Round-off Error Analysis of Explicit One-Step Numerical Integration Methods

RAIM 2017, LIP, ÉNS Lyon

Sylvie Boldo[1] Florian Faissole[1] Alexandre Chapoutot[2]

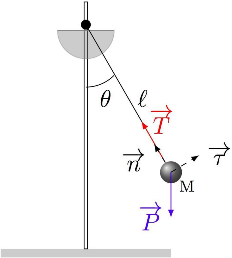
[1]Inria - LRI, Univ. Paris-Sud et CNRS - Univ. Paris-Saclay

[2]U2IS, ÉNSTA ParisTech

Table of contents

Ordinary differential equations (ODEs)

$$y'(t) = f(y, t).$$



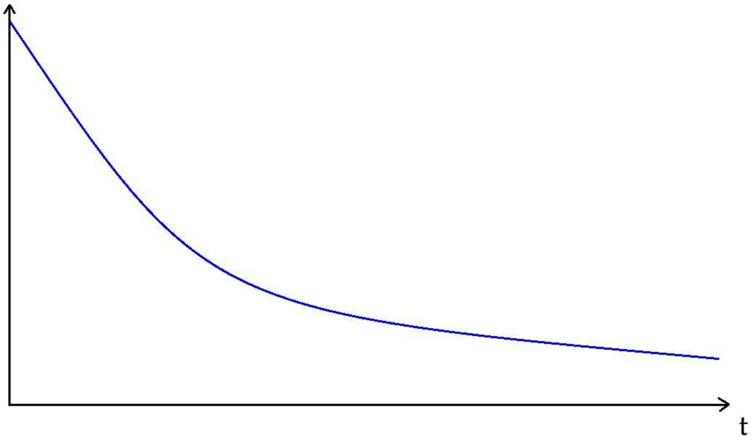
Motion of pendulum

Car shock absorber

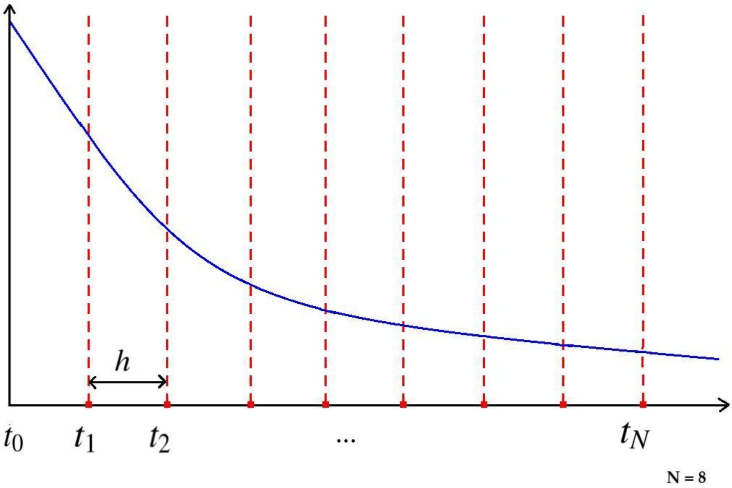
SpaceX Falcon 9 space rocket

Exact resolution is hard ⇒ numerical methods.

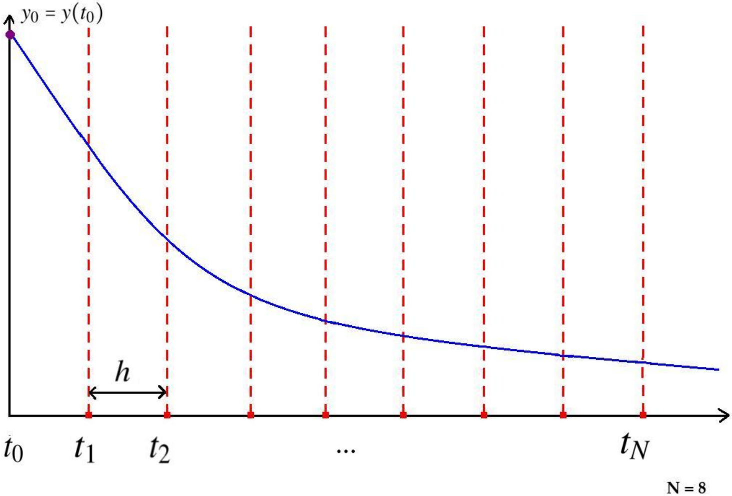
Numerical integration



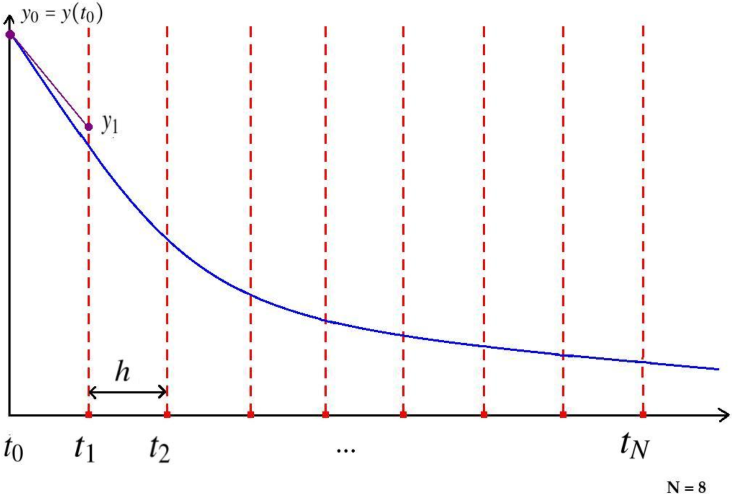
Numerical integration



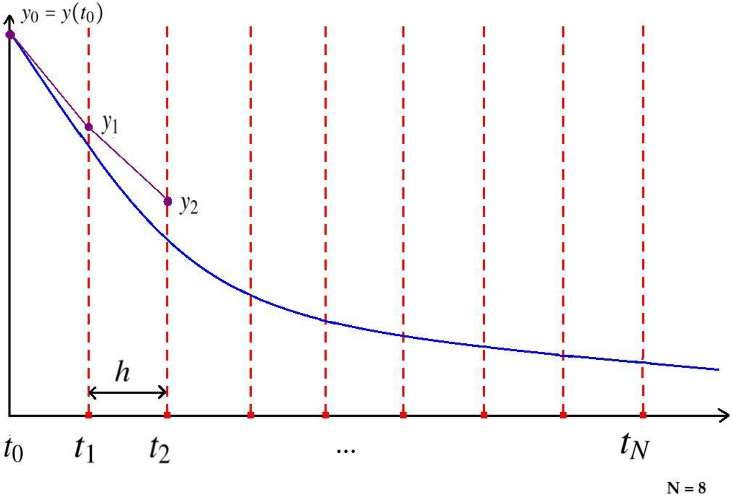
Numerical integration



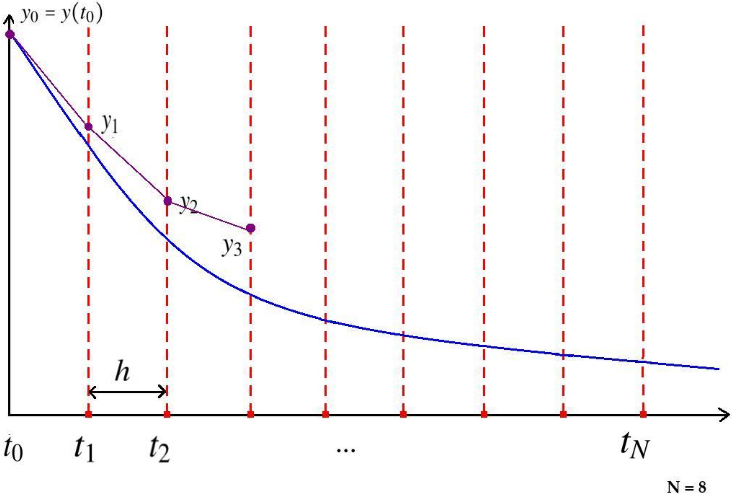
Numerical integration



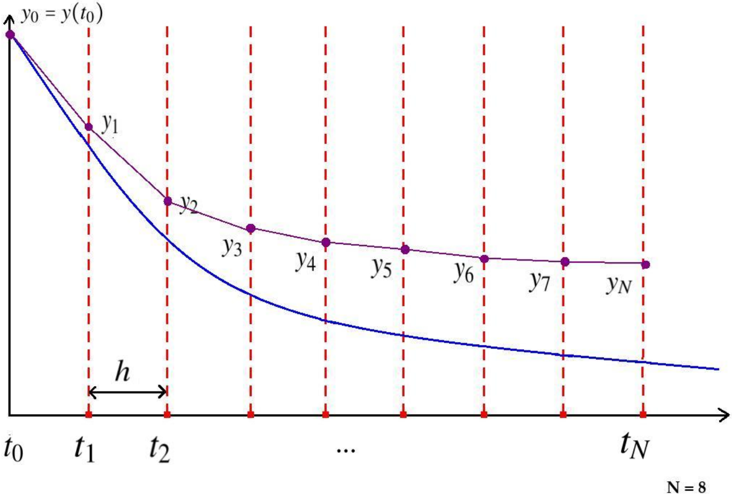
Numerical integration



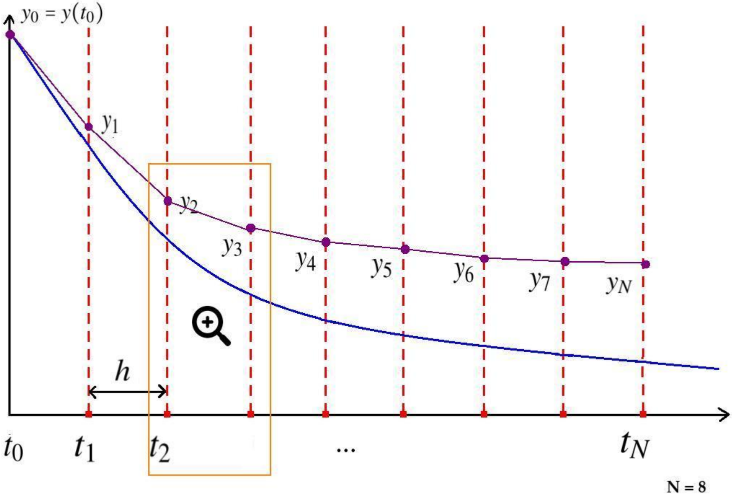
Numerical integration



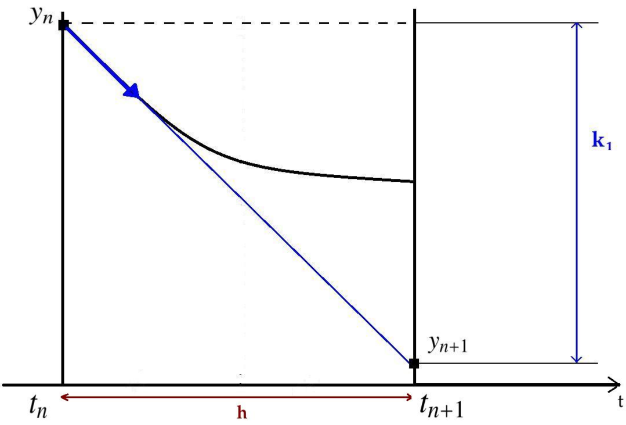
Numerical integration



Numerical integration

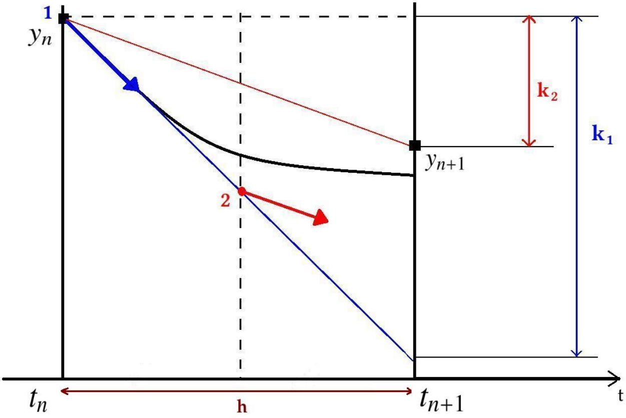


Euler method



$$k_1 = hf(t_n, y_n)$$
$$y_{n+1} = y_n + k_1 + O(h^2)$$

RK2 method



$$k_1 = h \times f(t_n, y_n) \qquad k_2 = h \times f(t_n + \frac{h}{2}, y_n + \frac{k_1}{2})$$
$$y_{n+1} = y_n + k_2 + O(h^3)$$

Working assumptions

- FP arithmetic;
  - neither underflow nor overflow,
  - radix 2 double precision, $u = 2^{-53}$

Working assumptions

- FP arithmetic;
  - neither underflow nor overflow,
  - radix 2 double precision, $u = 2^{-53}$
- ODEs;
  - first-order,
  - linear, $y' = \lambda y$
  - $y : \mathbb{R} \to \mathbb{R}$,

Working assumptions

- FP arithmetic;
  - neither underflow nor overflow,
  - radix 2 double precision, $u = 2^{-53}$
- ODEs;
  - first-order,
  - linear, $y' = \lambda y$
  - $y : \mathbb{R} \to \mathbb{R}$,
- Methods.
  - explicit,
  - one step,
  - constant step.

RK methods on linear problems: linear stability

$$\begin{cases} y_0 \in \mathbb{R} \\ y_{n+1} = R(h,\lambda) y_n \end{cases}$$

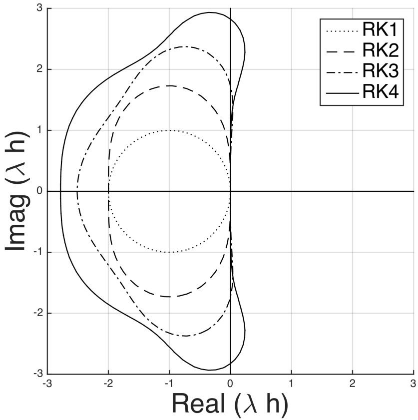
(R polynomial in $h\lambda$)

RK methods on linear problems: linear stability

$$\begin{cases} y_0 \in \mathbb{R} \\ y_{n+1} = R(h,\lambda) y_n \end{cases}$$

(R polynomial in $h\lambda$)

Stable $\Leftrightarrow$ $|R(h,\lambda)| < 1$:



RK methods on linear problems: FP implementation

$$\begin{cases} y_0 \in \mathbb{R} \\ y_{n+1} = R(h,\lambda) y_n \end{cases} \qquad \begin{cases} \widetilde{y_0} \simeq y_0 \\ \widetilde{y_{n+1}} = \widetilde{R}(\widetilde{h}, \widetilde{\lambda}, \widetilde{y_n}) \end{cases}$$

RK methods on linear problems: FP implementation

$$\begin{cases} y_0 \in \mathbb{R} \\ y_{n+1} = R(h,\lambda) y_n \end{cases} \qquad \begin{cases} \widetilde{y_0} \simeq y_0 \\ \widetilde{y_{n+1}} = \widetilde{R}(\widetilde{h}, \widetilde{\lambda}, \widetilde{y_n}) \end{cases}$$

Euler:

- $R(h,\lambda) = 1 + h\lambda$;
- $\widetilde{R}(\widetilde{h}, \widetilde{\lambda}, \widetilde{y_n}) = \widetilde{y_n} \oplus \widetilde{h} \otimes \widetilde{\lambda} \otimes \widetilde{y_n}$.

RK methods on linear problems: FP implementation

$$\begin{cases} y_0 \in \mathbb{R} \\ y_{n+1} = R(h,\lambda) y_n \end{cases} \qquad \begin{cases} \widetilde{y_0} \simeq y_0 \\ \widetilde{y_{n+1}} = \widetilde{R}(\widetilde{h}, \widetilde{\lambda}, \widetilde{y_n}) \end{cases}$$

Euler:

- $R(h,\lambda) = 1 + h\lambda$;
- $\widetilde{R}(\widetilde{h}, \widetilde{\lambda}, \widetilde{y_n}) = \widetilde{y_n} \oplus \widetilde{h} \otimes \widetilde{\lambda} \otimes \widetilde{y_n}$.

RK4:

- $R(h,\lambda) = 1 + h\lambda + \frac{1}{2}(h\lambda)^2 + \frac{1}{6}(h\lambda)^3 + \frac{1}{24}(h\lambda)^4$;
- $\widetilde{R}(\widetilde{h}, \widetilde{\lambda}, \widetilde{y_n}) =$

$$\begin{aligned} &\widetilde{y_n} \oplus \widetilde{h} \oslash 6 \otimes \widetilde{\lambda} \otimes \widetilde{y_n} \oplus \widetilde{h} \oslash 3 \otimes \widetilde{\lambda} \otimes \widetilde{y_n} \oplus \widetilde{h} \otimes \widetilde{h} \oslash 6 \otimes \widetilde{\lambda} \otimes \widetilde{\lambda} \otimes \widetilde{y_n} \oplus \widetilde{h} \oslash 3 \otimes \widetilde{\lambda} \otimes \widetilde{y_n} \otimes \\ &\widetilde{h} \otimes \widetilde{h} \oslash 6 \otimes \widetilde{\lambda} \otimes \widetilde{\lambda} \otimes \widetilde{y_n} \oplus \widetilde{h} \otimes \widetilde{h} \otimes \widetilde{h} \oslash 12 \otimes \widetilde{\lambda} \otimes \widetilde{\lambda} \otimes \widetilde{\lambda} \otimes \widetilde{y_n} \oplus \widetilde{h} \oslash 6 \otimes \widetilde{\lambda} \otimes \widetilde{y_n} \\ &\oplus \widetilde{h} \otimes \widetilde{h} \oslash 6 \otimes \widetilde{\lambda} \otimes \widetilde{\lambda} \otimes \widetilde{y_n} \oplus \widetilde{h} \otimes \widetilde{h} \otimes \widetilde{h} \oslash 12 \otimes \widetilde{\lambda} \otimes \widetilde{\lambda} \otimes \widetilde{\lambda} \otimes \widetilde{y_n} \\ &\oplus \widetilde{h} \otimes \widetilde{h} \otimes \widetilde{h} \otimes \widetilde{h} \oslash 24 \otimes \widetilde{\lambda} \otimes \widetilde{\lambda} \otimes \widetilde{\lambda} \otimes \widetilde{\lambda} \otimes \widetilde{y_n}. \end{aligned}$$

($>$ 60 flops!)

State-of-the-art

Roundoff errors in numerical methods (N = nb of iterations):

- probabilistic result: error in $\sqrt{N}$ [Henrici,1963];
- in practice (implicit RK): error in N [Hairer&al, 2008];
- interval analysis [Bouissou-Martel, 2006];
- numerical integration (fine-grained): Newton-Cotes, Gauss-Legendre, ... [Fousse, 2006].

State-of-the-art

Roundoff errors in numerical methods (N = nb of iterations):

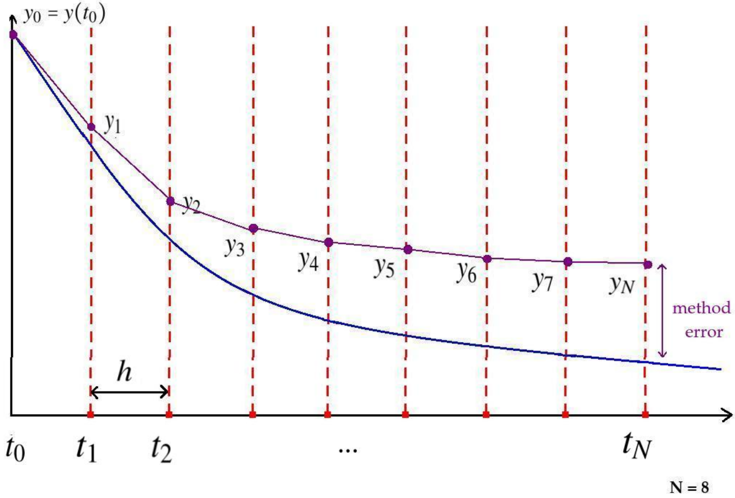
- probabilistic result: error in $\sqrt{N}$ [Henrici,1963];
- in practice (implicit RK): error in N [Hairer&al, 2008];
- interval analysis [Bouissou-Martel, 2006];
- numerical integration (fine-grained): Newton-Cotes, Gauss-Legendre, ... [Fousse, 2006].

Our approach:

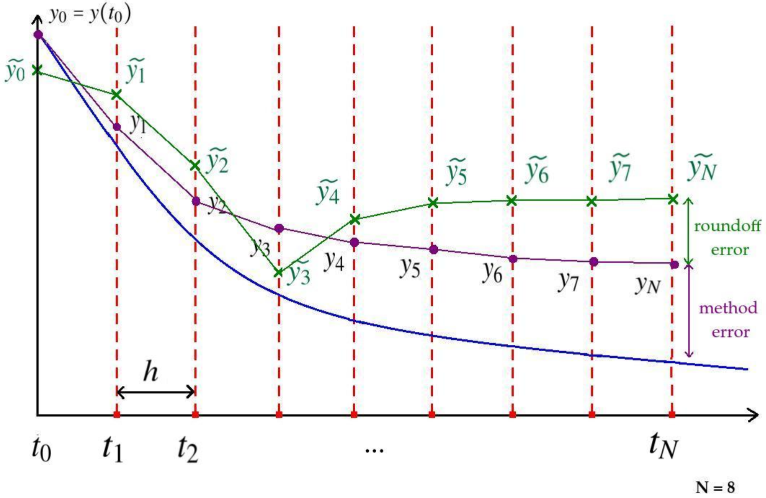
- fined-grained analysis;
- use of mathematical properties of the methods (stability).

Table of contents

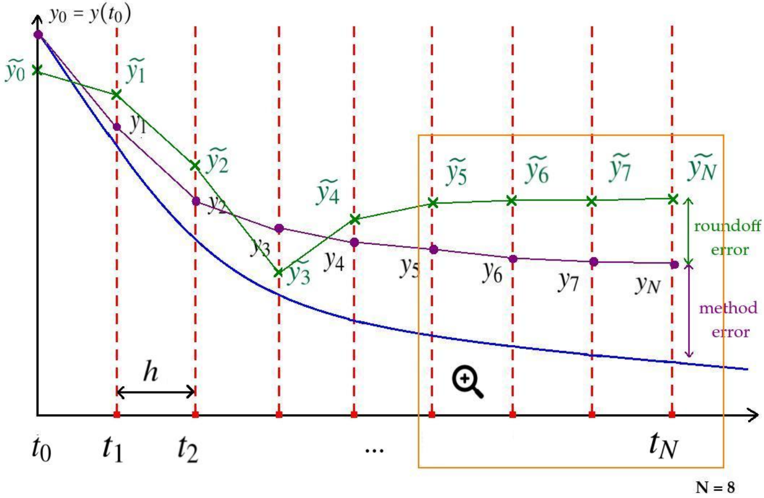
Method error vs roundoff error



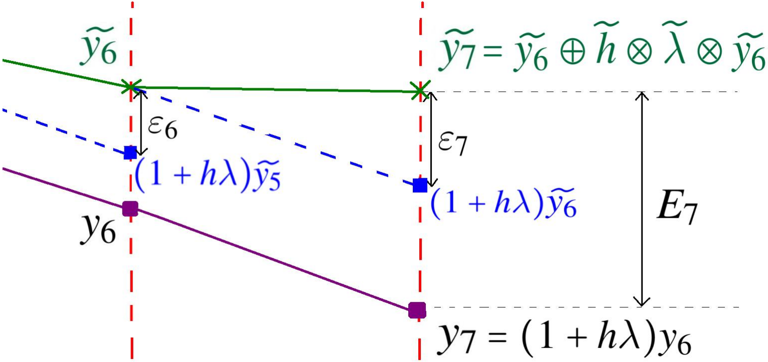
Method error vs roundoff error



Method error vs roundoff error



Local roundoff error vs global roundoff error



Local error:

$$\varepsilon_0 = |\widetilde{y_0} - y_0|$$
$$\forall n \in \mathbb{N}^*,\ \varepsilon_n = |\widetilde{R}(\widetilde{h}, \widetilde{\lambda}, \widetilde{y_{n-1}}) - R(h, \lambda)\widetilde{y_{n-1}}|.$$

Global error:

$$\forall n \in \mathbb{N},\ E_n = \widetilde{y_n} - y_n.$$

From local to global roundoff error

Local error:

$$\varepsilon_0 = |\widetilde{y_0} - y_0|$$
$$\forall n \in \mathbb{N}^*,\ \varepsilon_n = |\widetilde{R}(\widetilde{h}, \widetilde{\lambda}, \widetilde{y_{n-1}}) - R(h, \lambda)\widetilde{y_{n-1}}|.$$

Global error:

$$\forall n \in \mathbb{N},\ E_n = \widetilde{y_n} - y_n.$$

Theorem 1: Global absolute error of RK methods

Let $C \in \mathbb{R}_+^*$. *Suppose* $\forall n \in \mathbb{N}^*, \varepsilon_n \leqslant C|\widetilde{y_{n-1}}|$. *Then,* $\forall n \in \mathbb{N}$,

$$|E_n| \leqslant (C + |R(h,\lambda)|)^n \left(\varepsilon_0 + n\frac{C|y_0|}{C + |R(h,\lambda)|}\right).$$

Relative roundoff errors

Relative error:

$$\left|\frac{\widetilde{y_n} - y_n}{y_n}\right| \leqslant \left(\frac{C + |R(h,\lambda)|}{|R(h,\lambda)|}\right)^n \left(\frac{\varepsilon_0}{|y_0|} + n\frac{C}{C + |R(h,\lambda)|}\right).$$

Relative roundoff errors

Relative error:

$$\left|\frac{\widetilde{y_n} - y_n}{y_n}\right| \leqslant \left(\frac{C + |R(h,\lambda)|}{|R(h,\lambda)|}\right)^n \left(\frac{\varepsilon_0}{|y_0|} + n\frac{C}{C + |R(h,\lambda)|}\right).$$

If $C \ll |R(h,\lambda)|$, then:

$$\left|\frac{\widetilde{y_n} - y_n}{y_n}\right| \lesssim \frac{\varepsilon_0}{|y_0|} + n\frac{C}{|R(h,\lambda)|}.$$

In practice (Euler, RK2, RK4): $C \leq 200u$ and $200u \ll |R(h,\lambda)|$.

Table of contents

Technical lemma for local roundoff errors

Local error of stable Euler's method ($-2 \leq h\lambda < 0$):

$$\varepsilon_{n+1} = |\widetilde{y_n} \oplus (\widetilde{h} \otimes \widetilde{\lambda} \otimes \widetilde{y_n}) - (1 + h\lambda)y_n|.$$

Technical lemma for local roundoff errors

Local error of stable Euler's method ($-2 \leq h\lambda < 0$):

$$\varepsilon_{n+1} = |\underbrace{\widetilde{y_n}}_{X_1} \oplus (\underbrace{\widetilde{h} \otimes \widetilde{\lambda}}_{X_2} \otimes \underbrace{\widetilde{y_n}}_{y}) - (\underbrace{1}_{\alpha_1} + \underbrace{h\lambda}_{\alpha_2}) \underbrace{\widetilde{y_n}}_{y}|.$$

Lemma 2

Let $y \in \mathbb{R}$. *Let* $C_1, C_2 \in \mathbb{R}_+$. *Let* $\alpha_1, \alpha_2 \in \mathbb{R}$. *Let* $X_1, X_2 \in \mathbb{F}$ *s.t.* $|X_1 - \alpha_1 y| \leqslant C_1 |y|$ *and* $|X_2 - \alpha_2| \leq C_2$.

Then:

$$|X_1 \oplus (X_2 \otimes y) - (\alpha_1 + \alpha_2) y| \leqslant |y| \left(C_1 + C_2 + u\left(|\alpha_1| + 2|\alpha_2| + C_1 + 2C_2\right) + u^2 \left(C_2 + |\alpha_2|\right)\right).$$

Technical lemma for local roundoff errors

Local error of stable Euler's method ($-2 \leq h\lambda < 0$):

$$\varepsilon_{n+1} = |\underbrace{\widetilde{y_n}}_{X_1} \oplus (\underbrace{\widetilde{h} \otimes \widetilde{\lambda}}_{X_2} \otimes \underbrace{\widetilde{y_n}}_{y}) - (\underbrace{1}_{\alpha_1} + \underbrace{h\lambda}_{\alpha_2}) \underbrace{\widetilde{y_n}}_{y}|.$$

Lemma 2

Let $y \in \mathbb{R}$. *Let* $C_1, C_2 \in \mathbb{R}_+$. *Let* $\alpha_1, \alpha_2 \in \mathbb{R}$. *Let* $X_1, X_2 \in \mathbb{F}$ *s.t.* $|X_1 - \alpha_1 y| \leqslant C_1 |y|$ *and* $|X_2 - \alpha_2| \leq C_2$.
Then:

$$|X_1 \oplus (X_2 \otimes y) - (\alpha_1 + \alpha_2) y| \leqslant |y| \left(C_1 + C_2 + u\left(|\alpha_1| + 2|\alpha_2| + C_1 + 2C_2\right) + u^2\left(C_2 + |\alpha_2|\right)\right).$$

$$|\underbrace{\widetilde{y_n}}_{X_1} - \underbrace{1}_{\alpha_1} \underbrace{\widetilde{y_n}}_{y}| \leq \underbrace{0}_{C_1} |\underbrace{\widetilde{y_n}}_{y}|, \qquad |\underbrace{\widetilde{h} \otimes \widetilde{\lambda}}_{X_2} - \underbrace{h\lambda}_{\alpha_2}| \leq \underbrace{6u}_{C_2}$$

(Gappa [Melquiond]).

Local roundoff errors of higher-order methods

Stable RK4 method:

	C	
$\widetilde{y_n}$	0	
$\circ\left[\widetilde{h\frac{1}{6}\lambda}\right]$	$2u$	$(Gappa)$
$\circ\left[\widetilde{y_n + h\frac{1}{6}\lambda y_n}\right] = \widetilde{y_n} \oplus \widetilde{h} \oslash 6 \otimes \widetilde{\lambda} \otimes \widetilde{y_n}$	$4u$	$(Lemma\,2)$
$\circ\left[\widetilde{h\frac{1}{3}\lambda}\right]$	$4u$	$(Gappa)$
$\circ\left[\widetilde{hh\frac{1}{6}\lambda\lambda}\right]$	$12u$	$(Gappa)$
$\circ\left[\widetilde{hhh\frac{1}{12}\lambda\lambda\lambda}\right]$	$28u$	$(Gappa)$
$\circ\left[\widetilde{hhhh\frac{1}{24}\lambda\lambda\lambda\lambda}\right]$	$53u$	$(Gappa)$
$\ldots$	$\ldots$	$(7 \times Lemma\,2)$
$\circ\left[\widetilde{y_n + \cdots + hhhh\frac{1}{24}\lambda\lambda\lambda\lambda y_n}\right]$	$194u$	$(Lemma\,2)$

Local roundoff errors of classical methods

Lemma 3: Local error of Euler method

Suppose $-2 \leqslant h\lambda \leqslant -2^{-100}$ *and* $2^{-60} \leqslant h \leqslant 1$. *Then:*
$\forall n \in \mathbb{N}, \quad \varepsilon_{n+1}^{Euler} \leqslant 11.01u\,|\widetilde{y_n}|$.

Lemma 4: Local error of RK2 method

Suppose $-2 \leqslant h\lambda \leqslant -2^{-100}$ *and* $2^{-60} \leqslant h \leqslant 1$. *Then:*
$\forall n \in \mathbb{N}, \quad \varepsilon_{n+1}^{RK2} \leqslant 28.01u\,|\widetilde{y_n}|$.

Lemma 5: Local error of RK4 method

Suppose $-3 \leqslant h\lambda \leqslant -2^{-100}$ *and* $2^{-60} \leqslant h \leqslant 1$. *Then:*
$\forall n \in \mathbb{N}, \quad \varepsilon_{n+1}^{RK4} \leqslant 194u\,|\widetilde{y_n}|$.

Table of contents

Global roundoff errors of classical methods

Bounds on global errors:

- Euler: $C = 11.01u$

$$|E_n| \leqslant (11.01u + |R(h,\lambda)|)^n \left(\varepsilon_0 + n\frac{11.01u|y_0|}{11.01u + |R(h,\lambda)|}\right);$$

- RK2: $C = 28.01u$

$$|E_n| \leqslant (28.01u + |R(h,\lambda)|)^n \left(\varepsilon_0 + n\frac{28.01u|y_0|}{28.01u + |R(h,\lambda)|}\right);$$

- RK4: $C = 194u$

$$|E_n| \leqslant (194u + |R(h,\lambda)|)^n \left(\varepsilon_0 + n\frac{194u|y_0|}{194u + |R(h,\lambda)|}\right).$$

$\Rightarrow$ no compensation ☹ but reasonable bounds ☺.

Table of contents

General methodology

$$y_{n+1} = \sum_{i=0}^{M} \alpha_i y_n \qquad \widetilde{y_{n+1}} = \oplus_{i=0}^{M} \widetilde{\alpha_i} \otimes \widetilde{y_n} \qquad (\alpha_i = h\lambda, \frac{h\lambda}{3}, \frac{h\lambda}{6}, \frac{h^4\lambda^4}{24} \dots)$$

General methodology

$$y_{n+1} = \sum_{i=0}^{M} \alpha_i y_n \qquad \widetilde{y_{n+1}} = \oplus_{i=0}^{M} \widetilde{\alpha_i} \otimes \widetilde{y_n} \qquad \left(\alpha_i = h\lambda, \frac{h\lambda}{3}, \frac{h\lambda}{6}, \frac{h^4\lambda^4}{24} \dots\right)$$

Methodology to bound roundoff errors:

General methodology

$$y_{n+1} = \sum_{i=0}^{M} \alpha_i y_n \qquad \widetilde{y_{n+1}} = \oplus_{i=0}^{M} \widetilde{\alpha_i} \otimes \widetilde{y_n} \qquad \left(\alpha_i = h\lambda, \frac{h\lambda}{3}, \frac{h\lambda}{6}, \frac{h^4\lambda^4}{24} \dots\right)$$

Methodology to bound roundoff errors:

1) bound $\varepsilon_0 = |\widetilde{y_0} - y_0|$;

General methodology

$$y_{n+1} = \sum_{i=0}^{M} \alpha_i y_n \qquad \widetilde{y_{n+1}} = \oplus_{i=0}^{M} \widetilde{\alpha_i} \otimes \widetilde{y_n} \qquad \left(\alpha_i = h\lambda, \frac{h\lambda}{3}, \frac{h\lambda}{6}, \frac{h^4\lambda^4}{24} \dots\right)$$

Methodology to bound roundoff errors:

1) bound $\varepsilon_0 = |\widetilde{y_0} - y_0|$;
2) bound the error on each term α_i (Gappa + stability);

General methodology

$$y_{n+1} = \sum_{i=0}^{M} \alpha_i y_n \qquad \widetilde{y_{n+1}} = \oplus_{i=0}^{M} \widetilde{\alpha_i} \otimes \widetilde{y_n} \qquad \left(\alpha_i = h\lambda, \frac{h\lambda}{3}, \frac{h\lambda}{6}, \frac{h^4\lambda^4}{24} \dots\right)$$

Methodology to bound roundoff errors:

1) bound $\varepsilon_0 = |\widetilde{y_0} - y_0|$;
2) bound the error on each term α_i (Gappa + stability);
3) bound local errors by M applications of Lemma 2;

General methodology

$$y_{n+1} = \sum_{i=0}^{M} \alpha_i y_n \qquad \widetilde{y_{n+1}} = \oplus_{i=0}^{M} \widetilde{\alpha_i} \otimes \widetilde{y_n} \qquad \left(\alpha_i = h\lambda, \frac{h\lambda}{3}, \frac{h\lambda}{6}, \frac{h^4\lambda^4}{24} \dots\right)$$

Methodology to bound roundoff errors:

1) bound $\varepsilon_0 = |\widetilde{y_0} - y_0|$;
2) bound the error on each term α_i (Gappa + stability);
3) bound local errors by M applications of Lemma 2;
4) bound the global error by instantiating Theorem 1.

Conclusion and perspectives

Conclusion:

- fine and mechanical analysis of roundoff errors;
- results on useful and classical methods (Euler, RK2, RK4);
- linear growth of roundoff errors;
- no compensation (for explicit one-step methods).

Perspectives:

- overflows and underflows;
- formalization in Coq (proof assistant):
  - based on the Flocq library [Boldo-Melquiond],
  - based on the gappa and interval tactics [Melquiond],
- more general ODEs: complex, matricial, non-linear;
- more general methods: multi-step, variable step, implicit.