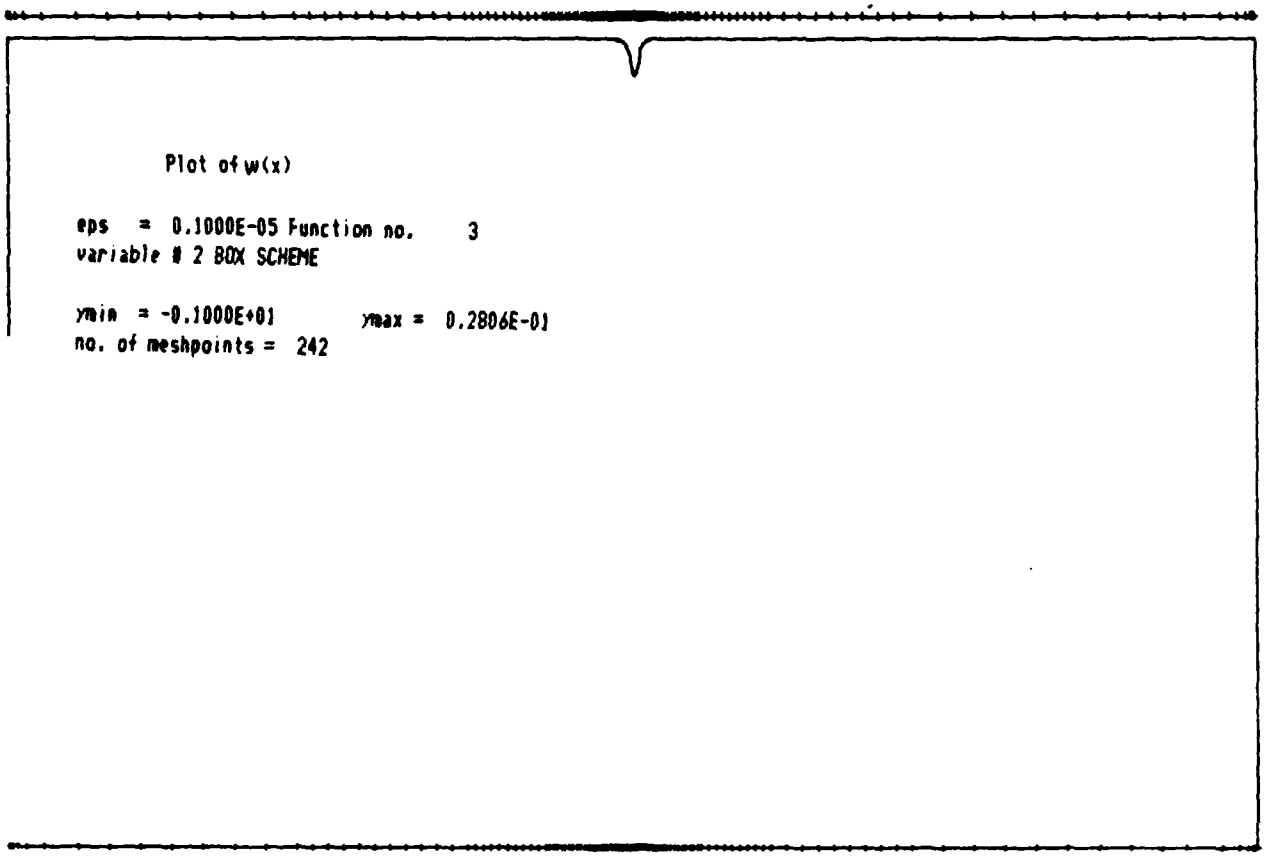

Plot of v(x)
eps = 0.1000E-05 Function no. 3
variable # 1 TRAP RULE
ymin = -0.2818E+00 ymax = 0.2000E+01
no. of meshpoints = 103
Plot of w(x)
eps = 0.1000E-05 Function no. 3
variable # 2 TRAP RULE
ymin = -0.1000E+01 ymax = 0.8572E-01
no. of meshpoints = 242

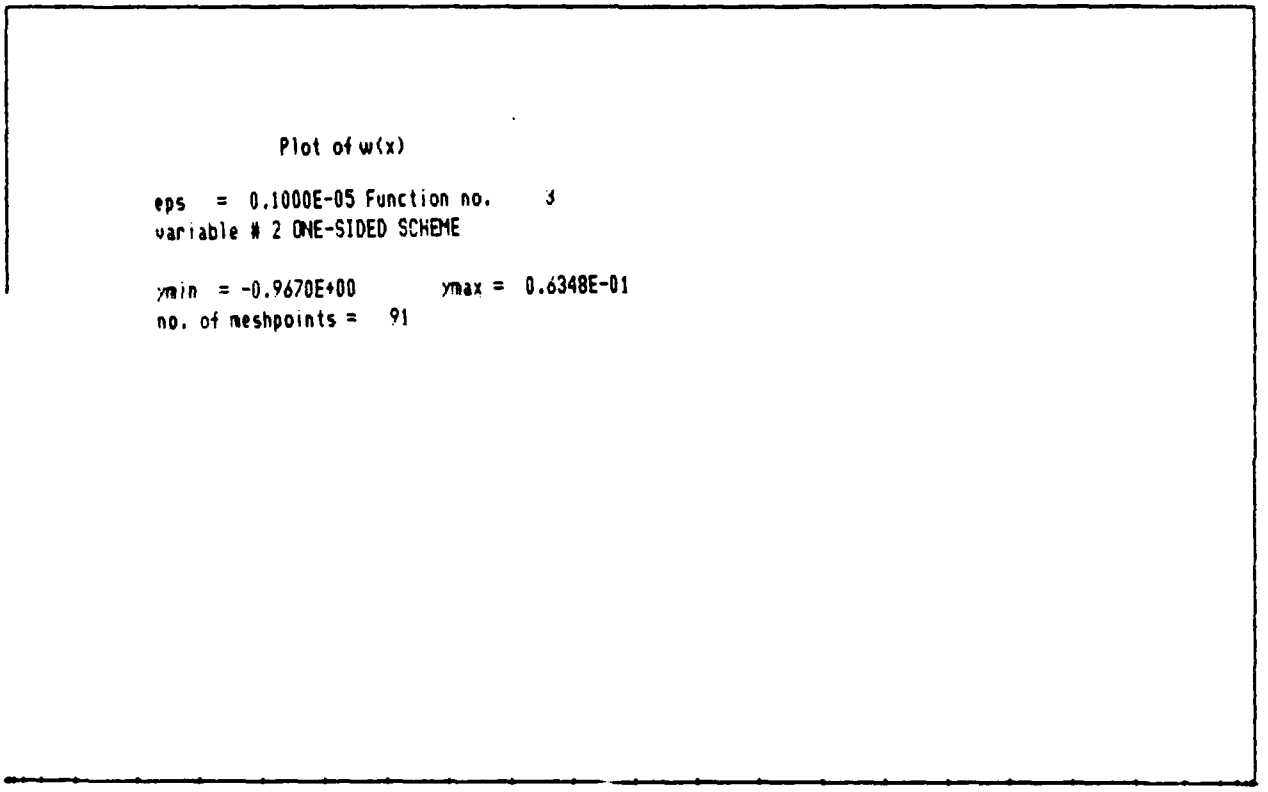
Figure 5.2

Figure 5.3



$x = -1$ $x = 1$

Figure 5.4


Plot of w(x)
eps = 0.1000E-05 Function no. 3
variable # 2 ONE-SIDED SCHEME
ymin = -0.9670E+00 ymax = 0.6348E-01
no. of meshpoints = 91

6. Difference approximations for scalar equations

In this section we start the discussion of our difference approximations. We divide the x-axis into subintervals of variable length h_j with gridpoints $x_0 = 0$, $x_\nu = \sum_{j=0}^{\nu-1} h_j$, $\nu = 1,2,\ldots,N$, $x_N = c$, and denote by $u_\nu = u(x_\nu)$ functions defined on the grid. We approximate (2.1) by methods of the form

$$\frac{u_{\nu+1}-u_\nu}{h_\nu} = \alpha_\nu a_\nu u_\nu + (1-\alpha_\nu) a_{\nu+1} u_{\nu+1} + \alpha_\nu f_\nu + (1-\alpha_\nu) f_{\nu+1}, \qquad (6.1)$$

$$u_0 = y_0.$$

We shall concentrate on two different methods: the Implicit Euler method and the Trapezoidal rule ($\alpha_\nu \equiv 0$ and $\alpha_\nu \equiv \frac{1}{2}$ respectively).

We assume that the conditions (2.2)-(2.5) are satisfied. Then it follows from the results of section 2 that the solution of the differential equation is smooth and we can obtain error estimates by standard truncation error analysis. For this we need stability estimates which we shall derive now.

Lemma 6.1: Let $\gamma_\nu \geq 0$, $\beta_\nu \geq 0$ be positive constants and consider a grid function v_ν satisfying

$$|v_{\nu+1}| \leq \frac{1}{1+\gamma_\nu}|v_\nu| + \frac{\gamma_\nu \beta_\nu}{1+\gamma_\nu}, \qquad \nu = 0,1,2,\ldots,$$

or for $0 \leq \gamma_\nu \leq 1$

$$|v_{\nu+1}| \leq \frac{1-\gamma_\nu}{1+\gamma_\nu}|v_\nu| + \frac{2\gamma_\nu \beta_\nu}{1+\gamma_\nu}, \qquad \nu = 0,1,2,\ldots.$$

Then

$$|v_\nu| \leq \max_{0 \leq j \leq \nu-1} \beta_j + \left(\prod_{j=0}^{\nu-1} \tau_j \right) |v_0|$$

with $\tau_j = 1/(1+\gamma_j)$, $\tau_j = (1-\gamma_j)/(1+\gamma_j)$ for the first and second case respectively.

Proof: is by induction on ν, it being trivially true for $\nu = 0$. For arbitrary ν, both inequalities state that $|v_{\nu+1}| \le \tau_\nu |v_\nu| + (1-\tau_\nu)\beta_\nu$, with $0 \le \tau_\nu \le 1$. Hence, using the induction hypothesis,

$$|v_{\nu+1}| \le \tau_\nu(\max_{j<\nu}\beta_\nu + |v_0|\prod_{j<\nu}\tau_j) + (1-\tau_\nu)\beta_\nu$$

$$= \tau_\nu \max_{j<\nu}\beta_j + (1-\tau_\nu)\beta_\nu + |v_0|\prod_{j\le\nu}\tau_j \le \max_{j\le\nu}\beta_j + |v_0|\prod_{j\le\nu}\tau_j .$$

Let us first consider the case that

$$\mathrm{Re}\,a \le -1 \tag{6.2}$$

The implicit Euler method ($d_\nu \equiv 0$) can be written in the form

$$u_{\nu+1} = \frac{u_\nu}{1-h_\nu a_{\nu+1}} + \frac{h_\nu f_{\nu+1}}{1-h_\nu a_{\nu+1}} = \frac{u_\nu}{1-h_\nu a_{\nu+1}} + \frac{h_\nu \mathrm{Re}a_{\nu+1}}{1-h_\nu a_{\nu+1}}\frac{f_{\nu+1}}{\mathrm{Re}a_{\nu+1}} . \tag{6.3}$$

Using lemma 6.1 with $\gamma_\nu = |h_\nu \mathrm{Re}a_{\nu+1}|$ we obtain therefore

$$|u_\nu| \le \max_{1\le j\le\nu}\left|\frac{f_j}{\mathrm{Re}a_j}\right| + \prod_{j=1}^{\nu}\frac{1}{1+\gamma_j}|u_0| . \tag{6.4}$$

We can now use (6.4) to obtain an error estimate. Assume that $p \ge 2$. Then $y(x)$ has two bounded derivatives and

$$(6.5)\quad \frac{y_{\nu+1}-y_\nu}{h_\nu} = a_{\nu+1}y_{\nu+1} + f_{\nu+1} + \tau_{\nu+1}$$

where

$$|\tau_{\nu+1}| \le \tfrac{1}{2}h_\nu \max_{x_\nu \le \xi \le x_{\nu+1}} |y''(\xi)|.$$

Thus the error $e_\nu = y_\nu - u_\nu$ satisfies

$$|e_\nu| \le \max_{0 \le j \le \nu} \left| \frac{\tau_j}{\mathrm{Re} a_j} \right|. \tag{6.6}$$

This error estimate is satisfactory if $|h_j \mathrm{Re} a_j| \ge 1$ because in that case it follows that $|e_\nu| = O(h^2)$. However if $|h_j \mathrm{Re} a_j| \ll 1$, then the method is not accurate enough for our purposes.

We consider now the trapezoidal rule ($d_\nu = \tfrac{1}{2}$):

$$u_{\nu+1} = \frac{1 + \tfrac{1}{2}h_\nu a_\nu}{1 - \tfrac{1}{2}h_\nu a_{\nu+1}} u_\nu + \frac{h_\nu g_{\nu+1}}{1 - \tfrac{1}{2}h_\nu a_{\nu+1}}, \quad g_{\nu+1} = \tfrac{1}{2}(f_\nu + f_{\nu+1}). \tag{6.7}$$

To estimate $|(1 + \tfrac{1}{2}h_\nu a_\nu)/(1 - \tfrac{1}{2}h_\nu a_{\nu+1})|$ we have to distinguish among a number of cases.

1) There is a constant $\sigma \le 2$ such that $|h_\nu \mathrm{Re} a_\nu| \le \sigma$, $|h_\nu \mathrm{Re} a_{\nu+1}| \le \sigma$ for all ν. In this case

$$\frac{1 + \tfrac{1}{2}h_\nu a_\nu}{1 - \tfrac{1}{2}h_\nu a_{\nu+1}} = \frac{1 + \tfrac{1}{2}h_\nu a_\nu}{1 - \tfrac{1}{2}h_\nu a_\nu} \frac{1}{1 - h_\nu \theta_\nu}$$

where

$$|\theta_\nu| = \tfrac{1}{2}\left| \frac{a_{\nu+1} - a_\nu}{1 - \tfrac{1}{2}h_\nu a_\nu} \right| \approx \left| \tfrac{1}{2} \frac{h_\nu a_\nu}{1 - \tfrac{1}{2}h_\nu a_\nu} \frac{da_\nu / dx}{a_\nu} \right| \le K_1$$

is uniformly bounded. Let $h_\nu a_\nu = b + ic$, b, c real, $|c| \le \rho |b|$. Then

$$\left| \frac{1 + \tfrac{1}{2}h_\nu a_\nu}{1 - \tfrac{1}{2}h_\nu a_\nu} \right| = \sqrt{\frac{1 + b + \tfrac{1}{4}(c^2 + b^2)}{1 - b + \tfrac{1}{4}(c^2 + b^2)}} = \sqrt{\frac{1 - 2\gamma_\nu}{1 + 2\gamma_\nu}} \le \frac{1 - \gamma_\nu}{1 + \gamma_\nu}, \tag{6.8}$$

where

$$\tfrac{1}{2}|h_\nu \mathrm{Re} a_\nu| \geq \gamma_\nu = \left|\frac{\tfrac{1}{2}b}{1 + \tfrac{1}{4}(c^2 + b^2)}\right| \geq \frac{2}{4 + \sigma^2(1+\rho^2)}|h_\nu \mathrm{Re} a_\nu|.$$

Thus

$$\left|\frac{1 + \tfrac{1}{2}h_\nu a_\nu}{1 - \tfrac{1}{2}h_\nu a_{\nu+1}}\right| \leq \frac{1-\gamma_\nu}{1+\gamma_\nu}\left|\frac{1}{1 - h_\nu \theta_\nu}\right| = \frac{1-\tilde{\gamma}_\nu}{1+\tilde{\gamma}_\nu} \tag{6.9}$$

where $\tilde{\gamma}_\nu \approx \gamma_\nu + \tfrac{1}{2}h_\nu \theta_\nu$. Also

$$\frac{h_\nu g_{\nu+1}}{1 - \tfrac{1}{2}h_\nu a_{\nu+1}} = \frac{h_\nu \mathrm{Re} a_{\nu+1}}{1 - \tfrac{1}{2}h_\nu a_{\nu+1}} \frac{g_{\nu+1}}{\mathrm{Re} a_{\nu+1}}.$$

Therefore we obtain from lemma 6.1 for sufficiently small $K_1 h_\nu$

$$|u_\nu| \leq \Gamma \max_{1 \leq j \leq \nu}\left|\frac{\tfrac{1}{2}(f_j + f_{j-1})}{\mathrm{Re} a_j}\right| + \prod_{j=0}^{\nu-1}\left[\frac{1-\tilde{\gamma}_j}{1+\tilde{\gamma}_j}\right]|u_0| \tag{6.10}$$

where

$$\Gamma = \max_\nu \frac{h_\nu \mathrm{Re} a_{\nu+1}}{2\tilde{\gamma}_\nu} \frac{1-\tilde{\gamma}_\nu}{1 - \tfrac{1}{2}h_\nu a_{\nu+1}} \leq (1 + \tfrac{1}{4}\sigma^2(1+\rho^2)) + O(h).$$

The last estimate gives us again an error estimate. Assume that $p \geq 3$. Then the solution of the differential equation has three bounded derivatives and by Taylor expansion we obtain

$$\frac{y_{\nu+1} - y_\nu}{h_\nu} = \tfrac{1}{2}(a_{\nu+1}y_{\nu+1} + a_\nu y_\nu) + \tfrac{1}{2}(f_{\nu+1} + f_\nu) + \tfrac{1}{2}(\tau_{\nu+1} + \tau_\nu),$$

where

$$\tfrac{1}{2}|(\tau_{\nu+1} + \tau_\nu)| \leq \frac{1}{12}h_\nu^2 \max_{x_\nu \leq \xi \leq x_{\nu+1}} |y'''(\xi)|.$$

Thus

$$|e_\nu| \leq \Gamma \max_{1 \leq j \leq \nu} \left| \frac{\frac{1}{2}(\tau_j + \tau_{j1})}{\mathrm{Re} a_j} \right|. \tag{6.11}$$

2) There is a constant $\sigma_1 > 2$ such that for all ν, $2 \leq |h_\nu \mathrm{Re} a_\nu| \leq \sigma_1$, $2 \leq |h_\nu \mathrm{Re} a_{\nu+1}| \leq \sigma_1$. This case can be reduced to the previous one by writing (6.7) in the form

$$u_{\nu+1} = \frac{-1}{1 - h_\nu \theta_\nu} \frac{1 + b_\nu}{1 - b_\nu} u_\nu - \frac{g_{\nu+1}}{a_{\nu+1}} \frac{1}{1 - c_\nu}$$

where

$$b_\nu = (\tfrac{1}{2} h_\nu a_\nu)^{-1}, \quad c_\nu = (\tfrac{1}{2} h_\nu a_{\nu+1})^{-1}.$$

Thus the same results as earlier hold, i.e. as long as $|h_\nu \mathrm{Re} a_\nu|$ stays bounded we obtain satisfactory error estimates.

3) $|h_\nu \mathrm{Re} a_\nu|$ can become arbitrarily large. In this case

$$(1 + \tfrac{1}{2} h_\nu a_\nu) / (1 - \tfrac{1}{2} h_\nu a_{\nu+1}) \to -a_\nu / a_{\nu+1}$$

and the exponential damping is lost. The solution of the homogeneous equation

$$v_{\nu+1} = -\frac{a_\nu}{a_{\nu+1}} v_\nu \tag{6.11}$$

can grow. We have

$$v_N = (-1)^{N-M} \frac{a_N}{a_M} v_M.$$

Thus if a_M is very large compared with a_N, then v_N is very large compared with v_M. This shows that even for a scalar equation the trapezoidal rule need not be stable.

The above error estimates suggest that we should use a combination of the trapezoidal rule and the implicit Euler method. The simplest way to do that would be to choose the coefficients d_ν in (6.1) in the following way:

$$d_\nu = \begin{cases} 0 & \text{if } |h_\nu a_\nu| > 1 \\ \frac{1}{2} & \text{if } |h_\nu a_\nu| \le 1. \end{cases} \tag{6.12}$$

However we would like to prevent the situation that we switch too often from one method to the other as a function of ν. Therefore the h_ν are only allowed to vary slowly and we replace (6.12) by

1) d_0 is chosen by (6.12)

2) For $\nu \ge 1$ we use

$$\begin{aligned} &\text{If } d_{\nu-1} = 0 \quad \text{choose} \quad d_\nu = \begin{cases} 0 & \text{if } |h_\nu a_\nu| > \frac{1}{2} \\ \frac{1}{2} & \text{otherwise} \end{cases} \\ &\text{If } d_{\nu-1} = \frac{1}{2} \quad \text{choose} \quad d_\nu = \begin{cases} \frac{1}{2} & \text{if } |h_\nu a_\nu \le 2 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{6.13}$$

Assume now that we have calculated the solution of (6.1) on some mesh $\{h_\nu\}$. Then we can divide the interval $0 \le x \le c$ into subintervals $c_i \le x \le c_{i+1}$, where c_i are meshpoints, $i = 0,1,\cdots,q-1$, $c_0 = 0$, $c_q = c$, such that on every subinterval we have used either the trapezoidal rule or the implicit Euler method. On every subinterval we can write down an error estimate. These local error estimates can be used to obtain global error estimates. Consider an interval $c_i \le x \le c_{i+1}$ and let $y(c_i)$ and $u(c_i)$ denote the solutions at $x = c_i$ of the differential equation and difference approximation respectively. Let u_P be the solution of the difference approximation with initial data $u_P(c_i) = y(c_i)$ and let u_H be the solution of the homogeneous difference equation with $u_H(c_i) = u(c_i) - u_P(c_i)$. Thus $u = u_P + u_H$. Our previous results tell us that

$$u_P(c_{i+1}) = y(c_{i+1}) + \varepsilon_i, \quad \varepsilon_i = O(h^2).$$

Also

$$u_H(c_{i+1}) = \lambda_i u_H(c_i) = \lambda_i(u(c_i) - y(c_i))$$

where $|\lambda_i| < 1$. Thus the error $e = y - u$ satisfies the relation

$$e(c_{i+1}) = y(c_{i+1}) - u(c_{i+1}) = y(c_{i+1}) - u_H(c_{i+1}) - u_P(c_{i-1}) \qquad (6.14)$$

$$= \lambda_i e(c_i) + \varepsilon_i.$$

Observing that $e(0) = 0$ we obtain a linear system of equations

$$A\underline{e} = \underline{\varepsilon}$$

where

$$A = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -\lambda_1 & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & -\lambda_{q-1} & 1 \end{bmatrix}, \quad \underline{e} = \begin{bmatrix} e(c_1) \\ e(c_2) \\ \vdots \\ e(c_{q-1}) \end{bmatrix}, \quad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{q-1} \end{bmatrix}.$$

The vector $\underline{e}$ represents the local error. The global error is obtained by inverting A. In this particular case all $|\lambda_i| < 1$ and the condition number is bounded by q.

To stress the interplay between local and global error we remove the restriction that $\mathrm{Re}\alpha \le -1$. We allow $\mathrm{Re}\alpha$ to become arbitrarily large and positive. We assume, however, that $y(x) = O(1)$ and slowly varying. Corresponding to (6.12) we choose the d_ν as follows:

$$d_\nu = \begin{cases} 0 & \text{if } |h_\nu a_\nu| > 1 \quad \mathrm{Re}\alpha_\nu < 0, \\ \tfrac{1}{2} & \text{if } |h_\nu a_\nu| \le 1, \\ -1 & \text{if } |h_\nu a_\nu| > 1, \quad \mathrm{Re}\alpha_\nu > 0. \end{cases} \qquad (6.15)$$

(In the same way as above, we would in practice modify (6.17) such that the d_ν do not change too often).

As before we divide the interval $0 \le x \le c$ into subintervals $c_i \le x \le c_{i+1}$ such that the parameters d_ν are constant on these subintervals. Also, if $d_\nu = \frac{1}{2}$ then we subdivide $c_i \le x \le c_{i+1}$ into subintervals $c_{ij} \le x \le c_{i,j+1}$ where one of the following conditions holds:

$$\mathrm{Re}\alpha \le -1, \quad |\mathrm{Re}\alpha| < 1, \quad \mathrm{Re}\alpha \ge 1.$$

Without restriction we can assume that the orginal c_i are chosen such that no subdivision is necessary. As earlier we obtain a relation of the type (6.14) for every interval $c_i \le x \le c_{i+1}$ with $\mathrm{Re}\alpha \le -1$. The same is true for intervals with $|\mathrm{Re}\alpha| \le 1$. This follows from well-known error estimates for nonstiff differential equations. If $\mathrm{Re}\alpha > 1$ then we use $y(c_{i+1})$ as initial data and integrate the differential equation in the direction of decreasing x. Correspondingly we solve the difference approximation in the direction of decreasing ν. This does not change the behavior of the trapezoidal rule but the explicit Euler method $d_\nu = 1$

$$(u_{\nu+1} - u_\nu)/h_\nu = a_\nu u_\nu + f_\nu$$

can be written as

$$u_\nu = \frac{u_{\nu+1}}{1 + h_\nu a_\nu} + \frac{f_\nu}{1 + h_\nu a_\nu}$$

and is the same as the implicit Euler method for decreasing values of ν. Thus we can apply our previous results and obtain

$$e(c_i) = \lambda_{i+1} e(c_{i+1}) + \varepsilon_i, \quad \varepsilon_i = O(h^2).$$

This shows that the local behavior of the difference approximation is

satisfactory also in the general case. However, it is well known that the initial value problem for (2.1) is not well posed if $\mathrm{Re}\,\alpha$ becomes arbitrarily large and positive. In this case the linear system of equations for the global error is not well conditioned. Observe that the λ_i can be computed and therefore also the condition number of A is available.

7. Difference approximations for diagonally dominant systems

In this section we consider systems of the type (3.2) and assume that all the conditions of section 3 are satisfied. We write the differential equations in the form

$$\frac{dy}{dx} = \Lambda y + G(y) , \tag{7.1}$$

where

$$G(y) = (A - \Lambda)y + F .$$

We approximate (7.1) by

$$\frac{u_{\nu+1} - u_\nu}{h_\nu} = D_\nu A(x_\nu)u_\nu + (I - D_\nu)A(x_{\nu+1})u_{\nu+1} + D_\nu F_\nu + (I - D_\nu)F_{\nu+1} \tag{7.2}$$

$$= D_\nu \Lambda(x_\nu)u_\nu + (I - D_\nu)\Lambda(x_{\nu+1})u_{\nu+1} + D_\nu G(u_\nu) + (I - D_\nu)G(u_{\nu+1})$$

Here D_ν is chosen by (6.17), i.e.

$$D_\nu = \begin{bmatrix} d_\nu^{(1)} & 0 & \cdots & 0 \\ 0 & d_\nu^{(2)} & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & d_\nu^{(m)} \end{bmatrix}, \tag{7.3}$$

$$d_\nu^{(i)} = \begin{cases} 0 & \text{if } |h_\nu a_{ii}(x_\nu)| > 1, \ \mathrm{Re}\, a_{ii} < 0 \\ \tfrac{1}{2} & \text{if } |h_\nu a_{ii}(x_\nu)| \le 1, \\ 1 & \text{if } |h_\nu a_{ii}(x_\nu)| > 1, \ \mathrm{Re}\, a_{ii} > 0 \end{cases}$$

In actual computations we modify (7.3) in the same way as earlier. For simplicity we assume that the $d_\nu^{(i)} = d^{(i)}$ do not depend on ν. Associated with the full system (7.2) is the diagonal system

$$\frac{v_{\nu+1} - v_\nu}{h_\nu} = D_\nu \Lambda(x_\nu)v_\nu + (I - D_\nu)\Lambda(x_{\nu+1})v_{\nu+1} + H_\nu . \tag{7.4}$$

By (6.4) and (6.10) we can find a constant $\Gamma \sim 2 + \rho^2$ such that

$$|v_\nu| \le 2\Gamma \|(\Lambda + \Lambda^*)^{-1} H\|_h + |v_0^I| + |v_N^{II}| \tag{7.5}$$

where we have used the notation

$$\|g\|_h = \max_{0 \le \nu \le N} |g_\nu|.$$

Instead of assuming that A is diagonally dominated we make a slightly stronger assumption:

Assumption 7.1: There is a constant δ with $0 \le \delta < 1$ such that

$$\|(\Lambda + \Lambda^*)^{-1}(A - \Lambda)\|_h \le \tfrac{1}{2}(\delta / \Gamma).$$

Remark: If we were only to use

$$\alpha^{(i)} = \begin{cases} 0 & \text{if } \mathrm{Re} a_{ii} < 0 \\ 1 & \text{if } \mathrm{Re} a_{ii} > 0 \end{cases} \tag{7.6}$$

then (6.4) would tell us that $\Gamma = 1$ and so assumption 7.1 is equivalent toassuming that A is diagonally dominated.

In the same way as lemma 3.1 we can now prove

Lemma 7.1: If assumption 7.1 is valid then the solutions of (7.2) satisfy the estimate

$$|u_\nu| \le \frac{2\Gamma}{\delta}\left\{\|(\Lambda + \Lambda^*)^{-1}(DF_\nu + (I - D)F_{\nu+1})\|_h + |u_0^I| + |u_N^{II}|\right\}. \tag{7.7}$$

Remark: In the same way as for the continuous problem one could estimate how theinfluence of $|u_0^I|$ and $|u_N^{II}|$ decreases away from the boundaries.

Lemma 7.1 gives us immediately an error estimate:

Theorem 7.1: Assume that $y(x)$ is a smooth and bounded solution of the differential equation. Then

$$\| y - u \|_h \leq \text{const.}(h^2 + |u_0' - y_0'| + |u_N' - y_N'|).$$

8. Difference approximations for essentially diagonally dominant systems In this section we consider difference approximations for the system (4.1) with boundary conditions (1.2). They are of the form

$$\frac{u_{\nu+1} - u_\nu}{h_\nu} = D_\nu A_{\parallel}(x_\nu)u_\nu + (I - D_\nu)A_{\parallel}(x_{\nu+1})u_{\nu+1} + D_\nu G_\nu + (I - D_\nu)G_{\nu+1}, \quad (8.1)$$

$$\frac{v_{\nu+1} - v_\nu}{h_\nu} = \tfrac{1}{2}(A_{22}(x_\nu)v_\nu + A_{22}(x_{\nu+1})v_{\nu+1}) + \tfrac{1}{2}(E_\nu + E_{\nu+1}) \quad (8.2)$$

with boundary conditions

$$B_0\begin{bmatrix}u_0\\v_0\end{bmatrix} + B_1\begin{bmatrix}u_N\\v_N\end{bmatrix} = g.$$

Here $G = A_{12}v + F$, $E = A_{21}u + H$ and D_ν is defined by (7.3). We assume again that $D_\nu \equiv D$ does not depend on ν. By assumption $|A_{22}h| \ll 1$ and it is therefore well-known that the solutions of (8.2) satisfy the estimate

$$\|v\|_h \le \tilde{K}_2(|v_0| + c\|E\|_h), \quad \tilde{K}_2 \sim \exp(\|A_{22}\|_{0,c}c).$$

Lemma 7.1 tells us that if assumption 7.1 holds then the solutions of (8.1) satisfy the estimate (7.7) with F replaced by G. Using the assumptions 4.1 and 4.2 we obtain the analog of lemma 1.1:

Lemma 8.1: If c is sufficiently small then there is a constant K_3 such that

$$(\|v\|_h + \|u\|_h) \le K_3(\|F\|_h + \|H\|_h + |u_0^I| + |u_N^{II}| + |v_0|)$$

Here K_3 depends only on K_1, $\tilde{K}_1$ and ρ.

We assume again that the solution of the differential equations is slowly varying and bounded. Lemma 8.1 leads then immediately to the error estimate

$$\begin{aligned}&\|y - u\|_h + \|w - v\|_h\\ &\le \text{const.}(h^2 + |y_0^I - u_0^I| + |y_N^{II} - u_N^{II}| + |w_0 - v_0|).\end{aligned} \quad (8.3)$$

This estimate gives us an error estimate for our problem provided that the interval $0 \le x \le c$ is sufficiently small. Since again this cannot be expected to be the case, we can obtain global error estimates in the same way as for the scalar equation by dividing the interval up into subintervals. We divide $0 \le x \le c$ into subintervals $c_i \le x \le c_{i+1}$, $i = 0,1,2,\ldots,q-1$ with the following properties:

1) D is constant on every subinterval.

2) $c_{i+1} - c_i$ is sufficiently small such that the estimate of lemma 8.1 holds.

3) The solution of the differential equation is bounded and slowly varying.

Let $y(x)$ be the solution of the differential equation. We write again $u = u_P + u_H$, $v = v_P + v_H$ where $(u_P, v_P)^T$ is the solution of the difference approximation with

$$u_P^I(c_i) = y^I(c_i), \quad u_P^{II}(c_{i+1}) = y^{II}(c_{i+1}), \quad v_P(c_i) = w(c_i) \tag{8.4}$$

and $(u_H, v_H)^T$ is the solution of the homogeneous difference equation with

$$u_H^I(c_i) = e^I(c_i), \quad u_H^{II}(c_{i+1}) = e^{II}(c_{i+1}), \quad v_H(c_i) = \tilde{e}(c_i). \tag{8.5}$$

Here $e = u - y$, $\tilde{e} = v - w$ denotes the error. Using the estimate (8.3) we obtain from (8.4)

$$\begin{aligned} u_P^{II}(c_i) &= y^{II}(c_i) + \varepsilon_i^{II}, \\ u_P^I(c_{i+1}) &= y^I(c_{i+1}) + \varepsilon_i^I, \\ v_P^{II}(c_{i+1}) &= w(c_{i+1}) + \tilde{\varepsilon}_i. \end{aligned} \tag{8.6}$$

Recalling that the difference equations are linear we can write

$$u_H^{II}(c_i) = L^{II}, \quad u_H^I(c_{i+1}) = L^I, \quad v_H(c_{i+1}) = \tilde{L} \tag{8.7}$$

Here L^{II}, L^I, $\tilde{L}$ are linear expressions in $e^I(c_i)$, $e^{II}(c_{i+1})$ and $\tilde{e}(c_i)$ whose coefficients can be estimated by K_3. The equations (8.6) and (8.7) give us n linear equations

$$
\begin{aligned}
e^{II}(c_i) - L^{II} &= \varepsilon_i^{II}, \\
e^{I}(c_{i+1}) - L^{I} &= \varepsilon_i^{I}, \\
\tilde{e}(c_{i+1}) - \tilde{L} &= \tilde{e}_i
\end{aligned}
$$

for every subinterval $c_i \le x \le c_{i+1}$. There are q intervals and $n(q+1)$ variables $e(c_i)$, $\tilde{e}(c_i)$. Therefore, using the boundary conditions we obtain a linear system of $n(q+1)$ equations in $n(q+1)$ variables. Again, global error estimates depend on the condition number of this system.

Remark: In section 4 we reduced the solutions of the differential equations to the solution of a corresponding linear system of equations. The two linear systems of equations obtained in that section and in this one need not be close, because we have not resolved the potential boundary layers near $x = c_i, c_{i+1}$ with an appropriate mesh. Thus in general the fundamental solutions of the homogeneous differential and difference equations respectively are not necessarily close. However, if the diagonal elements of A_{11} satisfy $|\mathrm{Re}\alpha_{\mu} h_{\nu}| \gg 1$ then one can show using asymptotic expansions that the corresponding relations are, in fact, close.

9. Normal form for the differential equation

We will now discuss how the general system (1.1) can be transformed to essentially diagonally dominant form. In this section we give a theoretical presentation of this procedure. The practical implementation of the transformation differs somewhat from the discussion in this section; those differences are discussed in section 10.

The procedure can be outlined as follows: We assume that away from a finite number of turning point regions the system is well-behaved. Then the transformation to essentially diagonally dominant form is effected in each subinterval of $0 \le x \le c$ through similarity transformations which put the matrix $A(x)$ into an appropriate "blocked" form, and a stretching of the independent variable x such that relative to the basic meshsize h, smoothness requirements similar to assumptions 3.1-3.2 and (2.2) are enforced. The results of section 4 then guarantee that we will get the appropriate error estimates when the difference approximation is applied.

First we calculate the eigenvalues $\kappa(x)$ of $A(x)$ and divide them into sets $\mathbf{M}^{(j)}$ containing eigenvalues which are of the same order of magnitude. This is done in the following way: Let K, $\delta > 0$ with $0 \le Kh \ll 1$ be constants. Then $\kappa \in \mathbf{M}^{(0)}$ if either $|\kappa| \le K$ or there exists a $\tilde{\kappa} \in \mathbf{M}^{(0)}$ such that

$$\left| |\kappa| - |\tilde{\kappa}| \right| \le \delta \, (|\kappa| + |\tilde{\kappa}|). \tag{9.1}$$

By choosing δ sufficiently small we can guarantee that all $\kappa \in \mathbf{M}^{(0)}$ satisfy $|\kappa h| \ll 1$. If all eigenvalues belong to $\mathbf{M}^{(0)}$ then the construction is complete. Otherwise let $\kappa_1, \kappa_2, \cdots, \kappa_p$ denote the eigenvalues not contained in $\mathbf{M}^{(0)}$ and let $|\kappa_j| = \min_{1 \le \nu \le p} |\kappa_\nu|$. Then the set $\mathbf{M}^{(1)}$ is formed recursively by taking $\kappa_j \in \mathbf{M}^{(1)}$, $\kappa \in \mathbf{M}^{(1)}$ if $(\mathrm{Re}\kappa_j)(\mathrm{Re}\kappa) \ge 0$ and there is a $\tilde{\kappa} \in \mathbf{M}^{(1)}$ such that (9.1) holds. Further sets are constructed correspondingly. We allow the number of sets $\mathbf{M}^{(j)}$ to

depend on x, i.e. as a function of x, sets can split up and recombine. Therefore the block-structure can be a function of x as well. We assume, however, that we can divide the interval $0 \le x \le c$ into a finite number of subintervals $c_i \le x \le c_{i+1}$ such that on every subinterval the block structure is constant. We will refer to such subintervals as "*blocking subintervals*".

The next step is to determine a transformation $S(x)$ such that

$$\mathcal{A}(x) = S^{-1}(x)A(x)S(x) = \begin{bmatrix} A_r(x) & 0 & \cdots & 0 \\ 0 & A_{r-1}(x) & 0 & 0 \\ \cdots & \cdots & \cdots & . \\ 0 & \cdots & 0 & A_0(x) \end{bmatrix}$$

is in block-diagonal form. Here the eigenvalues of every $A_j(x)$ are exactly the eigenvalues contained in $\mathbf{M}^{(j)}$ (counted according to their multiplicity). We must make a couple of assumptions about the blocks $A_j(x)$. For the matrix $A_0(x)$ we require that

$$h\,|A_0(x)| \ll 1. \tag{9.2}$$

We know that the eigenvalues of $A_0(x)$ satisfy $|\kappa h| \ll 1$. Therefore (9.3) says that if we were to transform $A_0(x)$ to upper triangular form by a unitary transformation, the off-diagonal elements a_{ij} would also satisfy $|a_{ij}h| \ll 1$. The following shows that this is a reasonable assumption: Consider the differential equation

$$\frac{dy}{dx} = U^*(x)\Lambda U(x)y$$

where $\Lambda = \begin{bmatrix} 0 & \gamma/\varepsilon \\ 0 & 0 \end{bmatrix}$ is a constant matrix and

$$U(x) = \begin{bmatrix} \cos x & \sin x \\ -\sin x & \cos x \end{bmatrix}$$

is a unitary matrix, $\gamma \neq 0$ and $0 < \varepsilon \ll 1$. The matrix Λ is in upper triangular form but the off-diagonal element is not small. If we change variables to $\tilde{y} = Uy$, then the system becomes

$$\tilde{y}' = \begin{bmatrix} 0 & \frac{\gamma}{\varepsilon} + 1 \\ -1 & 0 \end{bmatrix} \tilde{y} = \tilde{A}\tilde{y} .$$

The eigenvalues of $\tilde{A}$ are given by $\kappa = \pm \sqrt{-(\gamma/\varepsilon + 1)}$ and thus the solution of the equation with variable coefficients has nothing to do with the solution of

$$d\tilde{w} / dx = \Lambda\tilde{w} .$$

For the other blocks we assume correspondingly that

$$|\bar{\kappa}_j^{-1} A_j| = O(1), \quad \text{where } \bar{\kappa}_j = \frac{1}{n_j} \sum_{\kappa_i \in \mathbb{M}^{(j)}} \kappa_i = \frac{1}{n_j} \sum a_{ii}^{(j)}, \tag{9.3}$$

with n_j the order and $\sum a_{ii}^{(j)}$ the trace of A_j. If these conditions are not satisfied then we choose the basic meshsize small enough so that $h|A_j| \ll 1$ for all A_j that violate the assumption.

We are not interested in highly oscillatory problems, because they have to be treated differently (see Scheid [5]). Thus we make the assumption that the eigenvalues κ of $A(x)$ satisfy

$$|\mathrm{Im}\kappa| \leq \rho\, |\mathrm{Re}\kappa| + C_1 \tag{9.4}$$

where $\rho \sim O(1)$ and $0 < C_1 h \ll 1$ are threshold constants. Observe, however, that (9.4) allows highly oscillatory solutions provided that the meshsize is sufficiently small.

We now describe the construction of $S(x)$ in detail. We start with the interval $0 \leq x \leq c_1$. At $x = 0$ we construct a unitary transformation $U(0)$ such that

$U^*(0)A(0)U(0)$ is an upper triangular matrix in which the eigenvalues appear in the correct order, i.e.

$$U^*(0)A(0)U(0) = \begin{pmatrix} A_r & A_{r,r-1} & \cdots & A_{r,0} \\ 0 & A_{r-1} & \cdots & A_{r-1,0} \\ \cdots & \cdots & & \cdots \\ 0 & \cdots & 0 & A_0 \end{pmatrix}.$$

This can be done with a slightly modified version of the usual QR method. We then determine

$$\tilde{S}(0) = \begin{pmatrix} I & S_{r,r-1} & \cdots & S_{r,0} \\ 0 & I & \cdots & S_{r-1,0} \\ \cdots & \cdots & & \cdots \\ 0 & \cdots & 0 & I \end{pmatrix}$$

such that

$$\tilde{A}(0) = S^{-1}(0)A(0)S(0) = \begin{pmatrix} A_r & 0 & \cdots & 0 \\ 0 & A_{r-1} & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & A_0 \end{pmatrix}, \quad S(x) = U(0)\tilde{S}(x)$$

has the desired block form. Now consider the transformed matrix

$$\tilde{A}(x) = \tilde{A}(0) + B(x), \quad B(0) = 0,$$

$$B(x) = S^{-1}(0)(A(x) - A(0))S(0) = \begin{pmatrix} B_{rr} & B_{r,r-1} & \cdots & B_{r0} \\ B_{r-1,r} & B_{r-1,r-1} & \cdots & B_{r-1,0} \\ \cdots & \cdots & \cdots & \cdots \\ B_{1,r} & \cdots & \cdots & B_{00} \end{pmatrix}.$$

By assumption the eigenvalues of each block are well separated from the eigenvalues of all other blocks. Therefore in the neighborhood of $x = 0$ we can con-

struct an $\tilde{S}(x)$ such that

$$S^{-1}(x)A(x)S(x) = \begin{pmatrix} A_r(x) & 0 & \cdots & 0 \\ 0 & A_{r-1}(x) & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & A_0(x) \end{pmatrix}, \quad S(x) = \tilde{S}(0)\tilde{S}(x) .$$

To discuss this transformation in detail we need a couple of lemmata.

Lemma 9.1: Let A_{11}, A_{22}, E be $p \times p$, $q \times q$ and $p \times q$ matrices respectively. Assume that the eigenvalues $\lambda_i, i = 1,2,\ldots,p$ of A_{11} are disjoint from the eigenvalues μ_j $j = 1,2,\ldots,q$ of A_{22}. Then the matrix equation

$$A_{11}X - XA_{22} = E$$

has a unique solution and there is a constant C which depends only on p, q, $|A|$, $|B|$ and $\min_{i,j} |\lambda_i - \mu_j|$ such that

$$|X| \leq C\,|E| . \tag{9.5}$$

Proof: Without restriction we can assume that A_{11}, A_{22} are upper triangular; then (9.5) is of the form

$$\begin{pmatrix} \lambda_1 & a_{12} & \cdots & \cdots & a_{1p} \\ 0 & \lambda_2 & a_{23} & \cdots & a_{2p} \\ \cdots & \cdots & \cdots & \cdots & . \\ 0 & \cdots & \cdots & \cdots & \lambda_p \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1q} \\ x_{21} & x_{22} & \cdots & x_{2q} \\ \cdots & \cdots & \cdots & . \\ x_{p1} & \cdots & \cdots & x_{pq} \end{pmatrix} \tag{9.6}$$

$$- \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1q} \\ x_{21} & x_{22} & \cdots & x_{2q} \\ \cdots & \cdots & \cdots & . \\ x_{p1} & \cdots & \cdots & x_{pq} \end{pmatrix} \begin{pmatrix} \mu_1 & \tilde{a}_{12} & \cdots & \tilde{a}_{1q} \\ 0 & \mu_2 & \cdots & \tilde{a}_{2q} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \mu_q \end{pmatrix} = E$$

For the first column of X we obtain

$$(\lambda_p - \mu_1)\, x_{p1} = e_{p1}$$
$$(\lambda_{p-1} - \mu_1)\, x_{p-1,1} = -a_{p-1,p} x_{p-1} + e_{p2}$$
$$\cdots \text{etc.} \cdots$$

Thus the first column of X can be computed by back-substitution and it satisfies an estimate of the type (9.5). The other columns are calculated in a corresponding manner. This proves the lemma.

Let us use the above matrices A_{11} and A_{22} to form

$$\begin{bmatrix} A_{11} & B_{12} \\ B_{21} & A_{22} \end{bmatrix}.$$

We want to construct a matrix R such that

$$\begin{bmatrix} I & -R \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{11} & B_{12} \\ B_{21} & A_{22} \end{bmatrix} \begin{bmatrix} I & R \\ 0 & I \end{bmatrix} =$$

$$= \begin{bmatrix} A_{11} - RB_{21} & A_{11}R - RA_{22} - RB_{21}R + B_{12} \\ B_{21} & B_{21}R + A_{22} \end{bmatrix} = \begin{bmatrix} C_{11} & 0 \\ C_{21} & C_{22} \end{bmatrix} = C$$

is a lower block triangular matrix, i.e.

$$A_{11}R \quad RA_{22} \quad RB_{21}R + B_{12} = 0. \tag{97}$$

Lemma 9.1 gives us

Lemma 9.2: The iteration

$$A_{11}R^{(n)} - R^{(n)}A_{22} - R^{(n-1)}B_{21}R^{(n-1)} + B_{12} = 0 \tag{98}$$

converges to a locally unique solution of (9.7) provided $|B_{21}|$ and $|B_{12}|$ are sufficiently small.

We shall now transform C to block diagonal form.

Lemma 9.3: *If* $|B_{21}|$ *and* $|B_{12}|$ *are sufficiently small then there is a transformation* Q *such that*

$$\begin{bmatrix} I & 0 \\ -Q & I \end{bmatrix}\begin{bmatrix} C_{11} & 0 \\ C_{21} & C_{22} \end{bmatrix}\begin{bmatrix} I & 0 \\ Q & I \end{bmatrix} = \begin{bmatrix} C_{11} & 0 \\ -QC_{11} + C_{22}Q + C_{21} & C_{22} \end{bmatrix} =$$

$$\begin{bmatrix} A_{11} - RB_{21} & 0 \\ 0 & B_{21}R + A_{22} \end{bmatrix}.$$

Proof: If $|B_{12}|$, $|B_{21}|$ are sufficiently small then the sets of eigenvalues of C_{11} and C_{22} respectively are still disjoint. Therefore we need only to solve the linear system

$$-QC_{11} + C_{22}Q + C_{21} = 0$$

This proves the lemma.

Remark: In practice, before making the transformation of lemmata 9.1 and 9.2, we diagonally scale the matrix so that the off-diagonal blocks are of the same order of magnitude: Choosing d such that $d|B_{21}| + 1 = d^{-1}|B_{12}| + 1$, this transformation is given by

$$\begin{bmatrix} A_{11} & d^{-1}B_{12} \\ dB_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} I & \\ & dI \end{bmatrix}\begin{bmatrix} A_{11} & B_{12} \\ B_{21} & A_{22} \end{bmatrix}\begin{bmatrix} I & \\ & d^{-1}I \end{bmatrix}.$$

This extra transformation does not change the end result but guarantees that R and Q are of the same order of magnitude. The effect of this is to help make $S(x)$ change as slowly as possible.

The results above can be used to construct a transformation $\tilde{S}(x)$ which transforms $\tilde{A}(x)$ to block diagonal form.

Theorem 9.1: If $|A_i^{-1}(0)B_{ij}(x)|$, $i = 1,2,\ldots,r$ and $|B_{0j}(x)|$, $j = 0,1,2,\ldots,r$ are sufficiently small then we can construct $S(x)$ locally in a unique way.

Proof: We write $\tilde{A}(x)$ as

$$\tilde{A}(x) = \frac{1}{\varepsilon}\begin{bmatrix} \varepsilon A_{11} & \varepsilon \tilde{B}_{12} \\ \varepsilon \tilde{B}_{21} & \varepsilon A_{22} \end{bmatrix}, \quad \varepsilon^{-1} = \left|\frac{1}{n_r}\sum_{\kappa_i \in \mathbb{K}(r)} \kappa_i\right| = |\bar{\kappa}_r|$$

where

$$A_{11} = A_r + B_{rr}, \quad \tilde{B}_{12} = (B_{r,r-1}, \cdots, B_{r0}), \quad \tilde{B}_{21} = (B_{r-1,r}^T, \cdots B_{0r}^T)^T.$$

By assumption $|\varepsilon A_{11}|$, $|(\varepsilon A_{11})^{-1}|$ and $|\varepsilon A_{22}|$ are $O(1)$. Also the eigenvalues of εA_{11} are well separated from those of εA_{22}. Thus the above lemmata give us that if $|\varepsilon \tilde{B}_{12}| \sim |A_{11}^{-1}\tilde{B}_{12}|$, $|\varepsilon \tilde{B}_{21}| \sim |A_{22}^{-1}\tilde{B}_{21}|$ are sufficiently small, i.e. if x is sufficiently small then there is a unique transformation of the type

$$S_1 = \begin{bmatrix} I & R \\ 0 & I \end{bmatrix}\begin{bmatrix} I & 0 \\ Q & I \end{bmatrix} = \begin{bmatrix} I + RQ & R \\ Q & I \end{bmatrix}$$

such that

$$S_1^{-1}\tilde{A}(x)S_1 = \begin{bmatrix} \tilde{A}_{11} & 0 \\ 0 & \tilde{A}_{22} \end{bmatrix} = \begin{bmatrix} A_{11} - RB_{21} & 0 \\ 0 & B_{21}R + A_{22} \end{bmatrix}$$

Now, $B_{21}R + A_{22}$ has the same properties as $\tilde{A}(x)$ did, and so the same process can be applied again to it. This proves the theorem.

We have constructed $S(x)$ in a neighborhood of $x = 0$, but it is clear that we can continue the construction as long as the block structure does not change, i.e. for $0 \le x \le c_1$. Let $S_-(c_1) = \lim_{x \to c_1 - 0} S(x)$. At $x = c_1$ we change $S_-(c_1)$ to $S_+(c_1)$ in the following way:

1) If two sets $\mathbf{M}^{(j)}$ merge: S does not change (although the blocking does)

2) If a set $\mathbf{M}^{(j)}$ splits into subsets then construct a transformation which transforms the corresponding block into block diagonal form. If necessary, a permutation matrix can be applied to rearrange the blocks according to the size of their eigenvalues. Alternatively, $S_+(c_1)$ can be computed in the same way as $S(0)$.

We now use $S_+(c_1)$ as the starting transformation for the interval $c_1 \le x \le c_2$ and repeat the above procedure to obtain $S(x)$ at intermediate points in that subinterval. In this way we determine $S(x)$ everywhere and use it to transform the system (1.1) to

$$d\tilde{y}/dx = \hat{A}(x)y + H(x)y + G(x) \tag{9.9}$$

with

$$\hat{A} = \begin{pmatrix} A_r(x) & 0 & \cdots & \cdots & 0 \\ 0 & A_{r-1} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & A_0(x) \end{pmatrix}, \qquad \begin{aligned} H &= -S^{-1}dS/dx, \\ G &= S^{-1}F, \\ \tilde{y} &= S^{-1}y \end{aligned}$$

on every blocking subinterval $c_i \le x \le c_{i+1}$.

In the same way as in [4] we now "stretch" the independent variable. We divide each blocking subinterval $c_i \le x \le c_{i+1}$ into $s \ge 1$ ***"stretching subintervals"*** $c_{ij} \le x \le c_{i,j+1}$ with $c_i = c_{i0} \le c_{i1} \le \cdots \le c_{is} = c_{i+1}$. Suppose we have determined $c_{i0}, c_{i1}, \cdots, c_{ij}$. Then $c_{i,j+1}$ is determined as follows: We introduce a new variable $\tilde{x}$ by $x - c_{ij} = \alpha_{ij}\tilde{x}$, $0 \le \tilde{x} \le 1$ and obtain

$$\frac{d\tilde{y}}{dx} = \alpha\hat{A}(\alpha\tilde{x})\tilde{y} + \alpha H(\alpha\tilde{x}) + \alpha G(\alpha\tilde{x}). \tag{9.10}$$

Here $\alpha = \alpha_{ij}$ with $0 < \alpha \le c_{i+1} - c_i$ is (an approximation to) the largest value such that

$$\max_{0\le\tilde{x}\le 1}\left\{\frac{1}{\alpha|A_j| + 1}\left|\alpha\frac{d^\nu A_j}{d\tilde{x}^\nu}\right|\right\} \le K$$

$$\max_{0\le\tilde{x}\le 1}\left|\alpha\frac{d^\nu H}{d\tilde{x}^\nu}\right| \le K \tag{9.11}$$

$$\max_{0\le\tilde{x}\le 1}\left((\max(|\alpha G^{(j)}|,1))^{-1}\,|\alpha\frac{d^\nu G^{(j)}}{d\tilde{x}^\nu}|\right) \le K,\quad j = 0,1,2,\ldots,r$$

$$\frac{|\mathrm{Im}\alpha\kappa(x)|}{|\mathrm{Re}\alpha\kappa(x)| + C_1/\rho} \le \rho \quad \text{for all } \kappa(x) = \text{ an eigenvalue of } \hat{A}(x)$$
$$j = 0,1,2,\ldots,r,\ \nu = 0,1,2,\ldots,p.$$

Here the first three conditions correspond to assumptions (3.1) and (3.2). The last condition corresponds to (9.4). One α has been determined we set $c_{i,j+1} = c_{ij} + \alpha_{ij}$.

Now if $c_{i,j+1} < c_{i+1}$ this is repeated until the endpoint c_{i+1} of the blocking subinterval is reacbed. This procedure can obviously be repeated until all blocking subintervals are divided into an appropriate number of stretching subintervals.

On every stretching subinterval $c_{ij} \le x \le c_{i,j+1}$ the system is of the form (9.10) with α replaced by α_{ij} and $\tilde{y}$ replaced by $\tilde{y}_{ij}$. The variables $\tilde{y}_{ij}$ are related by

$$\tilde{y}_{ij}(c_{ij}) = \tilde{y}_{i,j+1}(c_{ij}) \quad \text{if} \quad c_{ij} \ne c_i \text{ and } c_{ij} \ne c_{i+1}$$
$$\tilde{y}_{i+1,0}(c_{ij}) = S_+(c_i)\tilde{y}_{ij}(c_{ij}) \quad \text{if} \quad c_{ij} = c_{i+1}.$$

On every subinterval we now use a uniform meshsize $\tilde{h}$ with $K\tilde{h} \ll 1$ and employ the difference approximation

$$\frac{u_{\nu+1} - u_\nu}{\alpha\tilde{h}} = D(\hat{A}(x_\nu)u_\nu + E_\nu) + (I - D)(\hat{A}(x_{\nu+1}) + E_{\nu+1}) \tag{9.12}$$

where

$$D = \begin{pmatrix} d_r I & 0 & \cdots & \cdots & 0 \\ 0 & d_{r-1} I & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & d_0 I \end{pmatrix} \quad \text{and} \quad E_\nu = H(x_\nu)u_\nu + G(x_\nu).$$

The matrix $\hat{A}$ is not in diagonally dominant form. However, in the neighborhood of an interior point x_0 of any subinterval we can find a constant transformation S of the form

$$S = \begin{pmatrix} U_r(x_0) & L_r & & \\ & U_{r-1}(x_0) & L_{r-1} & \\ & & \ddots & \ddots \\ & & & I \end{pmatrix}$$

such that $S^{-1}\hat{A}(x)S$ satisfies the conditions of assumption 7.1. Here U_j is a unitary transformation such that $U_j^* A_j(x_0)U_j$, $j = 1,2,\cdots,r$, is upper triangular for $x = x_0$. L_j is a diagonal scaling such that $L_j^{-1}U_j^* A_j(x_0)U_j L_j$ satisfies assumption 7.1 with $\frac{1}{2}(\delta/\Gamma)$ replaced by $\frac{1}{4}(\delta/\Gamma)$. Then the smoothness properties of the coefficients guarantee that assumption 7.1 is satisfied in a whole neighborhood of x_0. Thus the solution of the differential equation satisfies estimates of the type (4.3) in the interior of any subinterval. However, for interior subintervals the estimates can be extended up to the boundary provided that

$$\begin{aligned} &|S_+(c_i)|\ |S_+^{-1}(c_i)| = O(1) \\ &\sigma^{-1} \le \bar{\alpha}_-/\bar{\alpha}_+ \le \sigma\,, \quad \sigma = O(1)\,. \end{aligned} \tag{9.13}$$

Here $\bar{\alpha}_-$, $\bar{\alpha}_+$ denote the stretching factors of any consecutive subintervals. The reason the estimates can be extended is that the breakpoints c_{ij} are somewhat arbitrary. We could move them a distance $O(1/K)$. Then the old breakpoints would become interior points in the new subintervals and we could estimate the derivatives. Provided (9.13) holds, these estimates would not be destroyed if we

were to move the breakpoints back to the original position. Finally we have to resolve the boundary layers at $x = 0,c$. This is done as described at the end of section 3.

10. Numerical details of the transformation to normal form

In this section we discuss some of the details of the numerical implementation of the transformation of a stiff system to the normal form discussed in the last section. The blocking and stretching subintervals introduced there are in practice not determined separately as in the theoretical discussion, but are determined simultaneously. For this reason, we introduce some new notation. Starting with the left endpoint $x = 0 \equiv b_0$ and working to the right, we divide the interval $[0,c]$ into stretching subintervals with endpoints b_1, b_2, etc. The block structure of the matrix is monitored as this is done, and appropriate points b_j are designated as blocking subinterval endpoints when the structure changes. The stretching parameter (see (9.10)) for the subinterval $[b_j, b_{j+1}]$ is denoted by α_j.

An outline of the algorithm for the mesh construction and the determination of the transformation to diagonally dominant form follows: We first determine the eigenvalues κ_j of $A(0)$. Then α_0 is determined by

$$-\alpha_0 \min_j \mathrm{Re}\kappa_j \approx K_2 .$$

The stretching parameter α_N for the stretching subinterval nearest $x = c$ is determined analogously. Using these stretching parameters near the endpoints of the interval assures that assumption 4.3 will be satisfied, and any possible boundary layers will be resolved. In practice we construct a "reference mesh" $\{x_j\}_{j=0}^{N}$ with $x_0 = 0$, $x_{N+1} = c$, $x_1 - x_0 = \alpha_0$ and $x_N - x_{N-1} = \alpha_N$ where the subintervals $[x_i, x_{i+1}]$ increase in length exponentially towards the center of the interval according to the rule

$$x_i - x_{i-1} = \min(2\alpha_0^i, 2^{N-i}\alpha_N, \alpha_{max})$$

where α_{max} is the maximum length for a stretching subinterval that we wish to

allow. (Typically $\alpha_{max} \sim c/10$). Then given a stretching subinterval endpoint b_j the next endpoint b_{j+1} is determined as follows:

A) If b_j has been previously determined to be the left endpoint of a blocking subinterval, then compute $S_+(b_j)$ using QR and theorem 9.1 (see remark 10.1 below). Let $\tilde{\alpha} = \alpha_{j-1}$, $\tilde{\delta} = b_j + \tilde{\alpha}$ be trial values for α_j and b_{j+1} respectively.

B) Compute the eigenvalues $\kappa(\tilde{\delta})$ of $A(\tilde{\delta})$ and determine the sets $\mathbf{M}^{(i)}$ at $x = \tilde{\delta}$. Then

a) If no sets $\mathbf{M}^{(i)}$ change, go to **C**.

b) If a new set forms, mark $\tilde{\delta}$ as a possible blocking subinterval endpoint and go to **C**.

c) If two sets merge, then b_j was a blocking subinterval endpoint. (We do not, however, need to make a special computation of $S_+(b_j)$, since taking $S_+(b_j) = S_-(b_j)$ is acceptable in this case. The difference in the treatment of this subinterval is that $S(\tilde{\delta})$ will be computed by updating $S_-(b_j)$ taking the *new* block structure into account.) Go to **C**.

C) Compute the transformation matrix $S(\tilde{\delta})$ by updating $S(b_j)$ using theorem 9.1 (see remarks 10.1 and 10.2). We need this updated version of S even if $x = \tilde{\delta}$ is a blocking subinterval endpoint

D) Now compute the left-hand sides of the tests in (9.11). (The actual implementation of these tests is discussed in remark 10.3). During the determination of b_{j+1} from b_j it is possible to return to this step (**D**) several times. Using the reference mesh, we first determine $\bar{\alpha} = x_{i+1} - x_i$ where $b_j \in [x_i, x_{i+1}]$. The action taken can be different in each case:

D1) (First time): If any test fails, replace $\tilde{\alpha}$ with $\tilde{\alpha}/\sqrt{2}$ (i.e. decrease the length of the stretching subinterval) and try again (go to **B**). If no tests fail, then replace $\tilde{\alpha}$ with $\min(\sqrt{2}\,\tilde{\alpha}, \bar{\alpha})$ (increase $\tilde{\alpha}$) and go to **B**.

D2) (Second time): If no tests failed under **D1** but any test fails this time,

set $\alpha_{ij} = \tilde{\alpha}/\sqrt{2}$ and go to **E** (below). If no test failed this time, replace $\tilde{\alpha}$ with $\min(\sqrt{2}\tilde{\alpha}, \bar{\alpha})$ and go to **B**). If any tests failed under **D1** and any test fails this time, replace $\tilde{\alpha}$ by $\tilde{\alpha}/\sqrt{2}$ and go to **B**.

D3) (Third time): If no tests fail, set $\alpha_{ij} = \tilde{\alpha}$ and go to **E**. If any test fails at this point, but no test failed under **D2**), set $\alpha_{ij} = \tilde{\alpha}/\sqrt{2}$ and go to **E**. If any test fails at this point and any test failed under **D1**, then special action must be taken, because if we were to decrease $\tilde{\alpha}$ any further, we would have $\alpha_j/\alpha_{j-1} \le \frac{1}{2}$. We would like to avoid this situation for the reasons discussed at the end of the last section (with $\sigma = 2$). First, however, we must determine how small α actually needs to be near b_j. So replace $\tilde{\alpha}$ with $\tilde{\alpha}/2$ and go to **B**.

D4) (Fourth and succeeding times): If any test fails, replace $\tilde{\alpha}$ with $\tilde{\alpha}/2$ and goto **B**. If no tests fail, then we have a value for α which represents the proper stretching near $x = b_j$. In order to use this, however, we have to redistribute the previous endpoints $b_{j-1}, b_{j-2}, \cdots$ so that there is a smooth exponential grading of subintervals into the region around $x = b_j$. We do this by first looking for the minimum value of i with

$$\frac{\alpha_{k-1}}{2} \le 2^i\tilde{\alpha} \le 2\alpha_{k-1}$$

where k is determined by first calculating

$$x = b_j - \tilde{\alpha}\sum_{l=1}^{i} 2^l ,$$

and then finding k such that $x \in [b_k, b_{k+1}]$. Then the stretching subinterval endpoints $b_{k+1}, b_{k+2}, \cdots, b_j$ are replaced by

$$b_{k+l} = \begin{cases} b_k + \tilde{\alpha}/2^{l-1}, & l = 1,2,\cdots,i \\ b_k + \tilde{\alpha}/2^{i-1}, & l = i+1,\cdots,m \end{cases}$$

where m is chosen such that $b_{k+m-1} < b_j \le b_{k+m}$. Steps **E** and **F** must now be redone for all of these corrected subintervals $[b_l, b_{l+1}]$, $l = k+1, \cdots, k+m$. Then set $j = k+m$ and go to **A**.

E) The subinterval endpoint has now been determined, i.e. $b_{j+1} = \tilde{b}$. Stretch the interval $[b_j, b_{j+1}]$ to $0 \le \tilde{x} \le 1$. Put down a uniform mesh $\{\tilde{x}_\nu^{[j]}\}_{\nu=0}^{1/\tilde{h}}$ with meshwidth $\tilde{h}$ where $\tilde{h}$ is a meshwidth that would be considered appropriate for the resolution of a smooth function on the interval $0 \le \tilde{x} \le 1$. (This is somewhat vague; typically $\tilde{h} \sim 1/10$). The meshpoints in the original variable x are given by

$$x_\nu^{[j]} = b_j + \nu \tilde{h} \alpha_j \qquad \nu = 0,1,\ldots,1/\tilde{h}.$$

F) Now compute the transformation matrices $S(x_\nu^{[j]})$, $\nu = 1,2,\ldots,1/\tilde{h}$ by updating $S_+(b_j)$ using theorem 9.1 (again see remarks 10.1 and 10.2). The difference approximation can now be written down for the mesh intervals $[x_\nu^{[j]}, x_{\nu+1}^{[j]}]$, $\nu = 0,1,2,\ldots,1/\tilde{h}-1$. Suppressing the superscript $[j]$, the difference approximation is given by

$$\begin{aligned} \tilde{v}_{\nu+1} - \tilde{v}_\nu = {} & D_\nu(h_\nu \hat{A}_{\nu+1} - S_{\nu+1}^{-1}(S_{\nu+1} - S_\nu))\tilde{v}_{\nu+1} \\ & + (I - D_\nu)(h_\nu \hat{A}_\nu - S_\nu^{-1}(S_{\nu+1} - S_\nu))\tilde{v}_\nu \\ & + h_\nu D_\nu S_{\nu+1}^{-1} F_{\nu+1} + h_\nu (I - D_\nu) S_\nu^{-1} F_\nu \end{aligned} \tag{10.1}$$

where $S_\nu = S(x_\nu^{[j]})$ and $\tilde{v}_\nu$ is an approximation to $\tilde{y}(x_\nu)$. In practice we have found it more convenient to make our computations in terms of the *original* (untransformed) variables $y(x_\nu)$ which we approximate with v_ν. The difference

approximation then becomes

$$\begin{aligned} S_{\nu+1}^{-1}v_{\nu+1} - S_{\nu}^{-1}v_{\nu} &= D_{\nu}(h_{\nu}S_{\nu+1}^{-1}A_{\nu+1} + S_{\nu+1}^{-1} - S_{\nu}^{-1})v_{\nu+1} \\ &+ (I - D_{\nu})(h_{\nu}S_{\nu}^{-1}A_{\nu} + S_{\nu+1}^{-1} - S_{\nu}^{-1})v_{\nu} \\ &+ h_{\nu}D_{\nu}S_{\nu+1}^{-1}F_{\nu+1} + h_{\nu}(I - D_{\nu})S_{\nu}^{-1}F_{\nu} \end{aligned} \tag{10.2}$$

where $A_{\nu} = A(x_{\nu}^{[j]})$. This is done because it is usually the original variables that we are interested in. One should note, however, that if $|S_{\nu}| + |S_{\nu}^{-1}|$ is not of reasonable size, one can expect the system (10.1) to be somewhat better conditioned that (10.2). The transformed variables $\tilde{y}$ are in some sense the "correct" variables for the problem since they have been scaled in such a way that we can obtain estimates for the system.

At this point we can now increment j and return to step **A**.

Remarks 10.1: When computing the transformation $S_{+}(b_j)$ at the left endpoint of a *blocking* subinterval, we first transform $A(b_j)$ to upper triangular form using the QR method. If the eigenvalues do not appear in the correct order on the diagonal of the transformed matrix, in practice we repeat the QR iteration using the (now known) eigenvalues as shifts in the order in which we wish them to appear.

The resulting matrix is then transformed to block diagonal form using lemmata 9.2 and 9.3. Note that if the eigenvalue sets $\mathbf{M}^{(i)}$ are well separated then the iteration (9.8) can be replaced with

$$A_{11}R^{(n)} = R^{(n-1)}A_{22} + R^{(n-1)}B_{21}R^{(n-1)} - B_{12}. \tag{10.3}$$

i.e. we only need to invert A_{11}. (This remark also applies when updating S later on).

Note also that the off-diagonal blocks need not be completely eliminated as any $O(1)$ blocks can be absorbed into the matrix H of (9.9).

Remark 10.2 As long as a blocking subinterval endpoint does not lie between two points x_j, x_{j+1}, $S(x_{j+1})$ can always be computed from $S(x_j)$ by updating $S(x_j)$ using the blocking technique of lemmata 9.2,9.3. In practice the iteration (9.8) is replaced by

$$\tilde{A}_{11}\tilde{R}^{(n)} - \tilde{R}^{(n)}\tilde{A}_{22} - \tilde{R}^{(n-1)}\tilde{B}_{21}\tilde{R}^{(n-1)} + \tilde{B}_{12} = 0 \tag{10.4}$$

where

$$\tilde{A}_{11} = U_1^* A_{11} U_1 , \quad \tilde{A}_{22} = U_2^* A_{22} U_1$$

are upper triangular (U_1 and U_2 are unitary and are determined by QR). Here $\tilde{B}_{12} = U_1^* B_{12} U_2$ and $\tilde{B}_{21} = U_2^* B_{21} U_1$. R is then computed from $\tilde{R}$ using $R = U_1 \tilde{R} U_2^*$. This simplifies the computation since in particular the iteration (10.4) only involves the solution of systems of the form (9.6). The only exception is if the eigenvalue sets $\mathbf{M}^{(i)}$ are well enough separated so that (10.3) can be used in place of (9.8).

Remark 10.3: In practice we only compute the tests (9.11) with $p = 1$. Although this means that we do not necessarily get the smoothness required for good error estimates, our experience has been that we always obtain satisfactory results. Also taking $p = 1$ means that the difference approximations to (9.11) will only involve two adjacent subinterval endpoints b_j, which simplifies the algorithm considerably. The difference approximations for the smoothness tests are described as follows: Suppose we want to assure that the function $f(x)$ satisfies

$$(\max(|\alpha f(\alpha \tilde{x})|, 1))^{-1} \alpha d f(\alpha \tilde{x}) / d\tilde{x} \leq K. \tag{10.5}$$

In practice we replace this with the test

$$\frac{\tilde{\alpha}\,|f(b_j + \tilde{\alpha}) - f(b_j)|}{1 + \tilde{\alpha}\,|f(b_j)|} \leq \tilde{K} \tag{10.6}$$

for some appropriate value of $\tilde{K} \approx K$, where b_j is the previously determined subinterval endpoint and $\tilde{\alpha}$ is the stretching parameter that we are testing. (10.6) is obtained by stretching $[b_j, b_j + \tilde{\alpha}]$ to $[0,1]$ and then replacing the derivative of (10.5) with a divided difference over the whole interval $0 \le \tilde{x} \le 1$. The denominator of (10.5) has been replaced by a sum because it is cheaper to compute than the maximum but gives approximately the same effect.

We have computed several examples to test our procedures, and we present three of them here. Each example was chosen to test a different aspect of the transformation to normal form. For the first example we consider the system

$$\frac{d}{dx}\begin{bmatrix} y \\ v \end{bmatrix} + \begin{bmatrix} \frac{1}{\varepsilon}(x^3 - \frac{1}{2}x) & \frac{1}{\varepsilon} \\ \frac{1}{2} + 3x^2 & 0 \end{bmatrix}\begin{bmatrix} y \\ v \end{bmatrix} = 0\,, \quad -1 \le x \le 1$$

with $y(-1) = 1$, $y(1) = 2$. This problem has turning point behavior near $x = -\sqrt{2}, 0, \sqrt{2}$, and thus we are testing the aspect of the mesh construction that looks at the smoothness of the coefficients in order to determine the proper stretching. Figure (10.1) shows the results of a computation with $\varepsilon = 10^{-5}$. Only the approximation to $y(x)$ is shown. As with the earlier plots, the horizontal lines above and below the plots are used to indicate the locations of the the meshpoints.

The second example (figure 10.2) shows an example with both boundary layers and a possible turning point at $x = 0$. The system is given by

$$\frac{d}{dx}\begin{bmatrix} y \\ v \end{bmatrix} = \begin{bmatrix} \frac{x}{\varepsilon} & -\frac{1}{\varepsilon} \\ \frac{1}{2} & 0 \end{bmatrix}\begin{bmatrix} y \\ v \end{bmatrix}, \quad y(-1)=1, y(1) = 2, \varepsilon = 10^{-3}\,.$$

Note that the region of the possible turning point has been refined even though the solution is smooth there. In order that our code be robust we have chosen to "resolve" all possible unsmooth behavior even though in cases such as this

one it might be possible to tell a priori that the solution will be smooth there.

In the third example (figures 10.3, 10.4), we test the feature of the mesh refinement algorithm that resolves possible highly oscillatory behavior. The system we consider is given by

$$\frac{d}{dx}\begin{bmatrix} y \\ v \end{bmatrix} + \begin{bmatrix} \frac{x^2}{\varepsilon} & \frac{1}{\varepsilon} \\ 2x - 1 & 0 \end{bmatrix}\begin{bmatrix} y \\ v \end{bmatrix} = 0, \quad y(-1) = 1, y(1) = 2, \varepsilon = 10^{-4}.$$

In this example, the eigenvalues of the matrix become complex in a neighborhood of $x = 0$ of width $O(\varepsilon^{\frac{1}{4}})$. The mesh in this region is thus determined by imposing the fourth of conditions (9.11). Another interesting feature of this example is that it is not particularly well-posed. The solution becomes very large near the left boundary ($\|y(x)\|_\infty = 8.9\times10^9$). However, the method is able to handle the situation quite well. Figure 10.4 is a magnification of the region near $x = 0$. This shows the oscillations that occur near $x = 0$, and demonstrates that the mesh has been properly scaled to resolve these oscillations.

E dd(y)/dxdx + (x^3 - x/2) dy/dx - y = 0

E = epsilon = 0.1000E-04 No. of meshpoints = 124

ymin = -0.8565E-01 ymax = 0.1871E+01

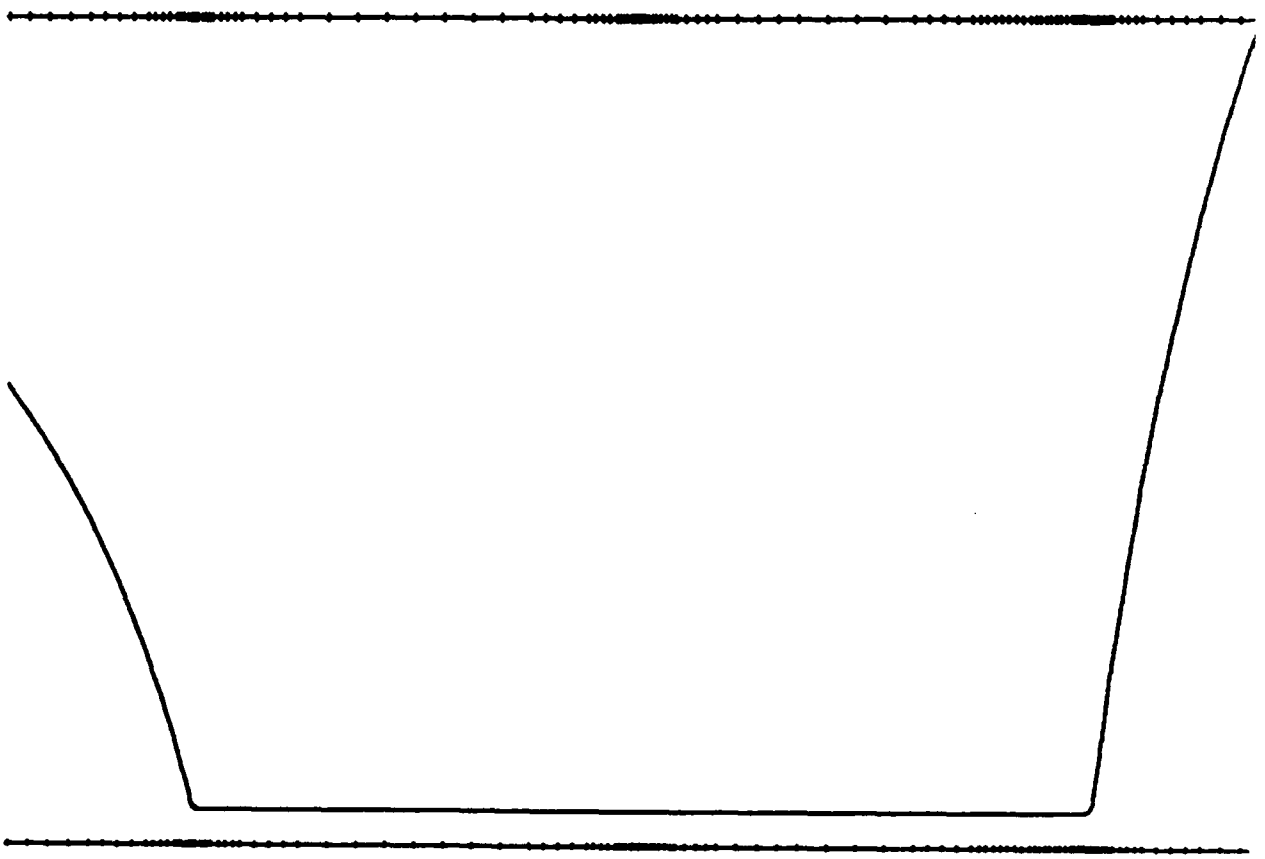


Figure 10.1

E dd(y)/dxdx - x dy/dx - y/2 = 0

E = epsilon = 0.1000E-02 No. of meshpoints = 106

ymin = -0.2345E-01 ymax = 0.2000E+01

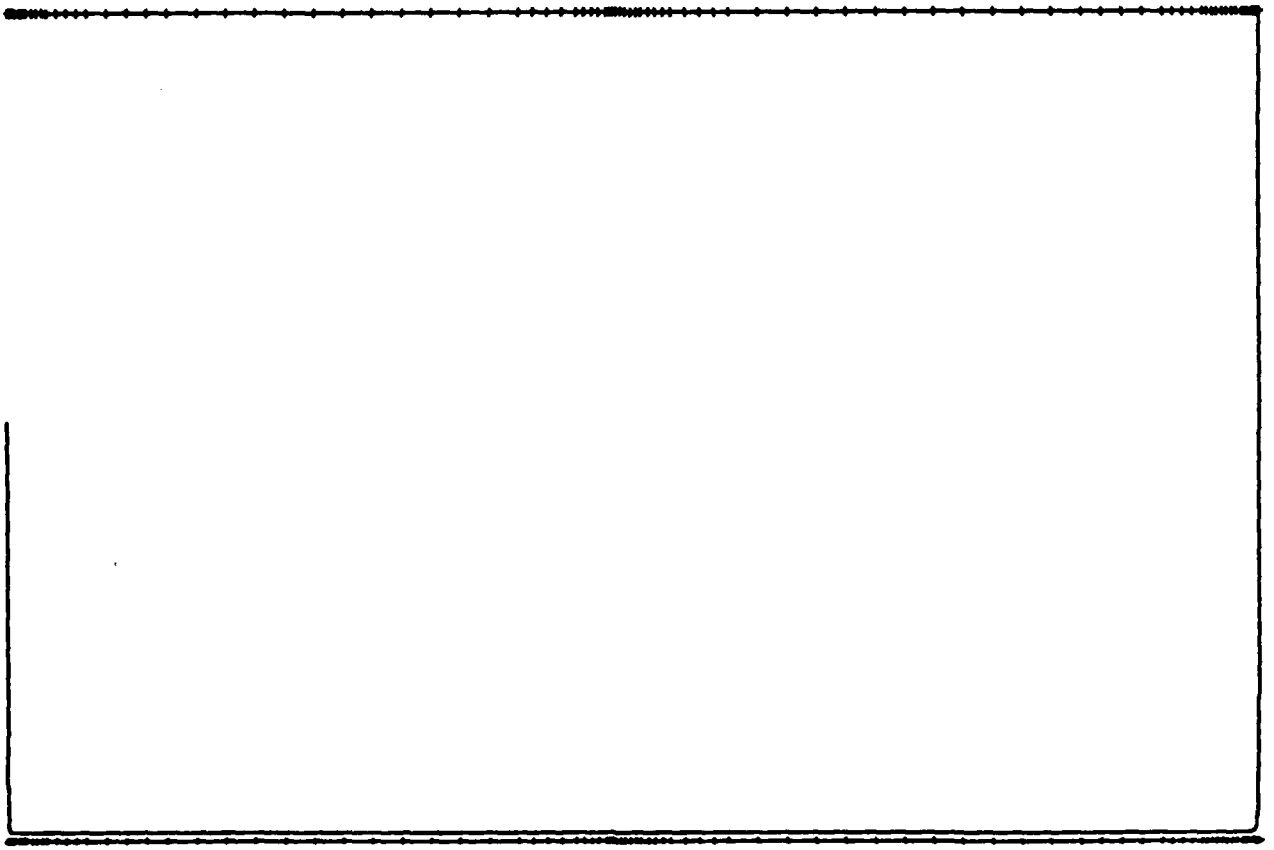


Figure 10.2

E dd(y)/dxdx + (x^2) dy/dx + y = 0

E = epsilon = 0.1000E-03 No. of meshpoints = 229

ymin = -0.8270E+09 ymax = 0.8912E+10

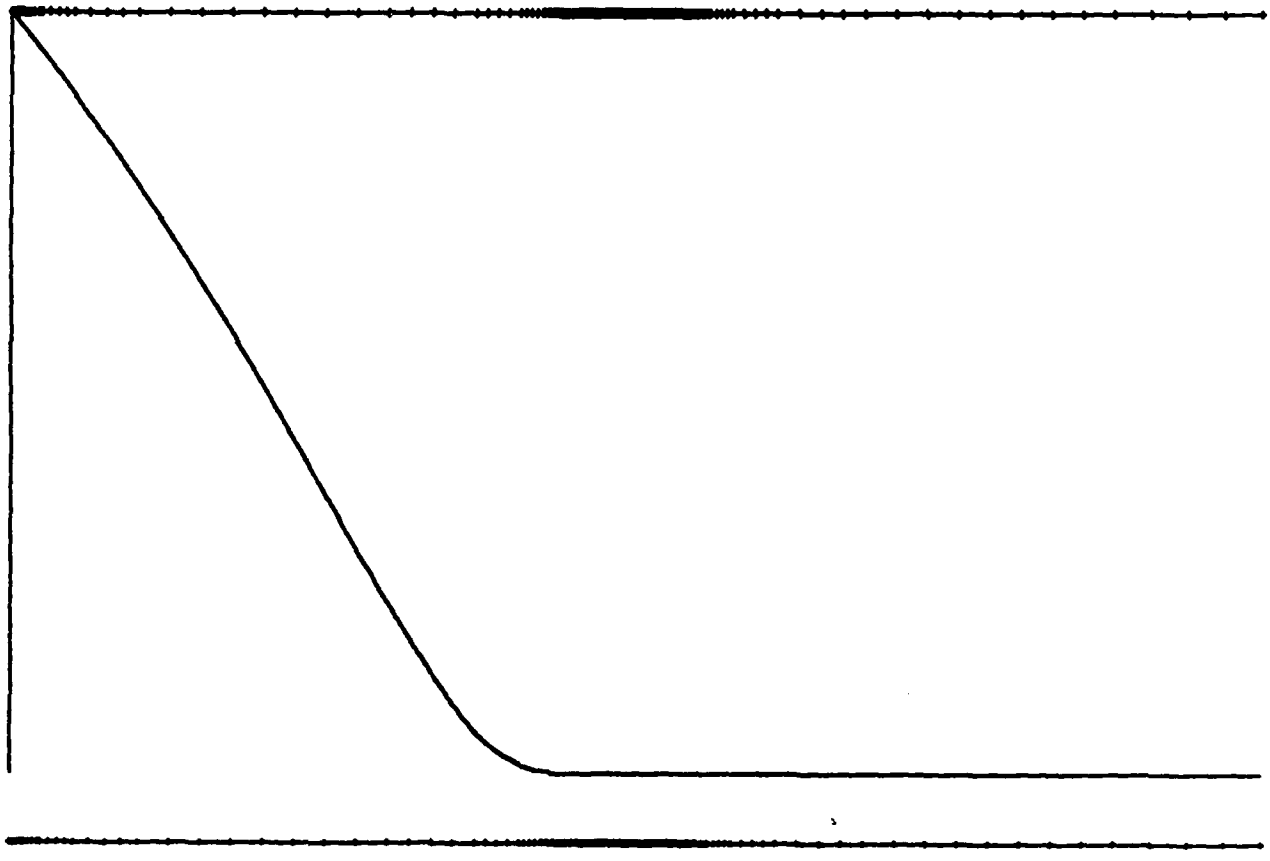


Figure 10.3

E dd(y)/dxdx + (x^2) dy/dx + y = 0

E = epsilon = 0.1000E-03 No. of meshpoints = 150
xmin = -0.1212E+00 xmax = 0.4671E+00
ymin = -0.9179E+06 ymax = 0.2673E+07

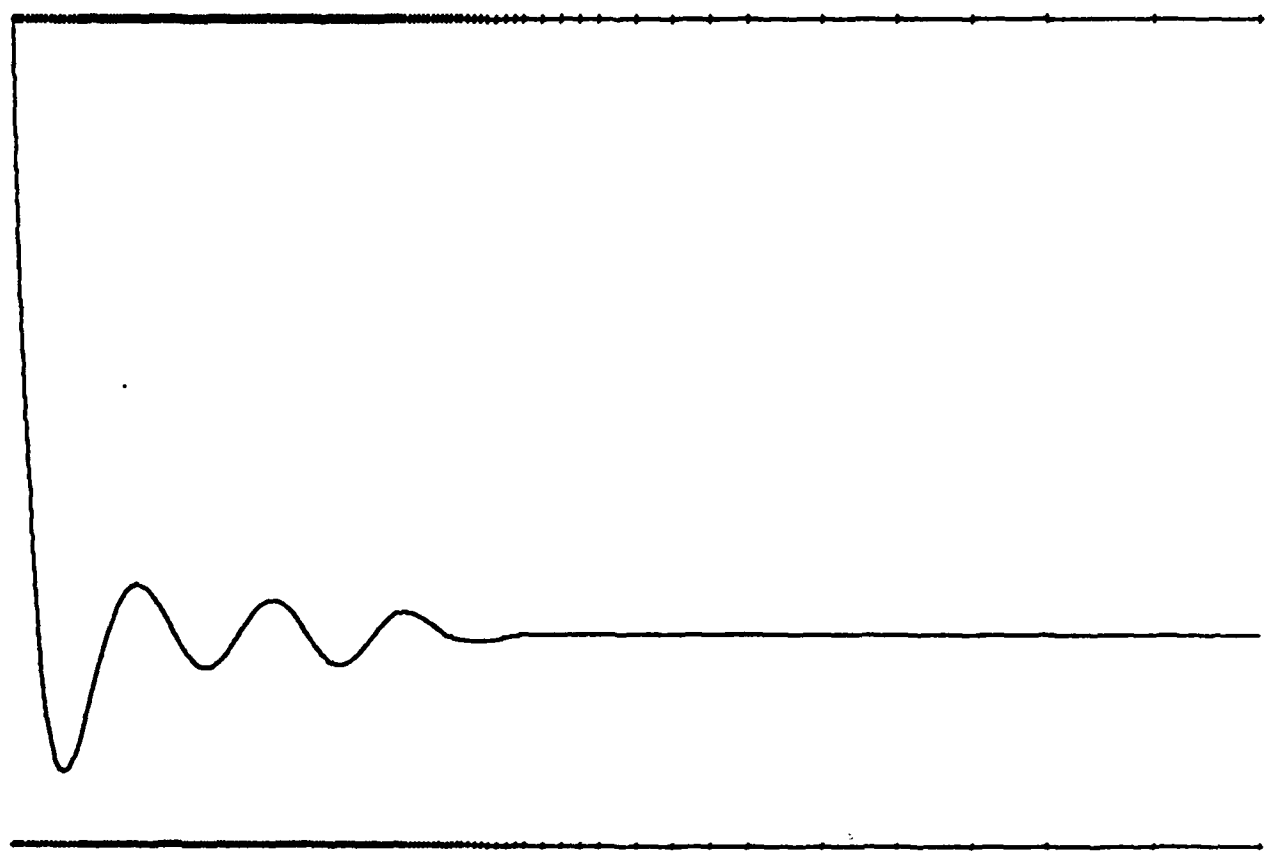


Figure 10.4

11. Solution of nonlinear systems

It is relatively straight-forward to apply our methods to nonlinear problems of the form

$$\frac{d}{dx}y(x) = f(y(x)) \qquad (11.1)$$

where for simplicity we specify linear boundary conditions (1.2). Here y and f are vector functions of dimension n. As in [3], we solve (11.1),(1.2) using a functional Newton iteration technique: We linearize (11.1) about a previous guess or approximate solution $y^n(x)$, obtaining

$$\frac{d}{dx}\hat{y}(x) = \frac{\partial f}{\partial y}(y^n(x))\hat{y}(x) + f(y^n(x)) - \frac{d}{dx}y^n(x)\,. \qquad (11.2)$$

Here $\hat{y}(x) = y^{n+1} - y^n$ is the correction to the guess y^n, and $\frac{\partial f}{\partial y}(y^n(x)) = A(x)$ is the Jacobian matrix of f. Letting $\hat{F} = f(y^n(x)) - \frac{d}{dx}y^n(x)$ we see that (11.2) is in the same form as (1.1) and so we can apply the methods discussed earlier in this paper for linear problems. We emphasize here that the linearization is done *before* the method is applied. This is in contrast to the usual approach for solving noninear ODEs, which is to first apply the difference method, and then use a Newton iteration to solve the resulting system of nonlinear algebraic equations. The reason for this is that we can only expect our method to apply to *linear* differential equations, and so to be certain that it works, we must first reduce the nonlinear ODEs to linear ones.

We have used this approach to solve a simple test problem, given by

$$\varepsilon y'' + \frac{1}{2}(y^2)' - y = 0, \quad -1 \le x \le 1, \quad y(-1) = 1, \quad y(1) = 2\,.$$

As in [3] we replace this with the 2×2 system of equations

$$\begin{aligned} \varepsilon y' &= -\tfrac{1}{2}y^2 + v \\ v' &= y \end{aligned} \tag{11.3}$$

and then linearize this system. The matrix transformations and mesh construction are the same as before. The difference approximation is essentially the same as (10.2) except that the term dy^n/dx is grouped with the term $d\hat{y}/dx$ when making the difference approximation. In terms of v_ν and v_ν^n, the approximations to $\hat{y}(x_\nu)$ and $y^n(x_\nu)$ respectively, and letting $F_\nu = f(y^n(x_\nu))$, we add the terms

$$(D_\nu(S_{\nu+1}^{-1} - S_\nu^{-1}) - S_{\nu+1}^{-1})y_{\nu+1}^n + ((I - D_\nu)(S_{\nu+1}^{-1} - S_\nu^{-1}) + S_\nu^{-1})y_\nu^n$$

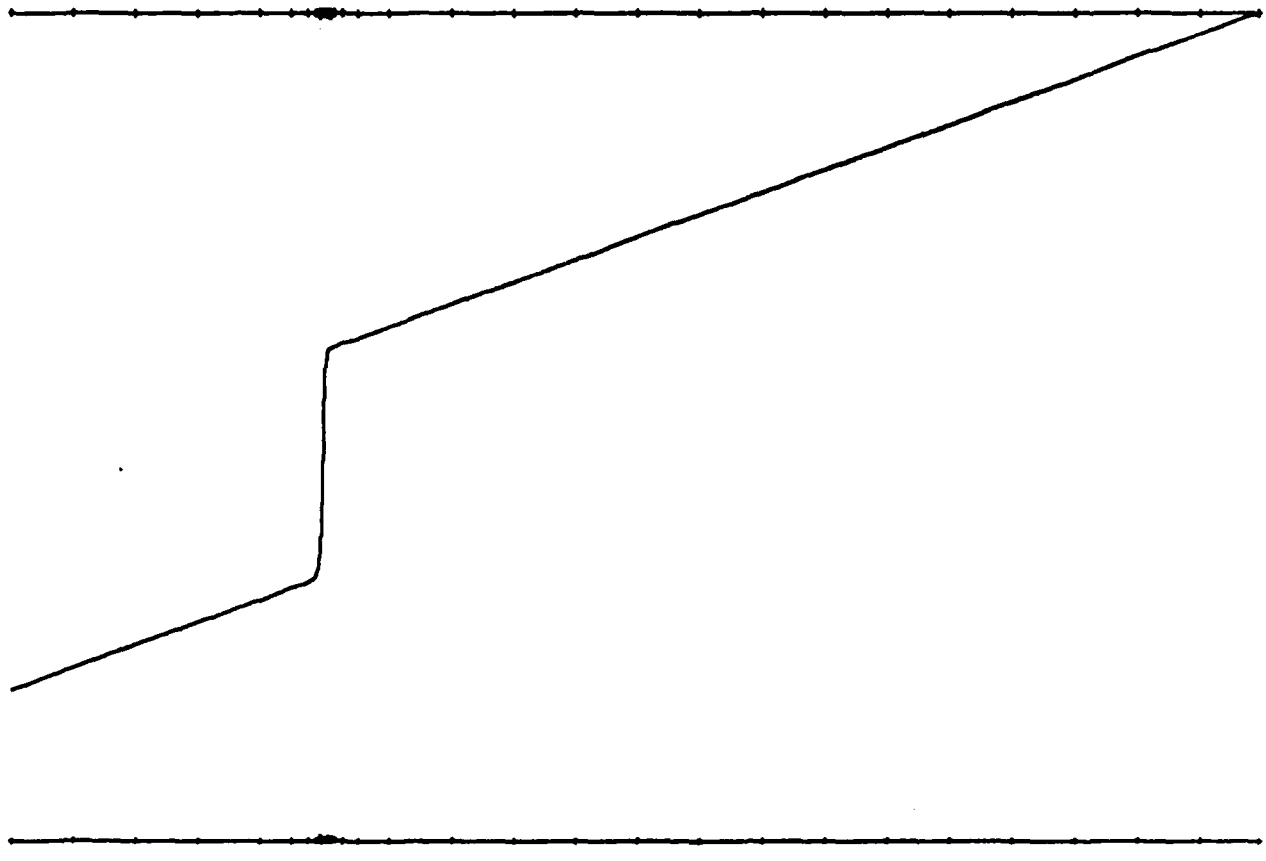
to the right-hand side of (10.2) to obtain our difference approximation.

Figure 11.1 shows the result of a numerical computation of the solution of (11.3) for $\varepsilon = 10^{-3}$. We started with an initial guess for y of a straight line between the boundary values and continued in ε with the values $\varepsilon =$ .1, .03, .015, .0075, .004, .001. Convergence was obtained to a tolerance of less than 10^{-3} in the maximum norm at each step.

Figure 11.1

E = epsilon = 0.1000E-02 No. of meshpoints = 47

ymin = -0.1673E+01 ymax = 0.2000E+01



References

[1] U. Ascher, J. Christiansen and R.D. Russell, *A collocation solver for mixed order systems of boundary value problems*, Math. Comp., 33, (1979), pp. 659-579.

[2] U. Ascher and R. Weiss, *Collocation for singular perturbation problems II: Linear first order systems without turning points*, University of British Columbia Dept. of Computer Science Technical Rept. 82-4, (1982).

[3] B. Kreiss and H.O. Kreiss, *Numerical methods for singular perturbation problems*, SIAM J. Num. Anal., 18 (1981), pp. 262-276.

[4] H.O. Kreiss, *Difference methods for stiff ordinary differential equations*, SIAM J. Num. Anal., 15, pp. 21-58.

[5] R. E. Schied, *The accurate numerical solution of highly oscillatory ordinary differential equations*, (1983), to appear in Math. Comp.

[6] R. Weiss, *An analysis of the box and trapezoidal schemes for linear singularly perturbed boundary value problems*, (1983), to appear in Math. Comp.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 2599	2. GOVT ACCESSION NO.: AD-A136427	3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): NUMERICAL METHODS FOR STIFF TWO-POINT BOUNDARY VALUE PROBLEMS	5. TYPE OF REPORT & PERIOD COVERED: Summary Report - no specific reporting period
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Heinz-Otto Kreiss, N. K. Nichols and David L. Brown	8. CONTRACT OR GRANT NUMBER(s): N00014-83-K-0422, DAAG29-80-C-0041, AFOSR-82-0321
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Mathematics Research Center, University of Wisconsin, 610 Walnut Street, Madison, Wisconsin 53706	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Work Unit Number 3 - Numerical Analysis and Scientific Computing
11. CONTROLLING OFFICE NAME AND ADDRESS: See Item 18 below.	12. REPORT DATE: November 1983
	13. NUMBER OF PAGES: 82
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office):	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

U. S. Army Research Office	Office of Naval Research	Air Force Office of Scientific Research
P. O. Box 12211	800 North Quincy St.	Building 410, Bolling AFB
Research Triangle Park	Arlington, VA 22217	Washington, DC 20332
North Carolina 27709		

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Stiff
ODE
Boundary value problem
Turning points
Difference equations
Essentially diagonally dominant form
Robust
Upwinding

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We consider the two-point boundary value problem for stiff systems of ordinary differential equations. For systems that can be transformed to essentially diagonally dominant form with appropriate smoothness conditions, a priori estimates are obtained. Problems with turning points can be treated with this theory, and we discuss this in detail. We give robust difference

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. ABSTRACT - cont'd.

approximations and present error estimates for these schemes. In particular we give a detailed description of how to transform a general system to essentially diagonally dominant form and then stretch the independent variable so that the system will satisfy the correct smoothness conditions. Numerical examples are presented for both linear and nonlinear problems.

E

ED

8