Numerical Analysis

Daming Li

Department of Mathematics, Shanghai JiaoTong University,
Shanghai, 200240, China
Email: lidaming@sjtu.edu.cn

October 14, 2014

Existence and Uniqueness of Solutions

Consider an initial-value problem

$$\begin{cases} x' = f(t, x) \\ x(t_0) = x_0 \end{cases}$$

Here x is an unknown function of t, where $x' = dx(t)/dt$. For example,

$$\begin{cases} x' = x\tan(t+3) \\ x(-3) = 1 \end{cases}$$

The analytic solution is $x(t) = \sec(t+3)$. Typically, for the above problem, analytic solutions are not available and numerical methods must be employed.

Existence and Uniqueness of Solutions

If f is continuous in a rectangle R centered at (t_0, x_0), say

$$R = \{(t, x) : |t - t_0| \leq \alpha, \quad |x - x_0| \leq \beta\}$$

then the initial-value problem has a solution $x(t)$ for $|t - t_0| \leq \min(\alpha, \beta/M)$, where M is the maximum of $|f(t, x)|$ in the rectangle R.

Existence and Uniqueness of Solutions

Prove that the initial-value problem

$$\begin{cases} x' = (t + \sin x)^2 \\ x(0) = 3 \end{cases}$$

has a solution on the interval $-1 \le t \le 1$. Taking $f(t,x) = (t + \sin x)^2$ and $(t_0, x_0) = (0, 3)$. The rectangle is

$$R = \{(t,x) : |t| \le \alpha, \quad |x - 3| \le \beta\}$$

The magnitude of f is bounded by

$$|f(t,x)| \le (\alpha + 1)^2 \equiv M$$

We want $\min(\alpha, \beta/M) \ge 1$, and so we can let $\alpha = 1$. Then $M = 4$, and our objective is met by letting $\beta \ge 4$.

Existence and Uniqueness of Solutions

If f is continuous in the strip $a \leq t \leq b$, $-\infty < x < \infty$ and satisfies the Lipschitz condition

$$|f(t, x_1) - f(t, x_2)| \leq L|x_1 - x_2|$$

then the initial-value problem has a unique solution in the interval $[a, b]$.

Taylor-Series Method

$$\begin{cases} x' = \cos t - \sin x + t^2 \\ x(-1) = 3 \end{cases}$$

$$x(t+h) = x(t) + hx'(t) + \frac{h^2}{2!}x^{(2)}(t) + + \frac{h^3}{3!}x^{(3)}(t) + \frac{h^4}{4!}x^{(4)}(t) + \cdots$$

$$x^{(2)} = -\sin t - x' \cos x + 2t$$

$$x^{(3)} = -\cos t - x^{(2)} \cos x + (x')^2 \sin t + 2$$

$$x^{(4)} = \sin t - x^{(3)} \cos x + 3x'x^{(2)} \sin x + (x')^3 \cos x$$

Substituting all the derivatives at t up to fourth order and truncating to the order h^4, we can calculate $x(t+h)$ with the truncation error $O(h^5)$.

Taylor-Series Method

Although the truncation error can be very high by Taylor-series method, there are many disadvantages in this method. First, the method depends on repeated differentiation of the given differential equation. Hence, the function $f(t, x)$ must possess partial derivatives in the region where the solution curve passes in the tx-plane. Such an assumption is, of course, not necessary for the existence of a solution. Secondly, various derivatives must be separately programmed.

Taylor-Series Method

The local truncation error is the error made in one step when we replace an infinite process by a finite one. In the Taylor-series method, we replace the infinite Taylor series for $x(t+h)$ by a partial sum. The local truncation error is inherent in any algorithm that we might choose, and is quite independent of roundoff error.

Taylor-Series Method

The accumulation of all local truncation errors gives rise to the global truncation error. Again, this error will be present even if all calculations are performed using exact arithmetic. It is an error that is associated with the method and is independent of the computer on which the calculations are performed. If the local truncation errors are $O(h^{n+1})$, then the global truncation error must be $O(h^n)$ because the number of steps necessary to reach an arbitrary point T, having started at t_0, is $(T - t_0)/h$.
If the global truncation error is $O(h^n)$, we say that the numerical procedure is of order n.

Taylor-Series Method

The Taylor-series method with $n = 1$ is called Euler's method

$$x(t + h) = x(t) + hf(t, x)$$

Exercises: P.495

14,16

Runge-Kutta Methods

The Runge-Kutta methods avoid this calculation of the derivatives although they do imitate the Taylor-series method by means of clever combinations of values of $f(t, x)$. We illustrate by deriving a second-order Runge-Kutta procedure.

$$x(t+h) = x(t) + hx'(t) + \frac{h^2}{2!}x^{(2)}(t) + +\frac{h^3}{3!}x^{(3)}(t) + \cdots$$

$$x'(t) = f$$

$$x^{(2)}(t) = f_t + f_x f$$

$$x^{(3)}(t) = f_{tt} + f_{tx}f + (f_t + f_x f)f_x + f(f_{xt} + f_{xx}f)$$

Thus

$$x(t+h) = x(t) + hf + \frac{1}{2}h^2(f_t + ff_x) + O(h^3)$$

Runge-Kutta Methods

Noting that

$$f(t+h, x+hf) = f + hf_t + O(h^2)$$

and substituting this to the expansion of $x(t+h)$, one has

$$x(t+h) = x(t) + \frac{h}{2}f + \frac{h}{2}f(t+h, x+hf) + O(h^3)$$

So the second-order Runge-Kutta (Heun's method) is

$$x(t+h) = x(t) + \frac{1}{2}(F_1 + F_2)$$

where

$$\begin{cases} F_1 = hf(t,x) \\ F_2 = hf(t+h, x+F_1) \end{cases}$$

Runge-Kutta Methods

In general, second-order Runge-Kutta formulae are of the form

$$x(t+h) = x + w_1 hf + w_2 hf(t+\alpha h, x+\beta hf) + O(h^3)$$

where w_1, w_2, α, β are parameters at our disposal.

$$x(t+h) = x + w_1 hf + w_2 h[f + \alpha hf_t + \beta hff_x] + 0(h^3)$$

Comparing this with the Taylor expansion of $x(t+h)$, we should impose these conditions:

$$\begin{cases} w_1 + w_2 = 1 \\ w_2\alpha = 1/2 \\ w_2\beta = 1/2 \end{cases}$$

One solution is $w_1 = w_2 = 1/2$, $\alpha = \beta = 1$, which is the one corresponding to Heun's method.

Runge-Kutta Methods

If $w_1 = 0$, $w_2 = 1$, $\alpha = \beta = 1/2$, the resulting formula is called modified Euler method:

$$x(t+h) = x(t) + F_2$$

where

$$\begin{cases} F_1 = hf(t, x) \\ F_2 = hf(t + \frac{1}{2}h, x + \frac{1}{2}F_1) \end{cases}$$

Runge-Kutta Methods

Fourth-order Runge-Kutta method:

$$x(t+h) = x(t) + \frac{1}{6}(F_1 + 2F_2 + 2F_3 + F_4)$$

where

$$\begin{cases} F_1 = hf(t,x) \\ F_2 = hf(t + \frac{1}{2}h, x + \frac{1}{2}F_1) \\ F_3 = hf(t + \frac{1}{2}h, x + \frac{1}{2}F_2) \\ F_4 = hf(t + h, x + F_3) \end{cases}$$

This is called a fourth-order method because it reproduces the terms in the Taylor series up to and including the one involving h^4. The error is therefore $O(h^5)$. Exact expressions for the h^5 error term are available.

Runge-Kutta Methods

At the very first step in the fourth-order Runge-Kutta procedure, a value of $x(t_0 + h)$ is computed by the algorithm. On the other hand, there is a correct solution, $x^*(t_0 + h)$, which we shall not know. The local truncation error in this step is, by definition,

$$x^*(t_0 + h) - x(t_0 + h)$$

The theory of the Runge-Kutta algorithm indicates that this truncation error behaves like Ch^5, for small values of h. Here C is a number independent of h.

Runge-Kutta Methods

Let v be value of the approximate solution at $t_0 + h$ obtained by taking one step of length h from t_0. Let u be the approximate solution at $t0 + h$ obtained by taking two steps of size $h/2$ from t_0. These are both computable. By the assumption made, we have

$$x^*(t_0 + h) = v + Ch^5$$

$$x^*(t_0 + h) = u + 2C(h/2)^5$$

By substraction, we obtain the local truncation error

$$Ch^5 = (u - v)/(1 - 2^{-4})$$

Thus the local truncation error is approximated by $u - v$.

Runge-Kutta Methods

In a computer, realization of the Runge-Kutta method, the approximate truncation error can be occasionally monitored, by computing $|u - v|$, to be sure that it remains below a specified tolerance. If it does not, the step size can be decreased (usually halved) to improve the local truncation error. On the other hand, if the local truncation error is far below a permitted threshold, then the step size can be doubled.

The number of required function evaluations increases more rapidly than the order of the Runge-Kutta methods.

number of function evaluations	1	2	3	4	5	6	7	8
maximum order of Rungr-Kutta method	1	2	3	4	4	5	6	6

Unfortunately, this makes the higher-order Runge-Kutta methods less attractive than the classical fourth-order method, since they are more expensive to use.

Exercises: P.506

2,5,6, 9

Multi-Step Methods

$$x(t_{n+1}) - x(t_n) + \int_{t^n}^{t_{n+1}} f(t, x(t))dt$$

The integral on the right can be approximated by a numerical quadrature scheme, and the result will be a formula for generating the approximate solution, step by step.
Suppose that resulting formula is of the following type:

$$x_{n+1} = x_n + af_n + bf_{n-1} + cf_{n-2} + \cdots$$

where f_i denotes $f(t_i, x_i)$. An equation of this type is called an Adams-Bashforth formula. Here is the Adams-Bashforth formula of order 5, based on equally spaced points $t_i = t_0 + ih$ for $0 \le i \le n$:

$$x_{n+1} = x_n + \frac{h}{720}[1901f_n - 2774f_{n-1} + 2616f_{n-2} - 1274f_{n-3} + 251f_{n-4}]$$

How were these coefficients determined?

Multi-Step Methods

We start with the intention of approximating the integral as

$$\int_{t_n}^{t_{n+1}} f(t,x(t))dt \approx h[Af_n + Bf_{n-1} + Cf_{n-2} + Df_{n-3} + Ef_{n-4}]$$

The coefficients A, B, C, D, E are determined by requiring that the above numerical integration formula be exact whenever the integrand is a polynomial of degree ≤ 4.

Consider the form

$$x_{n+1} = x_n + af_{n+1} + bf_n + cf_{n-1} + \cdots$$

Here is a formula of this type, known as the Adams-Moulton formula of order 5:

$$x_{n+1} = x_n + \frac{h}{720}[251f_{n+1} + 646f_n - 264f_{n-1} + 106f_{n-2} - 19f_{n-3}]$$

This formula also can be derived using the method of undetermined coefficients.

Multi-Step Methods

Notice that it cannot be used directly to advance the solution because x_{n+1} occurs on both sides of the equation! Remember that f_i stands for $f(t_i, x_i)$, and so the f_{n+1} can be computed only after x_{n+1} is known. However, a very satisfactory algorithm, called a predictor-corrector method, employs the Adams-Bashforth formula to predict a tentative value for x_{n+1}, say x^*_{n+1}, and then the Adams-Moulton formula of order 5 to compute a corrected value of x_{n+1}. So we evaluate f_{n+1} as $f(t_{n+1}, x^*_{n+1})$ using the predicted value x^*_{n+1} obtained.

Multi-Step Methods

The multi-step methods in general has the form

$$a_k x_n + a_{k-1}x_{n-1} + \cdots + a_0 x_{n-k} = h[b_k f_n + b_{k-1} f_{n-1} + \cdots + b_0 f_{n-k}]$$

This is called a k-step method. If $b_k = 0$, the method is said to be explicit. Otherwise, the method is said to be implicit.
We define a linear functional L by

$$Lx = \sum_{i=0}^{k} [a_i x(ih) - h b_i x'(ih)] = d_0 x(0) + d_1 h x'(0) + d_2 h^2 x^{(2)}(0) + \cdots$$

By the Taylor expansion,

$$x(ih) = \sum_{j=0}^{\infty} \frac{(ih)^j}{j!} x^{(j)}(0), \quad x'(ih) = \sum_{j=0}^{\infty} \frac{(ih)^j}{j!} x^{(j+1)}(0)$$

we have

$$d_0 = \sum_{i=0}^{k} a_i, \quad d_j = \sum_{i=0}^{k} \left(\frac{i^j}{j!} a_i - \frac{i^{j-1}}{(j-1)!} b_i \right) \quad (j = 1, 2, \cdots)$$

Multi-Step Methods

These properties of the multi-step method are equivalent: (i) $d_0 = d_1 = \cdots = d_m = 0$. (ii) $L_p = 0$ for each polynomial p of degree $\leq m$. (iii) Lx is $O(h^{m+l})$ for all $x \in C^{m+1}$.

Proof. If (i) is true, then

$$Lx = d_{m+1}h^{m+1}x^{(m+1)}(0) + \cdots$$

If x is a polynomial of degree $\leq m$, then $Lx = 0$. So (i) implies (ii). Assume that (ii) is true. If $x \in C^{m+1}$, then by Taylor's Theorem we can write $x = p + r$ where p is a polynomial of degree $\leq m$ and r is a function whose first m derivatives vanish at 0. Since $Lp = 0$,

$$Lx = Lr = d_{m+1}h^{m+1}r^{(m+l)}(0) = O(h^{m+1})$$

and (ii) implies (iii). Finally, assume that (iii) is true, we must have the condition $d_0 = d_1 = \cdots = d_m = 0$. Hence, (iii) implies (i).

The multi-step method in Equation is the unique natural number m such that

$$d_0 = d_1 = \cdots = d_m = 0 \neq d_{m+1}$$

Exercises: P.514

1,4,8,9

Local and Global Errors: Stability

We denote $x(h,t)$ the approximate solution obtained by using step size h. As usual, the exact solution The multi-step method is said to be convergent if

$$\lim_{h\to 0} x(h,t) = x(t) \quad (t \textit{ fixed})$$

for all t in some interval $[t_0, t_m]$, provided only that the starting values obeys the same equation, that is,

$$\lim_{h\to 0} x(h, t_0 + nh) = x_0 \quad (0 \leq n < k)$$

Local and Global Errors: Stability

Two other terms that are used are stable and consistent. The method is stable if all roots of

$$p(z) = a_k z^k + a_{k-1} z^{k-1} + \cdots + a_0$$

lie in the disk $|z| \leq 1$ and if each root of modulus one is simple. The method is consistent if $p(1) = 0$ and $p'(1) = q(1)$, where

$$q(z) = b_k z^k + b_{k-1} z^{k-1} + \cdots + b_0$$

For the multi-step method to be convergent, it is necessary and sufficient that it be stable and consistent.

Local and Global Errors: Stability

If the multi-step is of order m, if $x \in C^{m+2}$, and if $\partial f/\partial x$ is continuous, then under the hypotheses of the preceding paragraph

$$x(t_n) - x_n = (d_{m+1}/a_k)h^{m+1}x^{(m+1)}(t_{n-k}) + O(h^{m+2})$$

Proof. It suffices to prove the equation when $n = k$, since x_n can be interpreted as the value of a numerical solution that began at the point t_{n-k}. Using the linear functional L, we have

$$Lx = \sum_{i=0}^{k}[a_i x(t_i) - hb_i x'(t_i)] = \sum_{i=0}^{k}[a_i x(t_i) - hb_i f(t_i, x(t_i))]$$

Local and Global Errors: Stability

On the other hand, the numerical solution satisfies the equation

$$0 = \sum_{i=0}^{k} [a_i x_i - h b_i f(t_i, x_i)]$$

Since we have assumed that $x_i = x(t_i)$ for $i < k$, the result of subtracting the second equation from the first will be

$$\begin{aligned} Lx &= a_k[x(t_k) - x_k] - hb_k[f(t_k, x(t_k)) - f(t_k, x_k) \\ &= [a_k - hb_k \partial f(t_k, \xi)/\partial x][x(t_k) - x_k] \end{aligned}$$

Since that the method being used is of order m, then L_x will have the form

$$Lx = d_{m+1} h^{(m+l)} x^{(m+1)}(t_0) + O(h^{m+2})$$

Combining the above two results, we complete the proof.

Exercises: P.523

1(a)(b),2,3,6(1)