On high precision integration of stiff differential equations

Joris van der Hoeven

▶ **To cite this version:**

Joris van der Hoeven. On high precision integration of stiff differential equations. 2018. hal-01889324

HAL Id: hal-01889324
https://hal.science/hal-01889324v1

Preprint submitted on 6 Oct 2018

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

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

On high precision integration of stiff differential equations

JORIS VAN DER HOEVEN
CNRS, Laboratoire d'informatique
Campus de l'École polytechnique
1, rue Honoré d'Estienne d'Orves
Bâtiment Alan Turing, CS35003
91120 Palaiseau
France
Email: vdhoeven@lix.polytechnique.fr

Draft version, October 6, 2018

In this paper, we present various new algorithms for the integration of stiff differential equations that allow the step size to increase proportionally with time. We mainly focus on high precision integrators, which leads us to use high order Taylor schemes. We will also present various algorithms for certifying the numerical results, again with the property that the step size increases with time.

KEYWORDS: stiff differential equation, numerical method, reliable computation, rigorous integration, Taylor models, multiple precision computations
A.M.S. SUBJECT CLASSIFICATION: 65L04, 65L20, 65G20, 37-04

1. INTRODUCTION

Consider the differential equation

$$\dot{\varphi} + \Lambda \varphi \;=\; \Phi(\varphi), \tag{1}$$

where $\Lambda \in \mathbb{R}^{d \times d}$ is a diagonal matrix with positive real entries $0 \leqslant \lambda_1 \leqslant \cdots \leqslant \lambda_d$, and where $\Phi \in \mathbb{C}[F^{[1]}, \ldots, F^{[d]}]^d$ is a polynomial forcing term. We are interested in analytic solutions $\varphi : I \mapsto \mathbb{C}^d$, where I is either an interval of the form $[0, T]$ with $T > 0$, or a larger subset of $\mathbb{C}$ that contains an interval of this form. The d components of this mapping will be written $\varphi^{[1]}, \ldots, \varphi^{[d]} : I \mapsto \mathbb{C}$, whereas subscripts will be reserved for the coefficients of local power series solutions. A similar notation will be used for the components of other vectors and matrices.

If the largest eigenvalue λ_d of Λ gets large with respect to the forcing term Φ, then the equation (1) is said to be *stiff*. Generic numerical integration schemes experience difficulties with this kind of equations, whether we use Euler's method, a Runge–Kutta scheme, or high order algorithms based on Taylor expansions. Roughly speaking, the problem is that all generic methods become inaccurate when the step size exceeds λ_d^{-1}. One of the most difficult cases is when d is large, but each of the quotients $\lambda_{k+1} / \lambda_k$ remains reasonably close to one. This happens for instance when $\lambda_k = k^2$, a case that is naturally encountered when discretizing certain partial differential equations.

Implicit integration schemes allow for larger step sizes than generic methods, but involve the computation and inversion of Jacobian matrices, which may be expensive in high dimension d. But even such more expensive schemes only seem to deal with

the step size issue in a somewhat heuristic manner: one typically selects a scheme that is "stable" when applied to any equation $\varphi' + \lambda \varphi = 0$ for all $\lambda \geqslant 0$, and then hopes that the scheme continues to behave well for the actual target equation (1). We refer to [25, Section 17.5] and [4, Section 6] for more information on classical integration schemes for stiff differential equations.

We are mainly interested in high precision computations, in which case it is natural to use high order integration schemes that are based on Taylor expansions. In this context, we are not aware of any numerical method that allows the step size to increase proportionally with time. The main aim of this paper is to present such a method, together with several variants, as well as a way to compute rigorous bounds for the error. Our fastest method is an explicit scheme, but its convergence deteriorates when the eigenvalues λ_i are close. The slower and rigourous counterparts rely on the computation of Jacobian matrices.

In sections 3 and 4, we recall various approaches that can be used to compute truncated series solutions to initial value problems and how to derive high order integration schemes from this. More precisely, given a numerical approximation $\varsigma\langle t\rangle$ for $\varphi(t)$ at time t, we compute the first n terms of a power series solution $f\langle t\rangle$ to the initial value problem

$$\frac{\partial f\langle t\rangle}{\partial z} + \Lambda f\langle t\rangle = \Phi(f\langle t\rangle), \qquad f\langle t\rangle_0 = \varsigma\langle t\rangle \tag{2}$$

and return $f\langle t\rangle_0 + f\langle t\rangle_1 \delta + \cdots + f\langle t\rangle_{n-1} \delta^{n-1}$ as an approximation of $\varphi(t+\delta)$ for some suitable step size δ. The inaccuracy of such schemes for larger step sizes is due to the fact that the k-th coefficient of $f\langle t\rangle$ tends to change proportionally to $\lambda_d^k / k!$ for slight perturbations of $\varsigma\langle t\rangle$. If λ_d is large, then even small multiples of $\lambda_d^k / k!$ quickly dominate $f\langle t\rangle_k$, which severely limits the step sizes that we can use.

This numerical instability is an unpleasant artifact of traditional methods for the integration of differential equations. It is well know that solutions of stiff equations tend to become very smooth after a certain period of time (in which case one says that the system has reached its *steady state*), but this does not lead to larger step sizes, as we would hope for.

In order to make this idea of "smoothness after a certain period of time" more precise, it is instructive to study the analytic continuation of φ in the complex plane. In section 2, for a fixed initial condition $\varphi(0) = c$ and a fixed forcing term Φ, we show the existence of a compact half disk $H_R = \{t \in \mathbb{C} : |t| \leqslant R, \operatorname{Re} t \geqslant 0\}$, $R > 0$, on which the solution φ of (1) is analytic, for *any* choice of Λ. In particular, for $t \in [0, R/2]$, the radius of convergence $\varrho\langle t\rangle$ of the exact Taylor expansion $f^*\langle t\rangle$ of φ at time t is at least t. Setting $K = \max_{z \in H_R} |\varphi(z)|$, Cauchy's formula thus yields the bound $|f^*\langle t\rangle_k| \leqslant K / t^k$ for all $k \in \mathbb{N}$.

If $t \gg \lambda_d^{-1}$, then this means that efficient integration schemes should allow for step sizes of the order αt at time t, for some fixed constant $\alpha > 0$. In section 5, we will present various such schemes. They are all based on the computation of good approximations $f\langle t\rangle_{;n} = f\langle t\rangle_0 + f\langle t\rangle_1 z + \cdots + f\langle t\rangle_{n-1} z^{n-1}$ of the exact Taylor expansion $f^*\langle t\rangle_{;n}$ of φ at time t and order n. By what precedes, such an approximation cannot be obtained by solving (2). Instead, for a suitable so-called *critical index* $\chi \in \{0, \ldots, d\}$ that depends on t, we will solve the following so-called *steady-state problem*:

$$\frac{\partial f\langle t\rangle}{\partial z} + \Lambda f\langle t\rangle = \Phi(f\langle t\rangle), \qquad \begin{cases} f\langle t\rangle_0^{[i]} = \varsigma\langle t\rangle^{[i]} & \text{if } i \leqslant \chi \\ f\langle t\rangle_n^{[i]} = 0 & \text{if } i > \chi \end{cases}. \tag{3}$$

Intuitively speaking, the system has reached a steady state for the eigenvalues λ_i with $i > \chi$. Such eigenvalues λ_i are "large", so it is more accurate to determine coefficients $f\langle t\rangle_k^{[i]}$ as a function of later coefficients $f\langle t\rangle_{k+1}, \ldots, f\langle t\rangle_{n-1}$ rather than earlier ones. Furthermore, the coefficients $f^*\langle t\rangle_n^{[i]}$ are so "small" that they can be approximated by zeros; this explains the *steady-state conditions* $f\langle t\rangle_n^{[i]} = 0$. The remaining eigenvalues λ_i with $i \leqslant \chi$ are small enough for the step size under consideration that the dependence of $f\langle t\rangle$ on the initial condition $\varsigma\langle t\rangle^{[i]}$ is sufficiently moderate for (3) to admit an accurate solution. We say that the system (1) is in a *transient state* for such eigenvalues λ_i at time t.

A final topic is the computation of rigorous error bounds for the numerical integration process. For differential equations that are not stiff, this is a classical theme in interval analysis [21, 15, 5, 16, 18, 19, 11]. It has long been an open problem to develop efficient reliable integrators for stiff differential equations. We report on progress in this direction in section 6.

2. ANALYTIC CONTINUATION

Consider the equation (1) for a fixed initial condition $\varphi(0) = c \in \mathbb{R}^d$ and a fixed forcing term Φ, but where we will allow ourselves to vary Λ. Our main aim is to show that there exists a compact "half disk"

$$H_R = \{t \in \mathbb{C} : |t| \leqslant R, \operatorname{Re} t \geqslant 0\}$$

with $R > 0$ on which the solution φ to (1) exists and is analytic, independently of the choice of Λ. We will both prove a weak and simpler version of this result and a stronger version that allows us to take larger radii R and that is also more convenient if we want to actually compute a radius R that works.

2.1. Notations

We will need a few more notations. We denote the interior of H_R by

$$H_R^\circ = \{t \in \mathbb{C} : |t| < R, \operatorname{Re} t > 0\}.$$

Given a non-empty open subset $U \subseteq \mathbb{C}$, we write $\mathcal{A}(U)$ for the Banach space of analytic functions $g : U \to \mathbb{C}$ with finite sup-norm

$$\|g\| = \sup_{z \in U} |g(u)|.$$

We will often use vector notation: given $v \in \mathbb{C}^d$, $g \in \mathcal{A}(U)^d$, and $r, s \in (\mathbb{R}^{\geqslant})^d$, we write

$$\begin{aligned} |v| &= (|v^{[1]}|, \ldots, |v^{[d]}|) \\ \|g\| &= (\|g^{[1]}\|, \ldots, \|g^{[d]}\|) \\ r \leqslant s &\Leftrightarrow r^{[1]} \leqslant s^{[1]} \wedge \cdots \wedge r^{[d]} \leqslant s^{[d]}. \end{aligned}$$

In addition to such "componentwise" norms, we will also use more traditional sup-norms: given $v \in \mathbb{C}^d$, $g \in \mathcal{A}(U)^d$, and an $n \times n$ matrix $M \in \mathbb{C}^{d \times d}$, we define

$$\begin{aligned} |v|_\infty &= \max(|v^{[1]}|, \ldots, |v^{[d]}|) \\ \|g\|_\infty &= \max(\|g^{[1]}\|, \ldots, \|g^{[d]}\|) \\ |M|_\infty &= \sup_{v \in \mathbb{C}^d, |v|_\infty = 1} |Mv|_\infty. \end{aligned}$$

2.2. Reminders from complex analysis

We will also need two well known results from complex analysis, namely an explicit version of Cauchy-Kovalevskaya's theorem and a variant of Montel's theorem.

THEOREM 1. *Let $\Psi(F) = \Phi(F) - \Lambda F$, $m_c = |c|_\infty$, and consider the initial value problem*

$$\dot\varphi \;=\; \Psi(\varphi), \qquad \varphi(0) = c. \tag{4}$$

Writing $\Psi = \sum_{i_1,\ldots,i_d} \Psi_{i_1,\ldots,i_d} (F^{[1]})^{i_1} \cdots (F^{[d]})^{i_d}$, let

$$M_c = \max_{i_1,\ldots,i_d} |\Psi_{i_1,\ldots,i_d}| \, (2 m_c d)^{i_1 + \cdots + i_d}.$$

Then (4) admits a unique analytic solution on the open disk with center 0 and radius

$$\varrho_c = \frac{m_c}{4 M_c}.$$

Proof. Our assumptions imply that

$$\dot{\bar\varphi} \;=\; \frac{M_c}{1 - \frac{\bar\varphi^{[1]} + \cdots + \bar\varphi^{[d]}}{2 m_c d}}, \qquad \bar\varphi^{[1]}(0) = \cdots = \bar\varphi^{[1]}(0) = m_c$$

constitutes a "majorant equation" for (4) in the sense of Cauchy-Kovalevskaya. By symmetry, each of the components $\bar\varphi^{[i]}$ of this equation satisfies the simpler equation

$$\dot g \;=\; \frac{M_c}{1 - \frac{g}{2 m_c}}, \qquad g(0) = m_c,$$

which admits the explicit solution

$$g \;=\; \left(2 - \sqrt{1 - \frac{4 M_c t}{m_c}}\right) m_c,$$

of radius ϱ_c. Since the formal power series solution of (4) satisfies $|\varphi_k^{[i]}| \leqslant |\bar\varphi_k^{[i]}| = g_k$ for all $i \in \{1,\ldots,d\}$ and $k \in \mathbb{N}$, it follows that φ has radius of convergence at least ϱ_c. □

THEOREM 2. *Given a non-empty subset U of $\mathbb{C}$ and a bounded sequence $g_1, g_2, \ldots \in \mathcal{A}(U)^d$, we can extract a subsequence $g_{k_1}, g_{k_2}, \ldots$ that converges uniformly to a limit $g_\infty \in \mathcal{A}(U)^d$ on every compact subset of U.*

Proof. If $d = 1$, then this is Montel's theorem. The general case is obtained using an easy induction on d: we first extract a subsequence $g_{i_1}, g_{i_2}, \ldots$ such that $g_{i_1}^{[d]}, g_{i_2}^{[d]}, \ldots$ converges to a limit in $\mathcal{A}(U)$ and then apply the induction hypothesis to this subsequence for the remaining $d-1$ components. □

2.3. Analyticity on open half disks

Let $J_\Phi = \partial\Phi / \partial F$ denote the Jacobian matrix of Φ.

THEOREM 3. *Let $R > 0$, $B \in (\mathbb{R}^>)^d$, and $C < 1$ be such that for all $z \in \mathbb{C}$, we have*

$$|z - c| \leqslant B \implies |\Phi(z)| R \leqslant B. \tag{5}$$

$$|z - c| \leqslant B \implies |J_\Phi(z)|_\infty R \leqslant C. \tag{6}$$

Then, for any choice of Λ, the initial value problem

$$\dot\varphi + \Lambda\varphi = \Phi(\varphi), \qquad \lim_{t \to 0} \varphi(t) = c \tag{7}$$

admits a unique analytic solution on H_R°. In addition, we have $|\varphi(t)-c|\leqslant B$ for all $t\in H_R^\circ$.

Proof. Consider the operator $\Omega_R\colon \mathcal{A}(H_R^\circ)^d \to \mathcal{A}(H_R^\circ)^d$ with

$$\Omega_R(g)(t) \;=\; c + \mathrm{e}^{-\Lambda t}\int_0^t \mathrm{e}^{\Lambda u}\,\Phi(g(u))\,\mathrm{d}u.$$

The fixed points $\varphi\in\mathcal{A}(H_R^\circ)^d$ of this operator are precisely the solutions of (7). Let us prove the existence of such a fixed point. The uniqueness follows by analytic continuation from the well known uniqueness of the solution of the initial value problem (7) on a neighbourhood of the origin.

Let us first notice that Ω_R maps the ball

$$\mathcal{B}_R(c,B) = \{g\in\mathcal{A}(H_R^\circ) : \|g-c\|\leqslant B\}$$

with center c and radius B into itself. Indeed, given $g\in\mathcal{B}_R(c,B)$ and $u\in H_R^\circ$, our hypothesis (5) implies

$$|\mathrm{e}^{\Lambda u}\,\Phi(g(u))| \;\leqslant\; \mathrm{e}^{\Lambda \mathrm{Re}\, u}\,\frac{B}{R}.$$

For all $t\in H_R^\circ$, it follows that

$$\begin{aligned}\left|\int_0^t \mathrm{e}^{\Lambda u}\,\Phi(g(u))\,\mathrm{d}u\right| &\leqslant \frac{B}{R}\int_0^t \mathrm{e}^{\Lambda \mathrm{Re}\, u}\,\mathrm{d}u\\ &\leqslant \frac{B}{R}\,\mathrm{e}^{\Lambda \mathrm{Re}\, t}\int_0^t \mathrm{e}^{\Lambda \mathrm{Re}\,(u-t)}\,\mathrm{d}u\\ &\leqslant \frac{B}{R}\,\mathrm{e}^{\Lambda \mathrm{Re}\, t}|t| \;\leqslant\; B\,\mathrm{e}^{\Lambda \mathrm{Re}\, t},\end{aligned}$$

whence $|\Omega_R(g)(t)-c|\leqslant B$, as desired.

Let us next show that Ω_R is actually contracting on $\mathcal{B}_R(c,B)$. Given $g,h\in\mathcal{B}_R(c,B)$, consider the homotopy $\eta_\epsilon = (1-\epsilon)\,g+\epsilon h\in\mathcal{B}_R(c,B)$ with $\epsilon\in[0,1]$. From (6), we get

$$\begin{aligned}\|\Phi(h)-\Phi(g)\|_\infty &= \left\|\int_0^1 \frac{\Phi(\eta_\epsilon)}{\partial\epsilon}\,\mathrm{d}\epsilon\right\|_\infty\\ &\leqslant \int_0^1 \|J_\Phi(\eta_\epsilon)\,(h-g)\|_\infty\,\mathrm{d}\epsilon\\ &= \int_0^1 \sup_{t\in H_R^\circ} |J_\Phi(\eta_\epsilon(t))\,(h(t)-g(t))|_\infty\,\mathrm{d}\epsilon\\ &\leqslant \frac{C}{R}\int_0^1 \sup_{t\in H_R^\circ} |h(t)-g(t)|_\infty\,\mathrm{d}\epsilon\\ &= \frac{C}{R}\,\|h-g\|_\infty.\end{aligned}$$

It follows that

$$\begin{aligned}\left\|\int_0^t \mathrm{e}^{\Lambda u}\,(\Phi(h(u))-\Phi(g(u)))\,\mathrm{d}u\right\|_\infty &\leqslant \frac{C}{R}\,\|h-g\|_\infty\int_0^t \mathrm{e}^{\Lambda \mathrm{Re}\, u}\,\mathrm{d}u\\ &\leqslant \frac{C}{R}\,\|h-g\|_\infty\,\mathrm{e}^{\Lambda \mathrm{Re}\, t}\int_0^t \mathrm{e}^{\Lambda \mathrm{Re}\,(u-t)}\,\mathrm{d}u\\ &\leqslant \frac{C}{R}\,\|h-g\|_\infty\,\mathrm{e}^{\Lambda \mathrm{Re}\, t}|t| \;\leqslant\; C\,\|h-g\|_\infty\,\mathrm{e}^{\Lambda \mathrm{Re}\, t}\end{aligned}$$

and

$$\|\Omega_R(h)-\Omega_R(g)\|_\infty \;\leqslant\; C\,\|h-g\|_\infty.$$

This shows that Ω_R is indeed contracting on $\mathcal{B}_R(c, B)$. Since $\mathcal{B}_R(c, B)$ is complete as a closed bounded subspace of $\mathcal{A}(H_R^\circ)$, we conclude that $g_1, g_2, \ldots$ converges to a fixed point $\varphi \in \mathcal{B}_R(c,B)$ of Ω_R. □

2.4. A refinement for larger compact half disks

It turns out that the condition (6) on the Jacobian is not really needed. Moreover, the solution can further be extended to the closure H_R of H_R°.

THEOREM 4. *Let $R > 0$ and $B \in (\mathbb{R}^>)^d$ be such that for all $z \in \mathbb{C}$, we have*

$$|z - c| \leqslant B \implies |\Phi(z)| R \leqslant B. \tag{8}$$

Then, for any choice of Λ, the initial value problem

$$\dot\varphi + \Lambda \varphi = \Phi(\varphi), \qquad \varphi(0) = c \tag{9}$$

admits a unique analytic solution on H_R. In addition, we have $|\varphi(t) - c| \leqslant B$ for all $t \in H_R$.

Proof. Let $\Omega_R : \mathcal{A}(H_R^\circ)^d \to \mathcal{A}(H_R^\circ)^d$ be as in the proof of Theorem 3 and notice that Ω_R still maps $\mathcal{B}_R(c,B)$ into itself. Now consider the sequence $g_1, g_2, \ldots$ with $g_k := \Omega_R^k(c) \in \mathcal{B}_R(c,B)$. Applying Montel's theorem, we obtain a subsequence g_{k_i} that converges uniformly to a limit $g_\infty \in \mathcal{B}_R(c,B)$ on every compact subset of H_R°.

We claim that g_∞ is a fixed point of Ω_R. Indeed, for a sufficiently small $R' > 0$ with $R' \leqslant R$, we have $|J_\Phi(z)|_\infty R' \leqslant 1/2$ for all $z \in \mathbb{C}$ with $|z - c| \leqslant B$. As in the proof of Theorem 3, this means that $\Omega_{R'}$ is contracting on $\mathcal{B}_{R'}(c, B)$. Consequently, the restrictions $g_1^\flat, g_2^\flat, \ldots$ of the functions $g_1, g_2, \ldots$ to $H_{R'}^\circ$ with $g_k^\flat = \Omega_{R'}^k(c)$ tend to a fixed point $g_\infty^\flat$, and so does the subsequence $g_{k_i}^\flat$. Now this fixed point $g_\infty^\flat$ coincides with g_∞ on $H_{R'}^\circ$ and also solves the initial value problem (7). By analytic continuation, g_∞ therefore solves the same initial value problem on H_R°, which completes the proof of our claim.

It remains to be shown that g_∞ can be extended to an analytic function φ on H_R that satisfies $|\varphi(t) - c| \leqslant B$ for all $t \in H_R$. Since $g_\infty' = \Phi(g_\infty) - \Lambda g_\infty$ on H_R°, we first notice that $|g_\infty'|$ is bounded on H_R°, whence g_∞ continuously extends to a function φ that is defined on the whole of H_R, and we clearly have $|\varphi(t) - c| \leqslant B$ for all $t \in H_R$.

Now let us consider the unique solution $\psi\langle 0 \,|\, \varsigma\rangle$ to the initial value problem $\psi'(t) + \Lambda\, \psi(t) = \Phi(\psi(t))$ with $\psi(0) = \varsigma$. Theorem 1 provides us with a lower bound $\varrho\langle 0 \,|\, \varsigma\rangle > 0$ for the radius of convergence of $\psi\langle 0 \,|\, \varsigma\rangle$, which depends continuously on ς. By the compactness of H_R, it follows that there exists some $\varepsilon > 0$ with $\varrho\langle 0 \,|\, \varphi(t)\rangle \geqslant \varepsilon$ for all $t \in H_R$. Now consider t on the boundary ∂H_R of H_R and let $t' \in H_R^\circ$ be such that $|t' - t| < \varepsilon$. Then $\varrho\langle 0 \,|\, \varphi(t')\rangle \geqslant \varepsilon$ and $\psi\langle 0 \,|\, \varphi(t')\rangle(z) = \varphi(t' + z)$ on some neighbourhood of t'. We conclude that $\psi\langle 0 \,|\, \varphi(t')\rangle(z - t')$ is an analytic extension of φ to the open disk $\{u \in \mathbb{C} : |u - t'| < \varepsilon\}$ that contains t. □

3. COMPUTING FORMAL POWER SERIES SOLUTIONS

Before we present algorithms for the numeric integration of (1), let us first consider the question how to compute formal power series solutions. Since we will later be considering power series solutions at various times t, we will use the letter f instead of φ for such solutions. Notice that a vector $f \in \mathbb{C}[[z]]^d$ of d formal power series can also be regarded as a formal power series $f = f_0 + f_1 z + f_2 z^2 + \cdots \in \mathbb{C}^d[[z]]$ with coefficients in $\mathbb{C}^d$.

Let $c \in \mathbb{C}^d$ be the initial condition. Setting $\Psi(f) = \Phi(f) - \Lambda f$, we are thus interested in the computation of truncated power series solutions to the equation

$$\frac{\partial f}{\partial z} = \Psi(f), \qquad f_0 = c. \tag{10}$$

Alternatively, this equation can be rewritten in integral form

$$f = c + \int \Psi(f), \tag{11}$$

where the integration operator $\int$ sends $g \in \mathbb{C}[[z]]^d$ to $g_0 z + \frac{1}{2} g_1 z^2 + \frac{1}{3} g_2 z^3 + \cdots$.

In this section, we recall several approaches to the computation of power series solutions to such equations. The efficiency of each method can be measured in terms of the number of field operations in $\mathbb{C}$ that are required to compute the first n terms of the solution. For the time being, we assume that all computations are done at a fixed bit precision p, so that operations in $\mathbb{C}$ are done at unit cost. One may also take into account the number s of scalar operations that are necessary for the evaluation of Ψ. A further refinement is to separately count the number s_{mul} of multiplications that are needed. For instance, if $d = 1$ and $\Psi(f) = (f \times f + 3) \times f - 100 \times f$, then we have $s = 5$ and $s_{\mathrm{mul}} = 3$. We always assume that $d = O(s)$.

Iterative method. For the first method, we systematically compute with truncated power series of order n in $\mathbb{C}[z] / (z^n)$, which are also called *jets* of order n. One addition in $\mathbb{C}[\varepsilon] / (\varepsilon^n)$ reduces to $O(n)$ additions in $\mathbb{C}$ and similarly for scalar multiplications with an element in $\mathbb{C}$. A naive multiplication in $\mathbb{C}[\varepsilon] / (\varepsilon^n)$ requires $O(n^2)$ operations in $\mathbb{C}$, although this cost can be reduced to $O(n \log n)$ using FFT techniques. The integration operator $\int$ sends $(g_0 + \cdots + g_{n-1} z^{n-1}) \bmod z^n$ to $(g_1 + \cdots + g_{n-2} / (n-1)\, z^{n-1}) \bmod z^n$ and can be computed in linear time.

Now assume that $\tilde{f} \in (\mathbb{C}[z] / (z^n))^d$ is an approximate solution to (11) whose first $k < n$ terms are correct. In other words, if f is the actual solution and $\check{f} \in \mathbb{C}[z]^d$ is a preimage of $\tilde{f}$ with $\tilde{f} = (\check{f} \bmod z^n)$, then $\check{f} - f = O(z^k)$. Given such an approximation, one iteration

$$\tilde{f} := c + \int \Psi(\tilde{f}) \tag{12}$$

of (11) yields a new approximation whose first $k + 1$ terms are correct. Starting with $\tilde{f} := (0 \bmod z^n)$, we thus obtain a solution modulo z^n of (11) after at most n iterations. The total cost of this computation is bounded by $O(s_{\mathrm{mul}} n^3 + s n^2)$, or by $O(s_{\mathrm{mul}} n^2 \log n + s n^2)$ when using FFT-techniques.

Recurrence relations. Since $\Psi(f)$ is a polynomial, it is built from the components of f using additions, multiplications, and scalar multiplications with constants in $\mathbb{C}$. For sums, scalar products, general products, and integrals of power series, we may extract their coefficients in z^k using the following rules

$$(g + h)_k = g_k + h_k \tag{13}$$
$$(\alpha g)_k = \alpha g_k \tag{14}$$
$$(g h)_k = g_0 h_k + g_1 h_{k+1} + \cdots + g_k h_0 \tag{15}$$
$$(\textstyle\int g)_{k+1} = \frac{1}{k+1} g_k. \tag{16}$$

Applying this recursively to the polynomial expression $\Psi(f)$, the iteration (12) yields a recursion relation

$$f_{k+1} := \frac{1}{k+1} (\Psi(f))_k \tag{17}$$

that allows us to compute f_k from f_0, by induction on k. Proceeding in this way, the computation of the solution $f_0 + \cdots + f_{n-1} z^{n-1}$ at order n requires $O(s_{\mathrm{mul}} n^2 + s\, n)$ operations.

The lazy power series approach. The above approach of computing the coefficients $f_1, f_2, \ldots$ successively using the recursion relation (17) can be reformulated elegantly in the framework of lazy power series computations. The idea is to regard a power series $g \in \mathbb{C}[[z]]$ as given by its stream $g_0, g_1, g_2, \ldots$ of coefficients. Basic arithmetic operations on power series are implemented in a *lazy* manner: when computing the k-th coefficient of a sum $g + h$, a product $g\,h$, or an integral $\int g$, we compute "just the coefficients that are needed" from g and h. The natural way to do this is precisely to use the relations (13–16).

The lazy power series approach has the important property that the k-th coefficient $(g\,h)_k$ of a product (say) becomes available as soon as $g_0, \ldots, g_k$ and $h_0, \ldots, h_k$ are given. This makes it possible to let g and h depend on the result $g\,h$, as long as the computation of g_k and h_k only involves previous coefficients $(g\,h)_0, \ldots, (g\,h)_{k-1}$ of $g\,h$. As a consequence, the fixed point equation (11) can be solved simply by evaluating the right hand side using lazy power series arithmetic. Indeed, the k-th coefficient $(\int \Phi(f))_k$ only depends on previous coefficients $f_0, \ldots, f_{k-1}$ of f.

The relaxed power series approach. The main drawback of the lazy power series approach is that the computation of a product at order n using (15) requires $O(n^2)$ operations. Here we recall that FFT techniques allow us to compute a product in $\mathbb{C}[z]/(z^n)$ using only $O(n \log n)$ operations.

One essential observation is that, in order to solve (11) using the lazy power series approach, we only relied on the fact that each coefficient $(g\,h)_k$ becomes available as soon as $g_0, \ldots, g_k$ and $h_0, \ldots, h_k$ are given. In fact, it is possible to design faster multiplication methods that still have this property; such methods are said to be *relaxed* or *on-line*. A relaxed multiplication method that computes a product at order n in time $O(n \log^2 n)$ was presented in [8] and can be traced back to [6]. An even faster algorithm of time complexity $n \log n\, e^{O(\sqrt{\log \log n})}$ was given in [9].

Denoting by $\mathsf{R}(n) = n\ (\log n)^{1+o(1)}$ the cost of relaxed multiplication at order n, the resolution of (11) at order n now requires only $O(s_{\mathrm{mul}}\, \mathsf{R}(n) + s\, n)$ operations.

Newton's method. Yet another idea to speed up computations is to replace the iteration (12) with a Newton-style iteration with faster convergence. This idea was first used by Brent and Kung [2] to show that (11) can be solved at order n in time $O(n \log n)$. However, this complexity analysis does not take into account the dependence on d and s. In particular, the dependence on d of Brent and Kung's method is exponential [8]. Faster algorithms were proposed in [26, 1], based on the simultaneous computation of the solution f and its first variation. This allows for the computation of a solution of (10) at order n in time $O((s_{\mathrm{mul}} + d)\, d\, n \log n + s\, d\, n)$. Whenever $d = o(\log n)$, the computation time can be further reduced to $O((s_{\mathrm{mul}} + d)\, n \log n + s\, n)$ [10]. For small d, this leads to the asymptotically most efficient method for solving (10). For large d, the computation of the first variation induces a $\Theta(d)$ overhead, and the relaxed method usually becomes more efficient.

4. NUMERICAL INTEGRATION USING TAYLOR EXPANSIONS

Let $\Psi(\varphi) = \Phi(\varphi) - \Lambda\, \varphi$ as before. The aim of this section is to present a naive algorithm for the numerical integration of the differential equation

$$\dot{\varphi} \;=\; \Psi(\varphi), \tag{18}$$

based on the computation of truncated power series solutions at successive times $t_0 = 0 < t_1 < t_2, \ldots$. We will use a fixed expansion order n, a fixed bit precision $p \geqslant 24$, but an adaptive step size $t_1 - t_0, t_2 - t_1, \ldots$. At $t = 0$, we are given an initial condition $\varphi(0) = c \in \mathbb{C}^d$ and our aim is to find a good numeric approximation for $\varphi(t)$ at time $t = T > 0$. The algorithm is not designed to be efficient when the equation gets stiff and we will present a heuristic discussion on what goes wrong when this happens.

4.1. Naive integration using Taylor series

Let us write $f\langle t\rangle = \varphi(t + z) \in \mathbb{C}[[z]]^d \cong \mathbb{C}^d[[z]]$ for the Taylor series expansion of φ at time t and $f\langle t\rangle_{;n} \in \mathbb{C}^d[z] \cong \mathbb{C}[z]^d$ for its truncation at order n. In other words, setting

$$\mathbb{C}[z]_{;n} = \{g \in \mathbb{C}[z] : \deg g < n\},$$

we have $f\langle t\rangle_{;n} \in \mathbb{C}[z]_{;n}^d$ and

$$\varphi(t+z) \;=\; f\langle t\rangle_{;n} + O(z^n) \;=\; f\langle t\rangle_0 + f\langle t\rangle_1 z + \cdots + f\langle t\rangle_{n-1} z^{n-1} + O(z^n).$$

If the time t at which we expand φ is clear from the context, then we will simply drop the postfix $\langle t\rangle$ and write f instead of $f\langle t\rangle$. Conversely, for any other quantities that implicitly depend on t, we will use the postfix $\langle t\rangle$ to make this dependence explicit.

So let $f = f\langle t\rangle$ be the power series expansion of φ at a fixed time t. In view of (18), this power series satisfies the equation

$$\frac{\partial f}{\partial z} \;=\; \Psi(f). \tag{19}$$

In the previous section, we have recalled various methods for computing truncated power series solutions $f_{;n}$ as a function of the initial condition $f_0 = \varsigma = \varsigma\langle t\rangle = \varphi(t)$ at time t. In what follows, we will use any of these algorithms as a black box, and show how to device a numerical integration scheme from that.

Obviously, given $\varphi(t)$ and an appropriate step size $\delta = \delta\langle t\rangle$ at time t, the idea is to compute $f_{;n} = f\langle t\rangle_{;n}$ and simply evaluate

$$\varphi(t + \delta) \approx f\langle t\rangle_{;n}(\delta) = f\langle t\rangle_0 + f\langle t\rangle_1 \delta + \cdots + f\langle t\rangle_{n-1} \delta^{n-1}.$$

We next continue with $t + \delta$ in the role of t and with a suitably adapted step size δ. The main question is therefore to find a suitable step size δ. Now the expected order of magnitude of $f\langle t\rangle_{;n}(\delta)$ is given by

$$M \;=\; \max_{k<n} |f_k|_\infty \, \delta^k. \tag{20}$$

Since we are computing with p bits of precision, we wish to keep the relative error of our numeric integration scheme below 2^{-p}, approximately. Therefore, we need to ensure that the truncation error

$$E \;=\; \left| \sum_{k \geqslant n} f_k \delta^k \right|_\infty \tag{21}$$

remains bounded by $M\, 2^{-p}$. Although we have not computed any of the coefficients f_k for $k \geqslant n$, a reasonable approximate upper bound for E is given by

$$E_{\text{guess}} \;=\; \max_{n-\gamma \leqslant k < n} |f_k|_\infty \, \delta^k, \tag{22}$$

where $\gamma > 0$ is a small positive integer, called the *number of guard terms*. In order to protect ourselves against the occasional vanishing of f_{n-1}, it is wise to take $\gamma > 1$. Nevertheless, a small value such as $\gamma = 2$ or $\gamma = 3$ should usually provide acceptable upper estimates for E. We now simply take the step size δ to be maximal with $E_{\text{guess}} \leqslant M\, 2^{-p}$. This leads to the following algorithm for the numerical integration of (18):

Algorithm 1
Input: an initial condition $c \in \mathbb{C}^d$ and $T > 0$
Output: a numerical approximation for $\varphi^*(T)$, where φ^* satisfies (18) with $\varphi^*(0) = c$

$\varsigma := c, t := 0$
while $t < T$ **do**
  Compute the truncated solution $f_{;n}$ to (19) with $f_0 = \varsigma$
  Let δ be maximal such that $E_{\text{guess}} \leqslant M 2^{-p}$, with M and E_{guess} as in (20) and (22)
  $\delta := \min(\delta, T - t)$
  $\varsigma := f_{;n}(\delta), t := t + \delta$
return ς

4.2. Step size and error analysis

Let φ^* denote the exact solution of (18) with $\varphi^*(0) = c$ and let φ denote the computed approximation by Algorithm 1. In order to analyze various aspects of the algorithm, it is instructive to look at the radius of convergence ϱ^* of φ^* at time t. Roughly speaking, with the notations from the previous subsection, the coefficients f_k^* of the power series expansion $f^* = f^*\langle t \rangle$ of φ at time t grow as

$$|f_k^*|_\infty \approx M (\varrho^*)^{-k}.$$

Ideally speaking, if we were able to compute $f_{;n}$ with sufficient accuracy, then the coefficients f_k should grow in a similar way. Setting δ^* for the step size in this ideal situation, we would then expect that $E_{\text{guess}} \approx E \approx M (\delta^* / \varrho^*)^n \approx M 2^{-p}$, whence

$$\delta^* \approx \varrho^* 2^{-p/n}. \tag{23}$$

This suggests to take the expansion order n to be proportional to p, after which the step size δ^* should be proportional to the radius of convergence ϱ^*.

Let us pursue this line of wishful thinking a little further. Theorem 4 implies that φ^* is analytic on a compact half disk H_R that is independent of Λ. In particular, we get that $\varrho^*\langle t \rangle \geqslant t$ for $t \leqslant R/2$. It can also be shown that the radius of convergence of φ^* at the origin is of the order $\varrho^*\langle 0 \rangle \approx \lambda_d^{-1}$ or more. For stiff equations, we typically have $R \gg \lambda_d^{-1}$. In order to integrate the equation until a time $T \leqslant R/2$, we thus hope for a step size that increases geometrically from λ_d^{-1} to $T 2^{-p/n}$. The entire integration would then require approximately $2^{p/n} \log(T \lambda_d)$ steps. Using the most efficient algorithms from section 3, each step requires $\tilde{O}(s\ n)$ floating point operations (here the "flat Oh" notation $\tilde{O}(\xi)$ stands for $O(\xi\ (\log \xi)^{O(1)})$). Since additions and multiplications of p bit floating point numbers can be performed in time $\tilde{O}(p)$, the overall bit cost of the entire integration would thus be bounded by $\tilde{O}(s\, n\, p\, 2^{p/n} \log(T \lambda_d))$.

But are we indeed able to compute $f_{;n}$ with enough accuracy in order to ensure that the coefficients f_k grow according to $|f_k|_\infty \approx M (\varrho^*)^{-k}$? Let us carefully analyze each of the sources of error for one step of our integration scheme. By construction, we ensured the truncation error to be of the order $|(f - f_{;n})(\delta)|_\infty \approx M 2^{-p}$. One of the most intrinsic sources of error comes from the initial condition f_0: since we are computing with a precision of p bits, the mere representation of f_0 induces a relative error of the order of 2^{-p}. Even when computing $f_{;n}$ from f_0 with infinite precision, this intrinsic source of error cannot be removed.

Let us now study how the error in the initial condition affects the errors in the other coefficients f_k. For this, we need to investigate the first variation of φ, which describes the sensitivity of the flow to the initial condition. More precisely, let $f\langle t\,|\,\varsigma\rangle$ be the power solution of (19) at time t as a function of the initial condition $f\langle t\,|\,\varsigma\rangle_0 = \varsigma \in \mathbb{C}^d$. Then the first variation $V\langle t\,|\,\varsigma\rangle$ is the $d\times d$ matrix with entries $V\langle t\,|\,\varsigma\rangle^{[i,j]} = \partial f\langle t\,|\,\varsigma\rangle^{[i]} / \partial \varsigma^{[j]}$. Dropping $\langle t\,|\,\varsigma\rangle$ suffixes when they are clear from the context, the first variation satisfies the linear differential equation

$$\frac{\partial V}{\partial z} \;=\; (J_\Phi(f) - \Lambda)\, V, \qquad V_0 = \mathrm{Id}_d.$$

Here we recall that $J_\Phi = \partial\Phi / \partial F$ stands for the Jacobian matrix of Φ. If our equation is very stiff, then J_Φ is small with respect to Λ, which leads to the extremely crude approximation $V \approx \mathrm{e}^{-\Lambda z}$ for the first variation.

Now the relative error in the initial condition $\varsigma = f_0$ is at best 2^{-p}, as explained above, which means that $|\varsigma - \varsigma^*|_\infty \approx M\,2^{-p}$ for $\varsigma^* = f_0^*$. In combination with the relation $f - f^* \approx V\,(\varsigma - \varsigma^*) \approx \mathrm{e}^{-\Lambda z}\,(\varsigma - \varsigma^*)$, this leads to the following errors for the other coefficients:

$$|f_k - f_k^*|_\infty \;\approx\; \frac{\lambda_d^k}{k!} M\,2^{-p}.$$

Now if $\lambda_d \gg (\varrho^*)^{-1}$, then the error $|f_n - f_n^*|_\infty \approx M\,\lambda_d^n\,2^{-p}/n!$ dominates the actual value $|f_n^*|_\infty \approx M\,(\varrho^*)^{-n}$, which yields $E \approx M\,(\lambda_d\,\delta_t)^n\,2^{-p}/n!$ instead of $E \approx M\,(\delta/\varrho^*)^n\,2^{-p}$. When chosing our step size δ such that $E \approx M\,2^{-p}$, as in Algorithm 1, this yields $(\lambda_d\,\delta)^n \approx n!$ and

$$\delta \;\approx\; \frac{\sqrt[n]{n!}}{\lambda_d} \;\approx\; \frac{n}{\mathrm{e}\lambda_d}, \tag{24}$$

instead of the desired step size $\delta \approx \varrho^*\,2^{-p/n}$. The actual bit cost of the complete integration is therefore bounded by $\tilde{O}(s\,n\,p\,T/\lambda_d)$ instead of $\tilde{O}(s\,n\,p\,2^{p/n}\log(T\lambda_d))$.

An interesting aspect of this analysis is the fact that the step size δ still turns out to be n/e times larger than λ_d^{-1}, whence larger orders n allow for larger step sizes. However, this comes at the expense of a larger relative error than 2^{-p}. Indeed, we have

$$|f_k\,\delta^k - f_k^*\,\delta^k|_\infty \;\approx\; \frac{(\lambda_d\,\delta)^k}{k!} M\,2^{-p} \;\approx\; \left(\frac{n}{k}\right)^k M\,2^{-p}.$$

This error is maximal for $k_{\max} \approx n/\mathrm{e}$, in which case we have

$$|f_{k_{\max}}\,\delta^{k_{\max}} - f_{k_{\max}}^*\,\delta^{k_{\max}}|_\infty \;\approx\; \mathrm{e}^{n/\mathrm{e}} M\,2^{-p}.$$

This means that the relative error for one step of our integration method is $2^{n/(\mathrm{e}\log 2)-p}$ instead of 2^{-p}. In other words, we have "sacrificed" $n/(\mathrm{e}\log 2)$ bits of precision, so that the method admits an "effective precision" of only $p - n/(\mathrm{e}\log 2)$ bits.

The last source of errors for Algorithm 1 comes from rounding errors during the computation of $f_{;n}$ from f_0. The nature of these errors depends on the particular method that we used for computing the power series $f_{;n}$. Nevertheless, for most methods, the rounding errors only contribute marginally to the total error. This is due to the fact that $|f_k\,\delta^k|_\infty\,2^{-p} \ll |f_k\,\delta^k - f_k^*\,\delta^k|_\infty$ for $k > 0$, so the rounding errors are absorbed by the errors induced by the error in the initial condition f_0.

5. Fighting stiffness

Let us continue with the notations from section 4 and its subsection 4.2. In particular, we assume that the exact solution φ^* to (1) with initial condition $\varphi^*(0) = c$ is analytic on the compact half disk H_R. We also denote by K an upper bound for $|\varphi^*|_\infty$ on H_R, so that $\varrho\langle t\rangle^* \geqslant t$ and

$$|f\langle t\rangle_k^*|_\infty \leqslant K t^{-k}$$

for all $t \leqslant R/2$ and $k \in \mathbb{N}$, by Cauchy's theorem. From now on, we assume that $R \gg \lambda_d^{-1}$, so that Algorithm 1 only allows us to use a step size of the order (24) instead of (23). The aim of this section is to present a new way to compute truncated power series solutions of the equation

$$\frac{\partial f}{\partial z} + \Lambda f = \Phi(f) \tag{25}$$

at time t, with the property that $f_{;n}$ is a good approximation of the true truncated solution $f_{;n}^*$. In a similar way as in section 4, we will then use this to derive an algorithm for the numerical integration of (1). Contrary to before, the property that $|f_k|_\infty \approx |f_k^*| \leqslant K/t^k$ for $k < n$ allows us to take step sizes of the desired order (23) for $t \leqslant R/2$.

5.1. Integrating stiff equations using Taylor series

We stress once more that the reduced step size for Algorithm 1 is a consequence of our choice to compute the truncated solution $f_{;n}$ of (25) in terms of the initial condition $f_0 = \varsigma$ (that can only be known approximately) and *not* of the choice of the particular algorithm that is used for this computation.

Indeed, as explained in section 4.2, only a small change in ς of the order of $|\Delta\varsigma|_\infty \approx |\varsigma|_\infty 2^{-p} \approx M 2^{-p}$ can have a dramatic effect on the solution $f_{;n}$, since the coefficient f_n can change by as much as $|\Delta f_n|_\infty \approx M 2^{-p} \lambda_d^n / n!$. Since $|f_n^*|_\infty \leqslant K/t^n \ll M 2^{-p} \lambda_d^n / n!$ for large $t \gg \lambda_d^{-1}$, it follows that a minor change in ς leads to a completely incorrect computation of the coefficients f_k with k close to n.

For small t, the error $|\Delta f_n|_\infty$ generally does remain bounded by the actual value $|f_n^*|_\infty$. In the theory of stiff differential equations, it is customary to say that the system is in a *transient state* for such t. As soon as the error $|\Delta f_n|_\infty \approx M 2^{-p} \lambda_d^n / n!$ exceeds $|f_n^*|$, we say that the system has reached its *steady state* for the largest eigenvalue λ_d of Λ. When this happens, the system may still be in a transient state for some of the other eigenvalues $\lambda_i < \lambda_d$. This motivates the definition of the *critical index* $\chi \in \{0, \ldots, n\}$ as being the largest index i such that we are in a transient state for λ_i; we take $\chi = 0$ if we reached the steady state for all eigenvalues λ_i.

The concepts of transient state, steady state, and critical index are deliberately somewhat vague. As a tentative definition, we might say that we reached the steady state for λ_i if $|f_n^*|_\infty \leqslant M 2^{-p} \lambda_i^n / n!$. However, for computational purposes, it is convenient to interpret this inequality as an approximate one. The crucial property of reaching the steady state for λ_i is that the smallness of $|(f_n^*)^{[i]}|_\infty$ essentially determines the i-th component $\varsigma^{[i]}$ of the initial condition up to the last p-th bit.

The main idea behind our algorithm is to use this property as a "boundary condition" for the computation of $f_{;n}$. More precisely, we boldly impose the boundary conditions $f_n^{[\chi+1]} = \cdots = f_n^{[d]} = 0$ as a replacement for the initial conditions $f_0^{[\chi+1]} = \varsigma^{[\chi+1]}, \ldots, f_0^{[d]} = \varsigma^{[d]}$, while keeping the remaining initial conditions $f_0^{[1]} = \varsigma^{[1]}$, ..., $f_0^{[\chi]} = \varsigma^{[\chi]}$ for the transient states. In other words, we wish to find the truncated solution $f_{;n}$ of the system

$$\frac{\partial f}{\partial z} + \Lambda f = \Phi(f), \qquad \begin{cases} f_0^{[i]} = \varsigma^{[i]} & \text{if } i \leqslant \chi \\ f_n^{[i]} = 0 & \text{if } i > \chi \end{cases}. \tag{26}$$

We will call $f_n^{[i]} = 0$ the *steady-state condition* for λ_i and (26) a *steady-state problem*.

In order to solve (26), it is natural to adapt the iterative method from section 3 and introduce the operator $\Xi: \mathbb{C}[z]_{;n}^d \to \mathbb{C}[z]_{;n}^d$ with

$$\Xi^{[i]}(g) = \begin{cases} \varsigma^{[i]} + \int (\Phi^{[i]}(g) - \lambda_i g^{[i]}) & \text{if } i \leqslant \chi \\ \lambda_i^{-1} (\Phi^{[i]}(g) - \partial g^{[i]}) & \text{if } i > \chi \end{cases} \mod z^n. \tag{27}$$

The actual arithmetic is performed over $\mathbb{C}[z] / (z^n)$, but it will be more convenient to view Ξ as an operator from $\mathbb{C}[z]_{;n}^d$ into itself. The computation of $f_{;n}$ as a fixed point of Ξ leaves us with two questions: how to chose a suitable *ansatz* for the iteration $g := \Xi(g)$ and how to determine the critical index χ?

For the *ansatz*, we go back to the solution $f\langle t'\rangle_{;n}$ from the previous step at time $t' = t - \delta'$ and simply use $f\langle t'\rangle_{;n}(\delta' + z)$ as a first approximation of the solution $f\langle t\rangle_{;n}(z)$ at time t. For $t = 0$, we fall back to the traditional method from section 4. For the initial and steady-state conditions to "propagate to the other end", at least n iterations are required in order to find a fixed point of Ξ, whereas $2n$ iterations usually suffice. One may thus take $f\langle t\rangle_{;n} := \Xi^{2n}(f\langle t'\rangle_{;n}(\delta' + z))$.

As to the critical index χ, we determine it as a function of the step size δ through

$$\chi = \max \left\{ i \leqslant d : \lambda_i \delta \leqslant \frac{n}{\mathrm{e}} \right\}. \tag{28}$$

This choice is meant to ensure the fastest convergence of the iteration $g := \Xi(g)$ and we used n / e as an approximation of $\sqrt[n]{n!}$. The step size δ itself is chosen in the same way as in section 4. Since $f_{;n}$ is not yet available for the computation of δ, we may use the previous step size δ' as an approximation for δ in (28). Optionally, one may also wish to implement a few additional sanity checks in order to ensure convergence of the Ξ-iteration and decrease the step size in case of failure. One useful such check is to verify that $|f\langle t\rangle_0 - f\langle t'\rangle_{;n}(\delta')|_\infty \leqslant M 2^{-p/2}$.

Altogether, this leads to the following algorithm:

Algorithm 2
Input: an initial condition $c \in \mathbb{C}^d$ and $T > 0$
Output: a numerical approximation for $\varphi^*(T)$, where φ^* satisfies (1) with $\varphi^*(0) = c$

$\varsigma := c$, $t := 0$
Compute the truncated solution $f_{;n}$ to (19) with $f_0 = \varsigma$
while $t < T$ **do**
  Let δ be maximal such that $E_{\text{guess}} \leqslant M 2^{-p}$, with M and E_{guess} as in (20) and (22)
  $\delta := \min(\delta, T - t)$
  $\varsigma := f_{;n}(\delta)$, $t := t + \delta$
  $\chi := \max \{ i \leqslant d : \lambda_i \delta \leqslant \frac{n}{\mathrm{e}} \}$
  Replace $f_{;n}$ by the approximate fixed point $\Xi^{2n}(f_{;n}(\delta + z))$ of Ξ as in (27)
return ς

5.2. Convergence of the Ξ-iteration

Let us now study the convergence of Ξ as a mapping on the dn-dimensional vector space $\mathbb{C}[z]_{;n}^d$ over $\mathbb{C}$. We define a norm $\|\cdot\|_*$ on this space by

$$\begin{aligned} \|g\|_* &= \max_{k<n} \mu_k |g_k|_\infty = \max_{1 \leqslant i \leqslant d, k < n} \mu_k |g_k^{[i]}| \\ \mu_k &= \frac{k!}{\lambda_*^k} \\ \lambda_* &= \sqrt{\lambda_\chi \lambda_{\chi+1}}. \end{aligned}$$

For a linear map on this space, represented as a matrix M, we have the corresponding matrix norm

$$\|M\|_* \;=\; \sup_{\|g\|_*=1} \|Mg\|_*.$$

Recall that J_Φ and J_Ξ stand for the Jacobian matrices of Φ and Ξ.

THEOREM 5. *Assume that g is a fixed point of Ξ and that $A>0$ and $\varrho > n/\lambda_*$ are constants such that $|(J_\Phi(g))_k|_\infty \leqslant A\varrho^{-k}$ for all $i,j=1,\ldots,d$ and $k=0,\ldots,n-1$. Then*

$$\|J_\Xi(g)\|_* \;\leqslant\; \frac{\lambda_\chi}{\lambda_*} + \frac{A}{\lambda_*}\left(1-\frac{n}{\lambda_*\varrho}\right)^{-1}.$$

Proof. Consider an infinitesimal perturbation Δg of g with $\Xi(g+\Delta g) = \Xi(g) + \Delta\Xi(g)$. Given $i \leqslant \chi$ and $0<k<n$, we have

$$\begin{aligned}
\mu_k |(\Delta\Xi(g))_k^{[i]}| \;&=\; \frac{\mu_k}{k} |(J_\Phi(g)\,\Delta g)_{k-1}^{[i]} - \lambda_i\,(\Delta g)_{k-1}^{[i]}| \\
&\leqslant\; \frac{\mu_k}{k}\left(\sum_{0\leqslant\ell\leqslant k-1} |(J_\Phi(g)_\ell\,\Delta g)_{k-1-\ell}|_\infty + \lambda_i\,|(\Delta g)_{k-1}|_\infty\right) \\
&\leqslant\; \frac{\mu_k}{k}\left(\sum_{0\leqslant\ell\leqslant k-1} \frac{A}{\varrho^\ell}\frac{\|\Delta g\|_*}{\mu_{k-1-\ell}} + \lambda_i\frac{\|\Delta g\|_*}{\mu_{k-1}}\right) \\
&=\; \left(\sum_{0\leqslant\ell\leqslant k-1} \frac{A\,\mu_k}{k\,\mu_{k-1}}\frac{\mu_{k-1}}{\mu_{k-1-\ell}\,\varrho^\ell} + \frac{\lambda_i\,\mu_k}{k\,\mu_{k-1}}\right)\|\Delta g\|_* \\
&=\; \left(\sum_{\ell\leqslant k-1} \frac{A}{\lambda_*}\frac{(k-1)!}{(k-1-\ell)!\,(\lambda_*\varrho)^\ell} + \frac{\lambda_i}{\lambda_*}\right)\|\Delta g\|_* \\
&\leqslant\; \left(\sum_{\ell\leqslant k-1} \frac{A}{\lambda_*}\left(\frac{n}{\lambda_*\varrho}\right)^\ell + \frac{\lambda_i}{\lambda_*}\right)\|\Delta g\|_* \\
&\leqslant\; \left(\frac{A}{\lambda_*}\left(1-\frac{n}{\lambda_*\varrho}\right)^{-1} + \frac{\lambda_\chi}{\lambda_*}\right)\|\Delta g\|_*
\end{aligned}$$

Similarly, given $i>\chi$ and $k<n$, we have

$$\begin{aligned}
\mu_k |(\Delta\Xi(g))_k^{[i]}| \;&=\; \frac{\mu_k}{\lambda_i} |(J_\Phi(g)\,\Delta g)_k^{[i]} - (k+1)\,(\Delta g)_{k+1}^{[i]}| \\
&\leqslant\; \frac{\mu_k}{\lambda_i}\left(\sum_{\ell\leqslant k} |(J_\Phi(g)_\ell\,\Delta g)_{k-\ell}|_\infty + (k+1)\,|(\Delta g)_{k+1}|_\infty\right) \\
&\leqslant\; \frac{\mu_k}{\lambda_i}\left(\sum_{\ell\leqslant k} \frac{A}{\varrho^\ell}\frac{\|\Delta g\|_*}{\mu_{k-\ell}} + (k+1)\frac{\|\Delta g\|_*}{\mu_{k+1}}\right) \\
&=\; \left(\sum_{\ell\leqslant k} \frac{A}{\lambda_i}\frac{\mu_k}{\mu_{k-\ell}\,\varrho^\ell} + \frac{(k+1)\,\mu_k}{\lambda_i\,\mu_{k+1}}\right)\|\Delta g\|_* \\
&=\; \left(\sum_{\ell\leqslant k} \frac{A}{\lambda_i}\frac{k!}{(k-\ell)!\,(\lambda_*\varrho)^\ell} + \frac{\lambda_*}{\lambda_i}\right)\|\Delta g\|_* \\
&\leqslant\; \left(\sum_{\ell\leqslant k} \frac{A}{\lambda_i}\left(\frac{n}{\lambda_*\varrho}\right)^\ell + \frac{\lambda_*}{\lambda_i}\right)\|\Delta g\|_* \\
&\leqslant\; \left(\frac{A}{\lambda_*}\left(1-\frac{n}{\lambda_*\varrho}\right)^{-1} + \frac{\lambda_\chi}{\lambda_*}\right)\|\Delta g\|_*
\end{aligned}$$

Putting both relations together, we obtain $\|\Delta\Xi(g)\|_* \leqslant \left(\frac{A}{\lambda_*}\left(1-\frac{n}{\lambda_*\varrho}\right)^{-1}+\frac{\lambda_\chi}{\lambda_*}\right)\|\Delta g\|_*$. □

Assume that

$$\frac{\lambda_\chi}{\lambda_*}+\frac{A}{\lambda_*}\left(1-\frac{n}{\lambda_*\varrho}\right)^{-1}<1. \tag{29}$$

Then the theorem implies the existence a small neighbourhood of the fixed point g of Ξ on which the Ξ-iteration converges to g. Whenever this condition is met, the Ξ-iteration actually tends to converge on a rather large neighbourhood that includes our *ansatz*; see the next subsection for a more detailed discussion. Intuitively speaking, the condition requires the eigenvalues λ_i to be sufficiently separated with respect to the norm of the forcing term Φ. Even when the condition does not hold, the Ξ-iteration usually still displays an initial convergence for our *ansatz*, but the quality of the approximate solution to (26) ceases to improve after a while.

If the condition (29) is satisfied for all critical indices χ that we encounter when integrating from 0 until time T, then Algorithm 2 should produce an accurate result. The idealized analysis from section 4.2 then also applies, so the algorithm takes $O(2^{p/n}\log(\lambda_d T))$ steps. Since each step now requires $\tilde{O}(s\,n^2)$ floating point operations at bit precision p, we finally obtain the bound $\tilde{O}(s\,n^2 p\,2^{p/n}\log(\lambda_d T))$ for the total running time.

5.3. Quality of the computed solution

Let us now investigate in more detail why fixed points $f_{;n}$ of the Ξ-iteration indeed approximate the true solution $f^*_{;n}$ quite well. For this, we will determine a more precise small neighbourhood of the fixed point g on which the Ξ-iteration converges and show that this neighbourhood in particular contains $f^*_{;n}$. We start with two lemmas.

LEMMA 6. *Let $M(z)=\sum_{k\in\mathbb{N}} M_k z^k$ be a $d\times d$ matrix of analytic functions defined on the closed ball $\mathcal{B}(0,\varrho)$ with center zero and radius $\varrho>0$. Let $A=\sup_{|z|\leqslant\varrho}|M(z)|_\infty$. Then $|M_k|_\infty\leqslant A/\varrho^k$ for all $k\in\mathbb{N}$.*

Proof. Given $v\in\mathbb{C}^d$ with $|v|_\infty=1$, we have

$$|M_k v|_\infty=\left|\frac{1}{2\pi\mathrm{i}}\oint_{|z|=\varrho}\frac{M(z)\,v}{z^{k+1}}\mathrm{d}z\right|_\infty\leqslant\frac{1}{\varrho^k}\sup_{|z|=\varrho}|M(z)\,v|_\infty\leqslant\frac{A}{\varrho^k}.$$

It follows that $|M_k|_\infty=\sup_{|v|_\infty=1}|M_k v|_\infty\leqslant A/\varrho^k$. □

LEMMA 7. *Let $\varrho>0$, $\varepsilon>0$, $C<1$, $g\in\mathbb{C}[z]^d_{;n}$, and*

$$A=\sup_{|u|\leqslant\varrho,|\eta|\leqslant\frac{\varepsilon}{1-C}\mathrm{e}^{\varrho\lambda_*}}|J_\Phi(g(u)+\eta)|_\infty.$$

Then for all $\tilde{g}\in\mathbb{C}[z]^d_{;n}$ with $\|\tilde{g}-g\|_*\leqslant\frac{\varepsilon}{1-C}$ and $k\in\mathbb{N}$, we have

$$|J_\Phi(\tilde{g})_k|_\infty\leqslant\frac{A}{\varrho^k}.$$

Proof. Setting $\Delta g=\tilde{g}-g$, we first notice that

$$\sup_{|u|\leqslant\varrho}|\Delta g(u)|_\infty\leqslant\sum_{k\in\mathbb{N}}|\Delta g_k|_\infty\varrho^k\leqslant\sum_{k\in\mathbb{N}}\frac{\|\Delta g\|_*}{\mu_k}\varrho^k=\sum_{k\in\mathbb{N}}\frac{(\varrho\lambda_*)^k}{k!}\|\Delta g\|_*\leqslant\frac{\varepsilon}{1-C}\mathrm{e}^{\varrho\lambda_*}\|\Delta g\|_*.$$

It follows that

$$\sup_{|u|\leqslant\varrho} |J_\Phi(\tilde{g}(u))|_\infty \leqslant A$$

and we conclude using the previous lemma. □

THEOREM 8. *Let $\varrho > n/\lambda_*$ and $A > 0$ be such that*

$$C = \frac{\lambda_\chi}{\lambda_*} + \frac{A}{\lambda_*}\left(1 - \frac{n}{\lambda_*\varrho}\right)^{-1} < 1.$$

Let $g \in \mathbb{C}[z]^d_{;n}$ and $\varepsilon = \|\Xi(g) - g\|_*$ be such that

$$|u| \leqslant \varrho, |\eta| \leqslant \frac{\varepsilon}{1-C} e^{\varrho\lambda_*} \implies |J_\Phi(g(u) + \eta)|_\infty \leqslant A.$$

Then the sequence $g, \Xi(g), \Xi^2(g), \ldots$ tends to a unique fixed point g on the set

$$\mathcal{B}_*\left(g, \frac{\varepsilon}{1-C}\right) = \left\{h \in \mathbb{C}[z]^d_{;n} : \|h - g\|_* \leqslant \frac{\varepsilon}{1-C}\right\}.$$

Proof. A straightforward adaptation of the proof of Theorem 5 shows that $\|J_\Xi(g)\|_* \leqslant C$ on the ball $\mathcal{B}_*(g, \frac{\varepsilon}{1-C})$, which means that $\|\Xi(h_1) - \Xi(h_2)\|_* \leqslant C \|h_1 - h_2\|_*$ on this ball. By induction on $i \in \mathbb{N}$, it follows that $\|\Xi^{i+1}(g) - \Xi^i(g)\|_* \leqslant C^i \varepsilon$ and $\|\Xi^{i+1}(g) - g\|_* \leqslant \varepsilon(1 + C + \cdots + C^i)$. We conclude that g, $\Xi(g)$, $\Xi^2(g)$, ... converges to a fixed point $g + (\Xi(g) - g) + (\Xi(g^2) - \Xi(g)) + \cdots$ of Ξ in $\mathcal{B}_*(g, \frac{\varepsilon}{1-C})$. □

Returning to the Taylor series expansion $f^*_{;n}$ of the exact solution of (1) at time t, we notice that

$$\Xi^{[i]}(f^*_{;n}) = \begin{cases} (f^*_{;n})^{[i]} & \text{if } i \leqslant \chi \\ (f^*_{;n})^{[i]} + \frac{n}{\lambda_i}(f^*_n)^{[i]} z^{n-1} & \text{if } i > \chi \end{cases}.$$

It follows that

$$\|\Xi(f^*_{;n}) - f^*_{;n}\|_* \leqslant \frac{n}{\lambda_*}\mu_{n-1}\frac{K}{t^n} = K\frac{n!}{(\lambda_* t)^n} \approx K\left(\frac{n}{e\lambda_*\delta}\right)^n \left(\frac{\delta}{t}\right)^n \leqslant K\left(\frac{\delta}{t}\right)^n.$$

Now assuming that the aimed step size $\delta \approx t\ 2^{-p/n}$ was more or less achieved at the previous step, it follows that $\varepsilon := \|\Xi(f^*_{;n}) - f^*_{;n}\|_* \lesssim K\ 2^{-p}$ is of the desired order. If the condition (29) is indeed satisfied for $\varrho = t$, we thus should be able to apply Theorem 8 and conclude that the computed fixed point Ξ is within distance $\varepsilon/(1-C)$ for the $\|\cdot\|_*$ norm. Using a similar reasoning, we also see that the *ansatz* at the next step will be sufficiently close to the true solution for the Ξ-iteration to converge.

5.4. Solving steady-state problems through the first variation

A more robust but costly approach to solve the steady-state problem (26) is to compute the initial values $f_0^{[\chi+1]}, \ldots, f_0^{[d]}$ with sufficient precision using Newton's method. Given a tentative approximation $\varsigma = f_0$ for the initial condition, we both compute $f_{;n+1} = f\langle t \mid \varsigma\rangle_{;n+1}$ and the first variation $V_{;n+1} = V\langle t \mid \varsigma\rangle_{;n+1}$, after which we update $\varsigma := \varsigma + \Delta\varsigma$ by solving the linear system

$$f_n + V_n \Delta\varsigma = 0, \qquad \Delta\varsigma^{[1]} = \cdots = \Delta\varsigma^{[\chi]} = 0.$$

This method admits quadratic convergence, but it requires us to compute with a precision of $n \log_2 (\mathrm{e} \lambda_d / n)$ bits at least in order to be accurate. Indeed, this comes from the fact that $|V_n|_\infty$ grows roughly as $\lambda_d^n / n!$. On the upside, we may compute $f_{;n+1}$ and $V_{;n+1}$ using any of the algorithms from section 3. The total running time is therefore bounded by $\tilde{O}(d s\, n\, (p + n \log \lambda_d)\, 2^{p/n} \log(\lambda_d T))$. Notice also that it actually suffices to compute the last $d - \chi$ rows of V_n, due to the requirement that $\Delta \varsigma^{[1]} = \cdots = \Delta \varsigma^{[\chi]} = 0$.

The main disadvantage of the above method is that the computation of the first variation $V_{;n+1}$ along with $f_{;n+1}$ induces an overhead of d, which may be a problem for systems of high dimension. Let us now sketch how one might reduce this overhead by combining the variational approach with the Ξ-iteration. We intend to return to a more detailed analysis in a future work.

Instead of using a single critical index χ, the first idea is to use a range $\underline{\chi}, \ldots, \bar{\chi}$ of indices starting at $\underline{\chi} = \chi$, and such that the condition (29) holds for

$$\lambda_* = \sqrt{\lambda_{\underline{\chi}} \lambda_{\bar{\chi}+1}}.$$

The second idea is to only vary the components $\varsigma^{[\underline{\chi}+1]}, \ldots, \varsigma^{[\bar{\chi}]}$ of the initial condition and use the steady-state conditions for the indices $\bar{\chi}+1, \ldots, d$.

More specifically, for a tentative initial condition $f_0 = \varsigma$, we first solve the steady-state problem

$$\frac{\partial f}{\partial z} + \Lambda f = \Phi(f), \qquad \begin{cases} f_0^{[i]} = \varsigma^{[i]} & \text{if } i \leqslant \bar{\chi} \\ f_n^{[i]} = 0 & \text{if } i > \bar{\chi} \end{cases}$$

using the Ξ-iteration technique. In a similar way, we next solve the following variational steady-state problem for the $d \times (\bar{\chi} - \underline{\chi})$ matrix W:

$$\frac{\partial W}{\partial z} + \Lambda W = J_\Phi(f)\, W, \qquad \begin{cases} W_0^{[i,j]} = \mathrm{Id}^{[i, j+\underline{\chi}]} & \text{if } i \leqslant \underline{\chi} \\ W_n^{[i,j]} = 0 & \text{if } i > \bar{\chi} \end{cases}.$$

As our *ansatz*, we may use $W^{[i,j]} = \mathrm{Id}^{[i, j+\underline{\chi}]}$ for all i, j. Having computed $f_{;n+1}$ and $W_{;n+1}$ at precision $n+1$, we finally update $\varsigma := \varsigma + \mathrm{d}\varsigma$ by solving the linear system

$$f_n^{[i]} + W_n^{[i,1]} (\Delta \varsigma)^{[\underline{\chi}+1]} + \cdots + W_n^{[i, \bar{\chi} - \underline{\chi}]} (\Delta \varsigma)^{[\bar{\chi}]} \;=\; 0, \qquad i = \underline{\chi} + 1, \ldots, \bar{\chi}$$

and setting $\Delta \varsigma^{[1]} = \cdots = \Delta \varsigma^{[\underline{\chi}]} = \Delta \varsigma^{[\bar{\chi}+1]} = \cdots = \Delta \varsigma^{[d]} = 0$. We repeat this whole process until $f_n^{[i]}$ is sufficiently close to zero for $i = \underline{\chi} + 1, \ldots, \bar{\chi}$.

Remark 9. The algorithms in this section are reminiscent of implicit numerical schemes for the integration of (1). One interesting difference is that our second optimized method only needs to compute a small part of the full Jacobian matrix.

Remark 10. Instead of imposing the exact steady-state conditions $f_n^{[\chi+1]} = \cdots = f_n^{[d]} = 0$, yet another approach would be to minimize the norm $|f_n|_\infty$ under the condition that $|f_0 - \varsigma|_\infty \leqslant |\varsigma|_\infty\, 2^{-p}$. This approach admits the advantages that one does not need to know χ and that it might be applied to more general complex matrices Λ. However, it requires the computation of the full Jacobian matrix.

6. CERTIFICATION

In the previous two sections, we have described numerical algorithms for integrating (1). An interesting question is how to compute a tight error bound such that the distance between the true and the computed solutions lies within this error bound. Ball arithmetic provides a suitable framework for such computations. This variant of interval arithmetic is well suited for high precision computations with complex numbers and we will briefly recall its basic principles in section 6.1. As a first application, we show how to make Theorem 4 more effective.

The certified integration of differential equations that are not stiff (i.e. the robust counterpart of section 4) is a classical topic in interval arithmetic [20, 21, 15, 5, 24, 16, 7, 22, 18, 14, 17, 23, 19]. For recent algorithms of good complexity, much in the spirit of the present paper, we refer to [11]. Most of the existing algorithms rely on Taylor series expansions as we do, while providing rigorous tail bounds for the truncation error.

The effective counterpart of Theorem 4 provides suitable tail bounds in the case of stiff differential equations. In sections 6.3 and 6.4, we show how to use this for the certified resolution of steady-state problems using the Ξ-iteration from section 5.1. Unfortunately, our first algorithms are rather naive and typically lead to heavily overestimated error bounds. The classical way to reduce this overestimation is to also compute the first variation and apply the mean value theorem: see section 6.5 for how to do this in our context.

6.1. Ball arithmetic

Given $a \in \mathbb{C}$ and $r \in \mathbb{R}^{\geqslant} := \{x \in \mathbb{R} : x \geqslant 0\}$, we write $B(a,r) = \{z \in \mathbb{C} : |z-a| \leqslant r\}$ for the *closed ball* with center a and radius r. The set of such balls is denoted by $\mathcal{B}(\mathbb{C},\mathbb{R})$ or $\mathbb{C}^{\bullet}$. One may lift the ring operations $+,-,\times$ in $\mathbb{C}$ to balls in $\mathbb{C}^{\bullet}$, by setting:

$$\begin{aligned} \mathcal{B}(a,r) \pm \mathcal{B}(b,s) &:= \mathcal{B}(a \pm b, r+s) \\ \mathcal{B}(a,r) \times \mathcal{B}(b,s) &:= \mathcal{B}(a\,b, (|a|+r)\,s + |b|\,r). \end{aligned}$$

These formulas are simplest so as to satisfy the so called *inclusion principle*: given $* \in \{+,-,\times\}$, $u \in \mathcal{B}(a,r)$ and $v \in \mathcal{B}(b,s)$, we have $u * v \in \mathcal{B}(a,r) * \mathcal{B}(b,s)$. This arithmetic for computing with balls is called *exact ball arithmetic*. It extends to other operations that might be defined on $\mathbb{C}$, as long as the ball lifts of operations satisfy the inclusion principle. Any ordinary complex number $z \in \mathbb{C}$ can be reinterpreted as a ball $\mathcal{B}(z,0) \in \mathbb{C}^{\bullet}$. Given a ball $\mathcal{B}(a,r) \in \mathbb{C}^{\bullet}$, we also notice that $\lceil \mathcal{B}(a,r) \rceil = |a| + r$ and $\lfloor \mathcal{B}(a,r) \rfloor = \max(|a|-r, 0)$ provide us with reliable upper and lower bounds for $|\mathcal{B}(a,r)|$ in $\mathbb{R}^{\geqslant}$.

Another interesting operation on balls $\mathcal{B}(a,r), \mathcal{B}(b,s) \in \mathbb{C}^{\bullet}$ that we will need below is intersection. Assuming that the set intersection $I := \mathcal{B}(a,r) \cap \mathcal{B}(b,s)$ is non-empty, we define the ball intersection $J = \mathcal{B}(a,r) \sqcap \mathcal{B}(b,s) \in \mathbb{C}^{\bullet}$ to be the ball of smallest radius that contains I. In order to determine this ball intersection, we may assume without loss of generality that $r \geqslant s$. If $\mathcal{B}(a,r) \supseteq \mathcal{B}(b,s)$, then $J = \mathcal{B}(b,r)$. Otherwise, let $u, v \in \mathbb{C}$ be the two (possibly identical) intersections of the circles $\partial\mathcal{B}(a,r)$ and $\partial\mathcal{B}(b,s)$. If $|a - (u+v)/2| \geqslant |a-b|$, then we still have $J = \mathcal{B}(b,r)$. Otherwise, $J = \mathcal{B}((u+v)/2, |v-u|/2)$.

It will also be convenient to extend vector notations to balls. First of all, we identify vectors of balls $(\mathcal{B}(a^{[1]}, r^{[1]}), \ldots, \mathcal{B}(a^{[d]}, r^{[d]})) \in \mathcal{B}(\mathbb{C},\mathbb{R})^d$ with "ball vectors" $\mathcal{B}(a,r) \in \mathcal{B}(\mathbb{C}^d, \mathbb{R}^d)$. Given $z \in \mathbb{C}^d$ and $v^{\bullet} \in (\mathbb{C}^{\bullet})^d$, we also write $z \in v^{\bullet}$ if and only if $z^{[i]} \in (v^{\bullet})^{[i]}$ for $i = 1, \ldots, d$. Similar remarks apply to ball matrices, ball series, etc.

Remark 11. For effective computations, one can only work with approximations of real and complex numbers of finite precision. The IEEE 754 norm provides a standard for real floating point arithmetic based on the concept of "correct rounding". It naturally generalizes to floating point numbers with a mantissa of p bits and an exponent of e bits [3]. Let $\mathbb{R}_{p,e}$ and $\mathbb{C}_{p,e}$ the sets of real and complex floating point numbers of this type (strictly speaking, one also has $\{-\infty,\infty\}\in\mathbb{R}_{p,e}$ for the reliable rounding of overflowing operations). One may adapt the above arithmetic on exact balls to floating point balls in $\mathcal{B}(\mathbb{C}_{p,e},\mathbb{R}_{p,e})$ while preserving the inclusion principle: it suffices to slightly increase the radii of output balls so as to take care of rounding errors. For precise formulas and interesting variations, we refer to [13]. For simplicity, the sequel will be presented for exact ball arithmetic only, but it is not hard to adapt the results to take into account rounding errors.

6.2. Computing bounds on compact half disks

The polynomials $\Phi^{[i]}$ in (1) can either be represented as linear combinations of monomials, or using a dag (directed acyclic graph). In both cases, the ball arithmetic from the previous subsection allows us to reliably evaluate Φ at balls in $\mathbb{C}^\bullet$.

For the effective counterpart of Theorem 4, there is a trade-off between the qualities of the radius R (larger radii being better) and the bound $B\in(\mathbb{R}^>)^d$ (smaller values of B being better). For a given $B\in(\mathbb{R}^>)^d$, it is simple to compute the "best" corresponding R that satisfies the condition (8), by taking $R=\min_i \lfloor B/\Phi(\mathcal{B}(z,B))\rfloor^{[i]}$. Conversely, for a fixed $R\in\mathbb{R}^>$, let us recall a classical technique from interval analysis that can be used to compute an "almost best" corresponding bound $B\in(\mathbb{R}^>)^d$ that satisfies (8).

We simply construct a sequence $B_0\leqslant B_1\leqslant B_2\leqslant\cdots$ in $(\mathbb{R}^\geqslant)^d$ by taking $B_0=0$ and $B_{k+1}=\lceil\Phi(\mathcal{B}(c,B_k))\,R\rceil$ for all k. If there exists a finite B for which (8) holds, then the sequence B_k usually converges to a minimal such B quite quickly. After say ten iterations, we should therefore have a reasonable approximation. One then slightly inflates the last value by taking $B=B_{10}+(B_{10}-B_9)$. If $\lceil\Phi(\mathcal{B}(c,B))\,R\rceil\leqslant B$, then we have succeeded in finding a suitable B. If not, then we return "failed".

Using the above procedure, we may also compute a reasonably large compact half disk H_R on which φ is analytic, together with a bound for $|\varphi|$: we simply perform a dichotomic search for the largest R for which the computation of a corresponding B does not fail. If we also wish the bound B to remain reasonably small, one may simply divide R by two at the end of the search and compute the corresponding B.

6.3. Certified integration of stiff differential equations

Let us now return to the integration of (1) and let φ^* be the exact solution for the initial condition $\varphi^*(0)=c\in\mathbb{C}^d$. Let $R\in\mathbb{R}^>$ and $B\in(\mathbb{R}^>)^d$ be computed as above such that (8) holds. Assume that we were able to reliably integrate (1) until a given time $t\leqslant R/2$ and let $\varsigma^\bullet=\varphi^\bullet(t)\in(\mathbb{C}^\bullet)^d$, so that $\varphi^*(t)\in\varsigma^\bullet$.

In order to adapt the Ξ-iteration to ball arithmetic, we first deduce from Theorem 4 that $|f^*\langle t\rangle_n^{[i]}|\leqslant B^{[i]}/t^n$ for $i=\chi+1,\ldots,d$. This provides us with the required steady-state conditions $f_n^{[\chi+1]}=\mathcal{B}(0,B^{[\chi+1]}/t^n)$, ..., $f_n^{[d]}=\mathcal{B}(0,B^{[d]}/t^n)$, in addition to the initial conditions $f_0^{[1]}=\varsigma^{[1]},\ldots,f_0^{[\chi]}=\varsigma^{[\chi]}$. The ball counterpart of our steady-state problem thus becomes

$$\frac{\partial f^\bullet}{\partial z}+\Lambda f^\bullet=\Phi(f^\bullet),\qquad\begin{cases}(f_0^\bullet)^{[i]}=(\varsigma^\bullet)^{[i]} & \text{if } i\leqslant\chi\\ (f_n^\bullet)^{[i]}=\mathcal{B}(0,B^{[i]}/t^n) & \text{if } i>\chi\end{cases}\tag{30}$$

We correspondingly define the ball version of Ξ for $g \in \mathbb{C}^\bullet[z]^d_{;n}$ by

$$\Xi^{[i]}(g^\bullet) \;=\; \begin{cases} (\varsigma^\bullet)^{[i]} + \int (\Phi^{[i]}(g^\bullet) - \lambda_i (g^\bullet)^{[i]}) & \text{if } i \leqslant \chi \\ \lambda_i^{-1} (\Phi^{[i]}(g^\bullet) - \partial\, (g^\bullet)^{[i]} - \mathcal{B}(0, n B^{[d]}/t^n)\, z^{n-1}) & \text{if } i > \chi \end{cases} \mod z^n.$$

This map has the property that $f^*_{;n} \in g^\bullet \Rightarrow f^*_{;n} \in \Xi(g^\bullet)$. In our ball context, we may actually iterate using the following improved version $\Xi_\sqcap$ of Ξ:

$$\Xi_\sqcap(g^\bullet) \;=\; g^\bullet \sqcap \Xi(g^\bullet).$$

Here we understand that the intersections are taken coefficientwise. Since iterations using $\Xi_\sqcap$ can only improve our enclosures, we do not need to worry much about the *ansatz*. We may deduce a reasonably good ansatz from the power series expansion $f^\bullet\langle t'\rangle_{;n}$ at the previous time $t' < t$, and the Cauchy bounds $|f^\bullet\langle t'\rangle^{[i]}_k| \leqslant B^{[i]} / (t')^k$ for all $k \geqslant n$. But it is perfectly reasonable as well to use

$$f^{\text{an},\bullet}_{;n} = \varsigma^\bullet + \mathcal{B}(0, B/t)\, z + \cdots + \mathcal{B}(0, B/t^{n-1})\, z^{n-1}.$$

Applying $\Xi_\sqcap$ a sufficient number of times to this ansatz, we obtain the desired ball enclosure $f^\bullet_{;n} = \Xi_\sqcap^{2n}(f^{\text{an},\bullet}_{;n})$ of the truncated power series expansion of φ^* at time t. Assuming that $\delta < t$, we may then deduce the enclosure

$$\varphi^\bullet(t + \delta) = f^\bullet_{;n}(\delta) + \mathcal{B}\left(0, B\, \frac{t}{t - \delta} \left(\frac{\delta}{t}\right)^n\right) \tag{31}$$

of φ^* at time $t + \delta$. Notice that the cost of the computation of certified solutions in this way is similar to the cost of Algorithm 1, up to a constant factor.

6.4. Certification of approximate solutions

The above method relies on the property that the steady-state problem (30) for balls specializes into a numerical steady-state problem that admits f^* as a solution, since $f^*_0 \in \varsigma^\bullet$ and $f^*_n \in \mathcal{B}(0, B/t^n)$. Given a numerical approximation of $f^*_{;n}$, as computed in section 5, a related problem is whether we can certify the existence of an actual solution in a small neighbourhood of the approximate solution.

In order to make this more precise, let us introduce a few more notations. Given $\varsigma^{[1]}, \ldots, \varsigma^{[\chi]} \in \mathbb{C}$ and $\tau^{[\chi+1]}, \ldots, \tau^{[d]} \in \mathbb{C}$, consider the numeric steady-state problem

$$\frac{\partial f}{\partial z} + \Lambda f = \Phi(f), \qquad \begin{cases} f_0^{[i]} = \varsigma^{[i]} & \text{if } i \leqslant \chi \\ f_n^{[i]} = \tau^{[i]} & \text{if } i > \chi \end{cases}. \tag{32}$$

We will denote by $f^*\langle t \,|\, \varsigma, \tau\rangle$ any exact solution to this problem if such solution exists. It will be convenient to regard ς and τ as elements of $\mathbb{C}^d$, through padding with zeros. Now consider the operator $\Xi_{\varsigma,\tau} \colon \mathbb{C}[z]^d_{;n} \to \mathbb{C}[z]^d_{;n}$ defined by

$$\Xi^{[i]}_{\varsigma,\tau}(g) \;=\; \begin{cases} \varsigma^{[i]} + \int (\Phi^{[i]}(g) - \lambda_i g^{[i]}) & \text{if } i \leqslant \chi \\ \lambda_i^{-1} (\Phi^{[i]}(g) - \partial g^{[i]} - n\, \tau_i z^{n-1}) & \text{if } i > \chi \end{cases} \mod z^n,$$

Through the iterated application of $\Xi_{\varsigma,\tau}$ to a suitable *ansatz*, we may obtain a numerical approximation $f^{\approx}\langle t \,|\, \varsigma, \tau\rangle_{;n}$ of $f^*\langle t \,|\, \varsigma, \tau\rangle_{;n}$.

Now consider balls $(\varsigma^\bullet)^{[1]}, \ldots, (\varsigma^\bullet)^{[\chi]} \in \mathbb{C}^\bullet$ and $(\tau^\bullet)^{[\chi+1]}, \ldots, (\tau^\bullet)^{[d]} \in \mathbb{C}^\bullet$, again padded with zeros. Then we have the following ball analogue of (32):

$$\frac{\partial f^\bullet}{\partial z} + \Lambda f^\bullet = \Phi(f^\bullet), \qquad \begin{cases} (f_0^\bullet)^{[i]} = (\varsigma^\bullet)^{[i]} & \text{if } i \leqslant \chi \\ (f_n^\bullet)^{[i]} = (\tau^\bullet)^{[i]} & \text{if } i > \chi \end{cases}. \tag{33}$$

Let ς^{cen} and τ^{cen} denote the centers of $\varsigma^\bullet$ and $\tau^\bullet$. Starting from a numerical approximation $f^{\approx}\langle t\,|\,\varsigma^{\mathrm{cen}},\tau^{\mathrm{cen}}\rangle_{;n}$ of $f^*\langle t\,|\,\varsigma^{\mathrm{cen}},\tau^{\mathrm{cen}}\rangle_{;n}$, we wish to compute the truncation $f^\bullet_{;n}\in\mathbb{C}^\bullet[z]^d_{;n}$ of a solution to (33), with the property that $f^*\langle t\,|\,\varsigma,\tau\rangle_{;n}\in f^\bullet_{;n}$ for all $\varsigma\in\varsigma^\bullet$ and $\tau\in\tau^\bullet$.

In order to do this, the idea is again to use a suitable ball version of $\Xi_{\varsigma,\tau}$, given by

$$\Xi^{[i]}_{\varsigma^\bullet,\tau^\bullet}(g^\bullet) \;=\; \begin{cases} (\varsigma^\bullet)^{[i]} + \int (\Phi^{[i]}(g^\bullet) - \lambda_i\,(g^\bullet)^{[i]}) & \text{if } i\leqslant\chi \\ \lambda_i^{-1}\,(\Phi^{[i]}(g^\bullet) - \partial\,(g^\bullet)^{[i]} - n\,(\tau^\bullet)^{[i]}\,z^{n-1}) & \text{if } i>\chi \end{cases} \mod z^n,$$

and to keep applying $\Xi_{\varsigma^\bullet,\tau^\bullet}$ on the *ansatz* $f^{\approx}\langle t\,|\,\varsigma^{\mathrm{cen}},\tau^{\mathrm{cen}}\rangle_{;n}$ until we are sufficiently close to a fixed point. Using a similar inflation technique as in section 6.2, we finally obtain a truncated series $f^\bullet_{;n}$ with $\Xi(f^\bullet_{;n})\subseteq f^\bullet_{;n}$, or "failed". Now for any $\varsigma\in\varsigma^\bullet$ and $\tau\in\tau^\bullet$, we notice that this also yields $\Xi_{\varsigma,\tau}(f^\bullet_{;n})\subseteq f^\bullet_{;n}$. Thinking of $f^\bullet_{;n}$ as a compact set of series in $\mathbb{C}[z]^d_{;n}$, this means in particular that $\Xi_{\varsigma,\tau}$ admits a fixed point $g = f^*\langle t\,|\,\varsigma,\tau\rangle_{;n}\in f^\bullet_{;n}$, as desired.

6.5. Curbing the wrapping effect

If $\chi = d$, then the algorithm from subsection 6.3 specializes to a classical way to compute the enclosure (31) of a solution to (1) at time $t+\delta$ as a function of the enclosure $\varsigma^\bullet = \mathcal{B}(\varsigma^{\mathrm{cen}},\varsigma^{\mathrm{rad}}) = f_0$ at time t. However, it is well known that this method suffers from a large overestimation of the errors, due to the so-called "wrapping effect". A well know technique to reduce this overestimation is to first compute a certified solution for the initial condition $\mathcal{B}(\varsigma^{\mathrm{cen}},0)$ instead of $\varsigma^\bullet$ and then to bound the error due to this replacement, by investigating the first variation. Let us now show how to extend this technique to general critical indices χ.

With ς^{cen} in the role of $\varsigma^\bullet$ and $\mathcal{B}(0,B/t^n)$ in the role of $\tau^\bullet$, we start with the computation of a truncated certified solution $f^{\mathrm{acc},\bullet}_{;n}\in\mathbb{C}^\bullet[z]^d_{;n}$ to (33). Recall that this ball solution has the property that $f^*\langle t\,|\,\varsigma^{\mathrm{cen}},\tau\rangle_{;n}\in f^{\mathrm{acc},\bullet}_{;n}$ for all $\tau\in\tau^\bullet$. Since the initial conditions are exact, this solution is generally fairly accurate. Now for any $\varsigma'\in\varsigma^\bullet$, we may write

$$f^*\langle t\,|\,\varsigma^{\mathrm{cen}},\tau\rangle \;=\; f^*\langle t\,|\,\varsigma^{\mathrm{cen}},\tau\rangle + \int_0^1 \frac{\partial}{\partial\epsilon}\, f^*\langle t\,|\,\varsigma^{\mathrm{cen}} + (\varsigma'-\varsigma^{\mathrm{cen}})\,\epsilon,\tau\rangle\,\mathrm{d}\epsilon. \tag{34}$$

We next observe that

$$\frac{\partial}{\partial\epsilon}\, f^*\langle t\,|\,\varsigma^{\mathrm{cen}} + (\varsigma'-\varsigma^{\mathrm{cen}})\,\epsilon,\tau\rangle \;=\; V^*\langle t\,|\,\varsigma^{\mathrm{cen}} + (\varsigma'-\varsigma^{\mathrm{cen}})\,\epsilon,\tau\rangle\,(\varsigma'-\varsigma^{\mathrm{cen}}),$$

where $V^*\langle t\,|\,\varsigma,\tau\rangle$ denotes the exact solution of the following equation for the *first steady-state variation*:

$$\frac{\partial V}{\partial z} + \Lambda\, V = J_\Phi(f\langle t\,|\,\varsigma,\tau\rangle)\,V, \qquad \begin{cases} V_0^{[i,j]} = \mathrm{Id}_d^{[i,j]} & \text{if } i\leqslant\chi \\ V_n^{[i,j]} = 0 & \text{if } i>\chi \end{cases}.$$

This equation again admits a ball analogue

$$\frac{\partial V^\bullet}{\partial z} + \Lambda\, V^\bullet = J_\Phi(f^\bullet)\,V^\bullet, \qquad \begin{cases} (V_0^\bullet)^{[i,j]} = \mathrm{Id}_d^{[i,j]} & \text{if } i\leqslant\chi \\ (V_n^\bullet)^{[i,j]} = 0 & \text{if } i>\chi \end{cases},$$

where $f^\bullet$ is a solution to (30). We only compute a crude solution $V^{\mathrm{cru},\bullet}$ to this equation, for a crude solution $f^{\mathrm{cru},\bullet}$ to (30), both truncated to the order n. These solutions can be obtained using the techniques from section 6.3. From (34) we finally conclude that

$$f^\bullet_{;n} \;=\; f^{\mathrm{acc},\bullet}_{;n} + V^{\mathrm{cru},\bullet}_{;n}\,\mathcal{B}(0,\varsigma^{\mathrm{rad}})$$

is a ball enclosure for $f^*_{;n}$. This enclosure is usually of a much better quality than $f^{\mathrm{cru},\bullet}$. However, its computational cost is roughly d times higher, due to the fact that we need to determine the first variation.

6.6. Handling close eigenvalues

In the case when $\lambda_{\chi+1}/\lambda_\chi$ is too small for (29) to hold, we may generalize the above strategy and incorporate some of the ideas from section 5.4. With $\chi = \underline{\chi}$ and $\bar{\chi}$ from there, let us briefly outline how to do this. We proceed in four stages:

- We first compute a truncated numeric solution $f^{\approx}_{;n} \in \mathbb{C}[z]^d_{;n}$ to (26), using the technique from section 5.4.
- Let $\varsigma^{\mathrm{acc},\bullet} \in (\mathbb{C}^\bullet)^d$ be such that $(\varsigma^{\mathrm{acc},\bullet})^{[i]} = (f_0^{\approx})^{[i]}$ for $i \leqslant \bar{\chi}$ and $(\varsigma^{\mathrm{acc},\bullet})^{[i]} = 0$ for $i > \bar{\chi}$. Similarly, let $\tau^{\mathrm{acc},\bullet} \in (\mathbb{C}^\bullet)^d$ be such that $(\tau^{\mathrm{acc},\bullet})^{[i]} = 0$ for $i \leqslant \bar{\chi}$ and $(\tau^{\mathrm{acc},\bullet})^{[i]} = \mathcal{B}(0, B^{[i]}/t^n)$ for $i > \bar{\chi}$. We use $f^{\approx}_{;n}$ as an *ansatz* for its deformation into a reliable truncated solution $f^{\mathrm{acc},\bullet}_{;n} \in \mathbb{C}^\bullet[z]^d_{;n}$ to (33), but with $\bar{\chi}$ in the role of χ, with $\varsigma^{\mathrm{acc},\bullet}$ in the role of $\varsigma^\bullet$, and $\tau^{\mathrm{acc},\bullet}$ in the role of $\tau^\bullet$.
- We further deform $f^{\mathrm{acc},\bullet}_{;n}$ into a reliable truncated solution $f^{\mathrm{aux},\bullet}_{;n} \in \mathbb{C}^\bullet[z]^d_{;n}$ to (33), this time with $\underline{\chi}$ in the role of χ. We do this by computing $f^{\mathrm{cru},\bullet}_{;n}$ as above, together with a crude but reliable solution to the equation

$$\frac{\partial W^\bullet}{\partial z} + \Lambda W^\bullet = J_\Phi(f^{\mathrm{cru},\bullet})\, W^\bullet, \qquad \begin{cases} (W_0^\bullet)^{[i,j]} = 0 & \text{if } i \leqslant \underline{\chi} \\ (W_n^\bullet)^{[i,j]} = \mathrm{Id}_d^{[i,j]} & \text{if } \underline{\chi} < i \leqslant \bar{\chi} \\ (W_n^\bullet)^{[i,j]} = 0 & \text{if } i > \bar{\chi} \end{cases}.$$

  We then take

$$f^{\mathrm{aux},\bullet}_{;n} = f^{\mathrm{acc},\bullet}_{;n} + W^\bullet_{;n}\, \mathcal{B}(0, \tau^{\mathrm{aux}}),$$

  where $(\tau^{\mathrm{aux}})^{[i]} = B^{[i]}/t^n$ for $\underline{\chi} < i \leqslant \bar{\chi}$ and $(\tau^{\mathrm{aux}})^{[i]} = 0$ for $i \leqslant \underline{\chi}$ and $i > \bar{\chi}$.
- We finally compute the desired truncated solution $f^\bullet_{;n} \in \mathbb{C}^\bullet[z]^d_{;n}$ to (30) as

$$f^\bullet_{;n} = f^{\mathrm{acc},\bullet}_{;n} + V^{\mathrm{cru},\bullet}_{;n}\, \mathcal{B}(0, \varsigma^{\mathrm{rad}}),$$

  where $V^{\mathrm{cru},\bullet}_{;n}$ is computed as above.

7. CONCLUSION

There are several directions for generalizations and further improvements of the results in this paper. For simplicity, we assumed Λ to be a diagonal matrix with positive eigenvalues. It should not be hard to adapt the results to arbitrary matrices Λ with positive real eigenvalues only (as long as the forcing term Φ does not explode for the change of coordinates that puts Λ in Jordan normal form).

A more interesting generalization would be to consider complex eigenvalues λ_i with strictly positive real parts. In that case, the half disks H_R would need to be replaced by smaller compact sectors of the form $S_{R,\theta} = \{z \in \mathbb{C} : |z| \leqslant R, |\arg z| \leqslant \theta\}$ with $\theta < \pi/2$. Even more generally, one may investigate how to develop accurate Taylor series methods for integrating arbitrary differential equations with the property that the step size is proportional to the radius of convergence of the exact solution; see also Remark 10.

In this paper, we focussed on high precision computations using high order Taylor schemes. Another interesting question is whether it is possible to develop analogues of our methods in the same spirit as more traditional Runge–Kutta schemes. How would such analogues compare to implicit integration schemes?

Concerning certified integration, the theory of Taylor models [18, 23, 19] allows for higher order expansions of the flow $f\langle t \,|\, \varsigma \rangle$ as a function of the initial condition ς (that is, beyond the computation of the first variation as in this paper). We think that it should be relatively straightforward to adapt our techniques to such higher order expansions. In a similar vein, we expect that the techniques from [7, 11] to further curb the wrapping effect can be adapted to the stiff setting.

From a more practical point of view, it would finally be interesting to have more and better machine implementations. For the moment, we only did a toy implementation of the main algorithm from section 5.1 in the MATHEMAGIX system [12]. We found this implementation to work as predicted on various simple examples. Detailed machine experiments with larger systems and the algorithms from sections 5.4 and 6 should make it possible to further improve the new techniques.

BIBLIOGRAPHY

[1] A. Bostan, F. Chyzak, F. Ollivier, B. Salvy, É. Schost, and A. Sedoglavic. Fast computation of power series solutions of systems of differential equations. In *Proceedings of the 18th ACM-SIAM Symposium on Discrete Algorithms*, pages 1012–1021. New Orleans, Louisiana, U.S.A., January 2007.

[2] R. P. Brent and H. T. Kung. Fast algorithms for manipulating formal power series. *Journal of the ACM*, 25:581–595, 1978.

[3] R. P. Brent and P. Zimmermann. *Modern Computer Arithmetic*. Cambridge University Press, 2010.

[4] J. C. Butcher. Numerical methods for ordinary differential equations in the 20th century. *J. of Computational and Applied Mathematics*, 125:1–29, 2000.

[5] J.-P. Eckmann, H. Koch, and P. Wittwer. A computer-assisted proof of universality in area-preserving maps. *Memoirs of the AMS*, 47(289), 1984.

[6] M. J. Fischer and L. J. Stockmeyer. Fast on-line integer multiplication. *Proc. 5th ACM Symposium on Theory of Computing*, 9:67–72, 1974.

[7] T. N. Gambill and R. D. Skeel. Logarithmic reduction of the wrapping effect with application to ordinary differential equations. *SIAM J. Numer. Anal.*, 25(1):153–162, 1988.

[8] J. van der Hoeven. Relax, but don't be too lazy. *JSC*, 34:479–542, 2002.

[9] J. van der Hoeven. New algorithms for relaxed multiplication. *JSC*, 42(8):792–802, 2007.

[10] J. van der Hoeven. Newton's method and FFT trading. *J. Symbolic Comput.*, 45(8):857–878, 2010.

[11] J. van der Hoeven. Certifying trajectories of dynamical systems. In I. Kotsireas, S. Rump, and C. Yap, editors, *Mathematical Aspects of Computer and Information Sciences: 6th International Conference, MACIS 2015, Berlin, Germany, November 11-13, 2015, Revised Selected Papers*, pages 520–532. Cham, 2016. Springer International Publishing.

[12] J. van der Hoeven, G. Lecerf, B. Mourrain et al. Mathemagix. from 2002. http://www.mathemagix.org.

[13] Joris van der Hoeven and Grégoire Lecerf. Evaluating straight-line programs over balls. In Paolo Montuschi, Michael Schulte, Javier Hormigo, Stuart Oberman, and Nathalie Revol, editors, *2016 IEEE 23nd Symposium on Computer Arithmetic*, pages 142–149. IEEE, 2016.

[14] W. Kühn. Rigourously computed orbits of dynamical systems without the wrapping effect. *Computing*, 61:47–67, 1998.

[15] O. E. Lanfort. A computer assisted proof of the Feigenbaum conjectures. *Bull. of the AMS*, 6:427–434, 1982.

[16] R. Lohner. *Einschließung der Lösung gewöhnlicher Anfangs- und Randwertaufgaben und Anwendugen*. PhD thesis, Universität Karlsruhe, 1988.

[17] R. Lohner. On the ubiquity of the wrapping effect in the computation of error bounds. In U. Kulisch, R. Lohner, and A. Facius, editors, *Perspectives on enclosure methods*, pages 201–217. Wien, New York, 2001. Springer.

[18] K. Makino and M. Berz. Remainder differential algebras and their applications. In M. Berz, C. Bischof, G. Corliss, and A. Griewank, editors, *Computational differentiation: techniques, applications and tools*, pages 63–74. SIAM, Philadelphia, 1996.

[19] K. Makino and M. Berz. Suppression of the wrapping effect by Taylor model-based validated integrators. Technical Report MSU Report MSUHEP 40910, Michigan State University, 2004.

[20] R. E. Moore. Automatic local coordinate transformations to reduce the growth of error bounds in interval computation of solutions to ordinary differential equation. In L. B. Rall, editor, *Error in Digital Computation*, volume 2, pages 103–140. John Wiley, 1965.

[21] R. E. Moore. *Interval Analysis*. Prentice Hall, Englewood Cliffs, N.J., 1966.

[22] A. Neumaier. The wrapping effect, ellipsoid arithmetic, stability and confedence regions. *Computing Supplementum*, 9:175–190, 1993.

[23] A. Neumaier. Taylor forms - use and limits. *Reliable Computing*, 9:43–79, 2002.

[24] K. Nickel. How to fight the wrapping effect. In Springer-Verlag, editor, *Proc. of the Intern. Symp. on interval mathematics*, pages 121–132. 1985.

[25] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes, the art of scientific computing*. Cambridge University Press, 3rd edition, 2007.

[26] A. Sedoglavic. *Méthodes seminumériques en algèbre différentielle ; applications à l'étude des propriétés structurelles de systèmes différentiels algébriques en automatique*. PhD thesis, École polytechnique, 2001.