Rounding Error Analysis of Linear Recurrences Using Generating Series

Marc Mezzarobba

ROUNDING ERROR ANALYSIS OF LINEAR RECURRENCES USING GENERATING SERIES

MARC MEZZAROBBA

Abstract. We develop a toolbox for the error analysis of linear recurrences with constant or polynomial coefficients, based on generating series, Cauchy's method of majorants, and simple results from analytic combinatorics. We illustrate the power of the approach by several nontrivial application examples. Among these examples are a new worst-case analysis of an algorithm for computing Bernoulli numbers, and a new algorithm for evaluating differentially finite functions in interval arithmetic while avoiding interval blow-up.

1. Introduction

This article aims to illustrate a technique for bounding round-off errors in the floating-point evaluation of linear recurrence sequences that we found to work well on a number of interesting examples. The main idea is to encode as generating series both the sequence of "local" errors committed at each step and that of "global" errors resulting from the accumulation of local errors. While the resulting bounds are unlikely to be surprising to specialists, generating series techniques, curiously, do not seem to be classical in this context.

As is well-known, in the evaluation of a linear recurrence sequence, rounding errors typically cancel out to a large extent instead of purely adding up. It is crucial to take this phenomenon into account in the analysis in order to obtain realistic bounds, which makes it necessary to study the propagation of local errors in the following steps of the algorithm somewhat finely. In the classical language of sequences, this tends to involve complicated manipulations of nested sums and yield opaque expressions.

Generating series prove a convenient alternative for several reasons. Firstly, they lead to more manageable formulae: convolutions become products, and the relation between the local and the accumulated errors can often be expressed exactly as an algebraic or differential equation involving their generating series. Secondly, such an equation opens the door to powerful analytic techniques like singularity analysis or Cauchy's method of majorants. Thirdly, as illustrated in Section 10, a significant part of the laborious calculations involved in obtaining explicit constants can be carried out with the help of computer algebra systems when the calculation is expressed using series.

2010 *Mathematics Subject Classification.* 65G50; 65Q30; 65L70; 05A15.

In this article, we substantiate our claim that generating series are an adequate language for error analysis by applying it to a selection of examples from the literature. As detailed below, some of the examples yield results that appear to be new and may be of independent interest.

The text is organized as follows. In order to get a concrete feeling of the basic idea, we start in Section 2 with an elementary example, postponing to Section 3 the discussion of related work. We continue in Sections 4 to 7 with a review of classical facts about generating series, asymptotics, the Cauchy majorant method, and floating-point error analysis that constitute our basic toolbox. Building on this background, we then illustrate the approach outlined in Section 2 in situations involving polynomial coefficients (Legendre polynomials, Section 8) and floating-point arithmetic (with a revisit of the toy example in Section 9). A reader only interested in understanding the method can stop there.

The remaining sections present more substantial applications of the same idea. They can be read independently, referring to Sections 4 to 7 for basic results as necessary. Section 10 answers a question of R. P. Brent and P. Zimmermann on the floating-point computation of Bernoulli numbers. Section 11 discusses variations on Boldo's [1] error analysis of a finite difference scheme for the 1D wave equation, using series with coefficients in a normed algebra to encode a bivariate recurrence. Finally, in Section 12, we take a slightly different perspective and ask how to evaluate the sum of a series whose coefficients satisfy a recurrence when the recurrence is part of the input. Under mild assumptions, we give an algorithm that computes a rigorous enclosure of the sum while avoiding the exponential blow-up that would come with a naive use of interval arithmetic.

2. A Toy Example

Our first example is borrowed from Boldo [1, Section 2.1] and deals with the evaluation of a very simple, explicit linear recurrence sequence with constant coefficients in a simple generic model of approximate arithmetic. It is not hard to carry out the error analysis in classical sequence notation, cf. [1], and the reader is encouraged to duplicate the reasoning in his or her own favorite language.

Consider the sequence $(c_n)_{n\geqslant -1}$ defined by the recurrence

$$c_{n+1} = 2c_n - c_{n-1} \tag{1}$$

with $c_{-1} = 0$ and a certain initial value c_0. Let us assume that we are computing this sequence iteratively, so that each iteration generates a small *local error* corresponding to the evaluation of the right hand side. We denote by $(\tilde{c}_n)$ the sequence of computed values. The local errors accumulate over the course of the computation, and our goal is to bound the *global error* $\delta_n = \tilde{c}_n - c_n$.

We assume that each arithmetic operation produces an error bounded by a fixed quantity u. (This model is similar to fixed-point arithmetic, except that we ignore the fact that some fixed-point operations are exact in order to make the analysis more interesting.) Thus, we have

$$\tilde{c}_{n+1} = 2\tilde{c}_n - \tilde{c}_{n-1} + \varepsilon_n, \qquad |\varepsilon_n| \leqslant 2u \tag{2}$$

for all $n \geqslant 0$. We will also assume $|\delta_0| \leqslant u$. Subtracting (1) from (2) yields

$$\delta_{n+1} = 2\delta_n - \delta_{n-1} + \varepsilon_n, \quad n \geqslant 0, \tag{3}$$

with $\delta_{-1} = 0$.

A naive forward error analysis would have us write $|\delta_{n+1}| \leqslant 2|\delta_n| + |\delta_{n-1}| + 2u$ and conclude by induction that $|\delta_n| \leqslant 3^{n+1}u$, or, with a bit more effort,

$$|\delta_n| \leqslant (\lambda_+\alpha_+^n + \lambda_-\alpha_-^n - 4)u, \qquad \alpha_\pm = 1 \pm \sqrt{2}, \quad \lambda_\pm = 4 \pm 3\sqrt{2}. \tag{4}$$

Neither of these bounds is satisfactory. To see why, it may help to consider the propagation of the first few rounding errors. Writing $\tilde{c}_0 = c_0 + \delta_0$ and $\tilde{c}_1 = 2\tilde{c}_0 + \varepsilon_0 = c_1 + 2\delta_0 + \varepsilon_0$, we have

$$\tilde{c}_2 = 2(2c_0 + 2\delta_0 + \varepsilon_0) - (c_0 + \delta_0) + \varepsilon_1 = c_2 + 3\delta_0 + 2\varepsilon_0 + \varepsilon_1$$

and hence $|\delta_2| \leqslant 9u$. The naive analysis effectively puts absolute values in this expression, leading to $|\delta_2| \leqslant 5|\delta_0| + 2|\varepsilon_0| + |\varepsilon_1| \leqslant 11u$ instead. Overestimations of this kind compound as n increases. Somehow keeping track of the expression of δ_n as a linear combination of the ε_i (and δ_0) clearly should yield better estimates.

To do so, let us note that (3) is a linear recurrence with the same homogeneous part as (1) and the sequence of local errors on the right-hand side, and rephrase this relation in terms of generating series. Define the formal power series[1]

$$\delta(z) = \sum_{n \geqslant 0} \delta_n z^n, \qquad \varepsilon(z) = \sum_{n \geqslant 0} \varepsilon_n z^n.$$

The relation (3) implies

$$(z^{-1} - 2 + z)\delta(z) = \sum_{n \geqslant -1} \delta_{n+1} z^n - 2\sum_{n \geqslant 0} \delta_n z^n + \sum_{n \geqslant 1} \delta_{n-1} z^n = \delta_0 z^{-1} + \sum_{n \geqslant 0} \varepsilon_n z^n$$

that is,

$$\delta(z) = \frac{\delta_0 + z\varepsilon(z)}{(1-z)^2}.$$

Since $|\delta_0| \leqslant u$ and $|\varepsilon_n| \leqslant 2u$, we see that the absolute values of the coefficients of the numerator are bounded by those of the corresponding coefficients in the series expansion of $2u/(1-z)$. Denoting by $\ll$ this termwise inequality relation, it follows that

$$\delta(z) \ll \frac{2u}{1-z}\frac{1}{(1-z)^2} = \frac{2u}{(1-z)^3}.$$

Going back to the coefficient sequences, this bound translates into $|\delta_n| \leqslant (n+1) \cdot (n+2)u$, a much sharper result than (4).

3. Related Work

There is a large body of literature on numerical aspects of linear recurrence sequences, with a focus on stability issues and backward recurrence algorithms[2]. A good entry point is Wimp's book [38]. Comparatively little has been written on linear recurrences considered in the forward direction, in Wimp's words, "not because a forward algorithm is more difficult to analyze, but rather for the opposite reason—that its analysis was considered straightforward" [37]. The first completely explicit error analysis of general linear recurrences that we are aware of appears in the work of Oliver [29, Section 2] (see also [28]). However, the importance of using linearity to study the propagation of local errors was recognized well before.

[1] While the sequence (δ_n) naturally starts at $n = -1$, the fact that $\delta_{-1} = 0$ allows us to use the same summation range for both series.

[2] Where the recurrence relation is used for decreasing n and combined with asymptotic information on the sequence, typically to compute minimal solutions.

For example, the first of Henrici's books on numerical methods for differential equations [17, Section 1.4] uses the terms "local round-off error" and "accumulated round-off error" with the same meaning as we do in a closely related setting.

Furthermore, linear recurrences are a special cases of triangular systems of linear equations. For example, computing the first n terms of the sequence c_n of the previous section is the same as solving the banded Toeplitz system

$$\begin{pmatrix} 1 & & & & \\ -2 & 1 & & \mathbf{0} & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ \mathbf{0} & & 1 & -2 & 1 \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} c_0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

The triangular case is an almost trivial subproblem in von Neumann and Goldstine's [34][3] and (more explicitly) Turing's [32, Section 12, p. 306] landmark analyses of linear system solving, both concluding in a polynomial growth with n of the forward error when some quantities related to the inverse or the condition number of the matrix are fixed. We refer to the encyclopedic book by Higham [18, Chapter 8] for a detailed discussion of the error analysis of triangular systems and further historical perspective.

Because of their dependency on condition numbers, these results do not, in themselves, rule out an exponential buildup of errors in the case of recurrences. In the standard modern proof, the forward error bound results from the combination of a backward error bound and a perturbation analysis that could in principle be refined to deal specifically with recurrences. An issue with this approach is that, to view the numeric solution as the exact solution corresponding to a perturbed input, one is led to perturb the matrix in a fashion that destroys the structure inherited from the recurrence. It is plausible that one could nevertheless derive meaningful bounds for recurrences, from, e.g., Theorem 8.5 in [18]. Our claim is that the tools of the present paper are better suited to the task. Note that use of linearity to study error propagation can also be viewed as an instance of backward error analysis, where one chooses to perturb the right-hand side of the system instead of the matrix. From this perspective, the present paper is about a convenient way of carrying out the perturbation analysis that enables one to pass to a forward error bound.

Except for the earlier publication [21] of the example considered again in Section 8 of the present paper, we are not aware of any prior example of error analysis conducted using generating series in numerical analysis, scientific computing or computer arithmetic. A close analogue appears however in the realm of digital signal processing, with the use of the Z-transform to study the propagation of rounding errors in realization of digital filters starting with Liu and Kaneko [24]. The focus in signal processing is rarely on worst-case error bounds, with the notable exception of recent work by Hilaire and collaborators, e.g., [19].

[3] See also Grcar's commentary [16, Section 4.5].

4. Generating Series

Let R be a ring, typically $R = \mathbb{R}$ or $R = \mathbb{C}$. We denote by $R[[z]]$ the ring of formal power series

$$u(z) = \sum_{n=0}^{\infty} u_n z^n,$$

where $(u_n)_{n=0}^{\infty}$ is an arbitrary sequence of elements of R. A series $u(z)$ used primarily as a more convenient encoding of its coefficient sequence (u_n) is called the *generating series* of (u_n).

It is often convenient to extend the coefficient sequence to negative indices by setting $u_n = 0$ for $n < 0$. We can then write $u(z) = \sum_n u_n z^n$ with the implicit summation range extending from $-\infty$ to ∞ (keeping in mind that the product of series of this form does not make sense in general if the coefficients are allowed to take nonzero values for arbitrarily negative n).

Given $u \in R[[z]]$ and $n \in \mathbb{Z}$, we denote by u_n or $[z^n]u(z)$ the coefficient of z^n in $u(z)$. Conversely, whenever (u_n) is a numeric sequence with integer indices, $u(z)$ is its generating series. We occasionally consider sequences $u_0(z), u_1(z), \dots$ of series, with $u_{i,n} = [z^n]u_i(z)$ in this case. We often identify expressions representing analytic functions with their series expansions at the origin. For instance, $[z^n](1-\alpha z)^{-1}$ is the coefficient of z^n in the Taylor expansion of $(1-\alpha z)^{-1}$ at 0, that is, α^n.

We denote by S the forward shift operator on bilateral sequences, that is, the operator mapping a sequence $(u_n)_{n\in\mathbb{Z}}$ to $(u_{n+1})_{n\in\mathbb{Z}}$, and by S^{-1} its inverse. Thus, $S \cdot (u_n)_{n\in\mathbb{Z}}$ is the coefficient sequence of the series $z^{-1}u(z)$. More generally, it is well-known that linear recurrence sequences with constant coefficients correspond to rational functions in the realm of generating series, as in the toy example from Section 2.

It is also classical that the correspondence generalizes to recurrences with variable coefficients depending polynomially on n as follows. We consider recurrence relations of the form

$$p_0(n)u_n + p_1(n)u_{n-1} + \cdots + p_s(n)u_{n-s} = b_n, \quad n \in \mathbb{Z}, \tag{5}$$

where $p_0, \dots, p_s$ are polynomials. Given sequences expressed in terms of an index called n, we also denote by n the operator

$$(u_n)_{n\in\mathbb{Z}} \quad \mapsto \quad (nu_n)_{n\in\mathbb{Z}}.$$

We then have $Sn = (n+1)S$, where the product stands for the composition of operators.

Example 1. With these conventions,

$$(nS + n - 1) \cdot (u_n)_{n\in\mathbb{Z}} = (nu_{n+1} + (n-1)u_n)_{n\in\mathbb{Z}} = ((S+1)(n-1)) \cdot (u_n)_{n\in\mathbb{Z}}$$

is an equality of sequences that parallels the operator equality $nS + n - 1 = (S+1)(n-1)$.

Any linear recurrence operator of finite order with polynomial coefficients can thus be written as a polynomial in n and $S^{\pm 1}$. Denoting with a dot the action of operators on sequences, (5) thus rewrites as

$$L(n, S^{-1}) \cdot (u_n) = (b_n) \quad \text{where} \quad L = \sum_{k=0}^{s} p_k(X)Y^k \in R[X][Y].$$

When dealing with sequences that vanish eventually (or that converge fast enough) as $n \to -\infty$, we can also consider operators of infinite order

$$p_0(n) + p_1(n)S^{-1} + p_2(n)S^{-2} + \cdots = \sum_{i=0}^{\infty} p_i(n)S^{-i} = P(n, S^{-1})$$

where $P \in R[X][[Y]]$.

Similarly, we view the multiplication by z of elements of $R[[z]]$ as a linear operator that can be combined with the differentiation operator $\mathrm{d}/\mathrm{d}z$ to form linear differential operators with polynomial or series coefficients. For example, we have

$$\left(\frac{\mathrm{d}}{\mathrm{d}z}\frac{1}{1-z}\right) \cdot u(z) = \frac{\mathrm{d}}{\mathrm{d}z} \cdot \frac{u(z)}{1-z} = \left(\frac{1}{1-z}\frac{\mathrm{d}}{\mathrm{d}z} + \frac{1}{(1-z)^2}\right) \cdot u(z)$$

where the rational functions are to be interpreted as power series.

Lemma 2. *Let $(u_n)_{n\in\mathbb{Z}}, (v_n)_{n\in\mathbb{Z}}$ be sequences of elements of R, with $u_n = v_n = 0$ for $n < 0$. Consider a recurrence operator of the form $L(n, S^{-1})$ with $L(X, Y) \in R[X][[Y]]$. The sequences $(u_n), (v_n)$ are related by the recurrence relation $L(n, S^{-1}) \cdot (u_n) = (v_n)$ if and only if their generating series satisfy the differential equation*

$$L\left(z\frac{\mathrm{d}}{\mathrm{d}z}, z\right) \cdot u(z) = v(z).$$

Proof. This follows from the relations

$$\sum_{n=-\infty}^{\infty} n f_n z^n = z\frac{\mathrm{d}}{\mathrm{d}z}\sum_{n=-\infty}^{\infty} f_n z^n, \qquad \sum_{n=-\infty}^{\infty} f_{n-1} z^n = z\sum_{n=-\infty}^{\infty} f_n z^n,$$

noting that the operators of infinite order with respect to S^{-1} that may appear when the coefficients of the differential equation are series are applied to sequences that vanish for negative n. □

Generating series of sequences satisfying recurrences of the form (5)—that is, by Lemma 2, formal series solutions of linear differential equations with polynomial coefficients—are called *differentially finite* or *holonomic*. We refer the reader to [22, 31] for an overview of the powerful techniques available to manipulate these series and their generalizations to several variables.

5. Asymptotics

Beside their being a serviceable encoding for deriving identities between sequences, perhaps the main appeal of generating series is how they give access to the *asymptotics* of the corresponding sequences. The basic fact here is simply the Cauchy-Hadamard theorem stating that the inverse of the radius of convergence of $u(z)$ is the limit superior of $|u_n|^{1/n}$ as $n \to \infty$. Concretely, as soon as we have an expression of $u(z)$ (or an equation satisfied by it) that makes it clear that it has a positive radius of convergence and where the complex singularities of the corresponding analytic function are located, the exponential growth order of $|u_n|$ follows immediately.

Much more precise results are available when more is known on the nature of singularities. We quote here a simple result of this kind that will be enough for our purposes, and refer to the book of Flajolet and Sedgewick [14] for far-ranging generalizations (see in particular [14, Corollary VI.1, p. 392] for a statement containing the following lemma as a special case).

Lemma 3. *Assume that, for some $\rho > 0$, the series $u(z) = \sum_n u_n z^n$ converges for $|z| < \rho$ and that its sum has a single singularity $\alpha \in \mathbb{C}$ with $|\alpha| = \rho$. Let Ω denote a disk of radius $\rho' > \rho$, slit along the ray $\{t\alpha : t \in [\rho, \rho']\}$, and assume that $u(z)$ extends analytically to Ω. If for some $C \in \mathbb{C}$ and $m \in \mathbb{C}\backslash\mathbb{Z}_{\leqslant 0}$, one has*

$$u(z) \sim \frac{C}{(1-\alpha^{-1}z)^m}$$

as $z \to \alpha$ from within Ω, then the corresponding coefficient sequence satisfies

$$u_n \sim \frac{C}{\Gamma(m)} n^{m-1} \alpha^{-n}$$

as $n \to \infty$.

6. Majorant Series

While access to identities of sequences and to their asymptotic behavior is important for error analysis, we are primarily interested in *inequalities*. A natural way to express bounds on sequences encoded by generating series is by majorant series, a classical idea of "19th century" analysis.

Definition 4. Let $f = \sum_n f_n z^n \in \mathbb{C}[[z]]$.

(a) A formal series with nonnegative coefficients $\hat{f} = \sum_n \hat{f}_n z \in \mathbb{R}_{\geqslant 0}[[z]]$ is said to be a *majorant series* of f when we have $|f_n| \leqslant \hat{f}_n$ for all $n \in \mathbb{N}$. We then write $f \ll \hat{f}$.

(b) We denote by ${}^\sharp f$ the minimal majorant series of f, that is, ${}^\sharp f = \sum_n |f_n| z^n$.

We also write $f \gg 0$ to indicate simply that f has real, nonnegative coefficients. Series denoted with a hat always have nonnegative coefficients, and $\hat{f}$ is typically some kind of bound on f, though not necessarily a majorant series in the sense of the above definition.

The following properties are classical and easy to check (see, e.g., Hille [20, Section 2.4]).

Lemma 5. *Let $f, g \in \mathbb{C}[[z]]$, $\hat{f}, \hat{g} \in \mathbb{R}_{\geqslant 0}[[z]]$ be such that $f \ll \hat{f}$ and $g \ll \hat{g}$.*

(1) The following assertions hold, where $f_{N:}(z) = \sum_{n \geqslant N} f_n z^n$:

(a)	$f + g \ll \hat{f} + \hat{g}$,	*(b)*	$\gamma f \ll \lvert\gamma\rvert \hat{f}$ *for* $\gamma \in \mathbb{C}$,
(c)	$f_{N:}(z) \ll \hat{f}_{N:}(z)$ *for* $N \in \mathbb{N}$,	*(d)*	$f'(z) \ll \hat{f}'(z)$,
(e)	$\left(\int_0^z f\right) \ll \left(\int_0^z \hat{f}\right)$,	*(f)*	$fg \ll \hat{f}\hat{g}$.

(2) The disk of convergence $\hat{D}$ of $\hat{f}$ is contained in that of f, and when $\hat{g}_0 \in \hat{D}$, we have $f(g(z)) \ll \hat{f}(\hat{g}(z))$. In particular, $|f(\zeta)|$ is bounded by $\hat{f}(|\zeta|)$ for all $\zeta \in \hat{D}$.

While majorant series are a concise way to express some types of inequalities between sequences, their true power comes from Cauchy's *method of majorants* [8, 9][4]. This method is a way of computing majorant series of solutions of functional equations that reduce to fixed-point equations. The idea is that when the terms of a

[4]See Cooke [12] for an interesting account of the history of this method and its extensions, culminating in the Cauchy-Kovalevskaya theorem on partial differential equations.

series solutions can be determined iteratively from the previous ones, it is often possible to "bound" the equation by a simpler "model equation" whose solutions (with suitable initial values) then automatically majorize those of the original equation.

A very simple result of this kind states that the solution y of a *linear* equation $y = ay+b$ is bounded by the solution $\hat{y}$ of $\hat{y} = \hat{a}\hat{y}+\hat{b}$ when $a \ll \hat{a}$, $b \ll \hat{b}$ and $\hat{a}_0 = 0$. Let us prove a variant of this fact. The previous statement follows by applying the lemma to ${}^\sharp y$.

Lemma 6. *Let $\hat{a}, \hat{b}, y \in \mathbb{R}_{\geqslant 0}[[z]]$ be power series with $\hat{a}_0 = 0$ such that (note the $\ll$ sign)*

$$y(z) \ll \hat{a}(z)y(z) + \hat{b}(z).$$

Then one has

$$y(z) \ll \hat{y}(z) := \frac{\hat{b}(z)}{1 - \hat{a}(z)}.$$

Proof. Extracting the coefficient of z^n on both sides of the inequality on $y(z)$ yields

$$|y_n| \leqslant \sum_{i=1}^{n} \hat{a}_i |y_{n-i}| + \hat{b}_n, \tag{6}$$

where the sums starts at $i = 1$ due to the assumption that $\hat{a}_0 = 0$. Similarly, the $\hat{y}_n$ satisfy

$$\hat{y}_n = \sum_{i=1}^{n} \hat{a}_i \hat{y}_{n-i} + \hat{b}_n. \tag{7}$$

One sees by comparing (6) and (7) that $|y_k| \leqslant \hat{y}_k$ for all $k < n$ implies $|y_n| \leqslant \hat{y}_n$ (including the trivial case $|y_0| \leqslant \hat{b}_0 = \hat{y}_0$) so that, by induction, one has $|y_n| \leqslant |\hat{y}_n|$ for all n. □

Another classical instance of the method applies to nonsingular linear differential equations with analytic coefficients. In combination with Lemma 2 above, it allows us to derive bounds on linear recurrence sequences with polynomial coefficients, including recurrences of infinite order.

Proposition 7. *Let $a_0, \dots, a_{r-1}, b \in \mathbb{C}[[z]]$, $\hat{a}_0, \dots, \hat{a}_{r-1}, \hat{b} \in \mathbb{R}_{\geqslant 0}[[z]]$ be such that $a_k \ll \hat{a}_k$ for $0 \leqslant k < r$ and $b \ll \hat{b}$. Assume that $\hat{y} \in \mathbb{R}[[z]]$ is a solution of the equation*

$$\hat{y}^{(r)}(z) - \hat{a}_{r-1}(z)\hat{y}^{(r-1)}(z) - \cdots - \hat{a}_1(z)\hat{y}'(z) - \hat{a}_0(z)\hat{y}(z) = \hat{b}(z). \tag{8}$$

Then, any solution $y \in \mathbb{C}[[z]]$ of

$$y^{(r)}(z) - a_{r-1}(z)y^{(r-1)}(z) - \cdots - a_1(z)y'(z) - a_0(z)y(z) = b(z) \tag{9}$$

with $|y_0| \leqslant \widehat{y_0}, \dots, |y_{r-1}| \leqslant \hat{y}_{r-1}$ satisfies $y \ll \hat{y}$.

Proof. Write $y^{(k)}(z) = \sum_n (n+k)^{\underline{k}} y_{n+k} z^n$, where $n^{\underline{k}} = n(n-1)\cdots(n-k+1)$. The equation on $y(z)$ translates into

$$\sum_n (n+r)^{\underline{r}} y_{n+r} z^n - \sum_{k=0}^{r-1} \sum_n \sum_{j=0}^{\infty} a_{k,j} (n+k)^{\underline{k}} y_{n+k-j} z^n = \sum_n b_n z^n,$$

whence

$$n^{\underline{r}} y_n = \sum_{j=0}^{\infty} \sum_{k'=1}^{r} a_{k,j} (n-k')^{\underline{r-k'}} y_{n-k'-j} + b_{n-r},$$

and similarly for $\hat{y}(z)$. As with Lemma 2, these formulae hold for $n \in \mathbb{Z}$. The right-hand side only involves coefficients y_j with $j < n$, and the polynomial coefficients $(n-k')^{\underline{r-k'}}$, including $n^{\underline{r}}$, are nonnegative as soon as $n \geqslant r$. For $n \geqslant r$ and assuming $|y_k| \leqslant \hat{y}_k$ for all $k < n$, we thus have

$$\begin{aligned} n^{\underline{r}} |y_n| &\leqslant \sum_{j=0}^{\infty} \sum_{k'=1}^{r} |a_{k,j}| (n-k')^{\underline{r-k'}} |y_{n-k'-j}| + |b_{n-r}| \\ &\leqslant \sum_{j=0}^{\infty} \sum_{k'=1}^{r} \hat{a}_{k,j} (n-k')^{\underline{r-k'}} \hat{y}_{n-k'-j} + \hat{b}_{n-r} \\ &= n^{\underline{r}} \hat{y}_n. \end{aligned}$$

The result follows by induction starting from the initial inequalities $|y_0| \leqslant \hat{y}_0, \ldots, |y_{r-1}| \leqslant \hat{y}_{r-1}$. □

Like in the case of linear algebraic equations, this result admits variants that deal with differential inequalities. We limit ourselves to first-order equations here.

Lemma 8. *Consider power series $\hat{a}_0, \hat{a}_1, b, y \in \mathbb{R}_{\geqslant 0}[[z]]$ with nonnegative coefficients such that $\hat{a}_1(0) = 0$ and*

$$y'(z) \ll \hat{a}_1(z) y'(z) + \hat{a}_0(z) y(z) + \hat{b}(z). \tag{10}$$

The equation

$$\hat{y}'(z) = \hat{a}_1(z) \hat{y}'(z) + \hat{a}_0(z) \hat{y}(z) + \hat{b}(z) \tag{11}$$

admits a unique solution $\hat{y}$ with $\hat{y}(0) = y(0)$, and one has $y \ll \hat{y}$.

Proof. Since $\hat{a}_{1,0} = 0$, the right-hand side of the inequality

$$(n+1) y_{n+1} \leqslant \sum_{j=1}^{n} \hat{a}_{1,j} (n-j+1) y_{n-j+1} + \sum_{j=0}^{n} \hat{a}_{0,j} y_{n-j} + \hat{b}_n$$

corresponding to the extraction of the coefficient of z^n in (10) only involves coefficients y_j with $j \leqslant n$. Equation (11) corresponds to a recurrence of a similar shape (and therefore has a unique solution), and one concludes by comparing these relations. □

Solving majorant equations of the type (8), (11) yields majorants involving antiderivatives. The following observation can be useful to simplify the resulting expressions.

Lemma 9. *For $\hat{f}, \hat{g} \in \mathbb{R}_{\geqslant 0}[[z]]$, one has $\int_0^z (\hat{f}\hat{g}) \ll \hat{f} \int_0^z \hat{g}$. In particular, $\int_0^z \hat{f}$ is bounded by $z\hat{f}(z)$.*

Proof. Integration by parts shows that $\int_0^z (\hat{f}\hat{g}) - \hat{f} \int_0^z \hat{g} \in \mathbb{R}_{\geqslant 0}[[z]]$. □

It is possible to state much more general results along these lines, and cover, among other things, general implicit functions, solutions of partial differential equations, and various singular cases. We refer to [7, Chap. VII], [33], [35], and [15, Appendix A] for some such results.

7. Floating-Point Errors

The toy example from Section 2 illustrates error propagation in fixed-point arithmetic, where each elementary operation introduces a bounded absolute error. In a floating-point setting, the need to deal with the propagation of relative errors complicates the analysis. Thorough treatments of the analysis of floating-point computations can be found in the books of Wilkinson [36] and Higham [18]. We will use the following definitions and properties.

For simplicity, we assume that we are working in binary floating-point arithmetic with unbounded exponents. We make no attempt at covering underflows. (It would be interesting, though, to extend the methodology to this case.) Following standard practice, our results are mainly based on the following inequalities that link the approximate version $\tilde{*}$ of each arithmetic operation $* \in \{+, -, \times, /\}$ to the corresponding exact mathematical operation:

$$x\tilde{*}y = (x * y)(1 + \delta_1), \qquad |\delta_1| \leqslant u \quad \text{(“first standard model”, e.g., [18, (2.4)])},$$
$$x\tilde{*}y = (x * y)(1 + \delta_2)^{-1}, \quad |\delta_2| \leqslant u \quad \text{(“modified standard model”, e.g., [18, (2.5)])}.$$

In addition, we occasionally use the fact that multiplications by powers of two are exact.

The quantity u that appears in the above bounds is called the *unit roundoff* and depends only on the precision and rounding mode. For example, in standard t-bit round-to-nearest binary arithmetic, one can take $u = 2^{-t}$.

The two “standard models” are somewhat redundant, and most error analyses in the literature proceed exclusively from the first standard model. However, working under the modified standard model sometimes helps avoid assumptions that $nu < 1$ when studying the effect of a chain of n operations, which is convenient when working with generating series.

Suppose that a quantity $x \neq 0$ is affected by successive relative errors $\delta_1, \delta_2, \dots$ resulting from a chain of dependent operations, with $|\delta_i| \leqslant u$ for all i. The cumulative relative error η after n steps is given by

$$1 + \eta = \prod_{i=1}^{n} (1 + \delta_i).$$

This leads us to introduce the following notation.

Definition 10. When the roundoff error u is fixed and clear from the context:

(a) We write $\eta = \theta_n$ to indicate that $1 + \eta = \prod_{i=1}^{n}(1 + \delta_i)$ for some $\delta_1, \dots, \delta_n$ with $|\delta_i| \leqslant u$.

(b) We define $\hat{\theta}_n = (1 + u)^n - 1$ for $n \geqslant 0$. As usual, this sequence is extended to $n \in \mathbb{Z}$ by setting $\hat{\theta}_n = 0$ for $n < 0$, and we also consider its generating series

$$\hat{\theta}(z) = \frac{1}{1 - (1 + u)z} - \frac{1}{1 - z}.$$

(c) More generally, we set

$$\hat{\theta}^{(p,q)}(z) = \sum_{n=0}^{\infty} \hat{\theta}_{pn+q} z^n = \frac{(1 + u)^q}{1 - (1 + u)^p z} - \frac{1}{1 - z}.$$

Thus, a series $q(z)$ such that $q_n = \theta_n$ satisfies

$$q(z) \ll \hat{\theta}(z). \tag{12}$$

Similarly, the generating series corresponding to a regularly spaced subsequence of (q_n) (e.g., cumulative errors after every second operation) is bounded by $\hat{\theta}^{(p,q)}$ for appropriate p and q.

The inequality (12) is closely related to that between quantities $|\theta_n|$ and $\gamma_n = nu/(1-nu)$ used extensively in Higham's book, and one has $\hat{\theta}_n \leqslant \gamma_n$ for $nu < 1$. Note however that, compared to [18, Lemma 3.1], we do not allow for negative powers of $1+\delta_i$ in the definition of θ_n.

To rewrite the relation $\tilde{x}_n = x_n(1+\theta_n)$ in terms of generating series, we can use the *Hadamard product* of series, defined by

$$(f \odot g)(z) = \sum_n f_n g_n z^n.$$

An immediate calculation starting from Definition 10 yields a closed-form expression of the Hadamard product with $\hat{\theta}^{(p,q)}$.

Lemma 11. *For any power series $\hat{f}(z)$ with nonnegative coefficients, it holds that*

$$(\hat{\theta}^{(p,q)} \odot \hat{f})(z) = \sum_{n=0}^{\infty} \hat{\theta}_{pn+q} \hat{f}_n z^n = (1+u)^q \hat{f}((1+u)^p z) - \hat{f}(z).$$

8. Variable Coefficients: Legendre Polynomials

With this background in place, we now consider a "real" application involving a recurrence with polynomial coefficients that depend on a parameter. The example is adapted from material previously published in [21, Section 3]. Let P_n denote the Legendre polynomial of index n, defined by

$$\sum_{n=0}^{\infty} P_n(x) z^n = \frac{1}{\sqrt{1-2xz+z^2}}.$$

Fix $x \in [-1,1]$, and let $p_n = P_n(x)$. The classical three-term recurrence

$$p_{n+1} = \frac{1}{n+1}((2n+1)xp_n - np_{n-1}), \qquad n \geq 0 \tag{13}$$

allows us to compute p_n for any n starting from $p_0 = 1$ and an arbitrary $p_{-1} \in \mathbb{R}$. Suppose that we run this computation in fixed-point arithmetic, with an absolute error ε_n at step n, in the sense that the computed values $\tilde{p}_n$ satisfy

$$\tilde{p}_{n+1} = \frac{1}{n+1}((2n+1)x\tilde{p}_n - n\tilde{p}_{n-1}) + \varepsilon_n, \qquad n \geq 0. \tag{14}$$

Proposition 12. *Let $(\tilde{p}_n)_{n\geq -1}$ be a sequence of real numbers satisfying* (14)*, with $\tilde{p}_0 = 1$. Assume that $|\varepsilon_n| \leq \bar{\varepsilon}$ for all n. Then, for all $n \geqslant 0$, the global absolute error satisfies*

$$|\tilde{p}_n - p_n| \leq \frac{(n+1)(n+2)}{4}\bar{\varepsilon}.$$

Proof. Let $\delta_n = \tilde{p}_n - p_n$ and $\eta_n = (n+1)\varepsilon_n$. Subtracting (13) from (14) gives

$$(n+1)\delta_{n+1} = (2n+1)x\delta_n - n\delta_{n-1} + \eta_n, \tag{15}$$

with $\delta_0 = 0$. Note that (15) holds for all $n \in \mathbb{Z}$ if the sequences (δ_n) and (η_n) are extended by 0 for $n < 0$. By Lemma 2, it translates into

$$(1-2xz+z^2)z\frac{\mathrm{d}}{\mathrm{d}z}\delta(z) = z(x-z)\delta(z) + z\eta(z).$$

The solution of this differential equation with $\delta(0) = 0$ reads

$$\delta(z) = p(z) \int_0^z \eta(w)\, p(w)\, \mathrm{d}w, \qquad p(z) = \sum_{n=0}^{\infty} p_n z^n = \frac{1}{\sqrt{1 - 2xz + z^2}}.$$

It is well known that $|P_n|$ is bounded by 1 on $[-1, 1]$, so that $p(z) \ll (1-z)^{-1}$, and the definition of η_n implies $\eta(z) \ll \bar{\varepsilon}(1-z)^{-2}$. It follows by Lemma 5 that

$$\delta(z) \ll \frac{1}{1-z} \int_0^z \frac{\bar{\varepsilon}}{(1-w)^3}\, \mathrm{d}w = \frac{\bar{\varepsilon}}{2(1-z)^3}$$

and therefore $|\delta_n| \leq (n+1)(n+2)\bar{\varepsilon}/4$. □

9. Relative Errors: The Toy Example Revisited

Let us return to the simple recurrence

$$c_{n+1} = 2c_n - c_{n-1} \tag{16}$$

considered in Section 2, but now look at what happens when the computation is performed in floating-point arithmetic, using the observations made in Section 7.

We assume binary floating-point arithmetic with unit roundoff u, and consider the iterative computation of the sequence defined by (16) with $c_{-1} = 0$ and an some exactly representable initial value c_0. The exact solution is $c_n = c_0(n+1)$, that is,

$$c(z) = \frac{c_0}{(1-z)^2}.$$

The floating-point computation produces a sequence of approximations $\tilde{c}_n \approx c_n$ with $\tilde{c}_0 = c_0$. Using the standard model recalled in Section 7 and the fact that multiplication by 2 is exact, the analogue of the local estimate (2) reads

$$\tilde{c}_{n+1} = (2\tilde{c}_n - \tilde{c}_{n-1})(1 + \varepsilon_n), \qquad |\varepsilon_n| \leqslant u. \tag{17}$$

Let $\delta_n = \tilde{c}_n - c_n$. By subtracting $(1 + \varepsilon_n)$ times (16) from (17) and reorganizing, we get

$$\delta_{n+1} - 2\delta_n + \delta_{n-1} = \varepsilon_n(c_{n+1} + 2\delta_n - \delta_{n-1}),$$

which rewrites

$$(z^{-1} - 2 + z)\delta(z) = \varepsilon(z) \odot (z^{-1}c(z) + (2-z)\delta(z)).$$

We multiply this equation by z to get

$$(1-z)^2\delta(z) = (z\varepsilon(z)) \odot (c(z) + z(2-z)\delta(z)).$$

Denote $\gamma(z) = (1-z)^2\delta(z)$ and $b(z) = (1-z)^{-2} - 1$. Since $|\varepsilon_n| \leqslant u$ and $b(z) \gg 0$, we have

$$\gamma(z) = (z\varepsilon(z)) \odot (c(z) + b(z)\gamma(z)) \ll u \left({}^\sharp c(z) + b(z)\,{}^\sharp\gamma(z)\right),$$

and therefore

$${}^\sharp\gamma(z) \ll u \left({}^\sharp c(z) + b(z)\,{}^\sharp\gamma(z)\right)$$

where $b(0) = 0$. By Lemma 6, it follows that

$$\gamma(z) \ll {}^\sharp\gamma(z) \ll {}^\sharp c(z) \frac{u}{1 - ub(z)} = \frac{|c_0|}{(1-z)^2} \frac{(1-z)^2 u}{1 - 2(1+u)z + (1+u)z^2},$$

whence

$$\delta(z) \ll \frac{|c_0|}{(1-z)^2} \frac{u}{(1-\alpha z)(1-\beta z)} =: \hat{\delta}(z) \tag{18}$$

where $\alpha > \beta$ are the roots of $z^2 - 2(1+u)z + (1+u)$.

Remark 13. With no other assumption on the error of subtraction than the first standard model of floating-point arithmetic, the bound (18) is sharp: when $c_0 \geqslant 0$ and $\varepsilon_n \equiv u$, we have $\delta(z) = \hat{\delta}(z)$.

From (18), a trained eye immediately reads off the essential features the bound. Perhaps the most important information is its asymptotic behavior as the working precision increases. For a bound that depends on a problem dimension n, it is customary to focus (sometimes implicitly) on the leading order term as $u \to 0$ for fixed[5] n, and in some cases to further simplify it by looking at its asymptotic behavior for large n.

In the present case, the definition of α as a root of $z^2 - 2(1+u)z + (1+u)$ yields $\alpha = 1 + u^{1/2} + \mathrm{O}(u)$, and hence

$$\hat{\delta}(z) = \frac{|c_0|u}{(1-z)^4} + \mathrm{O}(u^{3/2}) \qquad \text{as } u \to 0, \tag{19}$$

or equivalently

$$\hat{\delta}_n = \frac{1}{6}(n+1)(n+2)(n+3)|c_0|u + \mathrm{O}(u^{3/2}), \qquad u \to 0, \quad n \text{ fixed}.$$

The fact that $\hat{\delta}_n \approx n^3|c_0|u/6$ in the sense that $\lim_{u\to 0}(u^{-1}\hat{\delta}_n) \sim_{n\to\infty} n^3|c_0|u/6$ also follows directly from (19) using Lemma 3. (One also sees that $\hat{\delta}_n = C\alpha^n + \mathrm{O}(n)$ as $n \to \infty$ for fixed u, where C could easily be made explicit. This is however less relevant for our purposes, cf. Remark 14 below.)

In order to obtain a bound that holds for all n, we can majorize $(1-z)^{-1}$ and $(1-\beta z)^{-1}$ by $(1-\alpha z)^{-1}$ in (18). We obtain $\hat{\delta}(z) \ll |c_0|u(1-\alpha z)^{-4}$, and therefore

$$|\delta_n| \leqslant |\hat{\delta}_n| \leqslant |c_0|\frac{(n+1)(n+2)(n+3)}{6}\alpha^n u. \tag{20}$$

Recalling that $c_n = c_0(n+1)$, we conclude that $\tilde{c}_n = c_n(1+\eta_n)$ where

$$|\eta_n| \leqslant \frac{(n+2)(n+3)}{6}\alpha^n u.$$

Note that the exponential growth with n for fixed u is unavoidable, at least under the hypotheses of Remark 13. However, the exponential factor only starts contributing significantly when n becomes extremely large compared to the working precision. It is natural to control it by tying the growth of n to the decrease of u. In particular, it is clear that $\alpha^n = \mathrm{O}(1)$ if $n = \mathrm{O}(u^{-1/2})$. One can check more precisely that $\alpha^n \leqslant e$ as long as $n \leqslant (\alpha-1)^{-1} \approx u^{-1/2}$, and $\alpha^n \leqslant 3$ for all $n \leqslant u^{-1/2}$ provided that $u \leqslant 2^{-7}$.

For an even more precise bound, one could also isolate the leading term of the series expansion of $\hat{\delta}(z)$ with respect to u and reason as above to conclude that $|\eta_n| \leqslant \frac{1}{6}(n+2)(n+3)u + p(n)\alpha^n u^2$ for some explicit $p(n)$.

[5]See Higham's classic book [18] for many examples, and in particular the discussion of linearized bounds at the beginning of Section 3.4. Already in equation 3.7, the implicit assumption $nu < 1$ is not sufficient to ensure that the neglected term is $\mathrm{O}(u^2)$, if n is allowed to grow while u tends to zero.

Remark 14. Starting back from (18), one may be tempted to write the partial fraction decomposition of $\hat{\delta}(z)$ and deduce an exact formula for $\hat{\delta}_n$. Doing so leads to

$$\hat{\delta}_n = (A\alpha^n - B\beta^n - u^{-1}n)|c_0|u, \qquad A = \frac{\alpha^3}{(\alpha-1)^2(\alpha-\beta)}, \quad B = A - 1 - u^{-1}.$$

This expression is misleading, for it involves cancellations between terms that tend to infinity as u goes to zero. In particular, we have $A \sim \frac{1}{2}u^{-3/2}$ as $u \to 0$.

10. Relative Errors, Infinite Order: Scaled Bernoulli Numbers

For a more sophisticated application of the same idea, let us study the floating-point computation of scaled Bernoulli numbers as described by Brent [4, Section 7] (see also Brent and Zimmermann [6, Section 4.7.2]). This section is the first to derive new results.

The scaled Bernoulli numbers are defined in terms of the classical Bernoulli numbers B_k by $b_k = B_{2k}/(2k)!$. Their generating series has a simple explicit expression:

$$b(z) = \sum_{k=0}^{\infty} b_k z^k = \frac{\sqrt{z}/2}{\tanh(\sqrt{z}/2)}. \tag{21}$$

One possible algorithm for computing b_k, suggested by Reinsch according to Brent [4, Section 12], is to follow the recurrence

$$b_k = \frac{1}{(2k)!4^k} - \sum_{j=0}^{k-1} \frac{b_j}{(2k+1-2j)!4^{k-j}}. \tag{22}$$

In [6, Exercise 4.35], it is asked to "prove (or give a plausibility argument for)" the fact that the relative error on b_k when computed using (22) in floating-point arithmetic is $\mathrm{O}(k^2u)$. The $\mathrm{O}(k^2u)$ bound is already mentioned without proof in [4], and again in [5, Section 2]. Paul Zimmermann (private communication, June 2018) suggested that the dependency in k may actually be linear rather than quadratic.

Our goal in this section is to prove a version of the latter conjecture. Like in the previous section, it cannot be true if the $\mathrm{O}(\cdot)$ is interpreted as uniform as $u \to 0$ and $k \to \infty$ independently. It does hold, however, when u and k are restricted to a region where their product is small enough, as well as in the sense that the relative error η_k for fixed k satisfies $|\eta_k| \leqslant C_k u$ when u is small enough, for a sequence C_k which itself satisfies $C_k = \mathrm{O}(k)$ as $k \to \infty$. We will in fact derive a fully explicit, non-asymptotic bound in terms of u and k.

Based on the form of (21), denote $w = \sqrt{z}/2$, and for any $f \in \mathbb{C}[[z]]$, define f^* by $f^*(w) = f(4w^2)$, so that $f^*(w) = f(z)$. In particular, we have $b^*(w) = w/\tanh w$. We will use the following classical facts about the numbers $|b_k|$.

Lemma 15. *The absolute values of the scaled Bernoulli numbers satisfy*

$$\sharp b^*(w) = 2 - \frac{w}{\tan w}, \qquad |b_k| \sim_{k\to\infty} \frac{2}{(2\pi)^{2k}}, \qquad \frac{2}{(2\pi)^{2k}} \leqslant |b_k| \leqslant \frac{4}{(2\pi)^{2k}}$$

where the last formula assumes $k \geqslant 1$.

Proof. The expression of $\sharp b^*(w)$ can be deduced from that of $b^*(w)$ and the fact that B_{2n} has sign $(-1)^{n+1}$ for $n \geqslant 1$, using the relation $\tanh(iw) = i\tan(w)$. The other

statements follow from the expression of Bernoulli numbers using the Riemann zeta function [13, (25.6.2)]. □

Let $\tilde{b}_k$ denote the approximate value of b_k computed using (22). We assume that the computed value of $n!$ is equal to $n!/(1+q_n)$ with $q_n = \theta_{n-2}$, in the notation of Section 7, for $n \geqslant 2$, and $q_0 = q_1 = 1$. According to the modified standard model of floating-point arithmetic, this holds true if $n!$ is computed as $(((2\times 3)\times 4)\times\cdots\times n)$ and the working precision is at least $\lceil \log_2 n \rceil$, taking into account that the first multiplication is exact since we are assuming binary floating-point arithmetic. The local error introduced by one step of the iteration (22) then behaves as follows.

Lemma 16. *At every step of the iteration* (22)*, the computed value* $\tilde{b}_k$ *has the form*

$$\tilde{b}_k = \frac{1+s_k}{(2k)!4^k} - \sum_{j=0}^{k-1} \frac{\tilde{b}_j(1+t_{k,j})}{(2k+1-2j)!4^{k-j}}, \qquad |s_k| \leqslant \hat{\theta}_{2k}, \quad |t_{k,j}| \leqslant \hat{\theta}_{3(k-j)+2}. \tag{23}$$

Proof. The computation of b_0 involves no rounding error. Assume $k \geqslant 1$, and first consider the term $a_k = 1/((2k)!4^k)$ outside the sum. By assumption, the computed value of $(2k)!$ is $(2k)!/(1+q_{2k})$, and the multiplication by 4^k that follows is exact. Inverting the result introduces an additional rounding error. The computed value of the whole term is hence $a_k(1+r'_k)$ where $r'_k = \theta_{2k-1}$. By the same reasoning, the term of index j in the sum is computed with a relative error $r_{k,j} = \theta_{2(k-j)+1}$ for all j, k. If $v_{k,i}$ denotes the relative error in the addition of the term of index i to the partial sum for $0 \leqslant j \leqslant i-1$, the computed value of the sum is hence

$$\sum_{j=0}^{k-1} \frac{\tilde{b}_j(1+r_{k,j})}{(2k+1-2j)!4^{k-j}} \prod_{i=j}^{k-1}(1+v_{k,i}).$$

Taking into account the relative error v'_k of the final subtraction leads us to (23), with

$$1+s_k = (1+r'_k)(1+v'_k), \qquad 1+t_{k,j} = (1+r_{k,j})(1+v'_k)\prod_{i=j}^{k-1}(1+v_{k,i}). \qquad \square$$

Let $\delta_k = \tilde{b}_k - b_k$. Comparison of (23) with (22) yields

$$\delta_k = \frac{s_k}{(2k)!4^k} - \sum_{j=0}^{k-1} \frac{\delta_j + \tilde{b}_j t_{k,j}}{(2k+1-2j)!4^{k-j}},$$

which rearranges into

$$\sum_{j=0}^{k} \frac{\delta_j}{(2k+1-2j)!4^{k-j}} = \frac{s_k}{(2k)!4^k} - \sum_{j=0}^{k-1} \frac{(b_j+\delta_j)t_{k,j}}{(2k+1-2j)!4^{k-j}}. \tag{24}$$

Using the bounds from Lemma 16, with $\hat{\theta}_{3(k-j)+2}$ replaced by $\hat{\theta}_{4(k-j)+2}$ to obtain slightly simpler expressions later, it follows that

$$\left|\sum_{j=0}^{k} \frac{\delta_j}{(2k+1-2j)!4^{k-j}}\right| \leqslant \frac{\hat{\theta}_{2k}}{(2k)!4^k} + \sum_{j=0}^{k-1} \frac{(|b_j|+|\delta_j|)\hat{\theta}_{4(k-j)+2}}{(2k+1-2j)!4^{k-j}}. \tag{25}$$

Let us introduce the auxiliary series

$$C(z) = \sum_{k=0}^{\infty} \frac{z^k}{(2k)!4^k} = \cosh w, \quad S(z) = \sum_{k=0}^{\infty} \frac{z^k}{(2k+1)!4^k} = \frac{\sinh w}{w}, \quad \check{S}(z) = \frac{w}{\sin w},$$

$$\tilde{C}(z) = \hat{\theta}^{(2,0)}(z) \odot C(z), \quad \tilde{S}(z) = \hat{\theta}^{(4,2)}(z) \odot (S(z) - 1).$$

The inequality (25) (note that the sum on the right-hand side stops at $k-1$) translates into

$$\delta(z)S(z) \ll \tilde{C}(z) + \tilde{S}(z)\left({}^\sharp b(z) + {}^\sharp \delta(z)\right).$$

Since $\check{S}(z)$ has nonnegative coefficients [13, (4.19.4)], we have $S(z)^{-1} = iw/\sin(iw) \ll \check{S}(z)$, hence

$$\delta(z) \ll \check{S}(z)\tilde{C}(z) + \check{S}(z)\tilde{S}(z)^\sharp b(z) + \check{S}(z)\tilde{S}(z)^\sharp \delta(z).$$

As $\tilde{S}(z) = \mathrm{O}(z)$, Lemma 6 applies and yields

$$\delta(z) \ll \hat{\delta}(z) := \frac{\check{S}(z)\tilde{C}(z) + \check{S}(z)\tilde{S}(z)^\sharp b(z)}{1 - \check{S}(z)\tilde{S}(z)}. \tag{26}$$

Using Lemma 11 to rewrite the Hadamard products, we have

$$\tilde{C}(z) = C(a^2 z) - C(z) = \cosh(aw) - \cosh(w), \tag{27}$$

$$\tilde{S}(z) = a^2 S(a^4 z) - S(z) - (a^2 - 1) = \frac{\sinh(a^2 w) - \sinh(w)}{w} - (a^2 - 1), \tag{28}$$

where $a = 1 + u$. In addition, Lemma 15 gives a formula for ${}^\sharp b(z)$. Thus (26) yields an explicit, if complicated, majorant series for $\delta(z)$.

It is not immediately clear how to extract a readable bound on δ_k in the style of (20). Yet we can already prove an asymptotic version of Zimmermann's observation. The calculations leading to Propositions 17 and 19 can be checked with the help of a computer algebra system. A worksheet that illustrates how to do that using Maple is provided in the supplementary material.

Proposition 17. *When k is fixed and for small enough u, the relative error $\eta_k = \delta_k/b_k$ satisfies $|\eta_k| \leqslant C_k u$ for some C_k. In addition, the constants C_k can be chosen such that $C_k = \mathrm{O}(k)$ as $k \to \infty$.*

Proof. As u tends to zero, we have

$$\begin{aligned}\hat{\delta}(z) &= \left(\frac{2(1-\cosh w)\cos(w)}{w^{-2}\sin(w)^2} + \frac{4(\cosh w - 1) + w\sinh w}{w^{-1}\sin w}\right) u + \mathrm{O}(u^2) \\ &=: \hat{\xi}(z)u + \mathrm{O}(u^2),\end{aligned}$$

where the coefficients are to be interpreted as formal series in z. Hence, for fixed k, the error δ_k is bounded by $\hat{\xi}_k u + \mathrm{O}(u^2)$ as $u \to 0$. The function $\hat{\xi}^*(w)$ is meromorphic, with double poles at $w = \pm\pi$, corresponding for $\hat{\xi}(z)$ to a unique pole of minimal modulus (also of order two) at $z = 4\pi^2$. This implies (by Lemma 3) that $(2\pi)^{2k}\hat{\xi}_k = \mathrm{O}(k)$ as $k \to \infty$. The result follows using the growth estimates from Lemma 15. □

In other words, there exist a constant A and a function R such that $\eta_k \leqslant Aku + R(k,u)$, where $R(k,u) = o(u)$ as $u \to 0$ for fixed k, but $R(k,u)$ might be unbounded if k tends to infinity while u tends to zero. (Remark 21 below shows that an exponential dependency in k is in fact unavoidable.) To get a bound valid for all u and k in a reasonable region, let us study the denominator of (26) more closely[6].

Lemma 18. *For small enough $u \geqslant 0$, the function $h : w \mapsto \tilde{S}^*(w) - \check{S}^*(w)^{-1}$ has exactly two simple zeros $\pm\alpha = \pm\pi/(1+\varphi(u))$ closest to the origin, with*

$$\varphi(u) = 2(\cosh(\pi) - 1)u + \mathrm{O}(u^2), \quad u \to 0. \tag{29}$$

Furthermore, if $u \leqslant 2^{-16}$, then one has $0 \leqslant \varphi(u) \leqslant 2(\cosh(\pi)-1)u$, and h has no other zero than $\pm\alpha$ in the disk $|w| < \rho := 6.2$.

Proof. To start with, observe that when $u = 0$, the term $\tilde{S}^*(w)$ in the definition of h vanishes identically, leaving us with $h(w) = \check{S}^*(w)^{-1} = w^{-1}\sin w$. The zeros of $1/\check{S}^*$ closest to the origin are located at $w = \pm\pi$, the next closest, at $w = \pm 2\pi$ and hence outside the disk $|w| < \rho$.

Let us focus on the zero at $w = \pi$. Since $h'(\pi) = 1/\pi$ for $u = 0$, the Implicit Function Theorem applies and shows that, locally, the zero varies analytically with u. One obtains the asymptotic form (29) by implicit differentiation.

We turn to the bounds on $\varphi(u)$. The power series expansion of $\tilde{S}^*$ with respect to w has nonnegative coefficients, showing that $h(\pi) = \tilde{S}^*(\pi) \geqslant 0$. Now, with $a = 1 + u$ as in (28), isolate the first nonzero term of that series and write

$$\tilde{S}^*(w) = \frac{a^6-1}{6}w^2 + \frac{G(a^2w) - G(w)}{w}, \qquad G(w) = \sinh(w) - w - \frac{w^3}{6}.$$

Let $K = 2(\cosh(\pi)-1)$ and $w_0 = \pi/(1+Ku)$. One has

$$G(a^2w_0) - G(w_0) \leqslant (a^2-1)w_0 \max_{1\leqslant t\leqslant a^2} |G'(tw_0)|. \tag{30}$$

Note that $a^2 w_0 \leq \pi$ for all $u \leqslant 1/2$. Since $G(w) \gg 0$, it follows that

$$G(a^2w_0) - G(w_0) \leqslant (a^2-1)w_0 G'(\pi),$$

therefore the term $\tilde{S}^*(w_0)$ in $h(w_0)$ satisfies

$$\tilde{S}^*(w_0) \leqslant \frac{a^6-1}{6}w_0^2 + (a^2-1)G'(\pi).$$

As for the other term, the inequality $\sin(\pi - v) \geqslant v - v^3/6$ $(u \geq 0)$ applied to $v = \pi - w_0 = Kuw_0$ yields

$$-\check{S}^*(w_0)^{-1} \leqslant -Ku + (Ku)^3\frac{w_0^2}{6}.$$

Collecting both contributions and substituting in the value $G'(\pi) = (K-\pi^2)/2$, we obtain

$$h(w_0) \leqslant \hat{h} := \frac{a^6-1}{6}\frac{\pi^2}{(1+Ku)^2} + \frac{a^2-1}{2}(K-\pi^2) - Ku + (Ku)^3\frac{w_0^2}{6}.$$

The right-hand side is an explicit rational function of u, satisfying

$$\hat{h} \sim (-2\pi^2(K-1) + K/2)u^2$$

[6]A similar argument applied to $\hat{\xi}$ in the place of $\hat{\delta}$ would make the constant A explicit.

as $u \to 0$. One can check that it remains negative for $0 < u \leqslant 2^{-16}$. Thus, one has $h(w_0) < 0 \leqslant h(\pi)$, and h has a zero in the interval $w \in [w_0, \pi]$, that is, a zero of the form $\alpha = \pi/(1+\varphi(u))$ with $0 \leqslant \varphi(u) \leqslant Ku$, as claimed. The corresponding statement for $-\alpha$ follows by parity.

It remains to show that $\pm\alpha$ are the only zeros of h in the disk $|w| < \rho$. We do it by comparing them to the zeros of $1/\check{S}^*$ using Rouché's theorem.

To this end, note first that the expression $|\sin(\rho e^{i\theta})|$ reaches its minimum for $\theta \in [0, \pi/2]$ when $\theta = 0$. Indeed, one has $|\sin(\rho e^{i\theta})|^2 = \cosh(\rho\sin\theta)^2 - \cos(\rho\cos\theta)^2$. The term $\cosh(\rho\sin\theta)^2$ is strictly increasing on the whole interval. For $\theta \leqslant \theta_0 := \arccos(3\pi/(2\rho))$, the term $-\cos(\rho\cos\theta)^2$ is increasing as well, so that $|\sin(\rho e^{i\theta})| \geqslant |\sin\rho|$ in that range, whereas for $\theta_0 \leqslant \theta \leqslant \pi$, we have $|\sin(\rho e^{i\theta})|^2 \geqslant \cosh(\rho\sin\theta_0)^2 - 1 > 1 \geqslant |\sin\rho|^2$. It follows that $|\sin w| \geqslant |\sin\rho|$, and hence $|\check{S}^*(w)^{-1}| \geqslant \rho^{-1}|\sin\rho|$, for $w = \rho e^{i\theta}$ with $0 \leqslant \theta \leqslant \pi/2$, and by symmetry on the whole circle $|w| = \rho$.

Turning to $\tilde{S}^*(w)$ and reasoning as in (30) starting from (28), one can write

$$|\tilde{S}^*(w)| \leqslant (a^2-1) \sup_{1 \leqslant t \leqslant a^2} |\cosh(tw) - 1| \leqslant (a^2-1)\cosh(a^2\rho)$$

for $|w| = \rho$. Comparing these bounds leads to

$$|\tilde{S}^*(w)| < 10^{-2} < |\check{S}^*(w)^{-1}|, \qquad |w| = \rho, \quad 0 \leqslant u \leqslant 2^{-16}.$$

Therefore, $h(w)$ has the same number of zeros as $w^{-1}\sin w$ inside the circle. □

Proposition 19. *For all $0 < u \leqslant 2^{-16}$, we have $\tilde{b}_k = b_k(1+\eta_k)$ where*

$$|\eta_k| \leqslant (1+21.2u)^k(1.1k+446)u.$$

Proof. We use the notation of the previous lemma. In the expression (26) of $\hat{\delta}(z)$, the series $\tilde{S}(z)$ and $\tilde{C}(z)$ define entire functions, while $\sharp b^*(w) = 2 - w/\tan w$ is a meromorphic function with poles at $w \in \pi\mathbb{Z}$. These observations combined with Lemma 18 imply that $\hat{\delta}^*$ is meromorphic in the disk $|w| < 6.2$, with exactly four simple poles located at $w = \pm\alpha$ and $w = \pm\pi$. Only the first two poles depend on u.

One has $\hat{\delta}^*(w) \sim -(1-w/\pi)^{-1}$ as $w \to \pi$. Let F^* denote the derivative of $w \mapsto w\tilde{S}^*(w)$. With the help of a computer algebra system, it is not too hard to determine that the singular expansion as $w \to \alpha$ reads

$$\hat{\delta}^*(w) \sim \frac{\tilde{C}^*(\alpha^2) - \cos\alpha + 2\alpha^{-1}\sin\alpha}{F^*(\alpha) - \cos\alpha} \frac{1}{1-w/\alpha} =: \frac{R(u)}{1-w/\alpha}$$

and one has $R(u) = 1 + (2\cosh\pi - 2 - \pi\sinh\pi)u + \mathrm{O}(u^2)$ as $u \to 0$. The expansions at $-\alpha$ and $-\pi$ follow since $\hat{\delta}^*$ is an even function. Set

$$\hat{\delta}^*(w) = \frac{R(u)}{1-w/\alpha} + \frac{R(u)}{1+w/\alpha} - \frac{1}{1-w/\pi} - \frac{1}{1+w/\pi} + g^*(u,w),$$

where $g^*(u,\cdot)$ now is analytic for $|w| < 6.2 \approx 1.97\pi$ and vanishes identically when $u = 0$. Since

$$\frac{R(u)}{1\pm w/\alpha} - \frac{1}{1\pm w/\pi} = \frac{R(u)-1}{1\pm w/\alpha} \mp \frac{\varphi(u)w/\pi}{(1\pm w/\alpha)(1\pm w/\pi)},$$

we have

$$\hat{\delta}^*(w) = \frac{2(R(u)-1)}{1-(w/\alpha)^2} + \frac{2\varphi(u)(2+\varphi(u))(w/\pi)^2}{(1-(w/\alpha)^2)(1-(w/\pi)^2)} + g^*(u,w),$$

that is,

$$\hat{\delta}(z) = \frac{2(R(u)-1)}{1-z/(2\alpha)^2} + \frac{2\varphi(u)(2+\varphi(u))z/(2\pi)^2}{(1-z/(2\alpha)^2)(1-z/(2\pi)^2)} + g(u,z).$$

By Cauchy's inequality, for any $\lambda < 1.9$, the Taylor coefficients g_k of $g(u,\cdot)$ satisfy

$$|g_k| \leqslant \frac{A_\lambda(u)}{(2\pi\lambda)^{2k}}, \qquad A_\lambda(u) := \max_{|w|=\lambda\pi} |g^*(u,w)|,$$

and therefore

$$\hat{\delta}(z) \ll \frac{2|R(u)-1|}{1-z/(2\alpha)^2} + \frac{2\varphi(u)(2+\varphi(u))z/(2\alpha)^2}{(1-z/(2\alpha)^2)^2} + \frac{A_\lambda(u)}{1-z/(2\pi\lambda)^2}, \tag{31}$$

where we have bounded $\nu z/(1-\nu z)$ where $\nu = 1/(2\pi)^2$ by the same expression with $\nu = 1/(2\alpha)^2$.

For $0 \leqslant u \leqslant 2^{-16}$ and using the enclosure of $\varphi(u)$ from Lemma 18, a brute force evaluation using interval arithmetic yields the bounds

$$|R(u)-1| \leqslant (\max_{0\leqslant v\leqslant u} |R'(v)|)u \leqslant 72u, \qquad \frac{2\varphi(u)(2+\varphi(u))}{(2\alpha)^2} \leqslant 2.2u.$$

SageMath code for computing these estimates can be found in the supplementary material. By the same method, choosing $\lambda = \sqrt{3/2}$ and making use of the fact that $g^*(0,w) = 0$, we get

$$A_\lambda(u) \leqslant u \max_{\substack{0\leqslant v\leqslant u \\ |w|=\lambda\pi}} \left|\frac{\partial g^*}{\partial u}(v,w)\right| \leqslant 747u.$$

We substitute these bounds in (31) to conclude that

$$\begin{aligned}
\hat{\delta}_k &\leqslant (2\times 72(2\alpha)^{-2k} + 2.2k(2\alpha)^{-2k} + 747(2\pi\lambda)^{-2k})u \\
&\leqslant \frac{1}{(2\pi)^{2k}}((2.2k+2\times 72)(1+2(\cosh\pi-1)u)^{2k} + 747(2/3)^k)u \\
&\leqslant \frac{(1+21.2u)^{2k}}{(2\pi)^{2k}}(2.2k+891)u.
\end{aligned}$$

The claim follows since, as noted in Lemma 15, $|b_k| \geqslant 2(2\pi)^{-2k}$ for all $k \geq 1$. □

Corollary 20. *For all u and k satisfying $0 < u \leqslant 2^{-16}$ and $43ku \leqslant 1$, one has $\tilde{b}_k = b_k(1+\eta_k)$ with $|\eta_k| \leqslant (3k+1213)u$.*

Proof. The assumption on ku implies $(1+21.2u)^{2k} \leqslant e$. □

Remark 21. Let us now prove our claim that no bound of the form $\eta_k = \mathrm{O}(ku)$ can hold uniformly with respect to u and k. In the notation of Lemma 16, suppose that we have $s_k = u$ and $t_{k,j} = \hat{\theta}_{k-j}$ for all k, j. These values are reached, e.g., by taking $v_{k,i} = u$ and $v'_k = r_{k,j} = r'_k = 0$ for all i, j, k. Since the $v_{k,i}$ and v'_k represent individual rounding errors, this corresponds to a feasible situation in our model. Then, (24) translates into $\delta(z)S(z) = C(z)u - (\hat{\theta}\odot S)(z)(b(z)+\delta(z))$, that is,

$$\delta(z) = \frac{C(z)u - (S((1+u)z) - S(z))b(z)}{S((1+u)z)}.$$

For small u, the denominator vanishes at $z = \beta := -4\pi^2/(1+u)$, while the numerator is analytic for $|w| < \pi$ and does not vanish at β (it tends to -1 as $u \to 0$, since

$S(\beta) \sim u/2$ and $\delta(\beta) \sim -2u^{-1}$). Thus, the radius of convergence of $\delta(z)$ is at most $4\pi^2/(1+u)$. This implies that η_k grows exponentially for fixed u.

11. Two Variables: The Equation of a Vibrating String

As part of an interesting case study of the use of formal program verification in scientific computing, Boldo *et al.* [1, 2, 3] give a full worst-case rounding error analysis of a simple explicit finite difference scheme for the one-dimensional wave equation

$$\frac{\partial^2 p}{\partial t^2} - c^2 \frac{\partial^2 p}{\partial x^2} = 0. \tag{32}$$

Such a finite difference scheme is nothing but a multi-dimensional linear recurrence—in the present case, a two-dimensional one, with a time index ranging over the natural numbers while the space index is restricted to a finite domain. We rephrase the relatively subtle error analysis using a slight extension of the language introduced in the previous sections. Doing so does not change the essence of the argument, but possibly makes it more palatable.

We use the notation and assumptions of [2, Section 3.2], [3, Section 5], except that we flip the sign of δ_i^k to keep with our usual conventions. Time and space are discretized into a grid with time step Δt and space step Δx. To the continuous solution $p(x,t)$ corresponds a double sequence (p_i^k) where $k \geqslant 0$ is the space index and $0 \leqslant i \leqslant n$ is the time index. Taking central differences for the derivatives in (32) and letting $a = (c\Delta t/\Delta x)^2$ leads to

$$\begin{cases} p_i^1 = p_i^0 + \dfrac{a}{2}(p_{i+1} - 2p_i + p_{i-1}), \\ p_i^{k+1} = 2p_i^k - p_i^{k-1} + a(p_{i+1}^k - 2p_i^k + p_{i-1}^k), \quad k \geqslant 1. \end{cases} \tag{33}$$

The problem is subject to the boundary conditions $p_0^k = p_n^k = 0$, $k \in \mathbb{N}$, and we are given initial data $(p_i^0)_{i=1}^{n-1}$. In accordance with the Courant-Friedrichs-Lewy condition, we assume $0 < a \leqslant 1$.

Boldo *et al.* study an implementation of (33) in floating-point arithmetic, but their focus is on the propagation of absolute errors. Their local error analysis shows that the computed values $(\tilde{p}_i^k)$ corresponding to (p_i^k) satisfy[7]

$$\begin{cases} \tilde{p}_i^0 = p_i^0 + \delta_i^0 \\ \tilde{p}_i^1 = \tilde{p}_i^0 + \dfrac{a}{2}(\tilde{p}_{i+1}^0 - 2\tilde{p}_i^0 + \tilde{p}_{i-1}^0) + \delta_i^1 - \left(\delta_i^0 + \dfrac{a}{2}(\delta_{i+1}^0 - 2\delta_i^0 + \delta_{i-1}^0)\right), \\ \tilde{p}_i^{k+1} = 2\tilde{p}_i^k - \tilde{p}_i^{k-1} + a(\tilde{p}_{i+1}^k - 2\tilde{p}_i^k + \tilde{p}_{i-1}^k) + \delta_i^{k+1}, \qquad k \geqslant 1 \end{cases} \tag{34}$$

with $|\delta_i^k| \leqslant \bar{\delta} := 78 \cdot 2^{-52}$ for all i and all $k \geqslant 1$. While the discussion of error propagation in [2, 3] assumes $\delta_i^0 = 0$, the machine-checked proof actually allows for initial errors $|\delta_i^0| \leqslant \bar{\delta}^0 := 14 \cdot 2^{-52}$ and gives the additional bound $|\delta_i^1| \leqslant \bar{\delta}^1 := 81 \cdot 2^{-53}$.

The iterations (33) and (34) are, a priori, valid for $0 < i < n$ only, but it is not hard to see that they can be made to hold for all $i \in \mathbb{Z}$ by extending the sequences (p_i^k), $(\tilde{p}_i^k)$, and (δ_i^k) by odd symmetry and $(2n)$-periodicity with respect

[7] The correcting term involving δ_i^0 in the expression of $\tilde{p}_i^1$ will help make the expression of the overall error more uniform.

to i. Viewing time as the main variable, we encode space-periodic sequences of period $2n$ by generating series of the form

$$f(x,t)=\sum_{k=0}^{\infty}\sum_{i=-n}^{n-1} f_i^k x^i t^k \in \Omega[[t]] \qquad \text{where } \Omega=\mathbb{R}[x]/\langle x^{2n}-1\rangle.$$

When f is an element of $\Omega[[t]]$, we denote

$$f^{u:}(x,t)=\sum_{k\geqslant u}\sum_{i} f_i^k x^i t^k, \qquad f_i(t)=\sum_k f_i^k t^k, \qquad f^k(x)=\sum_i f_i^k x^i.$$

Multiplication by x and t in $\Omega[[t]]$ respectively correspond to backward shifts of the indices i and k.

Let $\Delta(x,t)=\tilde{p}(x,t)-p(x,t)$ be the generating series of the global error. Our goal is to obtain bounds on the Δ_i^k. We start by expressing $\Delta(x,t)$ in terms of $\delta(x,t)$. This is effectively a more precise version of [3, Theorem 5.1] covering initial data with numeric errors.

Proposition 22. *One has*

$$\Delta(x,t)=\lambda(x,t)\eta(x,t) \tag{35}$$

where

$$\lambda(x,t)=\frac{1}{1-\varphi(x)t+t^2}, \quad \eta(x,t)=\delta(x,t)-\varphi(x)t\delta_0(x) \tag{36}$$

with

$$\varphi(x)=2+a(x^{-1}-2+x).$$

Proof. By comparing (34) with (33) and observing how Δ_i^1 simplifies thanks to the correcting term in $\tilde{p}_i^1$, we get

$$\begin{cases} \Delta_i^k=\delta_i^k, & k=0,1,\\ \Delta_i^{k+1}=2\Delta_i^k-\Delta_i^{k-1}+a(\Delta_{i+1}^k-2\Delta_i^k+\Delta_{i-1}^k)+\delta_i^{k+1}, & k+1\geqslant 2. \end{cases} \tag{37}$$

In terms of series, (37) translates into

$$\Delta^0(x)=\delta^0(x), \qquad \Delta^1(x)=\delta^1(x), \tag{38}$$

$$t^{-1}\Delta^{2:}(x,t)=2\Delta^{1:}(x,t)-t\Delta(x,t)+a(x^{-1}-2+x)\Delta^{1:}(x,t)+t^{-1}\delta^{2:}(x,t). \tag{39}$$

Taking into account (38), equation (39) becomes:

$$\Delta(x,t)=2t(\Delta(x,t)-\delta_0(x))-t^2\Delta(x,t)+a(x^{-1}-2+x)t(\Delta(x,t)-\delta^0(x))+\delta(x,t),$$

that is,

$$(1-(2+a(x^{-1}-2+x))t+t^2)\Delta(x,t)=\delta(x,t)-\varphi(x)t\delta_0(x)$$

The coefficient on the left-hand side is invertible in $\Omega[[t]]$, leading to $\Delta(x,t)=\lambda(x,t)\eta(x,t)$. □

Rather than by bivariate majorant series, we will control elements of $\Omega[[t]]$ by majorant series in a single variable relative to a norm on Ω. We say that $\hat{f}\in\mathbb{R}_{\geqslant 0}[[t]]$ is a majorant series of $f\in A[[t]]$ with respect to a norm $\|\cdot\|_s$ on an algebra A, and write $f\ll_s \hat{f}$, when $\|f_k\|_s\leqslant \hat{f}_k$ for all $k\in\mathbb{N}$. The basic properties listed in Section 6 extend in the obvious way. In particular, if $\|\cdot\|_q, \|\cdot\|_r, \|\cdot\|_s$ are norms such that $\|uv\|_q\leqslant\|u\|_r\|v\|_s$, then $f\ll_r\hat{f}$ and $g\ll_s\hat{g}$ imply $fg\ll_q\hat{f}\hat{g}$.

For $u = \sum_i u_i x^i \in \Omega$, we define

$$\|u\|_1 = \sum_{i=-n}^{n-1} |u_i|, \qquad \|u\|_2 = \left(\sum_{i=-n}^{n-1} |u_i|^2\right)^{1/2}, \qquad \|u\|_\infty = \max_{i=-n}^{n-1} |u_i|.$$

Note the inequality $\|uv\|_\infty \leqslant \|u\|_1 \|v\|_\infty$ for $u, v \in \Omega$ (an instance of Young's convolution inequality).

The local error analysis yields

$$\delta(x,t) \ll_\infty \bar{\delta}^0 + \bar{\delta}^1 t + \frac{\bar{\delta} t^2}{1-t}.$$

Using the notation of Proposition 22, one has $\|\varphi(x)\|_1 \leqslant 2$, hence $\|\varphi(x)\delta^0(x)\|_\infty \leqslant 2\bar{\delta}^0$ and

$$\eta(x,t) \ll_\infty \bar{\delta}^0 + (\bar{\delta}^1 + 2\bar{\delta}_0)t + \frac{\bar{\delta} t^2}{1-t} \ll \frac{\bar{\delta}}{1-t}. \tag{40}$$

We first deduce a bound on the global error in quadratic mean with respect to space. We will later get a second proof of this result as a corollary of Proposition 25; the main interest of the present one is that it does not rely on Lemma 24.

Proposition 23. *One has*

$$\Delta(x,t) \ll_2 \frac{\sqrt{2n}}{(1-t)^3}\bar{\delta}.$$

In other words, the root mean square error at time k satisfies

$$\left(\frac{1}{n}\sum_{i=0}^{n-1} (\Delta_i^k)^2\right)^{1/2} \leqslant \frac{(k+1)(k+2)}{2}\bar{\delta}.$$

Proof. Elements of Ω can be evaluated at $2n$-th roots of unity, and the collection $u^* = (u(\omega))_{\omega^{2n}=1}$ of values of a polynomial $u \in \Omega$ is nothing but the discrete Fourier transform of its coefficients. The coefficientwise Fourier transform of formal power series,

$$f(t) \mapsto f^*(t) = (f(\omega, t))_{\omega^{2n}=1},$$

is an algebra homomorphism from $\Omega[[t]]$ to $\mathbb{C}^{2n}[[t]]$. One easily checks Parseval's identity for the discrete Fourier transform:

$$\|u\|_2 = \frac{1}{\sqrt{2n}}\|u^*\|_2, \qquad u \in \Omega,$$

where the norm on the right-hand side is the standard Euclidean norm on $\mathbb{C}^{2n}$.

The uniform bound (40) resulting from the local error analysis implies

$$\eta(x,t) \ll_2 \frac{\sqrt{2n}}{1-t}\bar{\delta},$$

and Parseval's identity yields

$$\eta^*(t) \ll_2 \frac{2n\bar{\delta}}{1-t}.$$

At $x = \omega = e^{\mathrm{i}\theta}$, the factor λ in (35) takes the form

$$\lambda(\omega, t) = \frac{1}{1 - 2b_\omega t + t^2}, \qquad b_\omega = 1 + a(\cos\theta - 1),$$

where $-1 \leqslant b_\omega \leqslant 1$ due to the assumption that $0 < a \leqslant 1$. The denominator therefore factors as $(1-\zeta_\omega t)(1-\bar{\zeta}_\omega t)$ where $|\zeta_\omega| = 1$, so that we have

$$\lambda(\omega, t) = \frac{1}{(1-\zeta_\omega t)(1-\bar{\zeta}_\omega t)} \ll \frac{1}{(1-t)^2}$$

in $\mathbb{C}[[t]]$ and hence $\lambda^*(t) \ll_\infty (1-t)^{-2}$.

Since the entrywise product in $\mathbb{C}^{2n}$ satisfies $\|u^* v^*\|_2 \leqslant \|u^*\|_\infty \|v^*\|_2$, the bounds on λ^* and η^* combine into

$$\lambda^*(t)\eta^*(t) \ll_2 \frac{2n\bar{\delta}}{(1-t)^3}.$$

Using Parseval's identity again, we conclude that

$$\Delta(x,t) = \lambda(x,t)\eta(x,t) \ll_2 \frac{\sqrt{2n}\bar{\delta}}{(1-t)^3}$$

as claimed. The second formulation of the result comes from the symmetry of the data: one has $\Delta(x^{-1}, t) = -\Delta(x,t)$, hence $\|\Delta^k\|_2^2 = 2\sum_{i=0}^{n-1} (\Delta_i^k)^2$ for all k. □

Proposition 23 immediately implies $\Delta(x,t) \ll_\infty \bar{\delta}\sqrt{2n}(1-t)^{-3}$. However, this estimate turns out to be too pessimistic by a factor $\sqrt{2n}$. The key to better bounds is the following lemma, proved (with the help of generating series!) in Appendix C of [3]. The argument, due to M. Kauers and V. Pillwein, reduces the problem to a an inequality of Askey and Gasper via an explicit expression in terms of Jacobi polynomials that is proved using Zeilberger's algorithm. It remains intriguing to find a more direct way to derive the uniform bound on Δ.

Lemma 24. *The coefficients λ_i^k of $\lambda(x,t)$ are nonnegative.*

Strictly speaking, the nonnegativity result in [3] is about the coefficients, not of $\lambda(x,t) \in \Omega[[t]]$ as defined above, but of its lift to $\mathbb{R}[x, x^{-1}][[t]]$ obtained by interpreting (36) in the latter ring. It is thus slightly stronger than the above lemma.

From this lemma it is easy to deduce a more satisfactory bound on the global error, matching that of [3, Theorem 5.2].

Proposition 25. *One has the bound*

$$\Delta(x,t) \ll_\infty \frac{\bar{\delta}}{(1-t)^3},$$

that is,

$$|\Delta_i^k| \leqslant \frac{1}{2}\bar{\delta}(k+1)(k+2)$$

for all i and k.

Proof. Lemma 24 implies $\lambda(x,t) \ll_1 \lambda(1,t)$. This bound combines with (40) to give

$$\Delta(x,t) = \lambda(x,t)\eta(x,t) \ll_\infty \lambda(1,t)\frac{\bar{\delta}}{(1-t)} = \frac{\bar{\delta}}{(1-t)^3},$$

just like in the proof of Proposition 23. □

12. Solutions of Linear Differential Equations

For the last application, we return to recurrences of finite order in a single variable. Instead of looking at a specific sequence, though, we now consider a general class of recurrences with polynomial coefficients. It is technically simpler and quite natural to restrict our attention to recurrences associated to nonsingular differential equations under the correspondence from Lemma 2: thus, we consider a linear ordinary differential equation

$$p_r(z)y^{(r)}(z) + \cdots + p_1(z)y'(z) + p_0(z)y(z) = 0, \tag{41}$$

where $p_0, \dots, p_r \in \mathbb{C}[z]$ and assume that $p_r(0) \neq 0$. We expect that this assumption could be lifted by working along the lines of [27].

It is classical that (41) then has r linearly independent formal power series solutions and all these series are convergent in a neighborhood of 0. Suppose that we want to evaluate one of these solutions at a point lying within its disk of convergence. A natural way to proceed is to sum the series iteratively, using the associated recurrence to generate the coefficients. Our goal in this section is to give an error bound on the approximation of the partial sum[8] computed by a version of this algorithm.

We formulate the computation as an algorithm based on interval arithmetic that returns an enclosure of the partial sum. Running the whole loop in interval arithmetic would typically lead to enclosures of width that growths exponentially with the number of computed terms, and thus to a catastrophic loss of precision. Instead, the algorithm executes the body of the loop in interval arithmetic, which saves us from going into the details of the local error analysis, but "squashes" the computed interval to its midpoint after each loop iteration. It maintains a running bound on the discarded radii that serves to control the overall effect of the propagation of local errors. This way of using interval arithmetic does not create long chains of interval operations depending on each other and only produces a small overestimation.

The procedure is presented as Algorithm 1. In its description and the analysis that follows, we use the notation of Section 4 for differential and recurrence operators, with the symbol ∂ denoting $\mathrm{d}/\mathrm{d}z$. When $P = p_r\partial^r + \cdots$ is a differential operator, we denote by $\rho(P) = \min\{|\xi| : p_r(\xi) = 0\} \in [0, \infty]$ the radius of the disk centered at the origin and extending to the nearest singular point. Variable names set in bold represent complex intervals, or *balls*, and operations involving them obey the usual laws of midpoint-radius interval arithmetic (i.e., $\boldsymbol{x} * \boldsymbol{y}$ is a reasonably tight ball containing $x * y$, for all $x \in \boldsymbol{x}$, $y \in \boldsymbol{y}$, and for every arithmetic operation $*$). We denote by $\mathrm{mid}(\boldsymbol{x})$ the center of a ball $\boldsymbol{x}$ and by $\mathrm{rad}(\boldsymbol{x})$ its radius.

As mentioned, the key feature of Algorithm 1 is that step 5(a)1 computes $\boldsymbol{u}_n$ based only on the centers of the intervals $\boldsymbol{u}_{n-1}, \dots, \boldsymbol{u}_{n-s}$, ignoring their radii. Let us prove that thanks to the correction made at step 6, the enclosure returned when the computation succeeds is nevertheless correct. The algorithm may also fail at step 6, but that can always be avoided by increasing the working precision (provided

[8] We do not consider the *truncation* error here. Note, though, that the tails of the series $\hat{u}(z) = \hat{u}_0\hat{g}(z)$ determined by Algorithm 1 are majorant series of the tails of $u(z)$. This means that the algorithm can be modified to simultaneously bound the truncation and rounding error, most of the steps involved being shared between both bounds. See [26] and the references therein for more on computing tight bounds on truncation errors.

Algorithm 1

Input: An operator $P = p_r(z)\partial^r + \cdots + p_1(z)\partial + p_0(z)$, with $p_r(0) \neq 0$. A vector $(\boldsymbol{u}_n)_{n=0}^{r-1}$ of ball initial values. An evaluation point $\zeta \in \mathbb{C}$ with $|\zeta| < \rho(P)$. A truncation order N.

Output: A complex ball containing $\sum_{n=0}^{N-1} u_n\zeta^n$, where $u(z)$ is the solution of $P \cdot u = 0$ corresponding to the given initial values.

1. [Compute a recurrence relation.] Define $L \in \mathbb{K}[X][Y]$ by $z^r P = L(z, z\partial)$. Compute polynomials $b_0(n), \dots, b_s(n)$ such that
$$L(S^{-1}, n) = b_0(n) - b_1(n)S^{-1} - \cdots - b_s(n)S^{-s}.$$
2. [Compute a majorant differential equation.] Let $m = \max(1, \deg p_r)$. Compute a rational $\alpha \approx \rho(P)^{-1}$ such that $\rho(P)^{-1} < \alpha < |\zeta|^{-1}$. (If $\rho(P) = \infty$, take, for instance, $\alpha \approx |\zeta|^{-1}/2$.) Compute rationals $c \approx |p_r(0)|$ and M such that $0 < c \leqslant |p_r(0)|$ and $M \geqslant \max_{i=0}^{r-1} \sum_{j=0}^{\deg p_i} |p_{i,j}|\alpha^{-i-j-1}$.
3. [Initial values for the bounds.] Compute positive lower bounds on the first r terms of the series $\hat{g}(z) = \exp\left(Mc^{-1}\alpha \int_0^z (1-\alpha z)^{-m}\right)$. (This is easily done using arithmetic on truncated power series with ball coefficients.) Deduce rationals $\hat{u}_0 \geqslant \max_{n=0}^{r-1}(|\boldsymbol{u}_n|/\hat{g}_n)$ and $\hat{\delta}_0 \geqslant \max_{n=0}^{r-1}(|\delta_n|/\hat{g}_n)$.
4. [Initialization.] Set
$$(\tilde{u}_{r-s}, \dots, \tilde{u}_{-1}, \tilde{u}_0, \dots, \tilde{u}_{r-1}) = (0, \dots, 0, \mathrm{mid}(\boldsymbol{u}_0), \dots, \mathrm{mid}(\boldsymbol{u}_{r-1})),$$
$\boldsymbol{s}_0 = 0$, $\boldsymbol{t}_0 = 1$, $\bar{\eta} = 0$. (Although most variables are indexed by functions on n for ease of reference, only $(\tilde{u}_{n-i})_{i=0}^{s}$, $\boldsymbol{s}_n$, $\boldsymbol{t}_n$, and $\bar{\eta}$ need to be stored from one loop iteration to the next.)
5. For $n = 1, \dots, N$, do:
   (a) If $n \geqslant r$, then:
      1. [Next coefficient.] Compute
$$\boldsymbol{u}_n = \frac{1}{b_0(n)}(b_1(n)\tilde{u}_{n-1} + \cdots + b_s(n)\tilde{u}_{n-s})$$
      in ball arithmetic.
      2. [Round.] Set[a] $\tilde{u}_n = \mathrm{mid}(\boldsymbol{u}_n)$.
      3. [Local error bound.] Let $\mu = |\tilde{u}_{n-1}| + \cdots + |\tilde{u}_{n-s}|$. If $\mu \neq 0$, then update $\bar{\eta}$ to $\max(\bar{\eta}, \eta_n)$ where $\eta_n = \mathrm{rad}(\boldsymbol{u}_n)/\mu$.
   (b) [Next partial sum.] Compute $\boldsymbol{t}_n = \zeta \cdot \boldsymbol{t}_{n-1}$ and $\boldsymbol{s}_n = \boldsymbol{s}_{n-1} + \tilde{u}_{n-1}\boldsymbol{t}_{n-1}$ using ball arithmetic.
6. [Account for accumulated numerical errors.] Compute $\sigma \geqslant c(|\zeta| + \cdots + |\zeta|^s)$. If $\sigma\bar{\eta} \geqslant 1$, signal an error. Otherwise, compute
$$A \geqslant \frac{M\alpha|\zeta|}{(1-\alpha|\zeta|)^m}, \qquad \Delta_N \geqslant \frac{\hat{\delta}_0 + \hat{u}_0\sigma(1+A)\bar{\eta}}{1-\sigma\bar{\eta}} \exp\frac{A}{1-\sigma\bar{\eta}} \tag{42}$$
and increase the radius of $\boldsymbol{s}_N$ by Δ_N.
7. Return $\boldsymbol{s}_N$.

[a] If $\boldsymbol{u}_n$ contains 0, it can be better in practice to force $\tilde{u}_n$ to 0 even if $\mathrm{mid}(\boldsymbol{u}_n) \neq 0$ and increase $\mathrm{rad}(\boldsymbol{u}_n)$ accordingly.

that interval operations on inputs of radius tending to zero produce results of radius that tends to zero).

With the notation from the algorithm, let $u(z)$ be a power series solution of $P \cdot u = 0$ corresponding to initial values $u_0 \in \boldsymbol{u}_0, \dots, u_r \in \boldsymbol{u}_r$. The Cauchy existence theorem implies that such solution exists and that $u(z)$ converges on the disk $|z| < \rho(P)$. We recall basic facts about the recurrence obtained at step 1. Let $Q(X) = X(X-1)\cdots(X-r+1)$.

Lemma 26. *The coefficient sequence (u_n) of $u(z)$ satisfies*

$$b_0(n)u_n = b_1(n)u_{n-1} + \cdots + b_s(n)u_{n-s}, \tag{43}$$

where one has $b_0(n) = p_r(0)Q(n)$. (In particular, b_0 is not the zero polynomial.)

Proof. Observe that for $p \in \mathbb{C}[z]$, the operator $\partial p = p\partial + p'$ has p as a leading coefficient when viewed as a polynomial in ∂ with coefficients in $\mathbb{C}[z]$ written to the left. It follows that $z^k\partial^k = (z\partial)^k +$ (terms involving lower powers of ∂) and therefore that $z^r P$ can be written as a polynomial in z and $z\partial$, as implicitly required by the algorithm. That the operator $L(S^{-1}, n)$ annihilates (u_n) follows from Lemma 2.

The only term of $z^r P$, viewed as a sum of monomials $p_{i,j}z^j\partial^i$ that can contribute to b_0 is $p_r(0)z^r\partial^r$, for all others have $i - j < 0$. The relation $z^k\partial^k = (z\partial - k + 1)z^{k-1}\partial^{k-1}$ then shows that $b_0(n) = p_r(0)Q(n)$, where $p_0(0) \neq 0$ by assumption. □

Let $\hat{a}(z) = Mc^{-1}\alpha(1-\alpha z)^{-m}$ where M, c, and α are the quantities computed at step 2 of the algorithm.

Lemma 27. *Let $y, \hat{y}$ be power series such that $P \cdot y \ll \partial^{r-1}(\partial - \hat{a}) \cdot \hat{y}$. If one has $|y_n| \leqslant \hat{y}_n$ for $n < r$, then $y \ll \hat{y}$.*

Proof. Let $\hat{P} = \partial^{r-1}(\partial - \hat{a}(z))$, that is,

$$\hat{P} = \partial^{r-1}\left(\partial - \frac{Mc^{-1}\alpha}{(1-\alpha z)^m}\right) = \partial^r - \sum_{i=0}^{r-1} \hat{f}_i(z)\partial^{r-1-i}$$

where

$$\hat{f}_i(z) = \binom{r-1}{i}\frac{(m+i-1)!}{(m-1)!}\frac{Mc^{-1}\alpha^{i+1}}{(1-\alpha z)^{m+i}}.$$

The parameters c, m, and α are chosen so that $(1-\xi^{-1}z)^{-1} \ll (1-\alpha z)^{-1}$ for every root ξ of p_r, and hence $p_r(z)^{-1} \ll c^{-1}(1-\alpha z)^{-m}$. Since, additionally, one has $(\alpha z)^j(1-\alpha z)^{-1} \ll (1-\alpha z)^{-1}$, it follows that, for $0 \leqslant i < r$,

$$\frac{p_i(z)}{p_r(z)} \ll \frac{\sum_j |p_{i,j}|z^j}{c(1-\alpha z)^m} = \sum_j |p_{i,j}|\alpha^{-j}\frac{(\alpha z)^j}{c(1-\alpha z)^m} \ll \frac{Mc^{-1}\alpha^{i+1}}{(1-\alpha z)^m} \ll \hat{f}_i(z),$$

by definition of M. By Proposition 7, these inequalities and our assumptions on $\hat{y}$ imply $y \ll \hat{y}$. □

Lemma 27 applies in particular to series $y, \hat{y}$ with $P \cdot y = 0$ and $\hat{y}' = \hat{a}\hat{y}$. The solution $\hat{g}$ of the latter equation with $\hat{g}_0 = 1$ is the series

$$\hat{g}(z) = \exp\int_0^z \hat{a}(w)\mathrm{d}w \tag{44}$$

already encountered in step 3 of the algorithm. Observe that none of its coefficients vanishes. Therefore, step 3 runs without error, and ensures that $|u_n| \leqslant \hat{u}_0\hat{g}_n$ (and $|\delta_n| \leqslant \hat{\delta}_0\hat{g}_n$) for $n < r$. As $P \cdot u = 0$, the lemma implies $u \ll \hat{u}_0\hat{g}$.

Let us now turn to the loop. As usual, consider now the computed coefficient sequence $(\tilde{u}_n)$, and let $\delta_n = \tilde{u}_n - u_n$. Write

$$\tilde{u}_n = \frac{1}{b_0(n)}(b_1(n)\tilde{u}_{n-1} + \cdots + b_s(n)\tilde{u}_{n-s}) + \varepsilon_n, \qquad r \leqslant n \leqslant N,$$

so that $|\varepsilon_n| \leqslant \mathrm{rad}(\boldsymbol{u}_n)$. Let $\eta_n = \varepsilon_n/(|\tilde{u}_{n-1}| + \cdots + |\tilde{u}_{n-s}|)$ when $r \leqslant n \leqslant N$ and the denominator is nonzero, and $\eta_n = 0$ otherwise. We thus have, for all $n \in \mathbb{Z}$,

$$b_0(n)\tilde{u}_n - b_1(n)\tilde{u}_{n-1} - \cdots - b_s(n)\tilde{u}_{n-s} = b_0(n)\eta_n(|\tilde{u}_{n-1}| + \cdots + |\tilde{u}_{n-s}|) \tag{45}$$

and, thanks to step 5(a)3 of the algorithm, $|\eta_n| \leqslant \bar{\eta}$. By subtracting (43) from (45) and bounding η_n by $\bar{\eta}$, we obtain

$$|b_0(n)\delta_n - b_1(n)\delta_{n-1} - \cdots - b_s(n)\delta_{n-s}| \leqslant c\bar{\eta}Q(n)(|\tilde{u}_{n-1}| + \cdots + |\tilde{u}_{n-s}|). \tag{46}$$

Let $\varphi(z) = z + \cdots + z^s$.

Lemma 28. *Let $\hat{v} \in \mathbb{R}_{\geqslant 0}[[z]]$ be any majorant series of $\left(\varphi^\sharp u\right)'(z)$. The equation*

$$(1 - c\bar{\eta}\varphi(z))\hat{\delta}'(z) = (\hat{a}(z) + c\bar{\eta}\varphi'(z))\hat{\delta}(z) + c\bar{\eta}\hat{v}(z). \tag{47}$$

admits a solution $\hat{\delta}(z)$ with the initial value $\hat{\delta}_0$ computed at step 3, and this solution is a majorant series of $\delta(z)$.

Proof. In terms of generating series, (46) rewrites as $z^r P \cdot \delta(z) \ll c\bar{\eta}Q(z\partial)\varphi(z) \cdot {}^\sharp\tilde{u}(z)$. As already observed in the proof of Lemma 26, one has $Q(z\partial) = z^r\partial^r$, so that the previous equation is equivalent to

$$P \cdot \delta(z) \ll c\bar{\eta}\partial^r\varphi(z) \cdot {}^\sharp\tilde{u}(z). \tag{48}$$

Let $\hat{\gamma}$ be the solution of

$$(\partial - \hat{a}(z)) \cdot \hat{\gamma}(z) = c\bar{\eta}\partial\varphi(z) \cdot {}^\sharp\tilde{u}(z) \tag{49}$$

with $\hat{\gamma}_0 = \hat{\delta}_0$. By Proposition 7, we have $\hat{\delta}_0\hat{g}(z) \ll \hat{\gamma}(z)$ where $\hat{g}$ is given by (44). In addition, as noted when discussing step 3, we have $|\delta_n| \leqslant \widehat{\delta_0}\hat{g}_n$ for $n < r$, hence $|\delta_n| \leqslant \hat{\gamma}_n$ for $n < r$. As $\hat{\gamma}$ also satisfies $P \cdot \delta \ll \partial^r(\partial - \hat{a}) \cdot \hat{\gamma}$, we can conclude that $\delta \ll \hat{\gamma}$ using Lemma 27. But then, since $\tilde{u} = u + \delta$, we have ${}^\sharp\tilde{u} \ll {}^\sharp u + \hat{\gamma}$, hence $\left(\varphi^\sharp\tilde{u}\right)' \ll \hat{v} + (\varphi\hat{\gamma})'$, and (49) implies

$$\hat{\gamma}' = \hat{a}\hat{\gamma} + c\bar{\eta}\left(\varphi^\sharp\tilde{u}\right)' \ll c\bar{\eta}\varphi\hat{\gamma}' + (\hat{a} + c\bar{\eta}\varphi')\hat{\gamma} + c\bar{\eta}\hat{v}$$

where we note that $\varphi(0) = 0$. This inequality is of the form required by Lemma 8, which yields the existence of $\hat{\delta}$ and the inequality $\hat{\gamma} \ll \hat{\delta}$. We thus have $\delta \ll \hat{\gamma} \ll \hat{\delta}$. □

It remains to solve the majorant equation (47) to get an explicit bound on $\hat{\delta}$.

Proposition 29. *The generating series $\delta(z)$ of the global error on u_n committed by Algorithm 1 satisfies*

$$\delta(z) \ll \frac{\hat{\delta}_0 + c\bar{\eta}\hat{u}_0\varphi(z)(1 + z\hat{a}(z))}{1 - c\bar{\eta}\varphi(z)} \exp\left(\frac{z\hat{a}(z)}{1 - c\bar{\eta}\varphi(z)}\right), \quad \hat{a}(z) = \frac{Mc^{-1}\alpha}{(1 - \alpha z)^m}. \tag{50}$$

Proof. The solution $\hat{h}(z)$ of the homogeneous part of (47) with $\hat{h}_0 = 1$ is given by

$$\hat{h}(z) = \frac{1}{1 - c\bar{\eta}\varphi(z)} \exp \int_0^z \frac{\hat{a}(w)}{1 - c\bar{\eta}\varphi(w)} \mathrm{d}w.$$

Observe that

$$\left(\varphi^\sharp u\right)' \ll \hat{u}_0(\varphi\hat{g})' = \hat{u}_0(\varphi' + \varphi\hat{a})\hat{g} \ll \hat{u}_0(\varphi' + \varphi\hat{a})\hat{h},$$

so that, in Lemma 28, we can take $\hat{v}(z) = \hat{u}_0(\varphi' + \varphi\hat{a})\hat{h}$. The method of variation of parameters then leads to the expression

$$\hat{\delta}(z) = \hat{h}(z) \left(\hat{\delta}_0 + c\bar{\eta} \int_0^z \frac{\hat{v}(w)}{\hat{h}(w)} \mathrm{d}w \right) = \hat{h}(z) \left(\hat{\delta}_0 + c\bar{\eta}\hat{u}_0 \left(\varphi(z) + \int_0^z \varphi(w)\hat{a}(w) \mathrm{d}w \right) \right).$$

Using the bounds from Lemma 9

$$\int_0^z \frac{\hat{a}(w)}{1 - c\bar{\eta}\varphi(w)} \mathrm{d}w \ll \frac{z\hat{a}(z)}{1 - c\bar{\eta}\varphi(z)}, \qquad \int_0^z \varphi(w)\hat{a}(w) \mathrm{d}w \ll z\varphi(z)\hat{a}(z),$$

(where $z\hat{a}(z)$ could be replaced by $\int_0^z \hat{a}$ at the price of a slightly more complicated final bound) we see that $\hat{\delta}(z)$ is bounded by the right-hand side of (50). □

Step 6 of Algorithm 1 effectively computes an upper bound on $|\delta(\zeta)|$ using inequality (50). It follows that the returned interval $\boldsymbol{s}_N$ contains the exact partial sum $\sum_{n=0}^{N-1} u_n\zeta^n$ corresponding to the input data, as stated in the specification of the algorithm.

Remark 30. When the operator P, the series $u(z)$ and the evaluation point ζ are fixed, our bound Δ_N on the global error decreases linearly with $\hat{\delta}_0 + \bar{\eta}$. Suppose that we run the algorithm with a relative working precision of t bits. Under the reasonable assumptions that $\mathrm{rad}(\boldsymbol{u}_n) = \mathrm{O}(2^{-t})$ for $0 \leqslant n < r$ and that both η_n and $\mathrm{rad}(\boldsymbol{s}_n)$ are[9] $\mathrm{O}(n^d 2^{-t})$ for some d, we then have $\Delta_N = \mathrm{O}(N^d 2^{-t})$. As the truncation order N necessary for reaching an accuracy $|\boldsymbol{s}_N - u(\zeta)| \leqslant 2^{-q}$ is $N = \mathrm{O}(q)$, this means that, for fixed u and ζ, the algorithm needs no more than $q + \mathrm{O}(\log q)$ bits of working precision to compute an enclosure of $u(\zeta)$ of width 2^{-q}.

Remark 31. The majorant series of Proposition 29 was chosen to keep the algorithm simple, not to optimize the error bound.

One helpful feature it does have that the parameter α can be taken arbitrarily close to $\rho(p_r)^{-1}$ without forcing other parts of the bound to tend to infinity. (This is in contrast with the geometric majorant series typically found in textbook proofs of theorems on differential equations.) Nevertheless, the exponential factor in (42) can easily grow extremely large, and we expect Algorithm 1 to lead to unusable bounds in practice on moderately complicated examples. Even in simple cases, forcing a majorant series of finite radius of convergence when p_r is constant is far from optimal.

The same technique, though, can be used with a more sophisticated choice of majorant series. In particular, the algorithm adapts without difficulty if $\hat{a}(z)$ is a sharper rational majorant of the coefficients of the equation. We leave for future

[9]Such a growth for $\bar{\eta}$ is reasonable since the coefficients b_i of the recurrence are polynomials. Regarding $\boldsymbol{s}_n$, we can in fact expect to have $\mathrm{rad}(\boldsymbol{t}_n)/\zeta^n = \mathrm{O}(n2^{-t})$, and, since $u_n\zeta^n$ converges geometrically to zero, $\mathrm{rad}(\boldsymbol{s}_n) \approx \mathrm{rad}(\boldsymbol{s}_{n-1}) + \mathrm{O}(nu_n\zeta^n 2^{-t}) + \mathrm{O}(2^{-t})$, leading to $\mathrm{rad}(\boldsymbol{s}_n) = \mathrm{O}(n2^{-t})$.

work the question of developing a truly practical variant of the algorithm[10]. It would also be interesting to extend the analysis to the computation of "logarithmic series" solutions of linear ODEs at regular singular points.

Remark 32. Instead of $\varepsilon_n/(|\tilde{u}_{n-1}| + \cdots + |\tilde{u}_{n-s}|)$, one could have Algorithm 1 compute a run-time bound directly on $|\varepsilon_n/\hat{h}_n|$. Doing so leads to a somewhat simpler variant of the above analysis. We chose to present the version given here because it is closer to plain floating-point error analysis — and gives us an excuse to illustrate the generalization to recurrences with polynomial coefficients of the technique of Section 9. Another small advantage is that plugging in a sharper first-order majorant equation as suggested in Remark 31 requires no other change to the algorithm, whereas it may not be obvious how to compute a good lower bound on $\hat{h}_n$.

Acknowledgments

This work benefited from remarks by many people, including Alin Bostan, Richard Brent, Frédéric Chyzak, Thibault Hilaire, Philippe Langlois, and Nicolas Louvet. The initial impulse came from discussions with Fredrik Johansson, Guillaume Melquiond, and Paul Zimmermann. Frédéric Chyzak suggested to work with elements of $\Omega[[t]]$ instead of $\mathbb{R}[x^{\pm 1}][[t]]$ in Section 11. I am especially grateful to Guillaume Melquiond and Anne Vaugon for many insightful comments at various stages, and to Paul Zimmermann for his thorough reading of a preliminary version.

References

[1] Sylvie Boldo. Floats and Ropes: A Case Study for Formal Numerical Program Verification. In *Automata, Languages and Programming*, Lecture Notes in Computer Science, pages 91–102. Springer, Berlin, Heidelberg, 2009. doi:10.1007/978-3-642-02930-1_8.

[2] Sylvie Boldo, François Clément, Jean-Christophe Filliâtre, Micaela Mayero, Guillaume Melquiond, and Pierre Weis. Wave Equation Numerical Resolution: A Comprehensive Mechanized Proof of a C Program. *Journal of Automated Reasoning*, 50(4):423–456, 2013. doi:10.1007/s10817-012-9255-4.

[3] Sylvie Boldo, François Clément, Jean-Christophe Filliâtre, Micaela Mayero, Guillaume Melquiond, and Pierre Weis. Trusting computations: A mechanized proof from partial differential equations to actual program. *Computers & Mathematics with Applications*, 68(3):325–352, 2014. doi:10.1016/j.camwa.2014.06.004.

[4] Richard P. Brent. Unrestricted algorithms for elementary and special functions. In S. H. Lavington, editor, *Information Processing 80*, 1980. URL: https://maths-people.anu.edu.au/~brent/pub/pub052.html.

[5] Richard P. Brent and David Harvey. Fast Computation of Bernoulli, Tangent and Secant Numbers. In David H. Bailey, Heinz H. Bauschke, Peter Borwein, Frank Garvan, Michel Théra, Jon D. Vanderwerff, and Henry Wolkowicz, editors, *Computational and Analytical Mathematics*, Springer Proceedings in Mathematics & Statistics, pages 127–142. Springer, 2013. doi:10.1007/978-1-4614-7621-4_8.

[6] Richard P. Brent and Paul Zimmermann. *Modern Computer Arithmetic*. Cambridge University Press, 2010. URL: https://members.loria.fr/PZimmermann/mca/pub226.html.

[7] Henri Cartan. *Théorie élémentaire des fonctions analytiques d'une ou plusieurs variables complexes*. Hermann, 1961.

[10] As a first step in this direction, we have implemented a variant of Algorithm 1 based on the more flexible framework of [26], in the `ore_algebra` package [23, 25] for SageMath. While very effective in some cases, it does not at this stage run consistently faster than naive interval summation, due both to overestimation issues and to the computational overhead of computing good bounds.

[8] Augustin Cauchy. Résumé d'un mémoire sur la mécanique céleste et sur un nouveau calcul appelé calcul des limites. In *Exercices d'analyse et de physique mathématique*, volume II, pages 50–109. Bachelier, 1841. Reproduced in [11], p. 58-112.

[9] Augustin Cauchy. Mémoire sur l'emploi du nouveau calcul, appelé calcul des limites, dans l'intégration d'un système d'équations différentielles. *Comptes-rendus de l'Académie des Sciences*, 15:14, 1842. Reproduced in [10], §169, p. 5-17.

[10] Augustin Cauchy. *Œuvres complètes d'Augustin Cauchy, Ière série*, volume VII. Gauthier-Villars, 1892. URL: `https://gallica.bnf.fr/ark:/12148/bpt6k901870`.

[11] Augustin Cauchy. *Œuvres complètes d'Augustin Cauchy, IIe série*, volume XII. Gauthier-Villars, 1916.

[12] Roger Cooke. The Cauchy-Kovalevskaya theorem. URL: `https://web.archive.org/web/20100725150116/http://www.cems.uvm.edu/~cooke/ckthm.pdf`.

[13] Digital library of mathematical functions, 2010. Companion to the NIST Handbook of Mathematical Functions [30]. URL: `http://dlmf.nist.gov/`.

[14] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009. URL: `http://algo.inria.fr/flajolet/Publications/book.pdf`.

[15] M. Giusti, G. Lecerf, B. Salvy, and J.-C. Yakoubsohn. On Location and Approximation of Clusters of Zeros of Analytic Functions. *Foundations of Computational Mathematics*, 5(3):257–311, 2005. `doi:10.1007/s10208-004-0144-z`.

[16] Joseph F. Grcar. John von Neumann's Analysis of Gaussian Elimination and the Origins of Modern Numerical Analysis. *SIAM Review*, 53(4):607–682, 2011. `doi:10.1137/080734716`.

[17] Peter Henrici. *Discrete Variable Methods in Ordinary Differential Equations*. Wiley, 1962.

[18] Nicholas J. Higham. *Accuracy and stability of numerical algorithms*. SIAM, 2002.

[19] Thibault Hilaire and Benoit Lopez. Reliable implementation of linear filters with fixed-point arithmetic. In *SiPS 2013 Proceedings*, pages 401–406, 2013. `doi:10.1109/SiPS.2013.6674540`.

[20] Einar Hille. *Ordinary differential equations in the complex domain*. Wiley, 1976. Dover reprint, 1997.

[21] Fredrik Johansson and Marc Mezzarobba. Fast and rigorous arbitrary-precision computation of Gauss-Legendre quadrature nodes and weights. *SIAM Journal on Scientific Computing*, 40(6):C726–C747, 2018. URL: `https://arxiv.org/abs/1802.03948`, `doi:10.1137/18M1170133`.

[22] Manuel Kauers. The holonomic toolkit. *Computer Algebra in Quantum Field Theory: Integration, Summation and Special Functions, Texts and Monographs in Symbolic Computation*, 2013. URL: `http://www.risc.jku.at/people/mkauers/publications/kauers13.pdf`.

[23] Manuel Kauers, Maximilian Jaroschek, and Fredrik Johansson. Ore polynomials in Sage. In Jaime Gutierrez, Josef Schicho, and Martin Weimann, editors, *Computer Algebra and Polynomials*, pages 105–125. Springer, 2015. URL: `https://arxiv.org/abs/1306.4263`, `doi:10.1007/978-3-319-15081-9_6`.

[24] B. Liu and T. Kaneko. Error analysis of digital filters realized with floating-point arithmetic. *Proceedings of the IEEE*, 57(10):1735–1747, 1969. `doi:10.1109/PROC.1969.7388`.

[25] Marc Mezzarobba. Rigorous multiple-precision evaluation of D-finite functions in SageMath. Technical Report 1607.01967, arXiv, 2016. Extended abstract of a talk at the 5th International Congress on Mathematical Software. URL: `http://arxiv.org/abs/1607.01967`.

[26] Marc Mezzarobba. Truncation bounds for differentially finite series. *Annales Henri Lebesgue*, 2:99–148, 2019. `doi:10.5802/ahl.17`.

[27] Marc Mezzarobba and Bruno Salvy. Effective bounds for P-recursive sequences. *Journal of Symbolic Computation*, 45(10):1075–1096, 2010. URL: `http://arxiv.org/abs/0904.2452`, `doi:10.1016/j.jsc.2010.06.024`.

[28] J. Oliver. *The Numerical Solution of the Initial Value Problem for Linear Recurrence Relations*. PhD thesis, Ph. D. Thesis, University of Cambridge, 1965.

[29] J. Oliver. Relative error propagation in the recursive solution of linear recurrence relations. *Numerische Mathematik*, 9:323–340, 1967. URL: `https://doi.org/10.1007/BF02162423`.

[30] F. W. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, editors. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

[31] Bruno Salvy. Linear differential equations as a data-structure. *Foundations of Computational Mathematics*, 19(5):1071–1112, 2019. URL: `http://arxiv.org/abs/1811.08616`.

[32] A. M. Turing. Rounding-Off Errors in Matrix Processes. *The Quarterly Journal of Mechanics and Applied Mathematics*, 1(1):287–308, 1948. URL: https://academic.oup.com/qjmam/article/1/1/287/1883483, doi:10.1093/qjmam/1.1.287.

[33] Joris van der Hoeven. Majorants for formal power series. Technical Report 2003-15, Université Paris-Sud, 2003. URL: http://www.texmacs.org/joris/maj/maj-abs.html.

[34] John von Neumann and H. H. Goldstine. Numerical inverting of matrices of high order. *Bulletin of the American Mathematical Society*, 53(11):1021–1099, 1947. doi:10.1090/S0002-9904-1947-08909-6.

[35] P. G. Warne, D. A. P. Warne, J. S. Sochacki, G. E. Parker, and D. C. Carothers. Explicit a-priori error bounds and adaptive error control for approximation of nonlinear initial value differential systems. *Computers and Mathematics with Applications*, 52(12):1695–1710, 2006. doi:10.1016/j.camwa.2005.12.004.

[36] J. H. Wilkinson. *Rounding errors in algebraic processes*. Prentice-Hall, 1963.

[37] Jet Wimp. Forward computation in second order difference equations. *Applicable Analysis*, 1(4):325–329, 1972. doi:10.1080/00036817208839021.

[38] Jet Wimp. *Computation with Recurrence Relations*. Pitman, 1984.

Sorbonne Université, CNRS, LIP6, F-75005 Paris, France

Current address: LIX, CNRS, École polytechnique, Institut polytechnique de Paris, 91120 Palaiseau, France

Email address: marc@mezzarobba.net

URL: http://marc.mezzarobba.net/