Convergence of ODE solvers

Bård Skaflestad*
Email: <bardsk@math.ntnu.no>

We study the convergence of numerical integrators for ordinary differential equations (ODEs). The notions of order of convergence and local errors are particularly important.

Introduction

We assume given the (system of) ordinary differential equation(s) and initial value

$$x' = f(t, x), \quad x(t_0) = x_0. \tag{1}$$

In the the following we will also assume that a unique solution $x(t)$ to (1) exists and that the function $f(t, x)$ is "sufficiently smooth." Morever, we will assume that $f(t, x)$ satisfies the 'Lipschitz' condition (is "Lipschitz continuous")

$$|f(t, x) - f(t, \tilde{x})| \leq L|x - \tilde{x}|, \tag{2}$$

for all x and $\tilde{x}$. The constant $L > 0$ does not depend on x and $\tilde{x}$ directly, but the value of L will in general depend on the specifics of the function $f(t, x)$.

Given step size h we seek approximations $\{x_i\}_{i\geq 1}$ to $x(t)$ at the points $\{t_i\}_{i\geq 1}$ according to the rule

$$x_i \approx x(t_i), \quad t_i = t_0 + ih, \quad i = 1, 2, \ldots$$

This process is illustrated in Figure 1 below.

*Slightly adapted from an original manuscript by A. Kværnø.

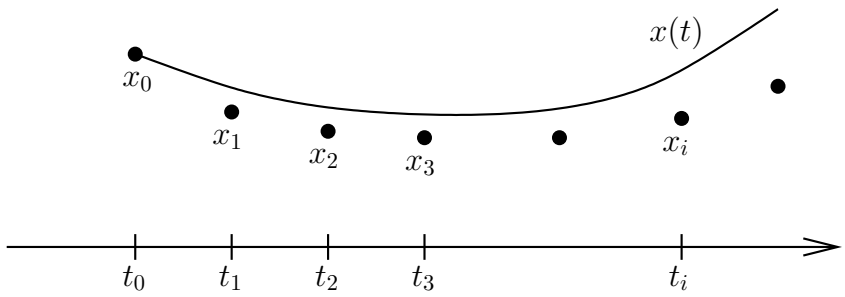

Figure 1: Numerical approximations to solution of ODE (1)

Definitions

Let $T > t_0$ be some fixed point and assume that we use n steps with our method to find and approximation to $x(T)$. The *global error* is the error

$$E_n = x(T) - x_n$$

with step size $h = (T - t_0)/n$. The method is *convergent* if

$$\lim_{n\to\infty} E_n = 0 \quad \text{(equivalently } h \to 0\text{)}$$

for all ODEs (1) satisfying the above assumptions. The method is of *order* p if

$$E_n = \mathcal{O}(h^p).$$

The global error depends on two factors, both of which are depicted in figure 2 below,

- The *local error*, d_i, which is the error made in each step, i.
- The propagation of errors.

Let us illustrate these concepts by the famous Euler method given by

$$x_{i+1} = x_i + hf(t_i, x_i), \quad i = 0, 1, 2, \ldots$$

The local truncation error is determined by comparing the numerical solution to the exact solution after a single step assuming exact initial values, $x_i = x(t_i)$. Thus

$$\begin{aligned} d_{i+1} &= x(t_i + h) - (x_i + hf(t_i, x_i)) \\ &= x(t_i) + hx'(t_i) + \tfrac{1}{2}h^2x''(t_i + \xi h) - x_i - hf(t_i, x_i) \end{aligned}$$

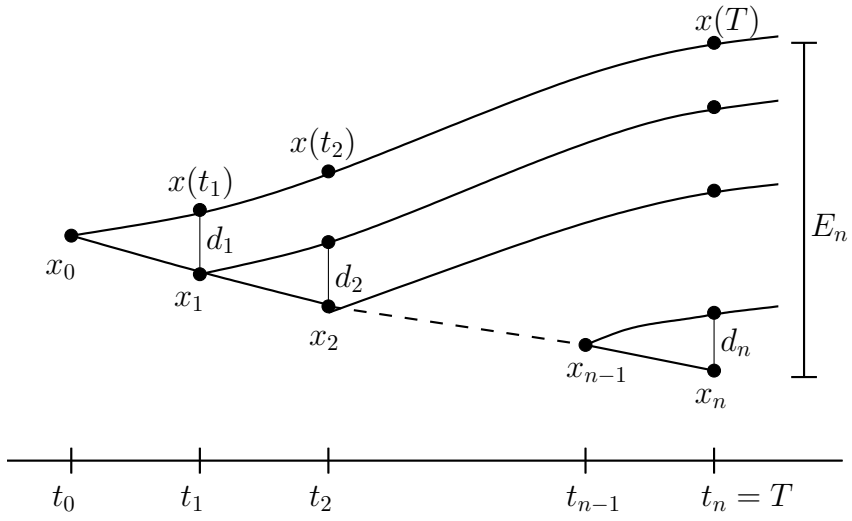

Figure 2: Local truncation errors and error propagation

for some $\xi \in (0,1)$. Using the ODE (1), we then get

$$d_{i+1} = \tfrac{1}{2}h^2 x''(t_i + \xi h)$$

whence, if $x''(t)$ is bounded in the region of interest, there exists a constant $C > 0$ such that

$$|d_i| \le C \cdot h^2 \tag{3}$$

for all $i = 0, 1, \ldots, n-1$.

Convergence of the Euler method

Let $E_i = x(t_i) - x_i$ be the global error after i steps. From the definitions of the local error d_i and the Euler formula, we get

$$\begin{aligned} x(t_i + h) &= x(t_i) + hf(t_i, x(t_i)) + d_i, \\ x_{i+1} &= x_i + hf(t_i, x_i) \end{aligned}$$

from which, after subtracting the latter equation from the former, we obtain

$$E_{i+1} = E_i + h\big(f(t_i, x(t_i)) - f(t_i, x_i)\big) + d_i$$

for all i. The Lipschitz condition (2) and the local error bound (3) then give

$$|E_{i+1}| \le (1 + hL)|E_i| + Ch^2,$$

for all i.

Repeated application of this relation now yields

$$\begin{aligned} |E_1| &\le (1 + hL)|E_0| + Ch^2 \\ |E_2| &\le (1 + hL)|E_1| + Ch^2 = (1 + hL)^2|E_0| + (1 + hL + 1)Ch^2 \\ &\vdots \\ |E_n| &\le (1 + hL)^n|E_0| + \sum_{i=0}^{n-1}(1 + hL)^i Ch^2 \\ &= (1 + hL)^n|E_0| + \frac{(1 + hL)^n - 1}{hL} \cdot Ch^2. \end{aligned}$$

Remember that E_n is the error $x(T) - x_n$ for the case where n steps of step size $h = (T - t_0)/n$ have been used. Also, as $1 + hL \le \mathrm{e}^{hL}$ for all h (since $hL > 0$), including in the limit $h \to 0$ (equivalently $n \to \infty$), we get

$$|E_n| \le \mathrm{e}^{hL \cdot n}|E_0| + \frac{\mathrm{e}^{hL \cdot n} - 1}{L} \cdot Ch = \mathrm{e}^{L(T - t_0)}|E_0| + \frac{\mathrm{e}^{L(T - t_0)} - 1}{L} \cdot Ch$$

The first term gives an upper bound for the propagation of any initial error $E_0 = x(t_0) - x_0$. If this is zero, which we will usually assume, we see that the method is of order 1 and thereby convergent.

In general, one step methods like Runge–Kutta methods can be written in terms of a "step function," $\Phi(t, x; h)$, with h denoting the step size as

$$x_{i+1} = x_i + h\Phi(t_i, x_i; h), \quad i = 0, 1, 2, \ldots$$

The step function $\Phi(t, x; h)$ depends on $f(t, x)$ and the specific numerical method being analysed. In this case the local truncation error is

$$d_i = x(t_{i+1}) - x(t_i) - h\Phi(t_i, x(t_i); h).$$

If Φ satisfies a Lipschitz condition of the form

$$|\Phi(t, x; h) - \Phi(t, \tilde{x}; h)| \le M|x - \tilde{x}|$$

and $|d_i| \le Ch^{p+1}$ for some $p > 0$, then the above argument can be repeated to prove that

$$|E_n| \le \frac{\mathrm{e}^{M(T - t_0)} - 1}{M} \cdot Ch^p$$

provided $E_0 = 0$.

Runge–Kutta methods

An s-stage Runge–Kutta (RK) method is defined as

$$k_i = f(t_0 + c_i h, x_0 + h\sum_{j=1}^{s} a_{ij} k_j), \quad i = 1, 2, \ldots, s$$

$$x_1 = x_0 + h\sum_{i=1}^{s} b_i k_i.$$

Particular methods are fully specified by the coefficients c_i, a_{ij}, and b_i which are often presented in a "Butcher tableau"

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array}.$$

A method is called *explicit* if $a_{ij} = 0$ for all $j \geq i$ (on and above the main diagonal of $A = (a_{ij})$), otherwise the method is called *implicit*.

Some example methods are listed below

- Euler's method

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

- Heun's method

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

- Trapezoidal rule (implicit)

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

- Kutta's classical 4th order method ('RK4')

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

An RK method is of order p, that is

$$|x(t_0 + h) - x_1| \leq Ch^{p+1},$$

if $c_i = \sum_{j=1}^{s} a_{ij}$ for all $i = 1, 2, \ldots, s$, and

$p = 1$

$$\sum_{i=1}^{s} b_i = 1$$

$p = 2$

$$\sum_{i=1}^{s} b_i c_i = \tfrac{1}{2}$$

$p = 3$

$$\sum_{i=1}^{s} b_i c_i^2 = \tfrac{1}{3}, \quad \sum_{i,j=1}^{s} b_i a_{ij} c_j = \tfrac{1}{6}$$

$p = 4$

$$\sum_{i=1}^{s} b_i c_i^3 = \tfrac{1}{4}, \quad \sum_{i,j=1}^{s} b_i c_i a_{ij} c_j = \tfrac{1}{8}$$

$$\sum_{i,j=1}^{s} b_i a_{ij} c_j^2 = \tfrac{1}{12}, \quad \sum_{i,j,k=1}^{s} b_i a_{ij} a_{jk} c_k = \tfrac{1}{24}$$

These "order conditions" constitute (non-linear) constraints on the method coefficients which must be satisfied if the method's local error is to be provably bounded by Ch^{p+1}. For higher order methods (i.e. larger values of p), the number of additional conditions grows quickly as shown in table 1. In other words, a method of order 10 must satisfy a total number of 1205 conditions.

p	5	6	7	8	9	10
add. cond.	9	20	48	115	286	719

Table 1: Number of additional conditions for higher order methods