MATHEMATICS OF COMPUTATION, VOLUME 29, NUMBER 131
JULY 1975, PAGES 761–776

A Collocation Method for Two-Point Boundary Value Problems*

By J. H. Ahlberg and T. Ito

Abstract. This article is concerned with the use of collocation by splines to numerically solve two-point boundary value problems. The problem is analyzed in terms of cubic splines first and then extended to the use of quintic and septic splines. Consideration is given both to convergences as the mesh is refined and to the bandwidth of the matrices involved. Comparisons are made to a similar approach using the Galerkin method rather than collocation.

1. Introduction. Given a two-point boundary value problem

$$(Lu)(x) = p(x)u''(x) + q(x)u'(x) + r(x)u(x) = f(x, u) \quad \text{in } (0, 1), \tag{1.1}$$

$$u(0) = u(1) = 0, \tag{1.2}$$

with sufficiently smooth coefficient functions p, q and r and the forcing function f, we attempt to solve it numerically by a collocation method using spline functions of degrees three, five and seven. Collocation schemes for such a problem have been analyzed by some Russian authors who used ordinary polynomials for approximating functions [11], [12], [14], [15]; and more recently, spline functions were introduced with more desirable results [1]–[3], [7]–[10]. This paper introduces yet another variation of the method, especially in the treatment of boundary conditions for the approximating splines, and the analysis is much more straightforward than [7], [9]. Working equations are also described for immediate application.

As is well known, the more general boundary condition,

$$u(0) = a, \qquad u(1) = b, \tag{1.3}$$

can be transformed to the homogeneous case (1.2) by setting $U(x) = u(x) - a(1 - x) - bx$; and the analysis here is not affected by the new forcing function.

2. Cubic Splines. We first impose a uniform partition on the interval $[0, 1]$ as

$$x_i = ih \qquad (i = 0, 1, \dots, n + 1) \quad \text{where } h = 1/(n + 1); \tag{2.1}$$

then the cubic B-splines of Schoenberg are

Received March 18, 1974.

AMS (MOS) subject classifications (1970). Primary 65L10, 41A15; Secondary 41A25, 65N35.

Key words and phrases. Collocation, two-point boundary value problems, spline, B-spline, diagonally dominant, order of convergence, nonlinear system, Lipschitz condition, quintic splines, spline of interpolation, septic splines, bandwidth.

*This work was supported by the Office of Naval Research under Contract N00014-67-A-0191-0015 at the Division of Applied Mathematics, Brown University, Providence, Rhode Island.

$$(2.2)\quad \widetilde{B}_i(x) = \frac{1}{h^3}\cdot\begin{cases} (x - x_{i-2})^3 & [x_{i-2}, x_{i-1}], \\ h^3 + 3h^2(x - x_{i-1}) + 3h(x - x_{i-1})^2 - 3(x - x_{i-1})^3 & [x_{i-1}, x_i], \\ h^3 + 3h^2(x_{i+1} - x) + 3h(x_{i+1} - x)^2 - 3(x_{i+1} - x)^3 & [x_i, x_{i+1}], \\ (x_{i+2} - x)^3 & [x_{i+1}, x_{i+2}], \\ 0 & \text{elsewhere}, \end{cases}$$

$$(i = -1, 0, \ldots, n + 2)$$

where subintervals are extended to the outside of [0, 1] with the same mesh size h. Here we modify these functions in order to accomodate the zero boundary condition at $x = 0$ and $x = 1$ as

$$(2.3)\quad \begin{cases} B_0(x) = \widetilde{B}_0(x) - 4\widetilde{B}_{-1}(x), & B_n(x) = \widetilde{B}_n(x) - \widetilde{B}_{n+2}(x), \\ B_1(x) = \widetilde{B}_1(x) - \widetilde{B}_{-1}(x), & B_{n+1}(x) = \widetilde{B}_{n+1}(x) - 4\widetilde{B}_{n+2}(x), \\ B_i(x) = \widetilde{B}_i(x) \quad (i = 2, \ldots, n - 1), \end{cases}$$

then we obtain a basis $\{B_i(x)\}_{i=0}^{n+1}$ for the space of cubic splines that automatically satisfies the boundary condition (1.2).

Now we assume that the coefficient functions p, q and r together with the forcing term f are smooth enough so that we have a unique solution $u(x)$ of the problems (1.1)–(1.2). We let

$$(2.4)\quad \bar{u}(x) = \sum_{i=0}^{n+1} \bar{c}_i B_i(x)$$

be the cubic spline of interpolation to the true solution $u(x)$ where $\bar{c}_i$'s are constants. We also consider another spline function,

$$(2.5)\quad \widetilde{u}(x) = \sum_{i=0}^{n+1} c_i B_i(x),$$

where constants c_i's are to be determined by the following collocation conditions,

$$(2.6)\quad L\widetilde{u}(x_i) = f(x_i, \widetilde{u}(x_i)) \qquad (i = 0, 1, \ldots, n + 1).$$

Explicitly, these are

$$(2.7.1)\quad \frac{p}{h^2}[-36c_0] + \frac{q}{h}[12c_0 + 6c_1] = f \qquad (x = x_0),$$

$$(2.7.2)\quad \begin{aligned} &\frac{p}{h^2}[6c_{i-1} - 12c_i + 6c_{i+1}] + \frac{q}{h}[-3c_{i-1} + 3c_{i+1}] \\ &\qquad + r[c_{i-1} + 4c_i + c_{i+1}] = f \qquad (x = x_i,\ i = 1, 2, \ldots, n), \end{aligned}$$

$$\frac{p}{h^2}[-36c_{n+1}] + \frac{q}{h}[-6c_n - 12c_{n+1}] = f \qquad (x = x_{n+1}). \tag{2.7.3}$$

Collecting these equations, we obtain

$$Ac = \mathbf{f}(\mathbf{c}) \tag{2.8}$$

where A is an $(n + 2)$ by $(n + 2)$ matrix, c is an $(n + 2)$-dimensional vector with components c_i and f(c) is the right-hand side vector of dimension $(n + 2)$.

We now examine the property of the matrix A to establish a convergence result later. From (2.7.2) we notice, for a sufficiently small h, the coefficient of c_i dominates others in absolute value if $p(x_i)r(x_i) < 0$ since

$$\left|\frac{-12p}{h^2} + 4r\right| - \left\{\left|\frac{6p}{h^2} - \frac{3q}{h} + r\right| + \left|\frac{6p}{h^2} + \frac{3q}{h} + r\right|\right\}$$

$$= \begin{cases} \left(\frac{12p}{h^2} - 4r\right) - \left\{\left(\frac{6p}{h^2} - \frac{3q}{h} + r\right) + \left(\frac{6p}{h^2} + \frac{3q}{h} + r\right)\right\} \\ \qquad\qquad = -6r > 0 \quad (\text{if } p(x_i) > 0), \\ -\left(\frac{12p}{h^2} - 4r\right) - \left\{\left(\frac{-6p}{h^2} + \frac{3q}{h} - r\right) + \left(\frac{-6p}{h^2} - \frac{3q}{h} - r\right)\right\} \\ \qquad\qquad = 6r > 0 \quad (\text{if } p(x_i) < 0). \end{cases} \tag{2.9}$$

From (2.7.1) we have

$$\left|\frac{-36p}{h^2} + \frac{12q}{h}\right| - \left|\frac{6q}{h}\right| > 0 \tag{2.10}$$

for a sufficiently small h, and also the positive quantity on the left-hand side is $O(h^{-2})$. Naturally, the similar results hold for the last equation (2.7.3). At these endpoints we need no restriction on the sign of $p(x)$ or $r(x)$. Thus we can conclude that the matrix A in (2.8) is diagonally dominant if h is sufficiently small, and

$$p(x_i)r(x_i) < 0 \qquad (i = 1, 2, \ldots, n),$$

which is automatically satisfied if

$$p(x)r(x) < 0, \quad x \in (0, 1); \tag{2.11}$$

moreover,

$$\|A^{-1}\|_\infty \leqslant \frac{1}{6 \min_{1 \leqslant i \leqslant n} |r(x_i)|} \equiv K. \tag{2.12}$$

Now consider the quantities $L\bar{u}(x_i)$ $(i = 0, 1, \ldots, n + 1)$. If we let $O(h^d)$ $(d > 2)$ be the order of convergence of the interpolating spline $\bar{u}$, i.e., $\|u - \bar{u}\|_\infty = O(h^d)$,** then we have $\|Lu - L\bar{u}\|_\infty = O(h^{d-2})$ where $\|\cdot\|_\infty$ indicates the maximum

**For $d = 2$, the rate of convergence is in terms of the modulus of continuity of $u''(x)$.

norm in $[0, 1]$, since L is a linear second-order differential operator. At the mesh points, in particular, we write

$$(L\bar{u})(x_i) = f(x_i, \bar{u}(x_i)) + g(x_i) \qquad (i = 0, 1, \ldots, n+1)$$

where $g(x)$ is an error function with the order of magnitude $O(h^{d-2})$. In the matrix form, this becomes

$$A\bar{\mathbf{c}} = \mathbf{f}(\bar{\mathbf{c}}) + \mathbf{g} \tag{2.13}$$

where A is the same matrix as in (2.8), $\bar{\mathbf{c}}$ is the $(n+2)$-dimensional vector with components $\bar{c}_i$, $\mathbf{f}(\bar{\mathbf{c}})$ and $\mathbf{g}$ are right-hand side vectors of dimension $(n+2)$ with components $f(x_i, \bar{u}(x_i))$ and $g(x_i)$, respectively.

With these preliminary results, we can proceed to analyze the convergence behavior of the collocating spline $\tilde{u}(x)$ to the true solution $u(x)$ as the mesh size h approaches zero. At this point we assume that h is small enough so that both of the nonlinear systems of equations (2.8) and (2.13) have unique solutions. We note that the degree of nonlinearity in these equations decreases as we take a smaller mesh size. We also assume a Lipschitz condition on the forcing function:

$$|f(x, u_1) - f(x, u_2)| \leqslant L|u_1 - u_2| \quad \text{for all } x \in [0, 1] \tag{2.14}$$

where the constant L is independent of x.

Let $\mathbf{e} = (e_0, e_1, \ldots, e_{n+1})^T$, where $e_i = \bar{c}_i - c_i$ $(0 \leqslant i \leqslant n+1)$. Then subtracting (2.8) from (2.13), we have

$$A\mathbf{e} = \mathbf{g} + f(\bar{\mathbf{c}}) - f(\mathbf{c}), \tag{2.15}$$

and by the Lipschitz condition (2.14),

$$\begin{aligned} f(x_i, \bar{u}(x_i)) - f(x_i, \tilde{u}(x_i)) &= L_i[\bar{u}(x_i) - \tilde{u}(x_i)] \\ &= \begin{cases} L_i\{e_{i-1} + 4e_i + e_{i+1}\} & (1 \leqslant i \leqslant n) \\ 0 & (i = 0 \text{ or } n+1) \end{cases} \end{aligned}$$

for some constants L_i where $|L_i| \leqslant L$ $(0 \leqslant i \leqslant n+1)$. The second equality above is derived using the property of the basis functions. Now we can rewrite (2.15) as

$$A\mathbf{e} = \mathbf{g} + \tilde{L}M\mathbf{e} \tag{2.16}$$

where $\tilde{L} = \mathrm{Diag}\{L_0, L_1, \ldots, L_{n+1}\}$,

$$M = \begin{pmatrix} 0 & & & & & \\ 1 & 4 & 1 & & & \\ & 1 & 4 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 4 & 1 \\ & & & & & 0 \end{pmatrix}.$$

Hence $\mathbf{e} = A^{-1}\mathbf{g} + A^{-1}\widetilde{L}M\mathbf{e}$ when A^{-1} exists and is bounded as in (2.12), so that

$$\|\mathbf{e}\|_\infty \leqslant K\|\mathbf{g}\|_\infty + 6KL\|\mathbf{e}\|_\infty = O(h^{d-2}) + 6KL\|\mathbf{e}\|_\infty, \tag{2.17}$$

where K is a bound on $\|A^{-1}\|_\infty$ and $\|M\|_\infty = 6$. Thus if

$$1 - 6KL > 0, \tag{2.18}$$

we have $\|\mathbf{e}\|_\infty = O(h^{d-2})$. This, in turn, means

$$\sup_{x\in[0,1]} |\widetilde{u}(x) - \overline{u}(x)| = O(h^{d-2}),$$

since at any point $x \in [0, 1]$, the values $\widetilde{u}(x)$ and $\overline{u}(x)$ are determined by only a finite number of coefficients $\{c_i\}$ and $\{\overline{c}_i\}$ due to the minimal support of B-splines. Because we already assumed $\|u - \overline{u}\|_\infty = O(h^d)$, we have

$$\|u - \widetilde{u}\|_\infty \leqslant \|u - \overline{u}\|_\infty + \|\overline{u} - \widetilde{u}\|_\infty = O(h^{d-2}).$$

Thus we have proved the following theorem.

THEOREM. *Let a two-point boundary value problem have the form* (1.1)–(1.2) *where the coefficient functions $p(x)$ and $r(x)$ and the forcing function $f(x, u)$ satisfy the conditions* (2.11), (2.14), *and* (2.18). *Let $\widetilde{u}(x)$ be the collocating approximate solution in* (2.5) *where the coefficients are defined by* (2.8). *If the true solution $u(x)$ of* (1.1)–(1.2) *is smooth enough so that the cubic spline function $\overline{u}(x)$ interpolating to $u(x)$ converges to $u(x)$ in the order of $O(h^d)$, then $\widetilde{u}(x)$ converges to $u(x)$ in the order of $O(h^{d-2})$ in the maximum norm over* $[0, 1]$.

For a linear problem, i.e. when the forcing term is a function of x alone, the Lipschitz constant L and the associated matrix $\widetilde{L}$ become zero; and we have instead of (2.17)

$$\|\mathbf{e}\|_\infty \leqslant \frac{\|\mathbf{g}\|_\infty}{6 \min_{1\leqslant i\leqslant n} |r(x_i)|} = O(h^{d-2}).$$

This error estimate holds when $r(x)$ is bounded away from zero on $(0, 1)$.

If we assume $p(x)$, $q(x)$, $r(x)$ and $f(x, y)$ for a fixed y are all $C^2[0, 1]$ functions, then the solution $u(x)$ is $C^4[0, 1]$. In such a case the cubic spline of interpolation $\overline{u}(x)$ converges to $u(x)$ in the order of $O(h^4)$; i.e. $d = 4$ in the theorem, hence our collocating spline function converges to $u(x)$ in the order of $O(h^2)$.

Similarly if $u(x) \in K_2^4[0, 1]$ where K_2^m denotes the collection of all real-valued functions $u(x)$ defined on $[0, 1]$ such that $u \in C^{m-1}[0, 1]$ and such that $u^{(m-1)}(x)$ is absolutely continuous, then $d = 3½$ [13], so the order of convergence of the collocating spline function is 3/2.

3. Quintic Splines.

The analysis here proceeds exactly as in the cubic case except for some minor modifications to incorporate the added degrees of polynomials. With the uniform mesh of (2.1), we have the quintic B-splines of Schoenberg

$$(3.1)\quad \widetilde{B}_i(x) = \frac{1}{h^5}\cdot\begin{cases} (x-x_{i-3})^5 & [x_{i-3}, x_{i-2}], \\ (x-x_{i-3})^5 - 6(x-x_{i-2})^5 & [x_{i-2}, x_{i-1}], \\ (x-x_{i-3})^5 - 6(x-x_{i-2})^5 + 15(x-x_{i-1})^5 & [x_{i-1}, x_i], \\ (x_{i+3}-x)^5 - 6(x_{i+2}-x)^5 + 15(x_{i+1}-x)^5 & [x_i, x_{i+1}], \\ (x_{i+3}-x)^5 - 6(x_{i+2}-x)^5 & [x_{i+1}, x_{i+2}], \\ (x_{i+3}-x)^5 & [x_{i+2}, x_{i+3}], \\ 0 & \text{elsewhere}, \end{cases}$$

$$(i = -2, -1, \ldots, n+3),$$

where subintervals are extended to outside of $[0, 1]$ with the same mesh size h. If we modify $\{\widetilde{B}_i(x)\}_{i=-2}^{n+3}$ to

$$(3.2)\quad \begin{cases} B_{-1} = \widetilde{B}_{-1} - 26\widetilde{B}_{-2}, & B_{n-1} = \widetilde{B}_{n-1} - \widetilde{B}_{n+3}, \\ B_0 = \widetilde{B}_0 - 66\widetilde{B}_{-2}, & B_n = \widetilde{B}_n - \widetilde{B}_{n+2}, \\ B_1 = \widetilde{B}_1 - \widetilde{B}_{-1}, & B_{n+1} = \widetilde{B}_{n+1} - 66\widetilde{B}_{n+3}, \\ B_2 = \widetilde{B}_2 - \widetilde{B}_{-2}, & B_{n+2} = \widetilde{B}_{n+2} - 26\widetilde{B}_{n+3}, \\ B_i = \widetilde{B}_i \quad (3 \leqslant i \leqslant n-2), & \end{cases}$$

we obtain a basis $\{B_i(x)\}_{i=-1}^{n+2}$ for the space of quintic splines that satisfy the homogeneous boundary condition (1.2).

We let

$$(3.3)\quad \bar{u}(x) = \sum_{i=-1}^{n+2} \bar{c}_i B_i(x)$$

be the quintic spline of interpolation to the true solution $u(x)$ of our original problem (1.1)–(1.2), where $\bar{c}_i$'s are constants. Henceforth we consider a linear equation $(Lu)(x) = f(x)$ for simplicity, although a mildly nonlinear case of (1.1) can be treated as in Section 2. We also consider another spline function,

$$(3.4)\quad \widetilde{u}(x) = \sum_{i=-1}^{n+2} c_i B_i(x),$$

where constants c_i's are to be determined by the following conditions,

$$(3.5)\quad L\widetilde{u}(x_i) = f(x_i) \quad (i = 0, 1, \ldots, n+1),$$

$$(3.6)\quad h\cdot\frac{d}{dx}(L\widetilde{u}) = h\cdot\frac{df}{dx} \quad \text{at } x = 0, 1.$$

At the points near $x_0 = 0$, the collocation condition (3.5) takes the form

$$(3.7.0)\quad \begin{aligned} &\frac{p(x_0)}{h^2}[-480c_{-1} - 1440c_0] \\ &\quad + \frac{q(x_0)}{h}[80c_{-1} + 330c_0 + 100c_1 + 10c_2] = f(x_0), \end{aligned}$$

$$\begin{aligned}(3.7.1)\quad &\frac{p(x_1)}{h^2}[20c_{-1} + 40c_0 - 140c_1 + 40c_2 + 20c_3]\\ &\qquad + \frac{q(x_1)}{h}[-5c_1 - 50c_0 + 5c_1 + 50c_2 + 5c_3]\\ &\qquad + r(x_1)[c_{-1} + 26c_0 + 65c_1 + 26c_2 + c_3] = f(x_1),\end{aligned}$$

and the boundary condition (3.6) becomes

$$\begin{aligned}(3.7.2)\quad &\frac{p(x_0)}{h^2}[1680c_{-1} + 3960c_0 - 240c_1 + 120c_2]\\ &\qquad + \frac{p'(x_0) + q(x_0)}{h}[-480c_{-1} - 1440c_0]\\ &\qquad + (q'(x_0) + r(x_0))[80c_{-1} + 330c_0 + 100c_1 + 10c_2] = h \cdot f'(x_0).\end{aligned}$$

The condition at x_i $(i = n, n + 1)$ can be similarly expressed. At each inner mesh point x_i $(2 \leqslant i \leqslant n - 1)$, (3.5) becomes

$$\begin{aligned}(3.7.3)\quad &\frac{p(x_i)}{h^2}[20c_{i-2} + 40c_{i-1} - 120c_i + 40c_{i+1} + 20c_{i+2}]\\ &\qquad + \frac{q(x_i)}{h}[-5c_{i-2} - 50c_{i-1} + 50c_{i+1} + 5c_{i+2}]\\ &\qquad + r(x_i)[c_{i-2} + 26c_{i-1} + 66c_i + 26c_{i+1} + c_{i+2}]\\ &\qquad = f(x_i) \qquad (i = 2, 3, \dots, n - 1).\end{aligned}$$

Collecting (3.7.0)–(3.7.3) and three other conditions near $x = 1$ similar to (3.7.0)–(3.7.2), we obtain

$$(3.8)\qquad A\mathbf{c} = \mathbf{f}$$

where A is an $(n + 4)$ by $(n + 4)$ matrix, c is an $(n + 4)$-dimensional vector with components c_i and f is also an $(n + 4)$-dimensional vector resulting from the right-hand sides of (3.7.0)–(3.7.3). We must now examine the property of the matrix A so that we may solve the system of Eqs. (3.8) in practice. From (3.7.3) we notice, for a sufficiently small h, the coefficient of c_i dominates others in absolute value if $p(x_i)r(x_i) < 0$ since

$$\begin{aligned}(3.9)\quad &\left|\frac{-120p}{h^2} + 66r\right| - \left\{\left|\frac{20p}{h^2} + \frac{5q}{h} + r\right| + \left|\frac{40p}{h^2} + \frac{50q}{h} + 26r\right|\right.\\ &\qquad \left. + \left|\frac{40p}{h^2} - \frac{50q}{h} + 26r\right| + \left|\frac{20p}{h^2} - \frac{5q}{h} + r\right|\right\}\\ &\qquad = \begin{cases} -120r > 0 & \text{if } p > 0, r < 0,\\ 120r > 0 & \text{if } p < 0, r > 0.\end{cases}\end{aligned}$$

From (3.7.0) we have

$$\begin{aligned} c_{-1} &= \frac{1}{-480p + 80qh}\,[(1440p - 330qh)c_0 - 100qhc_1 - 10qhc_2 + h^2 f] \\ &= -\,3c_0 + O(h). \end{aligned} \tag{3.10}$$

Now we can replace c_{-1} in (3.7.1)–(3.7.2) by the right-hand side of (3.10), and the coefficients of c_1 and c_0, respectively, dominate others in the sense of (3.9). Here we need no restriction on the sign of $p(x)$ or $r(x)$. Thus we know the system (3.8) has a unique solution **c** by the diagonal dominance property and

$$\|A^{-1}\|_\infty \leqslant \frac{1}{120 \min_{2 \leqslant i \leqslant n-1} |r(x_i)|}\,. \tag{3.11}$$

This guarantees that our approximating solution $\widetilde{u}$ in (3.4) exists uniquely.

The next step is to show that $\widetilde{u}$ is close to the true solution u. If we let $O(h^d)$ be the order of convergence of the interpolating quintic spline $\bar{u}$, i.e.

$$\|u - \bar{u}\|_\infty = O(h^d), \tag{3.12}$$

we have

$$\|Lu - L\bar{u}\|_\infty = O(h^{d-2}) \quad \text{and} \quad \left\| h \cdot \frac{d}{dx}(Lu) - h \cdot \frac{d}{dx}(L\bar{u}) \right\|_\infty = O(h^{d-2}) \tag{3.13}$$

where $\|\cdot\|_\infty$ indicates the maximum norm in $[0, 1]$. If we apply the conditions (3.5)–(3.6) to $\bar{u}(x)$, instead of $\widetilde{u}$, we have

$$A\bar{\mathbf{c}} = \mathbf{f} + \mathbf{g} \tag{3.14}$$

where A and **f** are the same quantities as in (3.8), $\bar{\mathbf{c}}$ is the counterpart of **c** in (3.8), and **g** is a vector whose components are of the order $O(h^{d-2})$ by (3.13). So from (3.8) and (3.14), we have

$$\|\mathbf{c} - \bar{\mathbf{c}}\|_\infty = \|A^{-1}\mathbf{g}\|_\infty \leqslant \|A^{-1}\|_\infty \cdot \|\mathbf{g}\|_\infty = O(h^{d-2}) \tag{3.15}$$

if $r(x)$ is bounded away from zero. This implies

$$\|\widetilde{u} - \bar{u}\|_\infty = \left\| \sum (c_i - \bar{c}_i) B_i \right\|_\infty \leqslant \|\mathbf{c} - \bar{\mathbf{c}}\|_\infty \cdot \left\| \sum B_i \right\|_\infty = O(h^{d-2}), \tag{3.16}$$

since each $B_i(x)$ has the support $[(i-3)h, (i+3)h]$, except for i near endpoints where the support is a little larger, and the value of $B_i(x)$ is bounded for any h. Thus by (3.12) and (3.16) we have

$$\|u - \widetilde{u}\|_\infty \leqslant \|u - \bar{u}\|_\infty + \|\bar{u} - \widetilde{u}\|_\infty = O(h^{d-2}), \tag{3.17}$$

which gives the convergence rate for our collocating splines. For example, for the case of $u \in K_2^6[0, 1]$, we have $d = 6 - ½ = 5½$ [13] and

$$\|u - \widetilde{u}\|_\infty = O(h^{3½}). \tag{3.17$'$}$$

4. **Septic Splines.** The analysis here is again similar to the preceding ones, and we still restrict ourselves to a linear boundary value problem as in Section 3. With

the uniform mesh of (2.1), the septic B-splines of Schoenberg are

$$\widetilde{B}_i(x) = \frac{1}{h^7} \cdot \begin{cases} (x - x_{i-4})^7 & [x_{i-4}, x_{i-3}], \\ (x - x_{i-4})^7 - 8(x - x_{i-3})^7 & [x_{i-3}, x_{i-2}], \\ (x - x_{i-4})^7 - 8(x - x_{i-3})^7 + 28(x - x_{i-2})^7 & [x_{i-2}, x_{i-1}], \\ (x - x_{i-4})^7 - 8(x - x_{i-3})^7 & \\ \qquad + 28(x - x_{i-2})^7 - 56(x - x_{i-1})^7 & [x_{i-1}, x_i], \\ (x_{i+4} - x)^7 - 8(x_{i+3} - x)^7 & \\ \qquad + 28(x_{i+2} - x)^7 - 56(x_{i+1} - x)^7 & [x_i, x_{i+1}], \\ (x_{i+4} - x)^7 - 8(x_{i+3} - x)^7 + 28(x_{i+2} - x)^7 & [x_{i+1}, x_{i+2}], \\ (x_{i+4} - x)^7 - 8(x_{i+3} - x)^7 & [x_{i+2}, x_{i+3}], \\ (x_{i+4} - x)^7 & [x_{i+3}, x_{i+4}], \\ 0 & \text{elsewhere}, \end{cases} \tag{4.1}$$

$$(i = -3, -2, \ldots, n + 4).$$

We again modify $\{\widetilde{B}_i(x)\}_{i=-3}^{n+4}$ so that the homogeneous boundary condition (1.2) is automatically satisfied as

$$\begin{cases} B_{-2} = \widetilde{B}_{-2} - 120\widetilde{B}_{-3}, \\ B_{-1} = \widetilde{B}_{-1} - 1191\widetilde{B}_{-3}, \\ B_0 = \widetilde{B}_0 - 2416\widetilde{B}_{-3}, \\ B_1 = \widetilde{B}_1 - \widetilde{B}_{-1}, \\ B_2 = \widetilde{B}_2 - \widetilde{B}_{-2}, \\ B_3 = \widetilde{B}_3 - \widetilde{B}_{-3}, \\ B_i = \widetilde{B}_i \qquad (4 \leqslant i \leqslant n - 3) \end{cases} \tag{4.2}$$

and similarly for $\{B_i(x)\}_{i=n-2}^{n+3}$. In order to determine the coefficients c_i's in $\widetilde{u} = \Sigma c_i B_i(x)$, we need an extra condition besides (3.5)–(3.6), which we set

$$h^2 \cdot \frac{d^2}{dx^2}(L\widetilde{u}) = h^2 \cdot \frac{d^2 f}{dx^2} \quad \text{at } x = 0, 1. \tag{4.3}$$

For points $\{x_i\}_{i=3}^{n-2}$, the collocation condition (3.5) takes the form

$$\begin{aligned} &\frac{p(x_i)}{h^2}[42c_{i-3} + 1008c_{i-2} + 630c_{i-1} - 3360c_i + 630c_{i+1} + 1008c_{i+2} + 42c_{i+3}] \\ &\quad + \frac{q(x_i)}{h}[-7c_{i-3} - 392c_{i-2} - 1715c_{i-1} + 1715c_{i+1} + 392c_{i+2} + 7c_{i+3}] \\ &\quad + r(x_i)[c_{i-3} + 120c_{i-2} + 1191c_{i-1} + 2416c_i + 1191c_{i+1} + 120c_{i+2} + c_{i+3}] \\ &\quad \equiv \sum_{j=i-3}^{i+3} \alpha_j c_j = f(x_i) \qquad (3 \leqslant i \leqslant n - 2). \end{aligned}$$

We again have the coefficient of c_i dominating others in absolute value, as in (3.9), if $p(x_i)r(x_i) < 0$ since

$$|\alpha_i| - \sum_{j\neq i} |\alpha_j| = \begin{cases} -5040r(x_i) > 0 & \text{if } p(x_i) > 0, r(x_i) < 0, \\ 5040r(x_i) > 0 & \text{if } p(x_i) < 0, r(x_i) > 0. \end{cases} \tag{4.4}$$

At $x = x_2$, and similarly at $x = x_{n-1}$, the collocation condition is

$$\begin{aligned} &\frac{p(x_2)}{h^2}[42c_{-1} + 1008c_0 + 588c_1 - 3360c_2 + 630c_3 + 1008c_4 + 42c_5] \\ &\quad + \frac{q(x_2)}{h}[-7c_{-1} - 392c_0 - 1708c_1 + 1715c_3 + 392c_4 + 7c_5] \\ &\quad + r(x_2)[c_{-1} + 120c_0 + 1190c_1 + 2416c_2 + 1191c_3 + 120c_4 + c_5] \\ &\quad = f(x_2), \end{aligned} \tag{4.5}$$

and the coefficient of c_2 is dominant by a quantity of order $O(h^{-2})$. Also, at $x = x_1$, and at $x = x_n$, the situation is similar since

$$\begin{aligned} &\frac{p(x_1)}{h^2}[42c_{-2} + 1008c_{-1} + 630c_0 - 4368c_1 + 588c_2 + 1008c_3 + 42c_4] \\ &\quad + \frac{q(x_1)}{h}[-7c_{-2} - 392c_{-1} - 1715c_0 + 392c_1 + 1722c_2 + 392c_3 + 7c_4] \\ &\quad + r(x_1)[c_{-2} + 120c_{-1} + 1191c_0 + 2296c_1 + 1190c_2 + 120c_3 + c_4] \\ &\quad = f(x_1). \end{aligned} \tag{4.6}$$

The collocation condition $x = x_0$ and two other conditions (3.6) and (4.3) must be treated in a modified manner as before. The collocation equation is

$$\begin{aligned} &\frac{p(x_0)}{h^2}[-4{,}032c_{-2} - 49{,}392c_{-1} - 104{,}832c_0] \\ &\quad + \frac{q(x_0)}{h}[448c_{-2} + 6{,}622c_{-1} + 16{,}912c_0 + 3{,}430c_1 + 784c_2 + 14c_3] \\ &\quad = f(x_0), \end{aligned} \tag{4.7}$$

and (3.6) and (4.3) are, respectively,

$$\begin{aligned} &\frac{p(x_0)}{h^2}[23{,}520c_{-2} + 254{,}100c_{-1} + 507{,}360c_0 - 7{,}980c_{-1} + 3{,}360c_2 + 420c_3] \\ &\quad + \frac{p' + q}{h}[-4{,}032c_{-2} - 49{,}392c_{-1} - 104{,}832c_0] \\ &\quad + (q' + r)[448c_2 + 6{,}622c_{-1} + 16{,}912c_0 + 3{,}430c_1 + 784c_2 + 14c_3] \\ &\quad = h \cdot f'(x_0), \end{aligned} \tag{4.8}$$

$$\begin{aligned}
&\frac{p(x_0)}{h^2}\,[-100{,}800c_{-2} - 1{,}008{,}000c_1 - 2{,}016{,}000c_0] \\
&\qquad + \frac{2p' + q}{h}\,[23{,}520c_{-2} + 254{,}100c_{-1} + 507{,}360c_0 - 7{,}980c_1 \\
&\qquad\qquad + 3{,}360c_2 + 420c_3] \\
&\qquad + (p'' + 2q' + r)[-4{,}032c_{-2} - 49{,}392c_{-1} - 104{,}832c_0] \\
&\qquad + h\cdot(q'' + 2r')[448c_{-2} + 6{,}622c_{-1} + 16{,}912c_0 + 3{,}430c_1 \\
&\qquad\qquad + 784c_2 + 14c_3] \\
&= h^2\cdot f''(x_0).
\end{aligned} \tag{4.9}$$

From (4.7) and (4.9), we can express c_{-2} and c_{-1} in terms of c_0 for a sufficiently small h, since the matrix,

$$\begin{pmatrix} -4{,}032 & -49{,}392 & -104{,}832 \\ -100{,}800 & -1{,}008{,}000 & -2{,}016{,}000 \end{pmatrix},$$

is of rank 2. So we can insert these expressions into (4.8), and it turns out that we have an equation dominant in the coefficient of c_0 for a sufficiently small h. The rest of the analysis is exactly the same as in the quintic case, and we have a unique approximating spline $\tilde{u}(x)$ with the rate of convergence,

$$\|u - \tilde{u}\|_\infty = O(h^{d-2}); \tag{4.10}$$

or for the case of $u \in K_2^8[0, 1]$, d becomes 7½ [13] and

$$\|u - \tilde{u}\|_\infty = O(h^{5\frac{1}{2}}). \tag{4.10$'$}$$

5. Numerical Examples. In this section results of some numerical examples are shown. All computations of the collocation method were done on the Hewlett-Packard 3000 at the Lafayette College computing center using BASIC single-precision mode. Some numerical comparisons were also made at United Aircraft Research Laboratories in 1966–1967, which indicated that accuracy-wise the collocation using cubic splines compared favorably to the finite difference method. Similar results were witnessed in more difficult two-dimensional elliptic problems [5].

Example 1. A simple linear problem,

$$u'' - 100u = 0, \qquad u(0) = u(1) = 1, \tag{5.1}$$

is solved by our collocation scheme. This problem appears in [2, p. 55] and also in [10], and the exact solution is given by

$$u(x) = \cosh(10(x - ½))/\cosh 5.$$

Error is computed at nineteen interior points uniformly spaced in (0, 1), rather than checking all the intermediate values, and maximum is taken over all such points. In the tables to follow, $1.32 - n$ means 1.32×10^{-n}. The columns entitled α indicate the quantities,

$$\ln\left(\frac{\|\tilde{u}_{h_1} - u\|\max}{\|\tilde{u}_{h_2} - u\|\max}\right) \Bigg/ \ln(h_1/h_2),$$

which give an estimate on the exponent of h for our error formula. The result (see Table 1) confirms the theory developed in this present article; i.e. the rate of convergence is close to $O(h^2)$, $O(h^4)$, and $O(h^6)$, respectively, for the cubic, quintic, and septic case, and the maximum error for each fixed step size h decreases significantly as the degree of spline polynomial increases.

TABLE 1

h	*Colloc.-Cubic*		*Colloc.-Quintic*		*Colloc.-Septic*	
	max | error |	α	max | error |	α	max | error |	α
1/5	1.00 − 1	- - -	7.88 − 3	- - -	4.60 − 4	- - -
1/10	1.69 − 2	2.71	2.91 − 4	4.76	4.47 − 6	6.66
1/15	7.30 − 3	2.07	4.87 − 5	4.41		
1/20	3.93 − 3	2.15	1.53 − 5	4.03		

Example 2. Another linear problem is

$$u'' = 4u + 4\cosh 1, \qquad u(0) = u(1) = 0, \tag{5.2}$$

which is treated in [4], [7], [9] comparing the results of various numerical methods. The exact solution is

$$u(x) = \cosh(2x - 1) - \cosh 1$$

which is symmetric at $x = ½$ as in Example 1. We also note that the condition (2.11) is trivially satisfied as it is in Example 1. The result (Table 2.1) again shows consistent increase in accuracy as h is decreased, the order of error being close to $O(h^4)$. The last column of Table 2.1 is obtained by the Galerkin method using cubic splines [4], which has the same order of accuracy and a slightly larger bandwidth of the matrix than our quintic spline collocation scheme (see Section 6). The collocation scheme is seen to give three to five times more accurate solutions in this particular problem.

TABLE 2.1

h	*Colloc.-Cubic*		*Colloc.-Quintic*		*Colloc.-Septic*	*Galerkin-Cubic**
	max | error |	α	max | error |	α		
1/3	1.53 − 2	- - -	1.01 − 4	- - -	1.18 − 6	
1/5	5.23 − 3	2.09	1.34 − 5	3.95		4.23 − 5
1/7	2.63 − 3	2.05	3.44 − 6	4.05		1.71 − 5
1/9	1.58 − 3	2.03	1.22 − 6	4.13		5.80 − 6

(*): Table 3.4 of [4].

TABLE 2.2

h	*Colloc.-Quintic*	*Collatz* [9]	*Bramble-Hubbard* [9]	*Numerov* [9]
1/5	1.34 − 5	2.56 − 5	2.06 − 3	3.88 − 5
1/10	7.73 − 7	1.65 − 6	1.64 − 4	4.83 − 6

TABLE 3

h	*Colloc.-Cubic*		*Colloc.-Quintic*		*Colloc.-Septic*	*Galerkin-Cubic**
	max | error |	α	max | error |	α		
1/3	9.59 − 4	- - -	5.89 − 6	- - -	9.07 − 8	
1/4	5.20 − 4	2.13	1.92 − 6	3.89		9.16 − 6
1/6	2.29 − 4	2.02	3.79 − 7	4.00		1.72 − 6
1/8	1.28 − 4	2.02	1.23 − 7	3.91		7.71 − 7

(*): Table 1.4 of [4].

Comparison is also made between the quintic collocation computations and some discrete methods having the same order of convergence (Table 2.2). These methods are Collatz's Mehrstellenverfahren, the Bramble and Hubbard five-point scheme, and Numerov's scheme which are all referred to in [9]. Our method again compares favorably.

Example 3. Now we turn to a nonlinear problem

$$u'' = \exp(u), \qquad u(0) = u(1) = 0, \tag{5.3}$$

which has the unique solution [4], [7], [9]

$$u(x) = \ln 2 + 2 \ln \left[c \cdot \sec\left(\frac{c(x - \frac{1}{2})}{2}\right)\right], \qquad c = 1.3360556949.$$

In this problem the key hypothesis in our proof (2.18) is not satisfied since $K = +\infty$. We may circumvent this difficulty by changing (5.3) to an equivalent form,

$$u'' - u = f(x, u) = e^u - u; \tag{5.4}$$

since, then we can find the Lipschitz constant $L \leqslant 0.11$ in (2.14) and all the conditions in the Theorem in Section 2 are satisfied. It is interesting to note, however, that in all the computations we experimented, we had no difficulty in solving (5.3) directly by our collocation scheme, and they rendered exactly the same result as the modified form (5.4). This partially supports our conjecture that collocation often works even when rigorous proofs are unavailable. The computational results are summarized in Table 3 which are similar to Example 2. To solve the nonlinear system of Eqs. (2.8), we used Newton's method and terminated the iteration when the successive iterates $\mathrm{c}^{(k)}$ satisfied the following criterion

$$\max|c_i^{(k+1)} - c_i^{(k)}| \leqslant 10^{-6}\cdot h^j$$

where

$$j = \begin{cases} 2 & \text{(cubic case)} \\ 4 & \text{(quintic case)} \\ 6 & \text{(septic case).} \end{cases}$$

In all the computations performed, the method converged after two to four iterations.

Example 4. Our final example is also a nonlinear problem [4]

$$u'' = ½(u + x + 1)^3, \qquad u(0) = u(1) = 0, \tag{5.5}$$

whose exact solution is

$$u(x) = 2/(2 - x) - x - 1.$$

We may modify (5.5) to

$$u'' - 10u = ½(u + x + 1)^3 - 10u$$

in order to satisfy the condition (2.8), although they both gave equivalent results computationally. The same approach was taken to solve the nonlinear system as in Example 3 (Table 4).

TABLE 4

h	*Colloc.-Cubic*		*Colloc.-Quintic*		*Colloc.-Septic*		*Galerkin-Cubic**
	max | error |	α	max | error |	α	max | error |	α	
1/4	5.04 − 3	- - -	9.91 − 5	- - -	3.31 − 6	- - -	9.10 − 5
1/6	2.13 − 3	2.12	1.56 − 5	4.57	2.40 − 7	6.48	2.68 − 5
1/8	1.18 − 3	2.05	5.24 − 6	3.79			7.96 − 6

(*): Table 2.5 of [4].

6. Discussion. Numerical methods for solving two-point boundary value problems are considered thoroughly in Keller [6]. Among these methods are the shooting method, the well-known finite-difference method, and the integral equation method. Aside from these classical approaches there is another important class of numerical schemes, which Keller calls the function space approximation methods, that includes the Rayleigh-Ritz-Galerkin method and the collocation method. The former of these two has been studied rigorously in the past several years, especially in connection with spline-type function spaces [4], [16].

The most significant virtue of the collocation procedure is its ease in application; e.g. matrix elements of the defining equation are evaluated directly, rather than by numerical integration as in the Galerkin method, and the bandwidth of the matrix A is smaller than that of the Galerkin method when the same degree splines are used. For the collocation method, the number of nonzero terms in a row of A is equal to the number of nonzero basis functions at the corresponding mesh point, and we have the bandwidths 1, 2, and 3, respectively, for the cubic, the quintic, and the septic cases. In the Galerkin method, however, we must integrate the products of basis functions to compute elements of the defining matrix. So if we were to use the same spline functions which appear in the present paper, the bandwidths become 3, 5, and 7, respectively. In general, polynomial splines of odd degree $2n + 1$ render the bandwidth of n in the case of collocation as compared to $2n + 1$ for the Galerkin case. Thus we see that the higher-order convergence of the Galerkin method is obtained at the expense of higher-order computational complexity.

Finally we remark that extensions of the present scheme are possible in several directions. As noted at the beginning of Section 3, the mildly nonlinear problem (1.1)–(1.2) may be treated for the quintic and the septic case. Some other possible extensions are: use of higher degree splines, treatment of higher-order differential equations and partial differential equations. Some of these problems are treated in the references cited in Section 1, though their theoretical justifications are more difficult than the present argument. For two-dimensional elliptic problems, we mention the work of one of the authors [5]. In practical computations, however, it has been our experience that a collocation scheme such as the one discussed here may be applied to a wide variety of problems with satisfactory results, even when its convergence cannot be proved rigorously (see Section 5).

Division of Applied Mathematics
Brown University
Providence, Rhode Island 02912

Mathematics Department
Lafayette College
Easton, Pennsylvania 18042

1. E. L. ALBASINY & W. D. HOSKINS, "Cubic spline solutions to two-point boundary value problems," *Comput. J.*, v. 12, 1969/70, pp. 151–153. MR 39 #3710.

2. J. H. AHLBERG, E. N. NILSON & J. L. WALSH, *The Theory of Splines and Their Applications,* Academic Press, New York and London, 1967. MR 39 #684.

3. W. G. BICKLEY, "Piecewise cubic interpolation and two-point boundary problems," *Comput. J.*, v. 11, 1968/69, pp. 206–208. MR 37 #6036.

4. P. G. CIARLET, M. H. SCHULTZ & R. S. VARGA, "Numerical methods of high-order accuracy for nonlinear boundary value problems. I. One dimensional problem," *Numer. Math.*, v. 9, 1966/67, pp. 394–430. MR 36 #4813.

5. T. ITO, *A Collocation Method for Boundary Value Problems Using Spline Functions,* Doctoral Thesis, Brown University, Providence, R. I., 1972.

6. H. B. KELLER, *Numerical Methods for Two-Point Boundary-Value Problems,* Blaisdell, Waltham, Mass., 1968. MR 37 #6038.

7. T. R. LUCAS & G. W. REDDIEN, JR., "Some collocation methods for nonlinear boundary value problems," *SIAM J. Numer. Anal.*, v. 9, 1972, pp. 341–356. MR 46 #8443.

8. J. L. PHILLIPS, "The use of collocation as a projection method for solving linear operator equations," *SIAM J. Numer. Anal.*, v. 9, 1972, pp. 14–28. MR 46 #6636.

9. R. D. RUSSELL & L. F. SHAMPINE, *A Collocation Method for Boundary Value Problems,* Univ. of New Mexico Tech. Rep. 205, October 1970; Also: *Numer. Math.*, v. 19, 1972, pp. 1–28. MR 46 #4737.

10. M. SAKAI, "Spline interpolation and two-point boundary value problems," *Mem. Fac. Sci. Kyushu Univ. Ser. A*, v. 24, 1970, pp. 17–34. MR 42 #8702.

11. A. A. ŠINDLER, "Certain theorems in the general theory of approximate methods of analysis and their application to the methods of collocation, moments and Galerkin," *Sibirsk. Mat. Ž.*, v. 8, 1967, pp. 415–432 = *Siberian Math. J.*, v. 8, 1967, pp. 302–314. MR 35 #5120.

12. A. A. ŠINDLER, "The rate of convergence of an enriched collocation method for ordinary differential equations," *Sibirsk. Mat. Ž.*, v. 10, 1969, pp. 229–233 = *Siberian Math. J.*, v. 10, 1969, pp. 160–163. MR 39 #2340.

13. M. H. SCHULTZ & R. S. VARGA, "*L*-splines," *Numer. Math.*, v. 10, 1967, pp. 345–369. MR 37 #665.

14. G. M. VAĬNIKKO, "On convergence and stability of the collocation method," *Differencial'nye Uravnenija*, v. 1, 1965, pp. 244–254 = *Differential Equations*, v. 1, 1965, pp. 186–194. MR 32 #8514.

15. G. M. VAĬNIKKO, "On convergence of the collocation method for nonlinear differential equations," *Ž. Vyčisl. Mat. i Mat. Fiz.*, v. 6, 1966, no. 1, pp. 35–42 = *U. S. S. R. Comput. Math. and Math. Phys.*, v. 6, 1966, no. 1, pp. 47–58. MR 33 #5129.

16. R. S. VARGA, *Functional Analysis and Approximation Theory in Numerical Analysis*, Conference Board of the Mathematical Sciences Regional Conference Series in Appl. Math., no. 3, SIAM, Philadelphia, Pa., 1971. MR 46 #9602.