HAL
open science

Multiple precision evaluation of the Airy Ai function with reduced cancellation

Sylvain Chevillard, Marc Mezzarobba

To cite this version:

Sylvain Chevillard, Marc Mezzarobba. Multiple precision evaluation of the Airy Ai function with reduced cancellation. 21st IEEE Symposium on Computer Arithmetic, 2013, Austin, TX, United States. pp.175-182, 10.1109/ARITH.2013.33 . hal-00767085v2

HAL Id: hal-00767085

https://inria.hal.science/hal-00767085v2

Submitted on 28 Apr 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multiple-precision evaluation of the Airy Ai function with reduced cancellation

Sylvain Chevillard
Inria, Apics Project-Team, Sophia Antipolis, France
sylvain.chevillard@inria.fr

Marc Mezzarobba
Inria, LIP (CNRS-ENS-Inria-UCBL), ENS de Lyon, France
marc@mezzarobba.net

***Abstract*—The series expansion at the origin of the Airy function $\mathrm{Ai}(x)$ is alternating and hence problematic to evaluate for $x > 0$ due to cancellation. Based on a method recently proposed by Gawronski, Müller, and Reinhard, we exhibit two functions F and G, both with nonnegative Taylor expansions at the origin, such that $\mathrm{Ai}(x) = G(x)/F(x)$. The sums are now well-conditioned, but the Taylor coefficients of G turn out to obey an ill-conditioned three-term recurrence. We use the classical Miller algorithm to overcome this issue. We bound all errors and our implementation allows an arbitrary and certified accuracy, that can be used, e.g., for providing correct rounding in arbitrary precision.**

***Index Terms*—Special functions; algorithm; numerical evaluation; arbitrary precision; Miller method; asymptotics; correct rounding; error bounds.**

Many mathematical functions (e.g., trigonometric functions, erf, Bessel functions) have a Taylor series of the form

$$y(x) = x^s \sum_{n=0}^{\infty} y_n\, x^{dn}, \quad y_n \sim (-1)^n \lambda \frac{\alpha^n}{n!^\kappa} \tag{1}$$

with $d, s \in \mathbb{Z}$ and $\alpha, \kappa > 0$. For large $x > 0$, the computation in finite precision arithmetic of such a sum is notoriously prone to *catastrophic cancellation*. Indeed, the terms $|y_n x^n|$ are first growing before the series "starts to converge" when $n^\kappa \geq \alpha x$. In particular, when $n^\kappa \approx \alpha x$, the terms $y_n x^n$ usually get much larger than $y(x)$. Eventually, their leading bits cancel out while lower-order bits that actually contribute to the first significant digits of the result get lost in roundoff errors.

This cancellation phenomenon makes the direct computation by Taylor series impractical for large values of x. Often, the function $y(x)$ admits an asymptotic expansion as $x \to +\infty$ that can be used very effectively to obtain numerical approximations when x is large, but might not provide enough accuracy (at least without resorting to sophisticated resummation methods) for intermediate values of x.

In the case of the error function $\mathrm{erf}(x)$, a classical trick going back at least to Stegun and Zucker [18] is to compute $\mathrm{erf}(x)$ as $G(x)/F(x)$ where $F(x) = e^{x^2}$ and [1, Eq. 7.6.2]

$$G(x) = e^{x^2}\,\mathrm{erf}(x) = \frac{2x}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{2^n}{1 \cdot 3 \cdots (2n+1)} x^{2n}. \tag{2}$$

The benefit of this transformation is that F and G are power series with nonnegative coefficients, and can thus be computed without cancellation. Algorithms based on (2) tend to behave well in some range $a < x < b$ where x is large enough for cancellation to be problematic but small enough to make the use of asymptotic expansions at infinity inconvenient. Note that the obvious way to compute $y(x) = e^{-x}$ for $x > 0$ fits into the same framework, now with $G(x) = 1$ and $F(x) = e^x$.

Gawronski, Müller and Reinhard [7], [14] provide elements to understand where these rewritings "come from". They relate the amount of cancellation in the summation of a series (1) to the shape of the Phragmén–Lindelöf indicator of y, a classical tool from the theory of entire functions [9]. This description allows them to state criteria for choosing auxiliary series suitable for the evaluation of a given entire function in a given sector of the complex plane. They apply their method (called the "GMR method" in what follows) to obtain "reduced cancellation" evaluation algorithms for the error function and other related functions in various sectors.

In this article, we are interested in the evaluation for positive x of the Airy function Ai [1, Chap. 9]. The function $\mathrm{Ai}(x)$ satisfies the linear ordinary differential equation (LODE)

$$\mathrm{Ai}''(x) - x\,\mathrm{Ai}(x) = 0 \tag{3}$$

with initial values

$$\mathrm{Ai}(0) = A := 3^{-2/3}\Gamma(\tfrac{2}{3})^{-1}, \quad \mathrm{Ai}'(0) = -B := -3^{-1/3}\Gamma(\tfrac{1}{3})^{-1}.$$

The classical existence theorem for LODE with complex analytic coefficients implies that $\mathrm{Ai}(x)$ is an entire function; and solving (3) by the method of power series yields the Taylor expansion $\mathrm{Ai}(x) = Af(x^3) - Bxg(x^3)$, where

$$f(x) = \sum_{n=0}^{\infty} \frac{1 \cdot 4 \cdots (3n-2)}{(3n)!} x^n, \quad g(x) = \sum_{n=0}^{\infty} \frac{2 \cdot 5 \cdots (3n-1)}{(3n+1)!} x^n.$$

Observe that while f and g are easy enough to evaluate individually, the difference $Af(x^3) - Bxg(x^3)$ causes catastrophic cancellation when computed in approximate arithmetic.

Using the GMR method, we derive a reduced cancellation algorithm for computing $\mathrm{Ai}(x)$. To our best knowledge, our algorithm for evaluating $\mathrm{Ai}(x)$ is new, and is the most efficient multiple-precision evaluation of $\mathrm{Ai}(x)$ when x is neither too small nor too large, while the precision is not large enough to make methods based on binary splitting [3] competitive.

Besides the new application, the main difference between the present article and the work of Gawronski *et al.* is our setting of multiple-precision arithmetic "à la MPFR [6]". Specifically, on the one hand, we are interested in *arbitrary precision* arithmetic rather than machine precision only. This makes it impossible, for instance, to tabulate the coefficients of auxiliary functions when these turn out to be hard to compute. Also, we are looking for *rigorous error bounds* instead of experimental

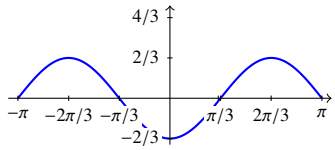

Figure 1. The indicator function h of Ai.

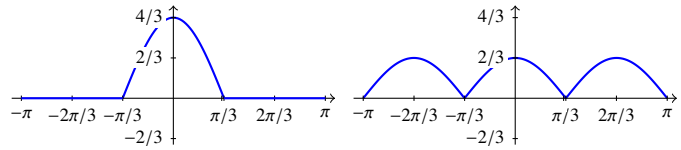

Figure 2. Indicator functions for F (left) and G (right).

error estimates. On the other hand, we restrict ourselves to numerical evaluation on a half line instead of a complex sector (though, in principle, the basic ideas generalize).

This article focuses on providing a complete algorithm in the specific case of $\mathrm{Ai}(x)$, $x > 0$. Yet, it should also be seen a *case study*, part of an effort to understand what the GMR method can bring in the context of multiple-precision computation, and, perhaps more importantly, how general and systematic it can be made. We discuss this last point further in Sec. VIII.

The rest of this text is organized as follows. In Sec. I, we use the GMR method to choose the functions F and G. Then, in Sec. II and III, we derive a few mathematical properties of these functions, including recurrences for their series expansions and various bounds. Sections IV to VI contain the details of our algorithm and its error analysis. Finally, in Sec. VII, we briefly describe our implementation of the algorithm.

I. The GMR method

We now review the GMR method and apply it to obtain candidate auxiliary series for the evaluation of the Ai function. Since the method itself is not crucial for our results, we summarize it in intuitive terms and refer the reader to the original works [7], [14] for more careful statements.

The starting point of the GMR method is the following observation. Let $y(z) = \sum_{n\geq 0} y_n z^n$ be an entire function. Assume that we have, in some intentionally vague sense,

$$y(re^{i\theta}) \approx \exp(h(\theta) r^{\rho}) \tag{4}$$

for large r. (To make things precise, we would assume that y has finite *order* ρ, and that h is its *indicator function* with respect to ρ [9].)

We consider the computation of $y(z)$ in floating-point arithmetic using its series expansion. It is well-known [4] that, if the sum is performed in floating-point arithmetic of precision t, the relative error between $y(z)$ and the computed sum is roughly given by $2^{-t}\,(\sum_{n\geq 0} |y_n z^n|)/|y(z)|$. The sum $\sum_{n\geq 0} |y_n z^n|$ is larger than $\max_{n\geq 0} |y_n z^n|$, and usually of the same order of magnitude. Therefore, the number of significant binary digits "lost by cancellation" is roughly

$$\log_2(\max_{n\geq 0} |y_n z^n|) - \log_2 |y(z)|.$$

Denote $M(r) = \sup_{|z|=r} |y(z)|$ for all $r > 0$. Cauchy's formula implies $\max_n(|y_n| r^n) \leq M(r)$, and under a suitable version of hypothesis (4), one can actually show that $\max_n(|y_n| r^n) \approx M(r)$. Hence, the loss of precision by cancellation in the evaluation of $y(re^{i\theta})$ is about

$$\log_2 \frac{M(r)}{|y(re^{i\theta})|} \approx [(\max h) - h(\theta)]\, r^{\rho}.$$

For instance, when the y_n all have the same complex argument, the maximum of h is reached for $\theta = 0$, in accordance with the fact that the sum is optimally conditioned.

In the case of the Airy Ai function, the following asymptotic equivalent holds as z tends to complex infinity in any open sector that avoids the negative real axis [1, Eq. 9.7.5]:

$$\mathrm{Ai}(z) \sim \widetilde{\mathrm{Ai}}(z) := \frac{\exp(-\frac{2}{3} z^{3/2})}{2\sqrt{\pi} z^{1/4}}. \tag{5}$$

Additionally, $\mathrm{Ai}(x)$ is bounded for $x < 0$. Hence, we may take $\rho = 3/2$ and

$$h(\theta) = -\tfrac{2}{3}\cos(\tfrac{3}{2}\theta), \qquad -\pi < \theta \leq \pi \tag{6}$$

(see Figure 1). The loss of precision is roughly proportional to $1 + \cos(\frac{3}{2}\theta)$. It is minimal in the directions of fastest growth $\theta = \pm\frac{2}{3}\pi$, and maximal for $\theta = 0$.

If now two entire functions y and F both satisfy conditions of the form (4) with the same ρ but different h (say h_y and h_F, respectively), we may expect that

$$G(z) = F(z)y(z) \approx \exp([h_y(\theta) + h_F(\theta)] r^{\rho}). \tag{7}$$

The GMR method consists in reducing the summation of the series y for z in some given sector to that of an *auxiliary series* $F(z)$ and a *modified series* $G(z)$ related by (7). The value of $y(z)$ is then recovered as $G(z)/F(z)$. The auxiliary series is chosen, based on the shape of h_y, so that both h_F and $h_G = h_y + h_F$ take values close to their maximum in the sector of interest.

There may be multiple choices, and it is not clear in general which one is better, except that the coefficients of F and G should be as easy to compute as possible. Gawronski *et al.* usually take $F(z) = \exp(az^{\rho})$ and search for a value of a that makes $(\max h_G) - h_G$ as small as possible on a whole complex sector. The choice of exponentials as auxiliary series is not appropriate in the case of Ai, since $\exp(z^{\rho})$ is an entire function of z only for integer ρ.

However, as we are interested in one direction only, we can easily build a suitable auxiliary series from Ai itself. Indeed, we may "shift" the indicator function of Ai by $2\pi/3$ to the left or to the right by changing z to $j^{\pm 1} z$, where $j = e^{2\pi i/3}$. (Note that this is *not* the same as changing θ to $\theta \pm \frac{2}{3}\pi$ in (6).) When we add such a shifted indicator to the original h_{Ai}, one of the humps of the curve cancels out with the valley in the middle.

Using this idea, we set

$$F(x) = \mathrm{Ai}(jx)\,\mathrm{Ai}(j^{-1}x), \quad G(x) = F(x)\,\mathrm{Ai}(x). \tag{8}$$

The indicator functions of F and G are pictured on Figure 2. Based on their shapes, we expect that both series are optimally conditioned on the positive real axis. We shall prove that this is indeed the case in the next two sections.

This second method of constructing auxiliary series seems to be new, and applies to many cases. For instance, applying it to the error function leads to

$$F(x) = -i\,\mathrm{erf}\, ix = \frac{2x}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{1\cdot 3\cdots(2n-1)}{(2n+1)!} 2^n x^{2n}, \tag{9}$$

$$G(x) = F(x)\operatorname{erf} x = \frac{8x^2}{\pi}\sum_{n=0}^{\infty}\frac{2\cdot 4\cdots(4n)}{(4n+2)!}4^n x^{4n}, \qquad (10)$$

a slightly worse alternative to (2). The advantage of (2) comes from the fact that e^{x^2} is faster to evaluate than (9).

II. The auxiliary series F

The functions F and G being chosen, we need to establish appropriate formulas to evaluate them, along with error bounds. Much of our analysis will be based on the following simple estimate [12, Chap. 4, §4.1], where $\widetilde{\mathrm{Ai}}$ was defined in Eq. (5).

Lemma 1: The Airy function Ai satisfies

$$|\mathrm{Ai}(re^{i\theta})/\widetilde{\mathrm{Ai}}(re^{i\theta}) - 1| \le \eta_1(\theta) r^{-3/2}, \qquad |\theta| < \pi,$$

where $\eta_1(\theta) = \frac{5}{48}(\cos\frac{\theta}{2})^{-7/2}$.

Now consider the power series expansion of F at the origin:

$$F(x) = \mathrm{Ai}(jx)\,\mathrm{Ai}(j^{-1}x) = \sum_{n=0}^{\infty} F_n x^n. \qquad (11)$$

Proposition 1: The coefficients F_n are positive and satisfy the two-term recurrence relation

$$(n+1)(n+2)(n+3)F_{n+3} - 2(2n+1)F_n = 0 \qquad (12)$$

with initial values

$$F_0 = 3^{-4/3}\Gamma(\tfrac{2}{3})^{-2}, \qquad F_1 = (2\sqrt{3}\pi)^{-1}, \qquad F_2 = 3^{-2/3}\Gamma(\tfrac{1}{3})^{-2}.$$

Proof: As a general fact, if two functions w and y each satisfy a homogeneous LODE with coefficients in $\mathbb{Q}(x)$, then their product wy satisfies an equation of the same class that can be explicitly computed [17, Sec. 6.4]. The functions $\mathrm{Ai}(j^{\pm 1}x)$ satisfy the same differential equation (3) as $\mathrm{Ai}(x)$ itself. Applying the procedure mentioned above to two copies of that equation yields $F^{(3)}(x) - 4xF'(x) - 2F(x) = 0$.

Similarly, when an analytic function y satisfies a homogeneous LODE over $\mathbb{Q}(x)$, we can compute a recurrence relation with coefficients in $\mathbb{Q}(n)$ on the coefficients y_n of its power series expansion. In the case of F, we get (12). Finally, we compute the initial values F_1, F_2, F_3 from the first few terms of the Taylor expansion of $\mathrm{Ai}(x)$:

$$\begin{aligned} F(x) &= (A - Bjx + O(x^3))(A - Bj^{-1}x + O(x^3)) \\ &= A^2 - ABx + B^2x^2 + O(x^3). \end{aligned}$$

It is then apparent from (12) that $F_n > 0$ for all $n \in \mathbb{N}$. ■

Thus, the coefficients of $F(x)$ obey a two-term recurrence whose coefficients do not vanish for $n \ge 0$. This allows one to compute them in a numerically stable way (see Sec. VI).

III. The modified series G

Recall that $G(x) = \mathrm{Ai}(x)\,\mathrm{Ai}(jx)\,\mathrm{Ai}(j^{-1}x)$, and set

$$\tilde{G}(x) = \widetilde{\mathrm{Ai}}(x)\widetilde{\mathrm{Ai}}(jx)\widetilde{\mathrm{Ai}}(j^{-1}x) = \frac{\exp(\frac{2}{3}x^{3/2})}{8\pi^{3/2}x^{3/4}}. \qquad (13)$$

Proposition 2: The function G is an entire function with power series expansion of the form $G(z) = \sum_{n=0}^{\infty} G_n z^{3n}$. The coefficient sequence $(G_n)_{n\in\mathbb{N}}$ is determined from its first terms

$$G_0 = A^3 = \frac{1}{9\Gamma(2/3)^3}, \quad G_1 = \frac{A^3}{2} - B^3 = \frac{1}{18\Gamma(2/3)^3} - \frac{1}{3\Gamma(1/3)^3}$$

by the recurrence relation

$$(n+1)(3n+4)(3n+5)(n+2)G_{n+2} - 10(n+1)^2 G_{n+1} + G_n = 0. \quad (14)$$

Proof: First, observe that $G(jz) = G(z)$ and $G(\bar{z}) = \overline{G(z)}$, so that the Taylor expansion of G at the origin is a power series in z^3 with real coefficients. The same routine reasoning as in the proof of Prop. 1 yields the LODE

$$G^{(4)}(x) - 10xG''(x) - 10G'(x) + 9x^2G(x) = 0,$$

and from there the recurrence (14). The coefficients of (14) do not vanish for $n \ge 0$, so that the sequence G_n is indeed determined by G_0, G_1, and (14). ■

Nonzero solutions of (14) decrease roughly as $n!^{-2}$ for large n. Setting $c_n = n!^2 G_n$ yields the "normalized" recurrence

$$\frac{(3n+4)(3n+5)}{(n+1)(n+2)}c_{n+2} - 10c_{n+1} + c_n = 0. \qquad (15)$$

Letting n go to infinity in the coefficients of (15), we get a limit recurrence with constant coefficients whose characteristic polynomial $9\alpha^2 - 10\alpha + 1$ has two roots of distinct absolute value, namely $\frac{1}{9}$ and 1. By the Perron–Kreuser theorem [21, Theorem B.10], it follows that any solution $(v_n)_{n\in\mathbb{N}}$ of (14) satisfies $v_{n+1}/v_n \sim \alpha n^{-2}$ with either $\alpha = 1$ or $\alpha = \frac{1}{9}$. Solutions (v_n) such that $v_{n+1}/v_n \sim \frac{1}{9n^2}$ are called *minimal* and form a linear subspace of dimension 1 of the solutions of (14).

We shall prove that (G_n) actually is such a minimal solution of (14). But our analysis uses a bit more than the rough estimate $G_n \approx n!^{-2}9^{-n}$. Prop. 3 below provides a more precise estimate which implies the minimality. Before turning to it, we recall a standard bound on the tails of incomplete Gaussian integrals [1, Eq. 7.12.1] and state a second technical lemma.

Lemma 2: The complementary error function $\operatorname{erfc} x = 1 - \operatorname{erf} x = \frac{2}{\sqrt{\pi}}\int_x^{\infty} e^{-t^2}\mathrm{d}t$ satisfies $0 \le \operatorname{erfc} x \le \frac{1}{\sqrt{\pi}x}e^{-x^2}$ for $x > 0$.

Lemma 3: The expression

$$I(r) = r^{9/8}\int_{r^{-5/8}}^{\pi/3} e^{\frac{2}{3}r^{3/2}(-1+\cos\frac{3\theta}{2})}\mathrm{d}\theta$$

satisfies $0 \le I(r) \le 0.51$ for all $r \ge 10$.

Proof: We first use the inequality $\cos(\frac{3}{2}\theta) \le 1 - \frac{9}{10}\theta^2$ (valid for $0 \le \theta \le \pi/3$) followed by the change of variable $\varphi = \sqrt{3/5}\, r^{3/4}\,\theta$ to get

$$I(r) \le r^{9/8}\int_{r^{-5/8}}^{\pi/3} e^{-\frac{3}{5}r^{3/2}\theta^2}\mathrm{d}\theta = \frac{\sqrt{5}}{\sqrt{3}}\, r^{3/8}\int_{a(r)}^{b(r)} e^{-\varphi^2}\mathrm{d}\varphi,$$

where $a(r) = \sqrt{3/5}\, r^{1/8}$ and $b(r) = \frac{\pi}{\sqrt{15}}r^{3/4}$. Now we have

$$\int_{a(r)}^{b(r)} e^{-\varphi^2}\mathrm{d}\varphi \le \int_{a(r)}^{\infty} e^{-\varphi^2}\mathrm{d}\varphi = \frac{\sqrt{\pi}}{2}\operatorname{erfc}\left(\frac{\sqrt{3}r^{1/8}}{\sqrt{5}}\right),$$

and, by Lemma 2, $I(r) \le \frac{5}{6}r^{1/4}e^{-\frac{3}{5}r^{1/4}}$. When $r \ge 10$, the last bound is decreasing and hence less than 0.51. ■

The following proposition is the main result of this section and the starting point of much of the error analysis that follows. Though somewhat technical, the proof is mostly routine.

Proposition 3: The sequence (G_n) satisfies

$$G_n \sim \gamma_n = \frac{1}{4\sqrt{3}\pi 9^n n!^2}, \quad n \to \infty,$$

with relative error $|G_n/\gamma_n - 1| \le 2.2n^{-1/4}$ $\quad (n \ge 1)$.

Notation 1: We write $E(\varepsilon_1, \ldots, \varepsilon_n) = \sum_{\varnothing \ne I \subset [\![1,n]\!]} \prod_{i\in I}\varepsilon_i$, so that, if $|\theta_i| \le \varepsilon_i$ for all i, then $\prod_{i=1}^n(1+\theta_i) = 1 + \theta$ with $|\theta| \le E(\varepsilon_1, \ldots, \varepsilon_n)$. Note that we obviously have $E(\alpha\,\varepsilon_1, \ldots, \alpha\,\varepsilon_n) \le \alpha\, E(\varepsilon_1, \ldots, \varepsilon_n)$ for $0 \le \alpha \le 1$.

Proof: The proof is a standard application of the *saddle point method* [5, §VIII.3], which we work out in some detail in order to get an explicit error bound.

We fix $n > 10$. We shall write $z = re^{i\theta}$ in the following. The method prescribes to choose $r = (3n + 3/4)^{2/3}$, so that

$$\frac{\tilde{G}(z)}{z^{3n}} = \frac{\tilde{G}(r)}{r^{3n}} e^{O((z-r)^2)} \tag{16}$$

(i.e., $\frac{\mathrm{d}}{\mathrm{d}z} \log(\tilde{G}(z) z^{-3n})_{|z=r} = 0$).

Now, guided by (16) and $G(z) \approx \tilde{G}(z)$, we write

$$G_n = \frac{1}{2\pi i} \oint \frac{G(z)}{z^{3n+1}} \mathrm{d}z = \frac{3}{2\pi} \frac{\tilde{G}(r)}{r^{3n}} I_1, \quad I_1 = \int_{-\pi/3}^{\pi/3} \frac{G(re^{i\theta})}{\tilde{G}(r)} e^{-3ni\theta} \mathrm{d}\theta.$$

Most of the weight of I_1 is concentrated around 0. We set

$$\theta_0 = r^{-5/8} = (3n + 3/4)^{-5/12}. \tag{17}$$

Since $n > 10$, we remark that $r \geq 10$ and $0 < \theta_0 < 1/4 < \pi/3$. We further let

$$I_2 = \int_{-\theta_0}^{\theta_0} \frac{G(re^{i\theta})}{\tilde{G}(r)} e^{-3ni\theta} \mathrm{d}\theta.$$

We first bound the error between I_1 and I_2. When $\theta \in [0, \pi/3]$, using Lemma 1 and the connection formula [1, Eq. 9.2.12]

$$\mathrm{Ai}(z) + j\,\mathrm{Ai}(jz) + j^{-1}\,\mathrm{Ai}(j^{-1}z) = 0,$$

we see that $|G(z)/\tilde{G}(z)|$ is bounded by

$$\begin{aligned}
&\left|\frac{\mathrm{Ai}(z)}{\widetilde{\mathrm{Ai}}(z)}\right| \left|\frac{\mathrm{Ai}(j^{-1}z)}{\widetilde{\mathrm{Ai}}(j^{-1}z)}\right| \left(\left|\frac{\widetilde{\mathrm{Ai}}(z)}{\widetilde{\mathrm{Ai}}(jz)}\right| \left|\frac{\mathrm{Ai}(z)}{\widetilde{\mathrm{Ai}}(z)}\right| + \left|\frac{\widetilde{\mathrm{Ai}}(j^{-1}z)}{\widetilde{\mathrm{Ai}}(jz)}\right| \left|\frac{\mathrm{Ai}(j^{-1}z)}{\widetilde{\mathrm{Ai}}(j^{-1}z)}\right| \right) \\
&\leq (1 + \eta_1(\theta) r^{-3/2})(1 + \eta_1(\theta - \tfrac{2\pi}{3}) r^{-3/2}) \\
&\quad \left(e^{-\frac{4}{3}\cos\frac{3\theta}{2}} (1 + \eta_1(\theta) r^{-\frac{3}{2}}) + 1 + \eta_1(\theta - \tfrac{2\pi}{3}) r^{-\frac{3}{2}} \right).
\end{aligned}$$

It follows that $|G(z)| \leq 2.1 |\tilde{G}(z)|$ when $|z| \geq 10$. By symmetry, this inequality holds for $-\pi/3 \leq \theta \leq 0$ too. Finally, we have

$$\begin{aligned}
|I_1 - I_2| &\leq 2 \int_{\theta_0}^{\pi/3} \frac{|G(re^{i\theta})|}{\tilde{G}(r)} \mathrm{d}\theta \leq 4.2 \int_{\theta_0}^{\pi/3} \frac{|\tilde{G}(re^{i\theta})|}{\tilde{G}(r)} \mathrm{d}\theta \\
&= 4.2 \int_{\theta_0}^{\pi/3} e^{\frac{2}{3} r^{3/2} (-1 + \cos\frac{3\theta}{2})} \mathrm{d}\theta
\end{aligned}$$

and hence, using Lemma 3,

$$|I_1 - I_2| \leq 2.15 r^{-9/8} \tag{18}$$

We now have to estimate I_2. We write

$$I_2 = \int_{-\theta_0}^{\theta_0} \frac{G(re^{i\theta})}{\tilde{G}(re^{i\theta})} \cdot \frac{\tilde{G}(re^{i\theta})}{\tilde{G}(r)} e^{-3ni\theta} \mathrm{d}\theta.$$

On the one hand, Lemma 1 gives $G(z) = \tilde{G}(z)(1 + \delta(\theta))$ with $|\delta(\theta)| \leq \eta_2(\theta) r^{-3/2}$ where $\eta_2(\theta) = E(\eta_1(\theta), \eta_1(\theta + \frac{2\pi}{3}), \eta_1(\theta - \frac{2\pi}{3}))$. Note that this bound increases with $|\theta|$.

On the other hand, thanks to (16), we can write

$$\tilde{G}(re^{i\theta}) e^{-3ni\theta} = \tilde{G}(r) e^{\frac{2}{3} r^{3/2} (\exp(\frac{3}{2} i\theta) - 1 - \frac{3}{2} i\theta)} = \tilde{G}(r) e^{(-\frac{3}{4}\theta^2 + u(\theta)) r^{3/2}}$$

where

$$|u(\theta)| = \tfrac{2}{3} \left| e^{\frac{3}{2} i\theta} - 1 - \tfrac{3}{2} i\theta + \tfrac{9}{8}\theta^2 \right| \leq \tfrac{3}{8} |\theta|^3. \tag{19}$$

Let

$$v(\theta) = \frac{G(re^{i\theta}) e^{-3in\theta}}{\tilde{G}(r) \exp(-\frac{3}{4} r^{3/2} \theta^2)} - 1.$$

Since $\theta_0^3 r^{3/2} = r^{-3/8}$ by the choice (17), using (19) and the inequality $|e^z - 1| \leq |z| e^{|z|}$, we have, for $|\theta| \leq \theta_0$ and $r \geq 10$:

$$|v(\theta)| \leq \frac{3}{8} r^{-3/8} e^{\frac{3}{8} r^{-3/8}} \leq 0.44 r^{-3/8}.$$

To sum up, we can now rewrite I_2 as

$$I_2 = \int_{-\theta_0}^{\theta_0} (1 + \delta(\theta))(1 + v(\theta)) \exp\left(-\frac{3}{4} r^{3/2} \theta^2\right) \mathrm{d}\theta.$$

Now, define

$$I_3 = \int_{-\theta_0}^{\theta_0} \exp\left(-\frac{3}{4} r^{3/2} \theta^2\right) \mathrm{d}\theta, \qquad I_4 = \frac{2\sqrt{\pi}}{\sqrt{3}} r^{-3/4}.$$

When $\theta \in [-\theta_0, \theta_0]$, we have

$$|(1 + \delta(\theta))(1 + v(\theta)) - 1| \leq E(\eta_2(\theta_0) r^{-3/2}, 0.44 r^{-3/8}) \leq 0.95 r^{-3/8},$$

so we get

$$|I_2 - I_3| \leq 0.95 r^{-3/8} I_3. \tag{20}$$

Finally, I_3 is an incomplete Gaussian integral:

$$I_3 = \frac{2\sqrt{\pi}}{\sqrt{3} r^{3/4}} \operatorname{erf}(\tfrac{1}{2}\sqrt{3} r^{3/4} \theta_0) = I_4 \operatorname{erf}(\tfrac{1}{2}\sqrt{3} r^{1/8}).$$

Lemma 2 yields

$$|I_3/I_4 - 1| \leq \frac{2 r^{-1/8} e^{-\frac{3}{4} r^{1/4}}}{\sqrt{3\pi}} \leq 0.66 r^{-1/8} e^{-\frac{3}{4} r^{1/4}}. \tag{21}$$

Putting together (18, 20, 21), we obtain $I_1 = I_4(1 + \epsilon)$ with

$$\begin{aligned}
|\epsilon| &\leq \frac{|I_1 - I_2|}{I_4} + \left| \frac{I_2 I_3}{I_3 I_4} - 1 \right| \\
&\leq 1.06 r^{-3/8} + E(0.95 r^{-3/8}, 0.66 r^{-1/8} e^{-\frac{3}{4} r^{1/4}}) \\
&\leq 2.45 r^{-3/8} \leq 1.9 n^{-1/4}.
\end{aligned}$$

Now, we can write $G_n = \frac{3}{2\pi} \cdot \frac{\tilde{G}(r)}{r^{3n}} I_4 (1 + \epsilon) = \gamma_n H (1 + \epsilon)$ with

$$H := \frac{\sqrt{3}}{\sqrt{\pi} r^{3/4} \gamma_n} \frac{\tilde{G}(r)}{r^{3n}} = \frac{n!^2 e^{2n}}{n^{2n} 2\pi n} \frac{\sqrt{e}}{(1 + \frac{1}{4n})^{2n+1}}.$$

Stirling's formula in the form [15] $\left| \frac{n! e^n}{n^n \sqrt{2\pi n}} - 1 \right| \leq s(n) = e^{\frac{1}{12n}} - 1$, combined with the bound $\left| \frac{\sqrt{e}}{(1 + 1/(4n))^{2n}} - 1 \right| \leq \frac{1}{16n}$, yields

$$|H - 1| \leq E\left(s(n), s(n), \tfrac{1}{16n}, \tfrac{1}{4n}\right) \leq \frac{1}{2n}, \qquad n \geq 11. \tag{22}$$

It follows that

$$\left| \frac{G_n}{\gamma_n} - 1 \right| \leq E(0.5 n^{-1}, 1.9 n^{-1/4}) \leq 2.2 n^{-1/4}, \qquad n \geq 11,$$

One easily checks that this bound is valid for $1 \leq n \leq 10$. ■

Note that the above bound is not the best we can get by this method. Indeed, by choosing the exponent of r in (17) closer to $-3/4$, we obtain $|G_n/\gamma_n - 1| = O(n^{-1/2+\epsilon})$ for any $\epsilon > 0$. This comes at the price of a larger constant factor and thus more terms to check separately.

A first consequence of Prop. 3 is that the series $G(x)$ has nonnegative coefficients, as stated below. We also deduce several other technical results that will be used to bound various error terms. Recall that $c_n = n!^2 G_n$, and let $\tau = 3/20$.

Corollary 1: The sequence (c_n) satisfies $0 \leq c_{n+1}/c_n \leq \tau$ for all $n \geq 0$. Accordingly, we have $0 < G_{n+1}/G_n \leq \tau(n + 1)^{-2}$. In particular, the G_n are positive.

Proof: Prop. 3 implies that $c_n = (1 + \theta_n)/(4\sqrt{3}\pi 9^n)$ with $|\theta_n| \leq 2.4 n^{-1/4}$. When $n \geq n_0 = 47610$, we have $|\theta_n| \leq 0.148936$. This implies $c_n \geq 0$ and $c_{n+1}/c_n \leq (1/9)(1 + \theta_{n+1})/(1 + \theta_n) \leq \tau$. The inequalities up to $n = n_0$ are checked by computing the corresponding terms with interval arithmetic. ■

Corollary 2: For any $n \geq 1$, it holds that $G_n \leq (e/(3n))^{2n}$.

Proof: Follows from Proposition 3 using $n! \geq (n/e)^n$. ■

Lemma 4: When $x > 0$ and $N + 1 \geq \sqrt{2\tau}x^{3/2}$, we have

$$\sum_{n=0}^{N} G_n x^{3n} \leq G(x) \leq \sum_{n=0}^{N-1} G_n x^{3n} + 2G_N x^{3N}.$$

Proof: The first inequality is obvious as $G_n \geq 0$. To show the second one, observe that a fortiori $\tau x^3/(n+1)^2 \leq \frac{1}{2}$ for all $n \geq N$. Using Corollary 1, we deduce $G_{n+1}x^3/G_n \leq \frac{1}{2}$, and hence $\sum_{n\geq N} G_n x^{3n} \leq G_N x^{3N}(1 + \frac{1}{2} + \frac{1}{4} + \cdots) = 2G_N x^{3N}$. ■

Lemma 5: The function G satisfies

$$0.01 e^{\frac{2}{3}x^{3/2}} x^{-3/4} \leq G(x) \leq 0.04 e^{\frac{2}{3}x^{3/2}} x^{-3/4}$$

for all $x \geq 1/2$.

Proof: Let $\mu(x) = E(\eta_1(0)x^{-3/2}, \sigma x^{-3/2}, \sigma x^{-3/2})$ where $\sigma = \eta_1(\frac{2\pi}{3}) = \frac{5}{6}\sqrt{2}$. Lemma 1 implies $|G(x)/\tilde{G}(x) - 1| \leq \mu(x)$ for $x \geq 0$. We can check that $\mu(x)$ is a decreasing function and

$$0.01 \leq \frac{1 - \mu(3)}{8\pi^{3/2}}, \qquad \frac{1 + \mu(3)}{8\pi^{3/2}} \leq 0.04,$$

whence the desired inequality for $x \geq 3$. For $0.5 < x \leq 3$, using the bounds $0 \leq e^x - (1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{24}x^4) \leq \frac{5}{48}x^4$ (valid for $0 < x \leq \frac{2}{3}3^{3/2}$) and Lemma 4, we are reduced to checking explicit polynomial inequalities. ■

IV. Roundoff error analysis

We now turn to the floating-point implementation of the functions $F(x)$ and $G(x)$. To make the algorithm rigorous, we will use classical techniques of error analysis that we briefly recall here. We refer the reader, e.g., to Higham [8], for proofs and complement of information.

We suppose that the precision of the floating-point format is t bits and that the exponent range is unbounded (in case it is bounded, it would probably be possible to rescale $F(x)$ and $G(x)$ by the same factor, to make them representable without changing the ratio $F(x)/G(x)$).

Notation 2: For $x \neq 0$, we denote $\mathrm{Exp}(x) = \lfloor \log_2 |x| \rfloor + 1$, so that $2^{\mathrm{Exp}(x)-1} \leq |x| < 2^{\mathrm{Exp}(x)}$.

Notation 3: If $x \in \mathbb{R}$, $\circ(x)$ denotes the floating-point number closest to x (ties can be decided either way). Circled operators such as $\oplus$ denote correctly rounded floating-point operations.

We always have $\circ(x) = x(1+\delta)$ and $x = \circ(x)(1+\delta')$ with $|\delta|, |\delta'| \leq 2^{-t}$. We will also extensively use the *relative error counter* notation $z\langle k \rangle$.

Notation 4: We write $\widehat{z} = z\langle k \rangle$ when there exist $\delta_1, \ldots, \delta_k$ such that $\widehat{z} = z\prod_{i=1}^{k}(1+\delta_i)^{\pm 1}$ with $|\delta_i| \leq 2^{-t}$ for all i.

Roughly speaking, each arithmetical operation adds one to the relative error counter of a variable. The overall error corresponding to an error counter can be bounded as follows.

Proposition 4: Suppose that we can write $\widehat{z} = z\langle k \rangle$ and that $k\,2^{-t} \leq 1/2$. Then $\widehat{z} = z(1+\theta)$ with $|\theta| \leq 2k \cdot 2^{-t}$.

V. Evaluation of the modified series

As we shall see in the next section, evaluating the auxiliary function F is fairly straightforward. The evaluation of G is more involved. Indeed, while $\sum G_n x^{3n}$ is well-conditioned as a sum for $x \geq 0$ (this is the whole point of the GMR method), the minimality of the *sequence* (G_n) among the solutions of (14) implies that its direct recursive computation from the initial values G_0 and G_1 is numerically unstable (cf. [21]).

Algorithm 1: Evaluation of G

Input: a target precision $p \geq 1$, a point $x \geq 0.5$
Output: s such that $|G(x) - s| \leq 3 \cdot 2^{-p}\, G(x)$

1 Choose $\alpha, \beta, \delta, \gamma$ s.t. $\alpha \lesssim 3\,e^{-1}x^{-3/2}$, $\beta \lesssim (2/3)\,\log_2(e)\,x^{3/2}$, $\gamma \gtrsim 1/\log_2(20/3)$, $\delta \gtrsim (2/3)\,\log_2(e)\,((20/3)^{1/2} - 1)\,x^{3/2}$;
2 $N_0 \leftarrow \max(1, \lceil (3/10)^{1/2}\, x^{3/2} - 1 \rceil)$;
3 Choose $N \geq N_0$ s.t. $\mathrm{Exp}((\alpha N)^{2N}) \geq p + 9 + \frac{3}{4}\,\mathrm{Exp}(x) - \lfloor \beta \rfloor$;
4 Choose $R \geq \max(N, (p + 2 + \delta)\,\gamma)$;
5 Choose t s.t. $128\,(N+3)\,2^{-t} \leq 2^{-p}$ and $(R+2)\,2^{-t} \leq 2^{-9}$;
6 **for** $(a \leftarrow 1, b \leftarrow 0, i \leftarrow R-1; i \geq 0; i \leftarrow i-1)$ **do**
7 $\quad c \leftarrow a$;
8 $\quad a \leftarrow (10 \otimes a) \ominus (3i+4)(3i+5) \otimes b \oslash ((i+1)(i+2))$;
9 $\quad b \leftarrow c$;
10 $\quad$ **if** $i = N-1$ **then** $s' \leftarrow a$;
11 $\quad$ **else if** $i < N-1$ **then** $s' \leftarrow a \oplus (x^3 \otimes s' \oslash (i+1)^2)$;
12 **return** $s = (s' \oslash a) \oslash \circ(9\,\Gamma(2/3)^3)$;

There is a standard tool to handle this situation, namely Miller's backward recurrence method [2], [21]. Miller's method allows one to accurately evaluate the minimal solution (c_n) of a recurrence of the form

$$a_2(n)\,u_{n+2} + a_1(n)\,u_{n+1} + a_0(n)\,u_n = 0, \quad a_0(n)a_2(n) \neq 0. \quad (23)$$

The idea is as follows: choose a starting index R and let (arbitrarily) $u_{R+1} = 0$ and $u_R = 1$. Then compute u_n as

$$-a_0(n)^{-1}\,(a_1(n)\,u_{n+1} + a_2(n)\,u_{n+2}), \quad n = R-1, \ldots, 1, 0.$$

It turns out that, for large R, the computed sequence (u_n) is close to a minimal solution of the forward recurrence. Since all minimal solutions are proportional to each other, we recover an approximation of c_n as $c_n \approx (c_0/u_0)\,u_n$.

We use Miller's method to evaluate the minimal solution (c_n) of the normalized recurrence (15), and we get an approximation of $G(x) = \sum_{n\geq 0} G_n\, x^{3n}$ as $S_N = (c_0/u_0)\,\sum_{n=0}^{N-1} u_n\, x^{3n}/(n!^2)$. The algorithm is summed up as Algorithm 1. The rest of this section is devoted to its proof of correctness, i.e., the proof that the value s it returns satisfies $|G(x) - s| \leq 3 \cdot 2^{-p} G(x)$.

Proposition 5: With N as in Algorithm 1, the truncation error satisfies $|\sum_{n=N}^{\infty} G_n x^{3n}| \leq 2^{-p}\, G(x)$ for all $x \geq 0.5$.

Proof: First, because of line 2 of the algorithm, we have $\frac{3}{10}x^3 \leq (N+1)^2$, so that

$$\Big|\sum_{n=N}^{\infty} G_n x^{3n}\Big| \leq 2G_N x^{3N} \leq 2\left(\frac{ex^{3/2}}{3N}\right)^{2N} \leq 2\,(\alpha N)^{-2N}$$

by Lemma 4 and Corollary 2. Line 3 then ensures that

$$2\,(\alpha N)^{-2N} \leq \frac{2^{-p}}{128}\, e^{\frac{2}{3}x^{3/2}} x^{-3/4} \leq 2^{-p} G(x),$$

the last inequality coming from Lemma 5. ■

There are two sources of error besides the truncation: first, (u_n) is not exactly proportional to (c_n), especially when n is close to R. Second, roundoff errors happen during the evaluation of (u_n). Rigorous bounds for both sources of error have been proposed by Mattheij and van der Sluis [10]. We combine them with classical techniques (well-explained, e.g., in [4]) to choose the starting index R and the working precision t so as to guarantee the final accuracy.

We now recall Mattheij and van der Sluis' main result (adapted to our particular case, which simplifies the statement quite a bit). Consider a recurrence of the form (23). Denote by (c_n) a minimal solution that we wish to evaluate, and let (d_n) be the solution such that $d_0 = d_1 = 1$. Assume that (d_n) is a dominant solution and that the sequences (c_n), (d_n) and (c_n/d_n) are decreasing. Define $H = \frac{d_0}{c_0}\sum_{i=0}^{R-1}\frac{c_i}{d_i}$ and, for $i \le R$,

$$U_i = \begin{pmatrix} c_i & c_i \\ c_{i+1} & c_i\,\frac{d_{i+1}}{d_i}\end{pmatrix} \text{ and } B_i = \begin{pmatrix} \frac{-a_1(i)}{a_0(i)} & \frac{-a_2(i)}{a_0(i)} \\ 1 & 0\end{pmatrix}.$$

Let $v_R \in \mathbb{R}^2$ be a column vector, and for $i \le R-1$, let v_i be the result of the floating-point evaluation of $B_i\, v_{i+1}$ at precision t. Write $v_i = (u_i, u_{i+1})^{\mathrm{T}}$. If all operations were exact, (u_i) would be the solution of the recurrence such that $(u_R, u_{R+1})^{\mathrm{T}} = v_R$. To take rounding errors into account, we write $v_i = (B_i + 2^{-t}\mathcal{G}_i)v_{i+1}$ for some matrix $\mathcal{G}_i$ instead. Define $y_R = (y_{R1}, y_{R2})^{\mathrm{T}} = U_R^{-1} v_R$. Let $\mathcal{F}_i = \|U_i^{-1}\mathcal{G}_i U_{i+1}\|$, the matrix norm being subordinate to the $\|\cdot\|_\infty$ norm for vectors, and let $\mathcal{F} \ge \max_i \mathcal{F}_i$.

Theorem 1: [10, Theorem 4.1] Provided that the quantities

$$\mathcal{F}R2^{-t}, \quad \frac{c_R\, d_0}{d_R\, c_0}\left|\frac{y_{R2}}{y_{R1}}\right|, \quad \frac{\|y_R\|_\infty}{|y_{R1}|}\,(R+H)(1.3\mathcal{F}2^{-t})$$

are all bounded by 0.1, the approximate value u_i computed by Miller's algorithm satisfies $(c_0/u_0)\,u_i = c_i\,(1+\theta_i)$ for some θ_i such that $|\theta_i| \le T_i + R_i$, where

$$T_i = 1.5\left(\frac{c_R\, d_i}{c_i\, d_R} + \frac{c_R\, d_0}{c_0\, d_R}\right)\left|\frac{y_{R2}}{y_{R1}}\right|, \quad R_i = 1.5\,\frac{\|y_R\|_\infty}{y_{R1}}\,\varepsilon\,(i+2H),$$

and $\varepsilon = 1.3\mathcal{F}2^{-t}$.

Turning back to the special case $c_n = n!^2 G_n$, Theorem 1 applied to (15) yields the following. Recall that $\tau = 3/20$.

Corollary 3: Set $v_R = (1,0)^{\mathrm{T}}$ and

$$B_i = \begin{pmatrix} 10 & -r(i) \\ 1 & 0\end{pmatrix} \quad \text{where} \quad r(n) = \frac{(3n+4)(3n+5)}{(n+1)(n+2)}.$$

Then, in the notation of Theorem 1, we have $T_i \le \tau^{R-i}$ and $R_i \le 76.5(i+4)2^{-t} \le 76.5(N+3)2^{-t}$ for all $i < N$.

Proving this corollary still requires some work. We postpone it for a bit to explore the consequences of this statement.

Observe that lines 8 and 9 of Algorithm 1 are equivalent to a floating-point evaluation of $B_i v_{i+1}$ where $v_{i+1} = (a,b)^{\mathrm{T}}$. Hence, at each loop turn, we have $a = u_i$ just after line 8. Lines 10 and 11 are a Horner-like evaluation scheme, so that $s' \approx \sum_{i=0}^{N-1} u_i x^{3i}/i!^2$ at the end of the loop. More precisely, assuming x^3 is approximated by $\circ(x)\otimes\circ(x)\otimes\circ(x)$ and division by $(i+1)^2$ is performed as two successive divisions by $(i+1)$, an easy induction shows that one can write

$$s' = \sum_{i=0}^{N-1}\left(u_i\,\frac{x^{3i}}{i!^2}\right)\langle 9(i+1)\rangle\,.$$

Line 12 adds 3 to all error counters. The choice of t on line 5 ensures that $(9(N+1)+3)\cdot 2^{-t} \le 1/2$. Using Prop. 4, we conclude that the sum s returned by Algorithm 1 satisfies

$$s = \sum_{i=0}^{N-1}\frac{c_0\,u_i}{u_0}\cdot\frac{x^{3i}}{i!^2}\,(1+\mu_i) = \sum_{i=0}^{N-1} G_i\,x^{3i}(1+\mu_i)(1+\theta_i),$$

where $|\mu_i| \le 2\,(9(i+1)+3)\cdot 2^{-t} \le 18(N+3)\,2^{-t}$ and $|\theta_i| \le T_i + R_i$.

Since $R \ge N$, we have $T_i \le \tau$ for all $i < N$ by Corollary 3. The choice of t also implies $R_i \le 76.5/256$. Altogether, this ensures that $|\theta_i| \le 1$. Writing $(1+\mu_i)(1+\theta_i) = 1+\delta_i$, we get

$$|\delta_i| \le 2|\mu_i| + |\theta_i| \le 112.5(N+3)\,2^{-t} + \tau^{R-i} \le 2^{-p} + \tau^{R-i},$$

and therefore $\left|s - \sum_{i=0}^{N-1} G_i x^{3i}\right| \le 2^{-p}G(x) + \tau^R G(\tau^{-1/3}x)$.

Lemma 6: For $x > 0.5$, we have $\tau^R\,G(\tau^{-1/3}x) \le 2^{-p}G(x)$.

Proof: It follows from Lemma 5 (and $\tau < 1$) that

$$\frac{G(x)}{G(\tau^{-1/3}x)} \ge \frac{1}{4}\exp\left(\frac{2}{3}x^{3/2}(1-\tau^{-1/2})\right),$$

and the algorithm ensures that

$$R\log_2(\tau^{-1}) \ge p + 2 + \frac{2}{3}x^{3/2}(\tau^{-1/2}-1)\log_2(e),$$

whence $\tau^R \le \frac{1}{4}\,2^{-p}\,\exp(\frac{2}{3}\,x^{3/2}(1-\tau^{-1/2}))$ and the result. ■

We can now prove the correctness of Algorithm 1:

Theorem 2: The value s returned by Algorithm 1 satisfies $|G(x) - s| \le 3\cdot 2^{-p}G(x)$.

Proof: It follows from Prop. 5 and the above discussion, since $|G(x) - s| \le |G(x) - \sum_{i=0}^{N-1} G_i x^{3i}| + |s - \sum_{i=0}^{N-1} G_i x^{3i}|$. ■

The remainder of this section is devoted to the proof of Corollary 3. We begin with a crucial lemma. Let (d_n) be the solution of (15) defined by $d_0 = d_1 = 1$, let $\eta(n) = 1/(3n^2)$, and let $r(n)$ be as in Corollary 3.

Lemma 7: For all $n \ge 1$, we have $d_{n+1} \le d_n \le (1+\eta(n))\,d_{n+1}$.

Proof: We proceed by induction. Since $d_2 = 9/10$, the property is true for $n = 1$. Now, supposing it for an arbitrary n, we get $(9-\eta(n))d_{n+1} \le 10d_{n+1} - d_n \le 9d_{n+1}$, so

$$\frac{9-\eta(n)}{r(n)}d_{n+1} \le d_{n+2} \le \frac{9}{r(n)}d_{n+1}.$$

We conclude by observing that $9/r(n) \le 1$ and $r(n)/(9-\eta(n)) \le 1+\eta(n+1)$ for $n \ge 1$. ■

Corollary 4: For all n, we have $0.783 \le d_n \le 1$.

Proof: By Lemma 7, (d_n) is decreasing and $d_0 = 1$: this proves the right-hand side. For $n \ge 100$, Lemma 7 implies

$$d_{100} \le d_n \prod_{i=100}^{n-1}(1+\eta(i)) \le d_n\frac{p_\infty}{p_{99}}, \quad p_k = \prod_{i=1}^{k}(1+\eta(i)).$$

Using the exact value $p_\infty = \frac{\sqrt{3}}{\pi}\sinh(\frac{\pi}{\sqrt{3}})$ [1, Eq. 4.36.1], we check that $d_n \ge d_{100}\,p_{99}\,p_\infty^{-1} \ge 0.783$ for $n \ge 100$. As (d_n) is decreasing, the inequality holds for $n < 100$ too. ■

This estimate, combined with Corollary 1 gives almost all we need to check the hypotheses of Theorem 1. We use the notation introduced for the statement of the theorem (specialized to the computation of $c_n = n!^2 G_n$ using (15), with (d_n) as above).

Corollary 5: The sequences (c_n), (d_n) and (c_n/d_n) are decreasing. Moreover, the following inequalities hold: $H \le 2$, $|y_{R2}/y_{R1}| \le \frac{1}{6}$, $\|y_R\|_\infty/|y_{R1}| = 1$, and $\det U_i \ge \frac{3}{4}c_i^2$.

Proof: Corollary 1 shows that (c_n) is decreasing and Lemma 7 shows that (d_n) is decreasing. Together, they imply $\frac{c_{n+1}}{d_{n+1}}/\frac{c_n}{d_n} \le \tau(1+\eta(n)) \le \frac{1}{5}$ for $n \ge 1$. We check separately that $c_1/d_1 \le c_0/d_0$. Hence (c_n/d_n) is decreasing.

Corollary 1 also implies $c_n \le \tau^n c_0$ for any n, and Corollary 4 shows that $d_0 = 1 \le d_n/0.783$ for any n. It follows that

$$H = \frac{d_0}{c_0}\sum_{i=0}^{R-1}\frac{c_i}{d_i} \le \frac{1}{0.783}\sum_{i=0}^{R-1}\tau^i \le 2.$$

We have $d_0/d_1 = 1$, $d_1/d_2 = 10/9$, and $d_i/d_{i+1} \le 1+1/(3i^2) \le 10/9$ for $i \ge 2$, and, by definition of U_R and v_R,

$$y_R = U_R^{-1} v_R = \frac{c_R}{\det U_R} \begin{pmatrix} d_{R+1}/d_R & -1 \\ -c_{R+1}/c_R & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{c_R}{\det U_R} \begin{pmatrix} d_{R+1}/d_R \\ -c_{R+1}/c_R \end{pmatrix}.$$

It follows that $|y_{R2}/y_{R1}| \le \frac{10}{9}\tau = \frac{1}{6}$, and hence $\|y_R\|_\infty = |y_{R1}|$.

Finally, from the expression $\det U_i = c_i^2\,(d_{i+1}/d_i - c_{i+1}/c_i)$, we obtain $c_i^{-2}\,\det U_i \ge (\frac{9}{10} - \tau) = \frac{3}{4}$. ■

Lemma 8: A suitable value for $\mathcal{F}$ is 39.

Proof: The multiplication $B_i v_i$ is performed on lines 8 and 9 of the algorithm. Denoting by a' and b' the new values of a and b, we can write $b' = a$ and

$$a' = 10a\,\langle 2\rangle - r(i)b\,\langle 5\rangle = 10a(1+\theta) - r(i)b(1+\theta')$$

where (since $t \ge 5$ by line 5) $|\theta| \le 4 \cdot 2^{-t}$ and $|\theta'| \le 10 \cdot 2^{-t}$ by Prop. 4. (This assumes that the multiplication by $r(i)$ is performed through four successive multiplications and divisions). Therefore, we have

$$|\mathcal{G}_i| \le \begin{pmatrix} 4 \cdot 10 & 10\,r(i) \\ 0 & 0 \end{pmatrix},$$

where, for matrices, $|A| \le |B|$ means that the inequality holds entrywise. Since moreover

$$|U_i^{-1}| \le \frac{|c_i|}{|\det U_i|} \begin{pmatrix} d_{i+1}/d_i & 1 \\ c_{i+1}/c_i & 1 \end{pmatrix} \le \frac{4}{3c_i} \begin{pmatrix} 1 & 1 \\ 3/20 & 1 \end{pmatrix},$$

$$|U_{i+1}| \le c_{i+1} \begin{pmatrix} 1 & 1 \\ c_{i+2}/c_{i+1} & d_{i+2}/d_{i+1} \end{pmatrix} \le c_{i+1} \begin{pmatrix} 1 & 1 \\ 3/20 & 1 \end{pmatrix},$$

and $\mathcal{F}_i \le |U_i^{-1}| \cdot |\mathcal{G}_i| \cdot |U_{i+1}|$, we get

$$\mathcal{F}_i \le \frac{4}{3} \cdot \frac{3}{20} \begin{pmatrix} 40 + \frac{3}{2}r(i) & 40 + 10r(i) \\ \frac{3}{20}\,(40 + \frac{3}{2}r(i)) & \frac{3}{20}\,(40 + 10r(i)) \end{pmatrix}.$$

Hence $\|\mathcal{F}_i\| \le 16 + \frac{23}{10}\,r(i)$. The result follows because $r(i) \le 10$ for all $i \ge 0$. ■

We can finally prove Corollary 3.

Proof: To apply Theorem 1, we need to check that

(i) $39\,R2^{-t} \le 0.1$, (ii) $\dfrac{c_R\,d_0}{d_R\,c_0} \le 0.1\,\left|\dfrac{y_{R1}}{y_{R2}}\right|$,

(iii) $(R+2)(50.7 \cdot 2^{-t}) \le 0.1$.

By definition of t, we have $R + 2 \le 2^{t-9}$, so (iii) is satisfied, and (i) follows immediately. Corollary 5 combined with the inequalities $d_0/d_R \le 1/0.783$ and $c_R/c_0 \le \tau^R \le \tau$ implies (ii).

In conclusion, we can apply Theorem 1, hence

$$T_i = 1.5\left(\frac{c_R\,d_i}{c_i\,d_R} + \frac{c_R\,d_0}{c_0\,d_R}\right)\left|\frac{y_{R2}}{y_{R1}}\right| \le \frac{1.5}{0.783}\,\left(\tau^{R-i} + \tau^R\right)\,\frac{1}{6}.$$

Therefore, $T_i \le 0.639 \cdot \tau^{R-i} \le \tau^{R-i}$ as announced. Corollary 5 yields $R_i = 1.5\,(50.7 \cdot 2^{-t})(i + 2H) \le 76.5 \cdot (i+4)\,2^t$. ■

VI. Evaluation of the auxiliary series

The implementation of the auxiliary series F is much easier than that of G. We limit ourselves to a sketch of the (fairly standard) algorithm.

A variable a_0 is used to successively evaluate F_0, F_3x^3, F_6x^6, etc., using the recurrence (12). Accordingly, two variables a_1 and a_2 are used to evaluate the successive values of $F_{3i+1}x^{3i+1}$ and $F_{3i+2}x^{3i+2}$. Each step adds at most 10 to the relative error counter of each variable. A variable s is used to accumulate the sum as the variables a_k are updated. Therefore, after step K, we can write $s = \sum_{i=0}^{3K-1} F_i x^i\,\langle 1 + 10\lfloor i/3\rfloor + 3K - i\rangle$ (the term $3K - i$ representing the errors due to additions). Bounding all errors uniformly, we get

$$s = \sum_{i=0}^{3K-1} F_i x^i\,\langle 10K\rangle\,.$$

It is easy to see that $q(i) = F_{i+3}/F_i$ decreases for $i \ge 1$, hence the loop can be stopped as soon as (i) $q(3K)x^3 < 1/2$, and (ii) $a_0, a_1, a_2 < 2^{\mathrm{Exp}(s)-p-4}$. These conditions ensure that the remainder $\sum_{i\ge 3K} F_i\,x^i$ is bounded by $2(F_{3K} + F_{3K+1} + F_{3K+2}) < 4(a_0 + a_1 + a_2) < \frac{12}{16}2^{-p}\,2^{\mathrm{Exp}(s)}$.

It is clear from Prop. 1 that $F_{n+3} \le 4F_n/n^2$ for any n, hence (using $n! \approx (n/e)^n$) we have $F_n \approx (4e^2/n^2)^{n/3}$. This is most likely an overestimation of the true value. Moreover, we can approximate $F(x)$ by $\frac{1}{32}x^{-1/2}\exp(\frac{4}{3}x^{3/2})$ for $x > 0.5$. These estimates are used to get a rough overestimation of K. The working precision t is then chosen so that $20K \cdot 2^{-t} \le 2^{-3-p}$. Hence, we have $|\sum_{i=0}^{3K-1} F_i x^i - s| \le 2^{-3-p}\,s \le 2^{\mathrm{Exp}(s)-3-p}$.

The initial estimation of the truncation rank is very unlikely to be underestimated. In the case it would be smaller than the actual truncation rank decided on-the-fly by the above criterion, this is checked *a posteriori* and, if necessary, the evaluation is run again with an updated working precision.

We conclude by writing $|F(x) - s| \le |F(x) - \sum_{i=0}^{3K-1} F_i x^i| + |s - \sum_{i=0}^{3K-1} F_i x^i| \le \frac{7}{8}2^{-p}\,2^{\mathrm{Exp}(s)} \le 2^{-p}s$.

VII. Complete algorithm

Assume $p \ge 3$. From the previous sections, it appears that we computed $\widehat{G}$ such that $\widehat{G} = G(x)(1+\delta_1)$ with $|\delta_1| \le 3 \cdot 2^{-p}$, and we computed $\widehat{F}$ such that $F(x) = \widehat{F}(1+\delta_2)$ with $|\delta_2| \le 2^{-p}$. Hence $\widehat{G}/\widehat{F} = (G(x)/F(x))(1+\delta_3)$ where $|\delta_3| = |\delta_1 + \delta_2(1+\delta_1)| \le 5 \cdot 2^{-p}$. Indeed, the division is performed in floating-point arithmetic at precision p, leading a final result $\widehat{A} = (\widehat{G}/\widehat{F})(1+\delta_4)$ with $|\delta_4| \le 2^{-p}$. Hence, finally, $\widehat{A} = \mathrm{Ai}(x)(1+\delta_5)$ with $|\delta_5| = |\delta_3 + \delta_4(1+\delta_3)| \le 7 \cdot 2^{-p} \le 2^{-(p-3)}$.

We developed a prototype implementation of this algorithm, based on the multiple-precision floating-point library MPFR [6]. For simplicity, we supposed $x \ge 1/2$ in the present paper, but our implementation is valid for any $x \ge 0$.

We compared the results with those of the implementation of $\mathrm{Ai}(x)$ available in MPFR, run with a larger precision. Random tests using thousands of points x and target precisions p yield relative errors smaller than $2^{-(p-3)}$ between the result of our implementation and the result of MPFR, as predicted by theory.

We also ran our implementation and MPFR with the same accuracy to compare their performance (see Fig. 3). When x is large, our method is faster than MPFR, which uses the Taylor expansion of Ai at the origin. This is all the benefit of reducing the cancellation: for large x, the working precision of MPFR has a large overhead because of the bad condition number of the series. In contrast, when p is large compared to x, our implementation pays the cost of evaluating two series instead of one, with little benefit in terms of working precision.

To be completely fair, we should mention that MPFR does not implement the asymptotic expansion of Ai at infinity, which should be a better choice than our algorithm for large x and

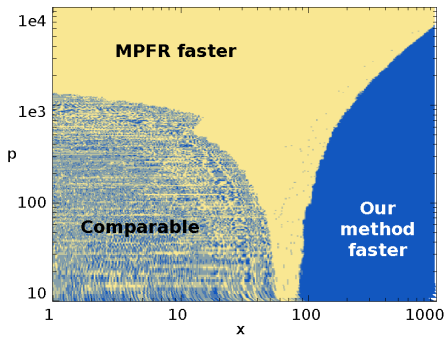

Figure 3. Fastest method as a function of x and p (scales are logarithmic). MPFR 3.1.1, GMP 4.3.1, on an Intel Xeon at 2.67GHz.

comparatively small p. On the other hand, our code currently uses only the naive series summation algorithm, while MPFR implements Smith's baby steps-giant steps technique [13], [16] that is more efficient. There is no theoretical obstacle to using Smith's method with our series: it only obfuscates a little bit the description of the algorithm and the roundoff error analysis. Once we will have implemented it, we will have a more complete picture with three areas, indicating what method between the Taylor series, the asymptotic series and our series is the most efficient, depending on (x, p).

VIII. Outlook: Towards a GMR algorithm?

As mentioned in the introduction, we see the present work as a case study. Indeed, in spite of the many technical details that occupy much of the space of this article, we really used few specific properties of the Airy function besides Equation (3). Looking back, our analysis essentially relies on the following ingredients.

(i) *The ability to find auxiliary series.* The indicator functions used in the GMR method depend only on the behaviour at (complex) infinity of the entire functions they are associated to. In the case of the solution of a LODE with analytic coefficients, this behaviour is entirely determined by the differential equation along with a finite number of "asymptotic initial values" [20], [19]. Once the indicator function is known, it remains to exhibit appropriate auxiliary series. Doing this in a truly general way remains an open problem. Yet, both the original GMR method and our variant apply to many cases, and it is likely that they can be combined and further generalized.

(ii) *An efficient way to compute their coefficients.* This seems to be considered a major limitation in the original GMR paper [7]. But, as already mentioned, recurrences with polynomial coefficients automatically exist as soon as both the original function to evaluate and the auxiliary series satisfy differential equations with polynomial coefficients. Numerical stability is not much of an issue in the case of three-term recurrences, thanks to Miller's method, though many technical details must be settled. The situation is more complicated in general for recurrences of higher order. Observe, though, that we proved the minimality of (G_n) using essentially the same asymptotic properties that were exploited by the GMR method in the first place. The minimality may hence not be fortuitous and might generalize.

(iii) *Upper and lower bounds on the coefficients and sums of the series F and G.* All these bounds were derived, in a pretty systematic way, from the asymptotic expansion of Ai at infinity combined with Lemma 1. Bounds similar to that from Lemma 1 can themselves often be obtained from a LODE [12].

(iv) *Roundoff error analyses.* Our error analyses follow a very regular pattern and could probably be abstracted to a more general case or automated.

In short, most steps of the present study could apparently be performed in a systematic way, starting from Eq. (3) plus a moderate amount of additional information. Systematizing the GMR method based on this observation seems a promising line of research. Solutions of LODE with polynomial coefficients are known in Computer Algebra as *D-finite*, or *holonomic*, functions. It would be interesting to isolate a subclass of D-finite functions to which the method applies in a truly systematic way, and attempt to *automate* it at least partially.

Acknowledgments: We thank Paul Zimmermann for pointing out the GMR article to us.

References

[1] *Digital library of mathematical functions*, 2010, Online companion to [11], http://dlmf.nist.gov/.
[2] W. G. Bickley, L. J. Comrie, J. C. P. Miller, D. H. Sadler, and A. J. Thompson, *Bessel functions*, Mathematical Tables, vol. X, British Association for the Advancement of Science, 1952.
[3] R. P. Brent and P. Zimmermann, *Modern computer arithmetic*, Cambridge University Press, 2010.
[4] S. Chevillard, *The functions erf and erfc computed with arbitrary precision and explicit error bounds*, Information and Computation **216** (2012), 72 – 95.
[5] P. Flajolet and R. Sedgewick, *Analytic combinatorics*, Cambridge University Press, 2009.
[6] L. Fousse, G. Hanrot, V. Lefèvre, P. Pélissier, and P. Zimmermann, *MPFR: A multiple-precision binary floating-point library with correct rounding*, ACM TOMS **33** (2007), no. 2, 13:1–13:15.
[7] W. Gawronski, J. Müller, and M. Reinhard, *Reduced cancellation in the evaluation of entire functions and applications to the error function*, SIAM Journal on Numerical Analysis **45** (2007), no. 6, 2564–2576.
[8] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, second ed., SIAM, 2002.
[9] B. Y. Levin, *Lectures on entire functions*, AMS, 1996.
[10] R. M. M. Mattheij and A. van der Sluis, *Error Estimates for Miller's Algorithm*, Numerische Mathematik **26** (1976), 61–78.
[11] F. W. Olver, D. W. Lozier, R. F. Boisvert, and C.W. Clark (eds.), *NIST handbook of mathematical functions*, Cambridge University Press, 2010.
[12] F. W. J. Olver, *Asymptotics and special functions*, A K Peters, 1997.
[13] M. S. Paterson and L. J. Stockmeyer, *On the number of nonscalar multiplications necessary to evaluate polynomials*, SIAM Journal on Computing **2** (1973), no. 1, 60–66.
[14] M. Reinhard, *Reduced cancellation in the evaluation of entire functions and applications to certain special functions*, Ph.D. thesis, Universität Trier, 2008.
[15] H. Robbins, *A remark on Stirling's formula*, The American Mathematical Monthly **62** (1955), no. 1, 26–29.
[16] D. M. Smith, *Efficient multiple-precision evaluation of elementary functions*, Mathematics of Computation **52** (1989), no. 185, 131–134.
[17] R. P. Stanley, *Enumerative combinatorics*, vol. 2, Cambridge University Press, 1999.
[18] I. A. Stegun and R. Zucker, *Automatic computing methods for special functions*, Journal of Research of the National Bureau of Standards **74B** (1970), 211–224.
[19] M. van der Put and M. F. Singer, *Galois theory of linear differential equations*, Springer, 2003.
[20] W. R. Wasow, *Asymptotic expansions for ordinary differential equations*, Wiley, 1965.
[21] J. Wimp, *Computation with recurrence relations*, Pitman, Boston, 1984.