Backward error analysis

Brynjulf Owren

July 28, 2015

Introduction. The main source for these notes is the monograph by Hairer, Lubich and Wanner [2]. Consider ODEs in $\mathbb{R}^d$ of the form

$$\dot{y} = f(y), \quad y(0) = y_0 \in \mathbb{R}^d \tag{1}$$

and let Φ_h be an integrator, i.e. $y_{n+1} = \Phi_h(y_n,)\ n = 0, 1, \ldots$.

Forward error analysis. Study the local and global errors. Let φ_t be the exact flow of (1).

$y_1 - \varphi_h(y_0)$	Local error
$y_n - \varphi_{nh}(y_0)$	Global error

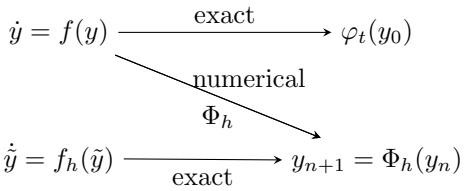
Backward error analysis. Find a modified ODE

$$\dot{\tilde{y}} = f_h(\tilde{y}) = f(\tilde{y}) + hf_2(\tilde{y}) + h^2 f_3(\tilde{y}) + \cdots \tag{2}$$

such that

$$y_n = \tilde{y}(nh)$$

So the idea is that the exact solution of the modified problem (2) equals the numerical solution of the original problem.



A couple of issues to notice at this point are

1. An important idea of backward error analysis is to obtain qualitative insight in the numerical solution. Does the modified equation inherit the structure we had in the original vector field?

2. The series (2) must be understand as a formal series at this point. In fact, the series will only converge in exceptional cases. There is a whole theory about how to truncate such asymptotic series at the optimal point.

Constructing the modified vector field. Formally there is a constructive method for finding the terms of (2). Fix t, and set $y := \tilde{y}(t)$. By Taylor expansion, surpressing the dependence of virtually all symbols on y,

$$\begin{aligned}\tilde{y}(t+h) &= y + hf_h + \frac{1}{2}h^2 f'_h f_h + \frac{1}{6}h^3(f''_h(f_h, f_h) + f'_h f'_h f_h) + \cdots \\ &= y + h(f + hf_2 + h^2 f_3 + \cdots) + \frac{1}{2}h^2(f' + hf'_2 + \cdots)(f + hf_2 + \cdots) \\ &+ \frac{h^3}{6}\big((f'' + \cdots)(f + \cdots, f + \cdots) + (f' + \cdots)(f' + \cdots)(f + \cdots)\big) + \cdots\end{aligned} \tag{3}$$

Now we just need to collect equal powers of h

$$\tilde{y}(t+h) = y + hf + h^2(f_2 + \frac{1}{2}f'f) + h^3(f_3 + \frac{1}{2}f'f_2 + \frac{1}{2}f'_2 f + \frac{1}{6}f''(f,f) + \frac{1}{6}f'f'f) + \cdots \tag{4}$$

Suppose that also the numerical method map $\Phi_h(y)$ has an expansion in powers of h with coefficients $d_k(y) =: d_k$

$$\Phi_h(y) = y + hf + h^2 d_2 + h^3 d_3 + \cdots \tag{5}$$

The functions $d_k(y)$ typically depends on the coefficients of the method, and for one-step methods one can often use the machinery of B-series to find them. Note the particularly simple Euler method, in which

$$d_k = 0, \qquad k \geq 2,$$

but this is a very special case, other Runge-Kutta methods, even if explicit, typical have an infinite series.

We formally set $\Phi_h(y) = \tilde{y}(t+h)$ and make an order by order comparison. This leads to

$$\begin{aligned}f_2(y) &= d_2 - \frac{1}{2}f'f \\ f_3(y) &= d_3 - \frac{1}{6}(f''(f,f) + f'f'f) - \frac{1}{2}(f'f_2 + f'_2 f)\end{aligned}$$

where one can substitute for f_2 in f_3 to obtain an expression depending only on d_2, d_3, f.

Example 1 From [2]. Apply the explicit Euler method to the problem

$$\dot{y} = y^2, \quad y(0) = 1. \quad \text{Exact solution: } y(t) = \frac{1}{1-t}$$

In this case, we get

$$\begin{aligned}f_2(y) &= d_2 - \frac{1}{2}f'f = 0 - \frac{1}{2}(2y)y^2 = -y^3 \\ f_3(y) &= 0 - \frac{1}{6}(2y^4 + (2y)(2y)y^2) - \frac{1}{2}(2y(-y^3) - 3y^2 y^2) = \frac{3}{2}y^4\end{aligned}$$

So this leads to the following first terms in the expansion for the modified equation

$$\dot{\tilde{y}} = \tilde{y}^2 - h\tilde{y}^3 + \frac{3}{2}h^2\tilde{y}^4 + \cdots$$

□

Exercise 1 Lotka-Volterra ([2]). We recall the Lotka-Volterra model used before with the following parameters

$$\dot{q} = q(p-1), \qquad \dot{p} = p(2-q)$$

We are interested in the modified vector field to the first order when applying

a. The explicit Euler method

b. The symplectic Euler method

To solve this, you may first want to verify that for the Lotka-Volterra equations

$$f'f = \begin{pmatrix} q(p^2 - pq + 1) \\ -p(pq - q^2 + 3q - 4) \end{pmatrix}$$

Now you are in a position to write down $f_2(q,p)$ for the Euler method. Show next that when the symplectic Euler method

$$\begin{aligned} q_1 &= q + hf^1(q_1, p) \\ p_1 &= p + hf^2(q_1, p) \end{aligned}$$

is applied to the Lotka-Volterra equations, it results in expressions that can be expanded as

$$\Phi_h(q,p) = \begin{pmatrix} q \\ p \end{pmatrix} + h \begin{pmatrix} q(p-1) \\ p(2-q) \end{pmatrix} + h^2 \begin{pmatrix} q(p-1)^2 \\ -qp(p-1) \end{pmatrix} + h^3 \begin{pmatrix} q(p-1)^3 \\ -qp(p-1)^2 \end{pmatrix} + \cdots$$

Use this to determine the modified equations for the symplectic Euler method to the first order. □

A theorem on convergence order. We consider which effect it has on the modified vector field if the method is of order p.

Theorem 1. *Suppose $y_1 = \Phi_h(y)$ is of order p, such that*

$$\Phi_h(y) = \varphi_h(y) + h^{p+1}\delta_{p+1}(y) + \mathcal{O}(h^{p+2}).$$

Then the modified vector field is of the form

$$f_h(\tilde{y}) = f(\tilde{y}) + h^p f_{p+1}(\tilde{y}) + h^{p+1} f_{p+2}(\tilde{y}) + \cdots$$

where $f_{p+1}(y) = \delta_{p+1}(y)$.

Proof. Looking at the construction of $f_k(y)$, we see that their expressions are of the form

$$f_k(y) = d_k(y) - \frac{1}{k!}\frac{\partial^k}{\partial h^k}\bigg|_{h=0} \varphi_h(y) - R_k(f, f_2, \ldots, f_{k-1})$$

where $R_k(f, 0, \ldots, 0) = 0$. Since for an order p method,

$$d_k(y) = \frac{1}{k!}\frac{\partial^k}{\partial h^k}\bigg|_{h=0} \varphi_h(y) \text{ for } k \le p,$$

it follows by induction that $f_2, \ldots, f_p$ all vanish, whereas

$$f_{p+1}(y) = d_{p+1}(y) - \frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial h^{p+1}}\bigg|_{h=0} \varphi_h(y) = \delta_{p+1}(y)$$

□

Modified vector fields for symplectic integrators. We shall prove the following theorem

Theorem 2. *If a symplectic method is applied to a Hamiltonian system with a smooth Hamiltonian $H : \mathbb{R}^{2d} \to \mathbb{R}$, then the modified equation is also Hamiltonian. There exist smooth functions $H_j : \mathbb{R}^{2d} \to \mathbb{R}$ for $j = 2, 3, \ldots$ such that*

$$f_j(y) = J^{-1}\nabla H_j(y)$$

Proof. (Benettin and Giorgilli 1994 [1], Tang 1994 [3]) We follow here the proof given in [2]. We use induction and assume that

$$f_j(y) = J^{-1}\nabla H_j(y), \quad j = 1, \ldots, r.$$

It clearly holds for $j = 1$ since $f_1(y) = f(y) = J^{-1}\nabla H(y)$. We must establish the existence of H_{r+1}. Consider a truncated version of the modified equation,

$$\dot{\tilde{y}} = f(\tilde{y}) + hf_2(\tilde{y}) + \cdots + h^r f_r(\tilde{y})$$

We already know that this truncated version is a Hamiltonian vector field with

$$H(y) + hH_2(y) + \cdots + h^{r-1}H_{r-1}$$

as Hamiltonian function. Its flow $\varphi_{r,t}$ satisfies

$$\Phi_h(y_0) = \varphi_{r,h}(y_0) + h^{r+1}f_{r+1}(y) + \mathcal{O}(h^{r+2}).$$

This is a direct consequence of Theorem 1. So its Jacobian

$$\Phi_h'(y_0) = \frac{\partial \Phi_h}{\partial y}(y_0) \text{ satisfies}$$

$$\Phi_h'(y_0) = \varphi_{r,h}'(y_0) + h^{r+1}f_{r+1}'(y_0) + \mathcal{O}(h^{r+2})$$

In the next passage we will use the fact that both Φ_h and $\varphi_{r,h}$ are symplectic maps, and also that since φ_h is differentiable in h and $\varphi_0(y) = y$, we have

$$\varphi_h' = I + \mathcal{O}(h)$$

We compute (omitting the argument y_0 for brevity)

$$\begin{aligned} J &= \Phi_h'^T J \Phi_h' \\ &= (\varphi_{r,h}' + h^{r+1} f_{r+1}' + \mathcal{O}(h^{r+2}))^T J (\varphi_{r,h}' + h^{r+1} f_{r+1}' + \mathcal{O}(h^{r+2})) \\ &= J + (I + \mathcal{O}(h))^T J h^{r+1} f_{r+1}' + h^{r+1} f_{r+1}'^T J (I + \mathcal{O}(h)) + \mathcal{O}(h^{r+1}) \\ &= J + h^{r+1}(J f_{r+1}' + f_{r+1}'^T J) + \mathcal{O}(h^{r+2}) \end{aligned}$$

This can only happen if

$$J f_{r+1}' - (J f_{r+1}')^T = 0$$

This means that $J f_{r+1}'$ is symmetric, so that by the integrability lemma (see e.g. [2, Ch. VI.2]), there exists a function we may call H_{r+1} such that

$$J f_{r+1}' = \nabla H_{r+1}$$

as sought. □

Exercise 2 Find the modified equation for the implicit midpoint rule applied to

$$\dot{y} = Ay, \quad A \in \mathbb{R}^{k \times k},\ y(0) = y_0$$

□

Backward error analysis for the adjoint of a method and for symmetric methods. Considering first the adjoint $\Phi_h^* = \Phi_{-h}^{-1}$ of the method Φ_h, we have the following result

Theorem 3. *[2] Suppose the modified equation for the method Φ_h are obtained from the functions f_j, $j = 2, 3, \ldots$. Then the corresponding functions f_j^* of the adjoint Φ_h^* are given by*

$$f_j^*(y) = (-1)^{j+1} f_j(y)$$

Proof. We spell out a few more details than what is given in [2]. First, let us define $\tilde{y}^*(t)$ to be the exact solution of the modified equation for Φ_h^*. By definition this means that

$$\Phi_h^*(\tilde{y}^*(t)) = \tilde{y}^*(t+h) \quad \Rightarrow \quad \tilde{y}^*(t) = \Phi_{-h}(\tilde{y}^*(t+h))$$

using the definition of the adjoint method. But the latter equality holds also if t is replaced by $t - h$ so that

$$\tilde{y}^*(t-h) = \Phi_{-h}(\tilde{y}^*(t)) := \Phi_{-h}(y)$$

Suppose it is already known that $f_j^*(y) = (-1)^{j+1} f_j(y),\ 1 \le j \le r-1$. For $r = 2$ this is true for any consistent method as seen in (2). From (5), we find

$$\Phi_{-h}(y) = y - hf + h^2 d_2 - h^3 d_3 + \cdots$$

In (3) we now get for the modified field $f_h^* = f + hf_2^* + \cdots$

$$\begin{aligned}
&\tilde{y}^*(t-h) = y - hf_h^* + \frac{h^2}{2}(f_h^*)' f_h^* - \frac{h^3}{6}\left((f_h^*)''(f_h^*, f_h^*) + (f_h^*)'(f_h^*)' f_h^*\right) + \cdots \\
&= y - h(f + hf_2^* + h^2 f_3^* + \cdots) \\
&+ \frac{h^2}{2}(f + hf_2^* + h^2 f_3^* + \cdots)'(f + hf_2^* + h^2 f_3^* + \cdots) \\
&- \frac{h^3}{6}(f + hf_2^* + h^2 f_3^* + \cdots)''(f + hf_2^* + h^2 f_3^* + \cdots, f + hf_2^* + h^2 f_3^* + \cdots) \\
&- \frac{h^3}{6}(f + hf_2^* + h^2 f_3^* + \cdots)'(f + hf_2^* + h^2 f_3^* + \cdots)'(f + hf_2^* + h^2 f_3^* + \cdots) \\
&+ \cdots
\end{aligned} \tag{6}$$

We make the following observations in this expression

1. Terms of order h^j may involve $f, f_2^*, \ldots, f_j^*$, but no coefficients with higher index than j.
2. The only term of order h^j which involves f_j^* is $-h^j f_j^*$
3. When we substitute $f_j^* = (-1)^{j+1} f_j,\ j \le r-1$ we see that all factors $h^{2i-1} f_{2i}^*$ change sign and become $-h^{2i-1} f_{2i}$ whereas $h^{2i} f_{2i+1}^*$ remain the same, i.e. $h^{2i} f_{2i+1}$. When we collect equal powers of h in (6) all the odd powers will have changed sign (up to $j = r$), and all the even powers are the same when we compare to (4).
4. All these terms involving $f, f_2, \ldots, f_{r-1}$ of order h^r in (4) are precisely equal to $d_r - f_r$.

In view of these observations, for the h^r-terms collected in (6) we get

$$-f_r^* + (-1)^r (d_r - f_r) = (-1)^r d_r \quad \Rightarrow \quad f_r^* = (-1)^{r+1} f_r$$

□

For symmetric methods we have $\Phi_h^* = \Phi_h$, and the following corollary is available almost for free

Corollary 1. *For symmetric methods $f_j(y) = 0$ when j is even*

References

[1] Giancarlo Benettin and Antonio Giorgilli. On the Hamiltonian interpolation of near-to-the-identity symplectic mappings with application to symplectic integration algorithms. *J. Statist. Phys.*, 74(5-6):1117–1143, 1994.

[2] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration*, volume 31 of *Springer Series in Computational Mathematics*. Springer, Heidelberg, 2010. Structure-preserving algorithms for ordinary differential equations, Reprint of the second (2006) edition.

[3] Y.-F. Tang. Formal energy of a symplectic scheme for Hamiltonian systems and its applications. I. *Comput. Math. Appl.*, 27(7):31–39, 1994.