

A PRIORI ESTIMATES FOR THE GLOBAL ERROR COMMITTED BY RUNGE–KUTTA METHODS FOR A NONLINEAR OSCILLATOR

JITSE NIESEN

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 nonlinear oscillators whose behaviour is described by the Emden–Fowler equation $y'' + t^{\nu}y'' = 0$. The result shows explicitly that later terms sometimes blow up faster than the leading term of order h^p , necessitating the whole computation. This is supported by numerical experiments.

1. Introduction

A numerical procedure for solving a given ordinary differential equation will induce an error; that is, the numerical solution differs from the exact solution. The goal of this paper is to give *a priori* estimates for this error. This is usually done by expanding the error in powers of the step size. Here, we are especially interested in terms beyond the leading term in this expansion. We shall discuss how to compute these terms, and we shall show by means of an example that these terms may in fact dominate the leading term. This example is the Emden–Fowler equation $y'' + t^{\nu}y'' = 0$. For the parameter range considered here, the solutions of this equation are oscillatory. It is a nonlinear variant of the Airy equation $y'' + ty = 0$, studied by Iserles in [10], which indeed formed the inspiration for the research described in this paper. Other research in this area is described in [2, 3, 6, 7].

Our approach is based on the Alekseev–Gröbner lemma, which describes the effect of perturbing a differential equation on its solution, and the theory of modified equations, which quantifies the idea that the numerical procedure can be considered as a perturbation of the original equation. It should be noted that computing the estimates for the global error derived below requires a good analytical approximation to the exact solution (which is available for the Emden–Fowler equation). In this case, calculating a numerical approximation to the solution might be of limited use. Of course, this does not imply that the work described here is worthless; compare it, for example, with the analysis of the trivial equation $y' = \lambda y$, which leads to the concept of A-stability.

The remainder of this paper is organized as follows. The next section derives a formula for estimating 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 is the order of the method. The Emden–Fowler equation is introduced in Section 3, where we also describe its asymptotic solution. Section 4 contains the actual calculation of global error estimates. The estimates are supported by numerical experiments, as described in Section 5. The last section contains a short discussion of the results.

Received 20 December 2001, revised 25 November 2002; published 17 February 2003.

2000 Mathematics Subject Classification 65L06, 65L70

© 2003, Jitse Niesen

2. Theory

In this section, we fix the notation, and we derive the basic formula (5) for estimating the global error.

Suppose that we are solving the ordinary differential equation

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), \quad \mathbf{y}(t_0) = \mathbf{y}_0, \quad (1)$$

where $\mathbf{y}(t) \in \mathbb{R}^d$ is a vector. We assume that \mathbf{f} is C^∞ to avoid technicalities that are irrelevant to our argument. Associated to this differential equation is the *flow map* $\Phi_s^t : \mathbb{R}^d \rightarrow \mathbb{R}^d$, defined by $\Phi_t^t(\mathbf{y}) = \mathbf{y}$ and $d/dt \Phi_s^t(\mathbf{y}) = \mathbf{f}(t, \Phi_s^t(\mathbf{y}))$. Its Jacobian matrix will be denoted by $D\Phi_s^t$, and is called the *variational flow*.

Now suppose that one uses a numerical one-step method with fixed step size h to solve the differential equation (1). Letting $\mathbf{y}_0 = \mathbf{y}(t_0)$, we denote the subsequent results of the method by $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \dots$, and we set $t_k = t_0 + kh$. We define the *global error* by $\mathbf{E}_h(t_k) = \mathbf{y}_k - \mathbf{y}(t_k)$. We say that a method is *of order* p if $\mathbf{E}_h(t) = \mathcal{O}(h^p)$.

As mentioned in the introduction, the general idea is to view the numerical ‘flow’ as a perturbation of the actual flow of the equation. This idea is made exact by the theory of *modified equations*. The same approach has been taken by Calvo and Hairer [2], and by Hairer and Lubich [7].

The theory of modified equations, analogous to Wilkinson’s backward error analysis in numerical linear algebra, tries to find a differential equation $\mathbf{z}' = \hat{\mathbf{f}}_h(t, \mathbf{z})$ whose solution is close to the numerical results. In fact, for any integer N , one can explicitly construct an $\hat{\mathbf{f}}_h$ such that the solution of the modified equation is $\mathcal{O}(h^N)$ -close to the numerical solution (see, for example, [1] or [8]).

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

$$\mathbf{E}_h(t) = \int_{t_0}^t D\Phi_s^t(\mathbf{y}(s)) \delta_h(s, \mathbf{y}(s)) ds + \mathcal{O}(h^{2p}), \quad (2)$$

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

Proof. The Alekseev–Gröbner lemma (see, for example, [9]) shows that

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

This can be proven immediately by differentiating both sides of the above equation. Now, the left-hand side is $\mathbf{E}_h(t) + \mathcal{O}(h^{2p})$. Furthermore, since the method is of order p , we have $\mathbf{y}(t) - \mathbf{z}(t) = \mathcal{O}(h^p)$. Finally, it follows from the theory of modified equations that $\delta_h(t, \mathbf{y}) = \mathcal{O}(h^p)$ (see, for example, [8, Section IX.1]). Together, these prove (2). \square

Note that the constant implied by the $\mathcal{O}(h^{2p})$ term in equation (2) depends on t . Hence the above result is valid only on bounded time intervals. The same goes for the results that we shall obtain later (Theorems 2 and 3).

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 [9] and [11], amongst others.

A B-series is a formal power series of the form

$$B(a, \mathbf{y}) = a(\emptyset)\mathbf{y} + \sum_{\tau \in \mathbb{T}} \frac{h^{\rho(\tau)}}{\rho(\tau)!} \alpha(\tau) a(\tau) \mathbf{F}(\tau)(\mathbf{y}). \quad (3)$$

Here \mathbb{T} denotes the set of rooted trees, $a : \mathbb{T} \rightarrow \mathbb{R}$ is a coefficient function, $\rho(\tau)$ is the order of the tree τ , $\alpha(\tau)$ denotes the number of monotone labellings of τ , and $\mathbf{F}(\tau)(\mathbf{y})$ is the elementary differential of the function \mathbf{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, Lubich and Wanner [8] use the concept of a B-series to derive a formula for the modified equation. Suppose that the outcome of the RK-method is described by the B-series with coefficient function a . Then the modified equation is given by $\mathbf{z}' = (1/h)B(b, \mathbf{z})$, where the coefficients are recursively defined by:

$$\begin{aligned} b(\emptyset) &= 0; \\ b(\bullet) &= 1; \\ b(\tau) &= a(\tau) - \sum_{j=2}^{\rho(\tau)} \frac{1}{j!} \partial_b^{j-1} b(\tau), \quad \text{for } \rho(\tau) \geq 2. \end{aligned} \quad (4)$$

In this formula, $\partial_b c$ refers to the Lie derivative of $B(c, \mathbf{y})$ with respect to the vector field $B(b, \mathbf{y})$, expressed as a B-series. An explicit formula is given in [8].

We now use this expression for the modified equation to rewrite the *a priori* estimate given in Theorem 1.

THEOREM 2. *Let $\mathbf{y}(t)$ be the solution of the differential equation $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$, 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 (4). If the numerical method has order p , then the global error satisfies:*

$$\mathbf{E}_h(t) = \sum_{\tau \in \mathbb{T}_{[2,2p]}} h^{\rho(\tau)-1} b(\tau) \frac{\alpha(\tau)}{\rho(\tau)!} I_\tau(t) + \mathcal{O}(h^{2p}),$$

where $I_\tau(t) = \int_{t_0}^t D\Phi_s^t(\mathbf{y}(s)) \mathbf{F}(\tau)(\mathbf{y}(s)) ds. \quad (5)$

Here, $\mathbb{T}_{[2,2p]}$ denotes the set of trees with order at least 2, and at most $2p$.

Proof. As mentioned above, the right-hand side of the modified equation is $(1/h)B(b, \mathbf{z})$, with b as in (4). Now $\mathbf{F}(\bullet) = \mathbf{f}$, so the first term in this B-series is the right-hand side of the original equation (1). Hence their difference $\delta_h(t, \mathbf{y})$ equals

$$\sum_{\tau \in \mathbb{T}_{[2,\infty)}} \frac{h^{\rho(\tau)-1}}{\rho(\tau)!} \alpha(\tau) b(\tau) \mathbf{F}(\tau)(\mathbf{y}).$$

We now substitute this expression in (2) 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 (5). \square

Note that the error estimate (5) nicely separates the problem and the method. The method enters the estimate only via the coefficients $b(\tau)$. On the other hand, the value of the integrals I_τ is completely determined by the particular equation that one is solving. As their role in the global error estimate (5) is similar to the role of the elementary differential $\mathbf{F}(\tau)(\mathbf{y})$ in the local error, we shall call I_τ the *elementary integral* associated with τ .

3. The Emden–Fowler equation

For the rest of the paper, we consider nonautonomous, nonlinear oscillators whose behaviour is described by the Emden–Fowler equation:

$$\begin{aligned} y'' + t^\nu y^n &= 0, & t \geq 0; \\ y(0) &= y_0; \\ y'(0) &= y'_0. \end{aligned} \tag{6}$$

This equation with $\nu = 1 - n$ (in which case it is commonly called the Lane–Emden equation) was originally proposed by Chandrasekhar [4] to model stars: considering a star as a radially symmetric gaseous polytrope of index n in thermodynamic and hydrostatic equilibrium, the relation between the density and the distance to the centre satisfies the Lane–Emden equation. The Emden–Fowler equation also arises in the fields of gas dynamics, fluid mechanics, relativistic mechanics, nuclear physics, and the study of chemically reacting systems (see [13] and the references therein). Note that the choice of $\nu = n = 1$ reduces the Emden–Fowler equation to the Airy equation $y'' + ty = 0$, studied by Iserles [10].

From now on, we shall assume that n is an odd integer above 1, and that $\nu > -\frac{1}{2}(n + 3)$. These conditions ensure that oscillatory solutions exists, as explained in the survey by Wong [13]. We remark incidentally that this remains true for any real $n > 1$, if we replace the equation (6) by $y'' + t^\nu \operatorname{sgn}(y) |y|^n = 0$, where $\operatorname{sgn}(y)$ denotes the sign of y . However, we then lose analyticity at $y = 0$.

The equation (6) has an obvious scaling symmetry, which can be used to reduce the order. Here we shall use a different approach, though, because the asymptotic solution as $t \rightarrow \infty$ will suffice for our purposes. From now on, we set

$$\gamma = \frac{\nu}{n + 3}.$$

The conditions on n and ν imply that $\gamma > -\frac{1}{2}$. Now consider the transformation given by

$$y(t) = (1 + 2\gamma)^{2/(n-1)} t^{-\gamma} u(t^{1+2\gamma}). \tag{7}$$

If we apply (7) to the differential equation (6), we get

$$\begin{aligned} (1 + 2\gamma)^{2n/(n-1)} t^{3\gamma} (u''(t^{1+2\gamma}) + u^n(t^{1+2\gamma})) + \gamma(1 + \gamma)(1 + 2\gamma)^{2/(n-1)} t^{-\gamma-2} u(t^{1+2\gamma}) \\ = 0. \end{aligned}$$

For large t , we can (hopefully) neglect the last term as $\gamma > -\frac{1}{2}$, and the above equation reduces to $u'' + u^n = 0$. The expression $(1/(n + 1))u^{n+1} + \frac{1}{2}(u')^2$ is an invariant of this equation; it is only an adiabatic invariant of the original equation. We shall denote the solution of $u'' + u^n = 0$ which satisfies the initial conditions $u(0) = 0$ and $u'(0) = 1$ by $w_n(t)$, and we note for further reference that this is an odd and periodic function. The general solution of $u'' + u^n = 0$ is then given by $u(t) = c_1^{2/(n+1)} w_n(c_1 t + c_2)$. Note that c_1 determines the amplitude of the oscillation, while c_2 determines the phase. In other, more sophisticated words, (c_1, c_2) are action-angle coordinates of the Hamiltonian system corresponding to the Emden–Fowler oscillator.

It follows that the solution of the Emden–Fowler equation (6) is asymptotically (as $t \rightarrow \infty$) given by

$$y(t) \approx (1 + 2\gamma)^{2/(n-1)} c_1^{2/(n-1)} t^{-\gamma} w_n(c_1 t^{1+2\gamma} + c_2). \tag{8}$$

We repeat our assumptions: $n > 1$ is an odd integer and $\nu > -\frac{1}{2}(n + 3)$.

4. Global error of Runge–Kutta methods

In this section, we use Theorem 2 to estimate the global error committed by RK-methods. We shall assume that the asymptotic solution (8) is in fact exact. The numerical experiments reported in Section 5 will show that this is a valid approximation.

The first task is the computation of elementary differentials. For this, we need to convert the differential equation to a system of autonomous first-order equations:

$$\begin{aligned} y_1' &= y_2; \\ y_2' &= -y_3^\nu y_1^n; \\ y_3' &= 1. \end{aligned} \quad (9)$$

Here y_1 , y_2 , and y_3 correspond to y , y' , and t , respectively, in the original equation (6).

It follows that the first component of the elementary differential $\mathbf{F}(\tau)(\mathbf{y})$ satisfies the following recurrence relations (where the argument \mathbf{y} is deleted):

$$\begin{aligned} F_1(\bullet) &= y_2; \\ F_1(\downarrow) &= F_2(\tau); \\ F_1(\downarrow \begin{array}{c} \tau_1 \dots \tau_k \\ \bullet \end{array}) &= 0 \quad (k \geq 2). \end{aligned}$$

The second component satisfies

$$\begin{aligned} F_2(\bullet) &= -y_1^n y_3^\nu; \\ F_2(\downarrow \begin{array}{c} \tau_1 \dots \tau_k \\ \bullet \end{array}) &= -\frac{n!}{(n-k)!} y_1^{n-k} y_3^\nu F_1(\tau_1) \dots F_1(\tau_k) \quad (\text{if } k \leq n); \\ F_2(\downarrow \begin{array}{c} \tau_1 \dots \tau_k \\ \bullet \end{array}) &= 0 \quad (\text{if } k \geq n + 1), \end{aligned}$$

where the derivatives with respect to y_3 are neglected, because they lead to terms that are smaller by a factor $t^{1+2\gamma}$. Finally, the third component F_3 always vanishes, except for $F_3(\bullet)$, which equals one. Hence we shall drop this component from now on.

By unwinding these recurrence relations, we find that the elementary differentials are given by:

$$\begin{aligned} F_1(\tau)(\mathbf{y}) &= C_{1,\tau} y_1^{(n+1)d-\rho+1} y_2^{\rho-2d} y_3^{d\nu}; \\ F_2(\tau)(\mathbf{y}) &= C_{2,\tau} y_1^{n\rho-(n+1)d} y_2^{2d-\rho+1} y_3^{(\rho-d)\nu}. \end{aligned}$$

Here ρ denotes the order of the tree τ (that is, the number of vertices), and d is the number of vertices with odd height (the height of a vertex is the distance to the root). Note that the constants $C_{1,\tau}$ and $C_{2,\tau}$ may be zero. In fact, the only trees for which both constants are nonzero are the trees without branches.

If we substitute the approximate solution (8), we find that the elementary differentials are

$$\mathbf{F}(\tau)(\mathbf{y}(t)) \approx \begin{bmatrix} C_{3,\tau} t^{\gamma(2\rho-1)} w_n^{(n+1)d-\rho+1}(\tilde{t}) w_n^{\rho-2d}(\tilde{t}) \\ C_{4,\tau} t^{\gamma(2\rho+1)} w_n^{n\rho-(n+1)d}(\tilde{t}) w_n^{2d-\rho+1}(\tilde{t}) \end{bmatrix}, \quad (10)$$

where $\tilde{t} = c_1 t^{1+2\gamma} + c_2$. The growth rate of the elementary differential is determined by the exponent of t . Note that the variable d does not enter in this exponent. The surprising

conclusion is that all trees of the same order contribute a term with the same growth rate, independent of their shape. This is in stark contrast to the linear Airy equation, where the differential corresponding to the branchless tree dominates (see [10]).

The next step is to calculate the elementary integral I_τ . 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 $X_t : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, defined by

$$X_t(\mathbf{c}) = \begin{bmatrix} (1 + 2\gamma)^{2/(n-1)} c_1^{2/(n-1)} t^{-\gamma} w_n(c_1 t^{1+2\gamma} + c_2) \\ (1 + 2\gamma)^{1+2/(n-1)} c_1^{1+2/(n-1)} t^\gamma w'_n(c_1 t^{1+2\gamma} + c_2) \end{bmatrix}. \quad (11)$$

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 (5) as:

$$I_\tau(t) \approx DX_t(\mathbf{c}) \int_0^t DX_s^{-1}(\mathbf{y}(s)) \mathbf{F}(\tau)(\mathbf{y}(s)) ds. \quad (12)$$

To find the integrand in the above expression, we multiply the inverse of the Jacobian matrix of (11) with (10). The result is

$$\begin{bmatrix} s^{2\gamma\rho} \left(C_{5,\tau} w_n^{(n+1)(d+1)-\rho} (w'_n)^{\rho-2d} + C_{6,\tau} w_n^{n\rho-(n+1)d} (w'_n)^{2d-\rho+2} \right) \\ s^{2\gamma\rho+2\gamma+1} \left(C_{7,\tau} w_n^{(n+1)(d+1)-\rho} (w'_n)^{\rho-2d} + C_{8,\tau} w_n^{n\rho-(n+1)d} (w'_n)^{2d-\rho+2} \right) \end{bmatrix}, \quad (13)$$

where the functions w_n and w'_n are evaluated at $\tilde{s} = c_1 s^{1+2\gamma} + c_2$.

In the calculation, we used the fact that

$$w''_n(t) = -w_n^n(t) \quad \text{and} \quad w_n'^2(t) = 1 - \frac{1}{n+1} w_n^{n+1}(t).$$

The first equality is indeed the definition of w_n , and the second one follows by multiplying the first one by $w'_n(t)$ and integrating, using the initial conditions $w_n(0) = 0$ and $w'_n(0) = 1$.

The next step is to integrate (13). But consider the exponents of w_n and w'_n . If ρ is even, these exponents are also even, and hence the integrand is nonnegative. Now recall that w_n is odd and periodic. So in the case when ρ is odd, the functions w_n and w'_n 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 ρ is even. We stress that this phenomenon does not occur in the linear case, analysed by Iserles in [10].

In fact, we have

$$\int_0^t w_n^\ell(s) w_n'^m(s) ds = \begin{cases} \tilde{C}_{\ell mn}(t), & \text{if either } \ell \text{ or } m \text{ is odd,} \\ C_{\ell mn} t + \tilde{C}_{\ell mn}(t), & \text{if both } \ell \text{ and } m \text{ are even,} \end{cases}$$

where $\tilde{C}_{\ell mn}(t)$ denotes an oscillatory function with the same period as $w_n(t)$, and $C_{\ell mn}$ is a constant. It follows that

$$\int_0^t s^k w_n^\ell(\tilde{s}) w_n'^m(\tilde{s}) ds = \begin{cases} \tilde{C}_{k\ell mn}(\tilde{t}) t^{k-2\gamma} + \mathcal{O}(t^{k-4\gamma-1}), & \text{if } \ell \text{ or } m \text{ is odd,} \\ C_{k\ell mn} t^{k+1} + \mathcal{O}(t^{k-2\gamma}), & \text{if } \ell \text{ and } m \text{ are even,} \end{cases}$$

where again $\tilde{C}_{k\ell mn}$ and $C_{k\ell mn}$ denote a periodic function and a constant, respectively.

We can use this result to integrate (13), and we find that

$$\int_0^t DX_s^{-1} \mathbf{F}(\tau)(\mathbf{y}) ds = \begin{cases} \begin{bmatrix} \tilde{C}_{9,\tau}(\tilde{t}) t^{2\gamma\rho-2\gamma} + \mathcal{O}(t^{2\gamma\rho-4\gamma-1}) \\ \tilde{C}_{10,\tau}(\tilde{t}) t^{2\gamma\rho+1} + \mathcal{O}(t^{2\gamma\rho-2\gamma}) \end{bmatrix}, & \text{if } \rho \text{ is odd,} \\ \begin{bmatrix} C_{9,\tau} t^{2\gamma\rho+1} + \mathcal{O}(t^{2\gamma\rho-2\gamma}) \\ C_{10,\tau} t^{2\gamma\rho+2\gamma+2} + \mathcal{O}(t^{2\gamma\rho+1}) \end{bmatrix}, & \text{if } \rho \text{ is even.} \end{cases} \quad (14)$$

To compute the elementary integral I_τ , we need to premultiply the integral (14) with DX_t ; see (12). But the expression (14) has an interpretation by itself. Remember that X_t maps the parameter space to the solution space. So the integral (14) represents the error in the parameter space. We conclude that the energy error associated with the tree τ grows as $t^{2\gamma\rho+1}$ if ρ is even, and as $t^{2\gamma\rho-2\gamma}$ if ρ is odd. The second component gives the phase error.

Multiplying the Jacobian matrix of the map X_t with the integral (14) gives us the elementary integrals

$$I_\tau(t) = \begin{cases} \begin{bmatrix} \tilde{C}_{11,\tau}(\tilde{t}) t^{2\gamma\rho-\gamma+1} + \mathcal{O}(t^{2\gamma\rho-3\gamma}) \\ \tilde{C}_{12,\tau}(\tilde{t}) t^{2\gamma\rho+\gamma+1} + \mathcal{O}(t^{2\gamma\rho-\gamma}) \end{bmatrix}, & \text{if } \rho \text{ is odd,} \\ \begin{bmatrix} \tilde{C}_{11,\tau}(\tilde{t}) t^{2\gamma\rho+\gamma+2} + \mathcal{O}(t^{2\gamma\rho-\gamma+1}) \\ \tilde{C}_{12,\tau}(\tilde{t}) t^{2\gamma\rho+3\gamma+2} + \mathcal{O}(t^{2\gamma\rho+\gamma+1}) \end{bmatrix}, & \text{if } \rho \text{ is even.} \end{cases}$$

Finally, we can find an estimate for the global error by adding the contributions of all the trees according to (5). For the first component, we find that

$$E_h(t) \approx \tilde{C}_1(\tilde{t}) h t^{5\gamma+2} + \tilde{C}_2(\tilde{t}) h^2 t^{5\gamma+1} + \tilde{C}_3(\tilde{t}) h^3 t^{9\gamma+2} + \tilde{C}_4(\tilde{t}) h^4 t^{9\gamma+1} + \dots + \mathcal{O}(h^{2p}).$$

More formally, we have the following result.

THEOREM 3. *Suppose that one is solving the Emden–Fowler equation $y'' + t^\nu y^n = 0$, with $\nu > -\frac{1}{2}(n+3)$ and n an odd integer greater than 1, with a numerical method of order p that can be expressed as a B-series. Then the global error has the form*

$$\mathbf{E}_h(t) = \sum_{p \leq 2r < 2p} h^{2r} \begin{bmatrix} \tilde{C}_{2r}^1(\tilde{t}) t^{4\gamma r + \gamma + 1} + \mathcal{O}(t^{4\gamma r - \gamma}) \\ \tilde{C}_{2r}^2(\tilde{t}) t^{4\gamma r + 3\gamma + 1} + \mathcal{O}(t^{4\gamma r + \gamma}) \end{bmatrix} + \sum_{p \leq 2r+1 < 2p} h^{2r+1} \begin{bmatrix} \tilde{C}_{2r+1}^1(\tilde{t}) t^{4\gamma r + 5\gamma + 2} + \mathcal{O}(t^{4\gamma r + 3\gamma + 1}) \\ \tilde{C}_{2r+1}^2(\tilde{t}) t^{4\gamma r + 7\gamma + 2} + \mathcal{O}(t^{4\gamma r + 5\gamma + 1}) \end{bmatrix} + \mathcal{O}(h^{2p}).$$

Here, $\tilde{C}_k^i(\tilde{t})$ denotes a function periodic in $\tilde{t} = c_1 t^{4/3} + c_2$, and $\gamma = \nu/(n+3)$. The remainder term $\mathcal{O}(h^{2p})$ is not uniform in t . \square

We would like to draw the reader's attention again to the difference between the even and the odd powers of h . Unfortunately, we do not know what causes this phenomenon, nor whether it also occurs for other differential equations.

5. 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 (8). Furthermore, it could be interesting to study whether

the $\mathcal{O}(h^p)$ term of the global error $\mathbf{E}_h(t)$ dominates, or whether other terms have to be taken into account.

All the experiments are performed with the parameters $n = 3$ and $\nu = 1$, so the differential equation that we are solving is

$$y'' + ty^3 = 0. \quad (15)$$

The reason for this particular choice is that the function $w_3(t)$, the solution of $u'' + u^3 = 0$ satisfying $u(0) = 0$ and $u'(0) = 1$ (see Section 3), can be expressed in terms of Jacobi elliptic functions (see, for example, [12]). In fact, we have $w_3(t) = \text{sd}(t \mid \frac{1}{2})$. The parameter $\frac{1}{2}$ will be dropped from now on. As a consequence, we can explicitly calculate the elementary integral associated with any given tree.

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

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}\left(t_n + \frac{1}{2}h, \mathbf{y}_n + \frac{1}{2}h\mathbf{f}(t_n, \mathbf{y}_n)\right). \quad (16)$$

To evaluate the error estimate (5), we need to express Runge's method as a B-series, and to compute the coefficients of the modified equation using (4). Next, we evaluate the elementary integrals I_τ and plug all this into the expression (5). After some laborious calculations, this yields

$$\mathbf{E}_h(t) \approx h^2 \begin{bmatrix} \frac{4}{15}\sqrt{2}c_1^4\chi t^{11/6}\text{sd}'(\tilde{t}) \\ -\frac{8}{45}\sqrt{2}c_1^5\chi t^{13/6}\text{sd}^3(\tilde{t}) \end{bmatrix} + h^3 \begin{bmatrix} \frac{256}{6237}\sqrt{2}c_1^6t^{7/2}\text{sd}'(\tilde{t}) \\ -\frac{512}{18711}\sqrt{2}c_1^7t^{23/6}\text{sd}^3(\tilde{t}) \end{bmatrix}. \quad (17)$$

Here, $\chi = (1/4K) \int_0^{4K} \text{sd}^2(s) ds = 0.91389$, where $4K$ is the period of the function sd , and $\tilde{t} = c_1 t^{4/3} + c_2$.

To check this estimate, we compute the solution of the nonlinear oscillator (15) with Runge's second-order method (16). 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 3, this would give an error of about 10^{-9} , so we can consider this to be the exact solution. The global error $\mathbf{E}_h(t)$ is computed by subtracting the result of Runge's method from the 'exact' solution. The first component is depicted in the top three rows of Figure 1. The left-hand column shows the time interval $[0, 50]$, and on the right the larger interval $[0, 2000]$ is displayed.

In the left-hand column of Figure 1, one can see that the global error $\mathbf{E}_h(t)$ oscillates, as is predicted by the estimate (17); this is the factor $\text{sd}'(\tilde{t})$. The amplitude of the oscillations, as predicted by (17), is shown by the thick curve in the left-hand column in Figure 1. We see that the estimate (17) describes the actual error accurately.

The right-hand column of Figure 1 shows a much larger time interval. Here the oscillations of the error are compressed so heavily that it appears as a grey blob. The dashed curve shows the first, leading term of the error estimate (17), 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 (17) 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, which is given by equation (18).

$$\begin{aligned} \mathbf{k}_1 &= \mathbf{f}(t_n, \mathbf{y}_n) \\ \mathbf{k}_2 &= \mathbf{f}\left(t_n + \frac{1}{3}h, \mathbf{y}_n + \frac{1}{3}h\mathbf{k}_1\right) \\ \mathbf{k}_3 &= \mathbf{f}\left(t_n + \frac{2}{3}h, \mathbf{y}_n + \frac{1}{3}h\mathbf{k}_2\right) \\ \mathbf{y}_{n+1} &= \mathbf{y}_n + h\left(\frac{1}{3}\mathbf{k}_1 + \frac{2}{3}\mathbf{k}_3\right) \end{aligned}$$

Butcher tableau:

$$\begin{array}{c|ccc} 0 & & & \\ \hline 1/3 & 1/3 & & \\ 2/3 & 0 & 2/3 & \\ \hline & 1/4 & 0 & 3/4 \end{array} \quad (18)$$

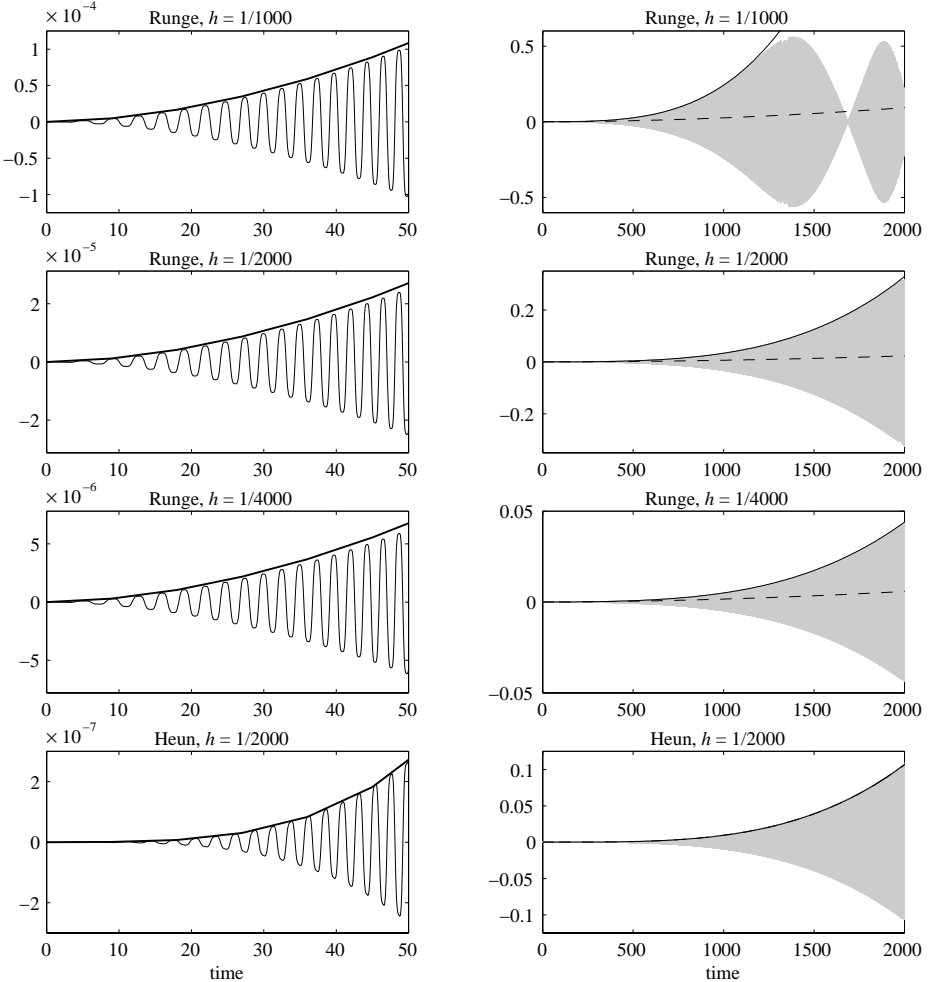


Figure 1: The top three rows show the first component of the global error committed by Runge's second-order method (16), and the estimate (17), for various step sizes. On the left, the oscillating curve is the global error $E_h(t)$, and the thick curve shows the amplitude of the oscillations as predicted by (17). On the right, the true error is shown in grey, the dashed curve shows the first, leading term of the error estimate (17), and the solid curve shows the sum of both terms. The bottom row shows the same for Heun's third-order method (18) and the estimate (19).

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

$$\mathbf{E}_h(t) \approx h^3 \begin{bmatrix} -\frac{512}{35721} \sqrt{2} c_1^6 t^{7/2} \text{sd}'(\tilde{t}) \\ \frac{1024}{107163} \sqrt{2} c_1^7 t^{23/6} \text{sd}^3(\tilde{t}) \end{bmatrix} + h^4 \begin{bmatrix} \frac{50208}{229635} \sqrt{2} c_1^6 t^{5/2} \text{sd}'(\tilde{t}) \\ -\frac{60416}{688905} \sqrt{2} c_1^7 t^{17/6} \text{sd}^3(\tilde{t}) \end{bmatrix} + h^5 \begin{bmatrix} \frac{557056}{34543665} \sqrt{2} c_1^8 \chi t^{25/6} \text{sd}'(\tilde{t}) \\ -\frac{1114112}{103630995} \sqrt{2} c_1^9 \chi t^{9/2} \text{sd}^3(\tilde{t}) \end{bmatrix}. \quad (19)$$

The actual error and the above estimate, for step size $h = 1/2000$, are displayed in the bottom row of Figure 1. We see that the error estimate (19) 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 1). For large values of t , however (see the right-hand column), Runge’s method behaves essentially as a third-order method, and we see that the difference between the two methods is indeed much smaller.

Remark. It turns out that the elementary integrals I_τ corresponding to trees of the same order are scalar multiples of each other. Following an idea put forward by B. Orelin a talk given in 2001 at the SciCADE event in Vancouver, we can use this to construct a Runge–Kutta method in which the dominating term of the global error vanishes. This yields a method which, for this particular equation, clearly outperforms other RK-methods of the same order: the error is several orders of magnitude lower. A deeper investigation into the mechanism behind this phenomenon might be useful.

6. Discussion

We have 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 Runge–Kutta methods for a class of nonlinear oscillators. A special feature is that we obtained some terms after the leading term of order h^p . The Emden–Fowler equation showed that this is sometimes necessary, as the later terms blow up faster than the leading term. Finally, some numerical experiments have been done, and have verified the accuracy of the estimates.

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 4 and 5 use the asymptotic solution as $t \rightarrow \infty$, 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 that we can make is that more research needs to be done on this issue.

The approach taken in this paper can be compared to that described by Hairer and Lubich [6], who built upon Gragg’s asymptotic expansion [5]. They expand the global error in powers of the step size h , and show how the terms can be found recursively by solving a differential equation involving the previous term. In principle, all the terms of the asymptotic expansion of the global error can be found in this way. Unfortunately, when we apply this approach to the Emden–Fowler oscillator (6), we find that the expressions quickly become too complicated to handle. Obviously, the first p terms of Hairer and Lubich’s expansion must equal the integral in (2). However, the connection has not, as yet, been found.

Acknowledgements. This work has greatly benefited from the author's discussions with Chris Budd, Stig Faltinsen, Arieh Iserles, Per Christian Moan, Matthew Piggott, and Divakar Viswanath, and from the remarks made by the referees. The financial support received from the EPSRC, Nuffic, VSB Funds, and others is gratefully acknowledged.

References

1. G. BENETTIN and A. GIORGILLI, 'On the Hamiltonian interpolation of near-to-the identity symplectic mappings with application to symplectic integration algorithms', *J. Statist. Phys.* 74 (1994) 1117–1143. 19
2. M. P. CALVO and E. HAIRER, 'Accurate long-term integration of dynamical systems', *Appl. Numer. Math.* 18 (1995) 95–105. 18, 19
3. 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 (1998) 57–75. 18
4. S. CHANDRASEKHAR, *An introduction to the study of stellar structure* (University of Chicago Press, 1939). 21
5. W. GRAGG, 'Repeated extrapolation to the limit in the numerical solution of ordinary differential equations', Ph.D. thesis, University of California, Los Angeles, 1964. 27
6. E. HAIRER and CH. LUBICH, 'Asymptotic expansions of the global error of fixed-size methods', *Numer. Math.* 45 (1984) 345–360. 18, 27
7. E. HAIRER and CH. LUBICH, 'Asymptotic expansions and backward analysis for numerical integrators', *Dynamics of algorithms (Minneapolis, MN, 1997)*, IMA Vol. Math. Appl. 118 (ed. R. de la Llave, L. R. Petzold and J. Lorenz, Springer, New York, 2000) 91–106. 18, 19
8. E. HAIRER, CH. LUBICH and G. WANNER, *Geometric numerical integration. Structure-preserving algorithms for ordinary differential equations*, Springer Ser. Comput. Math. 31 (Springer, Berlin, 2002). 19, 19, 20, 20
9. E. HAIRER, S. NØRSETT and G. WANNER, *Solving ordinary differential equations I. Nonstiff problems*, 2nd edn (Springer, Berlin, 1993). 19, 19
10. A. ISERLES, 'On the global error of discretization methods for highly-oscillatory ordinary differential equations', *BIT* 42 (2002) 561–599. 18, 21, 23, 23
11. J. LAMBERT, *Numerical methods for ordinary differential equations: the initial value problem* (John Wiley & Sons, Chichester, 1991). 19
12. E. NEVILLE, *Jacobian elliptic functions* (Clarendon Press, Oxford, 1944). 25
13. J. WONG, 'On the generalized Emden–Fowler equation', *SIAM Rev.* 17 (1975) 339–360. 21, 21

Jitse Niesen J.Niesen@damtp.cam.ac.uk

<http://www.damtp.cam.ac.uk/user/na/people/Jitse/index.html>

Department of Applied Mathematics and Theoretical Physics
University of Cambridge
Wilberforce Road
Cambridge CB3 0WA