SIAM J. SCI. COMPUT.
Vol. , No. , pp. 1–26,

EFFICIENT AND RELIABLE MULTIPRECISION IMPLEMENTATION OF ELEMENTARY AND SPECIAL FUNCTIONS

ANNIE CUYT, BRIGITTE VERDONK[1] AND HAAKON WAADELAND[2]

Abstract. Special functions are pervasive in all fields of science. The most well-known application areas are in physics, engineering, chemistry, computer science and statistics. Because of their importance several books and a large collection of papers have been devoted to the numerical computation of these functions.

The technique to provide a floating-point implementation of a function differs substantially when going from a fixed finite precision context to a finite multiprecision context. In the former, the aim is to provide an optimal mathematical model, valid on a reduced argument range and requiring as few operations as possible. Here optimal means that, in relation to the model's complexity, the truncation error is as small as it can get. The total relative error should not exceed a prescribed treshold, round-off error and possible argument reduction effect included. In the latter, the goal is to provide a more generic technique, from which an approximant yielding the user-defined accuracy, can be deduced at runtime. Hence best approximants are not an option since these models have to be recomputed every time the precision is altered and a function evaluation is requested. At the same time the generic technique should generate an approximant of as low complexity as possible.

In the current approach we point out how continued fraction representations of functions can be helpful in the multiprecision context. The newly developed generic technique is mainly based on the use of sharpened a priori truncation error estimates for real continued fraction representations of a real variable, developed in Section 3. As illustrated in Section 4, the technique is very efficient and even quite competitive when compared to the traditional fixed precision implementations. The implementation is reliable in the sense that it allows to return a sharp interval enclosure for the requested function evaluation, at the same cost.

The paper follows a recipe style. In Section 2 we gather the ingredients for the new results. In Section 3 we construct or prepare, for a general function $f(x)$, a continued fraction approximant satisfying all requirements of a proper implementation. In Section 4 the procedure is illustrated with results obtained for several specific functions $f(x)$.

Key words. continued fractions, special functions, verified, variable precision

AMS subject classifications. 11A55, 30B70, 41A20, 33B99, 33C99, 65G20

1. Introduction.

Virtually all present-day computer systems, from personal computers to the largest supercomputers, implement the IEEE 64-bit floating-point arithmetic standard, which provides 53 binary or approximately 16 decimal digits accuracy. For most scientific applications, this is more than sufficient. However, for a rapidly expanding body of applications, 64-bit IEEE arithmetic is no longer sufficient. These range from some interesting new mathematical investigations to large-scale physical simulations performed on highly parallel supercomputers. Moreover in these applications, portions of the code typically involve numerically sensitive calculations, which produce results of questionable accuracy using conventional arithmetic. These inaccurate results may in turn induce other errors, such as taking the wrong path in a conditional branch. Such blocks of code benefit enormously from a combination

[1] Dept of Mathematics and Computer Science, University of Antwerp, Middelheimlaan 1, B2020 Antwerp, Belgium, {annie.cuyt,brigitte.verdonk}@ua.ac.be

[2] Department of Mathematical Sciences, Norwegian University of Science and Technology, N7491 Trondheim, Norway, haakonwa@math.ntnu.no

of reliable numeric techniques and the use of high-precision arithmetic. Indeed, the aim of reliable numeric techniques is to deliver, together with the computed result, a guaranteed upper bound on the total error or, equivalently, to compute an enclosure for the exact result.

Up to this date, even environments such as Maple, Mathematica, MATLAB and libraries such as IMSL, CERN and NAG offer no routines for the reliable evaluation of special functions. The following quotes concisely express the need for new developments in the evaluation of special functions:

- *"Algorithms with strict bounds on truncation and rounding errors are not generally available for special functions. These obstacles provide an opportunity for creative mathematicians and computer scientists."* D. Lozier, general director of the Digital Library for Mathematical functions project, and F. Olver [3].
- *"The decisions that go into these algorithm designs — the choice of reduction formulae and interval, the nature and derivation of the approximations — involve skills that few have mastered. The algorithms that MATLAB uses for gamma functions, Bessel functions, error functions, Airy functions, and the like are based on Fortran codes written 20 or 30 years ago."* Cleve Moler, founder of MATLAB [14].

2. Ingredients. A lot of well-known constants in mathematics, physics and engineering, as well as elementary and special functions enjoy very nice and rapidly converging continued fraction representations. Among those we almost always find one or more limit-periodic continued fractions. In this paper we focus on the use of these limit-periodic continued fractions in the reliable multiprecision implementation of the functions they represent. We restrict ourselves to the real-valued case. This implementation is built on top of multiprecision IEEE 754-854 compliant computer arithmetic.

2.1. IEEE-based arithmetic. Let us assume to have at our disposal a multiprecision IEEE 754-854 compliant floating-point implementation of the basic operations, comparisons, base and type conversions. Such an implementation is characterised by four parameters: the internal base β, the precision t and the exponent range $[L, U]$. In the current generic context, we are at least aiming at implementations for $\beta = 2$ at non-standard precisions $t > 64$, and at implementations for use with $\beta = 2^i$ where $i > 1$ or $\beta = 10^j$ where $j \geq 1$. An IEEE compliant implementation offers 4 rounding modes: upward $\triangle$, downward $\triangledown$, truncation $\bowtie$ and round-to-nearest $\bigcirc$. We are especially interested in the first two rounding modes, because a correct implementation in these rounding modes provides a reliable enclosure of the exact function value.

To provide an implementation of a function $f(x)$ in a particular precision t, one first needs to develop a mathematical model or approximation $F(x)$ for $f(x)$ that enjoys a very small relative error, compared to the precision in use. This is usually a very time-consuming effort, because the model changes whenever the precision does. The sum of the truncation error $|f(x) - F(x)|/|f(x)|$ and the rounding error $|F(x) - \mathtt{F}(x)|/|f(x)|$, where $\mathtt{F}(x)$ denotes the machine implementation of the model $F(x)$, should preferably not exceed a few `ulp` where

$$1\ \mathtt{ulp} = \beta^{-t+1} \tag{1}$$

A typical double precision implementation ($\beta = 2, t = 53$) of the elementary functions achieves this in about 25 basic operations. When analyzing the efficiency of our multiprecision implementation, we compare the number of basic operations, required in the new approach when the precision is set to $t = 53$, to this reference.

Since the accumulated rounding error is included in the total error which we bound by only a few `ulp` (typically 1 or 2), the actual evaluation of $f(x)$ needs to take place in a slightly larger working precision $s > t$. The optimal working precision is determined dynamically, depending on the rounding error analysis and the required accuracy. If applicable (not for special functions), our rounding error analysis also takes the effect of argument reduction into account. In any case a fully reliable evaluation of $f(x)$ over the domain is guaranteed.

2.2. Continued fraction representations and tails. Let us consider a continued fraction representation of the form

$$f = \frac{a_1}{1+\dfrac{a_2}{1+\dots}} = \frac{a_1}{\left\lceil 1 \right\rfloor} + \frac{a_2}{\left\lceil 1 \right\rfloor} + \dots = \sum_{n=1}^{\infty} \frac{a_n}{\left\lceil 1 \right\rfloor} \qquad a_n := a_n(x) \qquad f := f(x) \tag{2}$$

Here a_n is called the n-th partial numerator. We use the notation f and $f(x)$ interchangeably. The latter is preferred when the dependence on x needs to be emphasized. We respectively denote by the N-th approximant $f_N(w)$ or $f_N(x; w)$, and N-th tail t_N or $t_N(x)$ of (2), the values

$$f_N(w) = f_N(x; w) = \sum_{n=1}^{N-1} \frac{a_n}{\left\lceil 1 \right\rfloor} + \frac{a_N}{\left\lceil 1+w \right\rfloor} \tag{3}$$

$$t_N = t_N(x) = \sum_{n=N+1}^{\infty} \frac{a_n}{\left\lceil 1 \right\rfloor} \tag{4}$$

We also need approximants of tails and therefore introduce the notation $f_N^{(k)}(w)$ or $f_N^{(k)}(x; w)$ for

$$f_N^{(k)}(w) = f_N^{(k)}(x; w) = \sum_{n=k+1}^{k+N-1} \frac{a_n}{\left\lceil 1 \right\rfloor} + \frac{a_{N+k}}{\left\lceil 1+w \right\rfloor} \qquad f_N^{(0)}(w) = f_N(w)$$

Sometimes the notation $f^{(k)}$ is used for the tail t_k. A continued fraction is said to converge if $\lim_{N\to\infty} f_N(0)$ exists. Note that convergence to ∞ is allowed. In the present paper we assume the continued fractions to converge. Moreover, we restrict ourselves to the case where some $w \neq 0$ can be chosen such that $\lim_{N\to\infty} f_N(w) = \lim_{N\to\infty} f_N(0)$.

The N-th approximant of a continued fraction can also be written as

$$f_N(w) = (s_1 \circ \dots \circ s_N)(w) \qquad s_n(w) = \frac{a_n}{1+w} \qquad n = N, \dots, 1$$

Using the linear fractional transformations s_n, one can define a sequence $\{V_n\}_{n\in\mathbb{N}}$ of value sets for f by:

$$s_n(V_n) = \frac{a_n}{1+V_n} \subseteq V_{n-1} \qquad n \geq 1 \tag{5}$$

The importance of such a sequence of sets lies in the fact that these sets keep track of where certain values lie. For instance, if $w \in V_N$ then $f_N(w) \in V_0$ and $f_{N-k}^{(k)}(w) \in V_k$. Also $t_N \in \overline{V}_N$ and $f \in \overline{V}_0$. An equally important role is played by a sequence of convergence sets $\{E_n\}_{n\in\mathbb{N}}$, of which the elements guarantee convergence of the continued fraction (2) as long as each partial numerator a_n belongs to the respective set E_n:

$$\forall n \geq 1 : a_n \in E_n \Rightarrow \sum_{n=1}^{\infty} \frac{a_n\rfloor}{\lceil 1} \text{ converges} \tag{6}$$

A sequence $\{V_n\}_{n\in\mathbb{N}}$ is called a sequence of value sets for a sequence $\{E_n\}_{n\in\mathbb{N}}$ of convergence sets if (5) holds for all $a_n \in E_n$. Value sets can also be defined for non-convergent continued fractions (then the E_n are called element sets), but in the current context this form of generality does not interest us.

It is well-known that the tail or rest term of a convergent Taylor series expansion converges to zero. It is less well-known that the tail of a convergent continued fraction representation does not need to converge to zero; it does not even need to converge at all. We give an example for each of the cases. Take for instance the continued fraction expansion

$$\frac{\sqrt{1+4x}-1}{2} = \sum_{n=1}^{\infty} \frac{x\rfloor}{\lceil 1} \qquad x \geq -\frac{1}{4}$$

Each tail t_N converges to $1/2(\sqrt{1+4x}-1)$ as well. More remarkable is that the even-numbered tails of the convergent continued fraction

$$\sqrt{2}-1 = \sum_{n=1}^{\infty} \left(\frac{(3+(-1)^n)/2\rfloor}{\lceil 1} \right) = \frac{1\rfloor}{\lceil 1} + \frac{2\rfloor}{\lceil 1} + \frac{1\rfloor}{\lceil 1} + \frac{2\rfloor}{\lceil 1} + \dots$$

converge to $\sqrt{2}-1$ while the odd-numbered tails converge to $\sqrt{2}$ (hence the sequence of tails does not converge), and that the sequence of tails $\{t_N\}_{N\geq 1} = \{N+1\}_{N\geq 1}$ of

$$1 = \sum_{n=1}^{\infty} \frac{n(n+2)\rfloor}{\lceil 1}$$

converges to $+\infty$.

When carefully monitoring the behaviour of these continued fraction tails, very accurate approximants $f_N(w)$ for f can be computed by making an appropriate choice for w. We call a continued fraction of the form (2) limit-periodic with period k, if

$$\lim_{p\to\infty} a_{pk+q} = \tilde{a}_q \qquad q = 1, \dots, k$$

More can be said about tails of limit-periodic continued fractions with period one, also called limit-periodic continued fractions. Let (2) converge and be limit-periodic with $a_n \geq -1/4$ and $\lim_{n\to\infty} a_n = \tilde{a} < \infty$. If $\tilde{w}$ is the in modulus smaller fixpoint of the linear fractional transformation $s(w) = \tilde{a}/(1+w)$, then

$$\tilde{w} = -\frac{1}{2} + \sqrt{\tilde{a} + \frac{1}{4}} = \lim_{N\to\infty} t_N$$

and, according to [16],

$$\lim_{N\to\infty}\left|\frac{f(x)-f_N(x;\tilde{w})}{f(x)-f_N(x;0)}\right|=0 \tag{7}$$

Hence a suitable choice of w in (3) may result in more rapid convergence of the approximants ($w=0$ is usually used as a reference).

In this paper we relax the condition that (2) converges, in the case of limit-periodic continued fractions, to the condition $a_n \geq -1/4$ and $\{a_n\}_{n\in\mathbb{N}}$ bounded [12, pp. 150–159]. This relaxed condition automatically implies that $\tilde{a} \geq -1/4$ and $\tilde{w}$ is real. It also simplifies the description of the results, which can however be generalized to the particular situations where $a_n \leq -1/4$, $\tilde{a}=-1/4$ and (2) converges because the convergence speed of the partial numerators can be termed safe (see [12, p. 159]).

2.3. The oval sequence theorem. Most truncation error upper bounds for $|f(x)-f_N(x;w)|$ are given for the classical choice $w=0$. For continued fractions with partial numerators of the form $a_n(x)=\alpha_n x$ with $\alpha_n>0$ we refer among others to the a priori Gragg-Warner bound

$$|f(x)-f_N(x;0)| \leq 2\frac{|a_1|}{\cos\phi}\prod_{k=2}^{N}\frac{\sqrt{1+4|a_k|/\cos^2(\phi)}-1}{\sqrt{1+4|a_k|/\cos^2(\phi)}+1} \qquad -\pi<2\phi=\arg(x)<\pi \tag{8}$$

which holds for $N\geq 2$ and the a posteriori Henrici-Pfluger bound

$$|f(x)-f_N(x;0)| \leq \begin{cases} |f_N(x;0)-f_{N-1}(x;0)| & |\arg(x)|\leq \pi/2 \\ \dfrac{|f_N(x;0)-f_{N-1}(x;0)|}{|\sin(\arg(x))|} & \pi/2<|\arg(x)|<\pi \end{cases}$$

A posteriori bounds are usually slightly sharper because they exploit the information contained in already computed approximants. But they are of no use in a variable precision context: one does not want to scan all approximants until the truncation error threshold is satisfied. The Gragg-Warner a priori bound does not suffer from this disadvantage, but it is only valid when $a_n(x)=\alpha_n x$ with $\alpha_n>0$. In this paper we develop a practical and sharp truncation error bound for the case $w\neq 0$, which is valid for all continued fractions with real partial numerators $a_n(x)$. Our point of departure for this is the oval sequence theorem from which a priori truncation error estimates can be obtained in case $w\neq 0$.

In the general formulation of the oval sequence theorem (see Theorem 1) [12, pp. 145–147], which holds in the complex plane, the value sets V_n are disks and the convergence sets E_n are cartesian ovals (the situation where the continued fraction does not necessarily converge can also be considered but is of no interest to us). We denote the complex conjugate of the center C_n by $\overline{C}_n$.

In the real case both V_n and E_n reduce to intervals. We reformulate the more specific real version in Theorem 2 (which allows for point intervals) and provide an elegant short proof.

THEOREM 1. *Let* $0<r_n<|1+C_n|$ *for* $n\geq 0$ *and* $|C_{n-1}|r_n<|1+C_n|r_{n-1}$ *for* $n\geq 1$. *Then* $\{V_n\}_{n\in\mathbb{N}}$ *with*

$$V_n=\{w\in\mathbb{R}:|C_n-w|<r_n\}$$

is a sequence of value sets for the sequence $\{E_n\}_{n\in\mathbb{N}_0}$ *of convergence sets given by*

$$E_n = \{a \in \mathbb{R} : |a(1+\overline{C}_n) - C_{n-1}(|1+C_n|^2 - r_n^2)| + r_n|a| \le r_{n-1}(|1+C_n|^2 - r_n^2)\}$$

For $w \in V_N$ *the truncation error* $|f(x) - f_N(x;w)|$ *is bounded by*

$$|f(x) - f_N(x;w)| \le 2r_N \frac{|C_0| + r_0}{|1+C_N| - r_N} \times \prod_{k=1}^{N-1} M_k \tag{9}$$

where $M_k = \max\{|u/(1+u)| : u \in \overline{V}_k\}$.

In the real case the set E_n given above is actually an interval. The set $\overline{V}_k = [C_k - r_k, C_k + r_k]$ and the maximum M_k is given by

$$M_k = \max\left(\left|\frac{C_k - r_k}{1 + C_k - r_k}\right|, \left|\frac{C_k + r_k}{1 + C_k + r_k}\right|\right)$$

since the function $w/(1+w)$ is a strictly increasing function. Let us simplify the statement and proof of the oval sequence theorem for the case of real interval sequences. At the same time, the new proof delivers a bound for the relative error instead of the absolute error, and we have $M_k \le 1$. In the sequel we refer to this theorem as the interval sequence theorem.

THEOREM 2. *Let for all* n *the values* L_n *and* R_n *satisfy* $-1/2 \le L_n \le R_n < \infty$ *and let*

$$b_n := (1 + \operatorname{sign}(L_n)\max(|L_n|, |R_n|))\, L_{n-1}$$
$$c_n := (1 + \operatorname{sign}(L_n)\min(|L_n|, |R_n|))\, R_{n-1}$$

satisfy $b_n \le c_n$ *and* $0 \le b_n c_n$. *Then the sequence* $\{V_n\}_{n\in\mathbb{N}}$ *with* $V_n = [L_n, R_n]$ *is a sequence of value sets for the sequence* $\{E_n\}_{n\in\mathbb{N}}$ *of convergence sets given by*

$$E_n = [b_n, c_n] = \begin{cases} [(1+R_n)L_{n-1}, (1+L_n)R_{n-1}] & b_n \ge 0 \\ [(1+L_n)L_{n-1}, (1+R_n)R_{n-1}] & b_n \le 0 \end{cases}$$

For $w \in V_N$ *the relative truncation error* $|f(x) - f_N(x;w)|/|f(x)|$ *is bounded by*

$$\left|\frac{f(x) - f_N(x;w)}{f(x)}\right| \le \frac{R_N - L_N}{1 + L_N} \prod_{k=1}^{N-1} M_k \tag{10}$$

where $M_k = \max\{|u/(1+u)| : u \in V_k\} = \max\{|L_k/(1+L_k)|, |R_k/(1+R_k)|\}$.

Proof. The relation between E_n and V_n is expressed in (5). For $E_n = [b_n, c_n] \subset \mathbb{R}^+$ this translates to

$$L_{n-1} \le \frac{b_n}{1+R_n}$$
$$R_{n-1} \ge \frac{c_n}{1+L_n}$$

while for $E_n \subset \mathbb{R}^-$ it translates to

$$L_{n-1} \le \frac{b_n}{1+L_n}$$
$$R_{n-1} \ge \frac{c_n}{1+R_n}$$

Using the notation $f^{(j)}(x)$ for the j-th tail of the continued fraction $f(x)$, we obtain

$$\begin{aligned} f(x) - f_N(x;w) &= \frac{a_1}{1+f^{(1)}(x)} - \frac{a_1}{1+f_{N-1}^{(1)}(x;w)} \\ &= \frac{-f^{(0)}(x)\left(f^{(1)}(x) - f_{N-1}^{(1)}(x;w)\right)}{1+f_{N-1}^{(1)}(x;w)} \\ &= -f(x)\frac{f^{(N)}(x)-w}{1+w}\prod_{j=1}^{N-1}\left(\frac{-f^{(j)}(x)}{1+f_{N-j}^{(j)}(x;w)}\right) \end{aligned}$$

Hence

$$\left|\frac{f(x)-f_N(x;w)}{f(x)}\right| \le \frac{R_N - L_N}{1+w}\left|\prod_{j=1}^{N-1}\frac{f^{(j)}(x)}{1+f_{N-j}^{(j)}(x;w)}\right|$$

Now for $1 \le k \le \lfloor (N-1)/2 \rfloor$, we have

$$\begin{aligned} &\frac{f^{(2k)}(x)}{1+f_{N-2k}^{(2k)}(x;w)}\,\frac{f^{(2k-1)}(x)}{1+f_{N-2k+1}^{(2k-1)}(x;w)} \\ &\quad= \frac{f^{(2k)}(x)}{1+f_{N-2k}^{(2k)}(x;w)}\,\frac{a_{2k}}{1+f^{(2k)}(x)}\,\frac{1}{1+f_{N-2k+1}^{(2k-1)}(x;w)} \\ &\quad= \frac{f^{(2k)}(x)}{1+f^{(2k)}(x)}\,\frac{f_{N-2k+1}^{(2k-1)}(x;w)}{1+f_{N-2k+1}^{(2k-1)}(x;w)} \end{aligned}$$

If $N-1$ is odd then the last factor can be combined with $1/(1+w)$ into

$$\frac{f^{(N-1)}(x)}{1+f_1^{(N-1)}(x;w)}\,\frac{1}{1+w} = \frac{f_1^{(N-1)}(x;w)}{1+f_1^{(N-1)}(x;w)}\,\frac{1}{1+f^{(N)}(x)}$$

Since $\max_{u\in V_k}|u/(1+u)| = M_k$ the theorem is proved. □

An upper bound for the truncation error $|f(x) - f_N(x;w)|$ is obtained by multiplying the right hand side of (10) by R_0 which is an upper bound for $|f(x)|$.

The smaller the sets V_n, in other words the smaller $R_n - L_n$, and the sharper one knows the M_k, the smaller the upper bound on the truncation error $|f(x) - f_N(x;w)|/|f(x)|$ becomes. We come back to this issue in Section 3.

In Section 3 we also combine the interval sequence theorem with the characteristic monotonicity behaviour of the tails of some limit-periodic continued fraction representations, which we derive now.

2.4. Monotonicity properties of tails. To prepare these results, we prove some easy lemmas. Lemma 1 mainly serves to establish computable upper bounds for tails of continued fractions, while Lemma 2 provides additional information about the relation between continued fraction approximants and tail estimates. In the Lemmas 3 and 4 we obtain some results on the boundedness and monotonicity behaviour of the sequence of tails. In the formulation of these lemmas the notions increasing and decreasing for sequences of numbers are never meant in the strict sense: a constant sequence can be considered to be decreasing as well as increasing.

To obtain monotonicity results about the continued fraction tails, we further distinguish between

- the fractions where $a_n \to \tilde{a}$ from one side, say $\{a_n\}_{n\in\mathbb{N}}$ is a decreasing (or increasing) sequence with $\lim_{n\to\infty} a_n = \tilde{a}$,
- and the fractions where $a_n \to \tilde{a}$ in an alternating fashion, say the sequences $\{a_{2n+1}\}_{n\in\mathbb{N}}$ and $\{a_{2n}\}_{n\in\mathbb{N}}$ respectively decrease and increase towards $\tilde{a}$.

For continued fractions with positive a_n it suffices to study the latter case in detail, while the former case is the basic building block when dealing with continued fractions with negative a_n. This becomes clear in Section 4 where all different types of behaviour for the partial numerators are illustrated and can be dealt with, starting from the two mentioned basic types. How to deal with a continued fraction containing both positive and negative partial numerators is explained in Section 3.4. From Section 3.4 it should be clear that the condition in most lemmas for the partial numerators to be either all positive or all negative, is not a true restriction but reflects merely a choice made by the authors to simplify the description of the results.

LEMMA 1.

1) *Let all $a_n > 0$ and let (2) converge. Then for $k > 0$ the sequences of even and odd approximants $\{f_{2n}^{(k)}(0)\}_{n\in\mathbb{N}}$ and $\{f_{2n+1}^{(k)}(0)\}_{n\in\mathbb{N}}$ are increasing and decreasing sequences respectively, satisfying*

$$f_{2n}^{(k)}(0) \le t_k \le f_{2n+1}^{(k)}(0)$$

2) *Let all $a_n < 0$ and let the sequence $\{a_n\}_{n\in\mathbb{N}}$ be decreasing with $\lim_{n\to\infty} a_n = \tilde{a} \ge -1/4$. Then for $k > 0$ the tail t_k of (2) is bounded by*

$$\frac{-1+\sqrt{4\tilde{a}+1}}{2} \le t_k \le \frac{-1+\sqrt{4a_{k+1}+1}}{2}$$

Proof. Part 1 is well-known from continued fraction literature [1, p. 223]. Remains to prove part 2. We give the proof for the upper bound in part 2 because the lower bound is proved analogously. Consider the approximant $f_n^{(k)}$ of t_k. Since for $j = 0, \ldots, n-1 : 0 \le 1 + a_{k+n-j} \le 1 + a_{k+1}$, we find that

$$\begin{aligned}
\frac{a_{k+n-1}}{1+a_{k+n}} &\le \frac{a_{k+1}}{1+a_{k+1}} \\
\frac{a_{k+n-2}}{1+\frac{a_{k+n-1}}{1+a_{k+n}}} &\le \frac{a_{k+1}}{1+\frac{a_{k+1}}{1+a_{k+1}}} \\
&\vdots \\
f_n^{(k)} &\le \cfrac{a_{k+1}}{\cfrac{}{1}} + \ldots + \cfrac{a_{k+1}}{\cfrac{}{1}}
\end{aligned}$$

In the limit the inequality is preserved because of the convergence of (2), leading to

$$t_k \le \frac{-1+\sqrt{4a_{k+1}+1}}{2}$$

which concludes this part of the proof. □

The condition that (2) converges is automatically satisfied in part 2 of Lemma 1 and in the Lemmas 2, 3 and 4, because of the boundedness of the partial numerators a_n.

LEMMA 2. *Let all $a_n \geq -1/4$ and let $\{a_n\}_{n\in\mathbb{N}}$ be bounded. Then:*

1) *For each $k : t_k \geq -1/2$ and $\operatorname{sign}(t_k) = \operatorname{sign}(a_{k+1})$.*
2) *If all $a_n > 0$ and $0 \leq w_{1,k} \leq t_k \leq w_{2,k}$, then*

$$f_{2n}(w_{1,2n}) \leq f \leq f_{2n+1}(w_{1,2n+1}) \tag{11a}$$

$$f_{2n}(w_{2,2n}) \geq f \geq f_{2n+1}(w_{2,2n+1}) \tag{11b}$$

3) *If all $a_n < 0$ and $-1/2 \leq w_{1,k} \leq t_k \leq w_{2,k}$, then*

$$f_n(w_{1,n}) \leq f \leq f_n(w_{2,n}) \tag{11c}$$

Proof. The proof is straightforward. We prove for instance the second inequality of part 3. From $0 \geq w_{2,n} \geq t_n \geq -1/2$ follows

$$\frac{a_n}{1+w_{2,n}} \geq \frac{a_n}{1+t_n}$$

$$\vdots$$

$$f_n(w_{2,n}) \geq f_n(t_n) = f$$

and similarly for the other equations. □

LEMMA 3.

1) *Let (2) converge and let all $a_n > 0$. If the sequences $\{a_{2n+1}\}_{n\in\mathbb{N}}$ and $\{a_{2n}\}_{n\in\mathbb{N}}$ are respectively decreasing and increasing, then the tail sequences $\{t_{2n}\}_{n\in\mathbb{N}}$ and $\{t_{2n+1}\}_{n\in\mathbb{N}}$ are respectively decreasing and increasing.*
2) *Let in addition for all integers k and $\ell : a_{2k+1} > a_{2\ell}$, then for all k and ℓ the tails of (2) satisfy $t_{2k} \geq t_{2\ell+1}$.*

Proof. Since $a_{2n+3} \leq a_{2n+1}$ and $a_{2n+4} \geq a_{2n+2}$, we find that

$$f_2^{(2n+2)}(0) = \frac{a_{2n+3}}{1+a_{2n+4}} \leq f_2^{(2n)}(0) = \frac{a_{2n+1}}{1+a_{2n+2}}$$

$$\vdots$$

$$f_k^{(2n+2)}(0) \leq f_k^{(2n)}(0) \qquad k > 2$$

In the limit the last inequality remains true, yielding $t_{2n+2} \leq t_{2n}$. The case $t_{2n+3} \geq t_{2n+1}$ follows in the same way and we proceed to part 2. Let us first prove the statement for $k = \ell$. Since $a_{2k+1} > a_{2k+2}, a_{2k+2} < a_{2k+3}$ and so on, we have

$$f_2^{(2k)}(0) > f_2^{(2k+1)}(0)$$

$$\vdots$$

$$f_s^{(2k)}(0) > f_s^{(2k+1)}(0) \qquad s > 2$$

Taking limits, the last inequality becomes $t_{2k} \geq t_{2k+1}$. For $k < \ell$, we use

$$t_{2\ell} \geq t_{2\ell+1} \geq t_{2k+1}$$

and for $k > \ell$, we use

$$t_{2\ell} \geq t_{2k} \geq t_{2k+1}$$

which concludes the proof. □

LEMMA 4. *Let all a_n satisfy $-1/4 \leq a_n < 0$ and let the sequence $\{a_n\}_{n\in\mathbb{N}}$ be decreasing. Then the tail sequence $\{t_n\}_{n\in\mathbb{N}}$ is decreasing.*

Proof. Consider for $k \geq 1$ and $n \geq 1$ the approximant $f_n^{(k)}$ of the tail t_k. Since $0 \leq 1 + a_{k+n+1} \leq 1 + a_{k+n}$ we find with $a_{k+n} < 0$ and $a_{k+n-1} < 0$ that

$$\frac{a_{k+n}}{1 + a_{k+n+1}} \leq \frac{a_{k+n-1}}{1 + a_{k+n}}$$

and analogously

$$f_n^{(k+1)} \leq f_n^{(k)}$$

In the limit this inequality is preserved and becomes

$$t_{k+1} \leq t_k$$

which concludes the proof. □

2.5. Backward recurrence algorithm. Several algorithms exist for the computation of $f_N(w)$. The easiest to use, because N can be determined concurrently with $f_N(w)$, is the forward recurrence algorithm:

$$\begin{aligned} A_{-1} &= 1 \qquad A_0 = 0 \\ B_{-1} &= 0 \qquad B_0 = 1 \\ A_n &= A_{n-1} + a_n A_{n-2} \qquad n = 1, \ldots, N-1 \\ B_n &= B_{n-1} + a_n B_{n-2} \qquad n = 1, \ldots, N-1 \\ f_N(w) &= \frac{(1+w)A_{N-1} + a_N A_{N-2}}{(1+w)B_{N-1} + a_N B_{N-2}} \end{aligned} \tag{12}$$

The most stable however [6] is the backward recurrence algorithm:

$$\begin{aligned} F_{N+1}^{(N)} &= w \\ F_n^{(N)} &= \frac{a_n}{1 + F_{n+1}^{(N)}} \qquad n = N, N-1, \ldots, 1 \\ f_N(w) &= F_1^{(N)} \end{aligned}$$

For the backward recurrence algorithm to be useful in a variable precision context, it must be possible to determine rather easily a priori which approximant to compute, in other words to decide which N guarantees a prescribed total error upper bound. For the determination of N in function of the truncation error upper bound, we refer to Section 3.1 and 3.2. The results follow from combining the interval sequence theorem with the Lemmas 1 to 4. Choosing a suitable estimate for the tail t_N is discussed in Section 3.3. The rounding error is subsequently controlled by determining a suitable working precision in which to execute the backward recurrence algorithm. The latter is dealt with in detail in Section 3.5.

3. Preparation. The realization of a machine implementation of $f(x)$ is a three-step process:

1) When given an argument x for a function f, the evaluation $f(x)$ is often reduced to the evaluation of f for another argument $\tilde{x}$ lying within specified bounds and for which there exists an easy relationship between $f(x)$ and $f(\tilde{x})$. For instance, for the exponential function in a base β implementation,

$$\exp(x) = \beta^k \exp(\tilde{x}) \qquad \tilde{x} = \mathrm{mod}(x, \ln \beta) \qquad |\tilde{x}| \leq \frac{\ln \beta}{2}$$

   Although the given argument x is known exactly, because it is a given floating-point number, usually the reduced argument $\tilde{x}$ cannot be computed exactly, but is subject to a rounding error.
2) After the reduced argument is determined, the mathematical model F for f is constructed and a truncation error comes into play.
3) When implemented, in other words evaluated, this mathematical model F is also subject to a rounding error.

Finally the effect of switching from the argument x to the reduced argument $\tilde{x}$ must be taken into account. This introduces a final additional error. Let us now have a look at how all these errors can be controlled.

The issue of argument reduction is a topic in itself and is not the subject of this paper. We mention it here because we want to emphasize that in our implementation this effect is taken into account, by which we mean that the error upper bound we guarantee holds for the entire argument range and not only for the reduced range. Detailed information on argument reduction can be found in [15, 4] and [13]. For each of the functions discussed in detail in Section 4, we briefly outline the argument reduction technique used.

For the use of continued fraction approximants in step 2 and step 3 we still need to address the following issues:

- how to determine N such that the upper bound for the relative error on $f_N(x; w)$ does not exceed a threshold ϵ_T?
- how to obtain an easily computable tail estimate $w \in V_N$ (or $w_N \in V_N$) for the evaluation of $f_N(x; w)$?

A rounding error analysis of the backward recurrence algorithm is given in Section 3.5.

3.1. Determination of N. The a priori Gragg-Warner bound is expressed entirely in terms of the partial numerators a_n and the function's argument x. Except for the result given in [12, p. 151], this cannot be said of the oval sequence or interval sequence bounds. Theorem 1 as well as Theorem 2 assume that the value sets are given and that convergence sets can be associated with them.

The aim of this section is slightly different from what we have obtained in Theorem 2, where the sets E_n are deduced from the intervals $V_n = [L_n, R_n]$ and the bounds of E_n are formulated in terms of L_n and R_n. Here we want to formulate L_n and R_n in terms of the bounds on a_n in E_n and hence associate intervals V_n with given intervals E_n, instead of the other way around. In Lemma 5 we further specify the bounds L_n and R_n introduced in Theorem 2, for some appropriately chosen convergence sets E_n.

We show that the bounds L_n and R_n are in fact tails.

Let $E_n = [b_n, c_n]$ with $-1/4 \leq b_n \leq c_n$ and $b_n c_n \geq 0$. The smallest useful E_n equals $[\bigtriangledown(a_n), \bigtriangleup(a_n)]$ where $\bigtriangledown(a_n)$ and $\bigtriangleup(a_n)$ denote the machine representations of the partial numerator a_n obtained using the downward and upward rounding modes respectively. The usual machine representation of a_n, namely $\bigcirc(a_n)$ which results from round-to-nearest, is contained in such E_n. In case a_n is itself the result of a computation, then the interval E_n can be taken slightly larger. The condition that b_n and c_n have the same sign means nothing more than that at least $\text{sign}(a_n)$ is kept fixed in E_n.

LEMMA 5. *If the sequence of convergence sets $\{E_n\}_{n\in\mathbb{N}}$ is given by $E_n = [b_n, c_n]$ with $b_n \geq -1/4$ and $0 \leq b_n c_n$, then the corresponding sequence of value sets $\{V_n\}_{n\in\mathbb{N}}$ is given by $V_n = [L_n, R_n]$ where L_n and R_n are particular tails of the continued fractions*

$$\begin{aligned} \hat{D} &= \frac{b_1\rfloor}{\lceil 1} + \frac{c_2\rfloor}{\lceil 1} + \frac{b_3\rfloor}{\lceil 1} + \frac{c_4\rfloor}{\lceil 1} + \dots \\ \hat{U} &= \frac{c_1\rfloor}{\lceil 1} + \frac{b_2\rfloor}{\lceil 1} + \frac{c_3\rfloor}{\lceil 1} + \frac{b_4\rfloor}{\lceil 1} + \dots \end{aligned} \tag{13}$$

$$\begin{aligned} \check{D} &= \frac{b_1\rfloor}{\lceil 1} + \frac{b_2\rfloor}{\lceil 1} + \frac{b_3\rfloor}{\lceil 1} + \frac{b_4\rfloor}{\lceil 1} + \dots \\ \check{U} &= \frac{c_1\rfloor}{\lceil 1} + \frac{c_2\rfloor}{\lceil 1} + \frac{c_3\rfloor}{\lceil 1} + \frac{c_4\rfloor}{\lceil 1} + \dots \end{aligned} \tag{14}$$

More precisely, denoting the tails of $\hat{D}, \check{D}$ and $\hat{U}, \check{U}$ respectively by $\hat{D}^{(n)}, \check{D}^{(n)}$ and $\hat{U}^{(n)}, \check{U}^{(n)}$ we have when all $b_n \geq 0$:

$$\begin{aligned} L_{2j} &= \hat{D}^{(2j)} \qquad & L_{2j-1} &= \hat{U}^{(2j-1)} \\ R_{2j} &= \hat{U}^{(2j)} \qquad & R_{2j-1} &= \hat{D}^{(2j-1)} \end{aligned} \tag{15}$$

and when all $b_n \leq 0$:

$$L_n = \check{D}^{(n)} \qquad R_n = \check{U}^{(n)} \tag{16}$$

Proof. A sequence of value sets $\{V_n\}_{n\in\mathbb{N}}$ associated with the sequence of convergence sets $\{E_n\}_{n\in\mathbb{N}}$ satisfies

$$\frac{a_n}{1+V_n} \subseteq V_{n-1} \qquad a_n \in E_n$$

Hence, if $V_n \subset \mathbb{R}^+$, expressions for the left and right endpoints L_n and R_n of the intervals V_n need to satisfy the condition

$$(1+L_n)R_{n-1} \geq (1+R_n)L_{n-1} \tag{17}$$

and if $V_n \subset \mathbb{R}^-$, they need to satisfy

$$R_{n-1}(1+R_n) \geq L_{n-1}(1+L_n) \tag{18}$$

Let us first write down explicit formulas for suitable bounds L_n and R_n in case we are dealing with a sequence of convergence sets that only allow positive a_n.

Each V_n contains the tail t_n of (2) since $V_n = \overline{V}_n$. The smallest possible value for the tail t_n of (2) is obtained when the partial numerators $a_{n+j} \in E_{n+j}$ with $j \geq 1$ alternatively take the smallest and largest values $b_{n+1} \in E_{n+1}, c_{n+2} \in E_{n+2}, \dots$ A similar result holds for the largest value of the tail which is attained when the

a_{n+j} alternatively take the largest and smallest values $c_{n+1}, b_{n+2}, \ldots$ Hence, for the convergence sets $E_n = [b_n, c_n]$, the corresponding value sets V_n are given by $V_n = [L_n, R_n]$ with

$$L_n = \cfrac{b_{n+1}}{\left\lceil 1 \right\rfloor} + \cfrac{c_{n+2}}{\left\lceil 1 \right\rfloor} + \cfrac{b_{n+3}}{\left\lceil 1 \right\rfloor} + \cfrac{c_{n+4}}{\left\lceil 1 \right\rfloor} + \ldots$$
$$R_n = \cfrac{c_{n+1}}{\left\lceil 1 \right\rfloor} + \cfrac{b_{n+2}}{\left\lceil 1 \right\rfloor} + \cfrac{c_{n+3}}{\left\lceil 1 \right\rfloor} + \cfrac{b_{n+4}}{\left\lceil 1 \right\rfloor} + \ldots$$

It is now easy to check that (17) holds.

When the convergence sets only allow negative a_n, then the proof goes as follows. Consider $f_n^{(k)}$. Since $0 \leq 1 + b_{k+n} \leq 1 + a_{k+n} \leq 1 + c_{k+n}$ we find that

$$\frac{b_{k+n-1}}{1 + b_{k+n}} \leq \frac{a_{k+n-1}}{1 + a_{k+n}} \leq \frac{c_{k+n-1}}{1 + c_{k+n}}$$

and analogously

$$\check{D}_n^{(k)} \leq f_n^{(k)} \leq U_n^{(k)}$$

In the limit this inequality is preserved and becomes

$$\check{D}^{(k)} \leq t_k \leq \check{U}^{(k)}$$

Again, (18) is easy to check. □

Given the accuracy with which the partial numerators can be obtained, in other words the width of the convergence sets E_n, we can now determine the bounds L_n and R_n in function of the bounds on E_n and use these in (10). This way, the right hand side of (10) can be evaluated for consecutive values of N until the relative error is sufficiently small. Note however that L_n and R_n are infinite expressions and hence not computable. The practical computation of $M_k = \max\{|L_k/(1+L_k)|, R_k/(1+R_k)\}$ in the right hand side of (10) is further detailed in Section 3.3.

We also prefer a bound that requires the computation of as few expressions as possible, for instance either the sequence $\{R_n\}_{n\in\mathbb{N}}$ or the sequence $\{L_n\}_{n\in\mathbb{N}}$ but not both. This can easily be guaranteed when the value sets V_n are small enough to ensure that $L_n R_n \geq 0$. Then in case $L_n \leq 0$, the bound $M_n = |L_n/(1 + L_n)|$, and in case $L_n \geq 0$, we have $M_n = R_n/(1 + R_n)$ since $L_n \leq R_n$ in all cases. According to Lemma 2, $L_n R_n \geq 0$ is guaranteed by $b_n c_n \geq 0$ because the sign of R_n and L_n is determined by the sign of either b_n or c_n.

In addition, when $0 \leq b = \lim_{n\to\infty} b_n$ and $0 \leq c = \lim_{n\to\infty} c_n$, the tails L_n and R_n respectively converge to

$$\begin{aligned} \lim_{n\to\infty} L_n = L &= \frac{b - c - 1 + \sqrt{(c-b)^2 + 2(c+b) + 1}}{2} \\ \lim_{n\to\infty} R_n = R &= \frac{c - b - 1 + \sqrt{(c-b)^2 + 2(c+b) + 1}}{2} \end{aligned} \tag{19}$$

with $L \leq R$. In case $b \leq 0$ and $c \leq 0$, we have

$$\begin{aligned} \lim_{n\to\infty} L_n = L &= \frac{-1 + \sqrt{4b + 1}}{2} \\ \lim_{n\to\infty} R_n = R &= \frac{-1 + \sqrt{4c + 1}}{2} \end{aligned} \tag{20}$$

In Section 4 we illustrate that the bound given by (10) combined with the formulas (15) and (16) is as sharp with respect to the true truncation error for the modified approximant $f_N(x;w)$, as the Gragg-Warner bound is when compared to the unmodified approximant $f_N(x;0)$. Both are also of the same a priori form.

3.2. Explicit formula for N. In this section we want to use formula (10) to deduce an explicit formula for N from the truncation error bound, rather than the bound from the knowledge of N. To this end we need to be able to bound the R_n from above by a value $\overline{R}$, from a certain $n_R \leq N$ on, and bound the L_n from below by $\overline{L}$ from a possibly different $n_L \leq N$ on. The continued fractions and convergence sets that satisfy such conditions are discussed in detail in Section 4. Sometimes we want to make use of Lemma 6.

LEMMA 6. *Let $\{x_n\}_{n\in\mathbb{N}}$ be an increasing sequence and $\{y_n\}_{n\in\mathbb{N}}$ be decreasing, with for all n: $x_n \geq -d$ and $y_n \geq -d$ where $d \in \mathbb{R}$, then the sequence*

$$\left\{\frac{y_n - x_n}{d + x_n}\right\}_{n\in\mathbb{N}}$$

is also decreasing.

Proof. For fixed ℓ we define the sequence $\{z_n^{(\ell)}\}_{n\in\mathbb{N}}$ with

$$z_n^{(\ell)} = \frac{y_\ell - x_n}{d + x_n}$$

Because for every ℓ it holds that $y_\ell \geq -d$, the function $(y_\ell - x)/(d+x)$ is an increasing function on $]-d, +\infty[$. Hence the sequence $\{z_n^{(\ell)}\}_{n\in\mathbb{N}}$ is a decreasing sequence and consequently

$$\frac{y_n - x_n}{d + x_n} \geq \frac{y_{n+1} - x_n}{d + x_n} = z_n^{(n+1)} \geq z_{n+1}^{(n+1)} = \frac{y_{n+1} - x_{n+1}}{d + x_{n+1}} \qquad \square$$

When the R_n and L_n satisfy the required monotonicity properties, we can bound the factor $(R_N - L_N)/(1 + L_N)$ for $N \geq \max(n_R, n_L)$ by $(\overline{R} - \overline{L})/(1 + \overline{L})$. In case the sequence $\{R_n\}_{n\in\mathbb{N}}$ is decreasing and the sequence $\{L_n\}_{n\in\mathbb{N}}$ is increasing, we can take $\overline{R} = R_K$ and $\overline{L} = L_K$ for $K \leq N$. In addition the product of the factors M_k can then be bounded above by

$$\prod_{k=1}^{N-1} M_k \leq \left(\prod_{k=1}^{K-1} M_k\right) \times M_K^{N-K} \qquad K \leq N \tag{21}$$

Although cruder, (21) allows to determine N explicitly in function of a given threshold ϵ_T. Taking the logarithm of the new bound (21) for the product $M_1 \ldots M_{N-1}$ in (10) and an imposed threshold ϵ_T results in

$$(N - K)(-\log M_K) \geq \log\left(\frac{R_K - L_K}{1 + L_K}\right) + \sum_{k=1}^{K-1} \log M_k - \log \epsilon_T \tag{22}$$

How $w \in [L_N, R_N]$ is obtained cheaply (in terms of numerical operations), is discussed together with some practical ways to obtain a computational upper bound for M_k.

3.3. Obtaining M_k and w. Lemmas 1 and 2 show the way to compute an upper bound for M_k. We discuss two cases in somewhat more detail.

When all partial numerators $a_n > 0$, then $L_k \geq 0$ and $M_k = R_k/(1+R_k)$. Let us denote the j-th approximants of R_k and L_k by $R_{k,j}$ and $L_{k,j}$ respectively and use part 1 of Lemma 1 (the correct notation for these approximants is actually $R_{k,j}(0)$ and $L_{k,j}(0)$ but we drop the suffix (0) to further simplify the notation). Because the tails R_k and L_k themselves cannot be computed, odd-numbered approximants $R_{k,2j+1} \geq R_k$ and an even-numbered approximant $L_{N,2j} \leq L_N$ are computed. In practice, for instance, formula (10) becomes

$$\frac{|f(x)-f_N(x;w)|}{|f(x)|} \leq \frac{R_{N,2j+1}-L_{N,2j}}{1+L_{N,2j}} \prod_{k=1}^{N-1} \frac{R_{k,2j+1}}{1+R_{k,2j+1}} \qquad w \in [L_N, R_N] \tag{23}$$

Here $R_{k,2j+1}$ equals in turn $\hat{U}_{2j+1}^{(2\ell)}$ or $\hat{D}_{2j+1}^{(2\ell+1)}$ depending on whether k is even ($k=2\ell$) or odd ($k=2\ell+1$). When replacing the odd-numbered approximants $R_{k,2j+1}$ in (23) by approximate values $\breve{R}_{k,2j+1}$ computed in rounded arithmetic, one should be careful not to violate the condition $\breve{R}_{k,2j+1} \geq R_k$.

When all $a_n < 0$ then $M_k = |L_k/(1+L_k)|$. If in addition, the sequences $\{a_n\}_{n\in\mathbb{N}}, \{b_n\}_{n\in\mathbb{N}}$ and $\{c_n\}_{n\in\mathbb{N}}$ from Lemma 5 decrease, then the lower bound for L_k given in part 2 of Lemma 1 can be used to obtain an upper bound for M_k, while the factor $(R_N-L_N)/(1+L_N)$ can be bounded by (here we keep the suffix (w) because it varies between R_N and L_N)

$$\frac{R_N-L_N}{1+L_N} \leq \frac{R_{N,j}\left(\frac{-1+\sqrt{4a_{N+1}+1}}{2}\right) - L_{N,j}(-1/2)}{1+L_{N,j}(-1/2)}$$

This results for (10) in

$$\frac{|f(x)-f_N(x;w)|}{|f(x)|} \leq \frac{R_{N,j}\left(\frac{-1+\sqrt{4a_{N+1}+1}}{2}\right) - L_{N,j}(-1/2)}{1+L_{N,j}(-1/2)} \prod_{k=1}^{N-1} \left|\frac{L_{k,j}(-1/2)}{1+L_{k,j}(-1/2)}\right|$$
$$w \in [L_N, R_N] \tag{24}$$

Similar bounds can be given when the sequences $\{a_n\}_{n\in\mathbb{N}}, \{b_n\}_{n\in\mathbb{N}}$ and $\{c_n\}_{n\in\mathbb{N}}$ are increasing.

The simplest choice for w, in case of a limit-periodic continued fraction with $\lim_{n\to\infty} a_n = \tilde{a} < \infty$, is given by

$$w = \frac{-1+\sqrt{4\tilde{a}+1}}{2} \tag{25}$$

Question is of course whether this w belongs to $V_N = [L_N, R_N]$. For the continued fractions discussed in Section 4, we indicate how this can be assured. It comes down to choosing the convergence sets E_n appropriately, by which we mean that the sequence of value sets $\{V_n\}_{n\in\mathbb{N}}$ has a nonempty intersection. When choosing the E_n too small, such that the sequence of value sets $\{V_n\}_{n\in\mathbb{N}}$ has an empty intersection, it may take some additional computations to obtain a suitable $w = w_N \in V_N$.

So in the ideal situation the sets V_n are small enough to bring the truncation error bound (10) down as much as possible, while at the same time the sets V_n are large enough to allow a nonempty intersection.

3.4. Fractions with positive and negative partial numerators. The condition that (2) has either only positive or only negative partial numerators a_n can be relaxed for $n < N$, as long as it is satisfied from N on. In that case Lemma 2, which gives us a tail estimate, still remains valid, while Lemma 5 is adapted as follows.

LEMMA 7. *If the sequence of convergence sets $\{E_n\}_{n\in\mathbb{N}}$ is given by $E_n = [b_n, c_n]$ with $b_n \geq -1/4$ and $0 \leq b_n c_n$, then the corresponding sequence of value sets $\{V_n\}_{n\in\mathbb{N}}$ is given by $V_n = [L_n, R_n]$ with*

$$L_n = \sum_{k=1}^{\infty} \frac{\beta_{n+k}\rfloor}{\lceil 1} \qquad R_n = \sum_{k=1}^{\infty} \frac{\gamma_{n+k}\rfloor}{\lceil 1}$$

where β_{n+k} and γ_{n+k} are determined by the following rule. If for k odd the product $a_{n+1}\cdots a_{n+k-1} > 0$ or if for k even the product $a_{n+1}\cdots a_{n+k-1} < 0$, then $\beta_{n+k} = b_{n+k}$ and $\gamma_{n+k} = c_{n+k}$. Else $\beta_{n+k} = c_{n+k}$ and $\gamma_{n+k} = b_{n+k}$.

Proof. The rule dictating whether β_{n+k} and γ_{n+k} equal either b_{n+k} or c_{n+k}, where $a_{n+k} \in [b_{n+k}, c_{n+k}]$, comes from the following considerations:

- Irrespective of the sign of a_{n+k} we find that the inequalities $b_{n+k} \leq a_{n+k} \leq c_{n+k}$ are flipped around by the division and result in
$$\frac{1}{1+c_{n+k}} \leq \frac{1}{1+a_{n+k}} \leq \frac{1}{1+b_{n+k}}$$
- For every negative a_{n+k-1} we know that $|c_{n+k-1}| \leq |a_{n+k-1}| \leq |b_{n+k-1}|$ and hence the inequalities are flipped around once more to
$$\frac{c_{n+k-1}}{1+c_{n+k}} \geq \frac{a_{n+k-1}}{1+a_{n+k}} \geq \frac{b_{n+k-1}}{1+b_{n+k}}$$

Hence the number of divisions plus the number of negative partial numerators encountered before one reaches a_{n+k}, determines whether the largest and smallest fraction value are attained with a_{n+k} replaced by either b_{n+k} or c_{n+k}. □

When the restriction that (2) has either only positive or only negative partial numerators a_n, holds from a certain $n \geq M$ on, one can also proceed as follows. A truncation error bound for $|f - f_N(w)|$ can be obtained from a truncation error bound for $\left|f^{(M)} - f^{(M)}_{N-M}(w)\right|$ by means of the relation [11, p. 420]

$$f - f_N(w) = \frac{(-1)^M a_1 \cdots a_M}{B^2_{M-1}\left(h_M + f^{(M)}\right)\left(h_M + f^{(M)}_{N-M}(w)\right)}\left(f^{(M)} - f^{(M)}_{N-M}(w)\right) \qquad (26)$$

Here the sequence $\{h_k\}_{k\in\mathbb{N}}$ is the so-called critical tail sequence given by

$$h_1 = 1$$
$$h_k = 1 + \frac{a_k\rfloor}{\lceil 1} + \ldots + \frac{a_2\rfloor}{\lceil 1} \qquad k \geq 2$$

and B_k is the k-th denominator of (2) which can be computed by means of (12). From (26) we obtain the estimate

$$|f - f_N(w)| \leq \frac{|a_1 \cdots a_M|}{|B_{M-1}|^2 \min^2\left(|L_M + h_M|, |R_M + h_M|\right)}\left|f^{(M)} - f^{(M)}_{N-M}(w)\right|$$

3.5. Rounding error. When implementing $f_N(w)$, we need to take into account that each basic operation $* \in \{+, -, \times, \div\}$ is being replaced by a machine operation $\circledast \in \{\oplus, \ominus, \otimes, \oslash\}$ and hence subject to a relative error of at most $1/2$ `ulp` [5]. Also each partial numerator a_n needs to be converted to a machine number $\breve{a}_n$, hence entailing a relative rounding error ϵ_n given by

$$\breve{a}_n = a_n(1 + \epsilon_n)$$

Here $|\epsilon_n|$ is mostly bounded by $1/2$ `ulp` if $a_n = a_n(x)$ is not a compound expression. Otherwise $|\epsilon_n|$ may be somewhat larger.

Without loss of generality, we assume that $w \in V_N$ is a machine number estimating t_N. So we only need to take into account the distance of w to t_N, which is present in the truncation error via the parameters C_N and R_N determining V_N. We do not have to take into account an additional rounding error on w representing t_N.

When executing the backward recurrence, each computed $\breve{F}_n^{(N)}$ differs from the true $F_n^{(N)}$ by a rounding error $\epsilon_n^{(N)}$, and this for $n = N, \ldots, 1$:

$$\begin{aligned}
\breve{F}_{N+1}^{(N)} &= w \qquad \epsilon_{N+1}^{(N)} = 0 \\
\breve{F}_n^{(N)} &= \breve{a}_n \oslash \left(1 \oplus \breve{F}_{n+1}^{(N)}\right) \qquad n = N, \ldots, 1 \\
&= \frac{\breve{a}_n}{1 + \breve{F}_{n+1}^{(N)}}(1 + \delta_n) \\
&= F_n^{(N)}(1 + \epsilon_n^{(N)}) \\
\breve{F}_1^{(N)} &= F_1^{(N)}(1 + \epsilon_1^{(N)})
\end{aligned}$$

Here δ_n is the relative rounding error introduced in step n of the algorithm. The main question is: how large is $|\epsilon_1^{(N)}|$? This question is answered in Lemma 8 and Theorem 3, the latter being a slight generalization of a result proved in [10] of which we omit the simple proof. Let us introduce the notation

$$\gamma_n^{(N)} = F_{n+1}^{(N)}/(1 + F_{n+1}^{(N)}) \qquad n = 1, \ldots, N$$

LEMMA 8. *Let $\{V_n\}_{n=1}^{\infty}$ be a sequence of value sets for (2). If $F_{N+1}^{(N)} = w \in V_N$, then for $1 \le n \le N$:*

$$|\gamma_n^{(N)}| = \left| \frac{F_{n+1}^{(N)}}{1 + F_{n+1}^{(N)}} \right| \le M = \max_{n=1,\ldots,N} M_n$$

THEOREM 3. *Let $F_{N+1}^{(N)} = w$ be a machine number and let for $n = 1, \ldots, N$*

$$\begin{aligned}
|\epsilon_n| &\le \epsilon \ \texttt{ulp} \\
|\delta_n| &\le \delta \ \texttt{ulp} \\
|\gamma_n^{(N)}| &\le M
\end{aligned}$$

Let the base β and precision t of the IEEE arithmetic in use satisfy

$$\beta^{-t+1}\left(1 + M(1 + 2\epsilon + 2\delta)\frac{M^{N-1} - 1}{M - 1}\right) < 1$$

Then $|\epsilon_1^{(N)}|$ *is bounded by*

$$|\epsilon_1^{(N)}| \leq \epsilon_R := \frac{1}{2}(1+2\epsilon+2\delta)\frac{M^N-1}{M-1}\,\texttt{ulp} \tag{27}$$

If the partial numerators a_n of the continued fraction (2) satisfy $a_n \geq -1/4$, then we know that:

- in case all $a_n > 0$ and $w \leq t_n$, the even approximants satisfy $f_{2k}(x;w) \leq f(x)$,
- in case all $a_n < 0$ and $w \leq t_n$, all approximants satisfy $f_n(x;w) \leq f$.

From Theorem 3 we then obtain for even $N = 2n$ with $F_1^{(N)} = f_{2n}(x;w)$ and $w \leq t_N$, or for general N with $F_1^{(N)} = f_N(x;w)$ and $w \leq t_N$:

$$\frac{|f_N(x;w) - \breve{F}_1^{(N)}|}{|f(x)|} = \left|\epsilon_1^{(N)}\right| \frac{|F_1^{(N)}|}{|f(x)|} \leq \epsilon_R$$

Clearly, ϵ_R is a function of the precision because of its dependence on the `ulp`. When targeting a threshold for ϵ_R, a suitable precision in which to evaluate the continued fraction by means of the backward recurrence algorithm can be computed from this condition. In the sequel this precision is called the working precision s, to distinguish it from the user-defined precision t from which the targeted threshold ϵ is usually obtained.

4. Savouring. In the following we distinguish a number of special cases for the partial numerators a_n:

- $\{a_{2n}\}_{n\in\mathbb{N}}$ increasing and $\{a_{2n+1}\}_{n\in\mathbb{N}}$ decreasing with
$$0 < a_n \qquad 0 < \lim_{n\to\infty} a_{2n} = \tilde{a} = \lim_{n\to\infty} a_{2n+1}$$
- $\{a_n\}_{n\in\mathbb{N}}$ decreasing with $\lim_{n\to\infty} a_n = \tilde{a} \geq 0$;
- $\{a_n\}_{n\in\mathbb{N}}$ decreasing with
$$a_n < 0 \qquad \lim_{n\to\infty} a_n = \tilde{a} \geq -1/4$$
- $\{a_n\}_{n\in\mathbb{N}}$ increasing with $\lim_{n\to\infty} a_n = \tilde{a} \leq 0$;

These cases cover almost all known limit-periodic continued fractions of the elementary functions and a large number of special functions.

4.1. The case $a_{2n} \nearrow \tilde{a}, a_{2n+1} \searrow \tilde{a}, \tilde{a} > 0$**.** Let the sequences of positive real numbers $\{a_{2n+1}\}_{n\in\mathbb{N}}$ and $\{a_{2n}\}_{n\in\mathbb{N}}$ be decreasing and increasing respectively. Let for any $p, q \in \mathbb{N}$,

$$a_{2p+1} > a_{2q}$$

If the a_n can be obtained from a single rounding then we can choose the convergence sets $E_n = [b_n, c_n] = [\bigtriangledown(a_n), \bigtriangleup(a_n)]$. If the a_n are more complicated, then we choose $[b_n, c_n]$ somewhat larger. In both cases the intervals $V_n = [L_n, R_n]$ with L_n and R_n given by (15) are the corresponding value sets, and the bounds (10) and (23) hold.

In addition, we now investigate the monotonicity properties of the sequences $\{L_n\}_{n\in\mathbb{N}}$ and $\{R_n\}_{n\in\mathbb{N}}$. To this end we reconsider the continued fractions $\hat{D}$ and $\hat{U}$ defined in (13) with tails $\hat{D}^{(n)}$ and $\hat{U}^{(n)}$.

Since the sequences $\{\bigtriangledown(a_{2n})\}_{n\in\mathbb{N}}$ and $\{\bigtriangleup(a_{2n+1})\}_{n\in\mathbb{N}}$ are respectively increasing and decreasing as a consequence of the monotonicity of the rounding functions $\bigtriangledown$ and $\bigtriangleup$, Lemma 3 guarantees that the even-numbered tails $\hat{U}^{(2n)} = R_{2n}$ and the odd-numbered tails $\hat{U}^{(2n+1)} = L_{2n+1}$ form a decreasing and increasing sequence respectively, and that for any $p, q \in \mathbb{N}, \hat{U}^{(2q+1)} \leq \hat{U}^{(2p)}$. The same conclusion holds for the tails $\hat{D}^{(2n)} = L_{2n}$ and $\hat{D}^{(2n+1)} = R_{2n+1}$. If the interval $[b_n, c_n]$ is somewhat larger, the same conclusion holds if we can ensure that the sequences $\{b_{2n}\}_{n\in\mathbb{N}}$ and $\{c_{2n+1}\}_{n\in\mathbb{N}}$ are respectively increasing and decreasing.

Following the line of proof of Lemma 3, we can show in addition that

$$R_{2n+1} \leq R_{2k} \qquad k, n \in \mathbb{N}$$
$$L_{2n+1} \leq L_{2k} \qquad k, n \in \mathbb{N}$$

Hence

$$\bigcap_{k=1}^{\infty} V_{2k-1} = L \qquad \bigcap_{k=1}^{\infty} V_{2k} = R$$

with L and R given by (19). When choosing $E_n = [b_n, c_n]$ equal to

$$[b_{2k}, c_{2k}] = [a_{2k}, \tilde{a}] \qquad [b_{2k-1}, c_{2k-1}] = [\tilde{a}, a_{2k-1}] \qquad k \geq 1 \tag{28}$$

inequalities (10) and (23) remain valid, while moreover

$$L = L_{2n} = R_{2n-1} = R = \frac{1}{2}\left(-1 + \sqrt{4\tilde{a}+1}\right)$$

Then, mathematically, a suitable $w \in V_N$ is easy to determine since for all N now $w = -1/2 + \sqrt{\tilde{a} + 1/4} \in V_N$. In practice, the above is applied with $\tilde{a}$ and a_{2k-1} in E_{2k-1} replaced by $\bigtriangledown(\tilde{a})$ and $\bigtriangleup(a_{2k-1})$ respectively, and $\tilde{a}$ and a_{2k} in E_{2k} by $\bigtriangleup(\tilde{a})$ and $\bigtriangledown(a_{2k})$ or similar lower and upper bounds. Then L and R differ slightly, because in an implementation they are given by (19) with $b = \bigtriangledown(\tilde{a})$ and $c = \bigtriangleup(\tilde{a})$ or similarly. Choosing w alternately equal to L and R depending on whether N is odd or even, is a valid choice (care must be taken during the actual computation of w that it is rounded in the appropriate direction, appropriate meaning inside V_N).

Examples of elementary functions illustrating this case include the hyperbolic arcsine,

$$\frac{\operatorname{asinh}(x)}{x\sqrt{1+x^2}} = \cfrac{1}{1} + \sum_{n=1}^{\infty} \cfrac{2\lfloor\frac{n+1}{2}\rfloor(2\lfloor\frac{n+1}{2}\rfloor - 1)x^2/(4n^2-1)}{1} \qquad \lim_{n\to\infty} a_n(x) = \frac{x^2}{4}$$

the hyperbolic arctangent,

$$x\,\operatorname{atanh}(x) = \cfrac{x^2/(1-x^2)}{1} + \sum_{n=1}^{\infty} \cfrac{2\lfloor\frac{n+1}{2}\rfloor(2\lfloor\frac{n+1}{2}\rfloor - 1)x^2/\left((4n^2-1)(1-x^2)\right)}{1}$$
$$\lim_{n\to\infty} a_n(x) = \frac{x^2}{4(1-x^2)}$$

and the natural logarithm,

$$\frac{x}{\ln(x+1)} - 1 = \sum_{n=1}^{\infty} \cfrac{nx/(4n-2)}{1} + \cfrac{nx/(4n+2)}{1} \qquad \lim_{n\to\infty} a_n(x) = x/4$$

We discuss the natural logarithm in more detail in Section 4.2.

Examples of special functions to which the technique applies are the hypergeometric functions

$$ {}_2F_1(a,1;c;x) \qquad a \neq 3/2 \qquad x<0 \qquad \tilde{a} = -x/4 $$

For $a = 3/2$, the technique described in Section 4.3 applies.

4.2. Natural logarithm. A useful continued fraction for which the results of Section 4.1 hold, is

$$ f(x) = \frac{x}{\ln(x+1)} - 1 = \sum_{n=1}^{\infty} \cfrac{nx/(4n-2)}{1} \Big\rfloor + \cfrac{nx/(4n+2)}{1} \Big\rfloor \qquad x+1 \geq 0 $$

Since $\ln(0) = -\infty$, $\ln(1) = 0$ and for $0 < x < 1$ the relation $\ln(x) = -\ln(1/x)$ holds, we only have to consider the case $x > 1$. For a given floating-point number

$$ x = +d_0 \,.\, d_1 \ldots d_{t-1} \times \beta^{e_x} = +m_x \times \beta^{e_x} \qquad d_0 \neq 0, 0 \leq d_i \leq \beta - 1, i = 0, \ldots, t-1 $$

as operand, and a choosen integer $\ell > 0$, we construct the reduced operand $\tilde{x}$,

$$ \begin{aligned} j_x &= \lfloor \log_{\beta^{1/\ell}} m_x \rfloor \\ \tilde{x} &= +d_0 \,.\, d_1 \ldots d_{t-1} \times \beta^{-j_x/\ell} - 1 \end{aligned} $$

which satisfies

$$ \ln(x) = \frac{\tilde{x}}{1+\ln(\tilde{x})} + \left(e_x + \frac{j_x}{\ell} \right) \ln(\beta) \tag{29} $$

For $\beta = 2$ and $\ell = 8$, the range of the reduced argument $\tilde{x}$ is then $[1, 2^{1/8}]$.

With $t = 53$ in (1) and $\epsilon_T = 2^{-52} = 1\texttt{ulp} \approx 2.22 \times 10^{-16}$, we obtain from (23) that $f_N(w)$ with $N = 9$, $w = {}^1\!/_2 \left(-1 + \sqrt{x+1}\right)$ and $2j+1 = 3$ satisfies

$$ \frac{|f(\sqrt[8]{2}-1) - f_9(w)|}{|f(\sqrt[8]{2}-1)|} = 7.98 \times 10^{-17} \leq \frac{R_{9,3} - L_{9,2}}{1+L_{9,2}} \prod_{k=1}^{8} \frac{R_{k,3}}{1+R_{k,3}} \leq 1.98 \times 10^{-16} < \epsilon_T $$

Already for double precision this compares nicely to the targeted complexity of 25 operations. The backward algorithm uses $2N = 18$ operations for the computation of $f_9(w)$, and the functional relationship (29) employs another 6. In the next section the comparison with the state of the art double precision algorithms is even more favourable. For the luxury of having a multiprecision version we have to pay the price of obtaining N from (10) or (22) though, while in a fixed precision implementation N is fixed a priori. Of course we do not have to add this additional work to the complexity count, when comparing the algorithm's complexity to that of a fixed precision algorithm.

Because of the specific behaviour of the sequences $\{R_n\}_{n\in\mathbb{N}}$ and $\{L_n\}_{n\in\mathbb{N}}$ when E_n is given by (28), we have with K even in (22) that $\overline{R} = R_K$ and $\overline{L} = L$. With $K = 4$ formula (22) also predicts $N = 9$. With $K = 2$ we obtain a slight overestimate of N, namely $N = 10$ instead of $N = 9$:

$$ \frac{|f(\sqrt[8]{2}-1) - f_{10}(w)|}{|f(\sqrt[8]{2}-1)|} \leq \left(\frac{R_{2,3} - L}{1+L} \right) M_1 \left(\frac{R_{2,3}}{1+R_{2,3}} \right)^4 M_2^4 \leq 1.06 \times 10^{-17} < \epsilon_T $$

When comparing these results to the classical results obtained for $w = 0$ using (8) (where we approximate $f(\sqrt[8]{2}-1)$ in the denominator of the relative error by $f_K(x;0)$ with K even), we see that

$$\frac{|f(\sqrt[8]{2}-1) - f_{11}(0)|}{|f(\sqrt[8]{2}-1)|} = 1.89\times10^{-17} \leq \frac{2a_1}{f_K(x;0)} \prod_{k=2}^{11} \frac{\sqrt{1+4a_k}-1}{\sqrt{1+4a_k}+1} = 3.77\times10^{-17} < \epsilon_T$$

This implies a gain of about 20% in complexity: $N = 9$ is predicted by (22) and (23) while $N = 11$ is predicted by the Gragg-Warner bound (8). From the multitude of experiments we have conducted we can confirm that this gain is typical when comparing the new technique for hardware precisions. In addition, the precision with which w needs to be computed, is determined by the width of V_N which is at least as wide as $R_{N,2j} - L_{N,2j+1}$. When t is increased while maintaining the rather wide sets E_n specified in (28), often w need not be computed to full precision.

Now let us consider $\epsilon_T = 2^{-511} \approx 1.50 \times 10^{-154}$ and slightly refine the sets E_n. Restricting the variation in the partial numerators to that of the double precision rounding of $a_n(x)$, amounts to choosing E_n like $E_n = [a_n(1-\delta), a_n(1+\delta)]$ with $|\delta| \leq 2^{-52}$. We find that for the same argument x and with $N = 84$ and $2j+1 = 11$:

$$\frac{|f(\sqrt[8]{2}-1) - f_{84}(w)|}{|f(\sqrt[8]{2}-1)|} \leq 5.82 \times 10^{-156} < \epsilon_T \qquad w \in V_{84}$$

Here all factors in the upper bound (23) are only computed in double precision. At the same time, the Gragg-Warner bound predicts $N = 94$. Hence the newly developed technique offers a gain of 20 high precision computations ($\Delta N = 94 - 84$). The practical problem here is to find $w_N \in V_N$. A possible choice is $w_{84} = L_{84,11}$ costing 22 double precision basic operations ($2j + 1 = 11$). This choice for w is only valid if $L_N \leq L_{N,2j+1}$ belongs to V_N, meaning that $L_{N,2j+1} \leq R_N$.

4.3. The case $a_n \searrow \tilde{a} \geq 0$. Let each partial numerator a_n satisfy $\tilde{a} \leq a_n$ where the sequence $\{a_n\}_{n\in\mathbb{N}}$ is a decreasing sequence of positive numbers and $0 \leq \tilde{a} = \lim_{n\to\infty} a_n$. So all a_n are strictly positive. For $E_n = [\bigtriangledown(a_n), \bigtriangleup(a_n)]$ the corresponding value sets are $V_n = [L_n, R_n]$ with L_n and R_n given by (15). For these V_n the bounds (10) and (23) hold. To guarantee that (25) is in $[L_N, R_N]$, we consider slightly larger convergence sets E_n.

For the convergence sets $E_n = [\tilde{a}, a_n]$ the value sets $V_n = [L_n, R_n]$ are bounded by

$$L_n = \cfrac{\tilde{a}}{1} + \cfrac{a_{n+2}}{1} + \cfrac{\tilde{a}}{1} + \cfrac{a_{n+4}}{1} + \dots$$
$$R_n = \cfrac{a_{n+1}}{1} + \cfrac{\tilde{a}}{1} + \cfrac{a_{n+3}}{1} + \cfrac{\tilde{a}}{1} + \dots$$

Following the line of proof of Lemma 3 we find that the sequence $\{R_n\}_{n\in\mathbb{N}}$ is decreasing and that $\{L_n\}_{n\in\mathbb{N}}$ is increasing. Now

$$\bigcap_{k\in\mathbb{N}} V_k = [L, R]$$

with $L = R = \left(-1 + \sqrt{4\tilde{a}+1}\right)/2$. Hence we can choose for all N the tail estimate $w = \left(-1 + \sqrt{4\tilde{a}+1}\right)/2$. In practice this is applied to $E_n = [\bigtriangledown(\tilde{a}), \bigtriangleup(a_n)]$, replacing

in L_n, R_n and (22) every occurrence of $\tilde{a}$ by $\bigtriangledown(\tilde{a})$ and every occurrence of a_n by $\bigtriangleup(a_n)$. Use of these roundings makes $L < R$ and consequently all $w \in [L, R]$ a valid choice.

With $\tilde{a} = 0$, we have $E_k = [0, a_k], R_k = a_{k+1}$ and $L_k = 0$. The truncation error bound simplifies to

$$\frac{|f(x) - f_N(x; w)|}{|f(x)|} \leq a_{N+1} \prod_{k=1}^{N-1} \frac{a_{k+1}}{1 + a_{k+1}} \tag{30}$$

Here $w = 0$ can be used which is a more classical choice.

Examples taken from the class of functions satisfying the conditions imposed in this section involve the arccosine function

$$\frac{x\ \mathrm{acos}(x)}{\sqrt{1-x^2}} = \cfrac{1}{1} + \sum_{n=1}^{\infty} \cfrac{n^2(1-x^2)/\left((2n-1)(2n+1)x^2\right)}{1} \qquad \lim_{n\to\infty} a_n(x) = \frac{(1-x^2)}{4x^2}$$

the arctangent function

$$x\ \arctan(x) = \cfrac{x^2}{1} + \sum_{n=1}^{\infty} \cfrac{n^2x^2/(4n^2-1)}{1} \qquad \lim_{n\to\infty} a_n(x) = \frac{x^2}{4}$$

and the arccotangent function,

$$\frac{\mathrm{acot}(x)}{x} = \cfrac{x^{-2}}{1} + \sum_{n=1}^{\infty} \cfrac{n^2x^{-2}/(4n^2-1)}{1} \qquad \lim_{n\to\infty} a_n(x) = \frac{x^{-2}}{4}$$

The arctangent is discussed in detail in Section 4.4. Examples where (30) applies include the hyperbolic tangent function,

$$x\ \tanh(x) = \cfrac{x^2}{1} + \sum_{n=1}^{\infty} \cfrac{x^2/(4n^2-1)}{1}$$

and the exponential function,

$$\frac{2x}{\exp(x)-1} = 2 - x + \cfrac{x^2/6}{1} + \sum_{n=2}^{\infty} \cfrac{x^2/\left(4(4n^2-1)\right)}{1}$$

Also several special functions belong to this class, among which the confluent hypergeometric functions ${}_1F_1(1; c; x)$.

4.4. Arctangent function. The conditions imposed in Section 4.3 hold for the following continued fraction representation of the arctangent function:

$$\arctan(x) = \cfrac{x}{1} + \sum_{n=1}^{\infty} \cfrac{n^2x^2/(4n^2-1)}{1} \qquad \lim_{n\to\infty} a_n(x) = \frac{x^2}{4}$$

Since $\arctan(-x) = -\arctan(x)$, we only need to focus on positive arguments. More-

over, by using the fact that

$$2-\sqrt{3} < x \le 1 \Rightarrow \left|\frac{x\sqrt{3}-1}{x+\sqrt{3}}\right| \le 2-\sqrt{3} \qquad \arctan(x) = \arctan\left(\frac{x\sqrt{3}-1}{x+\sqrt{3}}\right) + \frac{\pi}{6}$$

$$1 < x \le 2+\sqrt{3} \Rightarrow \left|\frac{\sqrt{3}-x}{x\sqrt{3}+1}\right| \le 2-\sqrt{3} \qquad \arctan(x) = -\arctan\left(\frac{\sqrt{3}-x}{x\sqrt{3}+1}\right) + \frac{\pi}{3}$$

$$2+\sqrt{3} < x \Rightarrow 1/x < 2-\sqrt{3} \qquad \arctan(x) = -\arctan(1/x) + \frac{\pi}{2}$$

one can reduce the argument of $\arctan(x)$ to the interval $[0, 2-\sqrt{3}]$. For $x \in [0, 2-\sqrt{3}]$ the value $\arctan(x) \in [0, \pi/12]$. The sequence of partial numerators is a decreasing positive sequence from the very first partial numerator on.

With $t = 53$ in (1) and $\epsilon_T = 2^{-52} = 1\texttt{ulp}$, we obtain from (23) that $f_N(w)$ with $N = 8$, $w = 1/2\left(-1+\sqrt{x^2+1}\right)$ and $2j+1 = 3$ satisfies

$$\frac{|f(2-\sqrt{3}) - f_8(w)|}{|f(2-\sqrt{3})|} = 4.74 \times 10^{-17} \le \frac{R_{8,3}-L_{8,2}}{1+L_{8,2}} \prod_{k=1}^{7} \frac{R_{k,3}}{1+R_{k,3}} \le 5.26 \times 10^{-17} < \epsilon_T \tag{31}$$

When using (22) with $K = 3$ we find $N = 9$. With $K = 4$ we obtain again $N = 8$. The upper bound for the truncation error is then estimated by

$$\frac{|f(2-\sqrt{3}) - f_8(w)|}{|f(2-\sqrt{3})|} \le \left(\frac{R_{8,3}-L_{8,2}}{1+L_{8,2}}\right) M_1\ M_2\ M_3\ M_4^4 < \epsilon_T$$

Compared to the classical approximant $f_N(0)$ for use with the Gragg-Warner bound, we find

$$\frac{|f(2-\sqrt{3}) - f_9(0)|}{|f(2-\sqrt{3})|} = 2.18 \times 10^{-16} \le \epsilon_T \tag{32}$$

$$\frac{|f(2-\sqrt{3}) - f_{11}(0)|}{|f(2-\sqrt{3})|} \le \frac{2a_1}{f_K(0)} \prod_{k=2}^{11} \frac{\sqrt{1+4a_k}-1}{\sqrt{1+4a_k}+1} = 7.56 \times 10^{-18} < \epsilon_T$$

Hence the new technique delivers $N = 8$ while the traditional Gragg-Warner bound estimates $N = 11$, which is an overestimate of N in (32) by 2. Again the use of a nonzero tail estimator in (31) together with the newly developed bounds offer a clear advantage.

Choosing $t = 113$ (as in the quadruple precision supported by Sun) and $\epsilon_T = 2^{-112} \approx 1.93 \times 10^{-34}$ we find with $N = 18, w = 1/2\left(-1+\sqrt{x^2+1}\right)$ and $2j+1 = 3$:

$$\frac{|f(2-\sqrt{3}) - f_{18}(w)|}{|f(2-\sqrt{3})|} = 5.65 \times 10^{-39} \le \frac{R_{18,3}-L_{18,2}}{1+L_{18,2}} \prod_{k=1}^{17} \frac{R_{k,3}}{1+R_{k,3}} \le 3.35 \times 10^{-35} < \epsilon_T$$

The traditional Gragg-Warner bound for use with $w = 0$ delivers $N = 21$ while

$$\frac{|f(2-\sqrt{3}) - f_{20}(0)|}{|f(2-\sqrt{3})|} = 9.35 \times 10^{-36}$$

Again a clear advantage of a nonzero choice for w over the standard $w = 0$.

4.5. The case $a_n \searrow \tilde{a} \geq -1/4$. For simplicity we take all $a_n < 0$. When this is only the case from an index $n = M$ on, then the technique described in Section 3.4 can be used to estimate the truncation error from the beginning of the fraction, given the truncation error from a_M on. Since all $L_n \geq -1/2$, the error bound (10) remains valid.

We choose $E_n = [a_{n+1}, a_n]$ and compute $V_n = [L_n, R_n]$ with L_n and R_n given by (16). Apparently $L_n = t_{n+1}$ and $R_n = t_n$. The sequences $\{R_n\}_{n\in\mathbb{N}}$ and $\{L_n\}_{n\in\mathbb{N}}$ are decreasing with

$$\lim_{n\to\infty} R_n = \frac{-1+\sqrt{4\tilde{a}+1}}{2} = \lim_{n\to\infty} L_n$$

Combining part 2 of Lemma 1 and part 3 of Lemma 2 provides the computable bounds

$$L_{n,j}\left(\frac{-1+\sqrt{4\tilde{a}+1}}{2}\right) \leq L_n \leq R_n \leq R_{n,j}\left(\frac{-1+\sqrt{4a_{n+1}+1}}{2}\right) \qquad j \geq 1$$

Under the condition that $L_N \leq R_{N,j}(1/2(-1+\sqrt{4\tilde{a}+1}))$, a valid choice for $w_N \in V_N$ is given by $w_N = R_{N,j}(1/2(-1+\sqrt{4\tilde{a}+1}))$. Similar computational remarks as in Section 3.3 and 4.1 apply, concerning the effect of rounding errors on the computed $L_{n,j}, R_{n,j}$ and w_N.

Special functions enjoying a limit-periodic continued fraction with $a_n \geq -1/4$ and $\tilde{a} = -1/4$ are:

$$\mathrm{E}_m(x) = \cfrac{\exp(-x)/(x+m)}{1} + \sum_{n=2}^{\infty} \cfrac{\frac{-n(m+n-2)}{(x+m+2(n-1))(x+m+2(n-2))}}{1} \qquad m \neq 1, x \neq 0$$

$$\mathrm{erfc}(x) = \cfrac{\frac{\exp(-x^2)}{2\sqrt{\pi}x(2x^2+1)}}{1} + \sum_{n=2}^{\infty} \cfrac{\frac{-2n(2n-1)}{(2x^2+1+4n)(2x^2+1+4(n-1))}}{1} \qquad x \neq 0 \tag{33}$$

As an illustration of the technique we compute the function $f(x)$, defined by the first tail of (33), up to 80 bits accuracy, in other words with $\epsilon_T = 2^{-79} \approx 1.65 \times 10^{-24}$. This function has only negative partial numerators. With E_n and V_n as given above and $x = 2$, the bound (10) is less than ϵ_T for $N \geq 59$. For $j = 12$ we obtain in addition that

$$t_{N+1} = L_N < L_{N,j}\left(\frac{-1+\sqrt{4a_{N+2}+1}}{2}\right) < R_{N,j}(-1/2) < R_N = t_N$$

and hence that all w satisfying

$$L_{N,j}\left(\frac{-1+\sqrt{4a_{N+2}+1}}{2}\right) < -3.7626 \times 10^{-1} \leq w \leq -3.7527 \times 10^{-1} < R_{N,j}(-1/2)$$

are valid choices for the approximation of $f(x)$ by $f_N(x; w)$, since they belong to V_N guaranteed.

4.6. The case $a_n \nearrow \tilde{a} \leq 0$. When the partial numerators a_n are negative and increasing with $\lim_{n\to\mathbb{N}} a_n = \tilde{a}$, we can choose $E_n = [a_n, \tilde{a}]$. The value sets V_n are then given by $V_n = [L_n, R_n] = \left[t_n, \frac{-1+\sqrt{4\tilde{a}+1}}{2}\right]$ where it can be proved as in Lemma

4, that the sequence $\{t_n\}_{n\in\mathbb{N}}$ is increasing. Following the line of proof of part 2 of Lemma 1, we find that

$$\frac{-1+\sqrt{4a_{n+1}+1}}{2} \le L_n$$

Hence the truncation error can be bounded by

$$|f(x) - f_N(x;w)| \le \frac{1-\sqrt{4a_{N+1}+1}}{1+\sqrt{4a_{N+1}+1}} \prod_{k=1}^{N-1} \frac{1-\sqrt{4a_{k+1}+1}}{1+\sqrt{4a_{k+1}+1}} \qquad w \in V_N$$

When $\tilde{a} = 0$ this is very similar to the Gragg-Warner bound on the real line ($\phi = \arg(x) = 0$). Note that $w = 0$ is a valid choice if $\tilde{a} = 0$. This truncation error bound applies for instance to the function $f(x)$ given by

$$f(x) = \frac{x}{\tan(x)} - 1 = \sum_{n=1}^{\infty} \cfrac{-x^2/(4n^2-1)}{1}$$

4.7. Conclusion. The truncation error bounds obtained here for $f_N(x;w)$, are more general than the ones given in [9] or [2] which contain more stringent conditions on the a_n that are not as easily satisfied for all n, and the ones found in [11] where $C_N = \tilde{w}$ with $\tilde{w}$ given by (7) is recommended for (9). Other truncation error bounds that can be found in the literature [8, 7] mainly hold for $w = 0$.

When classifying all these bounds with respect to sharpness, we have to distinguish between bounds for the unmodified approximant $f_N(x;0)$ and bounds for the modified approximant $f_N(x;w)$. As can be expected, the a posteriori error bounds, which take advantage of the information contained in continued fraction approximants computed for smaller values of N, are the most accurate truncation error bounds for $f_N(x;0)$. The newly developed a priori error bound is to the modified approximant $f_N(x;w)$ what the Gragg-Warner bound is to the unmodified approximant $f_N(x;0)$. Both are of comparable quality, which is precisely our goal.

An analysis of the complexity of the proposed implementation (counting additions/subtractions and multiplications/divisions) illustrates that the multiprecision continued fraction method presented here even compares well to the state-of-the-art fixed precision techniques which are in use. When choosing the radix $\beta = 2$ and the precision $t = 53$ (standard IEEE 754 double precision) then the proposed continued fraction model allows to evaluate the elementary functions in about 20 to 25 operations, including the computation of the square root modification $w = \tilde{w}$, which compares favourably to the current double precision implementations. Moreover, the proposed technique is generic in the sense that it can be used for any user-definable precision.

Acknowledgement. Thanks are due to Stefan Becuwe and Johan Vervloet, for implementing a lot of our ideas in either Maple or C++. The insight gained from the numerical illustrations guided us in the selection of the material for this paper and will assist us with the implementation of a reliable special function library in the future.

REFERENCES

[1] G.A. Baker, Jr. and P. Graves-Morris. *Padé approximants (2nd Ed.)*. Cambridge University Press, 1996.

[2] Ch. Baltus and W.B. Jones. Truncation error bounds for modified continued fractions with applications to special functions. *Numer. Math.*, 55:281–307, 1989.

[3] B.A. Cipra. A new testament for special functions? *SIAM News*, 31(2), 1998.

[4] W.J. Cody and W. Waite. *Software manual for the elementary functions*. Prentice Hall, New Jersey, 1980.

[5] A. Cuyt and B. Verdonk. *Computer arithmetic and Numerical techniques*. `www.cant.ua.ac.be/`.

[6] W. Gautschi. Computational aspects of three-term recurrence relations. *SIAM Rev.*, 9(1):24–82, 1987.

[7] W.B. Gragg and D.D. Warner. Two constructive results in continued fractions. *SIAM J. Numer. Anal.*, 20(6):1187–1197, 1983.

[8] Peter Henrici and Pia Pfluger. Truncation error estimates for Stieltjes fractions. *Numer. Math.*, 9:120–138, 1966.

[9] L. Jacobsen. Convergence acceleration and analytic continuation by means of modifications of continued fractions. In H. Waadeland et al., editors, *Padé Appr. and Cont. Fr.*, pages 19–33, 1983.

[10] W.B. Jones and W.J. Thron. Numerical stability in evaluating continued fractions. *Math. Comp.*, 28:795–810, 1974.

[11] L. Lorentzen. A priori truncation error bounds for continued fractions. *Rocky Mountain J. Math.*, 33:409–474, 2003.

[12] L. Lorentzen and H. Waadeland. *Continued fractions with applications*. North-Holland Publishing Co., Amsterdam, 1992.

[13] Sun Microsystems. *fdlibm*, 1995. `www.netlib.org/fdlibm`.

[14] C. Moler. The tetragamma function and numerical craftsmanship. *MATLAB News & Notes*, February 2002.

[15] J.-M. Muller. *Elementary functions: Algorithms and implementation*. Birkhäuser, 1997.

[16] W.J. Thron and H. Waadeland. Accelerating convergence of limit periodic continued fractions $K(a_n/1)$. *Numer. Math.*, 34:155–170, 1980.