A Priori Estimates for the Global Error Committed by Runge–Kutta Methods for a Nonlinear Oscillator

Jitse Niesen*

22 August 2001

Abstract

The Alekseev–Gröbner lemma is combined with the theory of modified equations to obtain an *a priori* estimate for the global error of numerical integrators. This estimate is correct up to a remainder term of order h^{2p}, where h denotes the step size and p the order of the method. It is applied to a class of nonautonomous linear oscillatory equations, which includes the Airy equation, thereby improving prior work which only gave the h^p term. However, the result is not very surprising.

Next, a single nonlinear oscillatory equation is considered, namely $y'' + ty^3 = 0$, and global errors of Runge–Kutta methods are calculated. Numerical experiments show that the resulting estimates are generally accurate. The main conclusion is that we need to do a full calculation to obtain good estimates: the behaviour is different from the linear case, it is not sufficient to look only at the leading term, and merely considering the local error does not provide an accurate picture either.

1 Introduction

This paper is about the numerical solution of ordinary differential equations. Unfortunately, the outcome of the calculation is almost invariably not exact, but carries an error. The error committed by a single step of the numerical method is called the *local (truncation) error*. But in general, a single step is not sufficient to cover the time interval of interest, so we are interested in the error at a specific time, after several steps. This quantity is called the *global error*.

It is important to have at least a rough idea of the size of the global error, so we need to *estimate* it. We can distinguish two approaches. One is to solve the equation numerically, and to use the result somehow to find an *a posteriori* estimate. The other approach is to use the analytic solution, or at least some knowledge about it. This gives rise to *a priori* estimates.

* Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Silver Street, Cambridge CB3 9EW, United Kingdom, e-mail J.Niesen@damtp.cam.ac.uk.

Both types of estimates have their respective strengths. If one has actually computed a numerical solution and wants to know how far off it is, one probably should use an *a posteriori* estimate. However, for the purpose of comparing different methods, or that of devising methods which are particularly suited for a certain class of problems, there is often no choice but to use *a priori* estimates.

Skeel [24] wrote a nice overview of various methods to obtain *a posteriori* estimates of the global error, but here we will concentrate on *a priori* estimates. Early research on this topic was done by Henrici [15] and Gragg [9], who expanded the global error in powers of the step size. This expansion was extended and simplified by Hairer and Lubich [12], and used by Cano and Sanz-Serna [4, 5] to analyse the behaviour of numerical methods for systems with periodic orbits.

A parallel path of research is based on the Alekseev–Gröbner lemma, to be introduced in the next section. We merely point to the work of Iserles and Söderlind [18], Calvo and Hairer [3], and Viswanath [25].

The recent paper of Iserles [17] gives an *a priori* estimate for the global error committed for certain linear oscillators. This estimate only concerns the first term in the expansion of the global error in powers of the step size. The goal of the research described here, is to extend this work in two directions. Firstly, more (but not all) terms of the expansion are derived. Numerical experiments will show that it is sometimes necessary to include these terms. Secondly, the effects of nonlinearity are studied by considering a specific example.

This example is an Emden–Fowler equation, namely $y'' + ty^3 = 0$. It was selected because it is neither linear nor autonomous, and the oscillatory behaviour of its solutions will probably lead to substantial cancellations causing more difficulties. Yet it is still amenable to analysis, in particular because we can derive its asymptotic solution for large t. Finally, it seems a logical sequel to the Airy equation $y'' + ty = 0$, considered by Iserles [17].

The remainder of this paper is organized as follows. The next section introduces the Alekseev–Gröbner lemma and the theory of modified equations, and shows how they can be used to estimate the global error. The resulting *a priori* estimates are exact up to a term of order h^{2p}, where h is the step size and p the order of the method. In Section 3, we show how these estimates look like for the Airy equation. We condense the calculations because many are already described by Iserles [17]. Our specific nonlinear example is introduced in Section 4, where we also describe its asymptotic solution. Section 5 contains the actual calculation of global error estimates. We require some integrals, which are computed in the appendix. The estimates are supported by numerical experiments, as described in Section 6. The last section discusses the results, draws them into comparison with the work done by others, and presents some pointers for further research.

2 Theory

Suppose that we are solving the ordinary differential equation

$$y'(t) = f(t, y(t)), \qquad y(t_0) = y_0, \tag{1}$$

where $y(t) \in \mathbf{R}^d$ is a vector. We assume that f is C^∞ to avoid technicalities which are irrelevant to our argument.

Associated to this differential equation is the *flow map*, which is denoted here by $\Phi_s^t(y_0)$, where s stands for the initial time, t the final time, and y_0 the starting position. The map $\Phi_s^t : \mathbf{R}^d \to \mathbf{R}^d$ is defined by

$$\Phi_t^t(y) = y \quad \text{and} \quad \tfrac{\mathrm{d}}{\mathrm{d}t}\Phi_s^t(y) = f\big(t, \Phi_s^t(y)\big). \tag{2}$$

Its Jacobian matrix will be denoted by $D\Phi_s^t$ and is called the *variational flow*. It plays an important role in this paper, since it describes how perturbations are propagated by the flow of (1). This follows from the Alekseev–Gröbner lemma (see eg. [14]), which extends the variation-of-constants technique to nonlinear equations.

Lemma 1 (Alekseev–Gröbner). *Denote by y and z the solutions of*

$$y' = f(t, y), \qquad\qquad y(t_0) = y_0, \tag{3a}$$

$$z' = f(t, z) + g(t, z), \qquad\qquad z(t_0) = y_0, \tag{3b}$$

respectively and suppose that $\partial f / \partial y$ exists and is continuous. Then the solutions of (3a) *and of the "perturbed" equation* (3b) *are connected by*

$$z(t) = y(t) + \int_{t_0}^{t} D\Phi_s^t(z(s))\, g\big(s, z(s)\big)\, \mathrm{d}s. \tag{4}$$

Now suppose that one uses a numerical one-step method with fixed step size h to solve the differential equation (1). Letting $y_0 = y(t_0)$, we denote the subsequent results of the method by $y_1, y_2, y_3, \ldots$, and set $t_k = t_0 + kh$.

Of course, we hope that numerical results are close to the exact solution, $y_k \approx y(t_k)$. The quality of this approximation is measured by the *global error*, which is defined by $E_h(t_k) = y_k - y(t_k)$. We say that a method is of *order p* if $E_h(t) = \mathcal{O}(h^p)$.

As mentioned in the introduction, the goal of the analysis presented in this paper is to obtain *a priori* estimates for the global error. The general idea is to view the numerical "flow" as a perturbation of the actual flow of the equation. As the global error is the difference between the numerical and the exact solution, the Alekseev–Gröbner lemma can be used to estimate it.

There is however a problem here: the numerical method provides only values at times t_k, but not a continuous flow. Iserles and Söderlind [18] interpolated the numerical results to obtain a function defined on the whole time interval. Instead, we will use the theory of *modified equations*, following Calvo and Hairer [3]. This theory, analogous to Wilkinson's backward error analysis in Numerical Linear Algebra, tries to find a differential equation $z' = \hat{f}_h(t,z)$ whose solution is close to the numerical results. Hairer and Lubich [13] proved that one can even arrange for the solution to be exponentially close (in the step size h) to numerical results. This is more than sufficient to apply the following theorem.

Theorem 2. *Let $y(t)$ be the solution of the differential equation* (1) *and $D\Phi_s^t$ denote its variational flow. Suppose that a numerical method of order p produces the values $\{y_k\}$. If the solution $z(t)$ of the modified equation $z' = \hat{f}_h(t,z)$ satisfies $z(t_k) - y_k = \mathcal{O}(h^{2p})$ for all k, then*

$$E_h(t) = \int_{t_0}^{t} D\Phi_s^t(y(s))\, \delta_h\big(s, y(s)\big)\, \mathrm{d}s + \mathcal{O}(h^{2p}), \tag{5}$$

where $\delta_h(t,y) = \hat{f}_h(t,y) - f(t,y)$.

Proof. If we take the perturbed equation (3b) from the Alekseev–Gröbner lemma to be the modified equation, we deduce

$$z(t) - y(t) = \int_{t_0}^{t} D\Phi_s^t(z(s))\, \delta_h\big(s, z(s)\big)\, \mathrm{d}s.$$

The left-hand side of this equation is $E_h(t) + \mathcal{O}(h^{2p})$. Furthermore, since the method is of order p, we have $y(t) - z(t) = \mathcal{O}(h^p)$. Finally, it follows from the theory of modified equations that $\delta_h(t,y) = \mathcal{O}(h^p)$ (see eg. [11]). Together this proves (5). $\square$

To compute the error estimate (5), we need to know the exact solution $y(t)$, the variational flow $D\Phi_s^t$, and the modified equation to compute δ_h. Note that these can all be determined *a priori*, hence the above theorem gives an *a priori* estimate for the global error.

In fact, if one is interested in the long-term behaviour of the global error, it is not essential to know the exact solution over the whole time interval. It often suffices to know only the asymptotic behaviour of the solution for large t, because this will allow one to compute the dominant contribution to the integral (5). This approach is actually taken in the computations that will follow in the next sections.

From now on, we restrict our attention to *Runge–Kutta methods.* We assume that the reader is familiar with the theory of Runge–Kutta methods and B-series, as explained in the text books [14] and [19], amongst others. A B-series is a formal power series of the form

$$B(a,y) = a(\emptyset)y + \sum_{\tau\in T} \frac{h^{\rho(\tau)}}{\rho(\tau)!}\, \alpha(\tau)\, a(\tau)\, F(\tau)(y). \tag{6}$$

Here T denotes the set of rooted trees, $a : T \to \mathbf{R}$ is a coefficient function, $\rho(\tau)$ is the order of the tree τ, $\alpha(\tau)$ denotes the number of monotone labellings of τ, and $F(\tau)(y)$ is the elementary differential of the function f associated with τ. The result of any Runge–Kutta method can be written as a B-series, and the coefficient function a depends on the coefficients of the method.

Hairer [11] uses the concept of B-series to derive a formula for the modified equation, taking a similar approach to Benettin and Giorgilli [2]. Suppose that the outcome of the Runge–Kutta method is described by the B-series with a coefficient function a. Then the modified equation is given by $z' = \frac{1}{h}B(b, z)$, where the coefficients are recursively defined by

$$b(\emptyset) = 0,\ b(\bullet) = 1, \text{ and } b(\tau) = a(\tau) - \sum_{j=2}^{\rho(\tau)} \frac{1}{j!}\partial_b^{j-1} b(\tau) \text{ for } \rho(\tau) \geq 2. \tag{7}$$

In this formula, ∂_b is a Lie derivative, defined as follows. Given two coefficient functions b and c, the Lie derivative of $B(c, y)$ with respect to $B(b, y)$ is defined as $\frac{\mathrm{d}}{\mathrm{d}t}B(c, y_b(t))$, where $y_b(t)$ is a (formal) solution of the differential equation $y'(t) = B(b, y(t))$. It turns out that this is again a B-series, and its coefficient function is denoted by $\partial_b c$. An explicit formula for $\partial_b c$ is given in [11].

We now use this expression for the modified equation to rewrite the *a priori* estimate given in Theorem 2.

Theorem 3. *Let $y(t)$ be the solution of the differential equation* (1) *and let $D\Phi_s^t$ denote its variational flow. Suppose that this equation is solved by a Runge–Kutta method with B-series coefficient function a. Let b be defined as in* (7). *If the numerical method has order p, then the global error satisfies:*

$$\begin{aligned} E_h(t) &= \sum_{\tau \in T_{\geq 2}} h^{\rho(\tau)-1} b(\tau) \frac{\alpha(\tau)}{\rho(\tau)!} \mathcal{I}_\tau(t) + \mathcal{O}(h^{2p}), \\ \textit{where } \mathcal{I}_\tau(t) &= \int_{t_0}^t D\Phi_s^t(y(s)) F(\tau)(y(s))\, \mathrm{d}s. \end{aligned} \tag{8}$$

Here $T_{\geq 2}$ denotes the set of trees with order at least 2.

Proof. As mentioned above, the right-hand side of the modified equation is $\frac{1}{h}B(b, z)$, with b as in (7). Now $F(\bullet) = f$, so the first term in this B-series is the right-hand side of the original equation (1). Hence their difference $\delta_h(t, y)$ equals

$$\sum_{\tau \in T_{\geq 2}} \frac{h^{\rho(\tau)-1}}{\rho(\tau)!}\, \alpha(\tau)\, b(\tau)\, F(\tau)(y).$$

We now substitute this expression in (5) and move the scalar factors out of the integral (remember that the variational flow $D\Phi_s^t$ is a linear operator). This results in (8). □

Note that the error estimate (8) nicely separates the problem and the method. The method only enters the estimate via the coefficients $b(\tau)$. On the other hand, the value of the integrals $\mathcal{I}_\tau$ is completely determined by the particular equation one is solving. As their role in the global error estimate (8) is similar to the role of the elementary differential $F(\tau)(y)$ in the local error, we will call $\mathcal{I}_\tau$ the *elementary integral* associated with τ.

3 The Airy equation

Iserles [17] studied in detail the leading term of the global error committed by various methods, including Runge–Kutta methods, for the Airy equation

$$y'' + ty = 0, \quad t \geq 0, \qquad y(0) = y_0, \quad y'(0) = y'_0. \tag{9}$$

Here we will use Theorem 3 to extend his analysis to include higher-order terms. This will prepare us for the study of a nonlinear variant of the Airy equation, viz. (14).

The asymptotic solution of (9) for large t can be found by the Liouville–Green approximation [21], which is also known, particularly to physicists, as the WKB(J)-approximation:

$$\begin{bmatrix} y(t) \\ y'(t) \end{bmatrix} \approx \Lambda(t) R(\theta(t)) S_0. \tag{10}$$

Here S_0 is a vector whose value is determined by the initial conditions, and $\theta(t) = \frac{2}{3} t^{3/2}$. Furthermore,

$$\Lambda(t) = \begin{bmatrix} t^{-1/4} & 0 \\ 0 & t^{1/4} \end{bmatrix} \quad \text{and} \quad R(\theta) = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}.$$

Note that $\Lambda(t)$ is a scaling matrix and $R(\theta)$ is a rotation.

We want to apply (8) to derive an estimate for the global error, so we have to calculate the elementary differentials $F(\tau)(y)$. This has already been done by Iserles [17]. He concludes that amongst all the trees of order k, the tall tree without branches dominates. We denote this tree by $\tau_k^{|}$ (anticipating our analysis of a nonlinear variant of (9), compare with the trees in Table 1). The elementary differential associated to this tree is $F(\tau_k^{|})(y) = \hat{F}_k y$, with

$$\hat{F}_{2r} \approx \begin{bmatrix} (-t)^r & 0 \\ 0 & (-t)^r \end{bmatrix} \quad \text{and} \quad \hat{F}_{2r+1} \approx \begin{bmatrix} 0 & (-t)^r \\ (-t)^{r+1} & 0 \end{bmatrix}. \tag{11}$$

The next step is to compute the elementary integral $\mathcal{I}_{\tau_k^{|}}$, defined in (8). We will use the shorthand $\mathcal{I}_k$ for this elementary integral. Hence we need the variational flow $D\Phi_s^t$. But since the Airy equation (9) is linear, the variational flow equals the flow map Φ_s^t, so we can deduce from (10) that

$$D\Phi_s^t = \Phi_s^t \approx \Lambda(t) R\big(\theta(t) - \theta(s)\big) \Lambda^{-1}(s). \tag{12}$$

Now we need to substitute (10), (11), and (12) in the elementary integral $\mathcal{I}_k = \int_0^t D\Phi_s^t \hat{F}_k y(s)\,\mathrm{d}s$. For even k we find

$$\begin{aligned} \mathcal{I}_{2r} &\approx \int_0^t \Lambda(t) R\big(\theta(t) - \theta(s)\big) \Lambda^{-1}(s) \cdot (-s)^r \Lambda(s) R(\theta(s)) S_0 \,\mathrm{d}s \\ &= \int_0^t (-s)^r \,\mathrm{d}s \cdot \Lambda(t) R(\theta(t)) S_0 = (-1)^r \frac{t^r}{r+1} y(t). \end{aligned}$$

For odd k, the calculation is only a bit more complicated. First we compute the matrix product that appears in the middle of the integrand:

$$\Lambda^{-1}(s)\hat{F}_{2r+1}\Lambda(s) = (-1)^r s^{r+\frac{1}{2}} R(\tfrac{1}{2}\pi).$$

Note that $R(\frac{1}{2}\pi) = \left[\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right]$. With this result, we can tackle the integral,

$$\mathcal{I}_{2r+1} \approx \int_0^t (-1)^r s^{r+\frac{1}{2}}\,\mathrm{d}s \cdot \Lambda(t) R\big(\theta(t) + \tfrac{1}{2}\pi\big) S_0 = (-1)^r \frac{t^{r+\frac{3}{2}}}{r+\frac{3}{2}} y_R(t).$$

Here $y_R(t)$ is the (asymptotic) solution of the Airy equation (9) with opposite phase. It is given by

$$y_R(t) = \Lambda(t) R\big(\theta(t) + \tfrac{1}{2}\pi\big) S_0. \tag{13}$$

To find an estimate for the error, we have to substitute the above integrals in (8). This gives

$$\begin{aligned} E_h(t) \approx{}& C_1 h t^2 y(t) + C_2 h^2 t^{5/2} y_R(t) + C_3 h^3 t^3 y(t) \\ &+ C_4 h^4 t^{7/2} y_R(t) + \cdots + \mathcal{O}(h^{2p}), \text{where } C_k = \frac{b(\tau^{|}_{k+1})}{(k+1)!\,(\frac{1}{2}k+1)}. \end{aligned}$$

Note that for a method of order p, the B-series coefficient of any tree whose order does not exceed p vanishes. Therefore, the h^k term vanishes as well for $k < p$, in agreement with the general theory.

Note furthermore that the nature of the particular Runge–Kutta method used, has no influence on the growth rate of the global error. It does not matter whether the method is for instance A-stable or symplectic. There is one exception to this, as pointed out by Orel [22]: for methods with $b(\tau^{|}_p) = 0$, the dominant term in the above estimate vanishes. Hence such methods should perform better than general Runge–Kutta methods.

It is possible to generalize the asymptotic solution (10) to a class of nonautonomous oscillators. Under the assumptions given by Iserles [17] and reproduced in Theorem 4 below, the asymptotic solution of the equation $y'' + g(t)y = 0$ is still given by (10), except that now $\theta(t) = \int_0^t \sqrt{g(s)}\,\mathrm{d}s$. The rest of the computation stays essentially the same. Its result is given by the following theorem.

Theorem 4. *Suppose we are solving the nonautonomous linear oscillator*

$$y'' + g(t)y = 0, \quad t \geq 0, \qquad y(0) = y_0, \quad y'(0) = y_0',$$

where $g(t)$ is positive for sufficiently large t, $|g^{(\ell)}(t)| = o\big(g(t)^{1/\ell}\big)$ as $t \to \infty$, and $\limsup_{t\to\infty} g(t) = +\infty$. If we use a Runge–Kutta method with B-series coefficient function a, and b is defined as in (7), then the global error is

$$\begin{aligned} E_h(t) \approx{}& \left[\sum_{r=1}^{\infty} (-1)^r \frac{b(\tau^{|}_{2r+1})}{(2r+1)!} h^{2r} \int_{t_0}^t g(s)^{r+\frac{1}{2}}\,\mathrm{d}s\right] y_R(t) \\ &+ \left[\sum_{r=0}^{\infty} (-1)^{r+1} \frac{b(\tau^{|}_{2r+2})}{(2r+2)!} h^{2r+1} \int_{t_0}^t g(s)^{r+1}\,\mathrm{d}s\right] y(t) + \mathcal{O}(h^{2p}). \end{aligned}$$

Here p denotes the order of the method, and $y(t)$ and $y_R(t)$ are given by (10) and (13) respectively, with $\theta(t) = \int_0^t \sqrt{g(s)}\,\mathrm{d}s$. $\square$

4 A nonlinear oscillator

The previous section should be considered as a mere warm-up to the real challenge: a nonlinear oscillator. The following equation seems to us to be the simplest nonlinear variant of the Airy equation (9),

$$y'' + ty^3 = 0, \quad t \geq 0, \qquad y(0) = y_0, \quad y'(0) = y'_0. \tag{14}$$

It is a particular instance of the Emden–Fowler equation $y'' + t^\nu y^\gamma = 0$, see eg. [26]. If $\nu = 1 - \gamma$, the Emden–Fowler equation reduces to the Lane–Emden equation which was originally proposed to model stars. Considering the star as a radially symmetric gaseous polytrope of index γ in thermodynamic and hydrostatic equilibrium, the relation between the density and the distance to the centre satisfies the Lane–Emden equation [6].

The equation (14) has an obvious scaling symmetry, which can be used to reduce the order. Here we will use a different approach though, because the asymptotic solution as $t \to \infty$ will suffice for our purposes. Consider the transformation

$$y(t) = \tfrac{2}{3}\sqrt{2}\, t^{-1/6} w(t^{4/3}). \tag{15}$$

If we apply (15) to the differential equation (14), we get

$$\tfrac{32}{27}\sqrt{2}\, t^{1/2}\big[w''(t^{4/3}) + \tfrac{1}{2}w^3(t^{4/3}) + \tfrac{7}{64}t^{-8/3}w(t^{4/3})\big] = 0.$$

For large t, we can (hopefully) neglect the $t^{-8/3}$ term and the equation is approximated well by to the biharmonic oscillator $w'' + \frac{1}{2}w^3 = 0$. The general solution of this equation is

$$w(t) = c_1 \operatorname{sd}(c_1 t + c_2 \mid \tfrac{1}{2}), \tag{16}$$

where $\operatorname{sd}(\,\cdot \mid \frac{1}{2})$ denotes the Jacobi elliptic function with parameter $m = \frac{1}{2}$, see eg. [20]. Note that c_1 determines the amplitude of the oscillation, while c_2 determines the phase. In other, more sophisticated terms, (c_1, c_2) are action-angle coordinates of the Hamiltonian system corresponding to the biharmonic oscillator (see eg. [1]).

From now on, we will drop the parameter $\frac{1}{2}$ in (16). The function $\operatorname{sd}(t)$, restricted to the real axis, is depicted in Figure 1. It is an odd, periodic function with period $t_{\text{per}} = 2\pi^{3/2}[\Gamma(\frac{3}{4})]^{-2} \approx 7.416$.

To obtain the solution of the original differential equation (14), we simply substitute (16) in (15). This yields an asymptotic solution of (14) for large t,

$$y(t) \approx \tfrac{2}{3}\sqrt{2}\, c_1 t^{-1/6} \operatorname{sd}(c_1 t^{4/3} + c_2). \tag{17}$$

So the amplitude of the oscillations attenuates, but the oscillations grow ever faster.

In Figure 2, we plot the Liouville–Green approximation (17), together with the exact solution of (14), using $y(0) = 1$ and $y'(0) = 0$ as initial conditions. Note that the approximation is already close for modest values of t. In the computation to be performed in the next section, we will in fact pretend that (17) gives the exact solution.

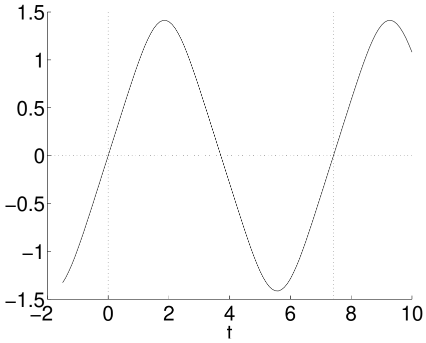


Figure 1: The function $\mathrm{sd}(t)$.

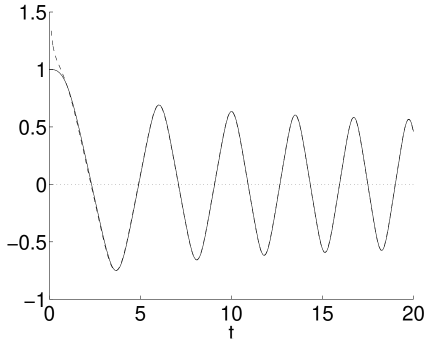


Figure 2: The exact solution (solid line), and the asymptotic approximation (dashed line).

5 Global error of Runge–Kutta methods

In this section, we repeat the computation of Section 3 for the nonlinear oscillator (14) with asymptotic solution (17).

The first task is the computation of elementary differentials $F(\tau)(y)$. For this we need to convert the differential equation to a system of autonomous first-order equations,

$$\begin{aligned} y_1' &= y_2, \\ y_2' &= -y_3 y_1^3, \\ y_3' &= 1. \end{aligned} \tag{18}$$

Here y_1, y_2, and y_3 correspond to y, y', and t in the original equation (14) respectively.

To get some intuition for the behaviour of the elementary differentials, we compute the differentials corresponding to the trees of order ≤ 4. The results are assembled in Table 1. Looking at the asymptotic solution (17), we see that $y_1 = \mathcal{O}(t^{-1/6})$ and $y_2 = \mathcal{O}(t^{1/6})$. This allows us to compute the order of magnitude of the elementary differentials, which we also list in Table 1. We see that the differentials corresponding to trees of equal order have the same order of magnitude. This is in stark contrast to the linear Airy equation, where the differential corresponding to the tall tree $\tau_k^|$ dominates.

Table 1 suggests that the elementary differential corresponding to a tree τ of order k is of the form

$$F(\tau)(y) = \begin{bmatrix} C_1 y_1^{4c+1-k} y_2^{k-2c} y_3^k + \mathcal{O}(y_3^{k-1}) \\ C_2 y_1^{4d-k} y_2^{k+1-2d} y_3^k + \mathcal{O}(y_3^{k-1}) \end{bmatrix}, \tag{19}$$

where c and d are integers such that none of the exponents is negative, and C_i denotes a constant, which may depend on the tree τ.

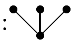
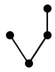
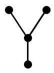
Tree τ	Elementary differential $F(\tau)(y)$	Elementary integral $\mathcal{I}_\tau(t) = \int_0^t D\Phi_s^t F(\tau)(y)\,\mathrm{d}s$
$\tau_2^{\vert}$:	$\begin{bmatrix} -y_1^3 y_3 \\ -3y_1^2 y_2 y_3 - y_1^3 \end{bmatrix} = \begin{bmatrix} \mathcal{O}(t^{1/2}) \\ \mathcal{O}(t^{5/6}) \end{bmatrix}$	$\begin{bmatrix} -\frac{128}{675}\sqrt{2}\,c_1^4 \chi t^{17/6}\,\mathrm{sd}'(\tilde{t}) \\ \frac{256}{2025}\sqrt{2}\,c_1^5 \chi t^{19/6}\,\mathrm{sd}^3(\tilde{t}) \end{bmatrix}$
τ_3^a: 	$\begin{bmatrix} 0 \\ -6y_1 y_2^2 y_3 - 6y_1^2 y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \mathcal{O}(t^{7/6}) \end{bmatrix}$	$\begin{bmatrix} -\frac{32}{135}\sqrt{2}\,c_1^4 \chi t^{11/6}\,\mathrm{sd}'(\tilde{t}) \\ \frac{64}{405}\sqrt{2}\,c_1^5 \chi t^{13/6}\,\mathrm{sd}^3(\tilde{t}) \end{bmatrix}$
$\tau_3^{\vert}$:	$\begin{bmatrix} -3y_1^2 y_2 y_3 - y_1^3 \\ 3y_1^5 y_3^2 \end{bmatrix} = \begin{bmatrix} \mathcal{O}(t^{5/6}) \\ \mathcal{O}(t^{7/6}) \end{bmatrix}$	$\begin{bmatrix} -\frac{208}{135}\sqrt{2}\,c_1^4 \chi t^{11/6}\,\mathrm{sd}'(\tilde{t}) \\ \frac{416}{405}\sqrt{2}\,c_1^5 \chi t^{13/6}\,\mathrm{sd}^3(\tilde{t}) \end{bmatrix}$
τ_4^a: 	$\begin{bmatrix} 0 \\ -6y_2^3 y_3 - 18y_1 y_2^2 \end{bmatrix} = \begin{bmatrix} 0 \\ \mathcal{O}(t^{3/2}) \end{bmatrix}$	$\begin{bmatrix} -\frac{4096}{14553}\sqrt{2}\,c_1^6 t^{7/2}\,\mathrm{sd}'(\tilde{t}) \\ \frac{8192}{43659}\sqrt{2}\,c_1^7 t^{23/6}\,\mathrm{sd}^3(\tilde{t}) \end{bmatrix}$
τ_4^b: 	$\begin{bmatrix} 0 \\ 6y_1^4 y_2 y_3^2 + 3y_1^5 y_3 \end{bmatrix} = \begin{bmatrix} 0 \\ \mathcal{O}(t^{3/2}) \end{bmatrix}$	$\begin{bmatrix} \frac{4096}{43659}\sqrt{2}\,c_1^6 t^{7/2}\,\mathrm{sd}'(\tilde{t}) \\ -\frac{8192}{130977}\sqrt{2}\,c_1^7 t^{23/6}\,\mathrm{sd}^3(\tilde{t}) \end{bmatrix}$
τ_4^c: 	$\begin{bmatrix} -6y_1 y_2^2 y_3 - 6y_1^2 y_2 \\ 0 \end{bmatrix} = \begin{bmatrix} \mathcal{O}(t^{7/6}) \\ 0 \end{bmatrix}$	$\begin{bmatrix} -\frac{4096}{43659}\sqrt{2}\,c_1^6 t^{7/2}\,\mathrm{sd}'(\tilde{t}) \\ \frac{8192}{130977}\sqrt{2}\,c_1^7 t^{23/6}\,\mathrm{sd}^3(\tilde{t}) \end{bmatrix}$
$\tau_4^{\vert}$:	$\begin{bmatrix} 3y_1^5 y_3^2 \\ 9y_1^4 y_2 y_3^2 + 3y_1^5 y_3 \end{bmatrix} = \begin{bmatrix} \mathcal{O}(t^{7/6}) \\ \mathcal{O}(t^{3/2}) \end{bmatrix}$	$\begin{bmatrix} \frac{16384}{43659}\sqrt{2}\,c_1^6 t^{7/2}\,\mathrm{sd}'(\tilde{t}) \\ -\frac{32768}{130977}\sqrt{2}\,c_1^7 t^{23/6}\,\mathrm{sd}^3(\tilde{t}) \end{bmatrix}$

Table 1: Trees of order ≤ 4, with their elementary differentials and integrals for the nonlinear oscillator (18). The third component is suppressed as it is always zero. In the last column, χ is as in (33), $\tilde{t} = c_1 t^{4/3} + c_2$, and only the term of leading order is displayed.

To prove this, we initially neglect the third component. From the definition of the elementary differentials, it follows that they obey the following recursive relations (where the argument y is deleted):

$$
\begin{aligned}
F_1(\bullet) &= y_2, & F_2(\bullet) &= -y_1^3 y_3,\\
F_1\left(\substack{\tau\\ |\\ \bullet}\right) &= F_2(\tau), & F_2\left(\substack{\tau\\ |\\ \bullet}\right) &= -3y_1^2 y_3 F_1(\tau),\\
F_1\left(\substack{\tau_1 \,\cdots\, \tau_k\\ \searrow\ \swarrow\\ \bullet}\right) &= 0 \quad (k \ge 2), & F_2\left(\substack{\tau_1 \quad \tau_2\\ \searrow\ \swarrow\\ \bullet}\right) &= -6y_1 F_1(\tau_1) F_1(\tau_2),\\
& & F_2\left(\substack{\tau_1\ \tau_2\ \tau_3\\ \searrow\ \downarrow\ \swarrow\\ \bullet}\right) &= -6F_1(\tau_1)F_1(\tau_2)F_1(\tau_3),\\
& & F_2\left(\substack{\tau_1 \,\cdots\, \tau_k\\ \searrow\ \swarrow\\ \bullet}\right) &= 0 \quad (k \ge 4).
\end{aligned}
$$

The result (19) easily follows by induction. Insofar as the third component (representing time) is concerned, careful reasoning shows that its effect is to add more terms to (19). But the extra terms have lower order of magnitude. To be precise, differentiating with respect to y_3 decreases the powers of y_2 and y_3 by one, and increases the power of y_1 by one. These extra terms are represented by the $\mathcal{O}$ terms in (19).

If we substitute the asymptotic solution (17) in the above formula for elementary differentials, we find

$$
F(\tau)(y(t)) = \begin{bmatrix} C_3\, t^{\frac{1}{3}k-\frac{1}{6}}\, \mathrm{sd}^{4c+1-k}(\tilde{t})\, (\mathrm{sd}')^{k-2c}(\tilde{t}) + \mathcal{O}(t^{\frac{1}{3}k-\frac{3}{2}}) \\ C_4\, t^{\frac{1}{3}k+\frac{1}{6}}\, \mathrm{sd}^{4d-k}(\tilde{t})\, (\mathrm{sd}')^{k+1-2d}(\tilde{t}) + \mathcal{O}(t^{\frac{1}{3}k-\frac{7}{6}}) \end{bmatrix}. \tag{20}
$$

Here we use the abbreviation $\tilde{t} = c_1 t^{4/3} + c_2$. Note that the growth rates of the elementary differentials are indeed solely determined by the order k.

The next step is to calculate the elementary integral $\mathcal{I}_\tau$. For this, we need to multiply the above differential with the variational flow matrix and integrate the resulting expression. To compute the variational flow, we introduce the map $X_t : \mathbf{R}^2 \to \mathbf{R}^2$, defined by:

$$
X_t(\mathbf{c}) = \begin{bmatrix} \frac{2}{3}\sqrt{2}\, c_1 t^{-1/6}\, \mathrm{sd}(c_1 t^{4/3} + c_2) \\ \frac{8}{9}\sqrt{2}\, c_1^2 t^{1/6}\, \mathrm{sd}(c_1 t^{4/3} + c_2) \end{bmatrix}. \tag{21}
$$

So X_t maps the parameter space to the solution space at time t. Neglecting lower-order terms, it follows that the flow map satisfies $\Phi_s^t = X_t \circ X_s^{-1}$. Hence we can write the elementary integral (8) as:

$$
\mathcal{I}_\tau(t) \approx DX_t(\mathbf{c}) \int_0^t DX_s^{-1}(y(s))\, F(\tau)(y(s))\, \mathrm{d}s. \tag{22}
$$

To find the integrand in the above expression, we multiply the inverse of the Jacobian matrix of (21) with (20). With the help of (32), we can write the integrand as

$$
\begin{bmatrix} s^{\frac{1}{3}k}\big(C_5\, \mathrm{sd}^{4c+4-k}(\tilde{s})\, (\mathrm{sd}')^{k-2c}(\tilde{s}) + C_6\, \mathrm{sd}^{4c-k}(\tilde{s})\, (\mathrm{sd}')^{k+2-2c}(\tilde{s})\big) + \mathcal{O}(s^{\frac{1}{3}k-\frac{4}{3}}) \\ s^{\frac{1}{3}k+\frac{4}{3}}\big(C_7\, \mathrm{sd}^{4d+4-k}(\tilde{s})\, (\mathrm{sd}')^{k-2d}(\tilde{s}) + C_8\, \mathrm{sd}^{4d-k}(\tilde{s})\, (\mathrm{sd}')^{k+2-2d}(\tilde{s})\big) + \mathcal{O}(s^{\frac{1}{3}k}) \end{bmatrix}, \tag{23}
$$

where $\tilde{s} = c_1 s^{4/3} + c_2$. Now consider the exponents of $\mathrm{sd}(\tilde{s})$ and $\mathrm{sd}'(\tilde{s})$ in the above expression. If k is even, these exponents are also even and hence the integrand is nonnegative. However, in the case when k is odd, the elliptic functions are raised to an odd power, which means that the integrand oscillates around zero. Thus we can expect cancellations in the latter case, but not if k is even. We stress that this phenomenon does not occur in the linear case, analysed in Section 3.

The calculations performed in Appendix A substantiate this claim. We can use the final result (35) to integrate (23). The outcome is

$$\int_0^t DX_s^{-1} F(\tau)(y)\,\mathrm{d}s = \begin{cases} \begin{bmatrix} C_9\, t^{\frac{1}{3}k+1} + \mathcal{O}(t^{\frac{1}{3}k-\frac{1}{3}}) \\ C_{10}\, t^{\frac{1}{3}k+\frac{7}{3}} + \mathcal{O}(t^{\frac{1}{3}k+1}) \end{bmatrix}, & \text{if } k \text{ is even}, \\ \begin{bmatrix} \tilde{C}_{11}(\tilde{t})\, t^{\frac{1}{3}k-\frac{1}{3}} + \mathcal{O}(t^{\frac{1}{3}k-\frac{5}{3}}) \\ \tilde{C}_{12}(\tilde{t})\, t^{\frac{1}{3}k+1} + \mathcal{O}(t^{\frac{1}{3}k-\frac{1}{3}}) \end{bmatrix}, & \text{if } k \text{ is odd}. \end{cases} \tag{24}$$

As explained in Appendix A, $\tilde{C}_i(\,\cdot\,)$ denotes a periodic function, so $\tilde{C}_i(\tilde{t})$ oscillates with the same "local period" as the solution of the differential equation.

Actually, the case when k is odd requires a more precise calculation. Due to the oscillations in the leading term of the integrand (23), the remainder term also needs to be considered. The first component has the form

$$C_{13} s^{\frac{1}{3}k-\frac{4}{3}}\, \mathrm{sd}^{\ell}(\tilde{s})\, (\mathrm{sd}')^{m}(\tilde{s}) + \mathcal{O}(s^{\frac{1}{3}k-\frac{8}{3}}),$$

where the exponents ℓ and m are even. Hence its integral is $C_{14} t^{\frac{1}{3}k} + \mathcal{O}(t^{\frac{1}{3}k-\frac{4}{3}})$, cf. (35). We conclude that it has the same order as the integral of the leading term in (23). The same goes for the second component.

To compute the elementary integral $\mathcal{I}_\tau$, we need to premultiply the integral (24) with DX_t, cf. (22). But the expression (24) has an interpretation by itself. Remember that X_t maps the parameter space to the solution space. So the integral (24) represents the error in the parameter space. We conclude that the energy error associated with the tree τ grows as $t^{\frac{1}{3}k+1}$ if k is even, and as $t^{\frac{1}{3}k-\frac{1}{3}}$ if k is odd. The second component gives the phase error.

Multiplying the Jacobian matrix of the map X_t with the integral (24) gives us the elementary integrals

$$\mathcal{I}_\tau(t) = \begin{cases} \begin{bmatrix} \tilde{C}_{15}(\tilde{t})\, t^{\frac{1}{3}k+\frac{13}{6}} + \mathcal{O}(t^{\frac{1}{3}k+\frac{5}{6}}) \\ \tilde{C}_{16}(\tilde{t})\, t^{\frac{1}{3}k+\frac{5}{2}} + \mathcal{O}(t^{\frac{1}{3}k+\frac{7}{6}}) \end{bmatrix}, & \text{if } k \text{ is even}, \\ \begin{bmatrix} \tilde{C}_{17}(\tilde{t})\, t^{\frac{1}{3}k+\frac{5}{6}} + \mathcal{O}(t^{\frac{1}{3}k-\frac{1}{2}}) \\ \tilde{C}_{18}(\tilde{t})\, t^{\frac{1}{3}k+\frac{7}{6}} + \mathcal{O}(t^{\frac{1}{3}k-\frac{1}{6}}) \end{bmatrix}, & \text{if } k \text{ is odd}. \end{cases}$$

The elementary integrals for trees whose order does not exceed 4 are listed in Table 1. These results include the actual value of the $\tilde{C}_i(\tilde{t})$ in the above equation.

Finally we can find an estimate for the global error by adding the contributions of all trees according to (8). For the first component, we find (where the $\tilde{C}_i$ are different from above)

$$E_h(t) \approx \tilde{C}_1(\tilde{t})ht^{17/6} + \tilde{C}_2(\tilde{t})h^2t^{11/6} + \tilde{C}_3(\tilde{t})h^3t^{7/2} + \tilde{C}_4(\tilde{t})h^4t^{5/2} + \cdots + \mathcal{O}(h^{2p}).$$

More formally, we have the following result.

Theorem 5. *Suppose we are solving the nonlinear oscillator equation*

$$y'' + ty^3 = 0, \quad t \geq 0, \qquad y(0) = y_0, \quad y'(0) = y_0',$$

with a numerical method of order p, that can be expressed as a B-series. Then the global error has the form

$$\begin{aligned} E_h(t) = \sum_{2r \geq p} h^{2r} &\begin{bmatrix} \tilde{C}^1_{2r}(\tilde{t})\, t^{\frac{2}{3}r+\frac{7}{6}} + \mathcal{O}(t^{\frac{2}{3}r-\frac{1}{6}}) \\ \tilde{C}^2_{2r}(\tilde{t})\, t^{\frac{2}{3}r+\frac{3}{2}} + \mathcal{O}(t^{\frac{2}{3}r+\frac{1}{6}}) \end{bmatrix} \\ &+ \sum_{2r+1 \geq p} h^{2r+1} \begin{bmatrix} \tilde{C}^1_{2r+1}(\tilde{t})\, t^{\frac{2}{3}r+\frac{17}{6}} + \mathcal{O}(t^{\frac{2}{3}r+\frac{3}{2}}) \\ \tilde{C}^2_{2r+1}(\tilde{t})\, t^{\frac{2}{3}r+\frac{19}{6}} + \mathcal{O}(t^{\frac{2}{3}r+\frac{11}{6}}) \end{bmatrix} + \mathcal{O}(h^{2p}). \end{aligned}$$

Here $\tilde{C}^i_k(\tilde{t})$ denotes a function that is periodic in $\tilde{t} = c_1t^{4/3} + c_2$. □

We would again like to draw the reader's attention to the difference between the even and the odd powers of h. Unfortunately, we do not know what causes this difference, and whether it also occurs for other differential equations.

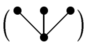
6 Numerical experiments

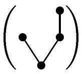
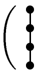
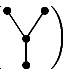
The purpose of this section is to supplement the calculations of the previous section with some numerical experiments. In particular, we want to see whether we were justified in using the asymptotic solution (17). Furthermore, it could be interesting to study whether the $\mathcal{O}(h^p)$ term of the global error $E_h(t)$ dominates, or whether other terms have to be taken into account.

Our first example concerns Runge's second order method, given by

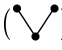
$$y_{n+1} = y_n + hf\big(t_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}hf(t_n, y_n)\big). \tag{25}$$

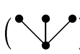
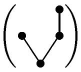
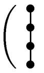
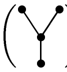
This is a Runge–Kutta method, hence it can be written as a B-series $B(a_R, y)$. Some values of the function a_R are

$a_R(\bullet) = 1$, $a_R($$) = 1$, $a_R($$) = \frac{3}{4}$, $a_R($$) = \frac{1}{2}$,

$a_R($$) = a_R($$) = a_R($$) = a_R($$) = 0.$

It follows from (7) that the B-series coefficients of the modified equation are

$b_R(\bullet) = 1$, $b_R($$) = 0$, $b_R($$) = -\frac{1}{4}$, $b_R($$) = -1$,

$b_R($$) = 0$, $b_R($$) = 0$, $b_R($$) = 3$, $b_R($$) = \frac{3}{2}.$

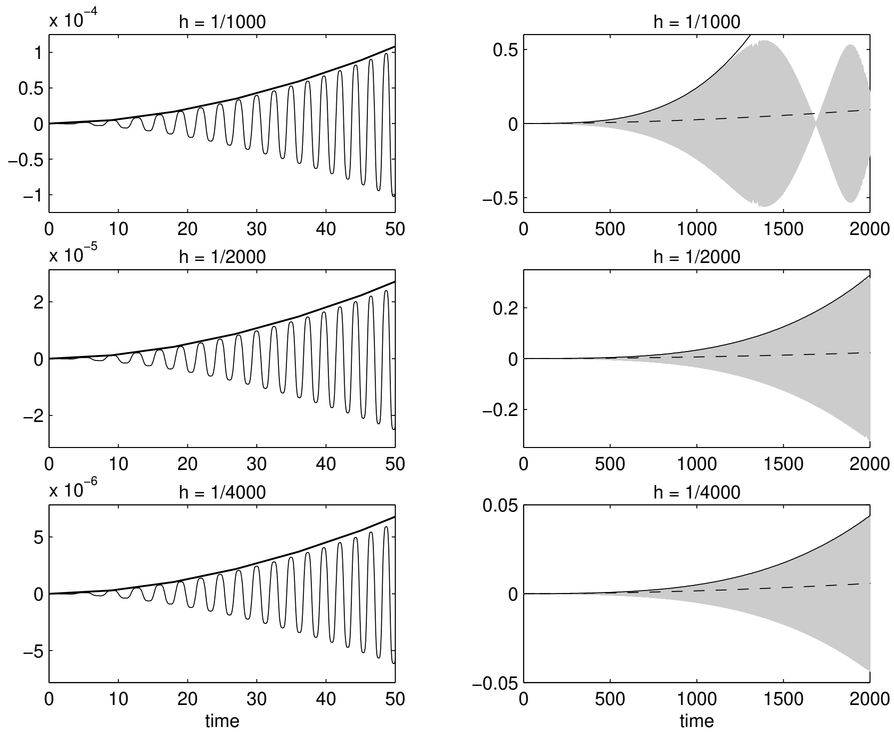


Figure 3: The first component of the global error committed by Runge's second-order method, and the estimate (26).

If we use these numbers in the error estimate (8), together with the elementary integrals from Table 1, we obtain the following estimate for the global error committed by Runge's method:

$$E_h(t) \approx h^2 \begin{bmatrix} \frac{4}{15}\sqrt{2}\, c_1^4 \chi t^{11/6}\,\mathrm{sd}'(\tilde{t}) \\ -\frac{8}{45}\sqrt{2}\, c_1^5 \chi t^{13/6}\,\mathrm{sd}^3(\tilde{t}) \end{bmatrix} + h^3 \begin{bmatrix} \frac{256}{6237}\sqrt{2}\, c_1^6 t^{7/2}\,\mathrm{sd}'(\tilde{t}) \\ -\frac{512}{18711}\sqrt{2}\, c_1^7 t^{23/6}\,\mathrm{sd}^3(\tilde{t}) \end{bmatrix}. \tag{26}$$

To check this estimate, we compute the solution of the nonlinear oscillator (14) with Runge's second-order method (25). The initial conditions are $y(0) = 1$ and $y'(0) = 0$, which lead to a solution with $c_1 \approx 0.7$. The solution is compared to the result of the standard fourth-order Runge–Kutta method with step size $h = 1/10000$. According to Theorem 5, this would give an error of about 10^{-9}, so we can consider this to be the exact solution. Hence the global error $E_h(t)$ is computed by subtracting the result of Runge's method from the "exact" solution. The first component is depicted in Figure 3. The left column shows the time interval $[0, 50]$, and on the right the larger interval $[0, 2000]$ is displayed.

In the left column of Figure 3, the oscillating curve shows the global error $E_h(t)$. The oscillating behaviour is correctly predicted by the error estimate (26); this is the factor $\mathrm{sd}'(\tilde{t})$. The amplitude of the oscillations, as predicted by (26), is shown by the thick curve in the left-hand column in Figure 3. We see that the estimate (26) describes the actual error accurately.

The right-hand column of Figure 3 shows a much larger time interval. Here the oscillations are compressed so heavily that the error appears as a grey blob. The dashed curve shows the first, leading term of the error estimate (26), and the solid curve shows the sum of both terms. We conclude that the leading h^2 term of the estimate does not describe the actual error correctly, but that the error is predicted accurately if the h^3 term is included. For $h = 1/1000$ the latter estimate breaks down around $t = 1200$. We note that at that point, the amplitude of the solution is approximately 0.3 and the error has about the same size, so the numerical solution has deviated considerably from the exact solution. For smaller values of the step size, the estimate (26) is accurate over the entire time interval $[0, 2000]$.

It follows that for large t (which here means: in the order of 1000), Runge's method behaves essentially as a third-order method. To check this, we compare it with Heun's classical third-order method:

$$\begin{aligned} k_1 &= f(t_n, y_n) \\ k_2 &= f(t_n + \tfrac{1}{3}h, y_n + \tfrac{1}{3}hk_1) \\ k_3 &= f(t_n + \tfrac{2}{3}h, y_n + \tfrac{1}{3}hk_2) \\ y_{n+1} &= y_n + h(\tfrac{1}{3}k_1 + \tfrac{2}{3}k_3) \end{aligned} \qquad \begin{array}{c} \text{Butcher tableau:} \\ \begin{array}{c|ccc} 0 & & & \\ 1/3 & 1/3 & & \\ 2/3 & 0 & 2/3 & \\ \hline & 1/4 & 0 & 3/4 \end{array} \end{array} \tag{27}$$

A similar calculation as for Runge's second-order method gives the following estimate for the global error committed by this method:

$$\begin{aligned} E_h(t) \approx h^3 & \begin{bmatrix} -\frac{512}{35721}\sqrt{2}\,c_1^6 t^{7/2}\,\mathrm{sd}'(\tilde{t}) \\ \frac{1024}{107163}\sqrt{2}\,c_1^7 t^{23/6}\,\mathrm{sd}^3(\tilde{t}) \end{bmatrix} + h^4 \begin{bmatrix} \frac{50208}{229635}\sqrt{2}\,c_1^6 t^{5/2}\,\mathrm{sd}'(\tilde{t}) \\ -\frac{60416}{688905}\sqrt{2}\,c_1^7 t^{17/6}\,\mathrm{sd}^3(\tilde{t}) \end{bmatrix} \\ & + h^5 \begin{bmatrix} \frac{557056}{34543665}\sqrt{2}\,c_1^8 \chi t^{25/6}\,\mathrm{sd}'(\tilde{t}) \\ -\frac{1114112}{103630995}\sqrt{2}\,c_1^9 \chi t^{9/2}\,\mathrm{sd}^3(\tilde{t}) \end{bmatrix}. \end{aligned} \tag{28}$$

The actual error and the above estimates, for step size $h = 1/2000$, are displayed in the second row of Figure 4. The first row displays Runge's second-order method. We see that the error estimate (28) again provides an excellent description of the actual error. Furthermore, the difference in order between Runge's and Heun's methods shows clearly for small values of t (see the left-hand column in Figure 4). For large values of t however (cf. the right-hand column), Runge's method behaves essentially as a third-order method and we see indeed that the difference between the two methods is much smaller.

The bottom row in Figure 4 shows a specially tuned method. Remember from Section 3 that for the Airy equation (9), the contribution of the tree $\tau_k^{|}$ to the global error dominates, and this is used by Orel [22] to construct specially tuned Runge–Kutta methods. We found in Section 5 that this is not the case for the nonlinear oscillator (14). However, if we carefully study the elementary

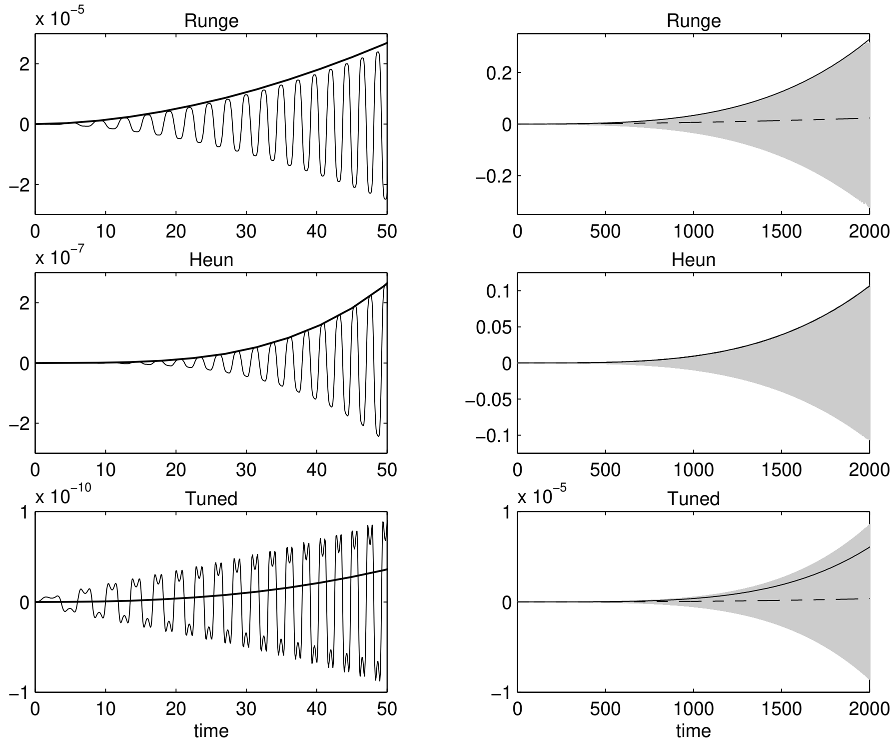
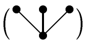
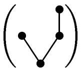
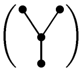


Figure 4: The first component of the global error committed by Runge's second-order method (25), Heun's third-order method (27), and the specially tuned third-order method (29), all with step size $h = 1/2000$, together with their respective error estimates (26), (28), and (30). The lines in the plot have the same meaning as in Figure 3.

integrals $\mathcal{I}_\tau$ in Table 1, we see that they are scalar multiples of each other:

$$4\mathcal{I}_{\tau_4^a}(t) = -12\mathcal{I}_{\tau_4^b}(t) = 12\mathcal{I}_{\tau_4^c}(t) = -3\mathcal{I}_{\tau_4^{|}}(t).$$

Therefore, bearing in mind the factor $\alpha(\tau)$ in Theorem 3, the h^3 term will be killed for a third-order method with

$3b($$) - 3b($$) + b($$) - 4b($$) = 0.$

From the general theory, it follows that a method has order three if

$b(\bullet) = 1$ and $b($$) = b($$) = b($$) = 0.$

These are five conditions together. An explicit 3-stage Runge–Kutta method has six free parameters, so there is some hope that we can find such a method satisfying these five conditions. Indeed, it turns out that there exists a one-parameter family of these methods. A particular instance is:

$$\begin{aligned} k_1 &= f(t_n, y_n) \\ k_2 &= f(t_n + h, y_n + hk_1) \\ k_3 &= f(t_n + \tfrac{3}{2}h, y_n + \tfrac{9}{4}hk_1 - \tfrac{3}{4}hk_2) \\ y_{n+1} &= y_n + h(\tfrac{7}{18}k_1 + \tfrac{5}{6}k_2 - \tfrac{2}{9}k_3) \end{aligned} \qquad \text{Butcher tableau:} \quad \begin{array}{c|ccc} 0 & & & \\ 1 & 1 & & \\ 3/2 & 9/4 & -3/4 & \\ \hline & 7/18 & 5/6 & -2/9 \end{array} \tag{29}$$

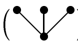
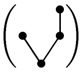
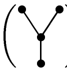
If we compute the estimate for the global error of this method, we find

$$E_h(t) \approx h^4 \begin{bmatrix} -\frac{5008}{25515}\sqrt{2}\,c_1^6 t^{5/2}\,\mathrm{sd}'(\tilde{t}) \\ \frac{10016}{76545}\sqrt{2}\,c_1^7 t^{17/6}\,\mathrm{sd}^3(\tilde{t}) \end{bmatrix} + h^5 \begin{bmatrix} -\frac{78848}{1279395}\sqrt{2}\,c_1^8 \chi t^{25/6}\,\mathrm{sd}'(\tilde{t}) \\ \frac{157696}{3838185}\sqrt{2}\,c_1^9 \chi t^{9/2}\,\mathrm{sd}^3(\tilde{t}) \end{bmatrix}. \tag{30}$$

As expected, the h^3 term has disappeared. This estimate, together with the actual error, is displayed in the bottom row of Figure 4. We see that the global error of this method is much smaller than that of Heun's method, though both methods are of third order. The other thing to note is that the error estimate (30) is not as accurate as for the other methods, even though it predicts the right order of magnitude.

We stress that we are not recommending the use of the method (29); it works well for *this* equation but we know of no reason whatsoever why it should also perform well for other equations. It may however be important to gain a better understanding of the mechanisms that cause this method to work so well.

The reader should keep in mind that the *local* error of the method (29) is still $\mathcal{O}(h^4)$, as for all third-order methods. In fact, the B-series coefficient function a_T of the method (29) satisfies

$a_T($$) = \frac{1}{3},\ a_T($$) = 2,\ a_T($$) = 2,\ a_T($$) = 0.$

So the local error of this method is

$$B(a_T - e, y) = h^4 \begin{bmatrix} -\frac{1}{8}y_1^5 y_3^2 - \frac{1}{4}y_1 y_2^2 y_3 - \frac{1}{4}y_1^2 y_2 \\ \frac{3}{8}y_1^4 y_2 y_3^2 + \frac{1}{4}y_1^5 y_3 + \frac{1}{6}y_2^3 y_3 + \frac{1}{2}y_1 y_2^2 \end{bmatrix} + \mathcal{O}(h^5).$$

where e denotes the function with $e(\emptyset) = 0$ and $e(\tau) = 1$ for $\tau \neq \emptyset$ (the B-series $B(e, y)$ gives the exact solution [11], so $B(a_T - e, y)$ is the local error).

It follows that the *global* error $E_h(t)$ must be $\mathcal{O}(h^3)$, in contrast to the error estimate (30). It is only the term of order $h^3[t^{7/2}, t^{23/6}]^\top$, which for generic third-order methods dominates the h^3 term, that disappears. But there still is an h^3 contribution to the global error. We believe that this contribution explains the substantial difference between the estimated and the actual error for this method, as displayed in Figure 4. The difference in the shape of the oscillations of the global error for the last method supports this conclusion.

7 Discussion

We used a combination of the Alekseev–Gröbner lemma, the theory of modified equations, and asymptotics to calculate an estimate for the global error committed by a Runge–Kutta method for a class of linear oscillators, and a specific nonlinear oscillator. Numerical experiments show that these estimates are in general accurate.

While in the linear case, the global error behaves more or less as we expected after the analysis by Iserles [17], the nonlinear case brings some surprises which indicate that the entire calculation of Section 5 is necessary. The specially tuned method (29) shows that it is not enough to consider only the local error. It neither suffices to study just the first term of order h^p; Figure 4 shows that the next term sometimes dominates. Finally, a comparison of Theorems 4 and 5 shows that linear and nonlinear oscillators can behave differently.

Let us highlight these differences. In the linear case, the contribution of one single tree (namely the tall, branchless tree) dominates the local error. This is no longer the case for the nonlinear oscillator (14). On the other hand, the contributions of all trees of the same order to the global error are constant multiples of one another (this has been checked for order ≤ 6). This allows us to kill the leading term of the global error. Experiments show that this strategy indeed works, but we still do not completely understand what is going on.

The other difference concerns the integral in the error estimate. The integration may bring down the time exponent due to oscillations in the integrand. For our equation this happens for even powers of h. As a consequence, the lowest order term in the expansion of the global error in powers of h does not always reveal the true behaviour of the global error, as the next term may dominate for sufficiently large values of t.

We stress that we have only considered one particular nonlinear oscillator, namely the Emden–Fowler equation (14). Hence we do not know how general these phenomena are. We plan to try and repeat the above analysis for other equations. An obvious target is the generalized Emden–Fowler equation, given by $y'' + a(t)y^\gamma = 0$ with $a(t) > 0$ [26]. A more ambitious plan is to try and connect our analysis to the recent interest in highly-oscillatory systems, triggered by García-Archilla, Sanz-Serna and Skeel [8], and Hochbruck and Lubich [16].

A delicate issue is the validity region of the estimates derived in this paper. If t is too large the $\mathcal{O}(h^{2p})$ term will probably dominate, rendering the estimates worthless. On the other hand, the estimates of Sections 3, 5, and 6 use the asymptotic solution as $t \to \infty$, so we need t to be sufficiently large. In any case, the step size h needs to be sufficiently small; but if it is very small only the h^p term will contribute. The final thing to keep in mind is that the expansion of the modified equation in powers of h usually diverges. The only definitive statement we can make is that more research needs to be done on this issue.

This paper has concentrated on Runge–Kutta methods. Of course, there are many other methods. If one knows that the solution will be oscillatory, one is probably better off with a method that is specifically designed for this case. The reader is refered to the review paper by Petzold, Jay and Yen [23] for examples. We hope that the framework of Section 2 is sufficiently flexible to include a wide variety of methods. This requires a theory of modified equations for these methods, as Hairer [10] and Faltinsen [7] have done for multistep and Lie-group methods respectively. Another challenge is presented by variable step size methods.

Let us end by comparing the general estimates from Section 2, especially Theorem 2, with some results from the literature. The estimate derived by Iserles [17] can be written as

$$E_h(t) = \frac{1}{h} \int_{t_0}^{t} D\Phi_s^t(y(s))\, L_h\big(s, y(s)\big)\, \mathrm{d}s + \mathcal{O}(h^{p+1}), \tag{31}$$

where $L_h(t, y)$ denotes the local error of the method. Indeed, this estimate follows from (5). We note that (31) gives only the $\mathcal{O}(h^p)$ term of the global error, and we know that this does not always paint an accurate picture. On the other hand, the local error is easier to determine than the modified equation.

Earlier, Viswanath [25] took quite a different approach but derived a bound on the global error which can be brought to almost the same form as (31), namely

$$E_h(t) \leq \int_{t_0}^{t} \|D\Phi_s^t(y(s))\|\, \mathrm{d}s \cdot \frac{1}{h} \max_t |L_h(t, y(t))| + \mathcal{O}(h^{p+1}).$$

This bound severly overestimates the actual error in some circumstances. On the other hand, Viswanath [25] showed that the first factor can be estimated for some large classes of problems, yielding an upper bound for the global error of any reasonable method.

The final note should be on the older but still highly relevant work of Hairer and Lubich [12], who built upon Gragg's asymptotic expansion [9]. They expanded the global error in powers of the step size h, and showed how the terms can be found recursively by solving a differential equation involving the previous term. In principle, all terms of the asymptotic expansion of the global error can be found in this way. Unfortunately, we are not able to solve the differential equations that appear when we apply their approach to the nonlinear oscillator (14). The first p terms of Hairer and Lubich's expansion must obviously equal the integral in (5). However, the connection has not been found yet.

Acknowledgements

This work has greatly benefited from discussions of its author with Chris Budd, Stig Faltinsen, Arieh Iserles, Per Christian Moan, Matthew Piggott, and Divakar Viswanath. The financial support from the EPSRC, Nuffic, VSB Funds, and others is gratefully acknowledged.

References

[1] R. Abraham and J. E. Marsden. *Foundations of Mechanics*. Benjamin/Cummings, Reading, MA, second edition, 1978.

[2] G. Benettin and A. Giorgilli. On the Hamiltonian interpolation of near-to-the identity symplectic mappings with application to symplectic integration algorithms. *J. Stat. Phys.*, 74(5/6):1117–1143, 1994.

[3] M. P. Calvo and E. Hairer. Accurate long-term integration of dynamical systems. *Appl. Numer. Math.*, 18:95–105, 1995.

[4] B. Cano and J. M. Sanz-Serna. Error growth in the numerical integration of periodic orbits, with application to Hamiltonian and reversible systems. *SIAM J. Numer. Anal.*, 34(4):1391–1417, 1997.

[5] B. Cano and J. M. Sanz-Serna. Error growth in the numerical integration of periodic orbits by multistep methods, with application to reversible systems. *IMA J. Numer. Anal.*, 18(1):57–75, 1998.

[6] S. Chandrasekhar. *An Introduction to the Study of Stellar Structure*. University of Chicago Press, 1939.

[7] S. Faltinsen. Backward error analysis for Lie-group methods. *BIT*, 40(4):652–670, 2000.

[8] B. García-Archilla, J. M. Sanz-Serna, and R. D. Skeel. Long-time-step methods for oscillatory differential equations. *SIAM J. Sci. Comput.*, 20(3):930–963, 1999.

[9] W. B. Gragg. *Repeated Extrapolation to the Limit in the Numerical Solution of Ordinary Differential Equations*. PhD thesis, University of California, Los Angeles, 1964.

[10] E. Hairer. Backward error analysis for multistep methods. *Numer. Math.*, 84(2):199–232, 1999.

[11] E. Hairer. Numerical geometric integration. Lecture notes, 1999. Available at http://www.unige.ch/math/folks/hairer/polycop.html.

[12] E. Hairer and C. Lubich. Asymptotic expansions of the global error of fixed-stepsize methods. *Numer. Math.*, 45:345–360, 1984.

[13] E. Hairer and C. Lubich. The life-span of backward error analysis for numerical integrators. *Numer. Math.*, 76:441–462, 1997.

[14] E. Hairer, S. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I. Nonstiff Problems.* Springer-Verlag, Berlin, second edition, 1993.

[15] P. Henrici. *Discrete Variable Methods in Ordinary Differential Equations.* John Wiley & Sons, New York, 1962.

[16] M. Hochbruck and C. Lubich. A Gautschi-type method for oscillatory second-order differential equations. *Numer. Math.*, 83(3):403–426, 1999.

[17] A. Iserles. On the global error of discretization methods for highly-oscillatory ordinary differential equations. To appear in *BIT*, 2001.

[18] A. Iserles and G. Söderlind. Global bounds on numerical error for ordinary differential equations. *J. Complexity*, 9:97–112, 1993.

[19] J. D. Lambert. *Numerical Methods for Ordinary Differential Equations: The Initial Value Problem.* John Wiley & Sons, Chichester, 1991.

[20] E. H. Neville. *Jacobian Elliptic Functions.* Clarendon Press, Oxford, 1944.

[21] F. W. Olver. *Asymptotics and Special Functions.* Academic Press, New York-London, 1974.

[22] B. Orel. Runge–Kutta and Magnus methods for oscillatory ODEs. Talk delivered at SciCADE '01, Vancouver, 2001.

[23] L. R. Petzold, L. O. Jay, and J. Yen. Numerical solution of highly oscillatory ordinary differential equations. *Acta Numerica*, 6:437–483, 1997.

[24] R. D. Skeel. Thirteen ways to estimate global error. *Numer. Math.*, 48(1):1–20, 1986.

[25] D. Viswanath. Global errors of numerical ODE solvers and Lyapunov's theory of stability. *IMA J. Numer. Anal.*, 21(1):387–406, 2001.

[26] J. S. W. Wong. On the generalized Emden–Fowler equation. *SIAM Review*, 17(2):339–360, 1975.

A Some integrals involving elliptic functions

In Section 5, we need the integral of expressions of the form $t^k \operatorname{sd}^\ell(\tilde{t})\,(\operatorname{sd}')^m(\tilde{t})$ for integer values of ℓ and m, where $\tilde{t} = c_1 t^{4/3} + c_2$. These integrals are calculated in this appendix.

First we need some general facts about the elliptic function $\operatorname{sd}(t)$. We refer the reader to Neville [20] for details. The derivatives of $\operatorname{sd}(t)$ satisfy

$$(\operatorname{sd}')^2(t) = 1 - \tfrac{1}{4}\operatorname{sd}^4(t) \quad \text{and} \quad \operatorname{sd}''(t) = -\tfrac{1}{2}\operatorname{sd}^3(t). \tag{32}$$

The latter formula shows that $\operatorname{sd}(t)$ indeed solves the biharmonic oscillator equation (14).

From the periodicity we deduce that (recalling that t_{per} denotes the period)

$$\int_0^t \operatorname{sd}^2(s)\,\mathrm{d}s = \chi t + \mathcal{O}(1), \text{ where } \chi = \frac{1}{t_{\mathrm{per}}}\int_0^{t_{\mathrm{per}}} \operatorname{sd}^2(s)\,\mathrm{d}s \approx 0.91389. \tag{33}$$

We use this and integration by parts (twice) to integrate $t^k \operatorname{sd}^2(\tilde{t})$, with $k \geq \frac{1}{3}$:

$$\int_0^t s^k \operatorname{sd}^2(\tilde{s})\,\mathrm{d}s = \frac{\chi}{k+1} t^{k+1} + \mathcal{O}(t^{k-\frac{1}{3}}), \quad \text{where } \tilde{s} = c_1 s^{4/3} + c_2.$$

Using (32) and again integration by parts, we find the following recurrence relation for $\ell \geq 4$, which allows us to integrate all expressions of the form $t^k \operatorname{sd}^\ell(\tilde{t})$ with ℓ even:

$$\int_0^t s^k \operatorname{sd}^\ell(\tilde{s})\,\mathrm{d}s = \frac{4(\ell-3)}{\ell-1}\int t^k \operatorname{sd}^{\ell-4}(\tilde{t})\,\mathrm{d}t + \mathcal{O}(t^{k-\frac{1}{3}}). \tag{34}$$

On the other hand, we have, using again integration by parts,

$$\int_0^t s^k \operatorname{sd}^\ell(\tilde{s})\ \operatorname{sd}'(\tilde{s})\,\mathrm{d}s = \frac{3}{4(\ell+1)c_1}\left[t^{k-\frac{1}{3}}\operatorname{sd}^{\ell+1}(\tilde{t}) - (k-\tfrac{1}{3})\int t^{k-\frac{4}{3}}\operatorname{sd}^{\ell+1}(\tilde{t})\,\mathrm{d}t\right].$$

Together with (34), this takes care of expressions of the form $s^k \operatorname{sd}^\ell(\tilde{s})\ \operatorname{sd}'(\tilde{s})$. Finally, an integral of the form $\int_0^t s^k \operatorname{sd}^\ell(\tilde{s})\,(\operatorname{sd}')^m(\tilde{s})\,\mathrm{d}s$ with $m \geq 2$ can be reduced to one of these forms by (32). This gives

$$\int_0^t s^k \operatorname{sd}^\ell(\tilde{s})\,(\operatorname{sd}')^m(\tilde{s})\,\mathrm{d}s = \begin{cases} C_{\ell,m}\, t^{k+1} + \mathcal{O}(t^{k-\frac{1}{3}}), & \text{if } \ell, m \text{ are even,} \\ \tilde{C}_{\ell,m}(\tilde{t})\, t^{k-\frac{1}{3}} + \mathcal{O}(t^{k-\frac{5}{3}}), & \text{if } \ell, m \text{ are odd.} \end{cases} \tag{35}$$

Here $C_{\ell,m}$ denotes a constant, and $\tilde{C}_{\ell,m}(\tilde{t})$ denotes a sum of terms of the form $\operatorname{sd}^{\ell'}(\tilde{t})\,(\operatorname{sd}')^{m'}(\tilde{t})$ where ℓ' and m' are integers. So, $\tilde{C}_{\ell,m}(\tilde{t})$ is periodic in $\tilde{t}$.

We do not need to consider the case that ℓ and m have different parity.