MATHEMATICS OF COMPUTATION, VOLUME 33, NUMBER 146
APRIL 1979, PAGES 659–679

A Collocation Solver for Mixed Order Systems of Boundary Value Problems

By U. Ascher*, J. Christiansen** and R. D. Russell**

Abstract. Implementation of a spline collocation method for solving boundary value problems for mixed order systems of ordinary differential equations is discussed.

The aspects of this method considered include error estimation, adaptive mesh selection, B-spline basis function evaluation, linear system solution and nonlinear problem solution.

The resulting general purpose code, COLSYS, is tested on a number of examples to demonstrate its stability, efficiency and flexibility.

1. Introduction. Recently there have been several efforts to develop high quality, general purpose software for the solution of boundary value problems for systems of ordinary differential equations. Most of the codes developed have been based on initial value methods, reflecting the current advanced state of such methods. In particular, multiple shooting codes have been developed by England, Nichols and Reid [19] and by Bulirsch, Stoer and Deuflhard [10]. Successful solution of some difficult nonlinear problems with the latter code is reported in [17]. Also, Scott and Watts have produced a superposition code with orthonormalization [35]. A comparison of some initial value type codes is given in [36].

A second approach has been implemented by Lentini and Pereyra [24], [25], where a finite difference method with deferred corrections is used.

A thorough theoretical analysis of finite element methods has been available for some time [13], [33], [7], but, to our knowledge, there has been no attempt, prior to this work, to write a general purpose code using these methods.

In this paper we discuss an implementation of a spline collocation method for solving boundary value problems for mixed order systems of ordinary differential equations. While not in polished form, our code COLSYS (COLlocation for SYStems) is sufficiently stabilized that we are able to present a number of its theoretical and practical aspects and demonstrate the power of this preliminary version.

There are a number of reasons for our choice of the collocation method. It is the most suitable method among the finite element ones, for a general purpose code. See [1], [30] and [31], where complexity comparisons are made which support the above claim and also show collocation, when efficiently implemented, to be competi-

Received November 28, 1977; revised June 12, 1978.
AMS (MOS) subject classifications (1970). Primary 65L10.
*Supported in part under NRC (Canada) Grant A4306.
**Supported in part under NRC (Canada) Grant A7871.

0025-5718/79/0000-0061/$06.25

tive with finite differences using extrapolation. The theoretical results on the convergence of the collocation method [11], [23], together with those on error estimation and mesh selection [32], [12] are more general than for the other methods mentioned. This, and the basic simplicity of the collocation procedure, also make programming of the method reasonably straightforward. COLSYS is designed to solve mixed order systems of nonlinear boundary value problems. This is in contrast to the other codes mentioned above which require conversion of a given problem to a first order system, thereby increasing the number of equations and changing the algebraic structure of the discretized problem. Numerous numerical experiments have demonstrated the stability of the collocation procedure, and recent attempts at adaptive mesh selection and error estimation have been quite successful [32]. For these reasons we feel that a robust, efficient collocation code can be developed to reliably solve a larger class of problems than has heretofore been possible.

Most of the points mentioned above are discussed and demonstrated in greater detail in the rest of the paper. In Section 2 the collocation theory for mixed order boundary value systems [11], [23] is extended to obtain an error expression useful for adaptive mesh selection, generalizing a result in [32] for a scalar equation. Also, a theoretical justification of the error estimation strategy as well as practical aspects of these features are given.

Section 3 considers the method used for evaluating the piecewise polynomial collocation solution, expressed in terms of a B-spline basis. This involves appropriate modification of de Boor's B-spline evaluation procedures [4].

Section 4 describes some aspects of solving the collocation equations. Newton's method is currently used for solving nonlinear problems. For each Newton iteration, the resulting linear algebraic system of equations is solved using a package developed by de Boor and Weiss [8], after first bringing the equations into a banded block structure.

Several representative test problems, demonstrating the stability and flexibility of COLSYS, are documented in Section 5. A more detailed set of problems are given in [2], where the linear examples are also tried with two other codes [35], [24] in order to put COLSYS in a perspective. From our considerable testing, COLSYS appears to be competitive in general and particularly suitable for mildly difficult and difficult problems. Several of the examples with a small parameter are best solved by COLSYS. It is also the only one which can solve some problems with singularities without any modification. The relative efficiency of the code increases for problems of higher order and more than one component.

2. Error Estimates and Mesh Selection. The class of problems treated by our code has the following general form: A system of d nonlinear differential equations of orders $m_1 \leqslant m_2 \leqslant \cdots \leqslant m_d$,

$$u_n^{(m_n)}(x) = F_n(x; u_1, u_1', \ldots, u_1^{(m_1-1)}, u_2, \ldots, u_d^{(m_d-1)}) \equiv F_n(x; \mathbf{z}(\mathbf{u})), \tag{2.1}$$

$$a \leqslant x \leqslant b, \quad n = 1, \ldots, d,$$

is subject to nonlinear side conditions,

$$(2.2)\qquad g_j(\zeta_j;\, \mathbf{z}(\mathbf{u})) = 0, \qquad \zeta_1 \leqslant \zeta_2 \leqslant \cdots \leqslant \zeta_{m^*},\ \zeta_j \in [a,\, b],\ j = 1, \ldots, m^*,$$

where $m^* = \Sigma_{n=1}^{d} m_n$. To conveniently facilitate an efficient implementation we require that $m_d \leqslant 4$ and that the side conditions (2.2) each involve only one point. Thus, for example, periodic boundary conditions are excluded. However, any problem with such nonseparated conditions (and even interface conditions) can be cast into form (2.1), (2.2) at the expense of increasing the size d of the problem, as shown by example in [2].

To be able to apply the collocation theory we need to have an isolated solution $\mathbf{u}$ to (2.1)–(2.2). This occurs if the linearized problem at $\mathbf{u}$ is uniquely solvable. Specifically, consider the curve $C \subset \mathcal{R}^{m^*+1}$ defined by

$$C \equiv \{[x,\, u_1(x), \ldots, u_1^{(m_1-1)}(x), \ldots, u_d^{(m_d-1)}(x)] : x \in [a,\, b]\},$$

and the linear problem

$$(2.3)\qquad L_n \mathbf{w} = 0, \qquad n = 1, \ldots, d,$$

$$(2.4)\qquad B_j \mathbf{w} = 0, \qquad j = 1, \ldots, m^*,$$

where $\mathbf{w} = (w_1, \ldots, w_d)$,

$$(2.3a)\qquad L_n \mathbf{w} \equiv L_n(\mathbf{u})\mathbf{w} = w_n^{(m_n)} - \sum_{l=1}^{m^*} \frac{\partial F_n(\cdot;\, \mathbf{z}(\mathbf{u}))}{\partial z_l} \cdot z_l(\mathbf{w}),$$

$$(2.4a)\qquad \beta_j \mathbf{w} \equiv \beta_j(\mathbf{u})\mathbf{w} = \sum_{l=1}^{m^*} \frac{\partial g_j(\zeta_j;\, \mathbf{z}(\mathbf{u}))}{\partial z_l} \cdot z_l(\mathbf{w}).$$

If the Green's function $G(x,\, t)$ for (2.3)–(2.4) exists (implying unique existence for the linearized problem) and $F_1, \ldots, F_d, g_1, \ldots, g_{m^*}$ are sufficiently smooth in some δ-neighborhood of C, this is sufficient to guarantee that there exists a $\sigma > 0$ such that $\mathbf{u}(x)$ is the unique solution of (2.1)–(2.2) in the sphere $B(D^m \mathbf{u},\, \sigma) = \{\mathbf{w}(x) : \|w_n^{(m_n)} - u_n^{(m_n)}\| \leqslant \sigma,\ n = 1, \ldots, d\}$ [11]. This also implies that Newton's method converges quadratically if the initial approximation is sufficiently close to $\mathbf{u}(x)$.

To solve (2.1)–(2.2) numerically, we apply collocation at Gaussian points, using piecewise polynomial functions. If π is a partition of $[a,\, b]$

$$(2.5)\qquad \begin{cases} \pi\colon a = x_1 < x_2 < \cdots < x_N < x_{N+1} = b, \\ I_i = (x_i,\, x_{i+1}), \quad h_i = x_{i+1} - x_i, \quad i = 1, \ldots, N, \\ h = \max\limits_{1 \leqslant i \leqslant N} h_i \end{cases}$$

and $\mathcal{P}_{k,\pi} = \{v \mid v$ is a polynomial of order k (degree $< k$) on $I_i,\ i = 1, \ldots, N\}$, then we seek an approximate solution $\mathbf{v} = (v_1, \ldots, v_d)$ such that $v_n \in \mathcal{P}_{k+m_n,\pi} \cap C^{(m_n-1)}[a,\, b]$, $n = 1, \ldots, d$, or $\mathbf{v} \in \mathcal{P}_{k+\mathrm{m},\,\pi} \cap C^{(\mathrm{m}-1)}[a,\, b]$. We require $k \geqslant m_d$, where k is the number of collocation points per subinterval. If $\{\rho_j\}_{j=1}^{k}$ are the Gauss-

Legendre points on $[-1, 1]$, then $\{x_{ij}\}_{i=1,j=1}^{N,k}$ are the collocation points, where

$$x_{ij} = \frac{x_i + x_{i+1}}{2} + \tfrac{1}{2}h_i\rho_j \equiv x_{i+1/2} + \tfrac{1}{2}h_i\rho_j. \tag{2.6}$$

The collocation equations which $\mathbf{v}$ has to satisfy are thus

$$v_n^{(m_n)}(x_{ij}) = F_n(x_{ij}; \mathbf{z}(\mathbf{v})), \qquad j = 1, \ldots, k,\ i = 1, \ldots, N,\ n = 1, \ldots, d, \tag{2.7}$$

and (2.2).

The theory and a priori error estimates for collocation have been presented in [11] (cf. also [23], [29], [7], [38]). Here, we merely quote the results that (assuming sufficient smoothness)

$$\|u_n^{(l)} - v_n^{(l)}\|_\infty = O(h^{k+m_n-l}), \qquad l = 0, \ldots, m_n,\ n = 1, \ldots, d, \tag{2.8}$$

and, at the mesh points, superconvergence occurs

$$|(u_n^{(l)} - v_n^{(l)})(x_i)| = O(h^{2k}), \qquad i = 1, \ldots, N,\ l = 0, \ldots, m_n - 1,\ n = 1, \ldots, d. \tag{2.9}$$

The phenomenon of higher order accuracy at the mesh points displayed in (2.9) may suggest (as has been noted in various places in the literature) using a posteriori high order interpolation of an approximate solution at the mesh points to improve the overall accuracy, at least when $k > m_d$. However, it has been the experience of these authors and others that this is generally not a very useful idea, as the asymptotic range of h, $0 < h \leqslant h_0$, where the superiority of the bound (2.9) over (2.8) is demonstrated, occurs very often for an h_0 which is effectively too small.

We feel that in practice it is more significant that the main term of the error expression is local if $k > m_d$. Below, we briefly describe this analysis which is similar to that in [32] (cf. also [5]).

It is known [11] that a collocation solution of the linearized problem

$$L_n\mathbf{w} = L_n\mathbf{u}, \qquad n = 1, \ldots, d, \tag{2.10a}$$

$$\beta_j\mathbf{w} = \beta_j\mathbf{u}, \qquad j = 1, \ldots, m^*, \tag{2.10b}$$

where L_n and β_j are defined in (2.3a), (2.4a), lies within $O(h^{2k})$ of the collocation solution of the original problem (2.1), (2.2). Therefore, for terms of order less than $2k$ in h, one need only consider the form of the error for linear problems. The Green's function $G(x, t)$ exists if the linear problem has a unique solution. If (2.10a) is cast as a system of m^* first order equations, with one component assigned to each of $u_n^{(l)}$, $l = 0, \ldots, m_n - 1$, $n = 1, \ldots, d$, then the Green's function $K(x, t)$ for this first order system can be constructed as in [29]. The Green's function $G(x, t)$ for the system (2.10a), (2.10b) then consists of a subset of the components of $K(x, t)$–see [11]. Using a general form for $K(x, t)$, it can be shown that as a function of t, $G_{ni}(x, t)$ is in $C^{(m_n-1)}[a, b]$ if $i \neq n$ and $G_{nn}(x, t)$ is in $C^{(m_n-2)}[a, b]$ with

$$\frac{\partial^{m_n-1}}{\partial t^{m_n-1}}G_{nn}(t, t^+) - \frac{\partial^{m_n-1}}{\partial t^{m_n-1}}G_{nn}(t, t^-) = (-1)^{m_n}.$$

If $\mathbf{r}(x) \equiv L(\mathbf{u} - \mathbf{v})(x)$, then the error is $\mathbf{e}(x) \equiv \mathbf{u}(x) - \mathbf{v}(x) = \int_a^b G(x, t)\mathbf{r}(t)\,dt$. Using (2.7) for the linear problem, we obtain (since $r_n(x_{ij}) = 0, j = 1, \ldots, k$)

$$r_n(t) = \frac{r_n^{(k)}(\sigma_{ni}(t))}{k!} \cdot \prod_{j=1}^{k} (t - x_{ij}) \quad \text{for } t \in [x_i, x_{i+1}],$$

for some $\sigma_{ni}(t) \in [x_i, x_{i+1}]$. So the error in the nth component is

$$e_n(x) = \sum_{i=1}^{N} \sum_{l=1}^{d} \int_{x_i}^{x_{i+1}} G_{nl}(x, t) \cdot r_l^{(k)}(\sigma_{li}(t)) \prod_{j=1}^{k} (t - x_{ij})/k!\,dt, \quad n = 1, \ldots, d.$$

The convergence results (2.8) and continuity arguments as in [32] imply, for $k > m_d$,

$$e_n^{(l)}(x) = \frac{u_n^{(k+m_n)}(x_i)}{2^{k+m_n-l}} P_n^{(l)}\left(\frac{2}{h_i}(x - x_{i+1/2})\right) h_i^{k+m_n-l} + O(h^{k+m_n+1-l}), \tag{2.11}$$

$$n = 1, \ldots, d,$$

for $x \in I_i$, where

$$P_n(\xi) = \int_{-1}^{\xi} \frac{(-1)^{m_n-1}(t - \xi)^{m_n-1}}{k!(m_n - 1)!} \prod_{j=1}^{k} (t - \rho_j)\,dt = \frac{d^{k-m_n}}{d\xi^{k-m_n}} \frac{(\xi^2 - 1)^k}{2k!} \tag{2.12}$$

for $\xi \in (-1, 1)$.

In arriving at error estimation and mesh selection schemes we assume that the local term in the error expression (2.11) is the dominant one. This, of course, can be guaranteed only when the mesh is quasiuniform, i.e. $h/\min_{1 \leq i \leq N} h_i$ is bounded, and h is small enough. If, for example, the solution behaves badly in one part of the domain and well in another, (2.11) indicates that h should still be taken small in the region of good behavior in order to keep the $O(h^{k+m_n+1-l})$ term relatively small. However, our experience has been that the mesh selection and error estimation schemes usually work well, supporting our above assumption.

A Posteriori Error Estimate. Suppose we have approximations $\mathbf{v}(\cdot)$ and $\mathbf{v}^*(\cdot)$ on the meshes $\{x_i\}_{i=1}^{N+1}$ and $\{x_i^*\}_{i=1}^{2N+1}$ respectively, with $x_{2i-1}^* = x_i$ and $x_{2i}^* = x_{i+1/2} = \frac{1}{2}(x_i + x_{i+1})$. We want to estimate the maximum of the error $e_n^*(x) = u_n(x) - v_n^*(x)$ for $x \in [x_i, x_{i+1}]$. If $k \geq m_d$, v_n and v_n^* can be compared at several points to estimate [32]

$$\|v_n - v_n^*\| \frac{1}{2^{k+m_n} + 1} + O(h^{k+m_n+1})$$

$$\leq \|e_n^*\| \leq \|v_n - v_n^*\| \frac{1}{2^{k+m_n} - 1} + O(h^{k+m_n+1}).$$

However, if $k > m_d$, we use the structure indicated by (2.11) as follows: Consider the points $x^*_{2i-2/3} = x_{i+1/6}$ and $x^*_{2i-1/3} = x_{i+1/3}$ (see Figure 2.1).

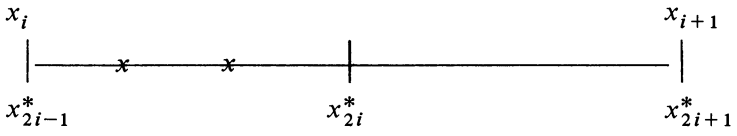

FIGURE 2.1

Let

$$\Delta_1 = |v_n(x_{i+1/6}) - v_n^*(x_{i+1/6})| = |e_n(x_{i+1/6}) - e_n^*(x_{i+1/6})|$$

$$= \frac{|u_n^{(k+m_n)}(x_i)|}{2^{k+m_n}} \left| P_n(-2/3) - \frac{1}{2^{k+m_n}} P_n(-1/3) \right| h_i^{k+m_n} + O(h^{k+m_n+1}), \tag{2.13}$$

and similarly

$$\Delta_2 = |v_n(x_{i+1/3}) - v_n^*(x_{i+1/3})|$$

$$= \frac{|u_n^{(k+m_n)}(x_i)|}{2^{k+m_n}} \left| P_n(-1/3) - \frac{1}{2^{k+m_n}} P_n(1/3) \right| h_i^{k+m_n} + O(h^{k+m_n+1}), \tag{2.14}$$

where P_n is defined in (2.12). From (2.11),

$$\max_{x \in [x^*_{2i-1}, x^*_{2i}]} |e_n^*(x)|$$

$$= \frac{\|P_n\|(\Delta_1 + \Delta_2)}{|2^{k+m_n} P_n(-2/3) - P_n(-1/3)| + |2^{k+m_n} P_n(-1/3) - P_n(1/3)|} \tag{2.15}$$

$$+ O(h^{k+m_n+1}).$$

When (2.15) is generalized to provide estimates of errors in all the components of $\mathbf{z}(\mathbf{v})$, then the weights multiplying $(\Delta_1 + \Delta_2)$ are given by

$$\omega_{k,\nu} = \frac{\|P^{(\nu)}\|}{|2^{2k-\nu} P^{(\nu)}(-2/3) - P^{(\nu)}(-1/3)| + |2^{2k-\nu} P^{(\nu)}(-1/3) - P^{(\nu)}(1/3)|}, \tag{2.16}$$

$$\nu = 0, \ldots, k-1,$$

with $P(\xi) \equiv P(k, \xi) = (\xi^2 - 1)^k/(2k)!$.

These weights are precomputed and stored as constant data in the program, and the error is then estimated by ignoring the $O(h^{k+m_n-l+1})$ term in

$$\max_{x \in [x^*_{2i-1}, x^*_{2i}]} |e_n^{*(l)}(x)| = \omega_{k,k-m_n+l}(\Delta_1 + \Delta_2) + O(h^{k+m_n-l+1}), \tag{2.17}$$

$$l = 0, \ldots, m_n - 1,$$

where Δ_1 and Δ_2 are taken for $v_n^{(l)}$, $n = 1, \ldots, d$.

Mesh Selection. The results below are a generalization of [5], [18], [12]. Given a set of tolerances TOL_j, $j = 1, \ldots, NTOL$, with a set of pointers $LTOL_j$, $j = 1, \ldots, NTOL$, COLSYS attempts to satisfy

$$\|z_l(\mathbf{u}) - z_l(\mathbf{v})\| \leqslant TOL_j, \qquad l = LTOL_j,\ j = 1, \ldots, NTOL. \tag{2.18}$$

The aim of the mesh selection algorithm is to meet the above requirements with the least number of mesh points.

As before we neglect the global term in (2.11) and write

$$\max_{x \in [x_i, x_{i+1}]} |e_n^{(l)}(x)| \doteq C_{k,k-m_n+l} |u_n^{(k+m_n)}(x_i)| h_i^{k+m_n-l}, \qquad l = 0, \ldots, m_n - 1, \tag{2.19}$$

where

$$C_{k,\nu} = \|P^{(\nu)}(k; \cdot)\|/2^{2k-\nu}, \qquad \nu = 0, 1, \ldots, k-1. \tag{2.20}$$

For each j ($1 \leqslant j \leqslant NTOL$), let $l = LTOL_j$, let $n = JTOL_j$ indicate the component of $\mathbf{u}$ that $z_l(\mathbf{u})$ is a derivative of, let WEIGHT_j be the appropriate $C_{k,\nu}$ divided by TOL_j, and let ROOT_j be the inverse of the expected rate of convergence of $z_l(\mathbf{v})$.

From (2.18)–(2.20), the goal is to pick a mesh $\{x_i^*\}_{i=1}^{N^*+1}$ for which

$$\max_{1 \leqslant j \leqslant NTOL} \mathrm{WEIGHT}_j \cdot |u_n^{(k+m_n)}(x_i^*)| h_i^{*1/\mathrm{ROOT}_j} \leqslant 1, \quad (n = JTOL_j); \qquad i = 1, \ldots, N^*, \tag{2.21}$$

for the smallest N^* possible. Actually finding this mesh is impossible since the $u_n^{(k+m_n)}(x_i^*)$ are unknown. Moreover, for COLSYS the final mesh is a halving of the one before last (so that an error estimate is at hand). If

$$S_j(x) = \mathrm{WEIGHT}_j |u_n^{(k+m_n)}(x)|, \tag{2.22}$$

and

$$S(x) = \max_{1 \leqslant j \leqslant NTOL} S_j^{\mathrm{ROOT}_j}(x), \tag{2.23}$$

then (2.21) is equivalent to

$$S(x_i^*) h_i^* \leqslant 1, \qquad i = 1, \ldots, N^*. \tag{2.24}$$

A collocation solution $\mathbf{v}$ on a mesh satisfying (2.24) would satisfy

$$\|z_l(\mathbf{u}) - z_l(\mathbf{v})\| \leqslant TOL_j(1 + O(h)), \qquad l = LTOL_j,\ j = 1, \ldots, NTOL, \tag{2.25}$$

the $O(h)$ term arising from neglecting higher order terms in (2.19). By requiring that

$$\int_{x_i^*}^{x_{i+1}^*} S(x)\, dx = 1 \tag{2.26}$$

instead of (2.24), (2.25) still holds [5]. To approximately satisfy (2.26), we still need to approximate $u_n^{(k+m_n)}(x)$, $n = 1, \ldots, d$. Given a mesh $\{x_i\}_{i=1}^{N+1}$ and an approximate collocation solution v, an accurate approximation for the higher order derivatives can be constructed as follows (cf. [12]): The polynomial in the error expression for the $(k + m_n - 1)$st derivative of the nth component is

$$\frac{1}{(2k)!}\frac{d^{2k-1}}{d\xi^{2k-1}}(\xi^2 - 1)^k = \xi.$$

Therefore, $e_n^{(k+m_n-1)}(x_{i+1/2}) = O(h^2)$ and

$$\begin{aligned}\hat{u}_n(x_{i+1}) &:= \frac{2|v_n^{(k+m_n-1)}(x_{i+1}) - v_n^{(k+m_n-1)}(x_i)|}{x_{i+2} - x_i} \\ &= |u_n^{(k+m_n)}(x_{i+1})| + O(h) = |u_n^{(k+m_n)}(x)| + O(h) \\ &\qquad \text{for } x \in [x_i, x_{i+2}],\ i = 1, \ldots, N-1.\end{aligned} \tag{2.27}$$

Define $\hat{u}_n(x)$ over the whole interval $[a, b]$ by

$$\hat{u}_n(x) = \begin{cases} \hat{u}_n(x_i), & x \in [x_i, x_{i+1}],\ i = 2, \ldots, N, \\ \hat{u}_n(x_2), & x \in [x_1, x_2], \end{cases} \tag{2.28}$$

so that $|u_n^{(k+m_n)}(x)| = |\hat{u}_n(x)| + O(h)$. Then

$$\hat{s}(x) := \max_{1 \le j \le NTOL} [\text{WEIGHT}_j \cdot \hat{u}_n(x)]^{\text{ROOT}_j}, \qquad n = JTOL_j, \tag{2.29}$$

is a piecewise constant computable function, and (2.25) is satisfied for $\{x_i^*\}_{i=1}^{N^*+1}$ by requiring

$$\int_{x_i^*}^{x_{i+1}^*} \hat{s}(x)\,dx = 1, \qquad i = 1, \ldots, N^*. \tag{2.30}$$

In practice (2.30) may lead to a very large N^*, compared to N, which could mean that N^* has been determined by premature data. Also, an error estimate is needed at the end to check whether the tolerances have been satisfied. So, we modify the criterion (2.30) to allow for these considerations by picking a new mesh (for some N^*), according to

$$\int_{x_i^*}^{x_{i+1}^*} \hat{s}(x)\,dx = \gamma \equiv \frac{1}{N^*}\int_a^b \hat{s}(x)\,dx = \frac{1}{N^*}\sum_{j=1}^{N} \hat{s}(x_j)h_j, \qquad i = 1, \ldots, N^*. \tag{2.31}$$

There are still two questions to be answered: When to redistribute the points, as opposed to just halving the current mesh, and how to choose N^*. When an approximate solution on the current mesh $\{x_i\}_{i=1}^{N+1}$ has been obtained, the diagnostics $r_1 = \max_i \hat{s}(x_i)h_i$, $r_2 = \Sigma_{i=1}^{N} \hat{s}(x_i)h_i$, and $r_3 = r_2/N$ are computed. The ratio r_1/r_3 gives

some idea of the gain to be achieved by redistribution. Specifically, the code feels it can reduce the error by as much in redistributing with $N^* = N$ as by taking $N^* = (r_1/r_3) \cdot N$ with the current distribution. Our present policy is to redistribute only when $r_1 \geqslant 2r_3$.

When redistributing, $r_2 = \gamma N^*$ predicts the number of points needed to satisfy the tolerances. If r_2 is much larger or much smaller than N, then we do not put much faith in this prediction. The current policy is to take

$$N^* = \min\{\tfrac{1}{2}\bar{N},\ N,\ \tfrac{1}{2}\max[N,\ r_2]\}, \tag{2.32}$$

where $\bar{N}$ is the maximum number of subintervals allowed by the storage specifications. This allows for changes up to a factor of 2 in N and for later halving of the mesh in order to obtain an error estimate. Also, restrictions are placed on the number of times a mesh can be redistributed before halving.

3. *B*-Spline Evaluation. For reasons of efficiency, stability, and flexibility in order and continuity, B-splines are chosen as the basis functions. Efficient algorithms for calculating with B-splines are given by deBoor [4], who implements these algorithms in a Fortran package [6]. Evaluation of the basis functions is a major cost for finite element methods, and careful implementation of the selected algorithms is necessary for the code to be competitive. Our use of B-splines is somewhat special because (i) we are solving a system of differential equations, so many repetitive calculations can be avoided, (ii) the continuity in the solution at the mesh points is more restricted here than in [6], allowing us to trade unneeded generality for an increase in speed, and (iii) on many occasions we evaluate the B-splines at points which are placed in a regular fashion in each subinterval. We take advantage of these special features in implementing restricted versions of de Boor's algorithms.

As we only outline the modifications to these algorithms, the interested reader is referred to [3] for the complete details.

A. *Evaluation of the B-Splines and the Solution.* Recall that $v_n(x) \in \mathcal{P}_{k+m_n,\pi} \cap C^{(m_n-1)}[a, b]$ $(1 \leqslant n \leqslant d)$ for a given mesh π: $a = x_1 < x_2 < \cdots < x_{N+1} = b$. If $N_{j,k}$ is the jth B-spline of order k [4], then

$$v_n(x) = \sum_{j=-k-m_n+2}^{Nk} \alpha_{j,n} N_{j,k+m_n}(x). \tag{3.1}$$

Defining the knot sequence

$$t_j = \begin{cases} x_1, & j \leqslant k + m_d, \\ x_{i+1}, & ik + m_d < j \leqslant (i+1)k + m_d,\ (1 \leqslant i \leqslant N-1), \\ x_{N+1}, & Nk + m_d < j \leqslant (N+1)k + 2m_d, \end{cases} \tag{3.2}$$

then only $k + m_n$ B-splines may be nonzero at $x \in [t_i, t_{i+1})$, viz.

$$v_n(x) = \sum_{j=-k-m_n+1}^{0} \alpha_{i+j,n} N_{i+j,k+m_n}(x). \tag{3.3}$$

The algorithm in [4] for the evaluation of these B-splines is

Algorithm I. Let $N_{i,1}(x) \equiv 1$.
Do for $l = 1, \ldots, k + m_d - 1$:
  $N_{i-l,l+1}(x) = 0$
  Do for $j = 1, \ldots, l$:
    $M_{i+j-l,l}(x) = N_{i+j-l,l}(x)/(t_{i+j} - t_{i+j-l})$,
    $N_{i+j-l-1,l+1}(x) = N_{i+j-l-1,l+1}(x) + (t_{i+j} - x)M_{i+j-l,l}(x)$,
    $N_{i+j-l,l+1}(x) = (x - t_{i+j-l})M_{i+j-l,l}(x)$.

From the recursive manner in which the B-splines are defined it is clear that Algorithm I need only be performed once for a given x to produce the B-splines needed to evaluate all components of $\mathbf{v}(x)$ by (3.3). Also, since the structure of the knot sequence is known in terms of the mesh π, there is no need to generate the t_j's. If $x \in [x_I, x_{I+1})$, we can make the changes in Algorithm I according to

$$t_{i+j} - t_{i+j-l} = \begin{cases} h_{I-1} + h_I & \text{for } j \leqslant k,\ j + k \leqslant l, \\ h_I & \text{for } j \leqslant k,\ l \leqslant j + k - 1, \\ h_I + h_{I+1} & \text{for } k + 1 \leqslant j, \end{cases} \tag{3.4}$$

and

$$t_{i+j} - x = \begin{cases} \rho h_I & \text{for } 1 \leqslant j \leqslant k, \\ \rho h_I + h_{I+1} & \text{for } k + 1 \leqslant j \leqslant k + m_d - 1\ (\leqslant 2k), \end{cases} \tag{3.5}$$

where ρ is chosen appropriately. The substitutions (3.4) and (3.5) have led to an algorithm about 50% faster than the general one [6] (when running on an IBM 370/168).

Some of the B-spline values at x depend only on their relative position in $[x_I, x_{I+1})$ and not on the subinterval itself. For example, the collocation points are located at the same relative positions in all subintervals, so it is only necessary to evaluate these mesh independent splines once for each relative position. The points at which the approximate solution is evaluated for the error estimate (2.17) are another instance where this saving may be made. Since $\frac{1}{2}(k + m_d)(k + m_d - 1)$ B-splines are needed for any x and only $m_d(m_d - 1)$ are subinterval dependent, a saving of at least 50% is obtained for $k \geqslant m_d$.

We have used two routines in the implementation of the modified version of Algorithm I. The first evaluates those B-splines which are mesh independent, while the second is for the splines whose values depend on I (where $x \in [x_I, x_{I+1})$).

We do not exploit the symmetry of the collocation points or the error estimation points; the saving is too small given the additional complexity. Also, we do not incorporate the nonconvex modification suggested in [30]. While it can save a multiplication in the last line of Algorithm I and our experiments have not yielded a case where accuracy was significantly affected, the improvement in efficiency proved small enough that we have decided to be conservative.

B. *Evaluation of Spline Derivatives.* Given an approximate solution component $v_n(x)$, as in (3.3), its derivatives are given by

$$(3.6)\quad v_n^{(r)}(x) = (k + m_n - 1) \cdots (k + m_n - r) \sum_{j=-k-m_n+r+1}^{0} \alpha_{i+j,n}^{(r)} N_{i+j,k+m_n-r}(x),$$

where

$$(3.7)\quad \alpha_{i+j,n}^{(r)} = \begin{cases} \alpha_{i+j,n} & \text{for } r = 0, \\ \dfrac{\alpha_{i+j,n}^{(r-1)} - \alpha_{i+j-1,n}^{(r-1)}}{t_{i+j+k+m_n-r} - t_{i+j}} & \text{for } r > 0. \end{cases}$$

The B-spline package in [6] contains a subroutine which prepares the divided difference table (3.7) (with $\alpha_{i+j,n}^{(r)}$ multiplied by $(k + m_n - 1) \cdots (k + m_n - r)$). We have written a similar routine which uses the particular form of $(t_{i+j+k+m_n-r} - t_{i+j})$.

To compute $\mathbf{z}(\mathbf{v}) = (v_1, v_1', \ldots, v_1^{(m_1-1)}, v_2, \ldots, v_d, \ldots, v_d^{(m_d-1)})$ we only need $v_n^{(r)}(x)$, $r = 0, \ldots, m_n - 1$, $n = 1, \ldots, d$. There are several occasions where evaluation of $\mathbf{z}(\mathbf{v})$ is necessary. Values of $\mathbf{z}(\mathbf{v})$ are needed for setting up the equations during the iterations on nonlinear problems and for the error estimation procedure. Also, when COLSYS has terminated successfully the user can evaluate $\mathbf{z}(\mathbf{v})$ for the final approximation. Two efficient ways to evaluate $\mathbf{z}(\mathbf{v})$ are:

Algorithm II. (a) Generate $\alpha_{i,n}^{(r)}$, $i = 1, \ldots, Nk + m_n$; $r = 1, \ldots, m_n - 1$; $n = 1, \ldots, d$,

(b) for $x \in (x_I, x_{I+1})$, form the $\frac{1}{2}(k + m_d)(k + m_d + 1)$ nonzero B-splines up to order $k + m_d$,

(c) form $v_n^{(r)}(x)$ $(r = 0, \ldots, m_n - 1; n = 1, \ldots, d)$ by (3.6).

Algorithm II′. (a) Generate $v_n^{(r)}(x_i)$, $r = 1, \ldots, k + m_n - 1$; $n = 1, \ldots, d$; $i = 1, \ldots, N$, by Algorithm II,

(b) for $x \in [x_I, x_{I+1})$,

$$(3.8)\quad v_n^{(r)}(x) = \sum_{j=r}^{k+m_n-1} \frac{v_n^{(j)}(x_I)}{(j-r)!}(x - x_I)^{j-r}.$$

While Algorithm II′ requires more than twice the storage and more initialization than Algorithm II, it is many times more efficient when $\mathbf{z}(\mathbf{v})$ is required for a large number of points. For the collocation example in [6], Algorithm II′ was used.

In [3] these algorithms are examined in our setting for two cases–when $\mathbf{z}(\mathbf{v})$ is to be evaluated at

(i) M_1 points irregularly distributed in $[a, b]$ and (ii) $M_1 = M_2 N$ points, consisting of M_2 regularly distributed points in each subinterval. The numbers of multiplications plus divisions required for Algorithm II are approximately

(i) $(m^* - d)(k + 2)N + [(k + m_d)(k + m_d - 1) + \bar{M}]M_1$,

(ii) $\{(m^* - d)(k + 2) + [2(m_d^2 - m_d) + \bar{M}]M_2\}N$,

and for Algorithm II′,

(i) and (ii)

$$\left[(m^* - d)(k+2) + kd(k+2) + 2(m_d^2 - m_d) + \bar{M} + \frac{d}{2}k(k+1)\right]N + [\bar{M} + 2(k + m_d - 2)]M_1,$$

where $\bar{M} = (k + \tfrac{1}{2})m^* + \tfrac{1}{2}\Sigma_{n=1}^{d} m_n^2$.

For case (i) Algorithm II is more efficient when $M_1 \leqslant \lambda N$ where, e.g., $\lambda \doteq 2.5$ if $d = 1$, $m = 2$, $k = 3$ and $\lambda \doteq 4.8$ if $d = 3$, $m_1 = 1$, $m_2 = m_3 = 3$, $k = 4$. In general, λ grows with d. In case (ii) Algorithm II is more efficient for most practical situations. Consequently, we use only Algorithm II.

C. *Derivatives of the B-Splines.* In order to generate the collocation equations an algorithm is needed to evaluate the B-spline derivatives. Formulas (3.6) and (3.7) could be used with $\alpha_{i+j,n}^{(0)} = \delta_{jl}$ for the function $N_{i+l,k+m_n}(x)$, but a number of savings can be made. First, if $m_n = m_{n+1}$, there is no need to repeat the computations, so COLSYS initially isolates the set of strictly increasing orders and deals only with them. Second, the algorithm avoids performing (3.7) on the many zero coefficients (as is also done in [6]). Third, the special form of $(t_{i+j+k+m_n-r} - t_{i+j})$ is used. The forth improvement arises from the fact that we have a system of differential equations. If the B-spline derivatives are evaluated for $n = d$ then a number of the $\alpha_{i+j,n}^{(r)}$ may be determined directly from

$$\alpha_{i+j,n}^{(r)} = \begin{cases} \alpha_{i+j+(m_d - m_n),d}^{(r)}, & 1 \leqslant j \leqslant k - (m_d - m_n) - r + 1, \\ \alpha_{i+j,d}^{(r)}, & m_d \leqslant j \leqslant k + m_n - r. \end{cases} \tag{3.9}$$

D. *Highest Order Derivatives.* Selecting a new mesh requires the values of the piecewise constant functions $v_n^{(k+m_n-1)}(x)$, $1 \leqslant n \leqslant d$. These are obtained by starting with the values $\alpha_{i+j,n}^{(m_n-1)}$ $(-k \leqslant j \leqslant 0)$, which have been obtained in Algorithm II, and repeatedly applying (3.7) with $t_{i+j+k+m_n-r} - t_{i+j} = h_I$ to get $\alpha_{i,n}^{(k+m_n-1)} = v_n^{(k+m_n-1)}(x)$ for $x \in [t_i, t_{i+1}) = [x_I, x_{I+1})$.

4. The Nonlinear Iteration and the Linear System Solver. In this section we briefly discuss the handling of nonlinear problems and the implementation of the linear system solver.

A. *Newton Iteration.* To solve (2.1), (2.2) we apply the Newton process of linearization and iteration. Specifically, choose an initial approximation $\mathbf{v}^{,0} \in \mathcal{P}_{k+\mathbf{m}} \cap C^{(\mathbf{m}-1)}[a, b]$. Then, for $s = 0, 1, 2, \ldots$ until a convergence criterion is satisfied, solve by collocation the problem

$$L_n(\mathbf{v}^{,s})\mathbf{w} = f_n, \qquad n = 1, \ldots, d, \tag{4.1}$$

$$\beta_j(\mathbf{v}^{,s})\mathbf{w} = \gamma_j, \qquad j = 1, \ldots, m^*, \tag{4.2}$$

for the solution $\mathbf{v}^{,s+1}$. Here L_n, β_j are defined in (2.3a), (2.4a) and

$$(4.3)\qquad f_n \equiv f_n(\cdot; \mathbf{v}'^s) = F_n(\cdot; \mathbf{v}'^s) - \sum_{l=1}^{m^*} \frac{\partial F_n(\cdot; \mathbf{v}'^s)}{\partial z_l} \cdot z_l(\mathbf{v}'^s), \qquad n = 1, \ldots, d,$$

$$(4.4)\qquad \gamma_j \equiv \gamma_j(\mathbf{v}'^s) = \sum_{l=1}^{m^*} \frac{g_j(\zeta_j; \mathbf{v}'^s)}{\partial z_l} \cdot z_l(\mathbf{v}'^s(\zeta_j)) - g_j(\zeta_j; \mathbf{v}'^s), \qquad j = 1, \ldots, m^*.$$

Most advantages and disadvantages of the Newton method are well known. Generally, if the initial approximation $\mathbf{v}'^0$ is close enough to $\mathbf{v}$, the method performs very satisfactorily. However, when $\mathbf{v}'^0$ is far from $\mathbf{v}$, the behavior of the algorithm is unpredictable (cf. [9], [16]). We are currently conducting an investigation to find more reliable fast algorithms to handle nonlinearities, and intend to report the results elsewhere.

Implementing the Newton iteration requires determining when the desired error tolerances (2.18) are satisfied. For a nonlinear problem, the error has two components, $\mathbf{v}'^{s+1} - \mathbf{v}$ and $\mathbf{v} - \mathbf{u}$, where $\mathbf{v}'^{s+1}$ is the Newton iterate which satisfies the convergence criterion to be specified and is thus taken as the approximation to $\mathbf{v} = \lim_{s\to\infty} \mathbf{v}'^s$. For any superlinearly convergent method,

$$\lim \frac{\|\mathbf{v}'^{s+1} - \mathbf{v}'^s\|_\infty}{\|\mathbf{v}'^s - \mathbf{v}\|_\infty} = 1$$

(see [15]), so that in the limit $\mathbf{v}'^{s+1} - \mathbf{v}'^s$ is a good measure for $\mathbf{v}'^s - \mathbf{v}$. Thus, the convergence criterion for the nonlinear iteration in (4.1), (4.2) is

$$(4.5)\qquad \|z_l(\mathbf{v}'^{s+1}) - z_l(\mathbf{v}'^s)\|_\infty \leqslant TOL_j, \qquad l = LTOL_j,\ j = 1, \ldots, NTOL.$$

To check (4.5) efficiently, recall that

$$(4.6)\qquad v_n^{(r)}(x) = \sum_i \alpha_{i,n}^{(r)} N_{i,k+m_n}(x), \qquad r = 0, \ldots, m_n - 1,\ n = 1, \ldots, d,$$

where the $\alpha_{i,n}^{(r)}$ in (3.6) are modified. The $N_{i,j}$ are normalized B-splines, so

$$(4.7)\qquad \|v_n^{(r)}\|_\infty \leqslant \max_i |\alpha_{i,n}^{(r)}|.$$

The $\alpha_{i,n} = \alpha_{i,n}'^s$ in (4.6) are precisely the coefficients computed when solving the linear system in each Newton iteration. Thus, the nonlinear iteration convergence criterion is as follows:

1. Having obtained $\alpha_{i,n}'^{s+1}$, compute $\alpha_{i,n}^{(r),s+1}$ for all i, $r = 0, \ldots, m_n - 1$, $n = 1, \ldots, d$.

2. For $j = 1, \ldots, NTOL$, let $l = LTOL_j$; and let (n, r) correspond to the coordinate l of $\mathbf{z}(\cdot)$.

If $\|\alpha_{(\cdot),n}^{(r),s+1} - \alpha_{(\cdot),n}^{(r),s}\|_\infty > TOL_j$, then go to step 4.

3. Dump $\alpha^{(\cdot),s+1}$ onto $\alpha^{(\cdot),s}$, signal success, and exit the Newton iteration.

4. Dump $\alpha^{(\cdot),s+1}$ onto $\alpha^{(\cdot),s}$, set $s = s + 1$, and reiterate.

In fact, steps 1 and 2 above are combined so that only the array $\alpha_{(\cdot),(\cdot)}^{(\cdot),s}$ is stored. Since the computation in step 1 is always needed to evaluate the approximate solution (see Section 3 and [2]), it is not wasteful. Finally note that the criterion is somewhat pessimistic and is scaling-resistant [14].

B. *The Linear System Solver.* Here we consider the method for the solution of the set of algebraic equations resulting from collocation applied to (4.1), (4.2). With $x_{i1}, \ldots, x_{ik}$ the k Gaussian points in the ith subinterval $I_i = (x_i, x_{i+1})$, $1 \leqslant i \leqslant N$ (cf. (2.5), (2.6)), write these equations as

$$L_n \mathbf{v}(x_{ij}) = f_n(x_{ij}), \qquad j = 1, \ldots, k,\ i = 1, \ldots, N,\ n = 1, \ldots, d, \tag{4.8}$$

$$\beta_j \mathbf{v} = \gamma_j, \qquad j = 1, \ldots, m^*. \tag{4.9}$$

The total number of equations in (4.8), (4.9) is $Nkd + m^*$, the dimension of the approximation space (or the number of parameters $\alpha_{i,n}$ to be determined).

Consider next the structure of the matrix obtained from (4.8), (4.9). Fixing i and n, $1 \leqslant i \leqslant N$, $1 \leqslant n \leqslant d$, there are m_n nonzero B-splines on $I_{i-1} \cup I_i$, $k - m_n$ B-splines which vanish outside I_i and m_n which vanish outside $I_i \cup I_{i+1}$. Giving $\alpha = (\alpha_{in})$ the natural ordering $\alpha_{11}, \alpha_{21}, \ldots, \alpha_{Nk+m_1,1}, \alpha_{12}, \ldots, \alpha_{Nk+m_d,d}$ causes an inconvenient zero structure, since it is desirable to have all nonzero elements concentrated around the main diagonal. Thus, we reorder the coefficient vector α in such a way that, for each i, all the columns in the matrix which contain nonzero entries corresponding to the ith subinterval are adjacent. This produces a block-structured matrix whose ith block, $1 \leqslant i \leqslant N$, is characterized as follows:

Rows: With l_i side conditions given at points ζ_l, $x_i \leqslant \zeta_l < x_{i+1}$ (when $i = N$, $x_N \leqslant \zeta_l \leqslant x_{N+1}$), there are $kd + l_i$ corresponding rows. (For each n, $1 \leqslant n \leqslant d$, k rows correspond to $x_{i1}, \ldots, x_{ik}$.) The numbering of the rows increases with the argument x.

Columns: For each $v_n(\cdot)$, $1 \leqslant n \leqslant d$, there are m_n B-splines which do not vanish on $I_i \cup I_{i-1}$. The corresponding columns, m_n for each n, will appear first in the order $m_1, m_2, \ldots, m_d$, totalling m^* columns. Then come $k - m_1, k - m_2, \ldots, k - m_d$ columns corresponding to the $kd - m^*$ B-splines which vanish outside I_i. The m^* columns of those B-splines which vanish outside $I_i \cup I_{i+1}$ appear last, ordered the same way as the first m^* columns. The total number of columns is, therefore, $kd + m^*$ (note that $m^* = \Sigma_{i=1}^{N} l_i$ and $kd + m^* \geqslant kd + l_i$).

Initial Coordinates: The upper left element of the ith block is the (i_1, i_2)th element of the matrix, where

$$i_1 = (i-1)kd + \sum_{j=1}^{i-1} l_j + 1, \quad i_2 = (i-1)kd + 1.$$

As an example, take $d = 2$, $m_1 = 1$, $m_2 = 2$, $k = 3$, $\zeta_1 = \zeta_2 = a$, $\zeta_3 = b$, and $N = 3$. Then the reordered collocation matrix has the form

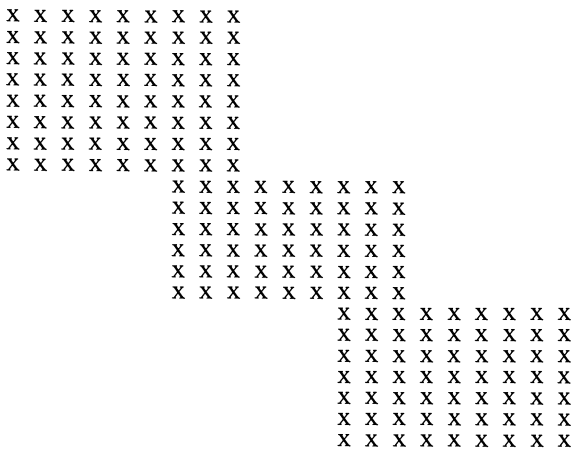

FIGURE 4.1

This is precisely the zero structure for the collocation method with $k \cdot d = 6$ points per subinterval applied to a problem of one differential equation of order $m^* = 3$. The matrix can be considered as banded (asymmetric), but this would almost double the amount of nonzero entries. It is better considered as almost block diagonal [8].

For the solution of the linear systems we have adopted the code developed in [8] which performs Gauss elimination with scaled row pivoting. This proceeds as follows:

For $i = 1, 2, \ldots, N$ do the following:

1. If $i > 1$, append the $\Sigma_{j=1}^{i-1} l_j$ rows of the $(i-1)$st block, not used as pivotal rows, to the beginning of the ith block, to form a block of $kd + \Sigma_{j=1}^{i} l_j$ rows.

2. Apply kd steps of Gauss elimination with scaled row pivoting, storing the resulting factorization in place of the original data.

3. If $i = N$ (the last block is square of size $kd + m^*$), apply $m^* - 1$ more elimination steps.

This produces an LU factorization of the original matrix. For a given right-hand side, the solution α is then obtained by a forward-backward substitution.

The permuted ordering in the solution vector $\alpha = (\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_{Ndk+m^*})^T$ is as follows: For $x \in [x_i, x_{i+1})$ and $1 \leqslant n \leqslant d$, let $\mu_n = \Sigma_{l=1}^{n-1} m_l$, and $\eta_n = (n-1)k - \mu_n$. Then

$$
\begin{aligned}
v_n(x) = &\sum_{l=1}^{m_n} \alpha_{(i-1)kd+\mu_n+l} N_{(i-1)kd+\mu_n+l,k+m_n}(x) \\
&+ \sum_{l=1}^{k-m_n} \alpha_{(i-1)kd+\eta_n+m^*+l} N_{(i-1)kd+\eta_n+m^*+l,k+m_n}(x) \\
&+ \sum_{l=1}^{m_n} \alpha_{ikd+\mu_n+l} N_{ikd+\mu_n+l,k+m_n}(x).
\end{aligned} \tag{4.14}
$$

The package [8] implementing the above method is used in COLSYS because of

its availability, stability, and its advantages over treating the system as merely banded. There is, however, some amount of fill-in generated. The ith block has $\Sigma_{j=1}^{i-1} l_i$ rows appended to it, and so the ratio of storage from fill-in to total storage is

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{j=1}^{i-1} l_j}{kd + l_i}.$$

For the example in Figure 4.1 this ratio is about 1/5. If N is large and we consider a two-point boundary value system with half of the boundary conditions at each end, the ratio tends to $\frac{1}{2}m^*/kd$. This is always less than ½ since $k \geqslant m_d \geqslant m_{d-1} \geqslant \cdots \geqslant m_1$, but the value ½ is obtained when $k = m_d = m_1$.

A method which generates no fill-in has been proposed in [37]. Here row and column pivoting are performed alternately. A comparison between the above two methods, regarding their efficiency and stability, is planned for the future.

5. Numerical Examples. COLSYS has been tested on a large variety of problems, and a representative selection of them to demonstrate the performance of the code is available in [2]. Some comparisons with the codes in [35], [25] are also made in [2] in order to gain a relative perspective. For brevity, we examine only three examples here.

The examples were run in double precision (14 hexadecimal digits) on the IBM 370/155 at Simon Fraser University, using the Fortran G1 compiler. Because of fluctuations in the computing environment, variations of 5–10% in the run times are meaningless.

In the examples below, the following notation is used:

$u_i(x)$–ith component of the exact solution.

$E(u_i^{(j)})$–uniform error in $u_i^{(j)}(x)$ (available when the exact solution is known).

est $E(u_i^{(j)})$–estimated uniform error in $u_i^{(j)}(x)$.

$TOL(u_i^{(j)})$–absolute error tolerance for the component $u_i^{(j)}(x)$. (COLSYS allows the user to specify different tolerances for different components, and the mesh selection algorithm considers only those components for which tolerances are specified.)

time–the actual solution time in seconds (not including error checking time).

$a \pm b$–$a \cdot 10^{\pm b}$.

k–number of collocation points per subinterval.

mesh sequence (iterations)–successive mesh sizes, i.e. numbers N of subintervals required, followed in parentheses by the number of Newton iterations performed on each mesh for nonlinear problems.

The errors $E(u_i^{(j)})$ are approximated by measuring the error at 4 equally spaced points in each subinterval. Unless otherwise stated, the initial mesh for COLSYS is uniform.

Example 1 [22].

$$\epsilon y'' + xy' = -\epsilon\pi^2 \cos(\pi x) - (\pi x)\sin(\pi x), \qquad -1 \leqslant x \leqslant 1,$$
$$y(-1) = -2, \quad y(1) = 0,$$
$$u(x) = \cos(\pi x) + \operatorname{erf}(x/\sqrt{2\epsilon})/\operatorname{erf}(1/\sqrt{2\epsilon}).$$

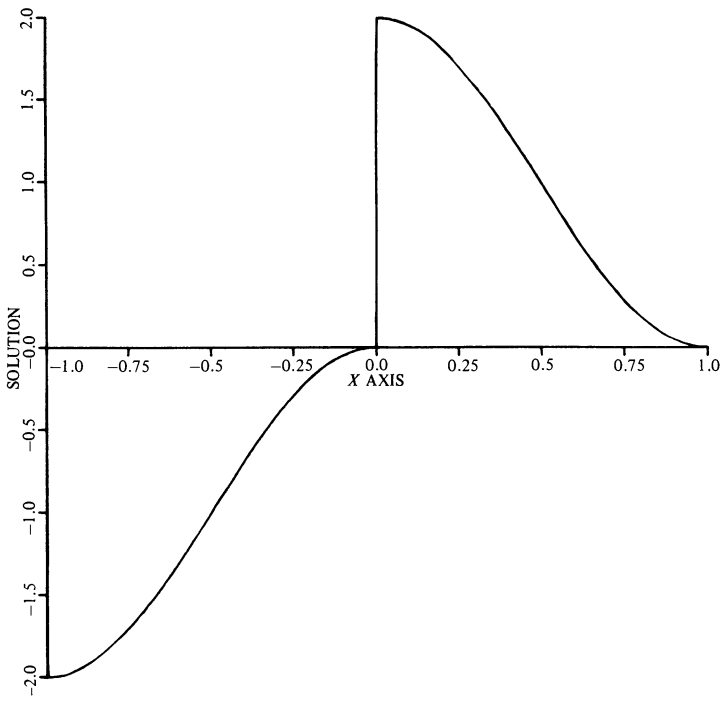

FIGURE 5.1. ($\epsilon = 10^{-6}$)

The solution has a spike at $x = 0$ (see Figure 5.1). Results for various values of ϵ are given in Table 1 below.

TABLE 1

ϵ	k	$TOL(u)$	$E(u)$	est $E(u)$	$TOL(u')$	$E(u')$	est $E(u')$	*time*	*mesh sequence*
.1−1	4	.1−1	.45−4	.19−4	.1−1	.22−2	.12−2	1.1	8, 16
.1−1	4	.1−5	.22−8	.18−8	.1−5	.24−6	.15−6	8.16	8, 16, 15, 30, 17, 34, 68
.1−3	4	.1−5	.16−7	.19−8	.1−5	.67−5	.25−6	25.9	8, 16, 32, 64, 35, 70, 35, 70, 35, 70, 140
.1−5	4	.1−5	.48−9	.43−9	.1−5	.23−6	.45−7	68.23	8, 16, 32, 64, 128, 128, 128, 128, 256, 128, 256, 128, 256

As ϵ gets smaller, the problem gets tougher, and the mesh selection algorithm has to use more mesh points in order to resolve the difficulty. In fact, if we choose a good initial mesh (concentrated around $x = 0$) which contains only a few points, the mesh selection algorithm often produces worse meshes at first, because the error is very much different from the assumed asymptotic form. In order to allow use of knowledge about where the region of fast variation is located (as many special purpose methods do) COLSYS has an option for choosing an initial mesh and repeatedly halving it until the tolerances are satisfied. Doing so for $\epsilon = 10^{-10}$, $TOL(u) = 10^{-7}$, $TOL(u') = 10^{-2}$, and the initial mesh $-1, -.1, -.01, -.001, -.0001, -.00001, 0, .00001, .0001, .001, .01, .1, 1$ has resulted in a final mesh of 384 subintervals, with

$E(u) = .30\text{–}8$, est $E(u) = .30\text{–}8$, $E(u') = .61\text{–}2$, est $E(u') = .98\text{–}2$. A large amount of storage, however, is necessary, and special methods such as expansion techniques [21], [22] for singular perturbation problems with very small ϵ will obviously be superior in many situations.

Example 2 [20]. Perhaps the greatest relative advantages of COLSYS is in solving problems in which the coefficients in the differential equations may contain singularities. These problems commonly arise when reducing partial to ordinary differential equations by physical symmetry [34], [20], [28]. Unlike other general purpose codes, no matching of the numerical solution to an analytic expansion in the neighborhood of a singularity is necessary.

As a simple example consider the equation

$$y'' = \frac{1}{x}y' - \left(\frac{8}{7}\right)^2 e^y, \quad y'(0) = y(1) = 0,$$

which has the solution $u(x) = 2\log(7/(8 - x^2))$ [20].

Results with the initial guess $u \equiv 0$ are tabulated below

TABLE 2

k	$TOL(u)$	$E(u)$	est $E(u)$	$TOL(u')$	$E(u')$	est $E(u')$	*time*	*mesh sequence*
4	$.1-5$	$.33-8$	$.24-8$	$.1-5$	$.77-7$	$.90-7$	.56	2(3), 4(1)

We have also solved more complicated singular problems such as the Ginsburg-Landau equations [28] with little difficulty. This will be reported in more detail elsewhere.

Example 3 [26]. The last example arises when considering the flow between two counter-rotating infinite plane disks. The equations which describe the motion can be cast into the form

$$\begin{aligned} &\epsilon G'' + HG' - H'G = 0, \\ &\epsilon H^{iv} + HH''' + GG' = 0, \end{aligned} \qquad -1 \leqslant x \leqslant 1,$$

and

$$G(-1) = \mu, \quad G(1) = 1, \quad H(-1) = H'(-1) = H(1) = H'(1) = 0$$

with $\mu = -1$ for the case where the disks are counter-rotating at the same speed. In this latter case, there exists an odd solution [26]. This solution has boundary layers near both ends and varies smoothly in between.

Various investigators have computed solutions to this problem (see references in [26], [27]) with conflicting results. The problem becomes very ill-conditioned for small values of $\epsilon > 0$. In [27], the antisymmetry for the particular case $\mu = -1$ is used to solve the problem on the half interval [0, 1] with $G(0) = 0$ and $H(0) = H''(0) = 0$ replacing the boundary conditions at -1. This improves the condition of the problem significantly, enabling the authors to obtain solutions for $\epsilon = 10^{-4}$.

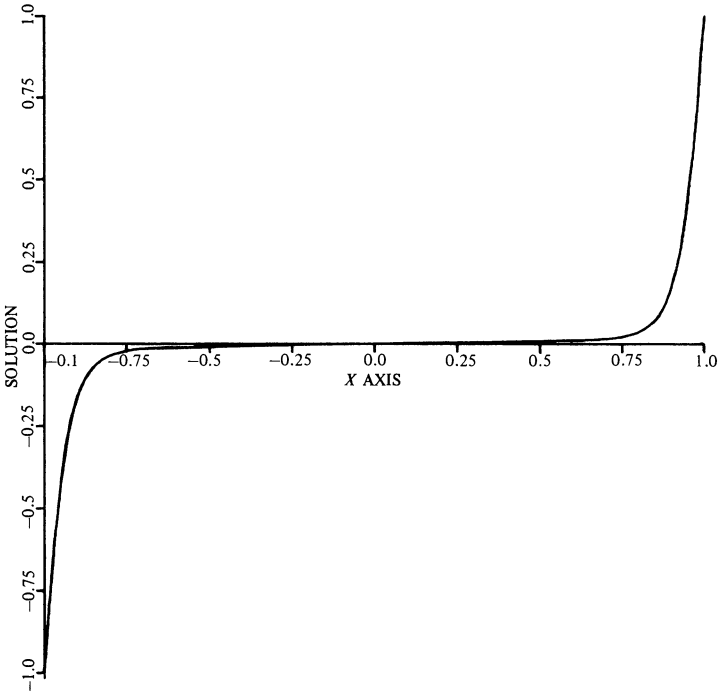

FIGURE 5.2

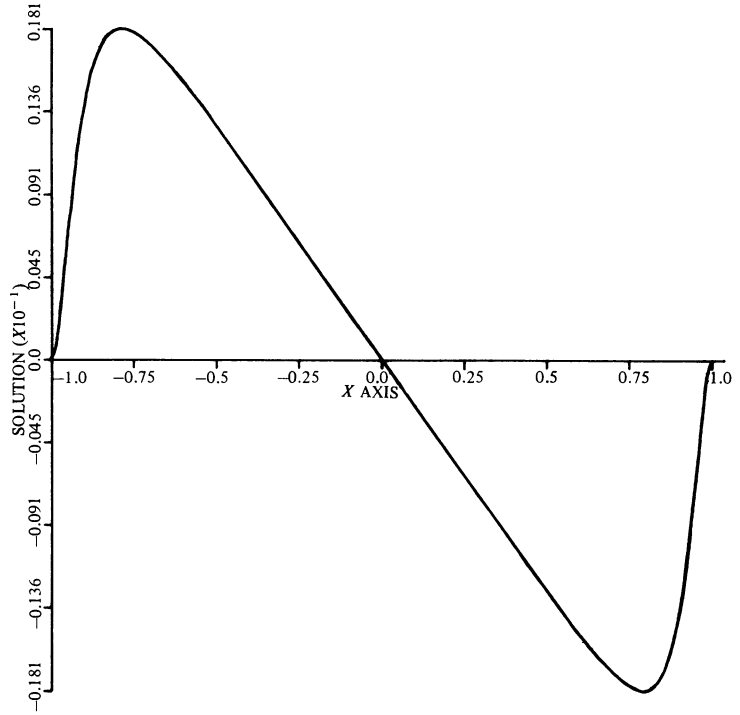

FIGURE 5.3

We solve the problem with $\mu = -1$ and $\epsilon = 10^{-3}$ on $[-1, 1]$, obtaining the odd solution without using antisymmetry. This enables us to demonstrate the stability and reliability of COLSYS. By comparing values of the obtained solution at points x and $-x$ (with the final mesh being nonuniform) we get an idea of how well the error estimates do.

For $k = 5$, $TOL(G) = TOL(H) = TOL(H') = .1-5$, and the initial guess $G = x^3$, $H = -x(x-1)^2(x+1)^2$, the results are listed in Table 3. Tests based on the symmetries of the solution support the error estimates.

TABLE 3

est $E(G)$	est $E(G')$	est $E(H)$	est $E(H')$	est $E(H'')$	est $E(H''')$	*time*	*mesh sequences*
$.78-6$	$.11-3$	$.79-8$	$.22-6$	$.15-4$	$.33-2$	56.45	10(10), 5(4), 10(5), 20(2)

The computed curves of G and H are plotted in Figures 5.2 and 5.3, respectively.

Computer Science Department
University of British Columbia
Vancouver, B. C., Canada V6T 1W5

Department of Mathematics
Simon Fraser University
Burnaby, B. C., Canada V5A 1S6

1. U. ASCHER, "Discrete least squares approximations for ordinary differential equations," *SIAM J. Numer. Anal.*, v. 15, 1978, pp. 478–496.

2. U. ASCHER, J. CHRISTIANSEN & R. D. RUSSELL, *A Collocation Solver for Mixed Order Systems of Boundary Value Problems*, Comp. Sci. Tech. Rep. 77-13, Univ. of British Columbia, 1977.

3. U. ASCHER & R. D. RUSSELL, *Evaluation of B-Splines for Solving Systems of Boundary Value Problems*, Comp. Sci. Tech. Rep. 77-14, Univ. of British Columbia, 1977.

4. C. DE BOOR, "On calculating with *B*-splines," *J. Approximation Theory*, v. 6, 1972, pp. 50–62.

5. C. DE BOOR, *Good Approximation by Splines with Variable Knots.* II, Lecture Notes in Math., vol. 363, Springer-Verlag, Berlin and New York, 1973.

6. C. DE BOOR, "Package for calculating with *B*-splines," *SIAM J. Numer. Anal.*, v. 14, 1977, pp. 441–472.

7. C. DE BOOR & B. SWARTZ, "Collocation at Gaussian points," *SIAM J. Numer. Anal.*, v. 10, 1973, pp. 582–606.

8. C. DE BOOR & R. WEISS, *Solveblok: A Package for Solving Almost Block Diagonal Linear Systems, with Applications to Spline Approximation and the Numerical Solution of Ordinary Differential Equations*, MRC TSR #1625, Madison, Wisconsin, 1976.

9. C. BROYDEN, "Recent developments in solving nonlinear, algebraic systems," in *Numerical Methods for Nonlinear Algebraic Equations*, P. Rabinowitz (ed.), Gordon & Breach, New York, 1977.

10. R. BULIRSCH, J. STOER & P. DEUFLHARD, *Numerical Solution of Nonlinear Two-Point Boundary Value Problems.* I, Numer. Math. Handbook Series Approximation, 1976.

11. J. CERUTTI, *Collocation for Systems of Ordinary Differential Equations*, Comp. Sci. Tech. Rep. 230, Univ. Wisconsin-Madison, 1974.

12. J. CHRISTIANSEN & R. D. RUSSELL, "Error analysis for spline collocation methods with application to knot selection," *Math. Comp.*, v. 32, 1978, pp. 415–419.

13. P. CIARLET, M. SCHULTZ & R. VARGA, "Numerical methods of high-order accuracy for nonlinear boundary value problems. I. One dimensional problem," *Numer. Math.*, v. 9, 1967, pp. 394–430.

14. J. DANIEL & A. MARTIN, "Implementing deferred corrections for Numerov's difference method for second-order two-point boundary-value problems," *SIAM J. Numer. Anal.*, v. 14, 1977, pp. 1033–1050.

15. J. DENNIS & J. MORÉ, "Quasi-Newton methods, motivation and theory," *SIAM Rev.*, v. 19, 1977, pp. 46–89.

16. P. DEUFLHARD, "A relaxation strategy for the modified Newton method," *Optimization and Optimal Control,* Bulirsch, Oettli and Stoer (eds.), Lecture Notes in Math., vol. 477, Springer, Berlin and New York, 1975, pp. 59–73.

17. H. J. DIEKOFF, P. LORY, H. J. OBERLE, H. J. PESCH, P. RENTROP & R. SEYDEL, "Comparing routines for the numerical solution of initial value problems of ordinary differential equations in multiple shooting," *Numer. Math.*, v. 27, 1977, pp. 449–469.

18. D. DODSON, *Optimal Order Approximation by Polynomial Spline Functions,* Ph. D. thesis, Purdue Univ., 1972.

19. R. ENGLAND, N. NICHOLS & J. REID, *Subroutine* D003AD, Harwell subroutine library, Harwell, England, 1973.

20. S. C. EISENSTADT, R. S. SCHREIBER & M. H. SCHULTZ, *Finite Element Methods for Spherically Symmetric Elliptic Equations,* Res. Rept. #109, Comp. Sci., Yale Univ., 1977.

21. J. E. FLAHERTY & R. E. O'MALLEY, JR., "The numerical solution of boundary value problems for stiff differential equations," *Math. Comp.*, v. 31, 1977, pp. 66–93.

22. P. HEMKER, *A Numerical Study of Stiff Two-Point Boundary Problems,* Math. Centrum, Amsterdam, 1977.

23. E. HOUSTIS, "A collocation method for systems of nonlinear ordinary differential equations," *J. Math. Anal. Appl.*, v. 62, 1978, pp. 24–37.

24. M. LENTINI & V. PEREYRA, "A variable order finite difference method for nonlinear multipoint boundary value problems," *Math. Comp.*, v. 28, 1974, pp. 981–1004.

25. M. LENTINI & V. PEREYRA, "An adaptive finite difference solver for nonlinear two point boundary problems with mild boundary layers," *SIAM J. Numer. Anal.*, v. 14, 1977, pp. 91–111.

26. J. B. McLEOD & S. V. PARTER, "On the flow between two counter rotating infinite plane disks," *Arch. Rational Mech. Anal.*, v. 54, 1974, pp. 301–327.

27. H. J. PESCH & P. RENTROP, *Numerical Solution of the Flow between Two-Counter-Rotating Infinite Plane Disks by Multiple Shooting,* Rep. #7621, Technische Universität München, 1976.

28. P. RENTROP, "Numerical solution of the singular Ginzburg-Landau equations by multiple shooting," *Computing,* v. 16, 1976, pp. 61–67.

29. R. D. RUSSELL, "Collocation for systems of boundary value problems," *Numer. Math.*, v. 23, 1974, pp. 119–133.

30. R. D. RUSSELL, "Efficiencies of B-splines methods for solving differential equations," *Proc. Fifth Conference on Numerical Mathematics,* Utilitas Math., Winnipeg, Manitoba, 1975, pp. 599–617.

31. R. D. RUSSELL, "A comparison of collocation and finite differences for two-point boundary value problems," *SIAM J. Numer. Anal.*, v. 14, 1977, pp. 19–39.

32. R. D. RUSSELL & J. CHRISTIANSEN, "Adaptive mesh selection strategies for solving boundary value problems," *SIAM J. Numer. Anal.*, v. 15, 1978, pp. 59–80.

33. R. D. RUSSELL & L. F. SHAMPINE, "A collocation method for boundary value problems," *Numer. Math.*, v. 19, 1972, pp. 1–28.

34. R. D. RUSSELL & L. F. SHAMPINE, "Numerical methods for singular boundary value problems," *SIAM J. Numer. Anal.*, v. 12, 1975, pp. 13–36.

35. M. L. SCOTT & H. A. WATTS, "Computational solutions of linear two-point boundary value problems via orthonormalization," *SIAM J. Numer. Anal.*, v. 14, 1977, pp. 40–70.

36. M. L. SCOTT & H. A. WATTS, "A systematized collection of codes for solving two-point boundary-value problems," in *Numerical Methods for Differential Systems,* Academic Press, New York, 1976, pp. 197–227.

37. J. M. VARAH, "Alternate row and column elimination for solving certain linear systems," *SIAM J. Numer. Anal.*, v. 13, 1976, pp. 71–75.

38. K. WITTENBRINK, "High order projection methods of moment and collocation type for nonlinear boundary value problems," *Computing,* v. 11, 1973, pp. 255–274.