Erreurs d'arrondi de méthodes d'intégration numérique explicites

Journées FASTRELAX, LIP6 - UPMC

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 des matières

1 Introduction

2 Méthodes et systèmes étudiés

3 Erreurs d'arrondi

4 Conclusion et Perspectives

Équations différentielles ordinaires

Équations différentielles ordinaires (ÉDOs) :

$$F(y, y', y'', \ldots, y^{(n)}, t) = 0.$$

Utilisation : aérospatial, modèles de populations, économie, astrophysique, microbiologie ...

Résolution exacte souvent difficile $\Rightarrow$ méthodes numériques.

Résolution numérique d'ÉDOs

Méthode numérique $\Longrightarrow$ solution approchée discrétisée :

$$y_0,\ y_1, \cdots,\ y_N\ (\forall\ 0 \leq i \leq N,\ y_i\ \simeq\ y(t_0 + ih)).$$

Méthodes explicites :

$$y_n \text{ calculé à partir de } y_{n-1},\ y_{n-2}, \cdots,\ y_0.$$

Implémentation (arithmétique FP) $\Longrightarrow$ deux sources d'erreurs :

- erreur de méthode ;
- erreur d'arrondi.

Analyse d'erreur : un état-de-l'art

Analyse d'erreurs sur méthodes numériques :

- résultat proba. : erreur proportionnelle à $\sqrt{N}$ [Henrici,1963] ;
- en pratique (RK implicite) : proportionnelle à N [Hairer, 2008] ;
- analyse par intervalles [Bouissou, Martel, 2006] ;
- intégration numérique (Newton-Cotes . . .) [Fousse, 2006].

Analyse d'erreur : un état-de-l'art

Analyse d'erreurs sur méthodes numériques :

- résultat proba. : erreur proportionnelle à $\sqrt{N}$ [Henrici,1963] ;
- en pratique (RK implicite) : proportionnelle à N [Hairer, 2008] ;
- analyse par intervalles [Bouissou, Martel, 2006] ;
- intégration numérique (Newton-Cotes . . .) [Fousse, 2006].

Notre approche :

- analyse en grain fin ;
- propriétés mathématiques des schémas (stabilité).

Table des matières

1 Introduction

2 Méthodes et systèmes étudiés

3 Erreurs d'arrondi

4 Conclusion et Perspectives

Méthodes de Runge-Kutta étudiées

Équations différentielles linéaires :

$$\dot{y} = \lambda\, y$$

Méthodes de Runge-Kutta étudiées

Équations différentielles linéaires :

$$\dot{y} = \lambda\, y$$

Méthodes de Runge-Kutta :

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i k_i, \qquad k_i = \lambda\left(t_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} k_j\right)$$

(a_{ij}, b_i, c_i consignés dans tableau de Butcher).

Méthodes de Runge-Kutta étudiées

Équations différentielles linéaires :

$$\dot{y} = \lambda\, y$$

Méthodes de Runge-Kutta :

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i k_i, \qquad k_i = \lambda\left(t_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} k_j\right)$$

(a_{ij}, b_i, c_i consignés dans tableau de Butcher).

Méthode exacte :

Implémentation :

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

Méthodes de Runge-Kutta étudiées

Équations différentielles linéaires :

$$\dot{y} = \lambda\, y$$

Méthodes de Runge-Kutta :

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i k_i, \qquad k_i = \lambda\left(t_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} k_j\right)$$

(a_{ij}, b_i, c_i consignés dans tableau de Butcher).

Méthode exacte :

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

Implémentation :

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

Méthodes de Runge-Kutta explicites classiques

Euler :

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

Méthodes de Runge-Kutta explicites classiques

Euler :

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

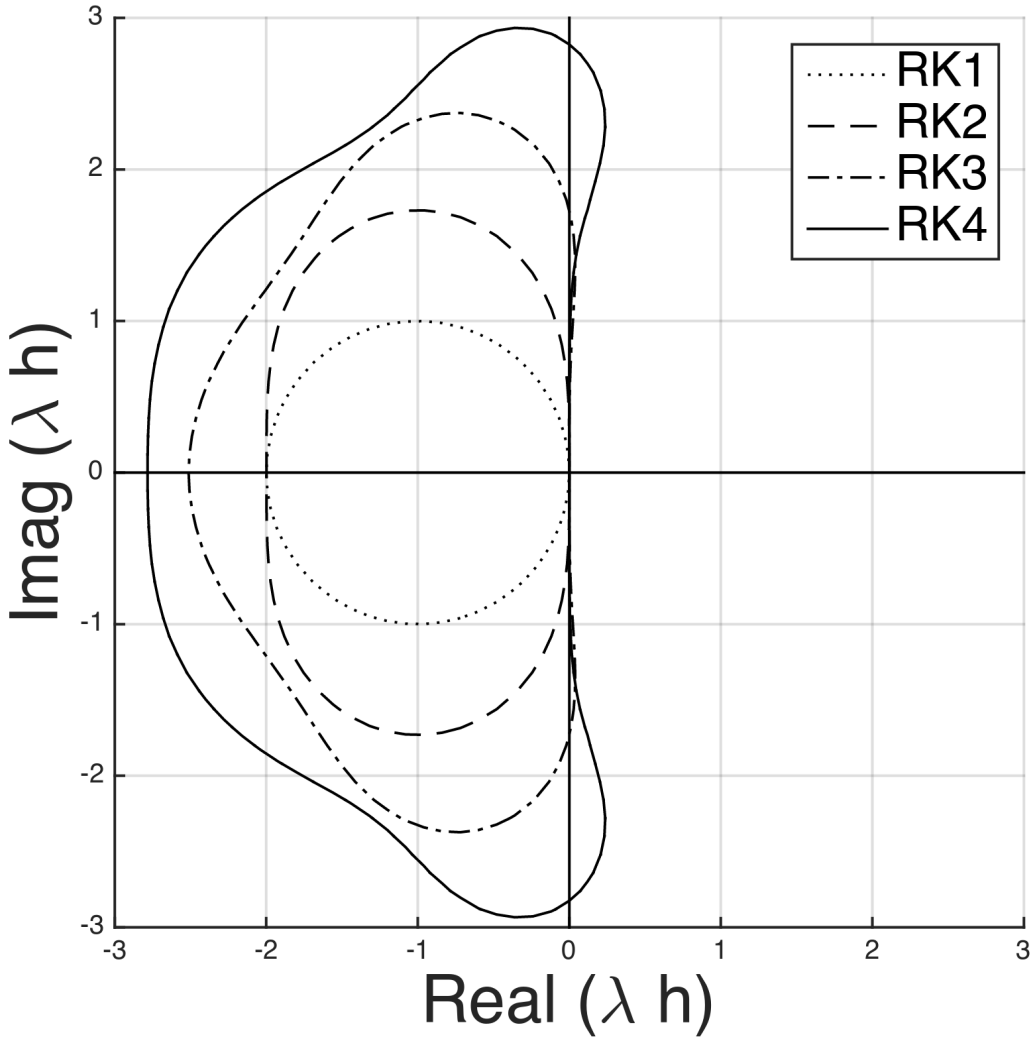
RK4 :

- $R(h,\lambda) = 1 + h\lambda + \frac{1}{2}(h\lambda)^2 + \frac{1}{6}(h\lambda)^3 + \frac{1}{24}(h\lambda)^4$;
- $\widetilde{R}(\widetilde{h},\widetilde{\lambda},\widetilde{y_n}) =$
  - $\left[\widetilde{y_n} + \frac{\widetilde{h}}{6}\widetilde{\lambda}\widetilde{y_n} + \frac{\widetilde{h}}{3}\widetilde{\lambda}\widetilde{y_n} + \frac{\widetilde{h}^2}{6}\widetilde{\lambda}^2\widetilde{y_n} + \frac{\widetilde{h}}{3}\widetilde{\lambda}\widetilde{y_n} + \right.$
    $\left. \frac{\widetilde{h}^2}{6}\widetilde{\lambda}^2\widetilde{y_n} + \frac{\widetilde{h}^3}{12}\widetilde{\lambda}^3\widetilde{y_n} + \frac{\widetilde{h}}{6}\widetilde{\lambda}\widetilde{y_n} + \frac{\widetilde{h}^2}{6}\widetilde{\lambda}^2\widetilde{y_n} + \frac{\widetilde{h}^3}{12}\widetilde{\lambda}^3\widetilde{y_n} + \frac{\widetilde{h}^4}{24}\widetilde{\lambda}^4\widetilde{y_n}\right]$.

(+ de 60 opérations flottantes !)

Stabilité linéaire

Méthode de Runge-Kutta linéaire stable ssi $|R(h,\lambda)| < 1$:



Hypothèses de travail

- ÉDOs ;
  - du premier ordre,
  - linéaires,
  - inconnue à valeur dans $\mathbb{R}$ (unidimensionnel),

Hypothèses de travail

- ÉDOs ;
  - du premier ordre,
  - linéaires,
  - inconnue à valeur dans $\mathbb{R}$ (unidimensionnel),
- Méthodes ;
  - explicites,
  - à un pas constant,

Hypothèses de travail

- ÉDOs ;
  - du premier ordre,
  - linéaires,
  - inconnue à valeur dans $\mathbb{R}$ (unidimensionnel),
- Méthodes ;
  - explicites,
  - à un pas constant,
- Arithmétique à virgule flottante.
  - ni underflow ni overflow,
  - radix 2 double précision ($u = 2^{-53}$),

Table des matières

1 Introduction

2 Méthodes et systèmes étudiés

3 Erreurs d'arrondi

4 Conclusion et Perspectives

Erreurs locales et globales

Erreur locale :

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

Erreur globale :

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

Erreurs locales et globales

Erreur locale :

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

Erreur globale :

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

Théorème 1 : Erreur globale absolue des schémas RK

Soit $C \in \mathbb{R}_+^*$. Supposons que $\forall n \in \mathbb{N}^*, \varepsilon_n \leqslant C|\widetilde{y_{n-1}}|$. Alors pour tout $n \in \mathbb{N}$,

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

Erreurs relatives

Erreur relative :

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

Erreurs relatives

Erreur relative :

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

Si $C \ll |R(h,\lambda)|$, alors :

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

En pratique (Euler, RK2, RK4) : $C \leq 200u$ et $200u \ll |R(h,\lambda)|$.

Erreurs locales

Lemme 2

Soit $y \in \mathbb{R}$. *Soit* $C_1, C_2 \in \mathbb{R}_+$. *Soit* $\alpha_1, \alpha_2 \in \mathbb{R}$. *Soit* $X_1, X_2 \in \mathbb{F}$ *t.q. :*
$|X_1 - \alpha_1 y| \leqslant C_1|y|$ *et* $|X_2 - \alpha_2| \leq C_2$.
Alors :

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

Erreurs locales

Lemme 2

Soit $y \in \mathbb{R}$. *Soit* $C_1, C_2 \in \mathbb{R}_+$. *Soit* $\alpha_1, \alpha_2 \in \mathbb{R}$. *Soit* $X_1, X_2 \in \mathbb{F}$ *t.q. :* $|X_1 - \alpha_1 y| \leqslant C_1|y|$ *et* $|X_2 - \alpha_2| \leq C_2$.
Alors :

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

Exemple (erreur relative de la méthode d'Euler) :

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

Erreurs locales

Lemme 2

Soit $y \in \mathbb{R}$. *Soit* $C_1, C_2 \in \mathbb{R}_+$. *Soit* $\alpha_1, \alpha_2 \in \mathbb{R}$. *Soit* $X_1, X_2 \in \mathbb{F}$ *t.q. :*
$|X_1 - \alpha_1 y| \leqslant C_1|y|$ *et* $|X_2 - \alpha_2| \leq C_2$.
Alors :

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

Exemple (erreur relative de la méthode d'Euler) :

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

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

Erreurs locales de méthodes d'ordre supérieur

Méthode RK4 stable :

FPterm	C	
$\widetilde{y_n}$	0	
$\circ\ \left[\widetilde{h}\widetilde{\frac{1}{6}}\widetilde{\lambda}\right]$	$2u$	$(Gappa)$
$\circ\ \left[\widetilde{y_n}+\widetilde{h}\widetilde{\frac{1}{6}}\widetilde{\lambda}\right]$	$4u$	
$\circ\ \left[\widetilde{h}\widetilde{\frac{1}{3}}\widetilde{\lambda}\right]$	$4u$	$(Gappa)$
$\circ\ \left[\widetilde{h\widetilde{h}}\widetilde{\frac{1}{6}}\widetilde{\lambda\widetilde{\lambda}}\right]$	$12u$	$(Gappa)$
$\circ\ \left[\widetilde{h\widetilde{h}\widetilde{h}}\widetilde{\frac{1}{12}}\widetilde{\lambda\widetilde{\lambda}\widetilde{\lambda}}\right]$	$28u$	$(Gappa)$
$\circ\ \left[\widetilde{h\widetilde{h}\widetilde{h}\widetilde{h}}\widetilde{\frac{1}{24}}\widetilde{\lambda\widetilde{\lambda}\widetilde{\lambda}\widetilde{\lambda}}\right]$	$53u$	$(Gappa)$
$\dots$	$\dots$	
$\circ\ \left[\widetilde{y_n}+\dots+\widetilde{h\widetilde{h}\widetilde{h}\widetilde{h}}\widetilde{\frac{1}{24}}\widetilde{\lambda\widetilde{\lambda}\widetilde{\lambda}\widetilde{\lambda}\widetilde{y_n}}\right]$	$194u$	

Erreurs locales de méthodes classiques

Lemme 3 : Erreur locale du schéma d'Euler

Supposons $-2 \leqslant h\lambda \leqslant -2^{-100}$ *et* $2^{-60} \leqslant h \leqslant 1$. *Alors :*

$\forall n \in \mathbb{N}, \quad \varepsilon_{n+1}^{Euler} \leqslant 11.01u \left|\widetilde{y_n}\right|.$

Lemme 4 : Erreur locale du schéma RK2

Supposons $-2 \leqslant h\lambda \leqslant -2^{-100}$ *et* $2^{-60} \leqslant h \leqslant 1$. *Alors :*

$\forall n \in \mathbb{N}, \quad \varepsilon_{n+1}^{RK2} \leqslant 28.01u \left|\widetilde{y_n}\right|.$

Lemme 5 : Erreur locale du schéma RK4

Supposons $-3 \leqslant h\lambda \leqslant -2^{-100}$ *et* $2^{-60} \leqslant h \leqslant 1$. *Alors :*

$\forall n \in \mathbb{N}, \quad \varepsilon_{n+1}^{RK4} \leqslant 194u \left|\widetilde{y_n}\right|.$

Erreurs globales de méthodes classiques

Bornes sur les erreurs globales de schémas classiques :

- Euler : $C = 11,01u$

$$|E_n| \leqslant (11,01u + |R(h,\lambda)|)^n \left(\varepsilon_0 + n \frac{11,01u|y_0|}{11,01u + |R(h,\lambda)|} \right) \quad ;$$

- RK2 : $C = 28,01u$

$$|E_n| \leqslant (28,01u + |R(h,\lambda)|)^n \left(\varepsilon_0 + n \frac{28,01u|y_0|}{28,01u + |R(h,\lambda)|} \right) \quad ;$$

- RK4 : $C = 194u$

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

$\Rightarrow$ pas de compensation ☹ mais bornes raisonnables ☺.

Table des matières

1 Introduction

2 Méthodes et systèmes étudiés

3 Erreurs d'arrondi

4 Conclusion et Perspectives

Méthodologie générale et bilan

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

Méthodologie générale et bilan

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

Méthodologie pour borner les erreurs d'arrondi :

Méthodologie générale et bilan

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

Méthodologie pour borner les erreurs d'arrondi :

1) calculer l'erreur $\varepsilon_0 = |\widetilde{y_0} - y_0|$;

Méthodologie générale et bilan

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

Méthodologie pour borner les erreurs d'arrondi :

1) calculer l'erreur $\varepsilon_0 = |\widetilde{y_0} - y_0|$;
2) borner les erreurs sur chaque terme α_i (Gappa + stabilité) ;

Méthodologie générale et bilan

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

Méthodologie pour borner les erreurs d'arrondi :

1) calculer l'erreur $\varepsilon_0 = |\widetilde{y_0} - y_0|$;
2) borner les erreurs sur chaque terme α_i (Gappa + stabilité) ;
3) borner l'erreur locale par M applications du Lemme 2 ;

Méthodologie générale et bilan

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

Méthodologie pour borner les erreurs d'arrondi :

1) calculer l'erreur $\varepsilon_0 = |\widetilde{y_0} - y_0|$;
2) borner les erreurs sur chaque terme α_i (Gappa + stabilité) ;
3) borner l'erreur locale par M applications du Lemme 2 ;
4) borner l'erreur globale par instantiation du Théorème 1.

Méthodologie générale et bilan

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

Méthodologie pour borner les erreurs d'arrondi :

1) calculer l'erreur $\varepsilon_0 = |\widetilde{y_0} - y_0|$;
2) borner les erreurs sur chaque terme α_i (Gappa + stabilité) ;
3) borner l'erreur locale par M applications du Lemme 2 ;
4) borner l'erreur globale par instantiation du Théorème 1.

Bilan :

- analyse fine et mécanique des erreurs d'arrondi ;
- pas de compensation (méthodes explicites à un pas) ;
- borne d'erreur linéaire.

Perspectives

- overflows et underflows à prendre en compte ;
- formalisation en Coq (assistant de preuves) :
  - utilisation bibliothèque Flocq [Boldo, Melquiond],
  - utilisation des tactiques Gappa et interval [Melquiond],
  - travail sur l'automatisation des preuves,
- ÉDOs + générales : complexes, matricielles, non linéaires ;
- méthodes + générales : multi-pas, pas variable, implicites ;
- somme erreurs de méthodes et erreurs d'arrondi ;
- équations aux dérivées partielles (éléments finis).