COLLOCATION FOR TWO-POINT BOUNDARY VALUE PROBLEMS REVISITED

by

Uri Ascher

Technical Report 84-17

November, 1984

COLLOCATION FOR TWO-POINT BOUNDARY VALUE PROBLEMS REVISITED

by

Uri Ascher

Technical Report 84-17

November, 1984

ABSTRACT

Collocation methods for two-point boundary value problems for higher order differential equations are considered. By using appropriate monomial bases, we relate these methods to corresponding one-step schemes for 1st order systems of differential equations. This allows us to present the theory for nonstiff problems in relatively simple terms, refining at the same time some convergence results and discussing stability. No restriction is placed on the meshes used.

Department of Computer Science, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W5 . Supported in part under NSERC (Canada) Grant A4306.

1. Introduction

More than a decade ago, de Boor and Swartz [9] analyzed a class of collocation methods for the numerical solution of boundary value problems (BVP) for a higher order ordinary differential equation (ODE). In that pioneering work, the authors put tools from functional analysis and approximation theory to good use, showing that under suitable assumptions there is a piecewise polynomial collocation solution which achieves a high order of convergence and may be obtained by an efficient and general implementation. This work was supplemented and extended in a number of ways, see Russell [18], Weiss [22], Cerutti [13], Wittenbrink [23], Russell and Christiansen [19], Christiansen and Russell [14] and others. Earlier, related work includes Vainikko [21] and Russell and Shampine [20]. A general purpose code called COLSYS [2,3] for mixed order systems of BVPs was written based on this knowledge. It proved to be a useful practical software package, and has increased interest in this type of collocation methods.

But the approach taken in those of the above cited papers which deal with higher order ODEs is not without its drawbacks. In the first place, the relationship between these collocation methods and some finite difference schemes, which was recognized already in [9] and [22], is not obvious. As a result, questions of stability and conditioning are typically not addressed at all in this literature. Moreover, the relatively sophisticated mathematics required, while being a source of excitement to some readers, is a source of anxiety to others.

In this note we present the theory for the class of collocation methods under consideration from an alternative point of view, as proposed by M. Osborne and discussed in Ascher, Pruess and Russell [4]. Since the nonlinear treatment is rather similar in both approaches (consisting of quasilinearization and application of some variant of the

Newton-Kantorovitch Theorem, cf. Keller [15] and [9]), we concentrate mainly on linear problems. Even though our approach is simpler, we obtain high order convergence and superconvergence results with a better localization of the error than in [9] and [14], without putting any restriction on the mesh used (see Theorem 11 below). Moreover, we relate the collocation method for a higher order ODE to a nontrivial, efficient finite difference scheme for the corresponding first order system of ODEs, and show that the condition number of the resulting linear systems of algebraic equations depends, under certain restrictions, on the number of mesh points and on the condition number of the equivalent 1st order BVP alone. In particular, no restriction is placed on the mesh. The latter fact was to some extent already pointed out in [4], but there this alternative approach was considered mainly in an implementation context, and no relation to the BVP condition number was made.

In order to be specific, let us consider a two-point BVP of the form

$$Nu \equiv u^{(m)} - f(x, u, u', \cdots, u^{(m-1)}) = 0, \qquad 0 < x < 1, \tag{1a}$$

$$\mathbf{g}(\mathbf{y}(0), \mathbf{y}(1)) = 0, \tag{1b}$$

where

$$\mathbf{y}(x) := (u(x), u'(x), \cdots, u^{(m-1)}(x))^T. \tag{2}$$

The basic idea of collocation is quite general: An approximate solution is sought in the form

$$u_\pi(x) = \sum_{j=1}^{M} \alpha_j \phi_j(x) \qquad 0 \le x \le 1 \tag{3}$$

where $\phi_j(x)$ are known linearly independent basis functions defined on [0,1], and α_j are

parameters. These parameters are determined by requiring $u_\pi(x)$ to satisfy the ODE or the BC at M points (the *collocation points*) in [0,1]. It is sometimes convenient to say that $u_\pi(x)$ is an element in a *linear space* of *dimension* M which is *spanned* by *basis functions* $\phi_1(x), \cdots, \phi_M(x)$. Here, this linear space is chosen to consist of piecewise polynomial functions. Thus, there is a partition π of [0,1]

$$\pi: 0 = x_1 < x_2 < \cdots < x_N < x_{N+1} = 1 \tag{4}$$
$$h_i := x_{i+1} - x_i, \qquad h := \max_{1 \le i \le N} h_i$$

such that any linear combination of the basis functions reduces to a polynomial on each subinterval $[x_i, x_{i+1}]$, $1 \le i \le N$. Furthermore, we restrict the functions $\phi_j(x)$, and therefore any of their linear combinations, to be in $C^{m-1}[0,1]$ (like the exact solution for piecewise continuous data). Also, the order of the polynomial pieces is restricted to be $k+m$, for some $k \ge m$.[1]

A k-stage collocation method under consideration is determined by a mesh π and a set of k points

$$0 \le \rho_1 < \rho_2 < \cdots < \rho_k \le 1. \tag{5}$$

Denoting, similarly to (2),

$$\mathbf{y}_\pi(x) := (u_\pi(x), u_\pi'(x), \cdots, u_\pi^{(m-1)}(x))^T, \tag{6}$$

an approximate solution $u_\pi(x)$ defined on [0,1] is determined such that

$$\mathbf{y}_\pi(x) \in C[0,1]\ ; \qquad u_\pi(x) \in \mathbf{P}_{k+m,\pi} \tag{7a}$$

[1] We say that v is in $\mathbf{P}_{k+m}$ if $v(x)$ is a polynomial of order $k+m$ (degree $< k+m$) on an appropriate interval, and that v is in $\mathbf{P}_{k+m,\pi}$ if $v(x)$ is a piecewise polynomial which is in $\mathbf{P}_{k+m}$ on each subinterval of the mesh π.

$$\mathbf{g}(\mathbf{y}_\pi(0),\mathbf{y}_\pi(1)) = 0 \tag{7b}$$

and

$$Nu_\pi(x_{ij}) \equiv u_\pi^{(m)}(x_{ij}) - f(x_{ij},\mathbf{y}_\pi(x_{ij})) = 0 \qquad 1 \le j \le k,\ 1 \le i \le N, \tag{7c}$$

where

$$x_{ij} := x_i + h_i\rho_j, \qquad 1 \le j \le k,\ 1 \le i \le N. \tag{7d}$$

One important question is how to choose the basis functions $\phi_j(x)$ so as to obtain an efficient, stable method. Choices of Hermite-type bases and of B-splines are discussed in [9,2,10,4,17]. But here we actually prefer not to explicitly specify any basis functions. Instead, we consider local representations of the polynomial pieces, called *monomial bases* in [4]. This enables us to relate more directly to similar collocation methods for first order ODEs, and to see that the method introduced here is just a fancy finite difference method. The reason for the importance of the local representation is that when using basis functions $\phi_j(x)$ as above, the continuity conditions on u_π are already imbedded in the basis functions, while the collocation equations are satisfied only later. In §5 of [4], on the other hand, we first imposed the collocation equations, followed by local parameter elimination, and only then connected to the action in adjacent subintervals. This is what Osborne had advocated in an unpublished manuscript. We proceed here with the latter, multiple shooting type approach, which allows us to capitalize on well-known theoretical results for one-step finite difference schemes, and to avoid introducing heavier functional analysis machinery.

The linear form of the two-point BVP (1) is

$$Lu \equiv u^{(m)} - \mathbf{c}^{\mathrm{T}}(\mathrm{x})\,\mathbf{y} = q(x), \qquad 0 < x < 1, \tag{8a}$$

$$B_0\,\mathbf{y}(0) + B_1\,\mathbf{y}(1) = \mathbf{b}, \tag{8b}$$

where

$$\mathbf{c}^{\mathrm{T}}(\mathrm{x}) = (\mathrm{c}_1(\mathrm{x}), \ldots, \mathrm{c}_{\mathrm{m}}(\mathrm{x}))$$

are given coefficients, assumed to be sufficiently smooth and of a moderate size, and B_0, B_1 are $m \times m$ well-scaled boundary matrices. Corresponding to (8a,b) there is the first order BVP

$$\mathbf{y}' = \begin{bmatrix} 0 & 1 & & \\ & & \ddots & \\ & & 0 & 1 \\ c_1(x) & c_2(x) & \ldots & c_m(x) \end{bmatrix} \mathbf{y} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ q(x) \end{bmatrix} \equiv A(x)\mathbf{y} + \mathbf{q}(x), \tag{9a}$$

$$B_0\,\mathbf{y}(0) + B_1\,\mathbf{y}(1) = \mathbf{b}. \tag{9b}$$

Let $H(x,t)$ be Green's function for (9) (assumed to exist) and $\Phi(x)$ the fundamental matrix satisfying

$$B_0\,\Phi(0) + B_1\,\Phi(1) = I.$$

Define the condition number

$$\kappa := ||H||_\infty + ||\Phi||_\infty. \tag{10}$$

We will show

Theorem 11.

Assume that there are integers $p \geq k \geq m$ such that

(a) the linear BVP of order m (8) is well-posed, in the sense that κ of (10) is of a moderate size; has coefficients in $C^p[0,1]$; and has a unique solution $u(x)$ in $C^{p+m}[0,1]$; and

(b) the k canonical collocation points $\rho_1, \ldots, \rho_k$ of (5) satisfy the orthogonality conditions

$$\int_0^1 \phi(t) \prod_{l=1}^{k} (t-\rho_l)\,dt = 0 \qquad \phi \in \mathbf{P}_{p-k}\,. \tag{11a}$$

Then for h small enough the following hold:

(a) The collocation method (7) for the linear BVP has a unique solution $u_\pi(x)$.

(b) There exists an implementation such that the solution scheme is stable, with a stability constant $\kappa O(N)$.

(c) The following error estimates hold at mesh points:

$$|u^{(j)}(x_i) - u_\pi^{(j)}(x_i)| = O(h^p) \qquad 0 \le j \le m-1,\ 1 \le i \le N+1. \tag{11b}$$

(d) At any point in [0,1], the error satisfies

$$u^{(j)}(x) - u_\pi^{(j)}(x) = h_i^{k+m-j} u^{(k+m)}(x_i) P^{(j)}\left(\frac{x-x_i}{h_i}\right) + O(h_i^{k+m-j+1}) + O(h^p) \tag{11c}$$

$$x_i \le x \le x_{i+1},\ 1 \le i \le N,\ 0 \le j \le k+m-1,$$

where

$$P(\xi) = \frac{1}{k!(m-1)!} \int_0^\xi (t-\xi)^{m-1} \prod_{l=1}^{k} (t-\rho_l)\,dt. \tag{11d}$$

□

For nonlinear problems, we define the linearization

$$L[u]z(x) \equiv z^{(m)}(x) - \sum_{l=1}^{m} c_l(x) z^{(l-1)}(x) \qquad 0 < x < 1 \tag{12a}$$

$$c_l(x) := \frac{\partial f(x, \mathbf{y}(x))}{\partial y_l} \tag{12b}$$

where u and $\mathbf{y}$ are related by (2). For the boundary conditions, define B_0, B_1 depending

on u by

$$B_0 := \frac{\partial \mathbf{g}}{\partial \mathbf{y}(0)}, \qquad B_1 := \frac{\partial \mathbf{g}}{\partial \mathbf{y}(1)}. \tag{12c}$$

The following theorem is obtained:

Theorem 13.

Let $u(x)$ be an isolated solution of the BVP (1), where f and $\mathbf{g}$ have continuous second partial derivatives and the assumptions of Theorem 11 hold for u and $L[u]$. Consider a k-stage collocation method, satisfying (11a), for (7). Then there are positive constants ρ and h_0 such that for all meshes with $h \leq h_0$,

(a) There is a unique solution $u_\pi(x)$ to the collocation equations (7) in a tube of radius ρ around $u(x)$, $S_\rho(u)$.

(b) This solution $u_\pi(x)$ can be obtained by Newton's method, which converges quadratically provided that the initial guess for $u_\pi(x)$ is sufficiently close to $u(x)$.

(c) For the linearized BVPs, there is a stable implementation with stability constants $\kappa O(N)$, where κ is given by (10) for the linearized problem at u.

(d) The error estimates (11b,c) hold.

□

Remarks.

(a) For reasons of brevity and simplicity, we concentrate here on one higher order ODE and, to the extent needed for this purpose, also on 1st order systems of ODEs. But our results can be extended for mixed order systems of ODEs and for multipoint boundary conditions, cf. [13,2]. Moreover, the above smoothness

assumptions can also be weakened. In particular, the coefficients may be only piecewise smooth if points of discontinuity are included in the mesh π used. For a 1st order system of ODEs, collocating at the points of (7d) with a continuous piecewise polynomial vector function of order $k+1$, Theorems 11 and 13 hold with $m=1$. Therefore, considering such a collocation process for (9) and comparing it to the collocation method for (8), a similar error estimate (11b) at mesh points is achieved (see example in §3). On the other hand, the result (11c) shows that, when x is not a mesh point, $u_\pi(x)$ is a *better approximation* to $u(x)$ than the corresponding collocation approximation for the equivalent 1st order system. High order convergence results away from mesh points are useful for approximating functional differential equations, see Bader[13], and have been used in a general purpose implementation [3] as well.

(b) Consider the approximation space to which $u_\pi(x)$ belongs. This is a linear space of piecewise polynomials of order $k+m$. A result from approximation theory states that, unless $u(x)$ itself is in the approximation space, we cannot get a global approximation order of more than $O(h^{k+m})$. Hence, for $p \geq k+m$, the collocation approximation is, by (11b), of *optimal* global convergence order. Furthermore, if $p > k+m$ then at mesh points we obtain a *superconvergence* order, i.e. an order of convergence higher than the best possible global order. It is interesting to note that the superconvergence result, which (as the name implies) is perhaps less natural from the point of view of approximation spaces, is most natural from the point of view of one-step difference schemes, as will become evident in the sequel. We recall that choosing $\rho_1,..,\rho_k$ to be Gaussian points yields $p = 2k$, Radau points give $p = 2k-1$, whereas Lobatto points give $p = 2(k-1)$, cf. [22].

(c) A restriction on the mesh in Theorems 11 and 13 is remarkably absent. We have

not assumed that the mesh is quasiuniform, nor that it has a locally bounded mesh ratio. Consider, in particular, the error bound (11c) when $p > k+m$ (Otherwise the error is simply $O(h^p)$ everywhere). The error is seen to consist of two contributions: A global, superconvergence order term, and a local term. The leading order of this local term is explicitly given - an unusually strong result. This explicit result was derived in [14] and used in [2], but under a quasiuniformity assumption. Our proof helps to better explain the robust performance of the mesh selection strategy in COLSYS [2,3], which is based on (11c). Also, the condition number of the implementation involved is $O(N)$. In many methods for higher order BVPs one finds condition numbers which depend on $\frac{h}{\min h_i}$ and/or on N^m. Even the collocation methods considered here, when implemented using B-splines or Hermite-type basis functions, involve a condition number of $O(\frac{h}{\min h_i})^{m-1}$, see [4] and Paine and Russell [17].

(d) Note that we are *not* dealing here with problems of singular perturbation type. While collocation at Gaussian points has been found to be rather useful for such problems, see [1-3,5-7], the type of analysis required there is different than the one described here. Specifically, we have already assumed that the BVP coefficients are of moderate size, so for h "small enough", h^{-1} is the large quantity in the numerical approximation. This allows us to use the good approximation to the fundamental solutions in each mesh subinterval, as in standard multiple shooting. These assumptions also allow us to relate to κ of (10) as the condition number of (9), see de Boor, de Hoog and Keller [11], Lentini, Osborne and Russell [16] and de Hoog and Mattheij [12] for refinements.

□

2. One-step schemes for first order linear systems

In this section we consider the BVP of size m

$$\mathbf{L}\mathbf{y}(x) \equiv \mathbf{y}'(x) - A(x)\mathbf{y}(x) = \mathbf{q}(x) \qquad 0 < x < 1 \tag{14}$$

and (9b), with the condition number κ of (10) and the smoothness and boundedness assumptions of §1. We do not assume that $A(x)$ necessarily has the form as in (9a). For this BVP, consider a one-step difference method on a mesh π,

$$\mathbf{L}_\pi \mathbf{y}_i \equiv \frac{\mathbf{y}_{i+1}-\mathbf{y}_i}{h_i} - \Psi(\mathbf{y}_i,\mathbf{y}_{i+1};h_i) = \mathbf{q}_i, \qquad 1 \le i \le N, \tag{15}$$

which we also write as

$$\mathbf{L}_\pi \mathbf{y}_i \equiv S_i\mathbf{y}_i + R_i\mathbf{y}_{i+1} = \mathbf{q}_i \qquad 1 \le i \le N, \tag{16a}$$

and

$$B_0\mathbf{y}_1 + B_1\mathbf{y}_{N+1} = \mathbf{b}. \tag{16b}$$

Here, $\mathbf{y}_i$ is to approximate $\mathbf{y}(x_i)$, $1 \le i \le N+1$, and $\mathbf{q}_i = \mathbf{q}(x_i) + O(h_i)$.

In (16) we have a system of $m(N+1)$ linear equations

$$\begin{pmatrix} S_1 & R_1 & & & \\ & S_2 & R_2 & & \\ & & \ddots & & \\ & & & S_N & R_N \\ B_0 & & & & B_1 \end{pmatrix} \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \\ \mathbf{y}_{N+1} \end{pmatrix} = \begin{pmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \vdots \\ \mathbf{q}_N \\ \mathbf{b} \end{pmatrix} \tag{17}$$

Denote the matrix in (17) by $\mathbf{A}$. Methods for the solution of (17) using the sparseness structure of $\mathbf{A}$ have been dicussed in the literature. The finite difference method (16) is said to be *stable* if there are constants K and $h_0 > 0$ such that for all meshes π with

$h \leq h_0$

$$\|\mathbf{A}^{-1}\|_\infty \leq K \tag{18}$$

and K is a constant of moderate size if κ is. Recall [15] that the *local truncation error* $\tau_i[\mathbf{y}]$ is defined as

$$\tau_i[\mathbf{y}] := \mathbf{L}_\pi \mathbf{y}(x_i) - \mathbf{q}_i \qquad 1 \leq i \leq N \tag{19}$$

and that the finite difference scheme (16) is said to be *consistent* of *order* p (p a positive integer) if for every (smooth) solution of (14) there exist constants c and $h_0 > 0$ such that for all meshes π with $h := \max_{1 \leq i \leq N} h_i \leq h_0$,

$$\tau[\mathbf{y}] := \max_{1 \leq i \leq N} |\tau_i[\mathbf{y}]| \leq ch^p. \tag{20}$$

Here $|\cdot|$ denotes the max vector norm. Note that when substituting the error

$$\mathbf{e}_i := \mathbf{y}(x_i) - \mathbf{y}_i \qquad 1 \leq i \leq N+1 \tag{21}$$

in the difference scheme, one obtains

$$|\mathbf{e}_i| \leq Kch^p \qquad 1 \leq i \leq N+1. \tag{22}$$

This is why we need K of (18) to be of a moderate size.

Theorem 23.

Suppose that the finite difference scheme (15) is consistent (of at least first order) and satisfies

$$|\Psi(\mathbf{u},\mathbf{v};h)| \leq c(|\mathbf{u}| + |\mathbf{v}|) \qquad h \leq h_0 \tag{23a}$$

c a constant independent of h, and that the BVP (14) is well-posed with condition number κ . Then the scheme is stable with stability constant

$$K = \kappa + O(h) \tag{23b}$$

Moreover, **A** of (17) can be equilibrated so that

$$cond(\mathbf{A}) \leq c\kappa N. \tag{23c}$$

□

The proof is very simple and is also contained in the results of [11] and [16]. We sketch it here following R.D. Russell (private communication): by (23a) we can write

$$h_i R_i = I + O(h_i), \qquad h_i S_i = I + O(h_i)$$

so (16a) can be written as

$$\mathbf{y}_{i+1} = \Gamma_i \mathbf{y}_i + \mathbf{r}_i \tag{24a}$$

$$\Gamma_i := -R_i^{-1} S_i, \qquad \mathbf{r}_i := R_i^{-1} \mathbf{q}_i = O(h_i) \tag{24b}$$

with $\Gamma_i = I + O(h_i)$ a well-defined $m \times m$ matrix. It then follows that

$$\Gamma_i = Y(x_i; x_{i+1}) + O(h_i^2)$$

where $Y(x;t)$ is a fundamental matrix defined by

$$\mathbf{L}\, Y(\cdot\,;t) = 0, \qquad Y(t;t) = I.$$

Thus, if we multiply the ith block of m rows of **A** by R_i^{-1}, we obtain a matrix approximating a standard multiple shooting matrix with an explicitly known inverse, first given by M. Osborne [16]. This inverse involves the Green's matrix $H(x,t)$, see (10), and the results (23b,c) follow.

□

Many one-step difference schemes are in the class of *Runge-Kutta schemes*

$$h_i^{-1}(\mathbf{y}_{i+1} - \mathbf{y}_i) = \sum_{l=1}^{k} \beta_l \{A(x_{il})\mathbf{y}_{il} + \mathbf{q}(x_{il})\} \tag{25a}$$

$$h_i^{-1}(\mathbf{y}_{ij} - \mathbf{y}_i) = \sum_{l=1}^{k} \alpha_{jl} \{A(x_{il})\mathbf{y}_{il} + \mathbf{q}(x_{il})\} \qquad 1 \le j \le k \tag{25b}$$

with x_{il} given in (7d), $1 \le i \le N$. The unknowns $\mathbf{y}_{i1}, \ldots, \mathbf{y}_{ik}$ can be eliminated locally for each subinterval of the mesh, and a difference scheme (16) results.

Collocation schemes are a subclass of Runge-Kutta schemes, whereby the coefficients α_{jl}, β_l appearing in (25) are given, for points ρ_j satisfying (5), by

$$\beta_l = \int_0^1 L_l(t)\,dt \qquad \alpha_{jl} = \int_0^{\rho_j} L_l(t)\,dt \qquad 1 \le j,\ l \le k. \tag{26}$$

Here L_l are Legendre polynomials. Recall that for any function $v \in C^k[x_i, x_{i+1}]$ we can write

$$v(x) = \sum_{l=1}^{k} v(x_{il}) L_l\left(\frac{x - x_i}{h_i}\right) + \psi(x) \qquad x_i \le x \le x_{i+1} \tag{27a}$$

where

$$L_l(t) := \frac{(t-\rho_1)\ldots(t-\rho_{l-1})(t-\rho_{l+1})\ldots(t-\rho_k)}{(\rho_l-\rho_1)\ldots(\rho_l-\rho_{l-1})(\rho_l-\rho_{l+1})\ldots(\rho_l-\rho_k)} \qquad 1 \le l \le k \tag{27b}$$

and the remainder term $\psi(x)$ is expressed in a (bounded) divided difference form

$$\psi(x) = v[x_{i1}, \ldots, x_{ik}, x] \prod_{l=1}^{k} (x - x_{il}). \tag{27c}$$

It follows that any such collocation method is consistent of order at least k, and so Theorem 23 applies to it.

3. Collocation for a linear ODE of order m

Recall the monomial basis of Osborne and [4]: Let us focus on one subinterval $[x_i, x_{i+1}]$ and express the polynomial $u_\pi(x)$ in terms of its Taylor series about x_i,

$$u_\pi(x) = \sum_{j=1}^{k+m} \frac{(x-x_i)^{j-1}}{(j-1)!} u_\pi^{(j-1)}(x_i) \qquad x_i \le x \le x_{i+1} .$$

This can be written as

$$u_\pi(x) = \sum_{j=1}^{m} \frac{(x-x_i)^{j-1}}{(j-1)!} y_{ij} + h_i^m \sum_{j=1}^{k} \psi_j(\frac{x-x_i}{h_i}) z_{ij} \tag{28}$$

where, corresponding to (2),

$$\mathbf{y}_\pi(x_i) \equiv \mathbf{y}_i = (y_{i1},\ldots,y_{im})^T,$$

and

$$z_{ij} = h_i^{j-1} u_\pi^{(m+j-1)}(x_i), \quad \mathbf{z}_i := (z_{i1},\ldots,z_{ik})^T.$$

The functions $\psi_j(t)$ are therefore defined as

$$\psi_j(t) = \frac{t^{m+j-1}}{(m+j-1)!} \qquad 0 \le t \le 1,\ 1 \le j \le k. \tag{29}$$

Note that $\psi_1(t),\ldots,\psi_k(t)$ are linearly independent polynomials of order $k+m$ on $[0,1]$, satisfying

$$\psi_j^{(l-1)}(0) = 0 \qquad 1 \le l \le m,\ 1 \le j \le k. \tag{30}$$

These functions are independent of i and can be used for each subinterval as in (28).

Now, we can write down the constraints (7) which define the approximate solution $u_\pi(x)$ in terms of the parameters $\mathbf{y}_i$ and $\mathbf{z}_i$ of the representation (28). For the linear problem (8) we write

$$Lu_\pi(x) = h_i^m \sum_{j=1}^{k} z_{ij} L[\psi_j(\frac{x-x_i}{h_i})] - \sum_{l=1}^{m} c_l(x) \sum_{j=l}^{m} \frac{y_{ij}(x-x_i)^{j-l}}{(j-l)!}$$

so the collocation conditions (7c) give

$$V\mathbf{y}_i + W\mathbf{z}_i = \mathbf{q}_i \qquad 1 \le i \le N \tag{31a}$$

where $\mathbf{q}_i = (q(x_{i1}),...,q(x_{ik}))^T$, V is a $k \times m$ matrix with entries

$$V_{rj} = -\sum_{l=1}^{j} \frac{c_l(x_{ir})(h_i\rho_r)^{j-l}}{(j-l)!} \qquad 1 \le r \le k,\ 1 \le j \le m \tag{32a}$$

and W is a $k \times k$ matrix with entries

$$W_{rj} = \psi_j^{(m)}(\rho_r) - \sum_{l=1}^{m} c_l(x_{ir}) h_i^{m+1-l} \psi_j^{(l-1)}(\rho_r) \qquad 1 \le r, j \le k. \tag{32b}$$

The continuity conditions in (7a) are even easier to write down. We evaluate $u_\pi(x)$ and its first $m-1$ derivatives at $x = x_{i+1}$ by (28) and equate to $\mathbf{y}_{i+1}$, the corresponding values at the (i+1)st subinterval. This yields

$$\mathbf{y}_{i+1} = C\mathbf{y}_i + D\mathbf{z}_i \qquad 1 \le i \le N \tag{31b}$$

where C is an $m \times m$ upper triangular matrix with entries

$$C_{rj} = \frac{h_i^{j-r}}{(j-r)!} \qquad j \ge r \tag{32c}$$

and D is an $m \times k$ matrix with entries

$$D_{rj} = h_i^{m+1-r} \psi_j^{(r-1)}(1) \qquad 1 \le r \le m,\ 1 \le j \le k. \tag{32d}$$

Note that the obvious dependence of C, D, W and V on i has been suppressed in the notation.

The specification of the collocation constraints (7) for the linear problem is completed by writing for (7b)

$$B_0\mathbf{y}_1 + B_1\mathbf{y}_{N+1} = \mathbf{b} \tag{31c}$$

Our next step is to eliminate the local unknowns $\mathbf{z}_i$. We note that as $h_i \to 0$, $W_{rj} \to \rho_r^{j-1}/(j-1)!$, so for h_i small enough W is nonsingular. Eliminating $\mathbf{z}_i$ from (31a) and substituting in (31b), we arrive at the form

$$\mathbf{y}_{i+1} = \Gamma_i\mathbf{y}_i + \mathbf{r}_i \qquad 1 \le i \le N \tag{33a}$$

with

$$\Gamma_i := C - DW^{-1}V, \qquad \mathbf{r}_i := DW^{-1}\mathbf{q}_i. \tag{33b}$$

The similarity between (33) and (24) is quite clear, noting that $C = I + O(h_i)$ and $D = O(h_i)$.

In (33a), (31c) we have, once again, a linear system of equations of the form (17) for $u_\pi(x)$ and its first $m-1$ derivatives at mesh points. After obtaining the values of $\mathbf{y}_i$ we can easily obtain $\mathbf{z}_i$ from (31a), and hence $u_\pi(x)$ from (28), if we store the values of $W^{-1}V$ and $W^{-1}\mathbf{q}_i$ which are computed while assembling Γ_i, for each i, $1 \le i \le N$.

Example

Consider the problem

$$u'' = -\frac{1}{x}u' + \left(\frac{8}{8-x^2}\right)^2$$

$$u'(0) = u(1) = 0.$$

The exact solution is

$$u(x) = 2\ln\left(\frac{7}{8-x^2}\right).$$

We have solved this problem numerically, using collocation at Gauss points with

$k = 2, 3$, and at Lobatto points with $k = 3$. Uniform meshes were used. The process was then repeated for the corresponding 1st order system (9) using collocation at the same mesh points. The maximum errors at mesh points for $u_\pi(\equiv y_1)$ and $u_\pi'(\equiv y_2)$ are listed in the table below. The results were ***identical*** (to the number of digits shown) in both computations.

□

Remarks.

(a) The results of the above example indicate a rather strong connection between the collocation method for the higher order BVP (8) and the corresponding collocation method using the same k points ρ_j $(k \geq m)$ for the transformed 1st order system (9). Indeed, the linear equations (33a) form a ***one-step difference scheme*** for (9). This will allow us to connect to the theory of one-step schemes, briefly reviewed in §2. But first we must ask if we have gained anything in the treatment here; indeed, is anything different than the corresponding collocation methods of §2? In answer, let us first point out that while here $u_\pi(x) \in \mathbf{P}_{k+m,\pi} \cap C^{m-1}[0,1]$, the

	2-Gauss		3-Lobatto		3-Gauss	
N	$e(y_1)$	$e(y_2)$	$e(y_1)$	$e(y_2)$	$e(y_1)$	$e(y_2)$
2	.20-3	.71-4	.17-4	.11-3	.14-6	.37-6
5	.64-5	.19-5	.57-6	.29-5	.70-9	.17-8
10	.46-6	.12-6	.37-7	.18-6	.13-10	.27-10
20	.33-7	.77-8	.23-8	.11-7	.27-12	.42-12
40	.23-8	.48-9	.15-9	.72-9	.60-14	.71-14
80	.16-9	.30-10	.91-11	.45-10	.13-14*	.94-15*

Higher order collocation schemes for a simple example
* - These values are mainly roundoff errors

corresponding collocation approximation of y_1 in (9) is in $\mathbf{P}_{k+1,\pi} \cap C[0,1]$, so generally the approximations are not exactly the same when $m>1$. Next we note that in (28) we have $k+m$ parameters per mesh subinterval, whereas the corresponding method for a 1st order system would have $m(k+1)$, i.e. $(m-1)k$ additional parameters. Of course the treatment of these parameters is slightly more cumbersome here; still, the matrix W for (9) is $mk \times mk$ and V is $mk \times m$, whereas in (32) W is merely $k \times k$ and V is $k \times m$. We can view (33) as *a sophisticated high order one-step scheme for (9), which takes advantage of the special structure of the coefficients in the transformed 1st order system.* We hasten to point out, however, that this view is beneficial only at mesh points.

(b) While the entire treatment here is done for h "sufficiently small", we point out in passing that collocation at symmetric points can be applied to advantage also when h is not so small compared to the coefficients in (8). In that case, however, the matrix W is not necessarily nonsingular, and the matrix

$$\begin{pmatrix} W \\ D \end{pmatrix}$$

has to be considered instead.

(c) The representation for the polynomial $u_\pi(x)$ in (28) is not unique. In particular, other choices for $\psi_j(t)$ can be considered, requiring (30) to hold [4]. Two such choices are mentioned: Requiring

$$\psi_j^{(m)}(\rho_r) = \delta_{jr} \qquad 1 \le j, r \le k \tag{34a}$$

implies $z_{ij} = u_\pi^{(m)}(x_{ij})$. It is easy to see that (34a), (30) defines $\psi_j(t)$ well. Moreover, for $m = 1$ we have

$$\psi_j(\rho_r) = \alpha_{rj}, \qquad \psi_j(1) = \beta_j$$

with α_{rj}, β_j given by (26). This is suitable for proving Theorem 11. Another choice which we mention is

$$\psi_j^{(r-1)}(1) = \delta_{j-k+m,r} \qquad 1 \le j,\, r \le k \tag{34b}$$

The importance of this is that D of (32d) becomes very simple, so this is the recommended variant for implementation.

□

We are now ready to prove the convergence results stated in §1.

Proof of Theorem 11.

Consider the representation (28) for $x_i \le x \le x_{i+1}$, with $\psi_j(x)$ defined by (30), (34a). We can write

$$u(x) = \sum_{j=1}^{m} \frac{(x-x_i)^{j-1}}{(j-1)!} u^{(j-1)}(x_i) + h_i^m \sum_{j=1}^{k} \psi_j\left(\frac{x-x_i}{h_i}\right) u^{(m)}(x_{ij}) + \mu(x). \tag{35}$$

It is easy to see that $\mu(x) = O(h_i^{k+m})$: In fact, recalling (27),

$$\mu^{(m)}(x) = u^{(m)}[x_{i1},\ldots,x_{ik},x] \prod_{l=1}^{k} (x-x_{il}) \tag{36a}$$

$$\mu(x_i) = \cdots = \mu^{(m-1)}(x_i) = 0. \tag{36b}$$

From this we obtain that $\mathbf{y}(x)$ of (2), which is the exact solution of the 1st order BVP (9), satisfies

$$\mathbf{y}(x_{i+1}) = \Gamma_i \mathbf{y}(x_i) + \mathbf{r}_i + O(h_i^{k+1}).$$

This means that the finite difference scheme (33a), (31c) is a one-step scheme for (9), which is consistent of order k, with (23a) holding. From Theorem 23 we obtain existence of a unique collocation solution and stability, as claimed in (a), (b). Moreover, the error

at mesh points satisfies, by (22),

$$|\mathbf{y}_i - \mathbf{y}(x_i)| = O(h^k) \qquad 1 \le i \le N+1. \tag{37a}$$

Let us use the notation $d_\pi(x) := u_\pi(x) - u(x)$ and write as in (28)

$$d_\pi(x) = \sum_{j=1}^{m} \frac{(x-x_i)^{j-1}}{(j-1)!} e_{ij} + h_i^m \sum_{j=1}^{k} \psi_j\left(\frac{x-x_i}{h_i}\right) g_{ij} - \mu(x) \tag{37b}$$

with $(e_{i1}, \ldots, e_{im})^T = \mathbf{e}_i = \mathbf{y}_i - \mathbf{y}(x_i)$. Equations (31a) for the error give the estimate

$$|g_{ij}| = O(h^k)$$

and so, taking derivatives in the expression (37b) for $d_\pi(x)$ we obtain

$$|u_\pi^{(j)}(x) - u^{(j)}(x)| = \begin{cases} O(h^k) & 0 \le j \le m \\ & x_i \le x \le x_{i+1} \\ O(h^{k+m-j})\theta_i^{j-m} & m \le j \le k+m-1 \end{cases} \tag{38a}$$

with

$$\theta_i := h/h_i. \tag{38b}$$

As in [9], let $G(x,t)$ be Green's function for (8) and denote $G_j(x,t) \equiv \dfrac{\partial^j}{\partial x^j} G(x,t)$. Then from the assumption on the problem coefficients, $G_j(x,t)$ is smooth as a function of t in [0,x) and in (x,1]; but at $t = x$, $G_j(x,t)$ has only $m-1-j$ derivatives. Writing for $0 \le j \le m-1$

$$u_\pi^{(j)}(x) - u^{(j)}(x) = \int_0^1 G_j(x,t)\, L(u_\pi(t) - u(t))\, dt = \sum_{n=1}^{N} E_n(x)$$

with

$$E_n(x) = \int_{x_n}^{x_{n+1}} G_j(x,t)(Lu_\pi(t) - q(t))\, dt,$$

we note that $Lu_\pi(t) - q(t) = 0$ at k collocation points $t = x_{nl}$. Thus, we can write the integrand as the remainder of its polynomial interpolant at these points,

$$G_j(x,t)\ L(u_\pi(t)-u(t)) = w(t)\prod_{l=1}^{k}(t-x_{nl})$$

where $w(t)$ involves the kth derivative of the integrand. From (38a) and the smoothness assumptions, $\theta_n^{-k+1}L(u_\pi(t)-u(t))$ has $p-k$ bounded derivatives. If $x \notin (x_n,x_{n+1})$ then we can therefore write

$$\theta_n^{-k+1}w(t) = \phi(t) + O(h_n^{p-k}), \quad \phi \in \mathbf{P}_{p-k}$$

and obtain, by (11a),

$$E_n(x) = O(h_n^{p+2-k}h^{k-1}) = O(h^{p+1})$$

Summing up on n, the superconvergence conclusion (11b) follows, because for a mesh point, $x_i \notin (x_n,x_{n+1})$, $1 \le n \le N$.

But, if $x \in (x_i,x_{i+1})$ then the limited smoothness of $G_j(x,t)$ allows us to conclude only that

$$E_i(x) = O(h_i^{k+m-j})\theta_i^{k-1}.$$

To obtain the error estimate (11c) for points other than mesh points, without imposing any mesh restriction, we use (31a) for the error, obtaining

$$|g_{ij}| = O(h^p) + O(h_i^k),$$

and substitute into (37b). This yields

$$|d_\pi^{(j)}(x)| = O(h^p) + O(h_i^{k+m-j}), \quad 0 \le j \le k+m-1.$$

Using the ODE (8a) which both u and u_π satisfy at the collocation points, we obtain that for the choice of $\psi_j(t)$ satisfying (34a),

$$|g_{ij}| = O(h^p) + O(h_i^{k+1}). \tag{39}$$

Hence, the leading error term when $p > k+m$ is $\mu(x)$. Now (36) yield (11c), because

$$u^{(m)}[x_{i1}, \ldots, x_{ik}, x] = \frac{1}{k!} u^{(k+m)}(x_i) + O(h_i).$$

This completes the proof.

□

Recall from [14,2] that as a consequence of (11c), there are additional "superconvergence" points, i.e. points other than mesh points where the order of convergence is higher than what is possible everywhere. In particular, any roots of $P^{(j)}(\xi)$ of (11d) in (0,1) correspond to such points for the j-th derivative of the error in (x_i, x_{i+1}), $1 \le i \le N$. We have already mentioned the roots of $P^{(m)}$ in (39). Let us also point out that $P^{(k+m-1)}(\bar{t}) = 0$ at

$$\frac{\bar{x}_i - x_i}{h_i} \equiv \bar{t} := \frac{1}{k} \sum_{l=1}^{k} \rho_l, \tag{40a}$$

so the piecewise constant $u_\pi^{(k+m-1)}$ satisfies

$$u_\pi^{(k+m-1)}(x) = u^{(k+m-1)}(\bar{x}_i) + O(h_i^2) \qquad x_i \le x \le x_{i+1},\ 1 \le i \le N \tag{40b}$$

(cf. [14,2]). If the collocation points are symmetric then $\bar{x}_i = x_{i+1/2}$.

4. Nonlinear problems

For the generally nonlinear BVP (1) we consider the method of quasilinearization. Recalling (12), the method of collocation with quasilinearization reads as follows: Given an initial approximate solution $u_\pi(x)$, *repeat* (i) solving by collocation the linearized problem

$$L[u_\pi]z(x) = -[u_\pi^{(m)}(x) - f(x,\mathbf{y}_\pi(x))] \qquad 0 < x < 1 \tag{41a}$$

$$B_0\mathbf{w}(0) + B_1\mathbf{w}(1) = -\mathbf{g}(\mathbf{y}_\pi(0),\mathbf{y}_\pi(1)) \tag{41b}$$

where $\mathbf{w}(x) := (z(x),z'(x),\ldots,z^{(m-1)}(x))$, and (ii) improving the approximate solution by

$$u_\pi(x) := u_\pi(x) + z_\pi(x) \qquad 0 \le x \le 1, \tag{41c}$$

until $|z_\pi|$ is below an error tolerance.

The quasilinearization method outlined above can be seen to be equivalent to Newton's method for solving (7). Note that, with the linear implementation advocated here, if we wish to use the same Jacobian matrix for more than one iteration (this is a *modified* Newton's method, see [3]) then we need to retain W^{-1} in some form, for each subinterval i.

The convergence results for the method of collocation and quasilinearization are contained in Theorem 13. We omit a full proof, since the nonlinear analysis is not different in principle from the usual, cf. [9,15,7]. Thus, one considers the collocation solution of the linearized BVP at the exact solution $u(x)$, for which Theorem 11 applies. This unknown-but-existing piecewise polynomial, call it say $\hat{u}_\pi(x)$, is then used as a starting guess for the quasilinearization process and the Newton-Kantorovitch conditions are verified. This yields conclusions (a) and (b) of the theorem. The proof is then completed by showing that the error in u_π is the same as that in $\hat{u}_\pi$, up to $O(h^{2k})$ terms.

References

1. U. Ascher, "On some difference schemes for singular singularly-perturbed boundary value problems", Numerische Mathematik, to appear.

2. U. Ascher, J. Christiansen and R.D. Russell, "A collocation solver for mixed order systems of boundary value problems", Math. Comp. 33 (1979), 659-679.

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

4. U. Ascher, S. Pruess and R.D. Russell, "On spline basis selection for solving differential equations", SIAM J. Numer. Anal. 20 (1983), 121-142.

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

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

7. U. Ascher and R. Weiss, "Collocation for singular perturbation problems III: Nonlinear problems without turning points", SIAM J. Scient. Stat. Comp. (1984).

8. G. Bader, "Solving boundary value problems for functional differential equations by collocation", in Proc. Workshop on Numerical Boundary Value ODEs, U. Ascher and R.D. Russell (eds.) Birkhauser (1985)

9. C. de Boor and B. Swartz, "Collocation at Gaussian points", SIAM J. Numer. Anal. 10 (1973), 582-606.

10. C. de Boor and B. Swartz, "Comment on "A comparison of global methods for linear two-point boundary value problems", Math, Comp. 31 (1977), 916-921.

11. C. de Boor, F. de Hoog and H.B. Keller, "The stability of one-step schemes for first-order two-point boundary value problems", SIAM J. Numer. Anal. 20 (1983), 1139-1146.

12. F. de Hoog and R.M.M. Mattheij, "The role of conditioning in shooting techniques ", in Proc. Workshop on Numerical Boundary Value ODEs, U. Ascher and R.D. Russell (eds.) Birkhauser (1985)

13. J. Cerutti, "Collocation for systems of ordinary differential equations", Comp. Sci. Tech. Rep. 230, Univ. Wiscinsin-Madison, 1974.

14. J. Christiansen and R.D. Russell, "Error analysis for spline collocation methods with application to knot selection", Math. Comp. 32 (1978), 415-419.

15. H.B. Keller, Numerical Solution of Two Point Boundary Value Problems, CBMS Regional Conference Series in Applied Mathematics, 24, SIAM, Philadelphia.

16. M. Lentini, M.R. Osborne and R.D. Russell, "The close relationships between methods for solving two-point boundary value problems ", SIAM J. Numer. Anal. (1985)

17. J. Paine and R.D. Russell, "Conditioning of collocation matrices and discrete Green's functions", manuscript, 1984.

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

19. R.D. Russell and J. Christiansen, "Adaptive mesh selection strategies for solving boundary value problems", SIAM J. Numer. Anal. 15 (1978), 59-80.

20. R.D. Russell and J. Shampine, "A collocation method for boundary value problems", Numer. Math. 19 (1972), 1-28.

21. G.M. Vainikko, "Convergence of the collocation method for nonlinear differential equations", USSR Comp. Math. and Math. Phys. 6 (1966), 47-58.

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

23. K. Wittenbrink, "High order projection methods of moment and collocation type for nonlinear boundary value problems", Computing 11 (1973), 255-274.