MATHEMATICS OF COMPUTATION, VOLUME 31, NUMBER 140
OCTOBER 1977, PAGES 916–921

Comments on the Comparison of Global Methods for Linear Two-Point Boundary Value Problems

By Carl de Boor* and Blair Swartz**

Abstract. A more careful count of the operations involved in solving the linear system associated with collocation of a two-point boundary value problem using rough splines reverses results recently reported by others in this journal. In addition, it is observed that the use of the technique of "condensation of parameters" can decrease the computer storage required. Furthermore, the use of a particular highly localized basis can also reduce the setup time when the mesh is irregular. Finally, operation counts are roughly estimated for the solution of certain linear systems associated with two competing collocation methods; namely, collocation with smooth splines and collocation of the equivalent first order system with continuous piecewise polynomials.

In a recent paper [1] in this journal, R. D. Russell and J. M. Varah carry out a comparison of various global methods for the numerical solution of the $(2m)$th order linear two-point boundary value problem

$$(1)\qquad Lu(x) := \sum_{i=0}^{m} (-D)^i(a_i(x)D^iu(x)) = f(x), \qquad a \leqslant x \leqslant b,$$

$$(2)\qquad D^iu(a) = D^iu(b) = 0, \qquad 0 \leqslant i < m.$$

We wish to take exception to their account of the computational effort required to solve (1)–(2) approximately by collocation at Gauss points with C^{2m-1} piecewise polynomials of degree less than $2n$.,

Suppose we collocate at

$$r := 2n - 2m$$

Gauss points per interval or polynomial piece, using splines of order $2n$ in $C^{2m-1}[a, b]$ with N polynomial pieces. Then, according to [2, p. 605, replacing m there by $2m$], the block structure of the linear system to be solved is

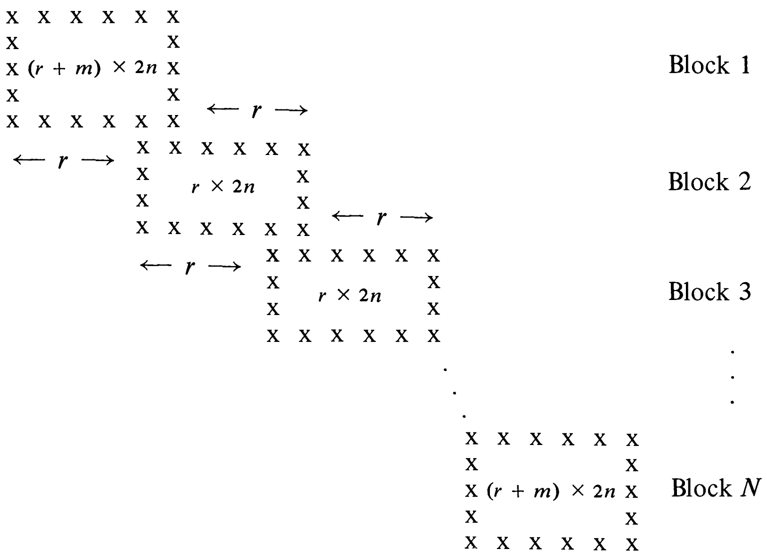
Received January 22, 1976; revised November 22, 1976.

AMS (MOS) subject classifications (1970). Primary 65L10.

*Work supported by the U.S. Army under contract No. DAAG29-75-C-0024.

**Work supported by the U.S. Energy Research and Development Administration under contract W-7405-ENG-36.

(3)


$(r + m) \times 2n$
Block 1
$\leftarrow r \rightarrow$
$\leftarrow r \rightarrow$
$r \times 2n$
Block 2
$\leftarrow r \rightarrow$
$\leftarrow r \rightarrow$
$r \times 2n$
Block 3
$(r + m) \times 2n$
Block N

if we use the basis of appropriate B-splines in their natural order. Russell and Varah view this as a block tridiagonal system (see (4.1)–(4.2) of [1]), with each block of size $r \times r$ (the first and last row of blocks being of somewhat different size because of the boundary conditions). They do take into account that the last $r/2$ rows of each of their subdiagonal blocks and the first $r/2$ rows of each of their superdiagonal blocks are zero and assume that no pivoting is required. Consequently, they obtain

$$(13r^3/12 + 2r^2)N \tag{4}$$

for the number of mulitplications/divisons necessary to solve the system.

This number is about right when $n = 2m$, i.e., $r = n$ (see (6) below). But, for $n > 2m$, i.e., for $r > n$, they treat $2(r - n)$ zero entries in each row as if they were nonzero; and therefore come to the incorrect conclusion that, for large n, collocation is twice as expensive as least squares.

It turns out to be more efficient not to impose a block tridiagonal structure on (3), but rather simply to carry out Gauss elimination *with* partial pivoting, paying attention to the zero structure of (3). This requires r steps of Gauss elimination in the first block, of size $(r + m) \times 2n$, after which the remaining m equations in that block do not involve the first r unknowns. Combining them with the r equations of the second block gives again a block of size $(r + m) \times 2n$ in which we carry out r steps of Gauss elimination with partial pivoting. The m equations of this block not used as pivotal equations now do not involve the first $2r$ unknowns, hence, together with the r equations of the next block, form again a block of size $(r + m) \times 2n$. Proceeding in this manner, we reach eventually a final block of size $2n \times 2n$, in which we carry out the full number of $2n - 1$ elimination steps. (See [3] for more computational detail and appropriate software.)

The required work is then that involved in performing

$$(N-1)\quad (r \text{ steps of G.E. for a } (r+m)\times 2n \text{ matrix})$$
$$+ \text{ G.E. for a } 2n \times 2n \text{ matrix.}$$

The required number of multiplications/divisions is

$$(r^3/3 + 3r^2m/2 + 2rm^2)N + \text{lower order terms.}$$

In particular, the necessary number of multiplications/divisions is given by

$$\begin{aligned} &[(19/12)r^3 + O(r^2)]N && \text{when } r = 2m = n, \\ &[r^3/3 + O(r^2)]N = [8n^3/3 + O(n^2)]N && \text{when } r \gg 2m, \text{ i.e., } n \gg 2m, \end{aligned} \tag{5}$$

if partial privoting is required.

If no pivoting for size is used (as is assumed in [1]), then one can take additional advantage of the fact that the last m equations of each block, after elimination in that block, involve only $2n - r$ unknowns. When these equations are adjoined to the next block, the first m steps of Gauss elimination for the block involve only $2n - r$ rather than $2n$ columns. This saves $r\,\Sigma_{i=1}^{m}(m + r - i)$ multiplications per block, even when partial pivoting is used in the remainder of each block. In particular, the necessary number of multiplications/divisions is given by

$$\begin{aligned} &[(23/24)r^3 + O(r^2)]N && \text{when } r = 2m = n, \\ &[r^3/3 + O(r^2)]N \sim 8n^3N/3 && \text{when } r \gg 2m, \text{ i.e., } n \gg 2m, \end{aligned} \tag{6}$$

if pivoting is avoided where it would produce fill-in.

If one now follows [1] in ignoring the crucially important constants in the order of convergence rates, then the conclusion on p. 1018 of [1] would have to be reversed to say that, for $n \gg 2m$, collocation takes about 2/3 (i.e., $(8/3)n^3N$ rather than $4n^3N$) of the computing time required for Galerkin and least squares when comparing "equal" global errors $O(h^{2n})$. This kind of comparison becomes even more lopsided if we take into account the $O(h^{4(n-m)})$ superconvergence at the knots [2, Theorem 4.1], i.e., if we regard the whole collocation process as a difference scheme for knot values, and interpolate [2, pp. 601–602] if a global approximation is really wanted. In this case it would be reasonable for $n \gg m$ to compare collocation results for $n/2$ with those of Galerkin or least squares for n. Now, collocation takes $(8/3)(n/2)^3N$ multiplications *vs* the $4n^3N$ required for Galerkin or least squares.

Finally, two small points: The proof of (3.7) on p. 1011 of [1] is considerably shorter than the corresponding proof in [2], due to the fact that the main difficulty in the proof in [2], viz. the fact that the functions $(\varphi q_i)^{(k)}$ (in the terminology of [1]) can be bounded appropriately in terms of local mesh sizes, is taken entirely for granted in [1]. The authors may have done this because they really only considered a ***uniform*** mesh throughout without ever saying so. This guess is supported by their remark on p. 1014 that "this work", i.e., the evaluation of the basis and its derivatives at the collocation points, "does not depend on N (i.e. on h) since the evaluations are always at the Gaussian points, and we assume these coefficients can be stored beforehand, no matter what h is." The other observation, as pointed out to us by John Rice, notes the fact that the authors have chosen, unnecessarily, to evaluate the differential equation at each collocation point repeatedly—once for each relevant basis function. See [1, p. 1014], where they say "for collocation, each matrix element involves an evaluation of (1.1)."

Stimulated by [1] to reconsider the problems in approximately solving (1)–(2), we would like to add a few remarks which are, in effect, suggestions for further work.

Our first comment considers the attractions of another basis—really, another set of unknowns—in connection with the collocation problem we initially considered. Beginning with the value of the piecewise polynomial and its first $2m - 1$ derivatives at each mesh point, we pick $2n - 4m = r - 2m$ additional parameters in each interval so as to obtain the $2m(N + 1) + (r - 2m)N = 2m + Nr$ independent parameters required to describe the general $2n$th order piecewise polynomial in C^{2m-1} with the given $(N - 1)$ interior mesh points. These additional "local" parameters could be the limiting values of the $2m$th through $(n - 1)$st derivatives at the two endpoints of each interval; the corresponding basis is easily derived from the Hermite basis. Other choices, such as the value of the function or the value of the $2m$th derivative at some of the collocation points, seem equally appropriate. Each such choice makes it possible to use in each interval the *same* information (properly scaled) about the basis functions even when a nonuniform mesh is used. The setup time is then given by

$$rN(\text{cost of evaluating the coefficients of the DE at one point} + 2mnM).$$

Further, one can use "condensation of parameters" as practiced in finite element calculations. Since the $r - 2m$ additional parameters per interval only involve that interval, they can be eliminated locally as part of the process of setting up the linear system to be solved, a strategy offering some potential for parallel processing. This procedure might or might not be stable. In any event, it would require $(r - 2m)$ steps of Gauss elimination with partial pivoting (for an $r \times 2n$ matrix) for each interval and would leave, finally, an almost block diagonal linear system, with each block (except for the first and last) of size $2m \times 4m$ instead of the original $(2n - 2m) \times 2n$. The unknowns in this final system are the approximate values of the function and its first $2m - 1$ derivatives at the mesh points, i.e., the quantities of most interest since they are $O(h^{4(n-m)})$ accurate. If, as would be reasonable, nothing else about the approximate solution is required, then this approach would make the storage requirements essentially independent of n.

The reader may have discovered that the structure of the linear system in the previous paragraph coincides with (3), the structure associated with the B-spline basis. We should note, then, that no similar savings in setup time for the B-spline system will be found in the case of a nonuniform mesh, since the required information about the basis functions must then be computed for each interval. And, finally, we observe that the application of the last paragraph's strategy to the solution of the linear system yields the following changes in the operation counts (5) and (6):

$m(r - 2m)(r + 6m)N/2$ savings over (5) (partial pivoting throughout);

$m(r - 2m)(r + m)N/2$ increase over (6) (pivoting partially avoided).

The second comment concerns collocation of (1)–(2) at the simple knots of smooth splines of order $2n + 2m$; a possibility not considered in [1]. According to [2], the order of accuracy attained is that associated with knot interpolation by smooth splines of order $2n$. The relevant linear system, assuming a B-spline basis, is about $N \times N$ with bandwidth $2n + 2m - 1$. The work involved in its solution is

$$O(n^2)N \tag{7}$$

as n becomes large; compare with the $O(n^3)N$ operations for all methods considered in [1]. Galerkin or least squares with these same smooth splines yields matrices of twice the bandwidth which should take about four times as much computational effort. Incidentally, assuming an irregular mesh, the bulk of the work in this approach is surely that involved in setting up the linear system.

The third comment concerns the computational work involved in the approach of Weiss [5] and Russell [4], who advocate (by example, at least) using the usual $2m$ first order equations $v' = Av + g$ equivalent to (1)–(2). Collocating at n Gauss points in each of the N intervals using a continuous piecewise polynomial of order $n + 1$ for each component, one obtains $O(h^{2n})$ accuracy at each mesh point (see Cerutti [8] to cover certain situations–like our example–not analyzed by [5] or by [4]). At first glance, it appears surprising that any resulting linear system could be solved in less than mn^3N operations in the case of an irregular mesh. Nevertheless, suppose the unknowns are of Runge-Kutta type; i.e., in each interval we represent the ith component of v by the values, $(v'_{ij})_{j=1}^n$, of its derivative at the n collocation points together with, say, the value of v_i at the midpoint. The linear system then consists of $2mnN$ collocation equations, $2m(N - 1)$ equations expressing the continuity of the piecewise polynomials at the mesh points, and the $2m$ boundary conditions. Now, using some initial preprocessing which is independent of the local mesh size, one may drop the rows corresponding to collocation of the first $2m - 1$ differential equations and express the $(v'_{ij})_{i=1,\ j=1}^{2m-1\ \ n}$, in the rows corresponding to (1)–(2), in terms of the $(v'_{2m,j})_1^n$ and the midpoint values using $O(mn^2)$ operations per interval. That is to say, one can locally eliminate most of the extra $(2m - 1)nN$ variables which were introduced, in the first place, by going over to the first order system. The block structure of the resulting linear system for the remaining $(2m + n)N$ unknowns is such that

$$[cn^3 + O(n^2)]N, \quad c \text{ independent of } m,$$

operations can suffice to solve it. On the other hand, we have no idea whether this numerical process is stable.

Following the publication of [1] and our composition of these comments, Dr. Russell kindly sent us manuscripts of [6] and [7]. Results (6) above together with an explicit estimate for (7) are among many other conclusions reached in these papers. We divert the diligent to their digestion.

Mathematics Research Center
University of Wisconsin
Madison, Wisconsin 53706

Theoretical Division
T-7, MS-233
Los Alamos Scientific Laboratory
University of California
Los Alamos, New Mexico 87545

1. R. D. RUSSELL & J. M. VARAH, "A comparison of global methods for linear two-point boundary value problems," *Math. Comp.*, v. 29, 1975, pp. 1007–1019. MR 52 #9622.

2. C. DE BOOR & B. SWARTZ, "Collocation at Gaussian points," *SIAM J. Numer. Anal.*, v. 10, 1973, pp. 582–606. MR 51 #9528.

3. C. DE BOOR & R. WEISS, *SOLVEBLOK: A Package for Solving Almost Block Diagonal Linear Systems*, MRC Tech. Report #1625, Mathematics Research Center, University of Wisconsin, Madison, Wisconsin, 1976.

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

5. R. WEISS, "The application of implicit Runge-Kutta and collocation methods to boundary-value problems," *Math. Comp.*, v. 28, 1974, pp. 449–464. MR 49 #6627.

6. R. RUSSELL, "Efficiences of B-spline methods for solving differential equations," *Proc.* 1975 *Manitoba Conference on Num. Math. and Computing*, pp. 599–617.

7. 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.

8. J. H. CERUTTI, *Collocation for Systems of Ordinary Differential Equations*, Tech. Reporl No. 230, Dept. of Computer Sciences, University of Wisconsin, Madison, Wisconsin, 1974; to appear in *BIT*.