Towards a Verified Library for Special Functions

Annie Cuyt[*] [1], **Brigitte Verdonk**[**] [1], **H. Waadeland**[***][2], and **Johan Vervloet**[†][1]

[1] Dept Mathematics and Computer Science, University of Antwerp, Middelheimlaan 1, B2020 Antwerp, Belgium

[2] Dept Mathematics, NTNU, NO-7491 Trondheim, Norway

Received 28 February 2003, accepted 21 March 2003

The technique to provide a floating-point implementation of a function differs substantially when going from a fixed precision context to a 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, with respect to the model's complexity, the truncation error is as small as it can get. The total relative error should not exceed a prescribed threshold, round-off error and argument reduction effect included. In the latter, the goal is to provide a more generic technique, from which an approximant with the user-defined accuracy can be obtained at runtime. Hence best approximants are not an option, since these models would have to be recomputed every time the precision is altered and a function is evaluated. At the same time the generic technique should generate an approximant of as low complexity as possible.

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. 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 evaluation of the function.

In this work we outline, as far as space restrictions allow, the tools needed to achieve the reliable implementation of a number of elementary and special functions.

1 Tools

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. We shall especially be interested in real-valued limit-periodic continued fractions and their use in the reliable multiprecision implementation of the functions they represent. This implementation is built on top of multiprecision floating-point arithmetic compliant with the principles of the IEEE 754-854 floating-point standards.

1.1 IEEE-based Arithmetic

We assume we have available a multiprecision floating-point implementation of the basic operations, comparisons, base and type conversions, which is compliant with the principles of the IEEE 754-854 standards. Such an implementation is characterised by four parameters: the base β, the precision t and the exponent range $[L, U]$. In the current context, we are at least aiming at non-standard precisions $t > 64$ when $\beta = 2$.

To provide an implementation of a function $f(x)$ in a particular precision, one first needs to develop an efficient mathematical model or approximation $F(x)$ for $f(x)$. This is usually a very time-consuming effort, because the model changes whenever the precision does. The sum of the truncation

[*] e-mail: annie.cuyt@ua.ac.be

[**] e-mail: brigitte.verdonk@ua.ac.be

[***] e-mail:haakonwa@math.ntnu.no

[†] e-mail:johan.vervloet@ua.ac.be

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 $\mathtt{ulp}$ where

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

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 shall compare the number of basic operations, required in our approach when the precision is set to $t = 53$, to this reference.

1.2 Continued Fractions

We consider continued fraction representations of the form

$$f(x) = \sum_{n=1}^{\infty} \frac{a_n \rfloor}{\lceil 1} \qquad a_n := a_n(x) \tag{1}$$

Here a_n is called the n-th partial numerator. Especially useful are continued fractions of the form (1) where $a_n(x) = a_n x$ with $a_n > 0$. Such continued fractions are called S-fractions. The N-th approximant $f_N(w)$ of (1) and the N-th tail t_N of (1) are given by

$$f_N(w) = \sum_{n=1}^{N-1} \frac{a_n \rfloor}{\lceil 1} + \frac{a_N \rfloor}{\lceil 1 + w} \tag{2}$$

$$t_N = \sum_{n=N+1}^{\infty} \frac{a_n \rfloor}{\lceil 1} \tag{3}$$

A continued fraction is said to converge if $\lim_{N\to\infty} f_N(0)$ exists. Note that convergence to ∞ is allowed. The N-th approximant of a continued fraction can also be written as

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

1.3 Useful Tails

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

$$s_n(V_n) = \frac{a_n}{1+V_n} \subseteq V_{n-1} \qquad n = N, \ldots, 1$$

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 $f_N(w) \in V_0$. More importantly, when $\{V_n\}_{n\in\mathbb{N}}$ is a sequence of value sets for a convergent continued fraction, $t_N \in \overline{V}_N$ and hence $f(x) \in \overline{V}_0$ [3, p. 111]. When carefully monitoring the behaviour of the continued fraction tails, very accurate approximants $f_N(w)$ for $f(x)$ can be computed by making an appropriate choice for w.

We call a continued fraction (1) limit-periodic with period k, if

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

More can be said about tails of limit-periodic continued fractions with period one, also called one-limit-periodic continued fractions. Let $\tilde{a} = \lim_{n\to\infty} a_n$ and let $\tilde{w}$ be the fixpoint with smallest modulus of the linear fractional transformation $s(w) = \tilde{a}/(1+w)$. It can be shown [3] that

$$\lim_{N\to\infty} t_N = \tilde{w}$$

and also

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

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).

1.4 Oval Sequence Theorem

Besides the sequence of value sets, an equally important role is played by the sequence of convergence sets $\{E_n\}_{n\in\mathbb{N}}$, of which the elements guarantee convergence of the continued fraction as long as each partial numerator a_n belongs to the respective set E_n:

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

Very sharp truncation error estimates can be obtained from the oval sequence theorem [3]. Here we cite only the real version of this theorem.

Theorem 1.1 *Let* $0 < R_n < |1 + C_n|$ *and* $|C_{n-1}|R_n < |1 + C_n|R_{n-1}$. *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 $\{E_n\}_{n\in\mathbb{N}}$ *of convergence sets given by*

$$E_n = \{a \in \mathbb{R} : |a(1 + C_n) - C_{n-1}((1 + C_n)^2 - R_n^2)| + R_n|a| \leq R_{n-1}((1 + C_n)^2 - R_n^2)\}$$

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

$$|f(x) - f_N(w)| \leq 2R_N \frac{|C_0| + R_0}{|1 + C_N| - R_N} \times \prod_{k=1}^{N-1} M_k$$

where $M_k = \max\{|\frac{w}{1+w}| : w \in \overline{V}_k\}$

The oval E_n given above actually reduces to an interval $[p_n, q_n]$ in the real case. It is clear that the smaller the sets V_n, the smaller the values M_n and hence the smaller the upper bound on the truncation error $|f(x) - f_N(w)|$. A key role herein is played by the radii R_n.

2 Results

When combining the above ingredients with the characteristic monotonicity behaviour of the partial numerators in a lot of continued fraction representations of elementary and special functions, we obtain extremely sharp truncation error bounds. The monotonicity properties of the partial numerators indeed make it possible to give explicit expressions for the radii R_k and the maxima M_k in the oval sequence theorem, and this for several classes of continued fraction representations. The truncation error bounds obtained are almost indistinguishable from the true truncation error. Other truncation error bounds which can be found in the literature, either only hold for $w = 0$ [2], or are not equally sharp [1].

Since the accumulated rounding error is included in the total error which we bound by only a few `ulp`, the actual evaluation of $f(x)$ needs to take place in a slightly larger working precision $s > t$. The optimal working precision shall be determined dynamically, depending on the target precision t, the rounding error analysis and the required accuracy. Our rounding error analysis also takes the effect of argument reduction into account and hence guarantees a fully reliable evaluation of $f(x)$ over the entire domain.

References

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

[2] W.B. Jones and W.J. Thron. *Continued fractions: analytic theory and application*, volume 11 of *Encyclopedia of mathematics and its applications*. Addison-Wesley Publishing Company, New York, 1980.

[3] Lisa Lorentzen and Haakon Waadeland. *Continued fractions with applications*. North-Holland Publishing Company, Amsterdam, 1992.