

Cours N°1 - Méthodes à un pas

Méthodes de Runge-Kutta.

M. TASSOT. - 1. SERIES.

Après avoir abordé les méthodes de type Euler dans le cours précédent et les enjeux associés, nous allons dans cette partie du cours aborder de manière plus générale les méthodes dites à un pas et la monter en ordre de manière à être plus performant en terme de coût calcul à précision fixée. Cette classe de méthodes diffère des méthodes appelées méthodes multiples que l'on abordera au prochain cours.

I Généralités:

Nous allons ici proposer une classe de méthodes pour résoudre le système d'équations différentielles ordinaires (EDO) suivant:

$$(1) \begin{cases} dy = f(t, y(t)) & t \in I_0 = [t_0, t_0 + 1] \\ y(t_0) = y_0 \end{cases}$$

avec $y \in \mathbb{R}^m$ et $y_0 \in \mathbb{R}^m$ donnée.

On se donne une discrétisation de l'intervalle I_0 au moyen de $N+1$ points.

$$I_0 = \bigcup_{i \in \{1, \dots, N\}} [t_{i-1}, t_i].$$

Lorsque le pas de discrétisation est constant, on a $t_{n+1} - t_n = \Delta t = \frac{1}{N}$.

On appelle méthode numérique d'intégration de (1) à un pas

une suite $(y_n)_{n \in \{0, \dots, N\}}$ telle que

$$(2) \quad y_{n+1} = y_n + \Delta t_n \phi(t_n, y_n; \Delta t_n)$$

(2)

où (Δt_n) est un vecteur de taille N contenant les divers pas de temps utilisés. Dans le cas d'un pas de temps fixe on a :

$$y_{n+1} = y_n + \Delta t \phi(t_n, y_n; \Delta t).$$

Dans le cadre de la méthode d'Euler explicite, on a

$$\phi(t, y; \Delta t) = f(t, y).$$

et l'on n'a pas de dépendance en Δt .

Pour Euler implicite, si l'on suppose que l'équation :

$$z - \Delta t f(t + \Delta t, z) = y$$

admet une unique solution $z = \psi(t, y; \Delta t)$ alors :

$$\phi(t, y; \Delta t) = f(t + \Delta t, \psi(t, y; \Delta t))$$

Nous dirons que la méthode est explicite si elle ne nécessite l'évaluation de la fonction f qu'en un nombre fini de points.

La méthode (2) est dite à un pas parce qu'elle n'implique aucune information antérieure à t_n et construit l'itéré y_{n+1} au temps $n+1$ à partir de la seule connaissance de y_n au temps t_n .

L'erreur locale de troncature est définie par

$$(3) \quad \varepsilon_n = y(t_{n+1}) - y(t_n) - \Delta t_n \phi(t_n, y(t_n); \Delta t_n).$$

que l'on peut réécrire de manière intéressante :

$$(4) \quad y(t_{n+1}) = y(t_n) + \Delta t_n \phi(t_n, y(t_n); \Delta t_n) + \varepsilon_n$$

La solution exacte est solution d'une méthode à un pas perturbée.

1) Définition: On appelle ordre de la méthode à un pas (2) le plus grand entier p pour lequel

$$(5) \quad y(t+h) - y(t) - \Delta t \phi(t, y(t); \Delta t) = O(\Delta t^{p+1})$$

au sens asymptotique, c'est-à-dire quand $\Delta t \rightarrow 0$
 ceci pour toute fonction f p fois dérivable et toute solution y de (1).

On dit que la méthode est constante si $p \geq 1$.

1.1. Conditions d'ordre. - Prévision.

On définit par récurrence les fonctions $f^{(k)} = f^{(k)}(t, y)$
 par récurrence: (f de classe C^p) y de classe C^{p+1}

$$(6) \quad \begin{cases} f^{(0)}(t, y) = f(t, y) \\ f^{(1)}(t, y) = \partial_t f^{(0)}(t, y) + \partial_y f^{(0)}(t, y) \cdot f(t, y) \\ f^{(2)} \dots \\ f^{(k)}(t, y) = \partial_t f^{(k-1)}(t, y) + \partial_y f^{(k-1)}(t, y) \cdot f(t, y) \end{cases}$$

On vérifiera que, pour $f \in C^k$ et y solution du système d'EDO

$$(7) \quad y^{(k+1)}(t) = f^{(k)}(t, y(t)) = d_t^k f(t, y(t)).$$

Pour le cas autonome $d_t y = f(y)$, on établit:

$$d_t^2 y = (\partial_y f, f)(y), \quad d_t^3 y = (\partial_y^2 f(f, f) + (\partial_y f, \partial_y f, f))(y)$$

(4)

$$d_{\epsilon}^4 y = \left(\partial_y^3 f(f, f, f) + 3 \partial_y^2 f(\partial_y f, f, f) + (\partial_y f \cdot \partial_y^2 f(f, f)) + \right. \\ \left. (\partial_y f, \partial_y f, (\partial_y f, f)) \right)(y).$$

Théorème : f de classe C^p et on suppose que

$(\phi, \partial_{\Delta t} \phi, \dots, \partial_{\Delta t}^p \phi)$ existent et sont continues. Alors une

condition nécessaire et suffisante pour que la méthode à un pas soit d'ordre p ou moins s'écrit :

$$(8) \quad \begin{cases} \phi(t, y; 0) = f(t, y) \\ \partial_{\Delta t} \phi(t, y; 0) = \frac{1}{2} f^{(1)}(t, y) \\ \vdots \\ \partial_{\Delta t}^{p-1} \phi(t, y; 0) = \frac{1}{p!} f^{(p-1)}(t, y). \end{cases}$$

pour tout couple $(t, y) \in \mathbb{I}_0 \times \mathbb{R}^m$.

Preuve : pour y une solution, on pose $E(t) = y(t + \Delta t) - y(t) - \Delta t \phi(t, y(t); \Delta t)$
 puisque y est C^{p+1} on peut effectuer un développement de Taylor :

$$y(t + \Delta t) = y(t) + \sum_{k=0}^{p-1} \frac{\Delta t^{k+1}}{(k+1)!} f^{(k)}(t, y(t)) + \frac{\Delta t^{p+1}}{(p+1)!} y^{(p+1)}(t + \theta \Delta t)$$

$0 < \theta < 1$

$$\phi(t, y(t); \Delta t) = \sum_{k=0}^{p-1} \frac{\Delta t^k}{k!} \partial_{\Delta t}^k \phi(t, y(t); 0) + \frac{\Delta t^p}{p!} \partial_{\Delta t}^p \phi(t, y(t); \sigma \Delta t)$$

$0 < \sigma < 1$.

d'où

$$E(t) = \sum_{k=0}^{p-1} \frac{\Delta t^{k+1}}{k!} \psi_k(t, y) + O(\Delta t^{p+1})$$

avec $\psi_k(t, y) = \frac{1}{k+1} f^{(k)}(t, y) - \partial_{\Delta t}^k \phi(t, y; 0)$

Il est alors clair que les conditions (8) sont suffisantes pour que la méthode soit d'ordre p ou moins. Réciproquement, si elles ne sont pas vérifiées, $\exists$ le plus petit entier tel que $\psi_k(t, y) \neq 0$. et par conséquent

$$\varepsilon(t) = \Delta t^{k+1} \psi_k(t, y(t)) + o(\Delta t^{k+2})$$

Soit (t_0, y_0) un couple tel que $\psi_k(t_0, y_0) \neq 0$ alors

$$\varepsilon(t_0) = o(\Delta t^{k+1}) \text{ ce qui aboutit à une contradiction.}$$

- On peut remarquer qu'il suffit que $\psi_p(t, y) \neq 0$ pour que la méthode soit exactement d'ordre p , c'est-à-dire

$$\partial_{\Delta t}^p \phi(t, y(t); 0) \neq \frac{1}{p+1} f^{(p)}(t, y).$$

- La méthode est constante si et seulement si $\phi(t, y; 0) = f(t, y)$ $\forall (t, y) \in I_0 \times \mathbb{R}^m$ et dans ce cas on voit facilement que

$$\lim_{\Delta t \rightarrow 0} \sum_{n=0}^{N-1} |E_n| = 0 \text{ où } E_n \text{ est l'erreur locale de troncature sous les hypothèses de continuité des fonctions } \phi \text{ et } f.$$

I.2 Stabilité et Convergence.

Après avoir une idée de la précision de la méthode, on voudrait préciser la notion de convergence, c'est-à-dire énoncer des conditions pour que

$$(g) \quad \lim_{\Delta t \rightarrow 0} \max_{0 \leq n \leq N} |e_n| = 0 \text{ voire } |e_n| = o(\Delta t^p);$$

pour cela on va commencer par étudier la stabilité des méthodes à un pas

La notion de stabilité peut être vue comme un résultat de continuité de la solution par rapport aux données et ceci uniformément par rapport au pas de discrétisation.

Si l'on considère le schéma perturbé :

$$(10) \quad z_{n+1} = z_n + \Delta t_n \phi(t_n, z_n; \Delta t_n) + \eta_n$$

DEFINITION : La méthode à un pas (2) est dite stable

s'il existe une constante $\Gamma > 0$ indépendante de Δt_n telle que

$$(11) \quad \max_{0 \leq n \leq N} |y_n - z_n| \leq \Gamma \left(|y_0 - z_0| + \sum_{n=0}^{N-1} |\eta_n| \right).$$

Cela revient à dire que le schéma n'amplifie pas les perturbations ou encore que lorsque la méthode est stable, une "petite" perturbation sur les données du schéma n'entraîne qu'une petite perturbation de la solution, y compris quand $\Delta t \rightarrow 0$.

Remarque : On construit un schéma dans l'optique de l'implémenter et la représentation des nombres réels en machine va forcément induire des erreurs d'arrondi et la propriété de stabilité sera alors essentielle au bon comportement et à la robustesse du schéma.

Le point clé est alors que **STABILITÉ + CONSISTANCE $\Rightarrow$ CONVERGENCE**

THÉORÈME : Si la méthode à un pas (2) est stable et consistante, alors elle converge. Si elle est stable et d'ordre p , alors si f est de classe C^p il existe une constante $K > 0$ indépendante de Δt telle que

$$(12) \quad \max_{0 \leq n \leq N} |e_n| \leq \Gamma \left(|e_0| + K \Delta t^p \right)$$

(7)

Preuve : Dans (10) on choisit $z_n = y(t_n)$ et $\eta_n = \varepsilon_n$
 La propriété de stabilité permet d'obtenir

$$\max_{0 \leq n \leq N} |\varepsilon_n| \leq \pi \left(|y_0 - y(t_0)| + \sum_{n=0}^{N-1} |\varepsilon_n| \right)$$

et comme $y(t_0) = y_0$, on en déduit que, comme par
 conséquence

$$\lim_{\Delta t \rightarrow 0} \sum_{n=0}^{N-1} |\varepsilon_n| = 0 \quad \text{alors} \quad \max_{0 \leq n \leq N} |\varepsilon_n| \rightarrow 0$$

Si la méthode est d'ordre p
 alors

$$|\varepsilon_n| \leq C |\Delta t_n|^{p+1}$$

$$\max_{0 \leq n \leq N} |\varepsilon_n| \leq \pi \left(|e_0| + C \Delta t^p \sum_{n=0}^{N-1} \Delta t_n \right) = \pi \left(|e_0| + C \Delta t^p \right)$$

THÉORÈME : On suppose que la fonction ϕ vérifie la condition de Lipschitz

$$|\phi(t, y; \Delta t) - \phi(t, z; \Delta t)| \leq \lambda |y - z| \quad \begin{array}{l} \forall t \in I_0 \\ \forall (y, z) \in \mathbb{R}^m \\ \forall \Delta t \in [0, \Delta_0] \end{array}$$

où la constante λ est indépendante de Δt . Alors la méthode
 à un pas est stable et $\pi = \exp(\lambda T)$

II. Méthodes de Runge-Kutta.

A partir de ce que nous avons étudié dans la partie précédente une
 première possibilité serait de prendre

$$\begin{aligned} \phi(t, y; \Delta t) = & f(t, y) + \frac{\Delta t}{2} f^{(1)}(t, y) + \frac{\Delta t^2}{2!} f^{(2)}(t, y) \dots \\ & + \frac{\Delta t^{p-1}}{p!} f^{(p-1)}(t, y) \end{aligned}$$

pour obtenir une méthode d'ordre $p > 1$ car la condition (8) est alors
 directement satisfaite

L'exemple puis précédemment sur le calcul des $f^{(k)}$ montrera au lecteur que la combinatoire exploise pour des ordres élevés et qu'il ne sera pas intéressant numériquement de suivre cette stratégie.

On va plutôt chercher à identifier une classe de méthodes à un peu ne reposant que sur un certain nombre d'évaluations de la fonction f , que sont les méthodes de Runge-Kutta.

On se rappelle que la solution exacte vérifie

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t', y(t')) dt'$$

et on va chercher à utiliser une famille de quadrature plus ou moins précise.

On se donne $c_1, \dots, c_q$, q nombres réels ordonnés compris entre 0 et 1 et on considère les formules de quadrature

$$\left| \begin{aligned} \int_0^{c_i} \varphi(z) dz &\approx \sum_{j=1}^q a_{ij} \varphi(z_j) \\ \int_0^1 \varphi(t) dz &\approx \sum_{j=1}^q b_j \varphi(z_j). \end{aligned} \right.$$

En introduisant les instant intermédiaires $t_{nj} = t_n + z_j \Delta t_n$ $1 \leq j \leq q$ et au moyen du changement de variable $t = t_n + z \Delta t_n$ on arrive à

$$y(t_{nj}) = y(t_n) + \int_{t_n}^{t_{nj}} f(t', y(t')) dt' \approx y(t_n) + \Delta t_n \sum_{j=1}^q a_{ij} f(t_{nj}, y(t_{nj}))$$

$$\text{et } y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t', y(t')) dt' = y(t_n) + \Delta t_n \sum_{j=1}^q b_j f(t_{nj}, y(t_{nj}))$$

d'où la méthode:

$$(13) \quad \begin{cases} y_{ni} = y_n + \Delta t_n \sum_{j=1}^q a_{ij} f(t_{nj}, y_{nj}) \\ y_{n+1} = y_n + \Delta t_n \sum_{j=1}^q b_j f(t_{nj}, y_{nj}) \end{cases}$$

(9)

Une forme équivalente, mieux adaptée à une mise en œuvre informatique s'est écrit :

On pose $k_{ni} = f(t_{ni}, y_{ni}) \quad 1 \leq i \leq q.$

et $(14) \quad \begin{cases} k_{ni} = f(t_{ni}, y_n + \Delta t_n \sum_{j=1}^q a_{ij} k_{nj}) \\ y_{n+1} = y_n + \Delta t_n \sum_{j=1}^q b_j k_{nj} \end{cases}$

(On peut remonter à la formulation précédente en posant $y_{ni} = y_n + \Delta t_n \sum_{j=1}^q a_{ij} k_{nj}$)

La première des deux équations de (13) ou (14) constitue un système non linéaire de q équations en les q inconnues k_{ni} ou y_{ni} .

On verra que pour Δt assez petit, ce système admet une unique solution.

En notant $k_i = k_i(t, y; \Delta t) \quad 1 \leq i \leq q$ la solution de

$$k_i = f\left(t + c_i \Delta t, y + \Delta t \sum_{j=1}^q a_{ij} k_j\right) \quad 1 \leq i \leq q$$

on trouve alors

$$\phi(t, y; \Delta t) = \sum_{j=1}^q b_j k_j(t, y; \Delta t)$$

qui est bien la forme des méthodes à 1 pas.

On représente la méthode en général par un tableau.

c_1	a_{11}	a_{12}	...	a_{1q}
c_2	a_{21}	a_{22}	...	a_{2q}
$\vdots$	$\vdots$			$\vdots$
c_q	a_{q1}	a_{q2}	...	a_{qq}
	b_1	b_2	...	b_q

Si la matrice de (a_{ij}) est carrée modulo une permutation est triangulaire strictement inférieure alors la méthode est dite explicite.

Pour les méthodes explicites, on se placera souvent dans le cadre où $C_1 = 0$ (on utilise l'information au point précédent comme au point de quadrature) et $C_i = \sum_{j=1}^{i-1} a_{ij}$. D'une part cela permet d'assurer que

$$k_{n,i} = f(t_n + C_i \Delta t_n, y(t_n + C_i \Delta t_n)) + O(\Delta t^2)$$

↳ on approche effectivement une évaluation de f au point $t_n + C_i \Delta t_n$ sur la solution exacte, et d'autre part permet de se limiter aux systèmes autonomes.

$$\text{En effet} \quad \left\{ \begin{array}{l} d_t y = f(t, y(t)) \Leftrightarrow d_t Y = \begin{pmatrix} 1 \\ f(t, y(t)) \end{pmatrix} = F(Y) \\ y(t_0) = y_0 \end{array} \right. \quad \left\{ \begin{array}{l} d_t Y = F(Y) \\ Y = \begin{pmatrix} t \\ y(t) \end{pmatrix} \quad Y(t_0) = \begin{pmatrix} t_0 \\ y_0 \end{pmatrix} \end{array} \right.$$

et le lecteur vérifiera alors que les méthodes de Runge Kutta sur les deux systèmes sont équivalentes.

II.1 Conditions d'ordre.

- Cas des méthodes explicites avec $C_1 = 0$ et $\sum_{j=1}^{i-1} a_{ij} = C_i$.

Dans ce cas, on va exprimer l'égalité du développement en série de Taylor de la solution exacte et de la solution numérique pour reproduire les égalités de (8) jusqu'à l'ordre 4.

On rappelle que: $d_t y = f(y)$.

$$\begin{aligned} (15) \quad y(t_0 + \Delta t) &= y(t_0) + \Delta t f_0 + \frac{\Delta t^2}{2!} (d_y f)_0 + \frac{\Delta t^3}{3!} (d_y^2 f(f, f)_0 + d_y f d_y f f)_0 \\ &+ \frac{\Delta t^4}{4!} (d_y^3 f(f, f, f) + 3 d_y^2 f(d_y f f, f) + d_y f d_y^2 f(f, f) + d_y f d_y f d_y f f)_0 \\ &+ O(\Delta t^5). \end{aligned}$$

où la notation $()_0$ signifie évalué en $y_0 = y(t_0)$.

(11)

Pour la solution numérique.

$$k_{ni} = f\left(y_n + \Delta t \sum_{j=1}^{i-1} a_{ij} k_{nj}\right)$$

Pour le calcul on choisit $n=0$ et on note $k_i = f\left(y_0 + \Delta t \sum_{j=1}^{i-1} a_{ij} k_j\right)$
par conséquent

$$k_i = f_0 + c_i \Delta t \left(\partial_y f f\right)_0 + o(\Delta t^2)$$

Et en itérant:

$$k_i = f\left(y_0 + \Delta t c_i f_0 + \Delta t^2 \sum_j a_{ij} c_j \left(\partial_y f f\right)_0 + o(\Delta t^3)\right)$$

$$k_i = f_0 + c_i \Delta t \left(\partial_y f f\right)_0 + \Delta t^2 \sum_j a_{ij} c_j \left(\partial_y^2 f f\right)_0 + \frac{\Delta t^2}{2} c_i^2 \left(\partial_y^2 f(f, f)\right)_0 + o(h^3)$$

et en insérant dans la formule de y_1 après encore une itération:

$$\begin{aligned} (16) \quad y_1 = & y_0 + \Delta t \left(\sum_i b_i\right) f_0 + \frac{\Delta t^2}{2!} \left(2 \sum_i b_i c_i \left(\partial_y f f\right)_0\right) \\ & + \frac{\Delta t^3}{3!} \left(\left(3 \sum_i b_i c_i^2\right) \left(\partial_y^2 f(f, f)\right)_0 + \left(6 \sum_{ij} b_i a_{ij} c_j\right) \left(\partial_y f \partial_y f f\right)_0 \right) \\ & + \frac{\Delta t^4}{4!} \left(\left(4 \sum_i b_i c_i^3\right) \left(\partial_y^3 f(f, f, f)\right)_0 + \left(8 \sum_{ij} b_i c_i a_{ij} c_j\right) 3 \left(\partial_y^2 f \left(\partial_y f f, f\right)\right)_0 \right. \\ & + \left. \left(12 \sum_{ij} b_i a_{ij} c_j^2\right) \left(\partial_y f \partial_y^2 f(f, f)\right)_0 + \left(24 \sum_{ijk} b_i a_{ij} a_{jk} c_k\right) \left(\partial_y f \partial_y f \partial_y f f\right)_0 \right) \\ & + o(\Delta t^5) \end{aligned}$$

On va donc pouvoir écrire des conditions d'ordre sur la méthode en comparant terme à terme les développements.

THEOREME. La méthode de Runge Kutta
4 si les coefficients satisfont

a un ordre de

$$(17) \quad C_1 = 0 \quad \sum_{j=1}^{i-1} d_{ij} = C_i$$

et si les conditions suivantes sont satisfaites.

$$(18) \quad \sum_i b_i = 1$$

$$(19) \quad \sum_i b_i C_i = 1/2$$

$$(20) \quad \sum_i b_i C_i^2 = 1/3$$

$$(21) \quad \sum_{ij} b_i a_{ij} C_j = 1/6$$

$$(22) \quad \sum_i b_i C_i^3 = 1/4$$

$$(23) \quad \sum_{ij} b_i C_i a_{ij} C_j = 1/8$$

$$(24) \quad \sum_{ij} b_i a_{ij} C_j^2 = 1/12$$

$$(25) \quad \sum_{ijk} b_i a_{ij} a_{jk} C_k = 1/24$$

$$b_1 + b_2 + b_3 + b_4 = 1$$

$$b_2 C_2 + b_3 C_3 + b_4 C_4 = 1/2$$

$$b_2 C_2^2 + b_3 C_3^2 + b_4 C_4^2 = 1/3$$

$$b_3 a_{32} C_2 + b_4 (a_{42} C_2 + a_{43} C_3) = 1/6$$

$$b_2 C_2^3 + b_3 C_3^3 + b_4 C_4^3 = 1/4$$

$$b_3 C_3 a_{32} C_2 + b_4 (a_{42} C_2 + a_{43} C_3) C_2 = 1/8$$

$$b_3 a_{32} C_2^2 + b_4 (a_{42} C_2^2 + a_{43} C_3^2) = 1/12$$

expression pour $q=4$
4 étages

$$(18) \Rightarrow \text{ordre 1}$$

$$(18+19) \Rightarrow \text{ordre 2}$$

$$(18 \rightarrow 21) \Rightarrow \text{ordre 3}$$

} Remarque.

Algorithme de construction de méthodes d'ordre 4 à 4 étages:

① On se donne les C_i avec $C_1 = 0$, $C_4 = 1$ et on garde deux degrés de liberté.

② A l'aide de (18) (19) (20) et (22) calculer b_1, b_2, b_3 et b_4 pour que la famille de quadratures soit d'ordre 4.

(3) Selon une idée de J. BUTCHER, on peut montrer que l'on peut remplacer (21) (24) et (25) par.

$$\begin{cases} (26) & b_2 (1 - c_2) = b_3 a_{32} + b_4 a_{42} \\ (27) & b_3 (1 - c_3) = b_4 a_{43} \\ (28) & b_4 (1 - c_4) = 0 \end{cases}$$

(27) permet alors de calculer a_{43} , (23) + (26) donne un système linéaire pour obtenir a_{42} et a_{32} et la condition (17) donne a_{21} , a_{31} et a_{41} .

Deux méthodes célèbres relèvent de cette construction :

"La" méthode de Runge-Kutta 4

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	2/6	2/6	1/6

La méthode dite règle 3/8

0				
1/3	1/3			
2/3	-1/3	1		
1	1	-1	1	
	1/8	3/8	3/8	1/8

Les méthodes d'Euler, de Runge et de Heun d'ordre respectif 1, 2 et 3 sont données par :

0	
1	

0		
1/2	1/2	
	0	1

0			
1/2	1/3		
2/3	0	2/3	
	1/4	0	3/4

Le lecteur vérifiera que ces méthodes satisfont les conditions d'ordre.

Exemple : α un paramètre $\in]0, 1]$

$$\begin{array}{ccc} 0 & & \\ \alpha & \alpha & \\ & 1 - \frac{1}{2\alpha} & \frac{1}{2\alpha} \end{array}$$

$$(29) \begin{cases} k_{n1} = f(t_n, y_n) \\ k_{n2} = f(t_n + \alpha \Delta t_n, y_n + \alpha \Delta t_n k_{n1}) \\ y_{n+1} = y_n + \Delta t_n \left(\left(1 - \frac{1}{2\alpha}\right) k_{n1} + \frac{1}{2\alpha} k_{n2} \right) \end{cases}$$

ou encore : $y_{n+\alpha} = y_n + \alpha \Delta t_n f(t_n, y_n)$

$$y_{n+1} = y_n + \Delta t_n \left\{ \left(1 - \frac{1}{2\alpha}\right) f(t_n, y_n) + \frac{1}{2\alpha} f(t_{n+\alpha}, y_{n+\alpha}) \right\}$$

avec $t_{n+\alpha} = t_n + \alpha \Delta t_n$. Pour $\alpha = \frac{1}{2}$ on retrouve la méthode dite d'Euler modifiée et pour $\alpha = 1$ la méthode de Heun.

Exemple : α un paramètre $\neq 0, \alpha > 0$.

$$c_1 = \frac{1}{2} \left(1 + \frac{1}{\alpha \sqrt{3}} \right) \quad c_2 = \frac{1}{2} \left(1 - \frac{\alpha}{\sqrt{3}} \right)$$

$c_1 > c_2$ pour $\alpha > 0$
la méthode définie par :

$$(30) \begin{cases} k_{n1} = f(t_n + c_1 \Delta t_n, y_n + c_1 \Delta t_n k_{n1}) \\ k_{n2} = f(t_n + c_2 \Delta t_n, y_n + \Delta t_n ((2c_2 - 1)k_{n1} + (1 - c_2)k_{n2})) \\ y_{n+1} = y_n + \frac{\Delta t_n}{1 + \alpha^2} (\alpha^2 k_{n1} + k_{n2}) \end{cases}$$

$$\begin{array}{ccc} c_1 & c_2 & 0 \\ c_2 & 2c_2 - 1 & 1 - c_2 \\ & \frac{\alpha^2}{1 + \alpha^2} & \frac{1}{1 + \alpha^2} \end{array}$$

est une méthode semi-implicite (triangulaire inférieure) d'ordre 3 à 2 étages

Exercice : le vérifier avec les conditions d'ordre (18)-(21)

• Exemple : θ -méthode

On choisit $\theta \in [0, 1]$ un nombre réel. La θ -méthode est définie par.

$$(31) \quad y_{n+1} = y_n + \Delta t_n \left((1-\theta) f(t_n, y_n) + \theta f(t_{n+1}, y_{n+1}) \right)$$

ou sa variante

$$(32) \quad \begin{cases} y_{n+1} = y_n + \Delta t_n f(t_{n+\theta}, y_{n+\theta}) \\ t_{n+\theta} = t_n + \theta \Delta t_n \quad y_{n+\theta} = \theta y_{n+1} + (1-\theta) y_n. \end{cases}$$

Chacune de ces méthodes rentre dans le cadre des méthodes de Runge Kutta.

$$\begin{cases} k_{n1} = f(t_n, y_n) \\ k_{n2} = f(t_{n+1}, y_n + \Delta t_n ((1-\theta) k_{n1} + \theta k_{n2})) \\ y_{n+1} = y_n + \Delta t_n ((1-\theta) k_{n1} + \theta k_{n2}) \end{cases}$$

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

$$\begin{cases} k_{n+\theta} = f(t_{n+\theta}, y_n + \theta \Delta t_n k_{n+\theta}) \\ y_{n+1} = y_n + \Delta t_n k_{n+\theta} \end{cases}$$

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

pour $\theta=0$, les deux méthodes donnent Euler-Forward et pour $\theta=1$ avec Euler impléité. Elles sont d'ordre 2 pour $\theta = \frac{1}{2}$.

Remarque : On a longtemps eu que l'utilisation de méthodes de Runge Kutta totalement impléites, du fait du système d'équations non linéaires à résoudre, ne pourrait mener à des classes de méthodes compétitives et exploitables en pratique. Nous verrons au prochain cours que les méthodes de type **ROUNUS** introduites par **HAIREN** et **WANNER** ont récemment prouvé le contraire.

II.2 Stabilité: introduction aux concepts

Nous définissons dans un premier temps la notion de domaine de stabilité absolue d'une méthode de Runge-Kutta.

Pour cela nous travaillons avec l'équation différentielle scalaire linéaire appelée équation test de Dahlquist

$$(33) \quad dy = \lambda y \quad \lambda \in \mathbb{C}$$

(où λ représente une valeur propre typique d'une matrice Jacobienne.)

En appliquant la méthode de RK à (33) on obtient:

$$(34) \quad \begin{cases} y_{ni} = y_n + \lambda \Delta t_n \sum_{j=1}^q a_{ij} y_{nj} & 1 \leq i \leq q \\ y_{n+1} = y_n + \lambda \Delta t_n \sum_{j=1}^q b_j y_{nj} \end{cases}$$

On peut écrire cela sous la forme vectorielle

$$y = y_n \underline{e} + \lambda \Delta t_n H y$$

avec $\underline{e} = (1, \dots, 1)^t$, $A = (a_{ij})_{i,j \in \{1, \dots, q\}} \in M_q(\mathbb{R})$.

$$y = (y_{n1} - y_{nq})^t \in \mathbb{R}^q$$

$$y_{n+1} = y_n + \lambda \Delta t_n \underline{b} \cdot y, \quad \underline{b} = (b_1, \dots, b_q)^t$$

Si l'on pose on a
$$r(z) = 1 + z \underline{b}^t (\mathbf{I} - zH)^{-1} \underline{e} \quad (35)$$

$$y_{n+1} = r(\lambda \Delta t_n) y_n$$

où $r(z)$ est appelé la fonction de stabilité de la méthode.

On voit que c'est une fraction rationnelle quotient de deux polynômes de degré $\leq q$, le dénominateur pouvant être pris égal à $\det(\mathbf{I} - zA)$

Lorsque la méthode est explícite, $\det(\mathbf{I} - zA) = 1$ et $r(z)$ est un polynôme de degré $\leq q$.

Exemple de polynôme de stabilité :

Euler impléite: $r(z) = \frac{1}{1-z}$

0 schema: $\frac{1+z(1-\theta)}{1-z\theta}$

RADAU II ordre 5:

$$\frac{1 + \frac{2}{5}z + \frac{z^2}{20}}{1 - \frac{3}{5}z + \frac{3z^2}{20} - \frac{z^3}{60}}$$

Euler explicite $1+z$.

LEMME

Si la méthode de RK est d'ordre p alors.

$$r(z) = e^z + C z^{p+1} + O(z^{p+2})$$

$r(z)$ est une approximation rationnelle de e^z d'ordre p et C est la constante d'erreur.

Preuve: pour (33) on a $\phi(t, y; \Delta t) = \frac{r(\lambda \Delta t) - 1}{\Delta t} y$.

et puisque la solution est $y(t) = y_0 e^{\lambda t}$

$$\exp(\lambda(\Delta t + t)) - \exp(\lambda t) - \left(\frac{r(\lambda \Delta t) - 1}{\Delta t} \right) \exp(\lambda t) = O(\Delta t^{p+1})$$

soit

$$\exp(\lambda \Delta t) - r(\lambda \Delta t) = O(\Delta t^{p+1})$$

■

En supposant que Δt_n est constant égal à Δt , $y_n = r(\lambda \Delta t)^n y_0$
et la solution reste bornée si $|r(\lambda \Delta t)| \leq 1$ quand $n \rightarrow +\infty$.

DEFINITION: On appelle domaine de stabilité absolue de la méthode RK l'ensemble

$$D = \{ z \in \mathbb{C} ; |r(z)| \leq 1 \}.$$

où $\mathbb{C}$ désigne le plan complexe complet par un point à l'infini ∞
L'intérieur de D noté $\overset{\circ}{D}$ est caractérisé par.

$$\overset{\circ}{D} = \{ z \in \mathbb{C} ; |r(z)| < 1 \}.$$

Lorsque le système d'équations est raide, un critère de stabilité heuristique consiste à évaluer le spectre de la matrice jacobienne du second membre en y $\partial_y f(t, y) = \left(\frac{\partial f_i}{\partial y_j} \right)_{1 \leq i, j \leq m} (t, y)$

et à choisir un pas de temps Δt_n tel que $\forall j=1, \dots, m$
 $\lambda_{ij}(t_n, y_n) \Delta t_n \in \mathcal{D}$.

Pour ce type de système différentiel raide, il ne s'agit que d'une règle qui n'est certainement pas une condition nécessaire et/ou suffisante, mais elle est simple d'utilisation et fonctionne plutôt bien.

On se référera à la PCH pour avoir une représentation graphique des principaux domaines de stabilité absolue pour les méthodes de RK explicites.

On aimerait bien pour des systèmes particulièrement raides ou pour lesquels on ne sait pas forcément évaluer la raideur à priori, disposer d'un domaine de stabilité qui contient le demi-plan complexe $\mathbb{C}^-$ des nombres imaginaires de partie réelle négative ou nulle.

DEFINITION: On dit qu'une méthode de RK est A-Stable si $\{z \in \mathbb{C}; \operatorname{Re} z \leq 0 \text{ ou } z = \infty\} \subset \mathcal{D}$.

On dit parfois qu'elle est fortement A-stable si le point à l'infini appartient à $\mathring{\mathcal{D}}$ intérieur de $\mathcal{D}$.

Remarque: En utilisant le principe du maximum sur $r(z)$ on peut montrer que la méthode est A-stablessi

$$\{z \in \mathbb{C}, \operatorname{Re} z < 0\} \subset \mathring{\mathcal{D}}$$

Remarque : Notons qu'une méthode de Runge Kutta explicite n'est jamais A-stable, puisque $r(z)$ est un polynôme et le domaine de stabilité absolue est forcément borné.

Il est important de bien réaliser que la notion de stabilité absolue et la A-stabilité concerne l'équation test de Dahlquist et se révèle en général insuffisante pour traiter de systèmes différentiels non linéaires. On doit alors introduire une notion un peu plus générale : la B-stabilité.

On va proposer une propriété à satisfaire pour une méthode de RK sur un ensemble de système non linéaires qui possède une propriété de "contraction" au niveau continu. Plus précisément, supposons f continue de $I_0 \times \mathbb{C}^m \rightarrow \mathbb{C}^m$ et :

$$(36) \quad \operatorname{Re} (f(t, y) - f(t, z), y - z) \leq 0 \quad \forall t \in I_0 \quad \forall y, z \in \mathbb{C}^m$$

pour un produit scalaire convenable dans $\mathbb{C}^m$. Alors pour deux solutions du système différentiel : de $y = f(t, y(t))$ on peut montrer que

$$(37) \quad |y(t) - z(t)| \leq |y(s) - z(s)| \quad \forall s, t \in I_0 \quad s \leq t$$

On va donc essayer de définir la B-stabilité comme l'analogue discret de cette propriété.

Pour ce faire on part de $z_n \neq y_n$ et en appliquant le schéma de RK :

$$(38) \quad \begin{cases} z_{n,i} = z_n + \Delta t_n \sum_{j=1}^q a_{ij} f(t_{n,j}, z_{n,j}) & 1 \leq i \leq q. \\ z_{n+1} = z_n + \Delta t_n \sum_{j=1}^q b_j f(t_{n,j}, z_{n,j}) \end{cases}$$

on obtient la définition :

DEFINITION : On dit que la méthode RK est B-stable si pour toute fonction vérifiant (36) et tout couple $(y_n, z_n) \in \mathbb{C}^m \times \mathbb{C}^m$ on a

$$|y_{n+1} - z_{n+1}| \leq |y_n - z_n|$$

où y_{n+1} et z_{n+1} sont données par (34) et (38).

Remarque : Le lecteur vérifiera que pour $m=1$ et $f(t, y) = -\lambda y$ on obtient qu'une méthode de RK B-stable est forcément A-stable.

THEOREME : On suppose que la fonction f vérifie (36) et que la méthode de RK est B-stable. Alors on a l'estimation d'erreur globale.

$$(39) \quad |e_n| \leq |e_0| + \sum_{p=0}^{n-1} |\varepsilon_p|$$

où ε_p est l'erreur locale de troncature.

Preuve : En utilisant la solution exacte

$$y(t_{n+1}) = y(t_n) + \Delta t_n \sum_{j=1}^q b_j f(t_{nj}, z_{nj}) + \varepsilon_n.$$

et en choisissant $z_n = y(t_n)$ et $z_{n+1} = y(t_{n+1}) - \varepsilon_n$

puis en appliquant la B-stabilité :

$$\begin{aligned} |z_{n+1} - y_{n+1}| &= |y(t_{n+1}) - \varepsilon_n - y_{n+1}| = |e_{n+1} - \varepsilon_n| \\ &\leq |y(t_n) - y_n| = |e_n| \end{aligned}$$

c'est à dire :

$$|e_{n+1}| \leq |e_{n+1} - \varepsilon_n| + |\varepsilon_n| \leq |e_n| + |\varepsilon_n|$$

□

Remarque : L'estimation obtenue est très générale mais dans certains cas particuliers, à partir de (39) on peut obtenir des estimations d'erreur plus fines.

On peut énoncer un critère de B-stabilité qui peut être utile dans un certain nombre de cas.

Si l'on note $B = \text{diag}(b_i)$ et $C = BA + A^t B - b \cdot b^t$ on a le théorème suivant.

THEOREME. Si les matrices B et C sont semi définies positives alors la méthode de RK est B-stable.

Éléments de preuve:

Notons $\varphi_i = \Delta t_n (f(t_n, y_n) - f(t_n, z_n))$ de manière à ce que

$$y_n - z_n = y_n - z_n + \sum_{j=1}^q a_{ij} \varphi_j$$

par ailleurs,

$$y_{n+1} - z_{n+1} = y_n + z_n + \sum_{j=1}^q b_j \varphi_j \quad \text{et alors}$$

$$|y_{n+1} - z_{n+1}|^2 = |y_n - z_n|^2 + \left| \sum_{j=1}^q b_j \varphi_j \right|^2 + 2 \operatorname{Re} \sum_{j=1}^q b_j (y_n - z_n, \varphi_j)$$

On remarque

$$(y_n - z_n, \varphi_i) = (y_n - z_n, \varphi_i) - \sum_{j=1}^q a_{ij} (\varphi_j, \varphi_i)$$

et donc

$$\begin{aligned} |y_{n+1} - z_{n+1}|^2 &= |y_n - z_n|^2 + \left| \sum_{j=1}^q b_j \varphi_j \right|^2 - 2 \operatorname{Re} \sum_{i,j=1}^q b_i a_{ij} (\varphi_i, \varphi_j) \\ &\quad + 2 \operatorname{Re} \sum_{j=1}^q b_j (y_n - z_n, \varphi_j) \end{aligned}$$

(semi définies positives implique)

$$\left| \sum_{j=1}^q b_j \varphi_j \right|^2 - 2 \operatorname{Re} \sum_{i,j=1}^q b_i a_{ij} (\varphi_i, \varphi_j) = - \sum_{i,j=1}^q C_{ij} (\varphi_i, \varphi_j) \leq 0$$

Pour conclure: $b_i \geq 0$. et donc

$$2 \operatorname{Re} \sum_{j=1}^q b_j (y_n - z_n, \varphi_j) = 2 \Delta t_n \sum_{j=1}^q b_j \operatorname{Re} (f(t_n, y_n) - f(t_n, z_n), \varphi_j)$$

$$y_n - z_n \leq 0$$

$$\text{d'où } |y_{n+1} - z_{n+1}|^2 \leq |y_n - z_n|^2$$

Remarque: Il est possible de montrer que les hypothèses du précédent théorème sont des conditions nécessaires et suffisantes, dès que les C_j sont distincts.

THEOREME Si f est continue et vérifie la condition (36), alors si la méthode de RK est semi-implé (matrice (a_{ij}) triangulaire inférieure avec des coefficients non nuls sur la diagonale) avec $a_{ii} > 0 \quad 1 \leq i \leq q$ alors les équations (34) définissent y de manière unique.

Revenons sur les méthodes présentées en pages (14) et (15).

• Par le bas de la page (14).

$$A = \begin{pmatrix} C_1 & 0 \\ 2C_2 - 1 & 1 - C_2 \end{pmatrix} \quad B = \frac{1}{1 + \alpha^2} \begin{pmatrix} \alpha^2 & 0 \\ 0 & 1 \end{pmatrix}$$

$$C = \frac{1}{(1 + \alpha^2)^2} \begin{pmatrix} \eta_1 & -\eta_2 \\ -\eta_2 & \eta_2 \end{pmatrix} \quad \text{avec} \quad \begin{cases} \eta_1 = 2\alpha^2 C_1 + \alpha^4 (2C_1 - 1) \\ \eta_2 = 1 - 2C_2 + 2\alpha^2 (1 - C_2) \end{cases}$$

On montre aisément que C est semi-définie positive pour $\alpha > 0$.
La remarque en haut de cette page, comme $C_1 \neq C_2$ permet de conclure que cette méthode est B-stable si $\alpha > 0$.

• 2 méthode - Page (15).

Pour ces deux méthode $r(z) = \frac{1 - (1 - \theta)z}{1 + \theta z}$.

pour $0 \leq \theta \leq \frac{1}{2}$ le domaine de stabilité absolue est le disque du plan complexe de centre $z = \frac{1}{1 - 2\theta}$ et de rayon $\frac{1}{1 - 2\theta}$.

La méthode est A-stable pour $\frac{1}{2} \leq \theta \leq 1$ pour (31) et (32).

Par contre concernant la B-stabilité, les deux diffèrent.
 Pour (31), on a $q=2$ et.

$$A = \begin{pmatrix} 0 & 0 \\ 1-\theta & 0 \end{pmatrix} \quad b = \begin{pmatrix} 1-\theta \\ 0 \end{pmatrix}$$

et

$$C = \begin{pmatrix} -(1-\theta)^2 & 0 \\ 0 & \theta^2 \end{pmatrix} \text{ n'est semi-définie positive}$$

que si $\theta = 1$, c'est à dire pour Euler impléte!

Pour (32). $q=1$, $A=\theta$, $b=1$ et $C=2\theta-1$

La méthode est B-stable dès que $\theta \geq \frac{1}{2}$

Par conséquent pour des systèmes différentiels non linéaires
 rigides, (32) sera préféré à (31)!

Il est important de noter pour ce dernier exemple
 le fait que deux méthodes, ayant le même polynôme
 de stabilité et donc la même conclusion de A-stabilité
 peuvent avoir des comportements très différents dans
 le cadre de la résolution de systèmes non linéaires
 pour lesquels la B-stabilité est essentielle.

Notons aussi que la méthode RADAUS étudiée
 en cours est A-stable et B-stable. (Voir les
 références de Hairer et Wanner à ce sujet).

L-stabilité: Dans le cadre des systèmes raides, on apprécie que la fonction de stabilité aie un comportement spécifique en ∞ .

Pour l'exemple, si la région de stabilité coïncide exactement avec le $\frac{1}{2}$ plan complexe $\mathbb{C}^-$, on a une méthode $\mathcal{H}$ -stable mais connue

$$\lim_{z \rightarrow -\infty} r(z) = \lim_{z \rightarrow -\infty} |r(z)| = \lim_{z=iy, y \in \mathbb{R}} |r(z)|$$

cette dernière doit valoir 1 en module car $|r(iy)| = 1$ par la forme du domaine de stabilité.

Par conséquent pour z réel négatif et très grand $|r(z)| < 1$ mais très proche de 1. Dans ce cadre les composantes raides ne décroissent vers zéro que très lentement.

Dans le cas d'Euler implémente, au contraire, ces composantes sont amorties immédiatement.

On définit alors la notion de L-stabilité:

DEFINITION. (Ehle 1969) une méthode est dite L-stable si elle est $\mathcal{H}$ -stable et

$$\lim_{z \rightarrow -\infty} r(z) = 0 \text{ aussi noté } r(\infty) = 0$$

PROPOSITION: Si une méthode de Runge-Kutta avec $\mathcal{H}$ non singulière satisfait

$$a_{qj} = b_j \quad j = 1, \dots, q$$

alors $r(\infty) = 0$ et si la méthode était $\mathcal{H}$ -stable elle demeure L-stable.

Preuve: Il est aisé de montrer que $r(\infty) = 1 - \underline{b}^T A^{-1} \underline{e}$
 la condition ci-dessus s'écrit: $\mathcal{H}^T \underline{\psi}_s = \underline{b}$ avec $\underline{\psi}_s = (0, \dots, 0, 1)^T \in \mathbb{R}^q$
 donc $r(\infty) = 1 - \underline{\psi}_s^T \underline{e} = 0.$

On parle surtout de méthode "stiffly accurate" et ces méthodes ont une importance cruciale dans le cadre de perturbations singulières. (cf. PCS).

Revenons une dernière fois sur l'exemple de la page 14

$$r(\infty) = 1 - \underline{b}^T A^{-1} \underline{e} = 1 - \frac{2}{\sqrt{3}} \frac{1 + \alpha\sqrt{3} + \alpha^2}{1 + \frac{4}{\sqrt{3}}\alpha + \alpha^2}$$

et on vérifie immédiatement

que $|r(\alpha)| < 1$ pour $\alpha > 0$. Les méthodes sont donc A-stable, fortement A-stable et même B-stable si $\alpha > 0$.

Par contre on voit que pour $\alpha = 1$, la méthode devient L-stable en plus et $|r(\alpha)| = 0$.