MATHEMATICS OF COMPUTATION, VOLUME 28, NUMBER 127, JULY 1974, PAGES 795-810

Numerical Stability in Evaluating Continued Fractions*

By William B. Jones and W. J. Thron

Abstract. A careful analysis of the backward recurrence algorithm for evaluating approximants of continued fractions provides rigorous bounds for the accumulated relative error due to rounding. Such errors are produced by machine operations which carry only a fixed number ν of significant digits in the computations. The resulting error bounds are expressed in terms of the machine parameter ν. The derivation uses a basic assumption about continued fractions, which has played a fundamental role in developing convergence criteria. Hence, its appearance in the present context is quite natural. For illustration, the new error bounds are applied to two large classes of continued fractions, which subsume many expansions of special functions of physics and engineering, including those represented by Stieltjes fractions. In many cases, the results insure numerical stability of the backward recurrence algorithm.

1. Introduction. The analytic theory of continued fractions provides a useful means for representation and continuation of special functions of mathematical physics [1], [2], [10]. Many applications of continued fractions and the closely related Padé approximants have recently been made in various areas of numerical analysis and of theoretical physics, chemistry and engineering [4], [5], [7]. Thus, it is important to establish a sound understanding of the basic computational problems associated with continued fractions. The present paper is written to help fulfill that aim.

A number of procedures for calculating the nth approximant f_n of a continued fraction

$$\frac{a_1}{b_1} + \frac{a_2}{b_2} + \frac{a_3}{b_3} + \cdots \tag{1.1}$$

are found in the literature. For example, the "forward recurrence algorithm" (F-R algorithm) consists in applying the well-known second-order linear difference equations

$$\begin{aligned} A_k &= b_k A_{k-1} + a_k A_{k-2}, \\ B_k &= b_k B_{k-1} + a_k B_{k-2}, \end{aligned} \qquad k = 1, \ldots, n, \tag{1.2a}$$

and initial conditions

$$A_{-1} = 1, \quad A_0 = 0, \quad B_{-1} = 0, \quad B_0 = 1 \tag{1.2b}$$

to give $f_n = A_n/B_n$.

Received August 24, 1973.

AMS (MOS) subject classifications (1970). Primary 30A22, 41A20; Secondary 30A22, 65G05.

Key words and phrases. Numerical stability, error analysis, continued fraction, special functions.

* Research sponsored by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under AFOSR Grant No. AFOSR-70-1888. The United States Government is authorized to reproduce and distribute reprints for Government purposes nonwithstanding any copyright notation hereon.

Copyright © 1974, American Mathematical Society

The so-called "backward recurrence algorithm" (B-R algorithm) consists of the following: Set

$$G_{n+1}^{(n)} = 0 \tag{1.3a}$$

and compute successively from "tail to head"

$$G_k^{(n)} = a_k/(b_k + G_{k+1}^{(n)}), \qquad k = n, n-1, \ldots, 1, \tag{1.3b}$$

to obtain $f_n = G_1^{(n)}$. Other algorithms have been based on the well-known series formula

$$f_n = \frac{A_n}{B_n} = \sum_{k=1}^{n} \left(\frac{A_k}{B_k} - \frac{A_{k-1}}{B_{k-1}}\right) = \sum_{k=1}^{n} \frac{(-1)^{k+1} \prod_{j=1}^{k} a_j}{B_k B_{k-1}}$$

(see, for example, [2], [3]). We omit a detailed description of these algorithms since they are not dealt with further in this paper.

The computation of a single approximant f_n by the F-R algorithm requires $4n + 1$ operations of multiplication or division, whereas only n such operations are used by the B-R algorithm. Thus, B-R is computationally more efficient if only one approximant is required. On the other hand, if one wishes to obtain n successive approximants $f_1, \cdots, f_n$, the F-R algorithm is more efficient since it requires only $5n$ operations of multiplication or division compared to $\frac{1}{2}n(n + 1)$ such operations for the B-R algorithm. This difference is due to the fact that the F-R algorithm has a carry-over of results from one approximant to the next which is not enjoyed by the B-R algorithm.

From the viewpoint of numerical stability, however, the F-R algorithm has inherent problems which the B-R algorithm does not appear to possess. One troublesome factor is that, although the sequence $\{f_n\}$ may converge to a finite limit, A_n and B_n may both tend to infinity or to zero, thus making it necessary to re-scale from time to time to prevent machine overflow or underflow. A more serious difficulty of the F-R algorithm is the tendency of rounding error to accumulate in successive application of the three-term recurrence relations (1.2). Some of the dangers of numerical instability associated with three-term recurrence relations have been discussed by Gautschi [3]. Blanch [2] has given an analysis of rounding errors which seems to indicate that the B-R algorithm is numerically more stable than the F-R algorithm. An illustration of this phenomenon is given by the numerical example in Section 2. Computations made for the convergent continued fraction

$$-.5 = \frac{-.25}{1} + \frac{-.25}{1} + \frac{-.25}{1} + \cdots \tag{1.4}$$

show that rounding error accumulates significantly from the F-R algorithm but not from the B-R algorithm.

The primary purpose of the present paper is to give explicit and precise upper bounds for the rounding error produced by the B-R algorithm. Our main results are contained in Theorems 3.1 and 4.3. The first of these is a general result which makes no special assumptions about the continued fraction. It evolved out of work included in [2]. (The problem has also been attacked in [8].) Theorem 4.3 provides methods for estimating a basic quantity $g_k^{(n)}$ needed to apply Theorem 3.1. The main assumption about continued fractions in Theorem 4.3 is the existence of a sequence of subsets $\{V_n\}$ of the extended complex plane such that, for all n,

$$0 \in V_n, \tag{1.5a}$$

and

$$a_n/(b_n + V_n) \subseteq V_{n-1} \tag{1.5b}$$

(see the discussion at the end of this section for the meaning of (1.5b)). Property (1.5) has played a fundamental role in developmg much of the known convergence theory of continued fractions ([6], [9]). Hence its occurrence here is quite natural. Some examples of applications of Theorems 3.1 and 4.3 are included in Section 5. Two large classes of continued fractions are considered: (a) Stieltjes fractions (subsection 5.1) and (b) a class which subsumes all convergent Stieltjes fractions and a larger subclass of the positive-definite continued fractions (subsection 5.2).

Before proceeding with the main body of the paper, we state for later use some definitions and notation employed. A *continued fraction* is an ordered triple of sequences $[\{a_n\}, \{b_n\}, \{f_n\}]$ such that, for each $n = 1, 2, 3, \ldots, a_n$ and b_n are complex numbers ($a_n \neq 0$) and f_n is defined as follows: Set

$$s_n(\zeta) = a_n/(b_n + \zeta), \qquad n = 1, 2, \ldots, \tag{1.6a}$$

and

$$S_1(\zeta) = s_1(\zeta); \quad S_n(\zeta) = S_{n-1}(s_n(\zeta)), \qquad n = 2, 3, \ldots. \tag{1.6b}$$

Then

$$f_n = S_n(0), \qquad n = 1, 2, \ldots.$$

The numbers a_n, b_n are called the *elements* and f_n the nth approximant of the continued fraction. A continued fraction is said to *converge* if its sequence of approximants converges. When convergent, a continued fraction has as its value $\lim f_n$. For convenience, other symbols are sometimes used to denote the continued fraction $[\{a_n\}, \{b_n\}, \{f_n\}]$ such as $K(a_n/b_n)$ and (1.1).

If g is a function and A is a subset of the extended complex plane, we mean by $g(A)$ the set $\{w: w = g(z), z \in A\}$. By $d(z, A)$, we mean the distance from point z to set A.

2. A Numerical Example. To illustrate numerical stability of the B-R algorithm and instability of the F-R algorithm, a numerical example is described in this section. Although such stability in the B-R algorithm cannot (at this point) be guaranteed for all continued fractions, the following sections show that it will occur in many cases. The continued fraction employed in the present example is (1.4). Values of the nth approximant f_n, $n = 1, \ldots, 15$, correctly rounded in the fifth decimal place, are given in Table 1. Also given are approximations to f_n obtained from the F-R algorithm (f_n^*) and from the B-R algorithm ($\hat{f}_n$), using floating-point arithmetic with 5-digit mantissas. It can be seen that the accumulative rounding error $f_n - f_n^*$ (F-R algorithm) grows steadily as n increases, starting at $n = 6$. At $n = 7$, f_n^* is correctly rounded only in the third decimal place. On the other hand, $\hat{f}_n$ obtained by the B-R algorithm is correctly rounded to 5 decimal places for $n = 1, \ldots, 15$, except for $n = 11$, where $\hat{f}_{11}$ is off by one unit in the fifth decimal place. Since the later values of $\hat{f}_n$ $(n > 11)$ are correctly rounded in the fifth place, the B-R algorithm appears to be self-correcting, at least in this example. Further calculations of f_n, f_n^* and $\hat{f}_n$, for $n = 1, 2, \ldots, 40$, showed that $f_n - f_n^*$ does not continue to increase indefinitely. A maximum error of .00031 is reached at $n = 22$. For $n > 22$, $f_n - f_n^*$ decreases to the value .00014 at $n = 40$. In the case of the B-R algorithm, $\hat{f}_n$ remains correctly rounded in the fifth decimal place for $1 \leqq n \leqq 40$, $n \neq 11$. This example is considered again in subsection 5.2. Using Theorems 3.1 and 4.3, we obtain rigorous bounds for the relative rounding error $|f_n - \hat{f}_n|/|f_n|$, which are consistent with those found numerically in the present example.

n	f_n	f_n^*	$f_n - f_n^*$	$\hat{f}_n$	$f_n - \hat{f}_n$
1	- .25000	- .25000	.00000	- .25000	.00000
2	- .33333	- .33333	.00000	- .33333	.00000
3	- .37500	- .37500	.00000	- .37500	.00000
4	- .40000	- .40000	.00000	- .40000	.00000
5	- .41667	- .41667	.00000	- .41667	.00000
6	- .42857	- .42859	.00002	- .42857	.00000
7	- .43750	- .43757	.00007	- .43750	.00000
8	- .44444	- .44452	.00008	- .44444	.00000
9	- .45000	- .45010	.00010	- .45000	.00000
10	- .45455	- .45467	.00012	- .45455	.00000
11	- .45833	-. 45849	.00016	- .45844	.00001
12	- .46154	- .46173	.00019	- .46154	.00000
13	- .46429	- .46449	.00020	- .46429	.00000
14	- .46667	- .46690	.00023	- .46667	.00000
15	- .46875	- .46900	.00025	- .46875	.00000

TABLE 1. *Computation of Approximants for the Convergent Continued Fraction* $K(-.25/1) = -.5$.

f_n equals the nth approximant correctly rounded in the fifth decimal place.

f_n^* equals the approximation to f_n by the F-R algorithm.

$\hat{f}_n$ equals the approximation to f_n by the B-R algorithm.

Both f_n^* and $\hat{f}_n$ are obtained with floating-point arithmetic using 5-digit mantissas.

3. Estimates of Relative Rounding Error. In this section, we establish general estimates of relative rounding error produced by the B-R algorithm in calculating an nth approximant. The following notation is used: For each $k = 1, \dots, n$, let $\hat{a}_k$ and $\hat{b}_k$ denote rounded values of the elements a_k and b_k, respectively, of a given continued fraction (1.1). Let α_k and β_k denote the relative error in $\hat{a}_k$ and $\hat{b}_k$, respectively, so that

$$\hat{a}_k = a_k(1 + \alpha_k), \quad \hat{b}_k = b_k(1 + \beta_k). \tag{3.1}$$

Similarly, let $\epsilon_k^{(n)}$ denote the relative error in $\hat{G}_k^{(n)}$, the approximation to $G_k^{(n)}$ obtained from (1.3) using "machine numbers" $\hat{a}_k$ and $\hat{b}_k$ and machine operations which carry only a fixed number of significant digits in the computations. Thus

$$\hat{G}_k^{(n)} = G_k^{(n)}(1 + \epsilon_k^{(n)}), \qquad k = 1, \dots, n, \tag{3.2a}$$

and

$$\hat{G}_{n+1}^{(n)} = G_{n+1}^{(n)} = \epsilon_{n+1}^{(n)} = 0. \tag{3.2b}$$

Further, let $\gamma_k^{(n)}$ denote the relative error produced in the computation of $\hat{G}_k^{(n)}$ from $\hat{a}_k$, $\hat{b}_k$ and $\hat{G}_{n+1}^{(n)}$, so that

$$\hat{G}_k^{(n)} = \hat{a}_k(1 + \gamma_k^{(n)})/(\hat{b}_k + \hat{G}_{k+1}^{(n)}), \qquad k = 1, \dots, n. \tag{3.3}$$

Combining (3.2) and (3.3) with

$$g_k^{(n)} = G_{k+1}^{(n)}/(b_k + G_{k+1}^{(n)}), \qquad k = 1, \dots, n, \tag{3.4}$$

one easily obtains the relation

$$\epsilon_k^{(n)} = \frac{(1 + \alpha_k)(1 + \gamma_k^{(n)})}{1 + \beta_k + g_k^{(n)}(\epsilon_{k+1}^{(n)} - \beta_k)} - 1,$$

or

$$\epsilon_k^{(n)} = \frac{\alpha_k - \beta_k + \gamma_k^{(n)} + \alpha_k\gamma_k^{(n)} - g_k^{(n)}(\epsilon_{k+1}^{(n)} - \beta_k)}{1 + \beta_k + g_k^{(n)}(\epsilon_{k+1}^{(n)} - \beta_k)}, \qquad k = 1, \dots, n. \tag{3.5}$$

Our interest is in estimating the number $\epsilon_k^{(n)}$ and, particularly, $\epsilon_1^{(n)}$, the relative error in the machine approximation $\hat{f}_n = \hat{G}_1^{(n)}$. Such estimates are provided by the following:

THEOREM 3.1. *For each $k = 1, \dots, n$, let $\epsilon_k^{(n)}$ satisfy* (3.5), *with $g_n^{(n)} = \epsilon_{n+1}^{(n)} = 0$. Further, let nonnegative numbers α, β, γ, η and ω be chosen such that, for $k = 1, \dots, n$,*

$$|\alpha_k| \leqq \alpha\omega, \quad |\beta_k| \leqq \beta\omega, \quad |\gamma_k^{(n)}| \leqq \gamma\omega, \quad |g_k^{(n)}| \leqq \eta, \tag{3.6a}$$

where

$$\begin{array}{lcl} \alpha = 0 & or & \alpha \geqq 1, \\ \beta = 0 & or & \beta \geqq 1, \\ \gamma \geqq 1, & & \eta > 0, \end{array} \tag{3.6b}$$

$$\alpha + \beta + \gamma \geqq 2.$$

Then

$$|\epsilon_k^{(n)}| \leqq \omega(1 + \alpha + \beta + \gamma + \beta\eta) \sum_{j=0}^{n-k} \eta^j, \tag{3.7}$$

provided that

$$0 \leqq \omega < 1/16(\alpha + \beta + \gamma)^2, \tag{3.8a}$$

and

$$0 \leqq \omega < \left(2\left[1 + \beta + \beta\eta + \eta(1 + \alpha + \beta + \gamma + \beta\eta) \sum_{j=0}^{n-2} \eta^j\right]^2\right)^{-1} \tag{3.8b}$$

Remarks. (1) Typically, ω will equal $(\frac{1}{2})10^{1-\nu}$, where ν is the number of significant decimal digits carried in the (machine) computation.

Then, for continued fractions of the form $K(a_n/b_n)$, one has $\alpha = \beta = 1$ and $\gamma = 2$ so that (3.7) gives

$$|\epsilon_1^{(n)}| \leqq (5 + \eta)\omega \sum_{j=0}^{n-1} \eta^j. \tag{3.9}$$

If η can be chosen such that $\eta \leqq 1$, then (3.7) gives

$$|\epsilon_1^{(n)}| \leqq 6n\omega, \tag{3.10}$$

implying that, at worst, the rounding error can grow fairly slowly. Moreover, if one can choose η such that $0 < \eta < 1$, then (3.7) gives

$$|\epsilon_1^{(n)}| < 6\omega/(1 - \eta),$$

insuring numerical stability of the B-R algorithm. It will be shown (subsection 5.1) that this is indeed the case for a great many continued fraction expansions.

(2) Slightly smaller error bounds can be obtained in certain special cases. For continued fractions of the form $K(a_n/1)$, we have $b_k = 1$, so that we may choose $\beta = 0$ and $\alpha = \gamma = 1$; whence (3.7) gives

$$|\epsilon_1^{(n)}| \leqq 3\omega \sum_{j=0}^{n-1} \eta^j. \tag{3.11}$$

Similarly, for continued fractions of the form $K(1/b_n)$, we have $a_k = 1$ so that $\alpha = 0$, $\beta = 1$ and $\gamma = 2$; thus, (3.7) gives

$$|\epsilon_1^{(n)}| \leqq (4 + \eta)\omega \sum_{j=0}^{n-1} \eta^j. \tag{3.12}$$

Hence, computationally, the form $K(a_n/1)$ appears to be somewhat preferable.

(3) The key obstacle in applying Theorem 3.1 is the determination of good estimates of the quantities $g_k^{(n)}$. Methods for obtaining such estimates are given in Section 4 and illustrated in Section 5.

Proof of Theorem 3.1. Set

$$C_k^{(n)} = (1 + \alpha + \beta + \gamma + \beta\eta) \sum_{j=0}^{n-k} \eta^j, \qquad k = 1, \dots, n. \tag{3.13}$$

The proof of the theorem then consists in showing that

$$|\epsilon_k^{(n)}| \leqq \omega C_k^{(n)}, \qquad k = 1, \dots, n. \tag{3.14}$$

It is convenient to define

$$h_k^{(n)} = \beta + \beta\eta + \eta C_k^{(n)}, \qquad k = 1, \dots, n. \tag{3.15}$$

Since, for all $2 \leqq k \leqq n$, we have $C_k^{(n)} \leqq C_2^{(n)}$, it follows from (3.8b) that

$$2\omega[h_k^{(n)}]^2 \leqq 1, \qquad k = 2, 3, \dots, n. \tag{3.16}$$

Next, note that, for $x \geqq 0$,

$$1/(1 - \omega x) \leqq 1 + \omega x + 2\omega^2 x^2 \tag{3.17}$$

is valid, provided $2\omega x \leqq 1$. This can be seen from the identity

$$(1 - \omega x)(1 + \omega x + 2\omega^2 x^2) = 1 + \omega^2 x^2 (1 - 2\omega x).$$

The proof of the theorem consists of a backward induction on k, starting with $k = n$. We have from (3.5) and $g_n^{(n)} = 0$ that

$$\epsilon_n^{(n)} = (1 + \alpha_n)(1 + \gamma_n^{(n)})/(1 - \beta_n) - 1.$$

Since, by (3.8a), $\beta\omega < 1$, we then have

$$|\epsilon_n^{(n)}| \leqq \omega(\alpha + \beta + \gamma + \alpha\beta\omega)/(1 - \beta\omega).$$

Agai ., oy (3.8a), $2\beta\omega < 1$; hence, by (3.17),

$$\begin{aligned} |\epsilon_n^{(n)}| &\leqq \omega(\alpha + \beta + \gamma + \alpha\beta\omega)(1 + \beta\omega + 2\beta^2\omega^2) \\ &= \omega(\alpha + \beta + \gamma) + \omega^2(\alpha\beta + \beta^2 + \gamma^2 + \alpha\gamma) \\ &\quad + \omega^3(\alpha\beta\gamma + 2\alpha\beta^2 + 2\beta^3 + 2\beta^2\gamma) + \omega^4(2\alpha\beta^2\gamma) \\ &\leqq \zeta + \zeta^2 + \zeta^3 + \zeta^4, \end{aligned}$$

where $\zeta = \omega(\alpha + \beta + \gamma)$. It follows that

$$|\epsilon_n^{(n)}| \leqq \omega(\alpha + \beta + \gamma + 1), \tag{3.18}$$

provided $\zeta + \zeta^2 + \zeta^3 + \zeta^4 \leqq \zeta + \omega$. But this is implied by (3.8a) and the hypothesis $\alpha + \beta + \gamma \geqq 2$. It follows from (3.18) that (3.7) is satisfied with $k = n$.

Now we assume that

$$(3.19)\qquad |\epsilon_{k+1}^{(n)}| \leqq \omega C_{k+1}^{(n)}$$

for some value of k such that $1 \leqq k \leqq n-1$. Then, from (3.5), one easily obtains

$$(3.20)\qquad \frac{|\epsilon_k^{(n)}|}{\omega} \leqq \frac{\alpha+\gamma+\alpha\gamma\omega+h_{k+1}^{(n)}}{1-\omega h_{k+1}^{(n)}},$$

since (3.8b) insures that $\omega h_{k+1}^{(n)} \leqq 1$. But since (3.8b) also implies that $2\omega h_{k+1}^{(n)} \leqq 1$, it follows from (3.17) and (3.20) that

$$\begin{aligned}|\epsilon_k^{(n)}|/\omega &\leqq (\alpha+\gamma+\alpha\gamma\omega+h_{k+1}^{(n)})(1+\omega h_{k+1}^{(n)}+2\omega^2[h_{k+1}^{(n)}]^2)\\ &= (\alpha+\gamma+h_{k+1}^{(n)})+\omega[\alpha\gamma+(\alpha+\beta)h_{k+1}^{(n)}+(h_{k+1}^{(n)})^2]\\ &\quad+\omega^2[\alpha\gamma h_{k+1}^{(n)}+2(\alpha+\gamma)(h_{k+1}^{(n)})^2+2(h_{k+1}^{(n)})^3]+\omega^3[2\alpha\gamma(h_{k+1}^{(n)})^2].\end{aligned}$$

Using (3.16), we then obtain

$$(3.21\text{a})\qquad |\epsilon^{(n)}|/\omega \leqq \alpha+\gamma+h_{k+1}^{(n)}+\Delta_k^{(n)},$$

where

$$(3.21\text{b})\qquad \begin{aligned}\Delta_k^{(n)} &\leqq 1/2+\frac{1}{\sqrt{2}}(\alpha+\gamma+1)\omega^{1/2}+(\alpha+\gamma+\alpha\gamma)\omega+\frac{1}{\sqrt{2}}\alpha\gamma\omega^{3/2}+\alpha\gamma\omega^2\\ &\leqq \tfrac{1}{2}+\omega^{1/2}+(\alpha+\gamma)\omega^{1/2}+(\alpha+\gamma)^2(\omega+\omega^{3/2}+\omega^2).\end{aligned}$$

It can be shown that $\Delta_k^{(n)} \leqq 1$. In fact, it follows from (3.8a) and the hypothesis $(\alpha+\beta+\gamma) \geqq 2$ that $\omega \leqq 2^{-6}$, so that $\omega^{1/2} \leqq 2^{-3}$. Moreover, (3.8a) implies that $(\alpha+\gamma)\omega^{1/2} \leqq 2^{-2}$, $(\alpha+\gamma)^2\omega \leqq 2^{-4}$, $(\alpha+\gamma)^2\omega^{3/2} \leqq 2^{-6}$ and $(\alpha+\gamma)^2\omega^2 \leqq 2^{-8}$. Hence, by (3.21a),

$$\frac{|\epsilon_k^{(n)}|}{\omega} \leqq 1+\alpha+\gamma+h_{k+1}^{(n)} = (1+\alpha+\beta+\gamma+\beta\eta)\sum_{j=0}^{n-k}\eta^j = C_k^{(n)}.$$

4. Methods for Estimating $g_k^{(n)}$. Application of Theorem 3.1 requires that estimates be found for the quantities $g_k^{(n)}$ defined by (1.3) and (3.4). Methods for obtaining such estimates are described in this section (Theorem 4.3). At the outset, we prove (Theorem 4.1) that $g_k^{(n)}$ is *invariant under equivalence transformations of continued fractions*. This significant property shows that there is no need to search for an optimal form of a continued fraction from the point of view of minimizing estimates of $g_k^{(n)}$.

THEOREM 4.1. *Let $K(a_n/b_n)$ and $K(a_n^*/b_n^*)$ be equivalent continued fractions, so that there exists a sequence of nonzero constants $\{r_n\}$ satisfying, for $n = 1, 2, \ldots,$*

$$(4.1)\qquad \begin{aligned}a_n &= r_n r_{n-1} a_n^* \qquad (r_0 = 1),\\ b_n &= r_n b_n^*.\end{aligned}$$

For $n = 1, 2, \ldots,$ and $k = 1, \ldots, n$, let

$$(4.2)\qquad g_k^{(n)} = G_{k+1}^{(n)}/(b_k+G_{k+1}^{(n)}) \quad \textit{and} \quad g_k^{(n)*} = G_{k+1}^{(n)*}/(b_k^*+G_{k+1}^{(n)*}),$$

where

$$(4.3)\qquad \begin{aligned} G_{n+1}^{(n)} &= 0, \qquad G_k^{(n)} = a_k/(b_k + G_{k+1}^{(n)}),\\ G_{n+1}^{(n)*} &= 0, \qquad G_k^{(n)*} = a_k^*/(b_k^* + G_{k+1}^{(n)*}), \end{aligned}$$

then

$$(4.4)\qquad G_k^{(n)} = r_{k-1} G_k^{(n)*}, \qquad k = 1, \ldots, n, \quad n = 1, 2, \ldots,$$

and

$$(4.5)\qquad g_k^{(n)} = g_k^{(n)*}, \qquad k = 1, \ldots, n, \quad n = 1, 2, \ldots.$$

Proof. First, we prove (4.4) for fixed n by a backward induction on k, starting with $k = n$. Using (4.1), we obtain

$$G_n^{(n)} = a_n/b_n = r_n r_{n-1} a_n^*/r_n b_n^* = r_{n-1} G_n^{(n)*}.$$

Now we assume that, for some k such that $0 < k < n - 1$, $G_{k+1}^{(n)} = r_k G_{k+1}^{(n)*}$. Then, again using (4.1) we obtain

$$G_k^{(n)} = \frac{a_k}{b_k + G_{k+1}^{(n)}} = \frac{r_k r_{k-1} a_k^*}{r_k b_k^* + r_k G_{k+1}^{(n)*}} = r_{k-1} G_k^{(n)*},$$

as asserted by (4.4). The proof of (4.5) follows immediately from (4.1), (4.2) and (4.4).

It was mentioned in the introduction that many of the known classes of convergent continued fractions satisfy properties of the general form

$$(4.6)\qquad s_n(V_n) \subseteq V_{n-1}, \qquad n = 1, 2, \ldots,$$

where $s_n(\zeta) = a_n/(b_n + \zeta)$ and $\{V_n\}$ is a sequence of subsets of the extended plane. It will now be seen that (4.6) also plays a basic role in obtaining estimates of $g_k^{(n)}$. We begin with the following:

LEMMA 4.2. *Let*

$$(4.7)\qquad f_n = \frac{a_1}{b_1} + \frac{a_2}{b_2} + \cdots + \frac{a_n}{b_n}$$

be given and let $V_1, \ldots, V_n$ *be subsets of the extended complex plane such that*

$$(4.8)\qquad 0 \in V_n$$

and

$$(4.9)\qquad s_k(V_k) = a_k/(b_k + V_k) \subseteq V_{k-1}, \qquad k = 2, \ldots, n.$$

If $G_k^{(n)}$ *is defined by* (1.3), *then*

$$(4.10)\qquad G_k^{(n)} \in V_{k-1}, \qquad k = 2, 3, \ldots, n, n + 1,$$

and hence

$$(4.11) \qquad |b_k + G_{k+1}^{(n)}| \geqq d(-b_k, V_k), \qquad k = 1, \ldots, n.$$

Proof. The proof of (4.10) is by a backward induction on k, starting with $k = n + 1$. By use of (4.8) and (1.3), we obtain $G_{n+1}^{(n)} = 0 \in V_n$. Now if we assume that, for some k such that $1 \leqq k \leqq n - 1$, $G_{k+1}^{(n)} \in V_k$, then again using (4.9), we obtain

$$G_k^{(n)} = s_k(G_{k+1}^{(n)}) \in s_k(V_k) \subseteq V_{k-1},$$

which proves (4.10). Assertion (4.11) follows from (4.10). This completes the proof.

THEOREM 4.3. *Let*

$$(4.12) \qquad f_n = \frac{a_1}{b_1} + \frac{a_2}{b_2} + \cdots + \frac{a_n}{b_n}$$

be given and let $V_1, \ldots, V_n$ be subsets of the extended complex plane such that

$$(4.13) \qquad 0 \in V_n$$

and

$$(4.14) \qquad s_k(V_k) = a_k/(b_k + V_k) \subseteq V_{k-1}, \qquad k = 2, \ldots, n.$$

Further, let

$$(4.15) \qquad A^{(n)} = \inf_{2 \leqq k \leqq n} |a_k|,$$

$$(4.16) \qquad \delta^{(n)} = \max_{1 \leqq k \leqq n} d(-b_k, V_k),$$

and

$$(4.17) \qquad M^{(n)} = \max\{|w| : w \in V_k/(b_k + V_k), k = 1, 2, \ldots, n\}.$$

If $g_k^{(n)}$ is defined by (3.4) *and* (1.3), *then*

$$(4.18) \qquad |g_k^{(n)}| \leqq A^{(n)}/(\delta^{(n)})^2, \qquad k = 1, 2, \ldots, n \quad (\textit{Method A})$$

and

$$(4.19) \qquad |g_k^{(n)}| \leqq M^{(n)}, \qquad k = 1, 2, \ldots, n \quad (\textit{Method B}).$$

Proof. By (3.4) and (1.3), we have

$$|g_k^{(n)}| = \left|\frac{G_{k+1}^{(n)}}{b_k + G_{k+1}^{(n)}}\right| = \left|\frac{a_{k+1}}{(b_k + G_{k+1}^{(n)})(b_{k+1} + G_{k+2}^{(n)})}\right|, \qquad k = 1, \ldots, n - 1.$$

Hence, from (4.15), (4.16) and Lemma 4.2, we obtain (4.18). Inequality (4.19) follows immediately from (3.4), (4.10) and (4.17).

Some examples of applications of Methods A and B will be given in the following section. It will be seen that for certain situations Method A is preferable to Method B and vice versa.

5. Applications. To illustrate the use of Theorem 4.3, we now obtain explicit bounds for $g_k^{(n)}$ for two important classes of continued fractions: (a) Stieltjes fractions and (b) a class associated with parabolic convergence regions.

5.1. *Stieltjes Fractions.* A Stieltjes fraction is a continued fraction of the form

$$\frac{a_1 z}{1} + \frac{a_2 z}{1} + \frac{a_3 z}{1} + \cdots \qquad (a_n > 0), \tag{5.1}$$

or one that can be put into the form (5.1) by an equivalence transformation. It is well known [9] that if (5.1) converges at a single point z ($z \neq 0$), then it converges at every point z in the cut plane $|\arg z| < \pi$. Its limit $f(z)$ can then be represented by a Stieltjes integral

$$f(z) = z \int_0^\infty \frac{d\Psi(t)}{1 + zt} \tag{5.2}$$

where $\Psi(t)$ is a bounded, nondecreasing real-valued function with infinitely many points of increase on $[0, \infty)$. Some examples of functions known to possess Stieltjes fraction representations include: exponential integrals, incomplete gamma functions, the logarithm of the gamma function, the error function, ratios of successive Bessel functions of the first kind and various elementary transcendental functions. The reader is referred to standard references [1], [2], [10] for explicit formulas and other examples. Our purpose here is to obtain bounds for $|g_k^{(n)}|$ in terms of the complex variable z and coefficients a_k of the Stieltjes fraction (5.1). The main results are summarized in the following:

THEOREM 5.1. *Let*

$$f_n = \frac{a_1 z}{1} + \frac{a_2 z}{1} + \cdots + \frac{a_n z}{1}, \tag{5.3}$$

where

$$0 < a_k \leqq A, \qquad k = 1, \ldots, n, \tag{5.4}$$

$$z = re^{i\theta}, \quad r > 0, \quad |\theta| < \pi. \tag{5.5}$$

Further, let

$$G_k^{(n)} = \frac{a_k z}{1} + \cdots + \frac{a_n z}{1}, \qquad k = 1, \ldots, n \quad (G_{n+1}^{(n)} = 0), \tag{5.6}$$

and

$$g_k^{(n)} = G_{k+1}^{(n)}/(1 + G_{k+1}^{(n)}), \qquad k = 1, \ldots, n. \tag{5.7}$$

(i) *If* $|\theta| \leqq \pi/2$, *then, for* $k = 1, \ldots, n$,

$$|g_k^{(n)}| \leqq Ar/(1 + 2Ar\cos\theta + A^2r^2)^{1/2} < 1. \tag{5.8}$$

(ii) *If* $\pi/2 < |\theta| < \pi$, *then, for* $k = 1, \ldots, n$,

$$|g_k^{(n)}| \leqq \frac{Ar}{1 + 2Ar\cos\theta + A^2r^2}, \quad \textit{provided } Ar < \cos(\pi - \theta), \tag{5.9a}$$

$$|g_k^{(n)}| \leqq \frac{Ar}{(1 + 2Ar\cos\theta + A^2r^2)^{1/2}}, \quad \textit{provided} \ \cos(\pi - \theta), \tag{5.9b}$$

$$\leqq Ar \leqq \sec(\pi - \theta),$$

$$|g_k^{(n)}| \leqq Ar\csc^2\theta, \ \textit{provided} \ \cos(\pi - \theta) \leqq Ar. \tag{5.9c}$$

Remarks. If $\pi/2 < |\theta| < \pi$ and $Ar < \frac{1}{2}\sec(\pi - \theta)$, then $|w_0| < 1$ (see (5.15)). For

$$\max\{\cos(\pi - \theta), \tfrac{1}{2}\sec(\pi - \theta)\} \leqq Ar \leqq \sec(\pi - \theta)$$

it is still true that $|g_k^{(n)}| \leqq |w_0|$, but then $|w_0| \geqq 1$.

Our proof of Theorem 5.1 is based on Theorem 4.3 and the following two lemmas:

LEMMA 5.2. *Let f_n and $G_k^{(n)}$ be defined as in Theorem* 5.1. *Then*

$$a_k z/(1 + V) \subseteq V, \qquad k = 1, \ldots, n, \tag{5.10}$$

where $V = V(A, r, \theta)$ is the convex lens-shaped region (Fig. 1) *(with interior angle) $|\theta|$, bounded by the ray issuing from the origin in the direction θ, and the circular arc starting at the origin, tangent to the real axis, and extending to the point $Are^{i\theta}$. Further,*

$$G_k^{(n)} \in V, \qquad k = 1, \ldots, n. \tag{5.11}$$

Proof. The region $1 + V$ has the two points 1 and $1 + Are^{i\theta}$ as vertices. Hence the lens-shaped region $1/(1 + V)$ has the vertices 1 and $1/(1 + Are^{i\theta})$; it is contained in the lens-shaped region X which is bounded by the real axis and the circular arc passing through 0 at an angle $-\theta$ with respect to the real axis and passing through 1. The point $1/(1 + Are^{i\theta})$ is located on this circular arc. Clearly $a_m zX \subseteq V$ for $m = 1, 2, \ldots, n$, since $z = -e^{i\theta}$ and $0 < a_m \leqq A$.

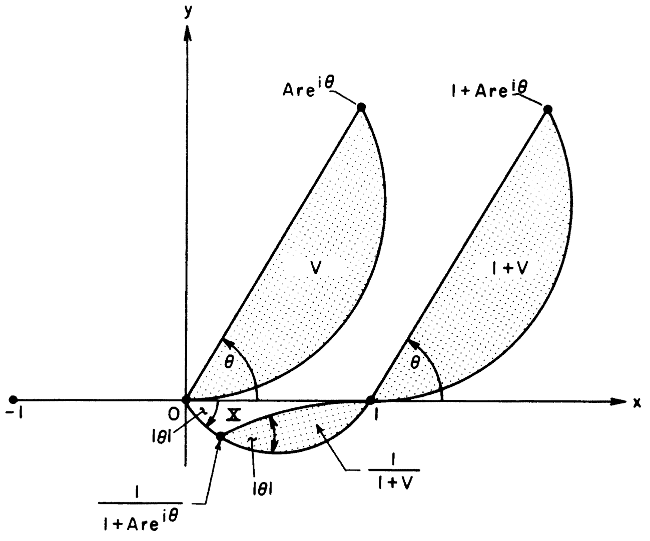

FIGURE 1. *Schematic Diagram of Regions V, $1 + V$, $1/(1 + V)$ and X*

Since $0 \in V$, the second assertion of the lemma follows from the first by Lemma 4.2. This completes the proof.

LEMMA 5.3. *Let f_n, a_k, z, $G_k^{(n)}$ and $g_k^{(n)}$ be defined as in Theorem* 5.1. *Then*

$$W = V/(1 + V) \tag{5.12}$$

is the convex lens-shaped region (Fig. 2) with the same interior angle $|\theta|$ as V, with vertices at 0 *and at*

$$w_0 = Are^{i\theta}/(1 + Are^{i\theta}), \tag{5.13}$$

and with one of its bounding circular arcs tangent to the real axis at 0. *Further,*

$$g_k^{(n)} \in W, \qquad k = 1, \dots, n. \tag{5.14}$$

Proof. Since V is a convex lens-shaped region, so is $1 + V$. Also $1/(1 + V)$ is a lens-shaped region with the same angular opening as V. That it is also convex follows from the fact that $1 + V$ passes through 1 and that its bounding circular arc is tangent to the real axis at 1. Thus,

$$W = V/(1 + V) = 1 - 1/(1 + V)$$

is also a convex lens-shaped region with the same interior angle as V. A simple calculation shows that 0 and w_0 are the vertices of W. Finally, (5.14) follows from (5.11) and (5.12). This completes the proof.

Proof of Theorem 5.1. Assertion (i) follows from the geometry of the region W described in Lemma 5.3 (Fig. 2), the fact that

$$|w_0| = Ar/\sqrt{1 + 2Ar\cos\theta + A^2r^2}, \tag{5.15}$$

and Method B of Theorem 4.3.

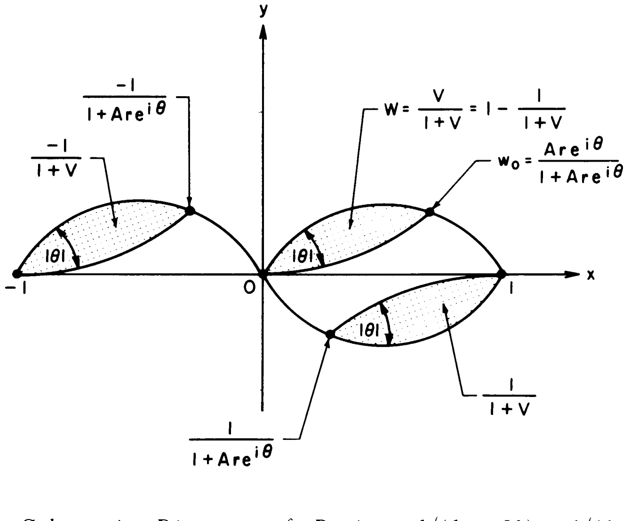
y
x
-1
0
1
$\frac{-1}{1+Are^{i\theta}}$
$\frac{-1}{1+V}$
$W = \frac{V}{1+V} = 1 - \frac{1}{1+V}$
$w_0 = \frac{Are^{i\theta}}{1+Are^{i\theta}}$
$|\theta|$
$|\theta|$
$|\theta|$
$\frac{1}{1+V}$
$\frac{1}{1+Are^{i\theta}}$

FIGURE 2. *Schematic Diagram of Regions* $1/(1 + V)$, $-1/(1 + V)$ *and* $X = V/(1 + V)$

The proof of (5.9b) is also based on (4.15), Method B of Theorem 4.3 and the following argument: When the convex lens-shaped region W has an interior angle greater than $\pi/2$, then it is possible that the distance $|w_0|$ between the vertices of W is smaller than the diameter of W. This will indeed be the case exactly when one of the angles which the straight line, passing through 0 and w_0, makes with one of the bounding arcs of W exceeds $\pi/2$. A simple geometric argument shows that one of the angles discussed above will exceed $\pi/2$ if and only if either $Ar > \sec(\pi - \theta)$ or $Ar < \cos(\pi - \theta)$.

The proofs of (5.9a) and (5.9c) follow from Method A of Theorem 4.3 and the following simple properties of the region V of Lemma 5.2 (see Fig. 1): If $\pi/2 < |\theta| < \pi$, then

$$d(-1, V) = \begin{cases} Ar\csc^2\theta, & \text{provided } \cos(\pi - \theta) \leqq Ar, \\ |1 + Are^{i\theta}|, & \text{provided } Ar < \cos(\pi - \theta). \end{cases}$$

This completes the proof of Theorem 5.1.

As a simple illustration of the use of Theorems 5.1 and 3.1, we consider the representation of the complementary error function [1]:

$$\text{(5.16)} \qquad \operatorname{erfc} w = \frac{2}{\sqrt{\pi}} \int_w^\infty e^{-t^2} dt = \frac{we^{-w^2}}{\sqrt{\pi}} F\left(\frac{1}{w^2}\right), \qquad \operatorname{Re}(w) > 0,$$

where $F(1/w^2)$ has the Stieltjes fraction representation

$$\text{(5.17)} \qquad F(z) = \frac{z}{1} + \frac{(1/2)z}{1} + \frac{(2/2)z}{1} + \frac{(3/2)z}{1} + \frac{(4/2)z}{1} + \cdots,$$

valid for all z such that $|\arg z| < \pi$. Thus, $F(z) = K(a_n z/1)$, where, for $n \geqq 2$,

$$\text{(5.18)} \qquad 0 < a_k < (n - 1)/2 = A, \qquad k = 1, \ldots, n.$$

Hence, for $|\arg z| \leqq \pi/2$ or, equivalently, for $|\arg w| \leqq \pi/4$, (5.8) implies that $|g_k^{(n)}| \leqq 1$, for all $k = 1, \ldots, n$ and $n = 1, 2, \ldots$. It follows from (3.11) that the relative rounding error $|\epsilon_1^{(n)}|$ in calculating the nth approximant of (5.17) is bounded by $3n\omega$, where $\omega = (\frac{1}{2})10^{1-\nu}$, ν equal to the number of significant decimal digits carried in the (machine) computation.

5.2 *Parabolic Convergence Regions*. As a second illustration, we consider the class of continued fractions to which the following general parabola theorem applies.

THEOREM 5.4 [6]. *Let the elements of a continued fraction $K(a_n/1)$ lie within parabolic regions defined by*

$$\text{(5.19)} \quad |a_n| - \operatorname{Re}[a_n \exp(-i(\Psi_n + \Psi_{n-1}))] \leqq 2p_{n-1}(\cos\Psi_n - p_n), \qquad n \geqq 1,$$

where $p_n > 0$, Ψ_n is real and

$$\text{(5.20)} \qquad |P_n - \tfrac{1}{2}| \leqq M < \tfrac{1}{2}, \quad P_n = p_n \exp(i\Psi_n).$$

Then the sequences of even and odd approximants both converge. The continued fraction $K(a_n/1)$ converges if and only if at least one of the series

$$(5.21)\qquad \sum\left|\frac{a_2\cdot a_4\cdots a_{2n}}{a_3\cdot a_5\cdots a_{2n+1}}\right|,\quad \sum\left|\frac{a_3\cdot a_5\cdots a_{2n+1}}{a_4\cdot a_6\cdots a_{2n+2}}\right|$$

diverges. If there exists a constant $K > 0$ *such that* $|a_n| \leqq K$, $n \geqq 1$, *then at least one of the series diverges so that the continued fraction converges.*

Remarks. (1) The region defined by (5.19) is bounded by a parabola with focus at the origin, vertex at the point $p_{n-1}(\cos \Psi_n - p_n)\exp[i(\Psi_n + \Psi_{n-1} + \pi)]$ and axis along the ray $\arg a_n = \Psi_n + \Psi_{n-1}$. Condition (5.20) implies that

$$0 < p_n \leqq \cos \Psi_n \quad\text{and}\quad -\pi/2 < \Psi_n < \pi/2.$$

Thus, it can be seen that the class of continued fractions covered by Theorem 5.4 subsumes all convergent Stieltjes fractions.

(2) It is shown in [6] that the class of continued fractions covered by Theorem 5.4 also subsumes a large subclass of positive definite continued fractions.

It is further shown in [6] that the elements a_n lying in the parabolic regions (5.19) satisfy

$$(5.22)\qquad s_n(V_n) = a_n/(1 + V_n) \subseteq V_{n-1}, \qquad n \geqq 1,$$

where the V_n are half-planes given by

$$(5.23)\qquad V_n = \{\zeta\colon \operatorname{Re}[\zeta \exp(-i\Psi_n)] \geqq -p_n\}, \qquad n \geqq 0.$$

A simple calculation shows that

$$(5.24)\qquad d(-1, V_n) = \cos \Psi_n - p_n \geqq \tfrac{1}{2} - M > 0, \qquad n \geqq 1,$$

where the V_n are given by (5.23) and p_n, Ψ_n are subject to (5.20). Thus, for a continued fraction $K(a_n/1)$ with elements a_n lying in parabolic regions defined by (5.19) and (5.20), we obtain from Method A of Theorem 4.3, (4.15) and (5.24) the bound

$$(5.25)\qquad |g_k^{(n)}| \leqq A^{(n)}/(\tfrac{1}{2} - M)^2, \qquad k = 1, 2, \dots, n, n \geqq 1.$$

For the important special case in which $\Psi_n = \Psi$, $|\Psi| < \pi/2$, $p_n = \frac{1}{2}\cos \Psi$, (5.24) and (5.25) yield the sharper result

$$(5.26)\qquad |g_k^{(n)}| \leqq A^{(n)}/(\tfrac{1}{2}\cos \Psi)^2, \qquad k = 1, 2, \dots, n, n \geqq 1.$$

If, $\Psi = 0$, the region (5.19) is bounded by the parabola with vertex at $-\frac{1}{4}$, focus at 0 and axis on the real axis. Thus, the continued fraction $K(-.25/1)$ considered in the numerical illustration in Section 2 is subsumed under the present class and one obtains from (5.26) the bounds $|g_k^{(n)}| \leqq 1$. Therefore, (3.11) gives a bound of $3n\omega$ for the relative rounding error in evaluating the nth approximant, where $\omega = (\frac{1}{2})10^{1-\nu}$, ν equal to the number of significant decimal digits carried in machine computation.

Application of Method B of Theorem 4.3 gives the bounds

$$(5.27)\qquad |g_k^{(n)}| \leqq M(W_k) = \max\{|w|\colon w \in W_k = V_k/(1 + V_k)\},$$

where V_k is defined by (5.23). To evaluate the right side of (5.27), one can easily verify the following:

$$1 + V_k = \{\zeta\colon \operatorname{Re}[(\zeta - 1)\exp(-i\Psi_k)] \geqq -p_k\},$$
$$\frac{1}{1+V_k} = \left\{\zeta\colon \left|\zeta + \frac{\exp(-i\Psi_k)}{2(\cos\Psi_k - p_k)}\right| \leqq \frac{1}{2(\cos\Psi_k - p_k)}\right\},$$

and hence

$$W_k = \frac{V_k}{1+V_k} = 1 - \frac{1}{1+V_k} = \left\{\zeta\colon \left|\zeta - 1 + \frac{\exp(-i\Psi_k)}{2(\cos\Psi_k - p_k)}\right| \leqq \frac{1}{2(\cos\Psi_k - p_k)}\right\}. \tag{5.28}$$

Thus, from (5.27) and (5.28), it follows that

$$|g_k^{(n)}| \leqq \left|1 - \frac{\exp(-i\Psi_k)}{2(\cos\Psi_k - p_k)}\right| + \frac{1}{2(\cos\Psi_k - p_k)} = \frac{1 - \sqrt{1 - 4p_k\cos\Psi_k + 4p_k^2}}{2(\cos\Psi_k - p_k)}. \tag{5.29}$$

It can be shown from (5.29) that $M(W_k) \leqq 1$ if and only if $\Psi_k = 0$ and $0 \leqq p_k \leqq \frac{1}{2}$. The continued fraction $K(-.25/1)$ is covered by Theorem 5.4 with $\Psi_k = 0$ and $p_k = \frac{1}{2}$. Hence, we obtain the bound $|g_k^{(n)}| \leqq 1$, which is the same that was given by Method A above.

Department of Mathematics
University of Colorado
Boulder, Colorado 80302

1. M. Abramowitz & I. A. Stegun (Editors), *Handbook of Mathematical Functions, With Formulas, Graphs and Mathematical Tables*, Nat. Bur. Standards Appl. Math. Series, 55, Superintendent of Documents, U. S. Government Printing Office, Washington, D.C., 1964. MR **29** #4914.

2. G. Blanch, "Numerical evaluation of continued fractions," *SIAM Rev.*, v. 6, 1964, pp. 383-421. MR **30** #1605.

3. W. Gautschi, "Computational aspects of three-term recurrence relations," *SIAM Rev.*, v. 9, 1967, pp. 24-82. MR **35** #3927.

4. P. R. Graves-Morris (Editor), *Padé Approximants and Their Applications*, Proc. Conference (University of Kent, Canterbury, England, July 1972), Academic Press, New York, 1973.

5. W. B. Gragg, "The Padé table and its relation to certain algorithms of numerical analysis," *SIAM Rev.*, v. 14, 1972, pp. 1-62. MR **46** #4693.

6. William B. Jones & W. J. Thron, "Convergence of continued fractions," *Canad. J. Math.*, v. 20, 1968, pp. 1037-1055. MR **37** #6446.

7. William B. Jones & W. J. Thron (Editors), *Proceedings of the International Conference on Padé Approximants, Continued Fractions and Related Topics*, Special issue of the Rocky Mountain J. Math. (To appear.)

8. N. Macon & M. Baskervill, "On the generation of errors in the digital evaluation of continued fractions," *J. Assoc. Comput. Mach.*, v. 3, 1956, pp. 199–202. MR 18, 337.

9. W. J. Thron, "A survey of recent convergence results for continued fractions," to appear in [7].

10. H. S. Wall, *Analytic Theory of Continued Fractions*, Van Nostrand, Princeton, N.J., 1948. MR 10, 32.