Lecture 30 (hrs. 60,61) - December 3, 2025, 11:30 - 13:30 A13

(4.12) Theorem (convergence of the TS(1) method).

Let t_0 be a real number, F be a function defined in $R \times R^n$ with values in R^n and let $x_0 \in R^n$, and consider the Cauchy Problem:

(§) $\quad x'(t) = F(\, t, x(t)\,) \quad , \quad x(t_0) = x_0 \quad , \quad t \in [t_0, t_f]$

If all first partial derivatives of F(t,x) are continuous functions of t and x and Problem (§) has only one solution, *then* for every $\lambda > 0$ the TS(1) method applied to Problem (§) is *convergent* as $E \to 0$ and:

- $N \to \infty$ *as* $1 / \sqrt{E}$;
- *For every* k: $ET(k) \to 0$ *as* $\sqrt{E}$.

(4.13) Realization in *Scilab* (TS_1_pv).

```
function [T, X, PASSO] = TS_1_pv(x0, t0, tf, F, G2, E, LAMBDA, HMIN)
//
// Numerically integrates, on [t0,tf], the Cauchy Problem
// in R(n):
//
// x' = F(t,x)
// x(t0) = x0
//
// using the TS(1)cmethod - explicit Euler - with variable step.
//
// x0: initial condition (column of n elements)
// t0: initial time (real number)
// tf: final time (real number)
// F: function which defines the differential equation; F(t,x) should
//    be a column of n real numbers
// G2: function whose values are the second derivative in t of the solution of the
//     differential equation whose value at time t is x; G2(t,x) should be
//     a column of n real numbers
// E: maximum value of the estimate of the local error (real number)
// LAMBDA: real number that sets the maximum value of the step
//         (OPTIONAL - predefined value: 1d-5)
// HMIN: minimal allowed value of the step
//       (OPTIONAL - predefined value: (tf - t0) / 1d6)
//
// T = [t(0),...,t(N)]: row whose elements are the integration instants
// X = [x(0),...,x(N)]: matrix n x (N+1) whose columns are the approximations
// PASSO = [h(0),...,h(N-1)]: row whose elements are the integration step
//
// Values for the optional input parameters
//
if ~exists('LAMBDA','l') then LAMBDA = 1d-5; end;
```

```
if ~exists('HMIN','l') then HMIN = (tf - t0) / 1d6; end;
//
// Initialization of the output variables
//
T(1,1) = t0;
X(:,1) = x0;
PASSO = [];
//
// main loop
//
while (T(1,$) < tf), // halt the construction if tf has been reached
  //
  // choice of the step
  //
  Nd2x = norm(G2(T(1,$),X(:,$)));
  d = max(LAMBDA, Nd2x);
  PASSO(1,$+1) = min(sqrt(2*E/d), tf - T(1,$));
  //
  // computation of the approximation and of the new integration instant
  //
  X(:,$+1) = X(:,$) + F(T(1,$),X(:,$)) * PASSO(1,$);
  T(1,$+1) = T(1,$) + PASSO(1,$);
  //
  // halt the construction if the computed step is too small and tf has not been reached
  //
  if (PASSO(1,$) < HMIN) & (T(1,$) < tf) then break; end;
  //
end;
//
// Verify if the integration reached tf
//
if T(1,$) < tf then
   printf("\n\nIntegrazione interrotta a T = %3.2e", T(1,$));
end;
//
endfunction
```

(4.14) Example (explained to the class on December 4th).

Consider a pendulum consisting of a heavy point of mass m connected to a fixed point by an inextensible string of length L. Assuming the motion of the point to be planar and adopting the angle x between the downward-sloping vertical and the string, measured counterclockwise, as the Lagrangian coordinate, the equation of motion is:

(ED) $$x''(t) = -\frac{g}{L}\sin x(t)$$

To approximate on the interval $[t_0,t_f] = [0,3]$ s the solution of the Cauchy problem which is obtained by considering the initial conditions:

(CI) $$x(0) = x_0 = \pi/4 \text{ rad} \quad , \quad x'(0) = 0$$

with *Scilab*, the TS_1_pv procedure is used. Using the procedure requires:

- *To determine a system of two first-order differential equations equivalent to equation* (ED). By introducing the variables $u_1(t) = x(t)$ and $u_2(t) = x'(t)$, we obtain:

  (ED') $$u_1'(t) = u_2(t) \quad , \quad u_2'(t) = -\frac{g}{L}\sin u_1(t)$$

  which is completed with the initial conditions:

  (CI') $$u_1(0) = x_0 \quad , \quad u_2(0) = 0$$

- *To write the* function *that defines the system* (ED'):

```
function y = F(t,u)

   y = [                  u(2)  ;
         - (g/L) * sin( u(1) ) ];

endfunction
```

- *To find the function that*, given t and u, *returns the value of the second derivative*, calculated at t, *of the solution of the system* (ED') that evaluates to u at time t:

$$u''(t) = \begin{bmatrix} u_2'(t) \\ -(g/L)\, u_1'(t) \cos(u_1(t)) \end{bmatrix} = \begin{bmatrix} -(g/L)\sin(u_1(t)) \\ -(g/L)\, u_2(t) \cos(u_1(t)) \end{bmatrix}$$

  and to write the relative function:

```
function y = G2(t,u)

   y = [        - (g/L) * sin( u(1) );
         - (g/L) * u(2) * cos( u(1) ) ];

endfunction
```

- *To assign the final instant* t_f (s):

```
tf = 3;
```

- *To assign the column of the initial conditions* (CI'):

```
u0 = [x0;0];
```

- *To assign values to the parameters*:

```
g = 9.82; // m/s^2
L = 1; // m
m = 1; // kg
```

- *To choice the maximum allowed value for the estimate of the local error*, E.

  To obtain an adequate value of E, we need a criterion to judge the accuracy of the approximation obtained by the procedure. For the physical system under consideration, we can proceed as follows.

  (A) Since the *mechanical energy*:

$$EN(x(t)) = mgL(1-\cos x_1(t)) + \frac{1}{2}mL^2(x_2(t))^2$$

  assumes along the motion the constant value $EN(t_0)$, as a *relative* measure of the accuracy of the approximation, we can choose *the relative change in energy during the approximate motion*:

$$Var\_EN = \frac{\max EN(u(t_k)) - \min EN(u(t_k))}{EN(u(t_0))}$$

  (B) Since the motion of the pendulum is periodic and:

$$\min x_1(t) = -\max x_1(t) \quad \Rightarrow \quad \max x_1(t) + \min x_1(t) = 0$$

  as a *relative* measure of the accuracy of the approximation, we can choose *the relative change of the amplitude of the oscillation during the approximate motion*:

$$Var\_A = \frac{\max u_1(t_k) + \min u_1(t_k)}{u_1(t_0)}$$

  This choice is reasonable if the interval $[t_0,t_f]$ includes at least one oscillation of the function $u_1(t_k)$.

  (C) We get the following table:

E	N	Var_EN (%)	Var_A (%)
10^{-3}	267	35.89	6.3
10^{-5}	2587	3.25	$5.99\ 10^{-1}$
10^{-7}	25779	0.32	$5.97\ 10^{-2}$

  What an appropriate value of E is depends on what the user wants to achieve. The table suggests that as E decreases, the accuracy of the approximation increases.

(4.15) Remark (variation of N and ET with E).

Let N and M, respectively, be the number of integration instants and the maximum value of ET(k) obtained using the TS_1_pv procedure with E = $\underline{E}$ and N' and M' be the corresponding values obtained with E = $\alpha\,\underline{E}$. By Theorem (4.12) we expect that:

$$N'/N \approx 1/\alpha^{1/2} \quad \text{and} \quad M'/M \approx \alpha^{1/2}$$

In the final table of the previous Example we have $\alpha = 10^{-2}$, so we expect:

$$N'/N \approx 10 \quad \text{and} \quad M'/M \approx 1/10$$

The relationship regarding the increase in the number of integration instants is evidently verified:

$$2587/267 = 9.69 \quad \text{and} \quad 25779/2587 = 9.96$$

Since we cannot access the total error, we simply note that for the relative variation of the energy we have:

$$Var\_EN'/Var\_EN = 3.25/35.89 \approx 0.90\ 10^{-1} \quad \text{and} \quad 0.32/3.25 \approx 0.98\ 10^{-1}$$

and for the relative variation of the amplitude:

$$Var\_A'/Var\_A = 5.99\ 10^{-1}/6.3 \approx 0.95\ 10^{-1} \quad \text{and} \quad 5.97\ 10^{-2}/5.99\ 10^{-1} \approx 0.99\ 10^{-1}$$

(4.B) TS(2) METHOD

(4.16) Hypothesis (regularity of the solutions).

Suppose that *all* solutions of the differential equation $x'(t) = F(t,x(t))$ have *continuous third derivatives*.

The condition is certainly satisfied if *all second* partial derivatives of the function $F(t,x)$ *exist* and are *continuous* functions of t and x.

(Indeed:

$$G_2(t,x) = \partial_t F(t,x) + \partial_x F(t,x) \cdot F(t,x)$$

has continuous first partial derivatives and therefore:

$$G_3(t,x) = \partial_t G_2(t,x) + \partial_x G_2(t,x) \cdot F(t,x)$$

is continuous. Then, if y(t) is a solution of the differential equation:

$$y^{(3)}(t) = ((y'(t))')' = ((F(t,y(t)))')' = (G_2(t,y(t)))' = G_3(t,y(t))$$

is continuous because $G_3(t,x)$ and y(t) are.)

(4.17) Definition (TS(2) method).

The TS(2) *method* is defined by the following procedures.

- CHOICE of h(k). Given $E > 0$ and $\lambda > 0$, for each k we set:

$$d(k) = \max \{ \lambda , \ \| y^{(3)}(t(k); x(k),t(k)) \| \ \}$$

then:

$$h(k) = \min \{ \ \sqrt[3]{\frac{6E}{d(k)}} \ , tf - t(k) \}$$

- CALCULATION of x(k+1). After choosing h(k) we set:

$$x(k+1) = x(k) + F(t(k),x(k)) \, h(k) + \frac{1}{2} \, G_2(t(k),x(k)) \, h(k)^2$$

The name of the method is a consequence of the fact that the function used to calculate x(k+1) is obtained by truncating the *Taylor series* of y(t(k) + h; x(k),t(k)) at the *second* term in h = 0.

(4.18) Remark (on the choice of h(k)).

Let y(t) be the solution y(t; x(k),t(k)) of the differential equation. The deviation s(h) between y(t(k) + h) and the approximation calculated by the method with a step of length h starting from (t(k),x(k)) is, using the Taylor Formula in h = 0 with Lagrange remainder:

$$s(h) = - \frac{1}{6} \, y^{(3)}(t(k)) \, h^3 + z(h) \, h^3 \qquad \text{where: } z(h) \to 0 \text{ as } h \to 0$$

If $y^{(3)}(t(k)) \neq 0$ *then*:

- *When* h *is small*: $- \frac{1}{6} \, y^{(3)}(t(k)) \, h^3$ *is a good estimate of* s(h)
- It is:

$$\| -\frac{1}{6} \, y^{(3)}(t(k)) \, h^3 \| = E \quad \Leftrightarrow \quad h = \sqrt[3]{\frac{6E}{\| y^{(3)}(t(k)) \|}}$$

The parameter λ is intended to prevent d(k) = 0 and also ensures that:

$$\text{for every k:} \quad d(k) \geqslant \lambda \quad \text{hence} \quad h(k) \leqslant \sqrt[3]{\frac{6E}{\lambda}}$$

(4.19) Theorem (convergence of the TS(2) method).

Let t_0 and $t_f > t_0$ be real numbers, let F be a function defined in $R \times R^n$ with values in R^n, let $x_0 \in R^n$ and consider the Cauchy Problem:

$$(\S) \qquad x'(t) = F(t,x(t)) \ , \quad x(t_0) = x_0 \ , \quad t \in [t_0,t_f]$$

If all second partial derivatives of F(t,x) are continuous functions of t and x and Problem (§) has only one solution, *then* for every $\lambda > 0$ the TS(2) method applied to Problem (§) is *convergent* as E → 0 and:

- N → ∞ *as* $1 / \sqrt[3]{E}$;
- *For every* k: ET(k) → 0 *as* $\sqrt[3]{E^2} = E^{2/3}$

(4.20) Remark.

Consider the Cauchy Problem (§). For each $E > 0$, let $N_1(E)$ and $ET_1(E)$ be the number of integration instants and the maximum total error generated by the TS(1) method, and let $N_2(E)$ and $ET_2(E)$ be the number of integration instants and the maximum total error generated by the TS(2) method. By Theorem (4.12) and Theorem (4.19), as $E \to 0$ we have:

- $N_1(E) / N_2(E) \to +\infty$ as $1 / \sqrt[6]{E}$, hence $N_1(E) \to \infty$ *faster than* $N_2(E)$
- $ET_1(E) / ET_2(E) \to +\infty$ come $1 / \sqrt[6]{E}$, dunque $ET_2(E) \to 0$ *faster than* $ET_1(E)$

We then expect that, with the same value of E:

- TS(2) generates a *smaller maximum total error* than that generated by TS(1)
- TS(2) reaches t_f with *fewer steps* than TS(1)