On the Construction of Discrete Approximations to Linear Differential Expressions

By C. Ballester and V. Pereyra*

Introduction. When solving differential equations numerically by means of finite differences it is often necessary to obtain formulae for approximating linear combinations of derivatives. While classically these formulae have been given in terms of linear combinations of differences of the function at nodal points it is a current trend to consider directly formulae in terms of ordinates. Moreover, for use on a high-speed computer it is more convenient to be able to generate these formulae as needed instead of having to store them in the form of a table.

The main objective of this note is to describe an efficient algorithm for generating discrete approximations to linear differential expressions in terms of ordinates. This is done in Section 2 after the problem and its analytic solution have been stated in Section 1. In Section 3 we give two applications. One of them takes advantage of one of the features of the procedure which, in certain cases, makes the use of rational arithmetic quite simple.

1. Discrete Approximations to Linear Differential Expressions. Consider the mth-order homogeneous linear differential expression

$$L[y] = \sum_{\nu=0}^{m} f_\nu(x) y^{(\nu)} \tag{1}$$

in $y(x)$ with given continuous coefficients $f_\nu(x)$.

The linear combination

$$A(x) = \sum_{r=0}^{n} C_r y(x + \alpha_r h) \tag{2}$$

where C_r and α_r are constants, is called a *discrete approximation* of order p for (1) at the point $x = x_i$ if for any sufficiently differentiable function $y(x)$ in the interval containing the points x_i, $(x + \alpha_r h)$ $(r = 0, \cdots, n)$ the Taylor expansion of $A(x) - L[y]$ (x) at x_i has its first nonzero term for $y^{(q)}(x_i)$, $q = m + p$.

It is shown in [3, pp. 161–162], that for any given $s \geqq 1$ and an arbitrary choice of $n + 1$ distinct points $\xi + \alpha_r h$ $(r = 0, \cdots, n)$, $n = m + s$, $n + 1$ quantities C_r can be found such that for every function $y(x)$ with $n + 1$ continuous derivatives

$$\sum_{r=0}^{n} C_r y(\xi + \alpha_r h) - L[y](\xi) = \frac{h^{s+m}}{(s+m)!} \sum_{p=0}^{n} \alpha_p{}^n C_p y^{(n)}(\eta) . \tag{3}$$

The coefficients C_k satisfy the Vandermonde system of equations

Received October 25, 1966. Revised January 18, 1967.

* Sponsored in part by the Mathematics Research Center, United States Army, Madison, Wisconsin, under Contract No. DA-11-022-ORD-2059, and in part by NASA Research Grant NGR-50-022-028.

$$\sum_{r=0}^{n} \alpha_r{}^k C_r = \frac{k!}{h^k} f_k(\xi) \quad \text{for} \quad 0 \leqq k \leqq m\,, \tag{4}$$
$$= \quad 0 \qquad \text{for} \quad m+1 \leqq k \leqq n\,.$$

There are applications in which several differential expressions of the type (1) have to be approximated by means of discrete formulae of diverse orders and with different configurations of nodal points (cf. Volkov [12], Pereyra [9]). In those cases it is of interest to solve the system of equations (4) in an efficient manner.

Closed formulae for the elements of the inverse of a Vandermonde matrix are well known (see for instance Gautschi [4], Macon and Spitzbart [8]). Traub [11] presents two algorithms for inverting Vandermonde matrices. The first of them seems to be the most efficient (in the number of operations required) of all algorithms known to us. By using the special structure of the matrix, Traub is able to reduce the number of operations to $\frac{1}{2}n(7n - 9)$ multiplications (divisions are counted as equal to multiplications) and $5n(n - 1)/2$ additions. Unfortunately, it is not clear to us how to obtain from this a similarly efficient method for solving systems of equations of the form (4) (with say $\sim$1/3 less operations).

Lynnes and Moler [7] discuss an algorithm based on Neville's formula which solves a Vandermonde system with a number of operations proportional to n^3.

The aim of this note is to provide an algorithm for solving Vandermonde systems of linear equations which applies directly to the solution of (4) and for which the number of operations is proportional to n^2.

2. Solution of Vandermonde System of Linear Equations. Given the $n + 1$ real and distinct numbers $(\alpha_0, \cdots, \alpha_n)$, let

$$\mathbf{V}(\alpha_0, \cdots, \alpha_n) = (v_{ij}) \qquad (i, j = 0, \cdots, n) \tag{5}$$

be a Vandermonde matrix, i.e.

$$v_{ij} = \alpha_j{}^i\,. \tag{6}$$

We will derive an algorithm for solving the system of linear equations

$$\mathbf{V}(\alpha_0, \cdots, \alpha_n)\mathbf{x} = \mathbf{b}\,, \tag{7}$$

where $\mathbf{b} = (b_0, \cdots, b_n)$ is given. This algorithm will take into account the special structure of the matrix $\mathbf{V}$.

We claim that the factorization of the matrix $\mathbf{V}$ as a product of an upper and a lower triangular matrix can be given explicitly and in a very simple fashion. Observe that this factorization is possible, since all the principal minors of $\mathbf{V}$ are also of the Vandermonde type and, since the α_i are distinct, this implies that those minors are nonzero.

Let us consider the n bidiagonal matrices $\mathbf{L}^{(i)} = (l_{jk}^{(i)})$ whose nonzero elements are

$$l_{j,j}^{(i)} = 1 \qquad (j = 0, \cdots, n) \tag{8}$$
$$l_{j,j-1}^{(i)} = -\alpha_i \qquad (i = 0, \cdots, n-1)\,, \qquad (j = i+1, \cdots, n)\,.$$

THEOREM. *Premultiplication of the system* (7) *by the matrices* $\mathbf{L}^{(0)}$, $\mathbf{L}^{(1)}$, $\cdots$, $\mathbf{L}^{(n-1)}$ *reduces it to upper triangular form. Moreover, this triangular form can be ex-*

plicitly written as $\mathbf{U} = (u_{ij})$ *with*

$$u_{0j} = 1 \qquad (j = 0, \cdots, n)$$
$$u_{ij} = 0 \qquad (i > j)$$
$$u_{ij} = \prod_{s=0}^{i-1} (\alpha_j - \alpha_s) \qquad (1 \leqq i \leqq j \leqq n) .$$

This last equation can also be written as

$$u_{ij} = (\alpha_j - \alpha_{i-1}) u_{i-1,j} \qquad (1 \leqq i \leqq j \leqq n) .$$

Proof. The proof is by induction. Call $\mathbf{V}^{(0)} = \mathbf{V}(\alpha_0, \cdots, \alpha_n)$,

$$\mathbf{V}^{(i+1)} = \mathbf{L}^{(i)} \mathbf{V}^{(i)} . \tag{9}$$

Thus

$$\mathbf{V}^{(1)} = \mathbf{L}^{(0)} \mathbf{V}^{(0)} = (v_{ij}^{(1)}) \quad \text{where} \quad v_{ij}^{(1)} = \alpha_j^{i-1} (\alpha_j - \alpha_0)$$

for $i \geqq 1$, and $v_{0j}^{(1)} = 1$ $(j = 0, \cdots, n)$.

In particular, $v_{i0}^{(1)} = 0$ for $i \geqq 1$ and all the elements of the first column but the first one have been eliminated. We will show now that, more generally, $\mathbf{V}^{(k)}$ has the following form

$$\begin{aligned} v_{ij}^{(k)} &= v_{ij}^{(k-1)} && (0 \leqq i < k; 0 \leqq j \leqq n) , \\ v_{ij}^{(k)} &= \alpha_j^{i-k} \prod_{s=0}^{k-1} (\alpha_j - \alpha_s) && (k \leqq i \leqq n; 0 \leqq j \leqq n) . \end{aligned} \tag{10}$$

Formula (9) shows that (10) is true for $k = 1$. Assume that it is valid for k, $1 < k < n$. It is clear that multiplication by $\mathbf{L}^{(k)}$ does not disturb the first $k + 1$ rows and k columns of $\mathbf{V}^{(k)}$. On the other hand, for $k < i \leqq n$, $k \leqq j \leqq n$ we have

$$v_{ij}^{(k+1)} = v_{ij}^{(k)} - \alpha_k v_{i-1,j}^{(k)} , \tag{11}$$

and from (10)

$$\begin{aligned} v_{ij}^{(k+1)} &= \alpha_j^{i-k} \prod_{s=0}^{k-1} (\alpha_j - \alpha_s) - \alpha_k \alpha_j^{i-k-1} \prod_{s=0}^{k-1} (\alpha_j - \alpha_s) \\ &= \alpha_j^{i-k-1} (\alpha_j - \alpha_k) \prod_{s=0}^{k-1} (\alpha_j - \alpha_s) = \alpha_j^{i-k-1} \prod_{s=0}^{k} (\alpha_j - \alpha_s) \end{aligned}$$

which is (10) for the step $k + 1$. If we put $k = n$ in (10) we obtain

$$v_{0,j}^{(n)} = 1 , \qquad v_{i,j}^{(n)} = \prod_{s=0}^{i-1} (\alpha_j - \alpha_s) \qquad (1 \leqq i, j \leqq n) \tag{12}$$

and it is clear that all elements below the main diagonal are zero and thus

$$\mathbf{U} = (u_{ij}) = \mathbf{L}\mathbf{V} = \left(\prod_{i=0}^{n-1} \mathbf{L}^{(i)} \right) \mathbf{V}(\alpha_0, \cdots, \alpha_n) \tag{13}$$

is an upper triangular matrix. Since $\mathbf{L}$ has ones on the main diagonal, this is the unique decomposition of $\mathbf{V}$ in terms of upper and lower triangular matrices having that property (cf. Householder [6]).

Observe that $\mathbf{L}$ has a very simple structure:

$$l_{i,i-k} = (-1)^k \sigma_{n,k}\,, \qquad (k = 1, \cdots, i; i = 2, \cdots, n)$$

where $\sigma_{n,k}$ is the kth elementary symmetric function of the (α_j). These $(n-1)$ numbers can be constructed by the recursion

$$\sigma_{m,j} = \sigma_{m-1,j} + \alpha_m \sigma_{m-1,j-1} \qquad (m = 2, \cdots, n; j = 1, \cdots, m)$$
$$\sigma_{m,0} = 1\,, \qquad \sigma_{11} = \alpha_1\,.$$

From (12) it follows that $\mathbf{U}$ can be constructed by means of the recursion formula

$$u_{0j} = 1\,, \qquad u_{ij} = (\alpha_j - \alpha_{i-1}) u_{i-1,j} \qquad (1 \leqq i \leqq j \leqq n)\,, \tag{14}$$

while the new right-hand side $\tilde{\mathbf{b}} = \mathbf{Lb}$ is to be obtained from:

$$\mathbf{b}^{(0)} = \mathbf{b}\,, \qquad b_j^{(i)} = b_j^{(i-1)} - \alpha_{i-1} b_{j-1}^{(i-1)} \qquad (1 \leqq i \leqq j \leqq n) \tag{15}$$
$$\tilde{\mathbf{b}} = \mathbf{b}^{(n)}\,.$$

Once (14) and (15) have been computed the $\mathbf{x}$ in (7) can be obtained by the standard backward substitution

$$x_s = \left(\tilde{b}_s - \sum_{i=s+1}^{n} u_{si} x_i\right) \Big/ u_{ss}\,. \tag{16}$$

If we consider division times as equivalent to multiplication times, then the algorithm (14)–(16) takes $\frac{1}{2}(3n+2)(n+1)$ multiplications and the same number of additions to produce the solution $\mathbf{x}$ of the system (7) with arbitrary right-hand side $\mathbf{b}$.

In Table 1 we give the number of operations corresponding to Traub's algorithm I (see [11]), and that of Lynnes and Moler [7]. In the former one we count also the operations involved in computing $\mathbf{x} = \mathbf{V}^{-1}\mathbf{b}$, after $\mathbf{V}^{-1}$ has been formed. Here $\mathbf{V}$ is a $n \times n$ matrix.

TABLE 1

Traub (I)	*Lynnes & Moler*	B-P
$\times$ $9n(n-1)/2$	$n(n+1)(n+2)$	$\frac{1}{2}n(3n-1)$
$+$ $\frac{1}{2}n(7n-5)$	$4n(n+1)(n+2)$	$\frac{1}{2}n(3n-1)$

A warning should be given about the possible instability of all these algorithms. This is expected since Vandermonde matrices are known to be very ill conditioned (cf. [4]). For our particular algorithm we refer the reader to the comments of Wilkinson [13, Chapter 3, §26, p. 108] on the errors that occur when inverting triangular systems of equations by elimination.

3. Applications. In applying the method of iterated deferred corrections (cf. [9]) to the solution of the two-point boundary value problem

$$y'' = f(x, y)\,, \qquad y(a) = \alpha\,, \qquad y(b) = \beta\,, \tag{17}$$

it is necessary to construct discrete approximations to linear differential expressions like

$$L[y](x_i) = \sum_{j=1}^{N} h^{2j+2} \frac{2}{(2j+2)!} y^{(2j+2)}(x_i)\,, \qquad N = 1, 2, \cdots, \tag{18}$$

with orders $2N + 4$ in h at all the nodal points $x_i = a + ih$, $i = 1, \cdots, n - 1$, $n = (b - a)/h$.

It is clear from our definition that what we need are expressions of the form (2) of order $p = 2$, that is $q = 2N + 4$. Since we would like to use values of $y(x)$ only at the nodal points, it is clear that we can use symmetrical formulae if we stay far enough from the boundary. For points close to the boundary it will be necessary to use unsymmetrical formulae. Let us examine the simplest case, $N = 1$. A symmetrical formula with five points will give the required accuracy for $y^{(4)}$ and it can be obtained from Bickley's table [1], [2]:

$$L[y](x_j) = \frac{h^4}{12} y^{(4)}(x_j) = \frac{1}{12} \sum_{i=0}^{4} \binom{4}{i} (-1)^i y(x_j + (i-2)h) - \frac{h^6}{72} y^{(6)}(\xi) \tag{19}$$

$$(j = 2, \cdots, n - 2)$$

or else it can be generated by solving a system of five linear equations like (7), with

$$\alpha_i = i - 2\,, \qquad b_i = 2\delta_{i2} \qquad (i = 0, \cdots, 4)\,. \tag{20}$$

For $j = 1$ and $j = n - 1$ we have to use unsymmetrical six-point formulae which still can be found in Bickley's table. If more terms in (18) are desired, then the boundary situation becomes more involved and the number of different formulae grows steadily. It is in this case that the use of our generating procedure becomes more advantageous, since by giving a few parameters we can obtain all the necessary coefficients quite rapidly. This is also true for more complicated situations like those arising in the solution of boundary value problems for partial differential equations.

As a second application of our procedure we have generated the coefficients for the n point approximations to the mth derivative of a function $y(x)$, for $m = 1(1)10$, $n = m + 1(1)11$ at every nodal point. This table superposes and complements that of Bickley, and a copy of it has been deposited in the UMT files. In the Microfiche Appendix we reproduce a part of it.

The procedure of Section 2 is particularly well suited for this purpose, since in this case the α_i are integers and the triangular decomposition does not change this situation. This can be clearly seen in formula (12). This means that only the backward substitution will have to be carried out in rational arithmetic in order to obtain the exact values of the coefficients. Another interesting feature, also shown in (12) is that the factorization tends to decrease the values of the entries. This has been proved to be of critical importance in the exact inversion of matrices with integer coefficients (cf. Rosser [10]), since otherwise the number of digits in intermediary calculations may grow beyond the word size of the computer being used.

Mathematics Research Center
University of Wisconsin
Madison, Wisconsin 53706

1. M. Abramowitz & I. A. Stegun, (Editors), *Handbook of Mathematical Functions*, Dover, New York, 1965.
2. W. G. Bickley, "Formulae for numerical differentiation," *Math. Gazette*, v. 25, 1941, pp. 19–27. MR **2**, 240.
3. L. Collatz, *The Numerical Treatment of Differential Equations*, 3rd. ed., Springer-Verlag, Berlin, 1960. MR **22** #322.
4. W. Gautschi, "On inverses of Vandermonde and confluent Vandermonde matrices," *Numer. Math.*, v. 4, 1962, pp. 117–123. MR **25** #3059.
5. R. T. Gregory, "A method for deriving numerical differentiation formulas," *Amer. Math. Monthly*, v. 64, 1957, pp. 79–82. MR **18**, 767.
6. A. Householder, *The Theory of Matrices in Numerical Analysis*, Blaisdell, New York, 1964. MR **30** #5475.
7. J. N. Lynnes & C. B. Moler, "Van der Monde systems and numerical differentiation," *Numer. Math.*, v. 8, 1966, pp. 458–464.
8. N. Macon & A. Spitzbart, "Inverses of Vandermonde matrices," *Amer. Math. Monthly*, v. 65, 1958, pp. 95–100. MR **21** #2129.
9. V. Pereyra, "On improving an approximate solution of a functional equation by deferred corrections," *Numer. Math.*, v. 8, 1966, pp. 376–391. (Also Stanford Univ. Tech. Rep. CS29, 1965.)
10. J. B. Rosser, "A method of computing exact inverses of matrices with integer coefficients," *J. Res. Nat. Bur. Standards*, v. 49, 1952, pp. 349–358. MR **14**, 1128.
11. J. F. Traub, "Associated polynomials and uniform methods for the solution of linear problems," *SIAM Rev.*, v. 8, 1966, pp. 277–301.
12. E. A. Volkov, "A method for improving the accuracy of grid solutions of the Poisson equation," *Vyčisl. Mat.*, v. 1, 1957, pp. 62–80; English transl., *Amer. Math. Soc. Transl.*, 2, v. 35, 1964, pp. 117–136. MR **22** #5131.
13. J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Prentice-Hall, Englewood Cliffs, N. J., 1963. MR **28** #4661.