MATHEMATICS OF COMPUTATION
VOLUME 50, NUMBER 182
APRIL 1988, PAGES 481-499

# Error Bounds for Linear Recurrence Relations*

By F. W. J. Olver

**Abstract.** Recurrence relations of the form

$$a_r p_{r+1} = b_r p_r + c_r p_{r-1}$$

are examined in two cases: (A) oscillatory systems, for which $b_r^2 + 4a_r c_r < 0$; (B) monotonic systems, for which $b_r^2 + 4a_r c_r \geq 0$. In both cases, a posteriori methods are supplied for constructing strict and realistic error bounds in $O(r)$ arithmetic operations. A priori bounds, also requiring $O(r)$ arithmetic operations, are supplied in Case B. Several illustrative numerical examples are included.

**1. Introduction.** The application of $m$th order linear recurrence relations

$$a_{r0}p_r + a_{r1}p_{r-1} + a_{r2}p_{r-2} + \cdots + a_{rm}p_{r-m} + d_r = 0, \tag{1.1}$$

in which $a_{r0} \neq 0$, all $r$, to generate a sequence of values $p_m$, $p_{m+1}, \ldots$ from prescribed values of $p_0, p_1, \ldots, p_{m-1}$ is a well-understood procedure in numerical analysis. See, for example, [1], [2], [3], [4] and, most recently, the monograph of Wimp [19]. If the corresponding homogeneous equation is regarded as a difference equation, then it has $m$ linearly independent solutions—the so-called complementary functions of (1.1). Each rounding error introduced in the recurrence process contaminates the wanted solution of (1.1) by small multiples of the complementary functions. This is of no concern if the wanted solution grows in size at least as fast as any of the complementary functions, that is, if it is a dominant solution. In other cases the process may fail, indeed fail disastrously, and in order to achieve stability it is necessary to apply the recurrence relation in a backward direction, or to solve a boundary value problem.

Perhaps because stability conditions are so well understood, comparatively little attention has been paid to the problem of constructing strict error bounds for the computed results. These bounds are to cover the effects of rounding errors introduced during the recurrence steps as well as inherent errors in the coefficients $a_{rj}$ and $d_r$ and the initial values $p_0, p_1, \ldots, p_{m-1}$. This is the problem treated in the present investigation. One obvious application is to the development of robust software for the generation of transcendental mathematical functions by recurrence.

The only relevant published work appears to be that for Miller's algorithm; see [7], [9], [16]. In fact, some results for the present problem could be found simply by specializing results given in these references, especially [7]. This approach leads to unnecessary complications, however, and a more direct attack is called for.

Received October 30, 1986, revised July 6, 1987.

1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 65Q05; Secondary 65G05.

*This research has been supported by the U. S. Army Research Office, Durham under Contract DAAG-29-84-K-0022, the U. S. National Science Foundation under Grant DMS-84-19820 and the U. K. Science and Engineering Research Council.

We first observe that the evaluation of $p_r$ for the range $r = m, m+1, \ldots, m+n-1$, say, is equivalent to the solution of a system of $n$ linear algebraic equations. Hence the required error bounds can be found by available algorithms in matrix algebra; see, for example, [13], [14]. A drawback to this approach is that it requires the inversion of a lower triangular band matrix. The number of arithmetic operations needed for the inversion is $O(n^2)$, for large $n$, compared with only $O(n)$ operations for the computation of the solution $p_r$. It can be argued that it suffices to have the norm of the inverse matrix. However, it is an upper bound for the norm that is really needed, and this is tantamount to the original problem.[1]

Another drawback to the matrix approach is that it usually fails to provide insight into the nature of the error bounds; in particular, it will not yield realistic bounds of a priori type unless, of course, bounds for the elements or norm of the inverse matrix are known.

A second general approach is to apply rounded interval arithmetic [8, Section 2.4]. Often this procedure is quite successful. In many cases, however, the computed intervals are absurdly unrealistic. We illustrate this observation by two simple examples.

*Example* 1.1.

$$12p_{r+1} = 25p_r - 13p_{r-1}; \qquad p_0 = 1, \ p_1 = 13/12. \tag{1.2}$$

Computed interval values of $p_2, p_3, \ldots, p_{16}$ are given in Table 1.1. For example, the entries for $r = 2$ mean that

$$1.17360 \le p_2 \le 1.17363.$$

Six-figure decimal arithmetic was employed, with directed rounding[2] applied immediately following each arithmetic operation at each recurrence step.

Clearly the interval widths grow rapidly as $r$ increases. After $r = 12$ the left endpoint begins to decrease and actually becomes negative at $r = 16$, even though the true solution $p_r = (13/12)^r$ is positive, increasing and dominant.

*Example* 1.2.

$$3p_{r+1} - \sqrt{22}p_r + 2p_{r-1} - 1 = 0; \qquad p_0 = p_1 = 1. \tag{1.3}$$

An interval solution was computed in the same manner as Example 1.1, and the results are presented in Table 1.2. Again the interval widths grow rapidly with $r$, even though the wanted solution is dominant and tends to the constant value $3.23013\ldots$ as $r \to \infty$. The actual solution is given by

$$p_r = \tfrac{1}{3}(5 + \sqrt{22}) - 2^{r/2}3^{-(r+2)/2}\{(2 + \sqrt{22})\cos r\omega + (\sqrt{176} - \sqrt{50})\sin r\omega\},$$

with $\omega = \tan^{-1}(1/\sqrt{11})$.

[1] Compare [5]. Here algorithms are supplied for computing the norm of the inverse of a tridiagonal matrix of order $n$ in $O(n)$ operations. The algorithms entail the application of three-term homogeneous recurrence relations.

[2] That is, towards $-\infty$ for left endpoints and towards $+\infty$ for right endpoints.

TABLE 1.1
*Interval solution of* (1.2)

| $r$ | $p_r$ | | $Ip_{r+1}/Ip_r$ |
|---|---|---|---|
| 0 | 1 | 1 | – |
| 1 | 1.08333 | 1.08334 | 3.0 |
| 2 | 1.17360 | 1.17363 | 3.333... |
| 3 | 1.27137 | 1.27147 | 2.5 |
| 4 | 1.37725 | 1.37750 | 2.6 |
| 5 | 1.49183 | 1.49248 | 2.523... |
| 6 | 1.61568 | 1.61732 | 2.524... |
| 7 | 1.74914 | 1.75328 | 2.514... |
| 8 | 1.89194 | 1.90235 | 2.515... |
| 9 | 2.04215 | 2.06834 | 2.514... |
| 10 | 2.19359 | 2.25945 | 2.514... |
| 11 | 2.32926 | 2.49487 | 2.514... |
| 12 | 2.40488 | 2.82127 | 2.514... |
| 13 | 2.30738 | 3.35430 | 2.514... |
| 14 | 1.75065 | 4.38285 | 2.514... |
| 15 | 0.0133583 | 6.63135 | 2.514... |
| 16 | – 4.72027 | 11.9189 | |

TABLE 1.2
*Interval solution of* (1.3)

| $r$ | $p_r$ | | $Ip_{r+1}/Ip_r$ |
|---|---|---|---|
| 0 | 1 | 1 | – |
| 1 | 1 | 1 | – |
| 2 | 1.23013 | 1.23014 | 3.0 |
| 3 | 1.58993 | 1.58996 | 2.333... |
| 4 | 1.99904 | 1.99911 | 2.142... |
| 5 | 2.39879 | 2.39894 | 2.133... |
| 6 | 2.75102 | 2.75134 | 2.0 |
| 7 | 3.03517 | 3.03581 | 2.015... |
| 8 | 3.24447 | 3.24576 | 1.968... |
| 9 | 3.38203 | 3.38457 | 1.933... |
| 10 | 3.45716 | 3.46207 | 1.930... |
| 11 | 3.48206 | 3.49154 | 1.916... |
| 12 | 3.46933 | 3.48750 | 1.915... |
| 13 | 3.42980 | 3.46460 | 1.913... |
| 14 | 3.37070 | 3.43730 | 1.912... |
| 15 | 3.29356 | 3.42094 | 1.912... |
| 16 | 3.19116 | 3.43477 | |

The explanation of the failure of interval arithmetic in these examples is the usual one: the process takes no account of the interdependence of errors at successive steps. In fact, in Example 1.1 the interval widths $Ip_r$, say, eventually grow in proportion to $\alpha^r$, where $\alpha = 2.514\ldots$ is the largest zero of the polynomial $12z^2 - 25z - 13$. This is confirmed by the numerical values of the ratio $Ip_{r+1}/Ip_r$ given in the final column of Table 1.1. Similarly in Example 1.2 the interval widths eventually grow in proportion to $\alpha^r$, where $\alpha = 1.912\ldots$ is the largest zero of $3z^2 - \sqrt{22}z - 2$.

To construct methods that entail no more than $O(r)$ arithmetic operations and yield realistic error bounds, we have to impose restrictions on the nature of the recurrence relation. Without such restrictions, we have only the general matrix approach, with its $O(r^2)$ operations, to fall back on for realistic bounds. The present paper treats only real second-order relations. We also restrict ourselves to homogeneous systems, mainly because inhomogeneous problems often require error bounds for the associated complementary functions as a necessary preliminary [1], [10], [19]. In some cases, however, our methods carry over straightforwardly to inhomogeneous systems. Admittedly, the problems that fall within our scope amount to only a small subclass of the general problem of solving linear difference equations; nevertheless, this subclass includes many important recurrence relations satisfied by the higher transcendental functions.

We standardize (1.1) for homogeneous second-order systems in the form

$$a_r p_{r+1} = b_r p_r + c_r p_{r-1}, \tag{1.4}$$

with $p_0$ and $p_1$ prescribed and $a_r \neq 0$, all $r$. We distinguish two cases: *oscillatory systems* in which $b_r^2 + 4a_r c_r$ is negative for all $r$, and *monotonic systems* in which $b_r^2 + 4a_r c_r$ is nonnegative for all $r$. This classification is suggested, of course, by the nature of the solutions when the $a_r$, $b_r$ and $c_r$ are constants. Oscillatory systems are treated in Section 2, and monotonic systems in Sections 3, 4 and 5. In both cases we provide methods for constructing error bounds of a posteriori type. For monotonic systems we also furnish a priori bounds. Some numerical examples are supplied in Section 6, and brief conclusions are drawn in Section 7.

**2. Oscillatory Systems.** In (1.4) we replace $c_r$ by $-c_r$ for convenience. The oscillatory case is then given by

$$a_r p_{r+1} = b_r p_r - c_r p_{r-1}, \tag{2.1}$$

with $b_r^2 < 4a_r c_r$, all $r$. Without loss of generality we may suppose that $a_r$ and $c_r$ are positive.

Example 1.2 is typical for systems of this kind in that interval arithmetic will generally yield unsatisfactory results. The error bounds, or interval widths, eventually grow at the same rate as the dominant solution of the equation

$$a_r p_{r+1} = |b_r| p_r + c_r p_{r-1}.$$

That this solution grows faster than the solutions of (2.1) can be inferred from the case in which the coefficients are constants.

In order to proceed, let $q_r$ be a solution of (2.1) that is independent of $p_r$ and (like $p_r$) is computed by forward recurrence from given values at $r = 0$ and 1. Denote the stored values of $p_r$, $q_r$ and other quantities by the addition of overbars. Also, let $\phi_r$ and $\psi_r$ be the aggregate errors introduced on the $(r-1)$st step in the computation of $p_r$ and $q_r$, as expressed by the formulae

$$a_{r-1}\bar{p}_r = b_{r-1}\bar{p}_{r-1} - c_{r-1}\bar{p}_{r-2} + \phi_r, \qquad a_{r-1}\bar{q}_r = b_{r-1}\bar{q}_{r-1} - c_{r-1}\bar{q}_{r-2} + \psi_r. \tag{2.2}$$

Thus $\phi_r$ includes the effects of all abbreviation errors[3] introduced in the computation of $\bar{p}_r$ from $\bar{p}_{r-1}$ and $\bar{p}_{r-2}$ as well as the effects of inherent errors in the given values of the coefficients $a_{r-1}$, $b_{r-1}$ and $c_{r-1}$. Similarly for $\psi_r$.

Bounds for $|\phi_r|$ and $|\psi_r|$ can be computed by standard methods of round-off error analysis, see for example [12], [18], or by interval arithmetic. For the initial values we set

$$\bar{p}_0 = p_0 + \phi_0, \quad \bar{p}_1 = p_1 + \phi_1, \quad \bar{q}_0 = q_0 + \psi_0, \quad \bar{q}_1 = q_1 + \psi_1. \tag{2.3}$$

The relationship of the stored values $\bar{p}_r$ and $\bar{q}_r$ to the true values $p_r$ and $q_r$ is easily verified to be

$$\bar{p}_r = p_r + B_r p_r - A_r q_r, \qquad \bar{q}_r = q_r + D_r p_r - C_r q_r, \tag{2.4}$$

where

$$A_r = -w_1 p_1 \phi_0 + \sum_{j=1}^{r} w_j p_{j-1}\phi_j, \qquad B_r = -w_1 q_1 \phi_0 + \sum_{j=1}^{r} w_j q_{j-1}\phi_j, \tag{2.5}$$

$$C_r = -w_1 p_1 \psi_0 + \sum_{j=1}^{r} w_j p_{j-1}\psi_j, \qquad D_r = -w_1 q_1 \psi_0 + \sum_{j=1}^{r} w_j q_{j-1}\psi_j, \tag{2.6}$$

and

$$w_1 = \frac{1}{p_1 q_0 - p_0 q_1}, \qquad w_r = \frac{1}{a_{r-1}(p_r q_{r-1} - p_{r-1} q_r)}, \qquad r \geq 2. \tag{2.7}$$

[3] By "abbreviation errors" we mean chopping or rounding errors.

The $w_r$ are finite since $p_r$ and $q_r$ are assumed to be independent solutions. We also have the recurrence relation

$$w_r = (a_{r-2}/c_{r-1})w_{r-1}, \qquad r \geq 3, \tag{2.8}$$

and $w_2 = w_1/c_1$. Let us denote the wanted errors by

$$\varepsilon_r = p_r - \bar{p}_r, \qquad \eta_r = q_r - \bar{q}_r. \tag{2.9}$$

Suppose that we have computed $\bar{p}_r$ and $\bar{q}_r$, together with bounds on $|p_j|$, $|q_j|$, $|\varepsilon_j|$, $|\eta_j|$, $|A_j|$, $|B_j|$, $|C_j|$, $|D_j|$ and $|w_j|$, for all $j \leq r-1$. We first compute bounds on $|\phi_r|$, $|\psi_r|$ and $|w_r|$; compare (2.2) and (2.8). Next, from (2.5) and (2.6) we have

$$A_r = A_{r-1} + w_r p_{r-1}\phi_r, \qquad B_r = B_{r-1} + w_r q_{r-1}\phi_r, \tag{2.10}$$

$$C_r = C_{r-1} + w_r p_{r-1}\psi_r, \qquad D_r = D_{r-1} + w_r q_{r-1}\psi_r, \tag{2.11}$$

provided that $r \geq 1$. Using these relations we compute bounds on $|A_r|$, $|B_r|$, $|C_r|$ and $|D_r|$. Then by substituting the results obtained so far into the identities

$$E_r\varepsilon_r = -\{B_r(1 - C_r) + A_r D_r\}\bar{p}_r + A_r\bar{q}_r, \tag{2.12}$$

$$E_r\eta_r = -D_r\bar{p}_r + \{(1 + B_r)C_r - A_r D_r\}\bar{q}_r, \tag{2.13}$$

in which

$$E_r = (1 + B_r)(1 - C_r) + A_r D_r, \tag{2.14}$$

we arrive at bounds for $|\varepsilon_r|$ and $|\eta_r|$. (These identities are obtained by solving Eqs. (2.4) for $p_r$ and $q_r$, and using (2.9).) Bounds for $p_r$ and $q_r$ follow from (2.9), and after computing $\bar{p}_{r+1}$ and $\bar{q}_{r+1}$ from (2.1) we are ready to repeat the cycle.

This is our method for constructing a posteriori error bounds. The magnitudes of the solutions $p_r$ and $q_r$ may rise or fall as $r$ increases, depending on whether $c_r \gtrless a_r$. However, provided that the rate of growth of the magnitudes of the solutions does not differ significantly from that of $(c_r/a_r)^{1/2}$, all terms in the sums in (2.5) and (2.6) will remain of comparable magnitude, owing to the presence of the factors $w_j$. That this growth condition is not unreasonable can be seen by analogy with the case in which the difference equation has constant coefficients. Nevertheless, the condition will not always be satisfied in the general case, and it may need to be examined by asymptotic analysis or other independent means.

When the growth condition just discussed is satisfied, the bounds for $|A_r|$, $|B_r|$, $|C_r|$ and $|D_r|$ may be expected to grow approximately linearly with $r$, which is an essential requirement for the bounds for $|\varepsilon_r|$ and $|\eta_r|$ to be realistic. The number of arithmetic operations needed is several times that required to compute the $\bar{p}_r$, of course, but is still only $O(r)$ for large $r$. Moreover, many of these computations could be performed in parallel: if this is arranged, then the total execution time will not greatly exceed that needed for the computation of the $\bar{p}_r$ alone. Lastly, the method can be extended easily to inhomogeneous oscillatory systems, as long as the wanted solution is not dominated by the complementary functions as $r$ increases.

**3. Monotonic Systems (i).** We now consider Eq. (1.4), that is,

$$a_r p_{r+1} = b_r p_r + c_r p_{r-1}, \tag{3.1}$$

with the condition $b_r^2 + 4a_r c_r \geq 0$, all $r$. We may suppose that $a_r > 0$, and we shall also suppose that $b_r \geq 0$.[4] In the present section we require $c_r \geq 0$, deferring the more difficult case of negative $c_r$ until Sections 4 and 5.

The essential behavior in this case is that for appropriately chosen solutions the relative errors are simply additive. To express this result precisely and conveniently, we use relative precision (rp) in place of relative error, that is, we work in terms of the absolute errors of the logarithms of approximations [12].

We assume that the stored values $\bar{a}_r$, $\bar{b}_r$ and $\bar{c}_r$ of $a_r$, $b_r$ and $c_r$, respectively, are correct to rp($\delta$), say, and the computations are performed in floating-point arithmetic with a working relative precision (wrp) of $\gamma$. (In other words, each arithmetic operation is accompanied by a chopping or rounding error not exceeding rp($\gamma$).) We also assume that the initial values satisfy

$$p_0 \simeq \bar{p}_0; \quad \mathrm{rp}(\varpi), \qquad p_1 \simeq \bar{p}_1; \quad \mathrm{rp}(\varpi), \tag{3.2}$$

where $\bar{p}_0$ and $\bar{p}_1$ are nonnegative, and $\varpi$, like $\delta$ and $\gamma$, is given. (Without these assumptions, $p_r$ might be recessive as $r \to \infty$.) By application of the rules of rp error analysis and a simple inductive argument we deduce that

$$p_r \simeq \bar{p}_r; \quad \mathrm{rp}\{\varpi + (2r-2)\delta + (3r-3)\gamma\}, \qquad r \geq 1. \tag{3.3}$$

This is the required result. Often it is improvable in minor ways. For example, if $a_r = 1$, all $r$, then the coefficients of $\delta$ and $\gamma$ can be reduced to $r-1$ and $2r-2$, respectively.

It should also be noted that if interval arithmetic is applied directly to (3.1), then it will yield realistic a posteriori bounds. However, in view of the simplicity and effectiveness of the a priori bounds just given, the extra computations entailed by use of interval arithmetic can be avoided.

**4. Monotonic Systems (ii).** In this and the next section we consider the equation

$$a_r p_{r+1} = b_r p_r - c_r p_{r-1}, \tag{4.1}$$

in which $b_r^2 \geq 4a_r c_r$, $a_r > 0$, $b_r > 0$ and $c_r \geq 0$, for all $r$. We seek a solution $p_r$ such that $p_r \geq 0$, for all $r$.

For reasons similar to those given in the oscillatory case (Section 2), interval arithmetic applied directly to (4.1) will yield unsatisfactory results. The method of Section 2 also fails. If $p_r$ is dominant and $q_r$ is recessive as $r \to \infty$, then in the second of (2.4) the term $D_r p_r$ soon overwhelms $q_r$. If $p_r$ and $q_r$ are both dominant, then the situation is even worse.

One way to proceed is to transform (4.1) into the nonlinear equation

$$a_r h_{r+1} = b_r - (c_r/h_r) \tag{4.2}$$

satisfied by the ratio $h_r = p_r/p_{r-1}$. Then interval arithmetic, or a running error analysis [12], [18], can be applied to the computation of the sequence $\{h_r\}$ by recurrence, and also to the subsequent recovery of the wanted solution from the product

$$p_r = h_r h_{r-1} \cdots h_1 p_0. \tag{4.3}$$

[4] Systems in which $a_r$ and $b_r$ have opposite signs for all $r$ are accommodated by replacing $p_r$ by $(-1)^r p_r$.

The reason these procedures are now more successful is that they make appropriate allowance for interactions of errors. In contrast, when (4.1) is computed in interval form, the upper (say) endpoint of $p_{r+1}$ depends on the upper endpoint of $p_r$ and the *lower* endpoint of $p_{r-1}$.

Another approach is to replace (4.1) by a pair of first-order linear equations with nonnegative coefficients; compare Section 3. For example, we can introduce a new variable $u_r$ defined by

$$u_r = p_{r+1} - \lambda_r p_r,$$

where $\lambda_r$ is a positive function of $r$ at our disposal, subject to the condition $u_r \geq 0$. Then (4.1) is equivalent to

$$a_r u_r = \nu_r p_r + \mu_{r-1} u_{r-1}, \qquad p_{r+1} = \lambda_r p_r + u_r, \tag{4.4}$$

where

$$\mu_{r-1} = c_r/\lambda_{r-1}, \qquad \nu_r = b_r - a_r\lambda_r - \mu_{r-1}. \tag{4.5}$$

By hypothesis, $\lambda_{r-1} > 0$, hence $\mu_{r-1}$ is finite and nonnegative. The remaining coefficient $\nu_r$ is nonnegative as long as $\lambda_r$ and $\lambda_{r-1}$ also satisfy

$$a_r\lambda_{r-1}\lambda_r - b_r\lambda_{r-1} + c_r \leq 0. \tag{4.6}$$

If the coefficients $a_r$, $b_r$ and $c_r$ are slowly-varying functions of $r$ such that $b_r^2 > 4a_rc_r$ and the starting values $p_0$, $p_1$ are chosen appropriately, then it will usually be possible to satisfy (4.6). This is because the zeros of the local characteristic polynomial $a_rz^2 - b_rz + c_r$ are real and distinct, and in effect (4.6) requires $\lambda_{r-1}$ and $\lambda_r$ to lie between them. For example, we might choose $\lambda_r$ to be the arithmetic mean of the zeros, given by

$$\lambda_r = b_r/(2a_r).$$

Then (4.6) is satisfied as long as

$$b_{r-1}b_r \geq 4a_{r-1}c_r, \quad \text{all } r.$$

Solutions of (4.4) may be generated by interval arithmetic or with a running error analysis. Considerable cancellation may occur in the computation of $\nu_r$ from the second of (4.5); in consequence, it may be necessary to employ higher precision on this step.

In the next section we describe a semianalytical method. This method provides greater insight into the actual error propagation, and leads to useful a priori bounds. It has some features in common with the valuable method used by Mattheij and van der Sluis for obtaining error bounds for Miller's algorithm [7].

**5. Monotonic Systems (iii).** As in Section 4 we consider the equation

$$a_r p_{r+1} = b_r p_r - c_r p_{r-1}, \tag{5.1}$$

but with the conditions on the coefficients modified to $b_r^2 > 4a_rc_r$, $a_r > 0$, $b_r > 0$ and $c_r > 0$, for all $r$. Again, we wish to compute a solution $p_r$ that is dominant as $r \to \infty$. We suppose that $p_r$ is positive when $r > 0$ and nonnegative when $r = 0$. To begin with, we denote by $q_r$ any positive solution that is independent of $p_r$.

As in earlier sections, we use overbars to indicate stored values. We first investigate the actual propagation of the aggregate abbreviation error $\phi_j$, say, introduced on the $(j-1)$st application of (5.1) according to the formula

$$a_{j-1}\bar{p}_j = b_{j-1}\bar{p}_{j-1} - c_{j-1}\bar{p}_{j-2} + \phi_j, \qquad j \geq 2; \tag{5.2}$$

compare (2.2). The solution $p_r^{(j)}$, say, of (5.1) that satisfies

$$p_{j-1}^{(j)} = 0, \qquad p_j^{(j)} = \phi_j/a_{j-1}, \qquad j \geq 2,$$

is expressible in the form

$$p_r^{(j)} = \left(1 - \frac{p_{j-1}q_r}{q_{j-1}p_r}\right)\frac{t_j\phi_j}{a_{j-1}p_j}p_r, \tag{5.3}$$

where

$$t_j = \left(1 - \frac{p_{j-1}q_j}{q_{j-1}p_j}\right)^{-1}, \qquad j \geq 1. \tag{5.4}$$

With the assumed conditions, $t_j$ is always finite.

Now suppose that $q_r$ is the recessive solution of (5.1), so that $q_r/p_r \to 0$ as $r \to \infty$. Although $q_r$ is unique only up to a constant factor, obviously from (5.4) the coefficients $t_j$ in (5.3) do not depend on this factor. Furthermore, from (5.3) we have

$$\frac{p_r^{(j)}}{p_r} \to \frac{t_j\phi_j}{a_{j-1}p_j}, \qquad r \to \infty, \quad j \text{ fixed}. \tag{5.5}$$

This means that the relative error $\phi_j/(a_{j-1}p_j)$ introduced on the $(j-1)$st application of (5.1) is magnified ultimately by the factor $t_j$. If it happens that $q_r/p_r$ is decreasing for all $r$, then we have, in addition,

$$\frac{|p_r^{(j)}|}{p_r} \leq \frac{t_j|\phi_j|}{a_{j-1}p_j}, \qquad r \geq j. \tag{5.6}$$

In other words, the actual propagated error is bounded by its limiting form. It also has the same sign.

For our purposes, it is not essential for $q_r$ to be the recessive solution. Suppose that we are computing $p_r$ over the range $r = 2, 3, \ldots, n$, where $n$ is arbitrary. Let $q_r$ now denote any solution of (5.1) that is positive when $0 \leq r \leq n-1$, nonnegative when $r = n$ and also has the property that $q_r/p_r$ is decreasing for $0 \leq r \leq n$. Then $q_r$ is independent of $p_r$; furthermore, if $t_j$ is defined by (5.4) in terms of the present $q_r$, then (5.6) applies for $j = 2, 3, \ldots, n$.

To investigate the effect of inherent errors in the starting values at $r = 0$ and 1, let

$$p_0 = \bar{p}_0 - \phi_0, \qquad p_1 = \bar{p}_1 - \phi_1, \tag{5.7}$$

as in (2.3). Then the solution $p_r^{(0)}$, say, of (5.1) that satisfies

$$p_0^{(0)} = -\phi_0, \qquad p_1^{(0)} = -\phi_1, \tag{5.8}$$

is given by

$$p_r^{(0)} = \left(1 - \frac{p_1q_r}{q_1p_r}\right)\frac{t_0\phi_0}{p_0}p_r - \left(1 - \frac{p_0q_r}{q_0p_r}\right)\frac{t_1\phi_1}{p_1}p_r, \tag{5.9}$$

where $t_1$ is defined as in (5.4) and

$$t_0 = \left( \frac{q_0 p_1}{p_0 q_1} - 1 \right)^{-1}. \tag{5.10}$$

With the assumed conditions we have

$$\frac{|p_r^{(0)}|}{p_r} \leq \frac{t_0 |\phi_0|}{p_0} + \frac{t_1 |\phi_1|}{p_1}, \qquad r \geq 1. \tag{5.11}$$

On combining the effects of all the errors $\phi_0, \phi_1, \ldots, \phi_r$ we arrive at

$$\frac{|p_r - \bar{p}_r|}{p_r} \leq t_0 \frac{|\phi_0|}{p_0} + t_1 \frac{|\phi_1|}{p_1} + \sum_{j=2}^{r} t_j \frac{|\phi_j|}{a_{j-1} p_j}, \qquad 2 \leq r \leq n. \tag{5.12}$$

In the relations (5.9) to (5.12) we have supposed that $p_0 \neq 0$. If $p_0 = 0$, then we suppose that $\bar{p}_0 = 0$. The inequalities (5.11) and (5.12) then apply without the term $t_0 |\phi_0| / p_0$ on their right-hand sides.[5]

In order to proceed, we need bounds on the coefficients $t_j$ defined by (5.4) and (5.10). In turn, this necessitates bounds on $p_{j-1}/p_j$ and $q_j/q_{j-1}$. Results of this kind have been supplied by the present writer [11], Mattheij [6] and van der Sluis [15]. For present purposes a simple and convenient result is provided by the following theorem. This result is included in that given by Theorem 4.1 of [6], but for simplicity we give a proof using our present notation.

THEOREM 5.1. *Let $\alpha_r$ and $\beta_r$ denote the (positive) zeros of the quadratic $a_r z^2 - b_r z + c_r$, chosen so that $\alpha_r > \beta_r$. Write*

$$\alpha = \min(\alpha_1, \alpha_2, \ldots, \alpha_{n-1}), \quad \mathrm{A} = \max(\alpha_1, \alpha_2, \ldots, \alpha_{n-1}), \\ \mathrm{B} = \max(\beta_1, \beta_2, \ldots, \beta_{n-1}), \tag{5.13}$$

*and assume that $\alpha \geq \mathrm{B}$. Also, let $v_r$ be any solution of (5.1) that is nonnegative when $r = 0$ and satisfies $v_1/v_0 \geq \mathrm{B}$. Then*

$$\hat{\alpha} \leq v_r / v_{r-1} \leq \hat{\mathrm{A}}, \qquad r = 1, 2, \ldots, n, \tag{5.14}$$

*where*

$$\hat{\alpha} = \min(\alpha, v_1/v_0), \qquad \hat{\mathrm{A}} = \max(\mathrm{A}, v_1/v_0). \tag{5.15}$$

(In the case $v_0 = 0$ the condition $v_1/v_0 \geq \mathrm{B}$ becomes $v_1 > 0$, $\hat{\alpha} = \alpha$ and $\hat{\mathrm{A}} = \infty$.)

To prove the theorem, write

$$f_r(z) = \frac{b_r}{a_r} - \frac{c_r}{a_r z},$$

so that

$$v_{r+1}/v_r = f_r(v_r/v_{r-1}).$$

We observe that for fixed $r$, $f_r(z)$ is increasing when $z > 0$ and

$$f_r(z) \leq z, \quad \text{if } z \geq \alpha_r; \quad f_r(z) \geq z, \quad \text{if } \beta_r \leq z \leq \alpha_r.$$

[5] An appropriate modification could be made, however, if $p_0 = 0$ but $\bar{p}_0 \neq 0$.

From these results and the identity $f_r(\alpha_r) = \alpha_r$ it follows that:

$$\begin{aligned} &\text{(a)} \quad \text{if } \alpha_r \le v_r/v_{r-1}, && \text{then } \alpha_r \le v_{r+1}/v_r \le v_r/v_{r-1}; \\ &\text{(b)} \quad \text{if } \beta_r \le v_r/v_{r-1} \le \alpha_r, && \text{then } v_r/v_{r-1} \le v_{r+1}/v_r \le \alpha_r. \end{aligned}$$

The result (5.14) is now proved by induction. Suppose that $v_r/v_{r-1} \ge \mathrm{B}$ and $\hat{\alpha} \le v_r/v_{r-1} \le \hat{\mathrm{A}}$, as is certainly the case when $r = 1$. Then $v_r/v_{r-1} \ge \beta_r$. Hence (a) or (b) applies. In either event we have $v_{r+1}/v_r \ge \mathrm{B}$ and $\hat{\alpha} \le v_{r+1}/v_r \le \hat{\mathrm{A}}$. □

Let us return to the bound (5.12). Defining $\alpha$ and B by (5.13) and applying Theorem 5.1, we find that

$$p_{r-1}/p_r \le 1/\rho, \qquad r = 1, 2, \ldots, n, \tag{5.16}$$

where

$$\rho = \min(\alpha, p_1/p_0), \tag{5.17}$$

provided that $\alpha \ge \mathrm{B}$ and $p_1/p_0 \ge \mathrm{B}$. To arrive at a similar bound for $q_r/q_{r-1}$, we now define $q_r$ to be the solution of (5.1) that satisfies

$$q_{n-1} = 1, \qquad q_n = 0.$$

This solution can be generated by backward recurrence:

$$c_r q_{r-1} = b_r q_r - a_r q_{r+1}, \qquad r = n-1, n-2, \ldots, 1.$$

By applying Theorem 5.1 to this form of the difference equation, we deduce that

$$q_r/q_{r-1} \le \mathrm{B}, \qquad r = 1, 2, \ldots, n. \tag{5.18}$$

If we now restrict $\alpha > \mathrm{B}$ and $p_1/p_0 > \mathrm{B}$, then $\rho > \mathrm{B}$, implying that $q_r/p_r$ is decreasing for $r = 0, 1, \ldots, n$. Accordingly, we may substitute in (5.4) and (5.10) by means of (5.16) and (5.18). This yields the required bounds in the form

$$t_0 \le \frac{\mathrm{B}}{\rho - \mathrm{B}}; \qquad t_j \le \frac{\rho}{\rho - \mathrm{B}}, \quad 1 \le j \le n. \tag{5.19}$$

It is now easy to see how to compute a posteriori bounds for $|p_r - \bar{p}_r|$ successively for $r = 2, 3, \ldots, n$. Write

$$T_r = t_0 \frac{|\phi_0|}{p_0} + t_1 \frac{|\phi_1|}{p_1} + \sum_{j=2}^{r} t_j \frac{|\phi_j|}{a_{j-1} p_j}, \qquad r \ge 1, \tag{5.20}$$

with the understanding that the term $t_0|\phi_0|/p_0$ is omitted in the case $p_0 = \bar{p}_0 = 0$ and the empty sum is zero in the case $r = 1$. From (5.12) we derive

$$p_r - \bar{p}_r = \vartheta_r T_{r-1} p_r + \vartheta_r t_r \frac{|\phi_r|}{a_{r-1}}, \qquad 2 \le r \le n,$$

where $\vartheta_r$ is some number in the interval $[-1, 1]$. Solving for $p_r$ we deduce that

$$|p_r - \bar{p}_r| \le \frac{1}{1 - T_{r-1}} \left( T_{r-1} \bar{p}_r + t_r \frac{|\phi_r|}{a_{r-1}} \right), \tag{5.21}$$

provided that $T_{r-1} < 1$. As in Section 2 write $\varepsilon_r = p_r - \bar{p}_r$, and suppose that we have arrived at a lower bound for $p_{r-1}$ and upper bounds for $|\varepsilon_{r-1}|$ and $|T_{r-1}|$, with $r \ge 2$. Inequalities (5.19) and (5.21) immediately yield an upper bound for $|\varepsilon_r|$. A lower bound for $p_r$ can then be obtained, for example, from the inequality

$$p_r \ge \bar{p}_r - |\varepsilon_r| \tag{5.22}$$

(as long as $|\varepsilon_r| < \bar{p}_r$). And since

$$T_r = T_{r-1} + \frac{t_r|\phi_r|}{a_{r-1}p_r}, \qquad r \geq 2, \tag{5.23}$$

we can also find an upper bound for $T_r$. The cycle is now ready to be repeated.

A more interesting problem is to extend the foregoing analysis to yield a priori bounds. As in Section 3, we suppose that the stored values of the coefficients $\bar{a}_r$, $\bar{b}_r$ and $\bar{c}_r$ are correct to rp($\delta$), the initial values $\bar{p}_0$ and $\bar{p}_1$ are correct to rp($\varpi$) and the computations are carried out in floating-point arithmetic with wrp($\gamma$).

THEOREM 5.2. *Let $p_r$ and $q_r$ be solutions of* (5.1) *such that $p_0 \geq 0$, $p_r > 0$ when $r > 0$, $q_r > 0$ when $0 \leq r \leq n-1$, $q_n \geq 0$, and $q_r/p_r$ is decreasing for $0 \leq r \leq n$. Assume also the conditions of the preceding paragraph, and let $\varpi_0 = \varpi_1 = \varpi$ and*

$$\varpi_r = 2\left[(t_0+t_1)\varpi + \sum_{j=2}^{r} t_j\left\{\delta + 2\gamma + \left(\frac{b_{j-1}}{a_{j-1}}\frac{p_{j-1}}{p_j} + \frac{c_{j-1}}{a_{j-1}}\frac{p_{j-2}}{p_j}\right)(\delta+\gamma)\right\}\right], \tag{5.24}$$

*$r \geq 2$, with $t_j$ defined by* (5.4) *and* (5.10).[6] *Then*

$$p_r \simeq \bar{p}_r; \qquad \mathrm{rp}(\varpi_r), \tag{5.25}$$

*provided that $\varpi_r \leq \varsigma$, where $\varsigma = 0.265\ldots$ is the positive root of the equation*

$$-\ln\left(1 - \frac{ze^{3z/2}}{2 - ze^{z/2}}\right) = z. \tag{5.26}$$

*Proof.* We first need an upper bound for the error term $\phi_j$ in (5.2). Since each arithmetic operation is accompanied by an abbreviation error of rp($\gamma$), we apply the rules of rp error analysis [12] to obtain

$$|\phi_j| \leq \{a_{j-1}\bar{p}_j(\delta+2\gamma) + (b_{j-1}\bar{p}_{j-1} + c_{j-1}\bar{p}_{j-2})(\delta+\gamma)\}e^{\delta+2\gamma}, \qquad j \geq 2.$$

Next, on comparing (5.7) with the given conditions we have

$$|\phi_0| \leq p_0\varpi e^{\varpi}, \qquad |\phi_1| \leq p_1\varpi e^{\varpi}.$$

Substituting in (5.12) by means of these inequalities, we derive

$$\frac{|p_r - \bar{p}_r|}{p_r} \leq \left[(t_0+t_1)\varpi + \sum_{j=2}^{r} t_j\left\{\frac{\bar{p}_j}{p_j}(\delta+2\gamma) + \left(\frac{b_{j-1}}{a_{j-1}}\frac{\bar{p}_{j-1}}{p_j} + \frac{c_{j-1}}{a_{j-1}}\frac{\bar{p}_{j-2}}{p_j}\right)(\delta+\gamma)\right\}\right]e^{\hat{\delta}+2\gamma}, \tag{5.27}$$

where

$$\hat{\delta} = \max(\varpi, \delta). \tag{5.28}$$

We shall establish (5.25) by induction. Suppose that

$$p_j \simeq \bar{p}_j; \qquad \mathrm{rp}(\varpi_j), \qquad j = 0, 1, \ldots, r-1, \tag{5.29}$$

[6] Again, when $p_0 = \bar{p}_0 = 0$ we set $t_0 = 0$.

as is certainly the case when $r = 1$ and 2. If we extract the term $t_r(\bar{p}_r/p_r)(\delta + 2\gamma)$ from within the square brackets of the right member of (5.27) and express it in the form

$$t_r\left(\frac{\bar{p}_r}{p_r} - 1\right)(\delta + 2\gamma) + t_r(\delta + 2\gamma),$$

then with the aid of (5.29) and the fact that each of $\varpi_0, \varpi_1, \ldots, \varpi_{r-1}$ is bounded by $\varpi_r$,[7] we see that

$$\frac{|p_r - \bar{p}_r|}{p_r} \le \left\{t_r(\delta + 2\gamma)\frac{|p_r - \bar{p}_r|}{p_r} + \frac{1}{2}\varpi_r e^{\varpi_r}\right\} e^{\hat{\delta}+2\gamma}.$$

Next, from (5.24), (5.28) and the inequalities $t_1 > 1$, $t_2 > 1$ it is easily seen that $t_r(\delta + 2\gamma)$ and $\hat{\delta} + 2\gamma$ are both bounded by $\frac{1}{2}\varpi_r$ when $r \ge 2$. It follows that

$$\frac{|p_r - \bar{p}_r|}{p_r} \le \frac{1}{2}\varpi_r \frac{|p_r - \bar{p}_r|}{p_r} e^{\varpi_r/2} + \frac{1}{2}\varpi_r e^{3\varpi_r/2},$$

and hence that

$$-\ln\left(1 - \frac{|p_r - \bar{p}_r|}{p_r}\right) \le -\ln\left(1 - \frac{\varpi_r e^{3\varpi_r/2}}{2 - \varpi_r e^{\varpi_r/2}}\right) \le \varpi_r,$$

the last step being a consequence of the assumption $\varpi_r \le \varsigma$; compare (5.26). Thus (5.29) holds when $j = r$. □

For the purpose of constructing a priori bounds, Theorem 5.2 possesses the essential feature that the error bound for $\bar{p}_r$ is expressed in terms of the true solution $p_r$ rather than the computed solution $\bar{p}_r$. With the notation of Theorem 5.1, and the assumptions $\alpha > \mathrm{B}$, $p_1/p_0 > \mathrm{B}$, the conditions of Theorem 5.2 on the solution $q_r$ are satisfied, and we may apply (5.16) and (5.19). From (5.1) we have

$$\frac{b_{j-1}p_{j-1}}{a_{j-1}p_j} = 1 + \frac{c_{j-1}p_{j-2}}{a_{j-1}p_j};$$

accordingly, (5.24) may be simplified into

$$\varpi_r = 2\left[(t_0 + t_1)\varpi + \sum_{j=2}^{r} t_j\left\{2\delta + 3\gamma + \frac{c_{j-1}}{a_{j-1}}\frac{p_{j-2}}{p_j}(2\delta + 2\gamma)\right\}\right]. \tag{5.30}$$

Then by making the indicated substitutions we arrive at

$$\varpi_r \le \frac{2}{\rho - \mathrm{B}}\left[(\mathrm{B} + \rho)\varpi + \rho\sum_{j=2}^{r}\left\{2\delta + 3\gamma + \frac{c_{j-1}}{a_{j-1}}\frac{1}{\rho^2}(2\delta + 2\gamma)\right\}\right].$$

If we now introduce the quantity

$$C = \max_{j\in[1,n-1]}(c_j/a_j),$$

[7] This follows from the definition (5.24) and the inequality $t_1 > 1$.

then we are led to the further simplification

$$\varpi_r \le \frac{2}{\rho - \mathrm{B}}\left[(\mathrm{B}+\rho)\varpi + (r-1)\left\{(2\delta+3\gamma)\rho + \frac{2C}{\rho}(\delta+\gamma)\right\}\right]. \tag{5.31}$$

*Remarks.* (a) The coefficient 2 outside the square brackets in the definition (5.24) of $\varpi_r$ is arbitrary, to some extent. In fact any constant in excess of unity could be used instead, provided that an appropriate change is made in the definition of $\varsigma$.

(b) By referring to the analysis in this section leading up to (5.12), it is easy to relate the terms on the right-hand side of (5.24) to the various errors introduced during the computations. Thus, the terms $(t_0+t_1)\varpi$ are contributed by the inherent errors in $\bar{p}_0$ and $\bar{p}_1$. In $\sum_{j=2}^{r}$ the terms $t_j(\delta+2\gamma)$ stem from the inherent error in $\bar{a}_{j-1}$ and the two errors introduced on abbreviating the difference $\overline{b_{j-1}p_{j-1}} - \overline{c_{j-1}p_{j-2}}$ and the quotient $\overline{b_{j-1}p_{j-1} - c_{j-1}p_{j-2}}/\bar{a}_{j-1}$. The remaining terms in $\sum_{j=2}^{r}$ stem from the inherent errors in $\bar{b}_{j-1}$ and $\bar{c}_{j-1}$, and the errors made in abbreviating the products $\bar{b}_{j-1}\bar{p}_{j-1}$ and $\bar{c}_{j-1}\bar{p}_{j-2}$.

(c) The bound (5.31) grows linearly with $r$, which is a necessary condition for it to be realistic. Moreover, if the coefficients $a_r$, $b_r$ and $c_r$ in the original equation are constants and $p_1/p_0 \ge \alpha$ (now the largest root of the characteristic equation), then $\rho = \alpha$ and the right-hand side of (5.31) becomes exactly twice the limiting value of the combined maximum effects of the inherent and abbreviation errors.

## 6. Numerical Examples.

*Example* 6.1. We compute the Legendre functions $P_r(x)$ and $Q_r(x)$ from the recurrence relation

$$(r+1)p_{r+1} = (2r+1)xp_r - rp_{r-1},$$

with the initial values

$$P_0(x) = 1, \quad P_1(x) = x, \quad Q_0(x) = \frac{1}{2}\ln\frac{1+x}{1-x}, \quad Q_1(x) = \frac{x}{2}\ln\frac{1+x}{1-x} - 1. \tag{6.1}$$

We take $x = 0.95$, with the understanding that this value may be in error by as much as $\pm 0.000001$, and use six-decimal floating-point arithmetic, with chopping, for the calculation of $p_r \equiv P_r(x)$ and $q_r \equiv Q_r(x)$. The computed values $\bar{p}_r$ and $\bar{q}_r$ are given for $r = 0, 1, \ldots, 16$ in the second and third columns of Table 6.1(i).

Error bounds have been computed from the formulae given in Section 2. It transpires, for example, that $w_r = 1$, all $r$. Upper bounds $|\varepsilon_r|^{\mathcal{A}}$ and $|\eta_r|^{\mathcal{A}}$ for the absolute errors $|\varepsilon_r|$ and $|\eta_r|$ in $\bar{p}_r$ and $\bar{q}_r$, respectively, appear in the fourth and fifth columns of Table 6.1(i). Some of the intermediate computations are shown in Table 6.1(ii). Here, and in subsequent examples, the superscript $\mathcal{A}$ ("above") is again used to signify upper bounds of the designated quantities, whereas in the final column the superscript $\mathcal{B}$ ("below") on $E_r$ indicates that entries in this column are lower bounds for $E_r$. These calculations were carried out by the methods of [12] using four-decimal floating-point arithmetic with chopping, except that in the cases $r = 0$ and 1 the values of $|\psi_r|^{\mathcal{A}}$ were found from Formulae (6.1) with the aid of high-precision values of the logarithmic function.

TABLE 6.1(i)

*Legendre functions $P_r(x)$ and $Q_r(x)$*

| $r$ | $\bar{p}_r$ | $\bar{q}_r$ | $10^6\lvert\varepsilon_r\rvert^A$ | $10^6\lvert\eta_r\rvert^A$ | $10^6\lvert\varepsilon_r\rvert$ | $10^6\lvert\eta_r\rvert$ |
|---|---|---|---|---|---|---|
| 0 | 1 | 1.83178 | 0 | 11.09 | 0 | 11.0... |
| 1 | 0.950000 | 0.740192 | 1 | 11.80 | 1 | 11.7... |
| 2 | 0.853750 | 0.138880 | 83.19 | 84.93 | 2.8... | 15.3... |
| 3 | 0.718436 | − 0.273566 | 138.8 | 102.8 | 6.7... | 14.9... |
| 4 | 0.554085 | − 0.558962 | 291.3 | 178.2 | 12.7... | 11.4... |
| 5 | 0.372736 | − 0.736972 | 441.6 | 315.9 | 17.8... | 8.1... |
| 6 | 0.187445 | − 0.817756 | 500.4 | 437.7 | 20.6... | 6.6... |
| 7 | 0.0112185 | − 0.811052 | 451.9 | 422.6 | 21.4... | 22.9... |
| 8 | − 0.144030 | − 0.729138 | 475.2 | 546.6 | 18.6... | 52.7... |
| 9 | − 0.268424 | − 0.587454 | 489.5 | 748.1 | 12.4... | 84.1... |
| 10 | − 0.354878 | − 0.404126 | 498.8 | 902.4 | 9.3... | 109.8... |
| 11 | − 0.399597 | − 0.198888 | 467.3 | 913.2 | 9.7... | 123.9... |
| 12 | − 0.402295 | 0.00830666 | 359.4 | 784.8 | 12.4... | 125.1... |
| 13 | − 0.366103 | 0.198763 | 535.5 | 903.5 | 19.8... | 112.6... |
| 14 | − 0.297193 | 0.356448 | 716.4 | 970.4 | 25.9... | 88.9... |
| 15 | − 0.204148 | 0.469164 | 834.4 | 994.6 | 29.9... | 57.4... |
| 16 | − 0.0971412 | 0.529380 | 839.7 | 924.6 | 31.5... | 16.4... |

TABLE 6.1(ii)

*Legendre functions (continued)*

| $r$ | $10^6\lvert\phi_r\rvert^A$ | $10^6\lvert\psi_r\rvert^A$ | $10^6\lvert A_r\rvert^A$ | $10^6\lvert B_r\rvert^A$ | $10^6\lvert C_r\rvert^A$ | $10^6\lvert D_r\rvert^A$ | $E_r^B$ |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 11.09 | 0 | 0 | 10.63 | 8.289 | − |
| 1 | 1 | 11.80 | 1.014 | 1.856 | 22.74 | 30.29 | − |
| 2 | 103.3 | 69.77 | 100.5 | 79.39 | 90.26 | 83.06 | 0.9998 |
| 3 | 150.6 | 46.09 | 232.2 | 101.6 | 131.4 | 90.70 | 0.9997 |
| 4 | 174.2 | 89.07 | 362.3 | 151.2 | 198.0 | 116.6 | 0.9996 |
| 5 | 169.3 | 189.2 | 462.4 | 249.2 | 307.0 | 225.4 | 0.9994 |
| 6 | 135.1 | 294.4 | 519.9 | 353.5 | 422.6 | 448.5 | 0.9992 |
| 7 | 74.25 | 378.2 | 541.2 | 419.9 | 500.5 | 768.5 | 0.9990 |
| 8 | 40.38 | 426.1 | 549.1 | 458.9 | 512.4 | 1129 | 0.9990 |
| 9 | 100.3 | 427.5 | 571.3 | 539.4 | 582.1 | 1460 | 0.9988 |
| 10 | 189.9 | 377.7 | 631.0 | 660.1 | 693.1 | 1705 | 0.9986 |
| 11 | 269.5 | 278.0 | 736.8 | 779.9 | 802.8 | 1842 | 0.9983 |
| 12 | 326.3 | 140.9 | 879.3 | 856.8 | 871.1 | 1896 | 0.9982 |
| 13 | 351.8 | 81.42 | 1034 | 871.9 | 916.4 | 1922 | 0.9981 |
| 14 | 340.4 | 212.6 | 1174 | 952.9 | 1008 | 1991 | 0.9980 |
| 15 | 291.0 | 383.5 | 1277 | 1070 | 1137 | 2157 | 0.9977 |
| 16 | 206.6 | 524.8 | 1337 | 1183 | 1261 | 2436 | 0.9975 |

Each of the quantities $|A_r|^A$, $|B_r|^A$, $|C_r|^A$ and $|D_r|^A$ appearing in Eqs. (2.4) grows monotonically with $r$, and very roughly in a linear fashion. The final error bounds $|\varepsilon_r|^A$ and $|\eta_r|^A$ exhibit some of the oscillatory character of the solutions $p_r$ and $q_r$. The overall sizes of $|\varepsilon_r|^A$ and $|\eta_r|^A$ are linked directly to the sizes of the bounds $|\phi_r|^A$ and $|\psi_r|^A$ for the abbreviation errors $\phi_r$ and $\psi_r$ in Eqs. (2.2).

Because of the uncertainty in the assumed value of $x$, the actual errors $\varepsilon_r$ and $\eta_r$ in $\bar{p}_r$ and $\bar{q}_r$ are unknown. However, their maximum absolute values can be found by taking $x = 0.95 \pm 0.000001$ in turn, and recalculating $p_r$ and $q_r$ using higher precision. The results are shown in the last two columns of Table 6.1(i). Of course, the bounds $|\varepsilon_r|^A$ and $|\eta_r|^A$ overestimate the actual values of $|\varepsilon_r|$ and $|\eta_r|$ considerably. This is caused partly by the stochastic nature of the actual abbreviation errors, and partly by the "radix effect". Had the computations been carried out in base 2, for example, instead of base 10, then the overestimation of the actual errors would be reduced by a factor of about 2 or 3 [12], [17].

*Example* 6.2. Let us solve the system (1.2) of Example 1.1 by the first method of Section 4, that is, by using the recurrence relation satisfied by the ratios $h_r \equiv p_r/p_{r-1}$. However, instead of assuming that the coefficients $a_r$, $b_r$ and $c_r$ in Eq. (4.1) are exactly 12, 25 and 13, respectively, we suppose that they are given in interval form $a_r = a$, $b_r = b$, $c_r = c$, all $r$, where

$$a = [11.9999, 12.0001], \quad b = [24.9998, 25.0002], \quad c = [12.9999, 13.0001].$$

The initial values $p_0 = 1$, $p_1 = 13/12$, are unchanged. (Of course, the method used in Example 1.1 would be just as unsuccessful with this modification.)

The recurrence formulae are given by

$$ah_r = b - (c/h_{r-1}), \qquad p_r = h_r p_{r-1}, \qquad r \geq 2. \tag{6.2}$$

Interval values of $p_r$ and $h_r$, computed with six-figure decimal arithmetic, are given in Table 6.2. These results obviously represent a considerable improvement on those found on Table 1.1. However, they are not entirely satisfactory for the following reason. The interval width $Ih_r$ of $h_r$ grows roughly in proportion to $r$: this can be seen from the entries in the penultimate column of Table 6.2. This linear growth in $Ih_r$ leads to an almost quadratic rate of growth in the corresponding relative errors of the $p_r$. This phenomenon is illustrated by the values of $Ip_r/(r^2\bar{p}_r)$ supplied in the last column of Table 6.2; here $\bar{p}_r$ denotes the midpoint of the interval value of $p_r$.

TABLE 6.2

*Interval solution of Eqs.* (6.2)

| $r$ | $h_r$ | | $p_r$ | | $Ih_r$ | $Ip_r/(r^2\bar{p}_r)$ |
|---|---|---|---|---|---|---|
| 0 | – | – | 1 | 1 | – | – |
| 1 | 1.08333 | 1.08334 | 1.08333 | 1.08334 | 0.00001 | 0.000009... |
| 2 | 1.08329 | 1.08338 | 1.17356 | 1.17367 | 0.00009 | 0.000023... |
| 3 | 1.08325 | 1.08342 | 1.27125 | 1.27158 | 0.00017 | 0.000028... |
| 4 | 1.08321 | 1.08346 | 1.37703 | 1.37771 | 0.00025 | 0.000030... |
| 5 | 1.08318 | 1.08349 | 1.49157 | 1.49274 | 0.00031 | 0.000031... |
| 6 | 1.08315 | 1.08352 | 1.61559 | 1.61742 | 0.00037 | 0.000031... |
| 7 | 1.08312 | 1.08355 | 1.74987 | 1.75256 | 0.00043 | 0.000031... |
| 8 | 1.08309 | 1.08357 | 1.89526 | 1.89903 | 0.00048 | 0.000031... |
| 9 | 1.08307 | 1.08360 | 2.05269 | 2.05779 | 0.00053 | 0.000030... |
| 10 | 1.08304 | 1.08362 | 2.22314 | 2.22987 | 0.00058 | 0.000030... |
| 11 | 1.08302 | 1.08364 | 2.40770 | 2.41638 | 0.00062 | 0.000029... |
| 12 | 1.08300 | 1.08366 | 2.60753 | 2.61854 | 0.00066 | 0.000029... |
| 13 | 1.08299 | 1.08368 | 2.82392 | 2.83766 | 0.00069 | 0.000028... |
| 14 | 1.08298 | 1.08370 | 3.05824 | 3.07518 | 0.00072 | 0.000028... |
| 15 | 1.08296 | 1.08371 | 3.31195 | 3.33261 | 0.00075 | 0.000027... |
| 16 | 1.08294 | 1.08372 | 3.58664 | 3.61162 | 0.00078 | 0.000027... |

*Example* 6.3. We solve the problem posed in Example 6.2 by the second method of Section 4. On omitting the suffix $r$ from the coefficients $a_r$, $b_r$, $c_r$, $\lambda_r$, $\mu_r$ and

$\nu_r$, we obtain the recurrence relations

$$(6.3) \qquad a u_r = \nu p_r + \mu u_{r-1}, \qquad p_{r+1} = u_r + \lambda p_r, \qquad r \geq 1,$$

in which

$$\lambda = b/(2a), \quad \mu = c/\lambda, \quad \nu = b - a\lambda - \mu.$$

The initial member of the sequence $\{u_r\}$ is given by $u_0 = p_1 - \lambda p_0$. Interval values of $\lambda, \mu$ and $\nu$ are found to be

$$\lambda = [1.04164, 1.04169], \quad \mu = [12.4797, 12.4803], \quad \nu = [0.0196000, 0.0204000],$$

and using six-figure decimal arithmetic we arrive at the interval values of $p_r$ and $u_r$ displayed in Table 6.3.

For large $r$, the intervals containing $p_r$ are narrower than those obtained in Table 6.2 but from the last column, in which $\bar{p}_r$ again denotes the mean value of $p_r$, it is evident that the growth of the relative error is still not linear in $r$.

TABLE 6.3

*Interval solution of Eqs.* (6.3)

| $r$ | $u_r$ | | $p_r$ | | $Ip_r/(r\bar{p}_r)$ |
|---|---|---|---|---|---|
| 0 | 0.0416400 | 0.0417000 | 1 | 1 | – |
| 1 | 0.0450735 | 0.0452113 | 1.08333 | 1.08334 | 0.00000... |
| 2 | 0.0487915 | 0.0490168 | 1.17350 | 1.17373 | 0.00009... |
| 3 | 0.0528176 | 0.0531412 | 1.27115 | 1.27169 | 0.00014... |
| 4 | 0.0571773 | 0.0576112 | 1.37689 | 1.37786 | 0.00017... |
| 5 | 0.0618983 | 0.0624557 | 1.49139 | 1.49293 | 0.00020... |
| 6 | 0.0670105 | 0.0677061 | 1.61538 | 1.61764 | 0.00023... |
| 7 | 0.0725463 | 0.0733965 | 1.74965 | 1.75279 | 0.00025... |
| 8 | 0.0785408 | 0.0795638 | 1.89504 | 1.89927 | 0.00027... |
| 9 | 0.0850317 | 0.0862483 | 2.05248 | 2.05803 | 0.00030... |
| 10 | 0.0920608 | 0.0934933 | 2.22297 | 2.23008 | 0.00031... |
| 11 | 0.0996716 | 0.101346 | 2.40759 | 2.41656 | 0.00033... |
| 12 | 0.107913 | 0.109856 | 2.60751 | 2.61866 | 0.00035... |
| 13 | 0.116838 | 0.119079 | 2.82399 | 2.83770 | 0.00037... |
| 14 | 0.126502 | 0.129076 | 3.05841 | 3.07509 | 0.00038... |
| 15 | 0.136967 | 0.139910 | 3.31226 | 3.33238 | 0.00040... |
| 16 | – | – | 3.58714 | 3.61122 | 0.00041... |

*Example* 6.4. We compute the absolute value of the Bessel function $Y_r(x)$ by forward recurrence from the relation

$$(6.4) \qquad p_{r+1} = (2r/x)p_r - p_{r-1}.$$

We take $x = 100$ and the initial values

$$(6.5) \quad p_{100} = -Y_{100}(100) = 0.166921\ldots, \quad p_{101} = -Y_{101}(100) = 0.200285\ldots.$$

Using six-decimal floating-point arithmetic, with chopping, we obtain the values $\bar{p}_r$ given in the second column of Table 6.4.

We shall compute both a posteriori and a priori error bounds by the methods of Section 5. These computations are carried out in four-decimal floating-point arithmetic with chopping. In the terminology of [12] this is the lower mode of computation ($\mathcal{L}$), and its associated wrp is $\gamma_\ell = 10^{-3}$. For the computation of the $\bar{p}_r$, the wrp is $\gamma = 10^{-5}$.

Both types of error bound require the evaluation of the bounds (5.19) for the coefficients $t_j$. The zeros of the local characteristic polynomial $z^2 - (2r/x)z + 1$ are given by

$$\alpha_r = (r/x) + \{(r/x)^2 - 1\}^{1/2}, \qquad \beta_r = (r/x) - \{(r/x)^2 - 1\}^{1/2}.$$

Consequently, for any $n$ exceeding 100, we have

$$\alpha = \alpha_{101} = 1.15177\dots, \qquad \mathrm{B} = \beta_{101} = 0.868225\dots;$$

compare (5.13). Also, from (5.17) and (6.5) we see that $\rho = \alpha$. From (5.19) we derive

$$t_{100} \le 3.061\dots; \qquad t_j \le 4.061\dots, \qquad j \ge 101. \tag{6.6}$$

For simplicity, however, we use the same bound for all $j$:

$$t_j < 4.062, \qquad j \ge 100. \tag{6.7}$$

For a posteriori error bounds we need to compute bounds for the quantities $\phi_r$ defined by

$$\bar{p}_r = \{(2r-2)/x\}\bar{p}_{r-1} - \bar{p}_{r-2} + \phi_r, \qquad r \ge 102;$$

compare (5.2). Since the coefficient $(2r-2)/x$ is exact, only two chopping errors are introduced at each recurrence step. Applying the methods of [12] we find that

$$|\phi_r| \le \bar{\bar{\chi}}_r \gamma e^{3\gamma\iota}, \qquad r \ge 102,$$

where

$$\chi_r = p_r + \{(2r-2)/x\}p_{r-1},$$

and the double bar signifies the value computed in $\mathcal{L}$. The rest of the computation proceeds in accordance with the relations (5.21), (5.22) and (5.23), as described in Section 5. The main steps are shown in columns 3, 4, 5 and 6 of Table 6.4: As in Example 6.1, the superscripts $\mathcal{A}$ and $\mathcal{B}$ signify upper and lower bounds respectively. Again, these bounds were computed using the methods of [12].

TABLE 6.4

*Bessel function* $-Y_r(x)$

| $r$ | $\bar{p}_r$ | $\bar{\bar{\chi}}_r$ | $10^5 T_r^{\mathcal{A}}$ | $p_r^{\mathcal{B}}$ | $10^5\lvert\epsilon_r\rvert^{\mathcal{A}}$ | $\lvert\epsilon_r/\bar{p}_r\rvert^{\mathcal{A}}$ | $\epsilon_r/\bar{p}_r$ | $\varpi_r^{\mathcal{A}}$ |
|---|---|---|---|---|---|---|---|---|
| 100 | 0.166921 | – | – | – | – | – | – | – |
| 101 | 0.200285 | – | 4.531 | – | – | – | – | – |
| 102 | 0.237654 | 0.6421 | 15.70 | 0.2375 | 3.755 | 0.00016 | 0.000004... | 0.00037 |
| 103 | 0.284529 | 0.7693 | 26.94 | 0.2844 | 7.710 | 0.00028 | 0.000007... | 0.00060 |
| 104 | 0.348475 | 0.9345 | 38.14 | 0.3482 | 13.35 | 0.00039 | 0.000011... | 0.00082 |
| 105 | 0.440299 | 1.165 | 49.20 | 0.4399 | 21.78 | 0.00050 | 0.000013... | 0.00104 |
| 106 | 0.576152 | 1.500 | 60.13 | 0.5757 | 34.86 | 0.00061 | 0.000016... | 0.00127 |
| 107 | 0.781141 | 2.002 | 70.94 | 0.7805 | 55.76 | 0.00072 | 0.000021... | 0.00149 |
| 108 | 1.09548 | 2.766 | 81.61 | 1.094 | 89.94 | 0.00083 | 0.000032... | 0.00172 |
| 109 | 1.58508 | 3.951 | 92.21 | 1.583 | 147.0 | 0.00093 | 0.000048... | 0.00194 |
| 110 | 2.35999 | 5.814 | 102.7 | 2.356 | 243.8 | 0.00104 | 0.000057... | 0.00216 |
| 111 | 3.60689 | 8.797 | 113.1 | 3.601 | 410.8 | 0.00114 | 0.000063... | 0.00239 |
| 112 | 5.64730 | 13.65 | 123.5 | 5.639 | 702.6 | 0.00125 | 0.000067... | 0.00261 |
| 113 | 9.04301 | 21.68 | 133.9 | 9.030 | 1218 | 0.00135 | 0.000074... | 0.00284 |
| 114 | 14.7899 | 35.21 | 144.3 | 14.75 | 2147 | 0.00146 | 0.000077... | 0.00306 |
| 115 | 24.6778 | 58.39 | 154.6 | 24.63 | 3840 | 0.00156 | 0.000084... | 0.00328 |
| 116 | 41.9690 | 98.71 | 164.8 | 41.89 | 6966 | 0.00166 | 0.000088... | 0.00351 |
| 117 | 72.6902 | 170.0 | 175.1 | 72.56 | 12810 | 0.00177 | 0.000090... | 0.00373 |
| 118 | 128.126 | 298.1 | – | – | 23920 | 0.00187 | 0.000092... | 0.00396 |

By way of comparison, the seventh column of Table 6.4 gives an upper bound $|\varepsilon_r/\bar{p}_r|^{\mathcal{A}}$ for the relative error. This is derived from the entries in the second and sixth columns. The next column gives the value of the actual relative error $\varepsilon_r/\bar{p}_r$ computed by use of high-precision values of $Y_r(100)$. Our bound overestimates the true error by a factor that ranges from about 35 at the beginning of the recurrences down to about 20 at $r = 118$. Two sources contribute to this factor. First, there is the radix effect associated with base 10. As we observed in Example 6.1, use of base 2 instead might save a factor of about 2 or 3. Secondly, we have used a uniform bound, given by (6.7), for the $t_j$. In fact, most of these coefficients are considerably less than 4.062. If desired, smaller bounds could be used without changing the $O(r)$ estimate of the total computing effort. For example, since the sequence $\beta_r$ is decreasing, it is easy to see that the second of the bounds (5.19) can be replaced by

$$t_j \leq \alpha/(\alpha - \beta_j), \qquad 101 \leq j \leq n.$$

The quantity $\alpha/(\alpha - \beta_j)$ has the values $2.258\ldots$ and $1.925\ldots$ at $j = 110$ and 118, respectively. Further sharpening is possible by application of the theorems given in [6, Section 5].

The final column of Table 6.4 gives a priori bounds $\varpi_r^{\mathcal{A}}$ for the relative precision of the approximation $\bar{p}_r$ to $p_r$. These were found as follows. Since the coefficients in (6.4) are exact, we have $\delta = 0$. Also, since $c_{j-1} = a_{j-1}$, all $j$, and only two chopping errors are made at each recurrence step, Eq. (5.24) may be replaced by

$$\varpi_r = 2\left\{(t_{100} + t_{101})\varpi + \gamma \sum_{j=102}^{r} t_j \left(2 + \frac{p_{j-2}}{p_j}\right)\right\}, \qquad j \geq 102.$$

On taking $\varpi = \gamma = 10^{-5}$, substituting for the $t_j$ by means of (6.6) and using the fact that $p_{j-2}/p_j \leq 1/\rho^2$, all $j$, we arrive at the numerical form

$$\varpi_r \leq \{14.25 + (22.40)(r - 101)\} \times 10^{-5}, \qquad r \geq 102.$$

As expected, the values of $\varpi_r^{\mathcal{A}}$ are approximately twice the size of the a posteriori relative error bounds $|\varepsilon_r/\bar{p}_r|^{\mathcal{A}}$.

**7. Conclusions.** We have described various methods for computing error bounds for solutions of difference equations of the form

$$a_r p_{r+1} = b_r p_r + c_r p_{r-1}$$

that are generated by forward recurrence. Two cases are considered: (A) oscillatory systems, in which $b_r^2 + 4a_r c_r < 0$, all $r$; (B) monotonic systems, in which $b_r^2 + 4a_r c_r \geq 0$, all $r$. In Case B methods have been provided for finding bounds of both a posteriori and a priori types. In Case A, only an a posteriori method is available, and there is a need for a method for constructing a priori bounds analogous to that of Section 5.

**Acknowledgments.** Helpful comments on this work have been made by J. R. Cash, E. Kaucher and D. W. Lozier, and especially the referee. I am also pleased to acknowledge the assistance of Dr. Lozier with the generation of high-precision values of the Legendre and Bessel functions needed in Examples 6.1 and 6.4.

University of Maryland
Institute for Physical Science and Technology
College Park, Maryland 20742

National Bureau of Standards
Mathematical Analysis Division
Gaithersburg, Maryland 20899

1. J. R. CASH, *Stable Recursions*, Academic Press, London, 1979.

2. W. GAUTSCHI, "Computational aspects of three-term recurrence relations," *SIAM Rev.*, v. 9, 1967, pp. 24–82.

3. W. GAUTSCHI, "Zur Numerik rekurrenter Relationen," *Computing*, v. 9, 1972, pp. 107–126. [Translated as Report ARL 73-0005, Aerospace Research Laboratories, Wright-Patterson Air Force Base, Ohio, 1973.]

4. W. GAUTSCHI, "Computational methods in special functions—a survey," in *Theory and Application of Special Functions* (R. A. Askey, ed.), Academic Press, New York, 1975, pp. 1–98.

5. N. J. HIGHAM, "Efficient algorithms for computing the condition number of a tridiagonal matrix," *SIAM J. Sci. Statist. Comput.*, v. 7, 1986, pp. 150–165.

6. R. M. M. MATTHEIJ, "Accurate estimates of solutions of second order recursions," *Linear Algebra Appl.*, v. 12, 1975, pp. 29–54.

7. R. M. M. MATTHEIJ & A. VAN DER SLUIS, "Error estimates for Miller's algorithm," *Numer. Math.*, v. 26, 1976, pp. 61–78.

8. R. E. MOORE, *Methods and Applications of Interval Analysis*, Society for Industrial and Applied Mathematics, Philadelphia, 1979.

9. F. W. J. OLVER, "Error analysis of Miller's recurrence algorithm," *Math. Comp.*, v. 18, 1964, pp. 65–74.

10. F. W. J. OLVER, "Numerical solution of second-order linear difference equations," *J. Res. Nat. Bur. Standards Sect. B*, v. 71, 1967, pp. 111–129.

11. F. W. J. OLVER, "Bounds for the solutions of second-order linear difference equations," *J. Res. Nat. Bur. Standards Sect. B*, v. 71, 1967, pp. 161–166.

12. F. W. J. OLVER, "Further developments of rp and ap error analysis, " *IMA J. Numer. Anal.*, v. 2, 1982, pp. 249–274.

13. F. W. J. OLVER & J. H. WILKINSON, "A posteriori error bounds for Gaussian elimination," *IMA J. Numer. Anal.*, v. 2, 1982, pp. 377–406.

14. S. M. RUMP, "Solving algebraic problems with high accuracy," in *A New Approach to Scientific Computation* (U. W. Kulisch and W. L. Miranker, eds.), Academic Press, New York, 1983, pp. 51–120.

15. A. VAN DER SLUIS, "Estimating the solutions of slowly varying recursions," *SIAM J. Math. Anal.*, v. 7, 1976, pp. 662–695.

16. R. TAIT, "Error analysis of recurrence relations," *Math. Comp.*, v. 21, 1967, pp. 629–638.

17. P. R. TURNER, "The distribution of leading significant digits," *IMA J. Numer. Anal.*, v. 2, 1982, pp. 407–412.

18. J. H. WILKINSON, *Rounding Errors in Algebraic Processes*, National Physical Laboratory Notes on Applied Science No. 32, Her Majesty's Stationery Office, London, 1963.

19. J. WIMP, *Computation with Recurrence Relations*, Pitman, Boston, 1984.