Towards Reliable Software for the Evaluation of a Class of Special Functions

Annie Cuyt and Stefan Becuwe

Universiteit Antwerpen
Departement Wiskunde en Informatica
Middelheimlaan 1, BE-2020 Antwerpen, Belgium
{annie.cuyt, stefan.becuwe}@ua.ac.be

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. But 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. Here the notion reliable indicates that, together with the function evaluation a guaranteed upper bound on the total error or, equivalently, an enclosure for the exact result, is computed.

We point out how limit-periodic continued fraction representations of these functions can be helpful in this respect. The newly developed (and implemented) scalable precision technique is mainly based on the use of sharpened a priori truncation error and round-off error upper bounds for real continued fraction representations of special functions of a real variable. The implementation is reliable in the sense that it returns a sharp interval enclosure for the requested function evaluation, at the same cost as the evaluation.

1 Basic Continued Fraction Material

Let us consider a continued fraction representation of the form

$$f = \frac{a_1}{1+\dfrac{a_2}{1+\ldots}} = \frac{a_1}{\left\lfloor 1 \right\rfloor} + \frac{a_2}{\left\lfloor 1 \right\rfloor} + \ldots = \sum_{n=1}^{\infty} \frac{a_n}{\left\lfloor 1 \right\rfloor}, \quad a_n := a_n(x), \quad f := f(x). \tag{1}$$

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 $f_N(w_N)$ or $f_N(x; w_N)$, and N-th tail t_N or $t_N(x)$ of (1), the values

$$f_N(w_N) = f_N(x; w_N) = \sum_{n=1}^{N-1} \frac{a_n}{\left\lfloor 1 \right\rfloor} + \frac{a_N}{\left\lfloor 1 + w_N \right\rfloor}, \tag{2}$$

A. Iglesias and N. Takayama (Eds.): ICMS 2006, LNCS 4151, pp. 308–319, 2006.

$$t_N = t_N(x) = \sum_{n=N+1}^{\infty} \frac{a_n}{\lceil 1}, \qquad t_0 = f. \tag{3}$$

We also need approximants of continued fraction tails and therefore introduce the notation $f_N^{(k)}(w_{N+k})$ or $f_N^{(k)}(x; w_{N+k})$ for

$$f_N^{(k)}(w_{N+k}) = f_N^{(k)}(x; w_{N+k}) = \sum_{n=k+1}^{k+N-1} \frac{a_n}{\lceil 1} + \frac{a_{N+k}}{\lceil 1 + w_{N+k}}.$$

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 (1) to converge. Moreover, we restrict ourselves to the case where some $w_N \neq 0$ can be chosen such that

$$\lim_{N\to\infty} f_N(w_N) = \lim_{N\to\infty} f_N(0).$$

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

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

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

$$s_n(V_n) = \frac{a_n}{1+V_n} \subseteq V_{n-1}, \qquad n \geq 1. \tag{4}$$

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 $w_N \in V_N$ then $f_N(w_N) \in V_0$ and $f_{N-k}^{(k)}(w_N) \in V_k$. 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 (1) as long as each partial numerator a_n belongs to the respective set E_n:

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

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 (4) 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

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

Each tail t_N converges to $\frac{1}{2}(\sqrt{1+4x}-1)$ 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 $+\infty$. Very accurate approximants $f_N(w_N)$ for f can be computed by making an appropriate choice for the tail estimate $w_N\approx t_N$.

We call a continued fraction of the form (1) 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 (1) 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 [8],

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

Hence a suitable choice of w in (2) may result in more rapid convergence of the approximants ($w=0$ is usually used as a reference).

In this paper we further restrict the condition that (1) 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 [7, pp. 150–159]. This condition automatically implies that $\tilde{a}\geq -1/4$ and $\tilde{w}$ is real.

2 Truncation Error

Most truncation error upper bounds for $|f(x)-f_N(x;w_N)|$ are given for the classical choice $w_N=0$. 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

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

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

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

In [4] we prove a practical and sharp truncation error bound for the case $w_N \neq 0$, which is valid for all continued fractions with real partial numerators $a_n(x)$. This result departs from the Oval Sequence Theorem [7, pp. 145–147], which holds in the complex plane, from which a priori truncation error estimates can be obtained in case $w_N \neq 0$. In the real case the involved value sets V_n and convergence sets E_n are intervals and the theorem can be simplified and sharpened to the real Interval Sequence Theorem [4], here Theorem 1.

Theorem 1. *Let for all n the values L_n and R_n satisfy $-1/2 \leq L_n \leq R_n < \infty$ 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 $b_n \leq c_n$ and $0 \leq b_n c_n$. 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

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

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

$$\left| \frac{f(x) - f_N(x;w_N)}{f(x)} \right| \leq \frac{R_N - L_N}{1 + L_N} \prod_{k=1}^{N-1} M_k \tag{5}$$

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

In Theorem 1 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. In the following Lemma 1 [4] we formulate L_n and R_n in terms of the bounds on a_n in E_n and associate intervals V_n with given intervals E_n, instead of the other way around. Let $E_n = [b_n, c_n]$ with $-1/4 \leq b_n \leq c_n$ and $b_n c_n \geq 0$. The condition that b_n and c_n have the same sign means nothing more than that at least $\operatorname{sign}(a_n)$ is kept fixed in E_n.

Lemma 1. *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*

$$\hat{D} = \frac{b_1}{\left|1\right.} + \frac{c_2}{\left|1\right.} + \frac{b_3}{\left|1\right.} + \frac{c_4}{\left|1\right.} + \ldots,$$
$$\hat{U} = \frac{c_1}{\left|1\right.} + \frac{b_2}{\left|1\right.} + \frac{c_3}{\left|1\right.} + \frac{b_4}{\left|1\right.} + \ldots,$$

and

$$\check{D} = \cfrac{b_1\rfloor}{1} + \cfrac{b_2\rfloor}{1} + \cfrac{b_3\rfloor}{1} + \cfrac{b_4\rfloor}{1} + \dots,$$
$$\check{U} = \cfrac{c_1\rfloor}{1} + \cfrac{c_2\rfloor}{1} + \cfrac{c_3\rfloor}{1} + \cfrac{c_4\rfloor}{1} + \dots,$$

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* $b_n \geq 0$*:*

$$\begin{aligned} L_{2j} &= \hat{D}^{(2j)}, \qquad & L_{2j-1} &= \hat{U}^{(2j-1)}, \\ R_{2j} &= \hat{U}^{(2j)}, \qquad & R_{2j-1} &= \hat{D}^{(2j-1)}, \end{aligned} \tag{6}$$

and when all $b_n \leq 0$*:*

$$L_n = \check{D}^{(n)}, \qquad R_n = \check{U}^{(n)}. \tag{7}$$

3 Round-Off Error

Several algorithms exist for the computation of $f_N(w)$, the most stable [5] being the backward recurrence algorithm

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

For the backward recurrence algorithm to be useful in a scalable precision context, it must be possible to determine N rather easily a priori, in other words which approximant to compute.

When actually implementing $f_N(w_N)$, we need to take into account that each basic operation $* \in \{+, -, \times, \div\}$ is being replaced by an (IEEE compliant) floating-point operation $\circledast \in \{\oplus, \ominus, \otimes, \oslash\}$. Such a floating-point implementation is characterized by four parameters, being the base β used for the internal number representation, the precision or amount p of β-digits, and the exponent range $[e_{\min}, e_{\max}]$ allowed in the floating-point notation. Usually the rounding mode in use is round-to-nearest (with the proper tie break). Each basic floating-point operation $x \circledast y$ is then subject to a relative error of at most $1/2$ `ulp` [3] where one `ulp` or unit-in-the-last-place equals β^{-p+1}. Also each partial numerator a_n needs to be converted to a floating-point number $\check{a}_n$, hence entailing a relative rounding error ϵ_n given by

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

Here $|\epsilon_n|$ is usually no more than a few `ulp`. Without loss of generality, we assume that $w_N \in V_N$ is a floating-point number estimating t_N. When executing the backward recurrence, each computed $\check{F}_n^{(N)}$ then differs from the true $F_n^{(N)}$ by a

rounding error $\epsilon_n^{(N)}$, and this for $n = N, \ldots, 1$, in other words

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

Here δ_n is the relative rounding error introduced in step n of the algorithm. The question how large $|\epsilon_1^{(N)}|$ is, is answered in Lemma 2 [6] and Theorem 2, the latter being a slight generalization of a result proved in [6]. Let us introduce the notation

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

Lemma 2. *Let $\{V_n\}_{n=1}^{\infty}$ be a sequence of value sets for (1). If $F_{N+1}^{(N)} = w_N \in V_N$, 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 2. *Let $F_{N+1}^{(N)} = w_N$ be a floating-point 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 p of the IEEE arithmetic in use satisfy

$$\left(1 + M(1 + 2\epsilon + 2\delta)\frac{M^{N-1} - 1}{M - 1}\right) \ \texttt{ulp} < 1.$$

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

$$|\epsilon_1^{(N)}| \le \frac{1}{2}(1 + 2\epsilon + 2\delta)\frac{M^N - 1}{M - 1} \ \texttt{ulp}.$$

From Theorem 2 we obtain for the relative round-off error:

$$\frac{|f_N(x; w_N) - \breve{F}_1^{(N)}|}{|f(x)|} = \left|\epsilon_1^{(N)}\right| \frac{|F_1^{(N)}|}{|f(x)|} \le \frac{1 + 2\epsilon + 2\delta}{2} \ \frac{M^N - 1}{M - 1} \frac{|F_1^{(N)}|}{|f(x)|} \beta^{-p+1}. \quad (8)$$

4 Towards a Reliable Implementation

Let us denote the right hand side of (5) by ϵ_T and the right hand side of (8) by ϵ_R. Clearly

$$\epsilon_T = \epsilon_T(N, b_1, \ldots, b_N, c_1, \ldots, c_N)$$

and

$$\epsilon_R = \epsilon_R(N, \beta, p, M_1, \ldots, M_N).$$

In order to guarantee that $\breve{F}_1^{(N)}$ has s significant β-digits, meaning that

$$\frac{\left| f - \breve{F}_1^{(N)} \right|}{|f|} \leq \epsilon_T + \epsilon_R \leq \sigma = \frac{\beta}{2}\beta^{-s}$$

we proceed as follows:

- we determine N (and $w_N \in V_N$) from the condition

$$\epsilon_T(N, b_1, \ldots, b_N, c_1, \ldots, c_N) \leq \tau < \sigma \tag{9}$$

- we determine a suitable precision p (for chosen β) from the condition

$$\epsilon_R(N, \beta, p, M_1, \ldots, M_N) \leq \rho < \sigma \tag{10}$$

with $\tau \geq 0, \rho \geq 0, \tau + \rho = \sigma$. The former condition directly involves the inaccuracy $|c_n - b_n|$ that we allow for the partial numerators a_n. The latter condition depends on the sequence of values M_n, hence on the L_n and R_n which can be obtained from the b_n and c_n.

Obtaining a useful value w_N is the remaining crucial step. To this end we need to establish a few new results. We further distinguish between

- limit-periodic continued 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 limit-periodic 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 their mutual limit $\tilde{a}$.

Let us denote the j-th approximants of R_k and L_k as given by (6) and (7) in Lemma 1, by $R_{k,j}(\omega_j)$ and $L_{k,j}(\omega_j)$ respectively. For the tail estimates in (6) and (7) we switch to the notation ω_j instead of the traditional w_j used in (2) in order to avoid confusion between the different tails. Detailed proofs of the new results will be given in future work [2]. For the time being we focus on the role of these results in a procedure for the reliable evaluation of special functions that allow a limit-periodic continued fraction representation (in a certain region of the real variable x).

For the accurate computation of a suitable N from (9) we need to know $|R_k - L_k|$ for $k = 1, \ldots, N$, in other words an upper bound for R_k and a lower bound for L_k. In order to obtain a suitable w_N, meaning a value $w_N \in V_N$, we need to know the interior of $[L_N, R_N]$ or an upper bound for L_N and a lower

bound for R_N. Both can be realized by computing enclosures for the values L_k and R_k. These upper and lower bounds for L_k and R_k are given in the Lemmas 3, 4 and 5. Some additional care needs to be taken, but for the moment we restrict ourselves to the headlines of the technique. More details will be given in [2].

4.1 Case a_n Positive

When (1) has positive partial numerators a_n, then the values M_k in Theorem 1 equal

$$M_k = \frac{R_k}{1+R_k}, \qquad k = 1, \ldots, N-1.$$

In Lemma 3 we explicit the bounds on L_k and R_k in case the partial numerators a_n show an oscillatory behaviour towards the limit $\tilde{a}$. In Lemma 4 we treat the case where the a_n decrease monotonically to $\tilde{a}$.

Lemma 3. *Let the sequences* $\{a_{2n+1}\}_{n\in\mathbb{N}}, \{b_{2n+1}\}_{n\in\mathbb{N}}, \{c_{2n+1}\}_{n\in\mathbb{N}}$ *and the sequences* $\{a_{2n}\}_{n\in\mathbb{N}}, \{b_{2n}\}_{n\in\mathbb{N}}, \{c_{2n}\}_{n\in\mathbb{N}}$ *respectively decrease and increase to their mutual limit* $\tilde{a}$. *With*

$$\begin{aligned}
2\omega &= -1 + \sqrt{4\tilde{a}+1},\\
2\omega^{(\ell)}_{k,2j-1} &= c_{k+2j} - b_{k+2j+1} - 1 + \sqrt{4c_{k+2j} + (c_{k+2j} - b_{k+2j+1} - 1)^2},\\
2\omega^{(\ell)}_{k,2j} &= b_{k+2j+1} - c_{k+2j+2} - 1 + \sqrt{4b_{k+2j+1} + (b_{k+2j+1} - c_{k+2j+2} - 1)^2},\\
2\omega^{(r)}_{k,2j} &= c_{k+2j+1} - b_{k+2j+2} - 1 + \sqrt{4c_{k+2j+1} + (c_{k+2j+1} - b_{k+2j+2} - 1)^2},\\
2\omega^{(r)}_{k,2j-1} &= b_{k+2j} - c_{k+2j+1} - 1 + \sqrt{4b_{k+2j} + (b_{k+2j} - c_{k+2j+1} - 1)^2},
\end{aligned}$$

the following bounds can be given for L_k *and* R_k *where* $\ell \geq 1$ *and* $j \geq 0$*:*

$$\begin{aligned}
L_{2\ell-1,2j}(\omega^{(\ell)}_{2\ell-1,2j}) \leq L_{2\ell-1} \leq L_{2\ell-1,2j}(\omega),\\
L_{2\ell-1,2j+1}(\omega^{(\ell)}_{2\ell-1,2j+1}) \leq L_{2\ell-1} \leq L_{2\ell-1,2j+1}(\omega),\\
L_{2\ell,2j}(\omega) \leq L_{2\ell} \leq L_{2\ell,2j}(\omega^{(\ell)}_{2\ell,2j}),\\
L_{2\ell,2j+1}(\omega) \leq L_{2\ell} \leq L_{2\ell,2j+1}(\omega^{(\ell)}_{2\ell,2j+1}),
\end{aligned}$$

and

$$\begin{aligned}
R_{2\ell-1,2j}(\omega^{(r)}_{2\ell-1,2j}) \leq R_{2\ell-1} \leq R_{2\ell-1,2j}(\omega),\\
R_{2\ell-1,2j+1}(\omega^{(r)}_{2\ell-1,2j+1}) \leq R_{2\ell-1} \leq R_{2\ell-1,2j+1}(\omega),\\
R_{2\ell,2j}(\omega) \leq R_{2\ell} \leq R_{2\ell,2j}(\omega^{(r)}_{2\ell,2j}),\\
R_{2\ell,2j+1}(\omega) \leq R_{2\ell} \leq R_{2\ell,2j+1}(\omega^{(r)}_{2\ell,2j+1}).
\end{aligned}$$

Lemma 4. *Let the sequences* $\{a_n\}_{n\in\mathbb{N}}, \{b_n\}_{n\in\mathbb{N}}, \{c_n\}_{n\in\mathbb{N}}$ *all decrease to* $\tilde{a} \geq 0$. *With*

$$\begin{aligned}
2\omega_{k,j}^{(01)} &= \tilde{a} - c_{k+j+2} - 1 + \sqrt{4\tilde{a} + (\tilde{a} - c_{k+j+2} - 1)^2},\\
2\omega_{k,j}^{(10)} &= b_{k+j+1} - \tilde{a} - 1 + \sqrt{4b_{k+j+1} + (b_{k+j+1} - \tilde{a} - 1)^2},\\
2\omega_{k,j}^{(20)} &= c_{k+j+1} - \tilde{a} - 1 + \sqrt{4c_{k+j+1} + (c_{k+j+1} - \tilde{a} - 1)^2},\\
2\omega_{k,j}^{(02)} &= \tilde{a} - b_{k+j+2} - 1 + \sqrt{4\tilde{a} + (\tilde{a} - b_{k+j+2} - 1)^2},
\end{aligned}$$

the following bounds can be given for L_k *and* R_k *where* $k \geq 1$ *and* $j \geq 0$*:*

$$\begin{aligned}
L_{k,2j}(\omega_{k,2j}^{(02)}) \leq L_k \leq L_{k,2j}(\omega_{k,2j}^{(10)}),\\
L_{k,2j+1}(\omega_{k,2j+1}^{(20)}) \leq L_k \leq L_{k,2j+1}(\omega_{k,2j+1}^{(01)}),
\end{aligned}$$

and

$$\begin{aligned}
R_{k,2j}(\omega_{k,2j}^{(01)}) \leq R_k \leq R_{k,2j}(\omega_{k,2j}^{(20)}),\\
R_{k,2j+1}(\omega_{k,2j+1}^{(10)}) \leq R_k \leq R_{k,2j+1}(\omega_{k,2j+1}^{(02)}),
\end{aligned}$$

4.2 Case a_n Negative

When (1) has negative partial numerators a_n, then the values M_k in Theorem 1 equal

$$M_k = \frac{|L_k|}{1 + L_k}, \qquad k = 1, \ldots, N-1.$$

In Lemma 5 we explicit the bounds on L_k and R_k in case the partial numerators a_n form a monotonic sequence towards the limit $\tilde{a}$, either decreasing or increasing.

Lemma 5. *Let* $k \geq 1, j \geq 0$ *and*

$$\begin{aligned}
2\omega &= -1 + \sqrt{4\tilde{a} + 1},\\
2\omega_{k,j}^{(\ell)} &= -1 + \sqrt{4b_{k+j+1} + 1},\\
2\omega_{k,j}^{(r)} &= -1 + \sqrt{4c_{k+j+1} + 1}.
\end{aligned}$$

If the sequences $\{a_n\}_{n\in\mathbb{N}}, \{b_n\}_{n\in\mathbb{N}}, \{c_n\}_{n\in\mathbb{N}}$ *are decreasing with* $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = \lim_{n\to\infty} c_n = \tilde{a}$*, then*

$$\begin{aligned}
L_{k,j}(\omega) \leq L_k \leq L_{k,j}(\omega_{k,j}^{(\ell)}),\\
R_{k,j}(\omega) \leq R_k \leq R_{k,j}(\omega_{k,j}^{(r)}).
\end{aligned}$$

If the sequences $\{a_n\}_{n\in\mathbb{N}}, \{b_n\}_{n\in\mathbb{N}}, \{c_n\}_{n\in\mathbb{N}}$ *are increasing with* $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = \lim_{n\to\infty} c_n = \tilde{a}$*, then*

$$\begin{aligned}
L_{k,j}(\omega_{k,j}^{(\ell)}) \leq L_k \leq L_{k,j}(\omega),\\
R_{k,j}(\omega_{k,j}^{(r)}) \leq R_k \leq R_{k,j}(\omega).
\end{aligned}$$

4.3 Mixed Case

The condition that (1) has either only positive or only negative partial numerators a_n can be relaxed, as long as it is satisfied from a certain n on. If the number of terms with mixed behaviour is small, we can proceed as in [4]. If it is larger, then an alternative technique based on a combination of a small number of predictions and corrections, can be used [1]. The latter uses the same estimates as given in the Lemmas 3, 4 and 5.

5 Numerical Illustration

The collection of functions that can be evaluated reliably using this technique is impressive. It essentially includes all functions that have a known limit-periodic continued fraction representation. If the behaviour of the partial numerators a_n (increasing, decreasing, oscillating) is known then the current technique can be applied. If the behaviour varies as n grows, like in some hypergeometric functions, the related technique explained in [1] can be applied.

Without the ambition of being exhaustive, we are currently working at implementations for:

- the (lower and upper) incomplete gamma functions $\gamma(a,x)$ and $\Gamma(a,x)$,
- the error and complementary error function, Dawson's integral, the exponential integrals and several probability distributions that can be expressed in terms of these functions,
- the hypergeometric and confluent hypergeometric functions ${}_2F_1(a,1;c;x)$ and ${}_1F_1(1;b;x)$, and several ratios of hypergeometric and confluent hypergeometric functions,
- particular ratios of Bessel, spherical Bessel, modified Bessel, modified spherical Bessel, Whittaker and parabolic cylinder functions.

Here we give two numerical examples, one where the continued fraction representation (1) has positive a_n and one where the partial numerators a_n are negative.

5.1 Positive a_n

We consider

$$f(a,x) = \frac{a\gamma(a,x)e^x}{x^a} = \frac{\frac{a}{a-x}}{1} + \sum_{n=2}^{\infty} \frac{\frac{(n-1)x}{(a+n-1-x)(a+n-2-x)}}{1} \tag{11}$$

where $\gamma(a,x)$ is the (lower) incomplete gamma function. The sequence $\{a_n\}_{n\in\mathbb{N}}$ is decreasing with $\tilde{a}=0$. Then L_k and R_k simplify to $L_k = 0$ and $R_k = a_{k+1}$ for $[b_n, c_n] = [0, a_n]$. We take $x = 1$ and $a = 9/2$ and require $f(a,x)$ to be evaluated with

$$\epsilon_T + \epsilon_R \leq 10^{-d+1}, \qquad d = 73, 74, \ldots, 80.$$

where τ and ρ in (9) and (10) are both taken equal to 5×10^{-d}. The results can be found in Table 1. Let us zoom in on the first line of output. For $d = 73$, the bound ϵ_T given by (5) is less than 2.0×10^{-73} if $N \geq 49$. Subsequently we choose our working precision p in (8) so as to keep ϵ_R below 5.0×10^{-73}. Here we take $\beta = 10$ because we are going to compare our evaluation with that given by the multiprecision implementation of Maple. From Theorem 1 we learn that all w_N satisfying

$$L_N = 0 \leq w_N \leq 1.812 \times 10^{-2} < R_N = a_{N+1}$$

are valid choices as a tail estimate, the easiest being $w_N = 0$.

In Table 2 we have set `Digits` in Maple to d and printed the result for the evaluation of $f(a, x)$ delivered by this computer algebra system. Clearly the evaluation in Maple is subject to a much larger error (2 or 3 trailing decimal digits are inaccurate in this case).

Table 1. Continued fraction library output

73	1.21400959177351261777749873464519839007959605662228349187716240969187970 0
74	1.21400959177351261777749873464519839007959605662228349187716240969187970 00
75	1.21400959177351261777749873464519839007959605662228349187716240969187969998
76	1.21400959177351261777749873464519839007959605662228349187716240969187969998 3
77	1.21400959177351261777749873464519839007959605662228349187716240969187969998 29
78	1.21400959177351261777749873464519839007959605662228349187716240969187969998 292
79	1.21400959177351261777749873464519839007959605662228349187716240969187969998 2919
80	1.21400959177351261777749873464519839007959605662228349187716240969187969998 29190

Table 2. Maple output

73	1.21400959177351261777749873464519839007959605662228349187716240969187977 4
74	1.21400959177351261777749873464519839007959605662228349187716240969187970 15
75	1.21400959177351261777749873464519839007959605662228349187716240969187969966
76	1.21400959177351261777749873464519839007959605662228349187716240969187970 0001
77	1.21400959177351261777749873464519839007959605662228349187716240969187969997 64
78	1.21400959177351261777749873464519839007959605662228349187716240969187969998 223
79	1.21400959177351261777749873464519839007959605662228349187716240969187969998 2930
80	1.21400959177351261777749873464519839007959605662228349187716240969187969998 29239

5.2 Negative a_n

Let us consider the function

$$f(x) = \frac{\exp(-x^2)}{2\sqrt{\pi}x(2x^2+1)\mathrm{erfc}(x)} - 1 = \sum_{n=1}^{\infty} \frac{\frac{-(2n+1)(2n+2)}{(2x^2+5+4n)(2x^2+1+4n)}}{1}$$

and $x = 2$. The partial numerators are negative and decrease to $\tilde{a} = -1/4$. We target $\epsilon_T \leq 2^{-79} \approx 1.65 \times 10^{-24}$ and use exact arithmetic for a change (hence $\epsilon_R = 0$).

The bound (5) is less than 2^{-79} 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_N satisfying

$$L_{N,j}\left(\frac{-1+\sqrt{4a_{N+2}+1}}{2}\right) < -0.37621 \leq w_N \leq -0.37527 < R_{N,j}(-1/2)$$

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

References

1. Colman, M., Cuyt, A.: Gauss and confluent hypergeometric functions accurate to the last digit. ACM TOMS (2006) (in preparation).
2. Cuyt, A., Becuwe, S.: Reliable software for the evaluation of several special functions. ACM TOMS (2006) (in preparation).
3. Cuyt, A., Verdonk, B.: Computer arithmetic: basic theory. SIAM, Philadelphia (2006) (in preparation).
4. Cuyt, A., Verdonk, B., Waadeland, H.: Efficient and reliable multiprecision implementation of elementary and special functions. SIAM Journal on Scientific Computing (2006) (to appear).
5. Gautschi, W.: Computational aspects of three-term recurrence relations. SIAM Rev. **9** (1987) 24–82
6. Jones, W.B., Thron, W.J.: Numerical stability in evaluating continued fractions. Math. Comp. **28** (1974) 795–810
7. Lorentzen, L., Waadeland, H.: Continued fractions with applications. North-Holland Publishing Company, Amsterdam (1992)
8. Thron, W.J., Waadeland, H.: Accelerating convergence of limit periodic continued fractions $K(a_n/1)$. Numer. Math. **34** (1980) 155–170