Méthodes numériques pour les équations différentielles

Notes[1] pour les cours MAT 5580 et 4785

Rémi VAILLANCOURT

Département de Mathématiques
Université d'Ottawa
Ottawa, ON, Canada K1N 6N5

18 septembre 1993

[1] Avec le concours du Conseil de l'éducation franco-ontarien.

Table des matières

Chapitre 1

PRÉLIMINAIRES

1.1 Problème de Cauchy pour les équations différentielles du premier ordre

Considérons l'équation différentielle du premier ordre aux valeurs initiales:

$$y' = f(x, y), \qquad y(a) = \eta. \tag{1.1}$$

On a le théorème qui suit.

Théorème 1.1 *Soit $f(x, y)$ continue en x et y sur $D = [a, b] \times (-\infty, \infty)$ et lipschitzienne en y sur D, c'est-à-dire,*

$$|f(x, y) - f(x, z)| \leq L|y - z| \qquad \forall (x, y), (x, z) \in D. \tag{1.2}$$

Alors (1.1) *admet une et une seule solution $y(x) \in C^1[a, b]$, pour chaque valeur initiale η.*

On trouve la démonstration dans les ouvrages sur les équations différentielles.

Définition 1.1 La constante L de (1.2) est la *constante de Lipschitz.*

Remarque 1.1 Si $f \in C^1(D)$, on peut prendre la valeur de L qui suit:

$$L = \sup_{(x,y)\in D} \left| \frac{\partial f(x, y)}{\partial y} \right|. \tag{1.3}$$

Exemple 1.1 L'équation différentielle $y' = \lambda y$, $y(a) = \eta$, admet la solution unique

$$y(x) = \eta\, e^{\lambda(x-a)}.$$

Il faut bien se rendre compte que

$$f(x,y) = \lambda y, \qquad \frac{\partial f}{\partial y} = \frac{\partial}{\partial y}(\lambda y) = \lambda.$$

Donc $L = |\lambda|$.

1.2 Équations aux différences

Considérons l'*équation aux différences* d'ordre k aux coefficients constants a_j:

$$a_k y_{n+k} + a_{k-1} y_{n+k-1} + \ldots + a_0 y_n = f_n, \qquad n = 1, 2, \ldots, \tag{1.4}$$

où

$$a_k a_0 \neq 0.$$

Une solution de (1.4) est une suite

$$\{y_n\} = \{y_1, y_2, \ldots\}$$

qui satisfait (1.4).
Soit maintenant $\{z_n\}$ la *solution générale* de l'*équation homogène*

$$a_k z_{n+k} + \cdots + a_0 z_n = 0, \qquad n = 1, 2, \ldots. \tag{1.5}$$

Si $\{w_n\}$ est une *solution particulière* de (1.4), la solution générale de (1.4) est

$$y_n = z_n + w_n.$$

Si le second membre de (1.4) est constant, c'est-à-dire $f_n = f$ est indépendante de n, et si

$$\sum_{j=0}^{k} a_j \neq 0,$$

alors

$$w_n = f \Big/ \sum_{j=0}^{k} a_j \tag{1.6}$$

est une solution particulière de (1.4).

Définition 1.2 Soient k solutions de (1.5),

$$\{y_{n,m}\}, \qquad m = 1, 2, \ldots, k;$$

celles-ci sont *linéairement indépendantes* si toute combinaison linéaire nulle:

$$b_1 y_{n,1} + b_2 y_{n,2} + \cdots + b_k y_{n,k} = 0, \qquad n = 1, 2, \ldots,$$

implique que

$$b_1 = b_2 = \cdots = b_k = 0.$$

Un ensemble de k solutions indépendantes de (1.5) se nomme un *système fondamental.*

Pour résoudre (1.5) on cherche des solutions de la forme

$$y_{n,m} = r_m^n. \tag{1.7}$$

En substituant $y_n = r^n$ dans (1.5) on obtient le *polynôme caractéristique*

$$p(r) = a_k r^k + a_{k-1} r^{k-1} + \cdots + a_0. \tag{1.8}$$

Si $p(r)$ admet k zéros *distincts* $r_m, m = 1, \ldots, k$, on peut montrer que $\{r_m^n\}, m = 1, \ldots, k$, forment un système fondamental de (1.5). Alors la solution générale de (1.4) est de la forme

$$y_n = \sum_{m=1}^{k} c_m r_m^n + \psi.$$

Lemme 1.1 *Si $r_1 = r_2$ est un zéro double et $r_3, \ldots, r_k$ sont des zéros simples de $p(r)$, alors nr_1^n est aussi une solution de (1.5).*

Démonstration: Par substitution,

$$\begin{aligned}
& a_k(n+k) r_1^{n+k} + a_{k-1}(n+k-1) r_1^{n+k-1} + \cdots + a_1(n+1) r_1^{n+1} + a_0 n r_1^n \\
&= n r_1^n [a_k r_1^k + a_{k-1} r_1^{k-1} + \cdots + a_1 r_1 + a_0] + \\
&\quad r_1^{n+1} [a_k k r_1^{k-1} + a_{k-1}(k-1) r_1^{k-2} + \cdots + a_2 2 r_1 + a_1] \\
&= n r_1^n p(r_1) + r_1^{n+1} p'(r_1) \\
&= 0 + 0,
\end{aligned}$$

puisque r_1 est une racine double de $p(r)$. □

Exemple 1.2 Montrer que les trois suites

$$\{r^n\}, \quad \{nr^n\}, \quad \{n(n-1)r^n\}, \qquad n = 0, 1, 2, \ldots, \tag{1.9}$$

sont linéairement indépendantes si $r \neq 0$.

Résolution. On a

$$\begin{aligned} b_1 r^n + b_2 n r^n + b_3 n(n-1) r^n &= 0, \\ (b_1 + n b_2 + n(n-1) b_3) r^n &= 0, \qquad r \neq 0, \\ q(n) := b_1 + n b_2 + n(n-1) b_3 &= 0, \qquad n = 0, 1, 2, \ldots . \end{aligned}$$

Or $q(n)$ est un polynôme du second degré en n. Puisqu'il admet plus de deux zéros distincts, alors $q(n) \equiv 0$. Donc $b_1 = b_2 = b_3 = 0$, et par conséquent, les trois fonctions en (1.9) sont linéairement indépendantes. □

En général, si $p(r) = 0$ admet s racines distinctes, $r_1, \ldots, r_s$, de multiplicité respective $m_1, \ldots, m_s$, $m_1 + \cdots + m_s = k$, la solution générale de (1.4), $\{y_n\}$, s'écrit de la forme qui suit:

$$\begin{aligned} y_n = {} & [c_{1,1} + c_{1,2} n + c_{1,3} n(n-1) + \cdots + c_{1,m_1} n(n-1) \cdots (n - m_1 + 2)] r_1^n \\ & + \cdots + \\ & [c_{s,1} + c_{s,2} n + c_{s,3} n(n-1) + \cdots + c_{s,m_s} n(n-1) \cdots (n - m_s + 2)] r_s^n \\ & + w_n . \end{aligned}$$

Exemple 1.3 Trouver la solution de l'équation aux différences

$$y_{n+4} - 4y_{n+3} + 5y_{n+2} - 4y_{n+1} + 4y_n = 4 \tag{1.10}$$

qui satisfait les conditions initiales

$$\begin{aligned} y_0 &= 5, \\ y_1 &= 0, \\ y_2 &= -4, \\ y_3 &= -12. \end{aligned} \tag{1.11}$$

Résolution: On obtient une solution particulière, w_n, par (1.6) puisque $f_n = 4$ est constante:

$$w_n = \frac{4}{1 - 4 + 5 - 4 + 4} = 2.$$

Le polynôme caractéristique

$$p(r) = r^4 - 4r^3 + 5r^2 - 4r + 4 = (r^2 + 1)(r - 2)^2$$

admet les zéros qui suivent:

$$i, -i, 2, 2.$$

La solution générale est donc

$$y_n = c_1 i^n + c_2 (-i)^n + c_3 2^n + c_4 n 2^n + 2. \tag{1.12}$$

Les conditions initiales donnent le système linéaire

$$\begin{aligned} c_1 + c_2 + c_3 + 0c_4 + 2 &= 5 \\ ic_1 - ic_2 + 2c_3 + 2c_4 + 2 &= 0 \\ -c_1 - c_2 + 4c_3 + 8c_4 + 2 &= -4 \\ -ic_1 + ic_2 + 8c_3 + 24c_4 + 2 &= -12 \end{aligned}$$

dont la solution s'obtient facilement:

$$\begin{aligned} c_1 &= 1 + i, \\ c_2 &= 1 - i = \bar{c}_1, \\ c_3 &= 1, \\ c_4 &= -1. \end{aligned}$$

Alors la solution de (1.10)–(1.11) est

$$y_n = (1+i)i^n + (1-i)(-i)^n + 2^n - n2^n + 2.$$

Cette solution est réelle puisque

$$(1+i)i^n + (1-i)(-i)^n = 2\Re(1+i)i^n.$$

En effet, puisque les coefficients de l'équation aux différences (1.10) sont réels de même que les valeurs initiales (1.11), on s'attend à une solution réelle. La substitution

$$i = e^{i\pi/2}$$

dans (1.12) donne

$$y_n = 2(\cos\frac{n\pi}{2} - \sin\frac{n\pi}{2}) + (1-n)2^n + 2. \tag{1.13}$$

sous forme réelle. □

1.3 Méthodes récursives pour les équations nonlinéaires

Pour résoudre l'équation nonlinéaire

$$y = g(y), \tag{1.14}$$

on peut procéder par la *récurrence*:

$$y^{[s+1]} = g(y^{[s]}), \qquad y^{[0]} \text{ arbitraire.} \tag{1.15}$$

Définition 1.3 On dit que g est *strictement contractante* si

$$\begin{aligned} |y^{[s+1]} - y^{[s]}| &= |g(y^{[s]}) - g(y^{[s-1]})| \\ &\leq M|y^{[s]} - y^{[s-1]}|, \qquad M < 1. \end{aligned}$$

Si g est lipschitzienne de constante $M < 1$, g est srictement contractante.

On a le résultat qui suit.

Théorème 1.2 *Si $g(y)$ est strictement contractante pour tout y, alors la récurrence* (1.15) *converge vers l'unique racine de l'équation* (1.14).

On peut trouver la démonstration dans Henrici [H1962], p. 216.

Chapitre 2

MÉTHODES DE RUNGE–KUTTA

2.1 Méthodes à un pas explicites

Soit le problème aux valeurs initiales, appelé aussi problème de Cauchy,

$$y' = f(x, y), \qquad y(a) = \eta, \tag{2.1}$$

où f est continue en x et lipschitzienne en y.
Considérons la méthode générale à un pas explicite:

$$y_{n+1} - y_n = h\phi(x_n, y_n, h) \tag{2.2}$$

où $x_n = a + nh$.
La fonction ϕ n'est pas, en général, linéaire en y ni en f; on atteint ainsi un ordre de précision supérieur. Les méthodes de Runge–Kutta sont des méthodes à un pas du type (2.2) où la fonction ϕ est d'une forme bien particulière. Le cas le plus simple est la *méthode d'Euler*:

$$y_{n+1} - y_n = hf(x_n, y_n) =: hf_n. \tag{2.3}$$

2.2 Ordre et convergence des méthodes à un pas explicites

Définition 2.1 La méthodes (2.2) est d'*ordre* p si p est le plus grand entier tel que la substitution de la solution exacte $y(x)$ de (2.1) dans (2.2) donne une

erreur d'ordre $p+1$:

$$y(x+h) - y(x) - h\phi(x, y(x), h) = O(h^{p+1}). \tag{2.4}$$

Notation 2.1 On note $g(h) = O(h^p)$ quand $h \to 0$ si

$$\left| \frac{g(h)}{h^p} \right| \leq M < \infty \quad \text{quand} \quad h \to 0,$$

et on dit que $g(h)$ est d'*ordre grand* O *de* h^p quand h tend vers 0.

Définition 2.2 La méthode (2.2) est *consistante* par rapport au problème (2.1) si

$$\phi(x, y, 0) = f(x, y), \tag{2.5}$$

où $f(x, y)$ est le second membre de l'équation différentielle (2.1).

On remarque que la consistance de (2.2) implique que (2.2) est d'ordre 1:

$$\begin{aligned}
& y(x+h) - y(x) - h\phi(x, y(x), h) \\
&= hy'(x) + \frac{1}{2}h^2 y'(x^*) - h\phi(x, y(x), 0) - h^2 \frac{\partial \phi}{\phi h}(x, y(x), h^*) \\
&= h[y'(x) - \phi(x, y(x), 0)] + O(h^2) \\
&= h[f(x, y(x)) - \phi(x, y(x), 0)] + O(h^2) \\
&= 0 + O(h^2), \qquad \text{par (2.5)}, \\
&= O(h^2).
\end{aligned}$$

On remarque aussi que la méthode d'Euler (2.3) est l'unique méthode linéaire de la forme (2.2); elle est consistante et d'ordre 1.

Définition 2.3 L'*algorithme de Taylor* d'ordre p est de la forme (2.2) avec $\phi = \varphi_T$:

$$\phi_T(x, y, h) = f(x, y) + \frac{h}{2!} f^{(1)}(x, y) + \cdots + \frac{h^{p-1}}{p!} f^{(p-1)}(x, y) \tag{2.6}$$

où

$$f^{(q)}(x; y) = \frac{d^q}{dx^q} f(x, y(x)), \qquad q = 1, 2, \ldots, p-1.$$

Théorème 2.1 *Soit $\phi(x, y, h)$ continue en x, y et h sur $D = [a, b] \times (-\infty, \infty) \times [0, h_0], h_0 > 0$ et lipschitzienne en y:*

$$|\phi(x, y^*, h) - \phi(x, y, h)| \leq M|y^* - y|$$

sur D. Alors la méthode (2.2) est convergente si elle est consistante.

La définition de la convergence d'une méthode numérique, soit à un pas, soit multipas, est l'objet du paragraphe 2.5 au chapitre 2. On peut trouver la démonstration du théorème dans Henrici [H1962], p.71.

2.3 Dérivation des méthodes classiques

Les méthodes classiques de Runge–Kutta ont précédé les ordinateurs. Le choix des coeficients convenait aux anciennes machines à calcul. On se restreint, dans cette section, aux méthodes d'ordre au plus quatre.

La *méthode de Runge–Kutta* explicite à s *stages* se définit par l'algorithme qui suit.

$$\begin{aligned} k_1 &= f(x,y) \\ k_i &= f\left(x + hc_i, y + h\sum_{j=1}^{i-1} a_{ij}k_j\right), \quad i = 2,3,\ldots,s, \end{aligned} \tag{2.7}$$

$$\phi(x,y,h) = \sum_{i=1}^{s} b_i k_i, \tag{2.8}$$

$$y_{n+1} = y_n + h\phi(x_n, y_n, h). \tag{2.9}$$

Dans ce chapitre, on adoptera l'hypothèse simplificatrice suivante:

$$c_i = \sum_{j=1}^{i-1} a_{ij}, \qquad i = 2,3,\ldots,s. \tag{2.10}$$

La consistance exige que

$$\sum_{i=1}^{s} b_i = 1. \tag{2.11}$$

On récrit (2.8) et (2.8) sous forme de tableau de Butcher:

$\vec{c}$	A					
0	0					
c_2	a_{21}					
c_3	a_{31}	a_{32}				
c_4	a_{41}	a_{42}	a_{43}			
$\vdots$						
c_s	a_{s1}	a_{s2}	a_{s3}	$\ldots$	$a_{s,s-1}$	
$\vec{b}^T$	b_1	b_2	b_3	$\ldots$	b_{s-1}	b_s

(2.12)

Tableau de Butcher d'une méthode de Runge–Kutta à s stage

On remarque que la matrice A est triangulaire inférieure stricte puisque la méthode est explicite.

Pour éviter les calculs fastidieux, on va restreindre la dérivation aux méthodes d'ordre 1, 2 et 3.

On introduit la notation

$$\begin{aligned} F &:= f_x + f f_y = f^{(1)}, \\ G &:= f_{xx} + 2 f f_{xy} + f^2 f_{yy}, \end{aligned} \tag{2.13}$$

où

$$f = f(x,y), \quad f_x = \frac{\partial f}{\partial x}(x,y), \quad \text{etc.}$$

Alors

$$\begin{aligned} f^{(2)} &= f_{xx} + 2 f f_{xy} + f^2 f_{yy} + f_x f_y + f f_y^2 \\ &= F f_y + G. \end{aligned}$$

La méthode de Taylor (2.6) devient donc

$$\phi_T(x,y,h) = f + \frac{1}{2} h F + \frac{1}{6} h^2 (F f_y + G) + O(h^3). \tag{2.14}$$

On développe k_2 et k_3:

$$\begin{aligned} k_2 &= f + h c_2 f_x + h a_{21} k_1 f_y + \\ &\quad \frac{1}{2} h^2 [c_2^2 f_{xx} + 2 c_2 a_{21} k_1 f_{xy} + a_{21}^2 k_1^2 f_{yy}] + O(h^3) \\ &\qquad\qquad \text{(on emploie } a_{21} = c_2 \text{ et } k_1 = f) \\ &= f + h c_2 [f_x + f f_y] + \frac{1}{2} h^2 c_2^2 [f_{xx} + 2 f f_{xy} + f^2 f_{yy}] + O(h^3), \end{aligned}$$

c'est-à-dire

$$k_2 = f + h c_2 F + \frac{1}{2} h^2 c_2^2 G + O(h^3); \tag{2.15}$$

$$\begin{aligned} k_3 &= f + h[c_3 f_x + \{a_{31} k_1 + a_{32} k_2\} f_y] + \\ &\quad \frac{1}{2} h^2 [c_3^2 f_{xx} + 2 c_3 \{a_{31} k_1 + a_{32} k_2\} f_{xy} + \{a_{31} k_1 + a_{32} k_2\}^2 f_{yy}] + O(h^3) \\ &\qquad\qquad \text{(on emploie } a_{31} = c_3 - a_{32}) \\ &= f + h[c_3 f_x + \{(c_3 - a_{32}) f + a_{32}(f + h c_2 F)\} f_y] + \end{aligned}$$

$$\frac{1}{2}h^2[c_3^2 f_{xx} + 2c_3\{(c_3 - a_{32})f + a_{32}f\}f_{xy} + \{(c_3 - a_{32})f + a_{32}f\}^2 f_{yy}] + O(h^3)$$

(on a rejeté dans $O(h^3)$ les termes en h^3)

$$= f + hc_3[f_x + ff_y] + h^2\{c_2 a_{32} F f_y + \frac{1}{2}c_3^2[f_{xx} + 2ff_{xy} + f^2 f_{yy}]\} + O(h^3),$$

c'est-à-dire

$$k_3 = f + hc_3 F + h^2\left(c_2 a_{32} F f_y + \frac{1}{2}c_3^2 G\right) + O(h^3). \tag{2.16}$$

Enfin

$$\begin{aligned} \phi(x,y,h) &= b_1 k_1 + b_2 k_2 + b_3 k_3 \\ &= (b_1 + b_2 + b_3)f + h(b_2 c_2 + b_3 c_3)F + \\ &\quad \frac{1}{2}h^2[2b_3 c_2 a_{32} F f_y + \{b_2 c_2^2 + b_3 c_3^2\}G] + O(h^3). \end{aligned} \tag{2.17}$$

Les coefficients de $1, h$ et h^2 doivent coïncider avec les termes correspondants de

$$\phi_T(x,y,h) = f + \frac{1}{2}hF + \frac{1}{6}h^2(Ff_y + G) + O(h^3), \tag{2.18}$$

suivant l'ordre de précision désiré.

Pour $s = 1$ (un stage), on a une seule équation de condition:

$$b_1 = 1,$$

puisque $b_2 = b_3 = \cdots = 0$. Alors

$$\phi(x,y,h) = f(x,y);$$

c'est la *méthode d'Euler*:

$$y_{n+1} = y_n + hf(x_n, y_n). \tag{2.19}$$

Le premier terme de l'erreur est $\frac{1}{2}h^2 F = O(h^2)$.

Pour $s = 2$ (deux stages) on a 2 équations de conditions à 3 inconnues:

$$\begin{aligned} b_1 + b_2 &= 1, \\ b_2 c_2 &= \frac{1}{2}, \end{aligned} \tag{2.20}$$

puisque $b_3 = b_4 = \cdots = 0$. Alors avec

$$b_1 = 0, \qquad b_2 = 1, \qquad c_2 = \frac{1}{2},$$

on a la *méthode d'Euler modifiée*:

$$y_{n+1} = y_n + hf\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}hf(x_n, y_n)\right). \tag{2.21}$$

Le premier terme de l'erreur est $\frac{1}{6}h^3(Ff_y + G) = O(h^3)$.
Avec

$$b_1 = b_2 = \frac{1}{2}, \qquad c_2 = 1,$$

on a la *méthode d'Euler améliorée*:

$$y_{n+1} = y_n + \frac{1}{2}h[f(x_n, y_n) + f(x_n + h, y_n + hf(x_n, y_n)]. \tag{2.22}$$

L'erreur est de l'ordre de $O(h^3)$.

Pour $s = 3$ (trois stages), on a 4 équations de condition à 6 inconnues puisque $b_4 = 0$.

$$\begin{aligned} b_1 + b_2 + b_3 &= 1, \\ b_2c_2 + b_3c_3 &= \frac{1}{2}, \\ b_2c_2^2 + b_3c_3^2 &= \frac{1}{3}, \\ b_3c_2a_{32} &= \frac{1}{6}. \end{aligned} \tag{2.23}$$

Avec

$$b_1 = \frac{1}{4}, \quad b_2 = 0, \quad b_3 = \frac{3}{4}, \quad c_2 = \frac{1}{3}, \quad c_3 = \frac{2}{3}, \quad a_{32} = \frac{2}{3},$$

on a la *méthode de Heun du* 3^e *ordre*:

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{4}(k_1 + 3k_3) \\ k_1 &= f(x_n, y_n), \\ k_2 &= f(x_n + \frac{1}{3}h, y_n + \frac{1}{3}hk_1), \\ k_3 &= f(x_n + \frac{2}{3}h, y_n + \frac{2}{3}hk_2). \end{aligned} \tag{2.24}$$

Avec

$$b_1 = \frac{1}{6}, \quad b_2 = \frac{2}{3}, \quad b_3 = \frac{1}{6}, \quad c_2 = \frac{1}{2}, \quad c_3 = 1, \quad a_{32} = 2,$$

on obtient la *méthode de Kutta du* 3^e *ordre*:

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{6}(k_1 + 4k_2 + k_3) \\ k_1 &= f(x_n, y_n), \\ k_2 &= f\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1\right), \\ k_3 &= f(x_n + h, y_n - hk_1 + 2hk_2). \end{aligned} \tag{2.25}$$

Pour $s = 4$, on donne sans dérivation et sous forme de tableau deux méthodes de Runge–Kutta du quatrième ordre, la classique (1/3) et une autre (3/8), la première utilisant Simpson 1/3 et la seconde Simpson 3/8.

c	A			
0	0			
1/2	1/2	0		
1/2	0	1/2	0	
1	0	0	1	0
b^T	1/6	2/6	2/6	1/6

(2.26)

Runge–Kutta classique d'ordre 4 (1/3)

c	A			
0	0			
1/3	1/3	0		
2/3	-1/3	1	0	
1	1	-1	1	0
b^T	1/8	3/8	3/8	1/8

(2.27)

Runge–Kutta d'ordre 4 (3/8)

2.4 Méthodes d'ordre supérieur à quatre

Notons l'ordre maximum, $p^*(s)$, des méthodes de Runge–Kutta à s stages. On a démontré que $p^*(s)$ satisfait les inégalités qui suivent.

$$\begin{aligned} p^*(s) &= s, \qquad s = 1, 2, 3, 4, \\ p^*(5) &= 4, \\ p^*(6) &= 5, \end{aligned}$$

$$\begin{aligned} p^*(7) &= 6, \\ p^*(8) &= 6, \\ p^*(9) &= 7, \\ p^*(s) &\leq s-3, \qquad s = 10, 11, \ldots . \end{aligned} \tag{2.28}$$

A titre d'exemple, on donne deux méthodes sous forme de tableaux. On ne répète pas le dénominateur de chacune des lignes de la matrice A ni celui du vecteur $\mathbf{b}^T$.

c	A					
0	0					
1/3	1/3	0				
2/5	4/25	6/	0			
1	1/4	-12/	15/	0		
2/3	6/81	90/	-50/	8/	0	
4/5	6/75	36/	10/	8/	0/	0
b^T	23/192	0/	125/	0/	-81/	125/

(2.29)

Kutta–Nyström d'ordre 5 à 6 stages

c	A							
0	0							
1/9	1/9	0						
1/6	1/24	3/	0					
1/3	1/6	-3/	4/	0				
1/2	-5/8	27/	-24/	6/	0			
2/3	221/9	-981/	867/	-102/	1/	0		
5/6	-183/48	678/	-472/	-66/	80/	3/	0	
1	716/82	-2079/	1002/	834/	-454/	-9/	72/	0
b^T	41/840	0	216/	27/	272/	27/	216/	41/

(2.30)

Huťa d'ordre 6 à 8 stages

2.5 Bornes de l'erreur de méthode

Définition 2.4 L'*erreur de méthode* à x_{n+1} de la méthode explicite à un pas (2.2):

$$y_{n+1} = y_n + h\varphi(x_n, y_n, h),$$

est notée et définie comme suit:

$$T_{n+1} := y(x_{n+1}) - y(x_n) - h\phi(x_n, y(x_n), h) \tag{2.31}$$

où $y(x)$ est la solution théorique de $y' = f(x,y), \quad y(a) = \eta$.

Définition 2.5 L'*erreur de méthode locale*, c'est-à-dire avec $y_n = y(x_n)$, est

$$\begin{aligned} T_{n+1} &= y(x_{n+1}) - y_n - h\phi(x_n, y_n, h) \\ &= y(x_{n+1}) - y_{n+1}. \end{aligned}$$

L'*erreur de méthode globale* est notée

$$e_{n+1} = y(x_{n+1}) - y_{n+1}.$$

Définition 2.6 Le *terme principal de l'erreur* de (2.2) d'ordre p est

$$\psi(x_n, y(x_n))h^{p+1}$$

où

$$T_{n+1} = \psi(x_n, y(x_n))h^{p+1} + O(h^{p+2}) \tag{2.32}$$

et $\psi(x,y)$ est la *fonction principale de l'erreur.*

Exemple 2.1 Pour $s = 2$ —cas général— par (2.31),

$$\begin{aligned} T_{n+1} &= y(x_{n+1}) - y(x_n) - h\phi(x_n, y(x_n), h) \\ &= y(x_n) + h\phi_T(x_n, y(x_n), h) \\ &\quad -y(x_n) - h\phi(x_n, y(x_n), h); \end{aligned} \tag{2.33}$$

alors par (2.14) et (2.18),

$$\begin{aligned} T_{n+1} &= h\left[f + \frac{1}{2}hF + \frac{1}{6}h^2(Ff_y + G)\right]_{x=x_n, y=y(x_n)} + O(h^4) \\ &\quad -h[(b_1+b_2)f + hb_2c_2F + \frac{1}{2}h^2b_2c_2^2G]_{x=x_n, y=y(x_n)} + O(h^4), \end{aligned}$$

et par (2.21) pour la méthode du second ordre,

$$T_{n+1} = h^3\left[\frac{1}{6}Ff_y + \left(\frac{1}{6} - \frac{1}{4}c_2\right)G\right]_{x=x_n, y=y(x_n)} + O(h^4).$$

Donc la fonction principale de l'erreur de la méthode générale de Runge–Kutta d'ordre 2 est

$$\psi(x,y) = \frac{1}{6}Ff_y + \left(\frac{1}{6} - \frac{1}{4}c_2\right)G. \tag{2.34}$$

Bornes supérieures de Lotkin sur ψ.

Supposons que sur $x \in [0,b], y \in (-\infty,\infty)$, on ait

$$|f(x,y)| < Q, \qquad \left|\frac{\partial^{i+j} f}{\partial x^i \partial y^j}\right| < \frac{P^{i+j}}{Q^{j-1}}, \qquad i+j \le p, \tag{2.35}$$

où P et Q sont des constantes > 0 et p est l'ordre de la méthode. Alors

$$|f_y| < P,$$

$$|F| = |f_x + ff_y| \le PQ + QP = 2PQ,$$

$$|G| = |f_{xx} + 2ff_{xy} + f^2 f_{yy}| \le P^2Q + 2QP^2 + Q^2P^2/Q = 4P^2Q.$$

Donc pour $s = 2$, (2.34) implique

$$\begin{aligned} |\psi(x,y)| &\le \frac{1}{6}|Ff_y| + \left|\frac{1}{6} - \frac{1}{4}c_2\right| |G| \\ &< \frac{1}{6}2PQ + \frac{1}{4}\left|\frac{2}{3} - c_2\right| 4P^2Q \\ &= \left(\frac{1}{3} + \left|\frac{2}{3} - c_2\right|\right) P^2Q; \end{aligned}$$

alors on a la borne pour l'erreur locale principale

$$|\psi(x_n y(x_n))h^3| < \left(\frac{1}{3} + \left|\frac{2}{3} - c_2\right|\right) h^3 P^2 Q.$$

Pour $s = 3$, on a une borne semblable

$$|T_{n+1}| < \text{const}_{(c_2,c_3)} h^4 P^3 Q. \tag{2.36}$$

Dans le cas classique, pour $s = 4$, on a

$$|T_{n+1}| < \frac{73}{720} h^5 P^4 Q. \tag{2.37}$$

Bornes supérieures de Bieberbach

Dans le cas classique, pour $s = 4$, on a

$$|T_{n+1}| < G h^5 Q N (1 + N + N^2 + N^3 + N^4) \tag{2.38}$$

où, dans le voisinage $|x - x_0| < A, |y - y_0| < B$,

$$|f(x,y)| < Q, \quad \left|\frac{\partial^{i+j} f(x,y)}{\partial x^i \partial y^j}\right| < \frac{N}{Q^{j-1}}, \quad i+j \leq 4,$$

$$|x - x_0|N < 1 \text{ et } AQ < B.$$

On peut démontrer que l'erreur globale est $O(h^p)$ si l'erreur locale est $O(h^{p+1})$, c'est-à-dire

$$|T_{n+1}| \leq Kh^{p+1} \tag{2.39}$$

implique

$$|e_n| := |y(x_n - y_n| \leq \frac{h^p K}{L}\left[e^{L(x_n - a)} - 1\right] \tag{2.40}$$

où L est la constante de Lipschitz de $f(x,y)$ par rapport à y.

Dans le cas classique RK4, Carr a obtenu une autre borne de l'erreur globale si $|T_n| < E$ et si dans une région D du plan xy,

$$-M_2 < \frac{\partial f}{\partial y} < -M_1 < 0,$$

alors, dans une autre région correspondante D^*,

$$|e_n| \leq \frac{2E}{hM_1} \tag{2.41}$$

pourvu que

$$h < \min\left\{\frac{M_1}{M_2^2}, \frac{4M_1^3}{M_2^4}\right\}.$$

On peut remarquer les faits suivants:

(1) Les bornes sont difficiles à appliquer.

(2) On ne peut contrôler h au moyen de (2.36) et (2.37).

(3) Les bornes sont utiles parce que les méthodes RK sont souvent utilisées pour le démarrage.

(4) On peut choisir les paramètres libres pour minimiser la borne de l'erreur locale.

Exemple 2.2 Considérons RK4 classique. Soit

$$y' = e^{10(x-y)}, \qquad y(0) = 0.1.$$

Démarrer à 10^{-8} près, en se référant à la borne de Lotkin pour déterminer h.

Résolution.

$$f(x,y) = e^{10(x-y)}, \quad \frac{\partial^{i+j} f(x,y)}{\partial x^i \partial y^i} = (-1)^j 10^{i+j} e^{10(x-y)}.$$

Si $|f| < Q$ près de 0, alors

$$|f_x|, \quad |f_y| < 10Q,$$

$$|f_{xx}|, \quad |f_{xy}|, \quad |f_{yy}| \leq 10^2 Q,$$

$$|f_{xxx}|, \quad |f_{xxy}|, \quad |f_{xyy}|, \quad |f_{yyy}| \leq 10^3 Q,$$

$$|f_{xxxx}|, \ldots, |f_{yyyy}| \leq 10^4 Q.$$

Par (2.35), on cherche P et Q telles que

$$|f| < Q,$$

$$|f_x| < QP, \quad |f_y| < P,$$

$$|f_{xx}| < QP^2, \quad |f_{xy}| < P^2, |f_{yy}| < P^2/Q,$$

$$|f_{xxx}| < QP^3,$$

$$|f_{xxxx}| < QP^4.$$

On voit que $P = 10$ et $0 < Q < 1$ font l'affaire. De plus $f(x_0, y_0) = f(0, 0.1) = e^{-1} = 0.368$. Faisons l'hypothèse gratuite: $|f(x,y)| < 0.368$ près de (0.0). Alors, par (2.37)

$$|T_{n+1}| < \frac{73}{720} 10^4 \times 0.368 h^5.$$

Pour $|T_{n+1}| < 10^{-8}$, on prend

$$|T_{n+1}| < \frac{73}{720} 10^4 \times 0.368 h^5 < 10^{-8}.$$

Donc

$$h^5 < \frac{72}{73} \times \frac{10^{-10}}{3.68}$$

ce qui donne

$$h \approx 0.8 \times 10^{-2}. \quad \square$$

Exemple 2.3 Résoudre le problème de l'exemple précédent sur $0 \leq x \leq 1$ avec RK4 classique (2.26) et $h = 0.01$. Comparer l'erreur avec l'erreur globale (2.40) et (2.41) et la borne (2.39).

Résolution. La solution analytique est

$$y(x) = \frac{1}{10}\log\left(e^{10x} + e - 1\right).$$

D'après la table 13 de Lambert, on voit que pour $x \in [0,1]$,

$$-0.1 \leq x - y < 0,$$

donc

$$e^{-1} \leq f(x,y) < 1. \tag{2.42}$$

Donc avec $P = 10, Q = 1$ (v. ex. 1), (2.35) implique

$$|T_{n+1}| < \frac{73}{720}10^4 h^5 = 1014h^5,$$

ce qui donne

$$K = 1014$$

dans (2.39).

Pour calculer la constante de Lipschitz L pour f, on voit que

$$f_y = -10e^{10(x-y)} = -10f;$$

alors $x \in [0,1]$ et (2.42) donnent

$$-10e^{-1} \geq f_y > -10, \tag{2.43}$$

c'est-à-dire

$$L = 10.$$

On voit que (2.40) donne

$$|e_n| < 1014 \times 10^{-9}[e^{10x_n} - 1] \leq 10^{-3}.$$

D'autre part, par (2.43) (v. table 13 col. 4) on peut appliquer (2.41) avec

$$M_1 = 10e^{-1}, \quad M_2 = 10$$

à condition que

$$h < \min\left[\frac{1}{10}e^{-1}, \frac{2}{5}e^{-3}\right] = 0.02.$$

Le choix de $h = 0.01$ satisfait cette exigence, donc (2.41) est valide avec

$$E = Kh^{p+1} = 1014 \times 10^{-10}.$$

On a donc la borne

$$|e_n| < \frac{2028 \times 0^{-8}}{10e^{-1}} = 5.5 \times 10^{-6}.$$

Donc les deux bornes sont trop fortes parce que l'erreur est de l'ordre de 10^{-8}.
□

2.6 Estimation de l'erreur de méthode

On considère quelques façons d'estimer l'erreur locale.

1. L'*extrapolation à la limite* de Richardson avec h et $2h$:

$$y(x_{n+1}) - y_{n+1} = \psi(x_n, y(x_n))h^{p+1} + O(h^{p+2}), \tag{2.44}$$

$$\begin{aligned} y(x_{n+1}) - y^*_{n+1} &= \psi(x_{n-1}, y(x_{n-1}))(2h)^{p+1} + O(h^{p+2}) \\ &= \psi(x_n, y(x_n))(2h)^{p+1} + O(h^{p+2}). \end{aligned} \tag{2.45}$$

On soustrait (2.44) de (2.45):

$$y_{n+1} - y^*_{n+1} = (2^{p+1} - 1)\psi(x_n, y(x_n))h^{p+1} + O(h^{p+2}),$$

ce qui donne l'erreur locale principale:

$$\psi(x_n, y(x_n))h^{p+1} = (y_{n+1} - y_{n+1}*)/(2^{p+1} - 1). \tag{2.46}$$

Cette méthode donne un contrôle adéquat de h mais coûte 50 pourcent de plus si on fait le contrôle à tous les 2 pas.

2. Moyenne de l'erreur locale proposée Scraton. Pour $RK4$ classique, on a

$$30T_{n+3} = 10y_n + 9y_{n+1} - 18y_{n+2} - y_{n+3} + 3h[f_n + 6f_{n+1} + 3f_{n+2}]. \tag{2.47}$$

Cette façon n'est pas très bonne quand l'erreur locale change rapidement, c'est-à-dire quand on veut vraiment contrôler le pas, en effet, cette estimation dépend des valeurs antérieures.

3. Merson à 5 stages du $4^{ème}$ ordre.

c	A				
1/3	1/3				
1/3	1/6	1/6			
1/2	1/8	0	3/8		
1	1/2	0	−3/2	2	
b^T	1/6	0	0	4/6	1/6

On a l'estimation

$$30T_{n+1} = h(-2k_1 + 9k_3 - 8k_4 + k_5).$$

Techniquement, cette estimation n'est valide que pour les équations différentielles de la forme

$$y' = ax + by + c.$$

4. Scraton à 5 stages du $4^{ème}$ ordre, valide pour les équations différentielles nonlinéaires.

c	A				
2/9	2/9				
1/3	1/12	1/4			
3/4	69/128	−243/	−270/		
9/10	−3105/10000	18225/	−11016/	4896/	
b^T	17/162	0	81/170	32/135	250/1377

On a l'estimation

$$T_{n+1} = hqr/s$$

où

$$\begin{aligned} q &= -1/18k_1 + 27/170k_3 - 4/15k_4 + 25/153k_5 \\ r &= 19/24k_1 - 27/8k_2 + 57/20k_3 - 4/15k_4 \\ s &= k_4 - k_1. \end{aligned}$$

Cette estimation est nonlinéaire dans les k_i; donc elle ne s'adapte pas aux systèmes d'équations différentielles.

5. Méthode d'England emboîtée à 6 stages d'ordre $(4, 5)$, c'est-à-dire $(\mathbf{c}, A, \mathbf{b})$ forment une méthode d'ordre 4 et $(\mathbf{c}, A, \widehat{\mathbf{b}})$ forment une méthode d'ordre 5. Voici le tableau de Butcher modifié de cette paire de méthodes.

c	A					
1/2	1/2					
1/2	1/4	1/4				
1	0	−1	2			
2/3	7/27	10/	0/	1/		
1/5	28/625	−125/	546/	54/	−378/	
$\widehat{\mathbf{b}}^T$	1/6	0	4/6	1/6	0	0
$\mathbf{b}^T$	1/24	0	0	5/48	27/56	125/336
$\mathbf{b}^T - \widehat{\mathbf{b}}^T = E^T$	−1/8	0	−2/3	−1/16	27/56	125/336

Remarque. Sans l'estimation de l'erreur, on a une méthode à 4 stages: k_1, k_2, k_3, k_4.

Résumé. Les estimations de l'erreur locale sont une moyenne sur un nombre de pas ou requièrent une ou plusieurs évaluations additionnelles de f.

2.7 Stabilité absolue des méthodes

On linéarise l'équation différentielle, c'est-à-dire on considère l'équation test:

$$y' = \lambda y.$$

On voit que

$$\frac{\partial f}{\partial y} = \lambda = \text{ constante.}$$

Appliquons RK générale avec $s = 3$ à l'équation test. Alors

$$\begin{aligned} k_1 &= f(x,y) = \lambda y \\ k_2 &= f(x + hc_2, y + hc_2k_1) = \lambda[y + hc_2\lambda y] = \lambda y[1 + c_2h\lambda] \\ k_3 &= f(x + hc_3, y + h(c_3 - a_{32})k_1 + ha_{32}k_2) \\ &= \lambda[y + h\lambda y(c_3 - a_{32}) + h\lambda y a_{32}(1 + c_2h\lambda)] \\ &= \lambda y(1 + c_3h\lambda + c_2a_{32}h^2\lambda^2); \end{aligned}$$

alors

$$\begin{aligned} \varphi(x,y,h) &= b_1k_1 + b_2k_2 + b_3k_3 \\ &= \lambda y[(b_1 + b_2 + b_3) + (b_2c_2 + b_3c_3)h\lambda + b_3c_2a_{32}h^2\lambda^2] \end{aligned}$$

et

$$y_{n+1} - y_n = h\lambda[(b_1 + b_2 + b_3) + (b_2c_2 + b_3c_3)h\lambda + b_3c_2a_{32}h^2\lambda^2]y_n.$$

Si on note $\bar{h} = h\lambda$, on obtient:

$$\frac{y_{n+1}}{y_n} = 1 + (b_1 + b_2 + b_3)\bar{h} + (b_2c_2 + b_3c_3)\bar{h}^2 + b_3c_2a_{32}\bar{h}^3.$$

La solution générale est

$$y_n = d_1r_1^n, \qquad d_1 \text{ une constante arbitraire,}$$

où

$$r_1 = 1 + (b_1 + b_2 + b_3)\bar{h} + (b_2c_2 + b_3c_3)\bar{h}^2 + b_3c_2a_{32}\bar{h}^3. \tag{2.48}$$

Définition 2.7 Une méthode RK à 3 stages est *absolument stable* sur (α, β) si

$$|r_1| < 1, \qquad \bar{h} \in (\alpha, \beta).$$

Si RK est consistante, alors (2.18) implique

$$b_1 + b_2 + b_3 = 1$$

Figure 2.1: Intervalle de stabilité absolue des méthodes RK explicites d'ordre 3.

et

$$r_1 = 1 + \bar{h} + O(\bar{h}^2).$$

Donc pour $\bar{h} > 0$ petit, $r_1 > 1$; donc l'intervalle de stabilité absolue est de la forme

$$(\alpha, 0), \qquad \alpha < 0.$$

Si RK à 3 stages est d'ordre 3, alors (2.23) implique

$$b_2 c_2 + b_3 c_3 = 1/2, \qquad b_3 c_2 a_{32} = 1/6,$$

et

$$r_1 = 1 + \bar{h} + \frac{1}{2}\bar{h}^2 + \frac{1}{6}\bar{h}^3.$$

Conclusion: Puisque (2.23) est satisfaite par toutes les méthodes RK d'ordre 3, l'intervalle de stabilité absolue $(-2.51, 0)$ est le même pour toutes les méthodes RK à 3 stages d'ordre 3. A la figure 2.1, le graphique de $r_1(\bar{h})$ montre que l'intervalle de la zéro stabilité est $(-2.51, 0)$. On généralise le résultat précédent par une autre méthode. Pour une méthode d'ordre p, de (2.8)–(2.9) il vient

$$\varphi_T(x, y, h) - \varphi(x, y, h) = O(h^p);$$

donc

$$\begin{aligned} y_{n+1} &= y_n + h\varphi_T(x_n, y_n, h) + O(h^{p+1}) \\ &= y_n + hf(x_n, y_n) + \frac{h^2}{2!}f^{(1)}(x_n, y_n) + \cdots + \frac{h^p}{p!}f^{(p-1)}(x_n, y_n) + O(h^{p+1}), \end{aligned}$$

et pour l'équation test $y' = \lambda y$.

$$f(x_n, y_n) = \lambda y_n, \quad f^{(q)}(x_n, y_n) = \lambda^{q+1} y_n, \qquad q = 1, 2, \ldots, p-1.$$

Donc

$$y_{n+1} = y_n + (h\lambda + \frac{1}{2!}\bar{h}^2\lambda^2 + \cdots + \frac{1}{p!}h^p\lambda^p)y_n + O(h^{p+1}),$$

et

$$\frac{y_{n+1}}{y_n} = r_1 = 1 + \bar{h} + \frac{\bar{h}^2}{2!} + \cdots + \frac{\bar{h}^p}{p!} + O(h^{p+1}). \tag{2.49}$$

Mais pour une méthode à s stages: r_1 est un polynôme de degré s en $\bar{h}$. Si $s = p = 1, 2, 3, 4$,

$$r_1 = 1 + \bar{h} + \frac{\bar{h}^2}{2!} + \cdots + \frac{\bar{h}^p}{p!} \tag{2.50}$$

pour une valeur quelconque des paramètres libres, qui satisfait les équations de condition. Donc pour $p = 1, 2, 3, 4$, toutes les méthodes de RK, dont l'ordre p est égal au nombre de stages s, ont la même région de stabilité absolue.

Si $p < s$,

$$r_1 = 1 + \bar{h} + \frac{\bar{h}^2}{2!} + \cdots + \frac{\bar{h}^p}{p!} + \sum_{q=p+1}^{s} \gamma_q \bar{h}^q,$$

où les γ_q sont fonctions des coefficients de la méthode RK mais ne sont pas déterminés par les conditions de l'ordre p. Alors la région de stabilité dépend du choix des paramètres libres.

Exemple Si $s = 3$, et l'ordre est 2, on doit satisfaire seulement les 2 premières équations (2.18). Alors (2.48) implique

$$r_1 = 1 + \bar{h} + 1/2\bar{h}^2 + \gamma_3 \bar{h}^3,$$

où $\gamma_3 = b_3 c_2 a_{32}$. Si $\gamma_3 = 0$, l'intervalle de stabilité absolue est $(-2, 0)$, si $\gamma_3 = 1/6$, il est $(-2.51, 0)$, et si $\gamma_3 = 1/12$ il est $(-4.52, 0)$.

Exemple 2.4 Illustrer l'effet de l'instabilité absolue avec RK 4 classique pour les problèmes suivants:

(i) $y' = -20y, \quad y(0) = 1$, avec $h = 0.1$ et 0.2.

(ii) $y' = -5xy^2 + 5/x - 1/x^2, \quad y(1) = 1$, avec $h = 0.2, 0.3$ et 0.4.

Pour (i), $f_y = -20, \bar{h} = 20h$ et l'intervalle de stabilité absolue est $(-2.78, 0)$. Avec

$$\begin{aligned} h &= 0.1, \qquad \bar{h} = -2 \in (-2.78, 0), \\ h &= 0.2, \qquad \bar{h} = -4 \notin (-2.78, 0). \end{aligned}$$

Voir la table 17 du manuel à la page 139.

Pour (ii), $f_y = -10xy = -10$ pour la solution théorique $y = 1/x$. On remarque que $f_y \neq$ constante le long des solutions voisines. Avec $\lambda = -10$ et $\bar{h} = -10h$, pour les valeurs de h données on a

$$\begin{aligned} h &= 0.2, \qquad \bar{h} = -2 \in (-2.78, 0), \\ h &= 0.3, \qquad \bar{h} = -3 \notin (-2.78, 0) \text{ mais assez près}, \\ h &= 0.4, \qquad \bar{h} = -4 \notin (-2.78, 0) \text{ et assez éloigné}. \end{aligned}$$

Voir la table 18 du manuel à la page 140.

2.8 Paires de formules de Runge–Kutta avec interpolants

On présente sous forme de tableaux la paire de formules bien connue de Runge–Kutta–Fehlberg d'ordre 4 et 5 à 6 stages, RKF(4,5)6, et quatre paires de formules de Dormand–Prince respectivement d'ordre 4 et 5. On modifie la notation pour suivre les articles sur le sujet: k_j devient f_j Les vecteurs $\hat{b}^T$ et b^T sont associés respectivement aux formules d'ordre inférieur et d'ordre supérieur. Les fractions sont réduites. On écrit

$$f_1^{n+1} = f(x_{n+1}, y_{n+1}).$$

Les paires sont DP(4,5)6M, DP(4,5)7M, DP(4,5)7C et DP(4,5)7S, d'ordre 4 et 5 à 6, respectivement, à 7 stages. Les lettres M, S et C indiquent resp. qu'on a *minimisé* le terme de l'erreur, maximisé le domaine de *stabilité* absolue, et fait un compromis entre les deux premiers cas.

2.8.1 La paire RKF(4,5)6 avec interpolant

	c	A					
f_1	0	0					
f_2	$\frac{1}{4}$	$\frac{1}{4}$	0				
f_3	$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$	0			
f_4	$\frac{12}{13}$	$\frac{1932}{2197}$	$-\frac{7200}{2197}$	$\frac{7296}{2197}$	0		
f_5	1	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$-\frac{845}{4104}$	0	
f_6	$\frac{1}{2}$	$-\frac{8}{27}$	2	$-\frac{3544}{2565}$	$\frac{1859}{4104}$	$-\frac{11}{40}$	0
$\hat{y}_{n+1}$	$\hat{b}^T$	$\frac{25}{216}$	0	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$	0
y_{n+1}	b^T	$\frac{16}{135}$	0	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50}$	$\frac{2}{55}$
$y_{n+0.6}$		$\frac{1559}{12500}$	0	$\frac{153856}{296875}$	$\frac{68107}{2612500}$	$-\frac{243}{31250}$	$-\frac{2106}{34375}$

(2.51)

Paire de Fehlberg d'ordre 4 et 5 à 6 stages

L'interpolant d'Hermite d'ordre 4 est alors

$$\begin{aligned} u_0(x_n+\tau h_n) &= d_0(\tau)y_n + d_1(\tau)h_n f_1 + d_2(\tau)y_{n+1} \\ &\quad + d_3(\tau)h_n f_1^{n+1} + d_4(\tau)y_{n+0.6} \end{aligned} \tag{2.52}$$

avec les polynômes d'interpolation d'Hermite

$$\begin{aligned} d_0(\tau) &= (\tau-1)^2\left(1-\frac{5}{3}\tau\right)\left(\frac{11}{3}\tau+1\right) \\ d_1(\tau) &= \tau(\tau-1)^2\left(1-\frac{5}{3}\tau\right) \\ d_2(\tau) &= \tau^2\left(\frac{3}{4}-\frac{5}{4}\tau\right)(9\tau-11) \\ d_3(\tau) &= \tau^2(\tau-1)\left(\frac{5}{2}\tau-\frac{3}{2}\right) \\ d_4(\tau) &= \frac{625}{36}\tau^2(\tau-1)^2. \end{aligned} \tag{2.53}$$

L'interpolant d'Hermite–Birkhoff d'ordre 5 s'obtient alors par le processus du boot-strapping d'Enright et al.:

$$\begin{aligned} u_1(x_n+\tau h_n) &= d_0(\tau)y_n + d_1(\tau)h_n f_1 + d_2(\tau)y_{n+1} \\ &\quad + d_3(\tau)h_n f_1^{n+1} + d_4(\tau)h_n f_7 + d_5(\tau)h_n f_8, \end{aligned} \tag{2.54}$$

où

$$f_7 = f(x_n+0.86h_n, u_0(x_n+0.86h_n))$$

et

$$\begin{aligned} u_1(x_n+0.86h_n) &= -\frac{396851}{11250000}y_n + \frac{3918031}{5000000}y_{n+1} + \frac{90601}{360000}y_{n+0.6} \\ &\quad -h_n\left(\frac{27391}{3750000}f_1 + \frac{168259}{2500000}f_1^{n+1}\right), \end{aligned}$$

$$f_8 = f(x_n+0.93h_n, u_0(x_n+0.93h_n))$$

et

$$\begin{aligned} u_1(x_n+0.93h_n) &= -\frac{237699}{20000000}y_n + \frac{75064671}{80000000}y_{n+1} + \frac{47089}{640000}y_{n+0.6} \\ &\quad -h_n\left(\frac{50127}{20000000}f_1 + \frac{1997919}{40000000}f_1^{n+1}\right), \end{aligned}$$

avec les polynômes d'Hermite–Birkhoff

$$\begin{aligned}
d_0(\tau) &= \left(\frac{375}{64}\tau^3 - \frac{8925}{1024}\tau^2 + 2\tau + 1\right)(\tau-1)^2 \\
d_1(\tau) &= \tau\left(\frac{5375}{3968}\tau^2 - \frac{19062325}{8189952}\tau + 1\right)(\tau-1)^2 \\
d_2(\tau) &= -\tau^2\left(\frac{375}{64}\tau^3 - \frac{20925}{1024}\tau^2 + \frac{12949}{512}\tau - \frac{11997}{1024}\right) \\
d_3(\tau) &= \tau^2\left(\frac{199625}{6272}\tau^2 - \frac{5385075}{100352}\tau + \frac{2291427}{100352}\right)(\tau-1) \\
d_4(\tau) &= \tau^2\left(\frac{78125}{1568}\tau - \frac{47953125}{1078784}\right)(\tau-1)^2 \\
d_5(\tau) &= -\tau^2\left(\frac{234375}{3038}\tau - \frac{8734375}{145824}\right)(\tau-1)^2.
\end{aligned} \tag{2.55}$$

On passe maintenant aux paires de Dormand–Prince. Martine Calvé, dans sa thèse de M.Sc., a étudié les méthodes DP(4,5)6M, 7M et 7S dans la norme uniforme.

2.8.2 La paire DP(4,5)6M avec interpolant

	c	A					
f_1	0	0					
f_2	$\frac{1}{5}$	$\frac{1}{5}$	0				
f_3	$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$	0			
f_4	$\frac{3}{5}$	$\frac{3}{10}$	$-\frac{9}{10}$	$\frac{6}{5}$	0		
f_5	$\frac{2}{3}$	$\frac{226}{729}$	$-\frac{25}{27}$	$\frac{880}{729}$	$\frac{55}{729}$	0	
f_6	1	$-\frac{181}{270}$	$\frac{5}{2}$	$-\frac{266}{297}$	$-\frac{91}{27}$	$\frac{189}{55}$	0
$\hat{y}_{n+1}$	$\hat{b}^T$	$\frac{31}{540}$	0	$\frac{190}{297}$	$-\frac{145}{108}$	$\frac{351}{220}$	$\frac{1}{20}$
y_{n+1}	b^T	$\frac{19}{216}$	0	$\frac{1000}{2079}$	$-\frac{125}{216}$	$\frac{81}{88}$	$\frac{5}{56}$
$y_{n+0.6}$		$\frac{16069}{187500}$	0	$\frac{9782}{20625}$	$-\frac{1931}{7500}$	$\frac{217161}{687500}$	$-\frac{1149}{62500}$

(2.56)

Paire de Dormand–Prince 6M d'ordre 4 et 5 à 6 stages

L'interpolant d'Hermite $u_0(x_n + \tau h_n)$ d'ordre 4 est le même que celui de la paire RKF(4,5)6 (2.52) puisqu'on interpole au mêmes points.

L'interpolant d'Hermite–Birkhoff d'ordre 5 s'obtient alors par le processus du boot-strapping d'Enright et al.:

$$\begin{aligned} u_1(x_n+\tau h_n) &= d_0(\tau)y_n + d_1(\tau)h_n f_1 + d_2(\tau)y_{n+1} \\ &\quad + d_3(\tau)h_n f_1^{n+1} + d_4(\tau)h_n f_7 + d_5(\tau)h_n f_8, \end{aligned} \tag{2.57}$$

où

$$f_7 = f(x_n + 0.05h_n, u_0(x_n + 0.05h_n))$$

et

$$\begin{aligned} u_1(x_n+0.05h_n) &= \frac{281941}{28800}y_n - \frac{2321}{128000}y_{n+1} + \frac{361}{9216}y_{n+0.6} \\ &\quad + h_n\left(\frac{3971}{96000}f_1 + \frac{209}{64000}f_1^{n+1}\right), \end{aligned}$$

$$f_8 = f(x_n + 0.95h_n, u_0(x_n + 0.95h_n))$$

et

$$\begin{aligned} u_1(x_n+0.95h_n) &= -\frac{1883}{28800}y_n + \frac{123823}{128000}y_{n+1} + \frac{361}{9216}y_{n+0.6} \\ &\quad - h_n\left(\frac{133}{96000}f_1 + \frac{2527}{64000}f_1^{n+1}\right), \end{aligned}$$

avec les polynômes d'Hermite–Birkhoff:

$$\begin{aligned} d_0(\tau) &= -\frac{1}{61}(480\tau^3 - 240\tau^2 - 122\tau - 61)(\tau-1)^2 \\ d_1(\tau) &= \frac{1}{1159}\tau(11440\tau^2 - 11820\tau + 1159)(\tau-1)^2 \\ d_2(\tau) &= \frac{1}{61}\tau^2(480\tau^3 - 1200\tau^2 + 838\tau - 57) \\ d_3(\tau) &= \frac{1}{1159}\tau^2(11440\tau^2 - 11060\tau + 779)(\tau-1) \\ d_4(\tau) &= -\frac{1000}{10431}\tau^2(144\tau - 133)(\tau-1)^2 \\ d_5(\tau) &= -\frac{1000}{10431}\tau^2(144\tau - 11)(\tau-1)^2. \end{aligned} \tag{2.58}$$

2.8.3 La paire DP(4,5)7M avec interpolant

	c	A						
f_1	0	0						
f_2	$\frac{1}{5}$	$\frac{1}{5}$	0					
f_3	$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$	0				
f_4	$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$	0			
f_5	$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$	0		
f_6	1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$	0	
f_7	1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	
$\hat{y}_{n+1}$	$\hat{b}^T$	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$
y_{n+1}	b^T	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0
$y_{n+0.5}$		$\frac{5783653}{57600000}$	0	$\frac{466123}{1192500}$	$-\frac{41347}{1920000}$	$\frac{16122321}{339200000}$	$-\frac{7117}{20000}$	$\frac{183}{10000}$

(2.59)

Paire de Dormand–Prince 7M d'ordre 4 et 5 à 7 stages

L'interpolant d'Hermite d'ordre 4 est alors

$$\begin{aligned} u_0(x_n+\tau h_n) &= d_0(\tau)y_n + d_1(\tau)h_n f_1 + d_2(\tau)y_{n+1} \\ &\quad + d_3(\tau)f_1^{n+1} + d_4(\tau)y_{n+0.5} \end{aligned} \tag{2.60}$$

avec les polynômes d'interpolation d'Hermite

$$\begin{aligned} d_0(\tau) &= (\tau-1)^2(1-2\tau)(4\tau+1) \\ d_1(\tau) &= \tau(\tau-1)^2(1-2\tau) \\ d_2(\tau) &= \tau^2(1-2\tau)(4\tau-5) \\ d_3(\tau) &= \tau^2(2\tau-1))(\tau-1) \\ d_4(\tau) &= 16\tau^2(\tau-1)^2. \end{aligned} \tag{2.61}$$

L'interpolant d'Hermite–Birkhoff d'ordre 5 s'obtient alors par le processus du boot-strapping d'Enright et al.:

$$\begin{aligned} u_1(x_n+\tau h_n) &= d_0(\tau)y_n + d_1(\tau)h_n f_1 + d_2(\tau)y_{n+1} \\ &\quad + d_3(\tau)h_n f_1^{n+1} + d_4(\tau)h_n f_8 + d_5(\tau)h_n f_9, \end{aligned} \tag{2.62}$$

où

$$f_8 = f(x_n + 0.05h_n, u_0(x_n + 0.05h_n))$$

et

$$\begin{aligned} u_1(x_n + 0.05h_n) &= \frac{9747}{10000}y_n - \frac{27}{2500}y_{n+1} + \frac{361}{10000}y_{n+0.5} \\ &\quad + h_n\left(\frac{3249}{80000}f_1 + \frac{171}{80000}f_1^{n+1}\right), \end{aligned}$$

$$f_9 = f(x_n + 0.95h_n, u_0(x_n + 0.95h_n))$$

et

$$\begin{aligned} u_1(x_n + 0.95h_n) &= -\frac{27}{2500}y_n + \frac{9747}{10000}y_{n+1} + \frac{361}{10000}y_{n+0.5} \\ &\quad - h_n\left(\frac{171}{80000}f_1 + \frac{3249}{80000}f_1^{n+1}\right), \end{aligned}$$

avec les mêmes poylnômes d'Hermite–Birkhoff (2.58) que pour DP6M puisqu'on interpole aux mêmes points.

2.8.4 La paire DP(4,5)7C avec interpolant

	c	A						
f_1	0	0						
f_2	$\frac{1}{5}$	$\frac{1}{5}$	0					
f_3	$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$	0				
f_4	$\frac{6}{13}$	$\frac{264}{2197}$	$-\frac{90}{2197}$	$\frac{840}{2197}$	0			
f_5	$\frac{2}{3}$	$\frac{932}{3645}$	$-\frac{14}{27}$	$\frac{3256}{5103}$	$\frac{7436}{25515}$	0		
f_6	1	$-\frac{367}{513}$	$\frac{30}{19}$	$\frac{9940}{5643}$	$-\frac{29575}{8208}$	$\frac{6615}{3344}$	0	
f_7	1	$\frac{35}{432}$	0	$\frac{8500}{14553}$	$-\frac{28561}{84672}$	$\frac{405}{704}$	$\frac{19}{196}$	
$\hat{y}_{n+1}$	$\hat{b}^T$	$\frac{11}{108}$	0	$\frac{6250}{14553}$	$-\frac{2197}{21168}$	$\frac{81}{176}$	$\frac{171}{1960}$	$\frac{1}{40}$
y_{n+1}	b^T	$\frac{35}{432}$	0	$\frac{8500}{14553}$	$-\frac{28561}{84672}$	$\frac{405}{704}$	$\frac{19}{196}$	0
$y_{n+0.5}$		$\frac{39\,893}{864\,000}$	0	$\frac{11\,654}{14\,553}$	$-\frac{106\,789\,579}{196\,344\,000}$	$\frac{455\,463}{1\,408\,000}$	$-\frac{47\,519}{1\,960\,000}$	$-\frac{39}{2\,500}$

(2.63)

Paire de Dormand–Prince 7C d'ordre 4 et 5 à 7 stages

L'interpolant d'Hermite d'ordre 4 est le même que pour DP7M:

$$\begin{aligned} u_0(x_n + \tau h_n) &= d_0(\tau)y_n + d_1(\tau)h_n f_1 + d_2(\tau)y_{n+1} \\ &\quad + d_3(\tau)f_1^{n+1} + d_4(\tau)y_{n+0.5} \end{aligned} \tag{2.64}$$

avec les mêmes polynômes d'interpolation d'Hermite (2.61).

L'interpolant d'Hermite–Birkhoff d'ordre 5 s'obtient alors par le processus du boot-strapping d'Enright et al.:

$$\begin{aligned} u_1(x_n+\tau h_n) &= d_0(\tau)y_n + d_1(\tau)h_n f_1 + d_2(\tau)y_{n+1} \\ &\quad + d_3(\tau)f_1^{n+1} + d_4(\tau)h_n f_8 + d_5(\tau)h_n f_9, \end{aligned} \tag{2.65}$$

où

$$f_8 = f(x_n + 0.90h_n, u_0(x_n + 0.90h_n))$$

et

$$\begin{aligned} u_1(x_n+0.90h_n) &= -\frac{23}{625}y_n + \frac{567}{625}y_{n+1} + \frac{81}{625}y_{n+0.5} \\ &\quad -h_n\left(\frac{9}{1250}f_1 + \frac{81}{1250}f_1^{n+1}\right), \end{aligned}$$

$$f_9 = f(x_n + 0.95h_n, u_0(x_n + 0.95h_n))$$

et

$$\begin{aligned} u_1(x_n+0.95h_n) &= -\frac{27}{2500}y_n + \frac{9\,747}{10\,000}y_{n+1} + \frac{361}{10\,000}y_{n+0.5} \\ &\quad -h_n\left(\frac{171}{80\,000}f_1 + \frac{3\,249}{80\,000}f_1^{n+1}\right), \end{aligned}$$

avec les polynômes d'Hermite–Birkhoff

$$\begin{aligned} d_0(\tau) &= \frac{1}{46}(240\tau^3 - 375\tau^2 + 92\tau + 46)(\tau-1)^2 \\ d_1(\tau) &= \frac{1}{15\,732}\tau(19\,440\tau^2 - 34\,975\tau + 15\,732)(\tau-1)^2 \\ d_2(\tau) &= -\frac{1}{46}\tau^2(240\tau^3 - 855\tau^2 + 1082\tau - 513) \\ d_3(\tau) &= \frac{1}{92}\tau^2(5\,840\tau^2 - 10\,635\tau + 4\,617)(\tau-1) \\ d_4(\tau) &= -\frac{125}{207}\tau^2(144\tau - 133)(\tau-1)^2 \\ d_5(\tau) &= -\frac{2000}{437}\tau^2(32\tau - 27)(\tau-1)^2. \end{aligned} \tag{2.66}$$

2.8.5 La paire DP(4,5)7S avec interpolant

	c	A						
f_1	0	0						
f_2	$\frac{2}{9}$	$\frac{2}{9}$	0					
f_3	$\frac{1}{3}$	$\frac{1}{12}$	$\frac{1}{4}$	0				
f_4	$\frac{5}{9}$	$\frac{55}{324}$	$-\frac{25}{108}$	$\frac{50}{81}$	0			
f_5	$\frac{2}{3}$	$\frac{83}{330}$	$-\frac{13}{22}$	$\frac{61}{66}$	$-\frac{9}{110}$	0		
f_6	1	$-\frac{19}{28}$	$-\frac{9}{4}$	$\frac{1}{7}$	$-\frac{27}{7}$	$\frac{22}{7}$	0	
f_7	1	$\frac{19}{200}$	0	$\frac{3}{5}$	$-\frac{243}{400}$	$\frac{33}{40}$	$\frac{7}{80}$	
$\hat{y}_{n+1}$	$\hat{b}^T$	$\frac{431}{5000}$	0	$\frac{333}{500}$	$-\frac{7857}{10000}$	$\frac{957}{1000}$	$\frac{193}{2000}$	$-\frac{1}{50}$
y_{n+1}	b^T	$\frac{19}{200}$	0	$\frac{3}{5}$	$-\frac{243}{400}$	$\frac{33}{40}$	$\frac{7}{80}$	0
$y_{n+0.5}$		$\frac{140621}{2000000}$	0	$\frac{150003}{200000}$	$-\frac{3797037}{4000000}$	$\frac{271887}{400000}$	$-\frac{1987}{800000}$	$-\frac{483}{10000}$

(2.67)

Paire de Dormand–Prince 7S d'ordre 4 et 5 à 7 stages

L'interpolant d'Hermite d'ordre 4 est le même que pour DP7M:

$$\begin{aligned} u_0(x_n+\tau h_n) &= d_0(\tau)y_n + d_1(\tau)h_n f_1 + d_2(\tau)y_{n+1} \\ &\quad + d_3(\tau)f_1^{n+1} + d_4(\tau)y_{n+0.5} \end{aligned} \tag{2.68}$$

avec les mêmes polynômes d'interpolation d'Hermite (2.61).

L'interpolant d'Hermite–Birkhoff d'ordre 5 s'obtient alors par le processus du boot-strapping d'Enright et al.:

$$\begin{aligned} u_1(x_n+\tau h_n) &= d_0(\tau)y_n + d_1(\tau)h_n f_1 + d_2(\tau)y_{n+1} \\ &\quad + d_3(\tau)f_1^{n+1} + d_4(\tau)h_n f_8 + d_5(\tau)h_n f_9, \end{aligned} \tag{2.69}$$

où

$$f_8 = f(x_n + 0.86h_n, u_0(x_n + 0.86h_n))$$

et

$$\begin{aligned} u_1(x_n+0.86h_n) &= -\frac{48951}{781250}y_n + \frac{648999}{781250}y_{n+1} + \frac{90601}{390625}y_{n+0.5} \\ &\quad - h_n\left(\frac{18963}{1562500}f_1 + \frac{116487}{1562500}f_1^{n+1}\right), \end{aligned}$$

$$f_9 = f(x_n + 0.94h_n, u_0(x_n + 0.94h_n))$$

et

$$\begin{aligned} u_1(x_n+0.94h_n) &= -\frac{11781}{781250}y_n+\frac{753269}{781250}y_{n+1}+\frac{19881}{390625}y_{n+0.5} \\ &\quad -h_n\left(\frac{4653}{1562500}f_1+\frac{72897}{1562500}f_1^{n+1}\right), \end{aligned}$$

avec les polynômes d'Hermite–Birkhoff

$$\begin{aligned} d_0(\tau) &= \frac{1}{521}(3000\tau^3-4500\tau^2+1042\tau+521)(\tau-1)^2 \\ d_1(\tau) &= \frac{1}{1052941}\tau(1406500\tau^2-2435375\tau+1052941)(\tau-1)^2 \\ d_2(\tau) &= -\frac{1}{521}\tau^2(2\tau-3)(1500\tau^2-3000\tau+2021) \\ d_3(\tau) &= \frac{1}{10941}\tau^2(406500\tau^2-690625\tau+295066)(\tau-1) \\ d_4(\tau) &= \frac{15625}{313642}\tau^2(880\tau-799)(\tau-1)^2 \\ d_5(\tau) &= -\frac{15625}{146922}\tau^2(720\tau-559)(\tau-1)^2. \end{aligned} \tag{2.70}$$

2.8.6 La paire DP(4,5)7M avec interpolant de Calvo–Montijano–Randez

M. Calvo, J. I. Montijano et L. Randez ont construit un interpolant d'Hermite $u_1(x_n+\tau h_n)$ d'ordre 5 pour les formules DP7M. On le donne sous forme de tableau:

$$\begin{array}{c|c} \vec{c} & A \\ \hline 1 & \vec{b}^T \\ \hline c_8 & \vec{a_8}^T \\ c_9 & \vec{a_9}^T \end{array} \tag{2.71}$$

Voici les valeurs des composantes des deux dernières lignes du tableau:

$$c_8=\frac{1277}{6000},\qquad c_9=\frac{643}{1500};$$

$$
\begin{array}{ll}
a_{81} = +.9883325946430815\,D-01 & b_1(\tau_1) = +.6980896127696493\,D-01 \\
a_{82} = +.6820075031771576\,D-01 & b_2(\tau_1) = 0 \\
a_{83} = +.6068825543094545\,D-01 & b_3(\tau_1) = -.1112396135095057\,D+00 \\
a_{84} = -.4711877279672667\,D-01 & b_4(\tau_1) = +.1467355407207904\,D-01 \\
a_{85} = +.3359191935282495\,D-01 & b_5(\tau_1) = -.1050001669088840\,D-01 \\
a_{86} = -.1840000000000000\,D-01 & b_6(\tau_1) = +.5664429110099508\,D-02 \\
a_{87} = +.1703792156426571\,D-01 & b_7(\tau_1) = -.4623134348070029\,D+00 \\
 & b_8(\tau_1) = +.2784380423115428\,D+00
\end{array}
$$

et

$$
\begin{array}{ll}
a_{91} = +.3224588250952791\,D-01 & b_1(\tau_2) = +.8452120256722232\,D-01 \\
a_{92} = +.2154582553348263\,D+00 & b_2(\tau_2) = 0 \\
a_{93} = -.5655554913580628\,D-01 & b_3(\tau_2) = +.1886217586996790\,D+00 \\
a_{94} = +.8148046202216559\,D-01 & b_4(\tau_2) = +.5008740324293839\,D+00 \\
a_{95} = -.4129025308081867\,D-01 & b_5(\tau_2) = -.2451081744470823\,D+00 \\
a_{96} = +.1666666666666667\,D-01 & b_6(\tau_2) = +.9864265620489278\,D-01 \\
a_{97} = -.2267213098322813\,D-01 & b_7(\tau_2) = -.5168670915601017\,D-01 \\
a_{98} = +.2033333333333333\,D+00 & b_8(\tau_2) = +.1326752337019144\,D+00 \\
 & b_9(\tau_2) = +.1734600000000000\,D+00
\end{array}
$$

Le polynôme d'interpolation d'Hermite cherché s'écrit:

$$
\begin{aligned}
u_1(x_n + \tau h_n) &= d_0(\tau) y_n + d_1(\tau) h_n f_1 + d_2(\tau) y_{n+1} \\
&\quad + d_3(\tau) h_n f_1^{n+1} + d_4(\tau) y_8 + d_5(\tau) y_9 .
\end{aligned} \tag{2.72}
$$

Les polynômes d'Hermite d_j interpolent à 0, $\tau_1 = 109/450$, $\tau_2 = 441/500$ et 1:

$$
\begin{aligned}
d_0(\tau) &= \frac{1}{2310628761}(\tau-1)^2(500\tau-441)(450\tau-109)(349088\tau+48069) \\
d_1(\tau) &= \frac{1}{48069}\tau(\tau-1)^2(500\tau-441)(450\tau-109) \\
d_2(\tau) &= -\frac{1}{404774161}\tau^2(500\tau-441)(450\tau-109)(237288\tau-257407) \\
d_3(\tau) &= \frac{1}{20119}\tau^2(500\tau-441)(450\tau-109)(\tau-1) \\
d_4(\tau) &= -\frac{369056250000}{3977438001119}\tau^2(\tau-1)^2(500\tau-441) \\
d_5(\tau) &= \frac{625000000000}{1949049491319}\tau^2(\tau-1)^2(450\tau-109)
\end{aligned} \tag{2.73}
$$

Le terme de l'erreur de la solution locale y_{n+1} d'une méthode d'ordre p est la différence

$$
y_{n+1} - y_n(x_{n+1}) = O(h_n^{p+1}), \qquad y_n = y_n(x_n). \tag{2.74}
$$

On borne le permier terme de l'erreur:

$$|h_n^{p+1}\psi(x_n,,y_n)| \leq \eta h_n^{p+1} M_{p+1}, \tag{2.75}$$

où la constante M_p dépend de la méthode et du problème et la constante η dépend de la méthode. On considère que la précision d'une méthode est meilleure si la constante η est plus petite. Dans le cas d'un interpolant continu, on minimise η sur $\tau \in [0,1]$. On présente dans le tableau 1 la valeur de η pour la norme maximum sur ψ. On indique aussi la longueur des intervalles de stabilité absolue $[\sigma, 0]$.

TABLEAU 1. Valeurs de η pour les solutions numériques et les interpolants.

	RKF6	*DP6M*	*DP7M*	*DP7C*	*DP7S*
$\hat{y}_{n+1}$	0.1538	0.0880	0.0970	0.0135	$0,0187$
y_{n+1}	0.6538	0.1500	0.2000	0.3077	0.4444
$y_{n+0.5}$	*s/o*	*s/o*	0.0123	0.0781	0.0644
$y_{n+0.6}$	0.0319	0.0366	*s/o*	*s/o*	*s/o*
$u_0, 0 \leq \tau \leq 1$	0.0436	0.0484	0.0179	0.0796	0.0662
$u_0, 0 \leq \tau \leq 2$	7.7115	8.1403	6.4089	11.0006	10.1213
$u_1, 0 \leq \tau \leq 1$	0.6574	0.1500	0.2000	0.3077	0.4444
$u_1, 0 \leq \tau \leq 2$	9.9533	11.2393	10.0135	17.0586	15.4927
$[-\sigma, 0]$	$[-3.7, 0]$	$[-3.8, 0]$	$[-3.3, 0]$	$[-4.4, 0]$	$[-5.7, 0]$

On remarque que pour DP6M, 7M, 7C et 7S l'erreur maximum en module de u_1 sur $[0,1]$ est à $\tau = 1$. On a la même chose pour DP7MH dans la norme euclidienne.

De plus on remarque que pour DP7M avec $\tau_1 = 0.90$ et $\tau_2 = .0.95$ les valeurs de η pour u_1 sont 0.2015 sur $0 \leq \tau \leq 1$, et 8.9227 sur $0 \leq \tau \leq 2$.

2.9 Mise en œuvre sur ordinateur

On présente quelques résultats numériques pour DETEST non raide pour les problèmes A1, A2, A4 et D3

Tableau 2. L'erreur globale de l'interpolant u_1 à 10 points entre x_n et x_{n+1} divisée par la plus grande des deux erreurs de la solution numérique y_n et y_{n+1} pour les problèmes non raides A1, A2, A4 et D3 de DETEST. Les astérisques indiquent l'erreur maximum de u_1 dans les unités de TOL.

		Tolérance							
Mét.	Prob.	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}	10^{-10}
F6	$A1$	6.114	4.487	2.586	1.473	1.484	1.437	1.391	1.301
	$*A1$	*0.644	*0.243	*0.265	*0.290	*0.313	*0.330	*0.343	*0.353
	$A2$	1.161	1.149	1.112	1.073	1.051	1.039	1.040	1.043
	$A4$	1.001	1.156	1.071	1.048	1.024	1.025	1.032	1.037
	$D3$	1.094	1.050	1.070	1.024	1.018	1.015	1.017	1.014
6M	$A1$	1.906	1.579	1.586	1.441	1.377	1.307	1.328	1.266
	$*A1$	*0.179	*0.037	*0.038	*0.040	*0.053	*0.065	*0.075	*0.082
	$A2$	1.047	5.361	23.401	941.899	59.511	50.751	51.931	58.409
	$*A2$	*0.092	*0.088	*0.133	*0.218	*0.270	*0.312	*1.014	*1.851
	$A4$	1.336	2.213	1.298	1.196	1.151	2.156	1.450	1.209
	$*A4$	*0.245	*0.136	*0.144	*0.261	*0.261	*0.233	*0.214	*0.193
	$D3$	1.078	1.029	1.054	1.043	1.031	1.019	1.013	1.008
7M	$A1$	1.861	1.613	1.618	1.284	1.208	1.164	1.135	1.118
	$*A1$	*1.404	*0.156	*0.161	*0.168	*0.172	*0.175	*0.1778	*0.179
	$A2$	1.097	1.073	1.089	1.117	1.404	1.911	2.659	3.301
	$*A2$	*8.563	*5.309	*0.853	*0.648	*0.481	*0.405	*0.332	*0.269
	$A4$	1.018	1.011	1.009	1.025	1.050	1.088	1.143	1.261
	$D3$	1.157	1.032	1.036	1.029	1.032	1.034	1.030	1.025
7C	$A1$	1.188	2.270	1.096	1.171	1.050	1.033	1.027	1.027
	$*A1$	*0.822	*0.868	*0.957	*1.183	*1.471	*1.764	*2.015	*2.229
	$A2$	1.228	1.092	1.212	1.168	1.118	1.079	1.052	1.034
	$A4$	1.117	1.104	1.482	1.075	1.072	1.411	1.050	1.062
	$D3$	1.258	1.059	1.032	1.024	1.016	1.010	1.006	1.004
7S	$A1$	1.993	1.172	1.169	1.171	1.170	1.171	1.166	1.144
	$*A1$	*1.214	*1.040	*1.094	*1.250	*1.430	*1.597	*1.730	*1.833
	$A2$	1.271	1.204	1.336	1.469	1.634	1.658	1.634	1.598
	$*A2$	*2.898	*2.402	*3.106	*0.526	*0.311	*0.286	*0.285	*0.289
	$A4$	1.133	5.492	1.097	1.056	1.031	1.048	1.031	1.033
	$*A4$	*0.388	*0.412	*1.499	*3.162	*4.338	*4.925	*4.667	*4.212
	$D3$	1.120	1.051	1.034	1.023	1.017	1.012	1.009	1.006
7MH	$A1$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$A2$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.009
	$A4$	1.000	1.029	1.024	1.043	1.017	1.026	1.016	1.033
	$D3$	1.158	1.042	1.036	1.012	1.005	1.002	1.001	1.007

Chapitre 3

THÉORIE DES MÉTHODES MULTIPAS

3.1 Méthodes multipas générales

Soit le problème à valeur initiale:

$$y' = f(x, y), \qquad y(a) = \eta, \tag{3.1}$$

où f est continue en x et lipschitzienne en y sur $[a, b] \times (-\infty, \infty)$. Alors la solution exacte $y(x)$ existe et est unique sur $[a, b]$.
On cherche une solution approchée $\{y_n\}$ de y aux points $x_n = a + nh$ où h est le pas et $n = (b - a)/h$.
Pour ce faire, on considére la *méthode linéaire à k pas*:

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f_{n+j}, \tag{3.2}$$

où $y_n \approx y(x_n)$ et $f_n := f(x_n, y_n)$. On normalise par la condition $\alpha_k = 1$ et l'on stipule que le nombre de pas est bien k par la condition $(\alpha_0, \beta_0) \neq (0, 0)$.
On choisit k *valeurs initiales* $y_0, y_1, \ldots, y_{k-1}$. La méthode est *explicite* si $\beta_k = 0$; on obtient alors y_{n+1} directement; elle est *implicite* si $\beta_k \neq 0$; dans ce cas il faut résoudre par la récurrence:

$$y_{n+k}^{[s+1]} = h\beta_k f\left(x_{n+k}, y_{n+k}^{[s]}\right) + g, \qquad y_{n+k}^{[0]} \text{ arbitraire.} \tag{3.3}$$

On a noté $g = g(x_n, \ldots, x_{n+k-1}, y_0, \ldots, y_{n+k-1})$. Par le théorème 1.3.2, p. 5, (3.3) converge lorsque $s \to \infty$, à condition que $0 \leq M < 1$ où M est la constante

de Lipschitz du second membre de (3.3) par rapport à y_{n+k}. Si la constante de Lipschitz de f par rapport à y est L, alors

$$M := Lh|\beta_k| < 1 \tag{3.4}$$

et il y a convergence si

$$h < \frac{1}{L|\beta_k|}.$$

3.2 Dérivation par un développement de Taylor

Soit le développement de Taylor:

$$y(x_n + h) = y(x_n) + hy^{(1)}(x_n) + \frac{h^2}{2!}y^{(2)}(x_n) + \cdots.$$

Les deux premiers termes donnent une *approximation* de la solution exacte:

$$y(x_n + h) \approx y(x_n) + hf(x_n, y(x_n)); \tag{3.5}$$

l'erreur locale de méthode est

$$\frac{h^2}{2!}y^2(x_n) + \frac{h^3}{3!}y^{(3)}(x_n) + \cdots. \tag{3.6}$$

D'autre part, (3.5) donne une *relation exacte* pour la solution approchée y_n:

$$y_{n+1} = y_n + hf_n. \tag{3.7}$$

C'est la méthode de *Euler* dont l'erreur de méthode est d'ordre $O\left(h^2\right)$.
On obtient la *méthode du point milieu*:

$$y_{n+2} - y_n = 2hf_{n+1} \tag{3.8}$$

de la même façon. Pour ce faire, on fait le développement:

$$y(x_n + h) = y(x_n) + hy^{(1)}(x_n) + \frac{h^2}{2!}y^2(x_n) + \frac{h^3}{3!}y^{(3)}(x_n) + \cdots$$

$$y(x_n - h) = y(x_n) - hy^{(1)}(x_n) + \frac{h^2}{2!}y^{(2)}(x_n) - \frac{h^3}{3!}y^{(3)}(x_n) + \cdots$$

et l'on soustrait:

$$y(x_n + h) - y(x_n - h) = 2hy^{(1)}(x_n) + \frac{1}{3}h^3y^3(x_n) + \cdots.$$

Si l'on tronque après le terme en h, on obtient (3.8) pour y_n et l'erreur locale est

$$\pm\frac{1}{3}h^3y^3(x_n)+\cdots.$$

Finalement on dérive la méthode à un pas implicite la plus précise:

$$y_{n+1}+\alpha_0 y_n = h(\beta_1 f_{n+1}+\beta_0 f_n).$$

On écrit

$$y(x_n+h)+\alpha_0 y(x_n) \approx h[\beta_1 y^{(1)}(x_n+h)+\beta_0 y^{(1)}(x_n)] \tag{3.9}$$

et l'on choisit les $\alpha_0, \beta_0, \beta_1$ qui donnent une approximation aussi précise que possible. On utilise les développements:

$$y(x_n+h) = y(x_n)+hy^{(1)}(x_n)+\frac{h^2}{2!}y^{(2)}(x_n)+\frac{h^3}{3!}y^{(3)}(x_n)+\cdots$$

et

$$y^{(1)}(x_n+h) = y^{(1)}(x_n)+hy^{(2)}(x_n)+\frac{h^2}{2!}y^{(3)}(x_n)+\cdots$$

dans (3.9); on regroupe les termes au premier membre

$$\begin{aligned}
&y(x_n)+hy^{(1)}(x_n)+\frac{h^2}{2!}y^{(2)}(x_n)+\cdots+\alpha_0 y(x_n)\\
&-h\beta_1[y^{(1)}(x_n)+hy^{(2)}(x_n)+\frac{h^2}{2!}y^{(3)}(x_n)+\cdots]\\
&-h\beta_0[y^{(1)}(x_n)]+\cdots=0,
\end{aligned}$$

c'est-à-dire

$$\begin{aligned}
&(1+\alpha_0)y(x_n)+(1-\beta_1-\beta_0)hy^{(1)}(x_n)+\\
&\left(\frac{1}{2}-\beta_1\right)h^2y^{(2)}(x_n)+\left(\frac{1}{6}-\frac{1}{2}\beta_1\right)h^3y^{(3)}(x_n)+\cdots=0.
\end{aligned}$$

Comme on a 3 paramètres, α_0, β_0 et β_1, on peut annuler au moins les 3 premiers termes. On obtient ainsi le système linéaire

$$\begin{aligned}
C_0 &= 1+\alpha_0=0,\\
C_1 &= 1-\beta_1-\beta_0=0,\\
C_2 &= \frac{1}{2}-\beta_1=0.
\end{aligned}$$

On voit facilement que

$$\alpha_0=-1,\quad \beta_1=\frac{1}{2},\quad \beta_0=\frac{1}{2}.$$

On a donc obtenu la méthode des trapèzes

$$y_{n+1} - y_n = \frac{h}{2}(f_{n+1} + f_n). \tag{3.10}$$

Le quatrième terme,

$$C_3 = \frac{1}{6} - \frac{1}{2}\beta_1 = -\frac{1}{12},$$

ne s'annule pas et l'erreur locale est

$$\pm 12h^3 y^{(3)}(x_n) + \cdots.$$

3.3 Dérivation par l'intégration numérique

On dérive *la méthode de Simpson* au moyen de la formule

$$y(x_{n+2}) - y(x_n) = \int_{x_n}^{x_{n+2}} y'(x)\,dx = \int_{x_n}^{x_{n+2}} f(x, y(x))\,dx. \tag{3.11}$$

Soit $P(x)$ l'unique polynôme de degré 2 passant par les trois points du plan:

$$(x_n, f_n), \quad (x_{n+1}, f_{n+1}), \quad (x_{n+2}, f_{n+2});$$

la formule d'interpolation de Newton–Gregory donne

$$P(x) = P(x_n + rh) = f_n + r\Delta f_n + \frac{r(r-1)}{2!}\Delta^2 f_n.$$

Donc, avec $x = hr + x_n$ et $dx = h\,dr$, on a

$$\begin{aligned}
\int_{x_n}^{x_{n+2}} y'(x)dx &\approx \int_{x_n}^{x_{n+2}} P(x)\,dx \\
&= h\int_0^2 [f_n + r\Delta f_n + \frac{1}{2}r(r-1)\Delta^2 f_n]\,dr \\
&= h[2f_n + 2\Delta f_n + \frac{1}{3}\Delta^2 f_n].
\end{aligned}$$

C'est la méthode de Simpson:

$$y_{n+2} - y_n = \frac{h}{3}[f_{n+2} + 4f_{n+1} + f_n]. \tag{3.12}$$

Si l'on remplace (3.11) par

$$y(x_{n+2}) - y(x_{n+1}) = \int_{x_{n+1}}^{x_{n+2}} y'(x)\,dx$$

et si l'on utilise $P(x)$ comme ci-dessus, on obtient la méthode à 2 pas d'Adams-Moulton:

$$y_{n+2} - y_{n+1} = \frac{h}{12}(5f_{n+2} + 8f_{n+1} - f_n). \tag{3.13}$$

3.4 Dérivation par l'interpolation

On dérive de nouveau la méthode de Simpson en approximant la solution $y(x)$ sur $[x_n, x_{n+2}]$ par un polynôme $I(x)$ tel que

$I(x)$ interpole les points (x_{n+j}, y_{n+j}), $\quad j = 0, 1, 2,$

et

$I'(x_{n+j})$ coïncide avec f_{n+j} pour $j = 0, 1, 2$. Un tel polynôme d'interpolation s'appelle un intepolant d'Hermite.

On a donc

$$I(x_{n+j}) = y_{n+j}, \quad I'(x_{n+j}) = f_{n+j}, \qquad j = 0, 1, 2. \tag{3.14}$$

Il y a six conditions en tout. On prend donc pour $I(x)$ un polynôme du $4^{ième}$ degré, qui contient cinq paramètres:

$$I(x) = ax^4 + bx^3 + cx^2 + dx + e;$$

ceci nous permettra d'éliminer les cinq paramètres des six équations. Alors $I(x)$ dans (3.14) donne (3.12). Voici le détail des calculs. On suppose sans perte que $x_{-1} = -h$, $x_0 = 0$ et $x_1 = h$; alors

$$\begin{array}{lllll}
(1) & I(x_{-1}) & = ah^4 - bh^3 + ch^2 - dh + e & = y_{-1} \\
(2) & I(x_0 & = e & = y_0 \\
(3) & I(x_1) & = ah^4 + bh^3 + ch^2 + dh + e & = y_1 \\
(4) & I'(x_{-1}) & = -4ah^3 + 3bh^2 - 2ch + d & = f_{-1} \\
(5) & I'(x_0) & = d & = f_0 \\
(6) & I'(x_1) & = 4ah^3 + 3bh^2 + 2ch + d & = f_1.
\end{array}$$

On combine ces équations de la façon suivante:

$$\begin{array}{llll}
(7) & = (3) - (1): & 2abh^3 + 2dh & = y_1 - y_{-1} \\
(8) & = (6) + 4(5) + (4): & 6bh^3 + 4d + 4d & = f_1 + 4f_0 + f_1.
\end{array}$$

On voit que $(7) - h(8)/3$ donne la méthode de Simpson:

$$y_2 - y_0 = \frac{h}{3}[f_2 + 4f_1 + f_0].$$

On remarque que l'on n'a pas utilisé l'équations (2). De plus, on remarque que si $y(x)$ n'est pas bien approchée par un tel polynôme $I(x)$, la méthode à pas multiples donnera un mauvais rendement.

Il existe un autre genre d'approximation de la solution, appelée *approximation spline*, c'est-à-dire une approximation par polynômes par morceaux, lisse aux nœuds. Par exemple, une approximation spline au moyen de polynômes du second degré donne la méthode des trapèzes sur chaque intervalle $[x_j, x_{j+1}]$. De même, la méthode de Simpson est la meilleure approximation spline par polynômes par morceaux de degré 3.

3.5 Convergence

On doit définir avec soin la convergence de la solution numérique $\{y_n\}$:

$$y_n \to y(x^*)$$

lorsque $h \to 0$ tel que $nh = b - a =$ constante.
Par exemple, si $x^* = (b-a)/2$ et $h_0 = (b-a)/2$, alors avec $h_n = h_0/n$ on dit que $y_n \to y(x^*)$ si

$$y_1(h_0), y_2(h_0/2), \cdots, y_n(h_0/n) \to y(x^*).$$

Il faut remarquer que

$$x^* = n(h_0/n) = h_0 = (b-a)/2$$

est un point fixé.

La définition de la convergence doit tenir compte des valeurs initiales additionnelles $y_1, y_2, \ldots, y_{k-1}$ pour $k \geq 2$. Enfin, la propriété de convergence de (3.2) doit être valide pour tous les problèmes à valeurs initiales (3.1):

$$y' = f(x,y), \qquad y(a) = \eta,$$

où f satisfait les hypothèses d'existence et d'unicité du théorème (1.1.1).

Définition 3.1 La méthode (3.2),

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f_{n+j},$$

est dite *convergente*, si pour tous problèmes à valeurs initiales (3.1), satisfaisant les hypothèses du théorème (1.1.1),

$$\lim_{h\to 0} y_n = y(x_n), \qquad nh = x - a, \tag{3.15}$$

pour tout $x \in [a,b]$ et pour toutes solutions de (3.2) satisfaisant les conditions initiales $y_\mu = \eta_\mu(h)$ pour lesquelles

$$\lim_{h\to 0} \eta_\mu(h) = \eta, \qquad \mu = 0, 1, 2, \cdots, k-1. \tag{3.16}$$

3.6 Ordre et constante de l'erreur

Avec la méthode linéaire à pas multiples (3.2), on associe l'opérateur aux différences:

$$\mathcal{L}[y(x);h] = \sum_{j=0}^{k} [\alpha_j y(x+jh) - h\beta_j y'(x+jh)] \tag{3.17}$$

où $y(x)$ est une function arbitraire infiniment dérivable sur [a,b]. Si l'on développe $y(x+jh)$ et $y'(x+jh)$ en séries de Taylor de centre x, $\mathcal{L}$ devient:

$$\mathcal{L}[y(x);h] = C_0 y(x) + C_1 h y^{(1)}(x) + \cdots + C_q h^q y^{(q)}(x) + \cdots, \tag{3.18}$$

où les C_q sont des constantes.

Définition 3.2 *L'opérateur* (3.17) et la méthode (3.2) aux différences sont d'*ordre* p si les coefficients $C_0, \ldots, C_{p+1}$ du développement (3.18) satisfont les conditions suivantes:

$$C_0 = C_1 = \cdots = C_p = 0, \quad C_{p+1} \neq 0.$$

Un simple calcul donne:

$$\begin{aligned} C_0 &= \alpha_0 + \alpha_1 + \cdots + \alpha_k \\ C_1 &= \alpha_1 + 2\alpha_2 + \cdots + k\alpha_k - (\beta_0 + \beta_1 + \cdots + \beta_k) \\ &\vdots \\ C_q &= \frac{1}{q!}(\alpha_1 + 2^q\alpha_2 + \cdots + k^q\alpha_k) \\ &\quad -\frac{1}{(q-1)!}(\beta_1 + 2^{q-1}\beta_2 + \cdots + k^{q-1}\beta_k) \quad q = 2, 3, \cdots \end{aligned} \tag{3.19}$$

Remarque 3.1 On peut utiliser (3.19) pour dériver des méthodes (3.2) de structure donnée et d'ordre maximal.

Exemple 3.1 Construire une méthode linéaire implicite à 2 pas d'ordre maximal, contenant un paramètre libre, et trouver son ordre.

Solution. Puisque $k = 2$, alors $\alpha_2 = +1$. Soit $\alpha_0 = a$ le paramètre libre. Il reste quatre paramètres indéterminés:

$$\alpha_1, \quad \beta_0, \quad \beta_1, \quad \beta_2.$$

On peut donc annuler les quatre premiers coefficients:

$$C_0 = a + \alpha_1 + 1 = 0,$$

$$\begin{aligned} C_1 &= \alpha_1 + 2 - (\beta_0 + \beta_1 + \beta_2) = 0, \\ C_2 &= \frac{1}{2!}(\alpha_1 + 4) - (\beta_1 + 2\beta_2) = 0, \\ C_3 &= \frac{1}{3!}(\alpha_1 + 8) - \frac{1}{2!}(\beta_1 + 4\beta_2) = 0. \end{aligned}$$

On résout:

$$\begin{aligned} \alpha_1 &= -1 - a, \\ \beta_0 &= -\frac{1}{12}(1 + 5a), \\ \beta_1 &= \frac{2}{3}(1 - a), \\ \beta_2 &= \frac{1}{12}(5 + a). \end{aligned}$$

Ainsi

$$y_{n+2} - (1+a)y_{n+1} + ay_n = \frac{h}{12}[(5+a)f_{n+2} + 8(1-a)f_{n+1} - (1+5a)f_n] \quad (3.20)$$

et

$$\begin{aligned} C_4 &= \frac{1}{4!}(\alpha_1 + 16) - \frac{1}{3!}(\beta_1 + 8\beta_2) = -\frac{1}{4!}(1+a), \\ C_5 &= \frac{1}{5!}(\alpha_1 + 32) - \frac{1}{4!}(\beta_1 + 16\beta_2) = -\frac{1}{3 \times 5!}(17 + 13a). \end{aligned}$$

Si $a \neq -1$, alors $C_4 \neq 0$ et (3.20) est d'ordre 3.
Si $a = -1$, alors $C_4 = 0$, $C_5 \neq 0$ et (3.20) est d'ordre 4: c'est la méthode de Simpson.

Remarque 3.2 Si $a = 0$, (3.20) est la méthode d'Adams-Moulton à 2 pas (3.13); si $a = -5$, la méthode est explicite.

Question. L'ordre de (3.17) et de (3.2) est-il indépendant du développement de Taylor autour de $x + th$, t réel, au lieu de x comme dans (3.18)?

Posons

$$\mathcal{L}[y(x), h] = D_0 y(x+th) + D_1 h y^{(1)}(x+th) + \cdots + D_q h^q y^{(q)}(x+th) + \cdots. \quad (3.21)$$

On développe le second membre à x:

$$y^{(q)}(x+th) = y^{(q)}(x) + thy^{(q+1)}(x) + \cdots + \frac{t^s h^s}{s!} y^{(q+s)}(x), \quad q = 0, 1, 2, \cdots,$$

où $y^{(0)}(x) := y(x)$.
On substitue dans (3.21) et l'on identifie les coefficients avec ceux de (3.18):

$$\begin{aligned} C_0 &= D_0 \\ C_1 &= D_1 + tD_0 \\ C_2 &= D_2 + tD_1 + \frac{t^2}{2!}D_0 \\ &\vdots \\ C_p &= D_p + tD_{p-1} + \cdots + \frac{t^p}{p!}D_0 \\ C_{p+1} &= D_{p+1} + tD_p + \cdots + \frac{t^{p+1}}{(p+1)!}D_0. \end{aligned} \tag{3.22}$$

Alors

$$C_0 = C_1 = \cdots = C_p = 0$$

si et seulement si

$$D_0 = D_1 = \cdots = D_p = 0.$$

Dans ce cas

$$D_{p+1} = C_{p+1}$$

et

$$D_{p+2} = C_{p+2} - tC_{p+1}.$$

On voit que le premier coefficient non nul, D_{p+1}, est indépendant de t.
On peut aussi voir, par un simple calcul, que

$$\begin{aligned} D_0 &= \alpha_0 + \alpha_1 + \cdots + \alpha_k \\ D_1 &= -t\alpha_0 + (1-t)\alpha_1 + (2-t)\alpha_2 + \cdots + (k-t)\alpha_k - (\beta_0 + \beta_1 + \beta_2 + \cdots + \beta_k), \\ &\vdots \\ D_q &= \frac{1}{q!}[(-t)^q\alpha_0 + (1-t)^q\alpha_1 + (2-t)^q\alpha_2 + \cdots + (k-t)^q\alpha_k] \\ &\quad - \frac{1}{(q-1)!}[(-t)^{q-1}\beta_0 + (1-t)^{q-1}\beta_1 + (2-t)^{q-1}\beta_2 + \cdots + (k-t)^{q-1}\beta_k], \\ &\qquad q = 2, 3, \cdots. \end{aligned} \tag{3.23}$$

Remarque 3.3 Un choix approprié de t peut faciliter la résolution de (3.23), et par conséquent, le développement de formules (3.2).

Exemple 3.2 Redévelopper la méthode de l'exemple précédent au moyen du développement de Taylor de centre $x + h$, c'est-à-dire avec $t = 1$.

Solution. On a $k = 2, \alpha_2 = 1, \alpha_0 = a$ et on prend $t = 1$ dans (3.23), alors:

$$\begin{aligned}
D_0 &= a + \alpha_1 + 1 = 0 \\
D_1 &= -a + 1 - (\beta_0 + \beta_1 + \beta_2) = 0 \\
D_2 &= \frac{1}{2!}(a+1) - (-\beta_0 + \beta_2) = 0 \\
D_3 &= \frac{1}{3!}(-a+1) - \frac{1}{2!}(\beta_0 + \beta_2) = 0 \\
D_4 &= \frac{1}{4!}(a+1) - \frac{1}{3!}(-\beta_0 + \beta_2) \\
D_5 &= \frac{1}{5!}(-a+1) - \frac{1}{4!}(\beta_0 + \beta_2).
\end{aligned}$$

Le système est plus facile à résoudre: α_1 et β_1 n'apparaissant que dans une seule équation. La solution est la même qu'à l'exemple 1.

Pour l'ordre 3: $D_4 = C_4$ mais $D_5 \neq C_5$

Pour l'ordre 4: $(a = -1), D_5 = C_5$.

Définition 3.3 C_{p+1} est la constante de l'erreur.

Rappel: Le choix du facteur de normalisation $\alpha_k = +1$ implique que la constante C_{p+1} est uniquement déterminée.

3.7 Erreurs de méthode locale et globale

Définition 3.4 L'erreur locale de la méthode (3.2),

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f_{n+j},$$

à x_{n+k}, est définie par (3.17):

$$\mathcal{L}[y(x_n); h] = \sum_{j=0}^{k} [\alpha_j y(x + jh) - h\beta_j y'(x + jh)] =: T_{n+k},$$

où $y(x)$ est la solution théorique de (3.1),

$$y' = f(x, y), \qquad y(a) = \eta.$$

Remarque 3.4 L'*erreur locale* T_{n+1} est l'erreur de la solution numérique de $y' = f(x, y)$, $y(x_n) = y_n$, c'est-à-dire on fait l'*hypothèse de localisation* qui consiste à considérer la valeur intiale y_n comme exacte.

De (3.17) il vient:

$$\begin{aligned}\sum_{j=0}^{k}\alpha_j y(x_n+jh) &= h\sum_{j=0}^{k}\beta_j y'(x_n+jh)+\mathcal{L}[y(x_n);h]\\ &= h\sum_{j=0}^{k}\beta_j f(x_n+jh,y(x_n+jh))+\mathcal{L}[y(x_n);h].\end{aligned}$$

Or

$$\sum_{j=0}^{k}\alpha_j y_{n+j}=h\sum_{j=0}^{k}\beta_j f(x_{n+j},y_{n+j}).$$

Donc, si l'on soustrait ces deux expressions, membre à membre, et l'on utilise l'hypothèse de localisation:

$$y_{n+j}=y(x_n+jh),\qquad j=0,\cdots,k-1,$$

on obtient

$$\begin{aligned}y(x_{n+k})-y_{n+k} &= h\beta_k[f(x_{n+k},y(x_{n+k})-f(x_{n+k},y_{n+k})]+\mathcal{L}[y(x_n);h]\\ &= h\beta_k\frac{\partial f}{\partial y}(x_{n+k},\eta_{n+k})[y(x_{n+k})-y_{n+k}]+\mathcal{L}[y(x_n);h].\end{aligned}$$

Ainsi

$$\left[1-h\beta_k\frac{\partial f}{\partial y}(x_{n+k},\eta_{n+k})\right][y(x_{n+k})-y_{n+k}]=\mathcal{L}[y(x_n);h]=T_{n+k}. \tag{3.24}$$

Remarque 3.5 Pour une méthode explicite $\beta_k=0$ et

$$T_{n+k}=y(x_{n+k})-y_{n+k}.$$

Remarque 3.6 Si $y(x)\in C^{p+2}$, (3.24) implique:

$$y(x_{n+k})-y_{n+k}=C_{p+1}h^{p+1}y^{(p+1)}(x_n)+O\left(h^{p+2}\right), \tag{3.25}$$

où $C_{p+1}h^{p+1}y^{(p+1)}(x_n)$ est le *terme principal de l'erreur* locale de la méthode d'ordre p et C_{p+1} est la constante de ce terme.

Sans l'hypothèse de localisation, l'erreur *globale* est

$$e_{n+k}:=y(x_{n+k})-y_{n+k}.$$

On rappelle que la convergence signifie: $e_{n+k}\to 0$ lorsque $h\to 0, n\to\infty$ et $nh=x_n-a$ est fixé.

Exemple 3.3 Vérifier (3.24) pour $y' = Ay$, $y(0) = 1$, à l'aide de

(i) la méthode d'Euler: $y_{n+1} - y_n = hf_n$;

(ii) la méthode des trapèzes: $y_{n+1} - y_n = \frac{h}{2}(f_{n+1} + f_n)$.

Solution théorique. La solution théorique est $y = e^{Ax}$. Pour la méthode d'Euler:

$$\begin{aligned}
\mathcal{L}[y(x);h] &= y(x_n+h) - y(x_n) - hy'(x_n) \\
&= \frac{1}{2!}h^2y^{(2)}(x_n) + \frac{1}{3!}h^3y^{(3)}(x_n) + \cdots + \frac{1}{q!}h^qy^{(q)}(x_n) + \cdots \\
&= e^{nhA}\left[\frac{1}{2!}h^2A^2 + \frac{1}{3!}h^3A^3 + \cdots + \frac{1}{q!}h^qA^q + \cdots\right] \\
&= e^{nhA}\left[e^{hA} - 1 - hA\right].
\end{aligned}$$

De plus, par l'hypothèse de localisation: $y_n = y(x_n)$, on a

$$y_{n+1} = y_n + hAy_n = (1 + hA)e^{nhA}.$$

Alors

$$\begin{aligned}
y(x_{n+1}) - y_{n+1} &= e^{(n+1)hA} - (1 + hA)e^{nhA} \\
&= e^{nhA}[e^{hA} - 1 - hA] \\
&= \mathcal{L}[y(x_n);h].
\end{aligned}$$

Puisque $\beta_1 = 0$, (3.24) est vérifiée.

Pour la méthode des trapèzes, lire Lambert, pp. 29–30.

3.8 Consistance et zéro stabilité

On définit la consistance et la zéro stabilité d'une méthode multipas et l'on cite le théorème de Dahlquist à l'effet que la méthode est convergente si et seulement si elle est consistante et stable.

Définition 3.5 La méthode (3.2) est *consistante* si elle est d'ordre $p \geq 1$.

Autrement dit, (3.2) est consistante si et seulement si les constantes C_0 et C_1 de (3.19) sont nulles:

$$C_0 = \sum_{j=0}^{k} \alpha_j = 0, \tag{3.26}$$

$$C_1 = \sum_{j=0}^{k} j\alpha_j - \sum_{j=0}^{k} \beta_j = 0. \tag{3.27}$$

Dans le corollaire qui suit, on souligne le rôle clé de la consistance.

Corollaire 3.1 La convergence implique la consistance.

Démonstration. Pour (3.26), si

$$y_{n+j} \to y(x), \qquad j = 0, \cdots, k,\ n \to \infty,\ nh = x - a;$$

alors

$$y(x) = y_{n+j} + \theta_{j,n}(h), \qquad j = 0, 1, \cdots, k,$$

où

$$\lim_{\substack{n \to \infty \\ nh = x-a}} \theta_{j,n}(h) = 0, \qquad j = 0, \cdots, k.$$

Ainsi

$$\sum_{j=0}^{k} \alpha_j y(x) = \sum_{j=0}^{k} \alpha_j y_{n+j} + \sum_{j=0}^{k} \alpha_j \theta_{j,n}(h),$$

c'est-à-dire

$$y(x) \sum_{j=0}^{k} \alpha_j = h \sum_{j=0}^{k} \beta_j f_{n+j} + \sum_{j=0}^{k} \alpha_j \theta_{j,n}(h).$$

A la limite chacun des termes du second membre s'annule. Donc $y(x) \sum \alpha_j = 0$ est indépendant de h, c'est-à-dire le premier membre $\equiv 0$. Puisque $y(x) \not\equiv 0$, alors

$$\sum \alpha_j = 0.$$

On remarque que (3.26) présuppose seulement que y_n tende vers une *certaine* function $y(x)$; on ne fait référence à aucune équations différentielle. Mais pour (3.27) on doit faire intervenir l'équation différentielle à résoudre. Si $y_n \to y(x)$ solution de (3.1), alors

$$\frac{y_{n+j} - y_n}{jh} \to y'(x), \qquad j = 1, 2, \cdots, k,$$

c'est-à-dire

$$y_{n+j} - y_n = jhy'(x) + jh\phi_{j,n}(h),$$

où

$$\phi_{j,n}(h) \to 0;$$

alors

$$\sum_{j=0}^{k} \alpha_j y_{n+j} - \sum_{j=0}^{k} \alpha_j y_n = h \sum_{j=0}^{k} j\alpha_j y'(x) + h \sum_{j=0}^{k} j\alpha_j \phi_{j,n}(h).$$

Par (3.2)

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f_{n+j}.$$

Alors on a

$$h\sum_{j=0}^{k}\beta_j f_{n+j} - y_n\sum_{j=0}^{k}\alpha_j = hy'(x)\sum_{j=0}^{k}j\alpha_j + h\sum_{j=0}^{k}j\alpha_j\phi_{j,n}(h).$$

On divise par h et on utilise la relation $\sum_{j=0}^{k}\alpha_j = 0$:

$$\sum_{j=0}^{k}\beta_j f_{n+j} = y'(x)\sum_{j=0}^{k}j\alpha_j + \sum_{j=0}^{k}j\alpha_j\phi_{j,n}(h)$$

A la limite

$$f_{n+j} \to f(x, y(x)) \quad \text{et} \quad \phi_{j,n}(h) \to 0.$$

On obtient

$$f(x, y(x))\left(\sum_{j=0}^{k}\beta_j\right) = y'(x)\left(\sum_{j=0}^{k}j\alpha_j\right).$$

Donc $y(x)$ satisfaisant (3.1) $\Rightarrow \sum j\alpha_j = \sum\beta_j$. □

Conclusion. Si la solution numérique y_n converge vers la solution exacte $y(x)$, alors les conditions (3.26) et (3.27) doivent être satisfaites, c'est-à-dire une méthode *convergente* (3.2) est nécessairement *consistante*.

La réciproque: consistance et stabilité $\Rightarrow$ convergence, est un théorème puissant dû à Dahlquist.

Il est utile d'introduire les 1^{ier} et $2^{ième}$ polynômes caractéristiques de (3.2):

$$\rho(\zeta) = \sum_{j=0}^{k}\alpha_j\zeta^j, \quad \sigma(\zeta) = \sum_{j=0}^{k}\beta_j\zeta^j. \tag{3.28}$$

Remarque. De (**??**) il vient que (3.2) est consistante si et seulement si

$$\rho(1) = 0 \qquad \rho'(1) = \sigma(1). \tag{3.29}$$

Alors pour une méthode consistante, $\rho(\zeta)$ admet toujours la racine $+1$, appelée *racine principale* et notée ζ_1. Les autres racines, $\zeta_s, s = 2, 3, \cdots k$, sont des racines parasites. Elles surviennent seulement si $k > 1$. Leur position doit être contrôlée avec soin pour assurer la convergence.

Exemple 3.4 $y' = 0, \quad y(0) = 0$, admet la solution $y(x) \equiv 0$. Puisque $f \equiv 0$, (3.2) devient

$$\sum_{j=0}^{k}\alpha_j y_{n+j} = 0. \tag{3.30}$$

Si toutes les racines ζ_s de $\rho(\zeta)$ sont réelles et distinctes, la solution de l'équation aux différences (3.30) est

$$y_n = h(d_1\zeta_1^n + \cdots + d_k\zeta_k^n)$$

et $y_\mu \to y(0)$ lorsque $h \to 0$ pour $\mu = 0, \cdots, k-1$. Puisque $nh = x$ est fixé,

$$\lim_{h\to 0} h\zeta_s^n = x \lim_{n\to\infty} \frac{\zeta_s^n}{n} = 0 \Leftrightarrow \mid \zeta_s \mid \leq 1.$$

Si $\rho(\zeta)$ admet une racine réelle de multiplicité $m > 1$, alors

$$y_n = h[d_{s,1} + d_{s,2}n + d_{s,3}n(n-1) + \ldots + d_{s,m}n(n-1)\cdots(n-m+2)]\zeta_s^n + \ldots.$$

Pour $q \geq 1$, et $nh = x$ fixé,

$$\lim_{h\to 0} hn^q\zeta_s^n = x \lim_{n\to\infty} n^{q-1}\zeta_s^n = 0 \Leftrightarrow \mid \zeta_s \mid < 1.$$

Ceci est valide aussi pour les racines complexes. □

Remarque 3.7 Puisque la consistance contrôle seulement la position de la racine principale, alors elle n'implique pas, en général, la convergence.

Définition 3.6 La méthode (3.2) est *zéro stable* si aucune racine du premier polynôme caractéristique $\rho(\zeta)$ n'a un module plus grand que 1 et toute racine de module 1 est simple. On dit aussi que (3.2) satisfait la *condition sur les racines.*

On dit *zéro* stable puisqu'il s'agit ici de la stabilité de la méthode quand $h \to 0$, et dans ce cas les solutions parasites $\to 0$ lorsque $h \to 0$.

Remarque 3.8 Pour les méthodes à un pas, $\zeta_1 = +1$ est l'unique racine de $\rho(\zeta)$; dans ce cas une méthode consistante à un pas est zéro stable.

Pour les méthodes zéro stables à k pas, $\sigma(1) \neq 0$, puisque par (3.29) $\rho'(1) = \sigma(1)$ et donc $\rho'(1) = 0 = \rho(1) \Rightarrow 1$ est une racine double de $\rho(\zeta)$.

Théorème 3.1 (Dahlquist) *Pour qu'une méthode linéaire à pas multiples soit convergente il faut et il suffit qu'elle soit consistante et zéro stable.*

Remarque. La consistance contrôle l'erreur locale; la zéro stabilité contrôle la propagation des erreurs.

Exemple 3.5 Illustrer l'effet de la zéro stabilité de la méthode

$$y_{n+2} - (1+a)y_{n+1} + ay_n = \frac{h}{2}[(3-a)f_{n+1} - (1+a)f_n]$$

avec

$$(i) \qquad a = 0, \qquad \text{donc d'ordre } 2,$$

et

$$(ii) \qquad a = -5, \qquad \text{donc d'ordre } 3,$$

appliquée à l'équation différentielle

$$y' = 4xy^{1/2}, \qquad y(0) = 1, \qquad 0 \leq x \leq 2,$$

dont la solution exacte est

$$y(x) = (1 + x^2)^2.$$

Résolution. On a

$$\rho(\zeta) = \zeta^2 - (1 - a)\zeta + a = (\zeta - 1)(\zeta - a).$$

On prend $y_0 = 1$, $y_1 = (1 + h^2)^2$ et on fait les calculs avec $h = 0.1$, $h = 0.05$ et $h = 0.025$. Lorsque h diminue pour x fixé, la solution s'améliore dans (i) mais elle s'appauvrit dans (ii).

Exemple 3.6 Montrer l'effet de la non-consistance en utilisant

$$y_{n+2} - y_{n+1} = \frac{1}{3}h(3f_{n+1} - 2f_n).$$

Résolution. Du fait que

$$\rho(1) := \sum_{j=0}^{k} \alpha_j = 1 - 1 + 0 = 0,$$

il suit que la solution numérique converge à une fonction, mais celle-ci n'est pas une solution; en effet l'inconsistance de la méthode provient du fait que $\rho'(1) \neq \sigma(1)$:

$$\sum_{j=0}^{k} j\alpha_j = 2 \times 1 - 1 \times 1 + 0 \times 0 = 1, \qquad \sum_{j=0}^{k} \beta_j = \frac{3}{3} - \frac{2}{3} = \frac{1}{3}.$$

D'autre part la méthode est stable et la solution n'explose pas du fait que

$$\rho(\zeta) = \zeta^2 - \zeta = \zeta(\zeta - 1).$$

3.9 Ordre maximum des méthodes zéro stables

Pour un ordre donné k, on choisit les α_j et les β_j pour atteindre un ordre raisonnablement élevé: on satisfait automatiquement la consistance mais la zéro stabilité est une barrière.

La méthode générale implicite à k pas possède $2k+2$ coefficients:

$$\alpha_j,\ \beta_j, \qquad j = 0, \cdots, k.$$

Puisque $\alpha_k = +1$, il reste $2k+1$ coefficients libres. Pour la méthode explicite, $\beta_k = 0$; il reste donc $2k$ coefficients libres.
Donc de la définition 3.2, pour avoir l'ordre p on doit satisfaires $p+1$ conditions linéaires.

Une méthode est d'*ordre maximal* si l'on a l'ordre le plus élevé: $2k$ si la méthode est implicite et $2k-1$ si elle est explicite.
Mais le théorème de Dahlquist (1956) établit une barrière:

Théorème 3.2 *L'ordre d'une méthode linéaire zéro stable à k pas ne peut excéder $k+1$ lorsque k est impair et $k+2$ lorsque k est pair.*

Définition 3.7 Une méthode linéaire zéro stable à k pas est d'*ordre optimal* si elle est d'ordre $k+2$.

Remarque 3.9 On peut montrer que les zéros de $\rho(\zeta)$ se situent sur le cercle unité si la méthode est optimale.

Remarque 3.10 La méthode de Simpson a une position unique: elle est à 2 pas $k = 2$; elle est maximale d'ordre $2k$ et optimale d'ordre $k+2$.

On verra au chapitre suivant qu'une méthode optimale souffre de certains désavantages pour les calculs, d'autre part les méthodes au nombre de pas impair restent utiles.

Exemple 3.7 Trouver la méthode linéaire à pas multiples optimale avec $k = 4$.

Solution. Toutes les racines de $\rho(\zeta)$ sont sur le cercle unité. Le degré de $\rho(\zeta) = 4$ et $\zeta_1 = +1$. Puisque les coefficients sont réels, la seconde racine réelle $\zeta_2 = -1$. Donc on a

$$\zeta_1 = +1, \quad \zeta_2 = -1, \quad \zeta_3 = e^{i\varphi}, \quad \zeta_4 = e^{-i\varphi}, \qquad 0 < \varphi < \pi,$$

et par conséquent

$$\begin{aligned} \rho(\zeta) &= (\zeta-1)(\zeta+1)(\zeta-e^{i\varphi})(\zeta-e^{-i\varphi}) \\ &= \zeta^4 - 2(\cos\varphi)\zeta^3 + 2(\cos\varphi)\zeta - 1. \end{aligned}$$

On note $\cos\varphi = \mu$; alors

$$\alpha_4 = +1, \quad \alpha_3 = -2\mu, \quad \alpha_2 = 0, \quad \alpha_1 = 2\mu, \quad \alpha_0 = -1.$$

Puisque l'ordre est $k+2$ et vu la symétrie des α_j, il est plus approprié de définir l'ordre au moyen des coefficients D_q donnés par (3.23) avec $t = 2$:

$$\begin{aligned}
D_0 &= \alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0, \\
D_1 &= -2\alpha_0 - \alpha_1 + \alpha_3 + 2\alpha_4 - (\beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4), \\
&\vdots \\
D_q &= \frac{1}{q!}[(-2)^q\alpha_0 + (-1)^q\alpha_1 + \alpha_3 + 2^q\alpha_4] \\
&\quad \frac{1}{(q-1)!}[(-2)^{q-1}\beta_0 + (-1)^{q-1}\beta_1 + \beta_3 + 2^{q-1}\beta_4], \\
&\qquad q = 2, 3, \ldots .
\end{aligned}$$

On voit que l'équation $D_0 = 0$ est satisfaite. On pose $D_q = 0, q = 2, 3, \ldots, 6$, ce qui donne 5 équations pour les 4 paramètres β_0, β_1, β_3, β_4:

$$\begin{aligned}
-2\beta_0 - \beta_1 + \beta_3 + 2\beta_4 &= 0 \\
2^2\beta_0 + \beta_1 + \beta_3 + 2^2\beta_4 &= \frac{2}{3}(2^3 - 2\mu) \\
-2^3\beta_0 - \beta_1 + \beta_3 + 2^3\beta_4 &= 0 \\
2^4\beta_0 + \beta_1 + \beta_3 + 2^4\beta_4 &= \frac{2}{5}(2^5 - 2\mu) \\
-2^5\beta_0 - \beta_1 + \beta_3 + 2^5\beta_4 &= 0.
\end{aligned}$$

Les $1^{\text{ière}}$, $3^{\text{ième}}$ et $5^{\text{ième}}$ équations sont satisfaites avec $\beta_1 = \beta_3$ et $\beta_0 = \beta_4$. Les $2^{\text{ième}}$ et $4^{\text{ième}}$ équations sont:

$$\begin{aligned}
4\beta_0 + \beta_1 &= \frac{1}{3}(8 - 2\mu), \\
16\beta_0 + \beta_1 &= \frac{1}{5}(32 - 2\mu).
\end{aligned}$$

On a donc les solutions:

$$\begin{aligned}
\beta_0 &= \frac{1}{45}(14 + \mu) \quad = \beta_4, \\
\beta_1 &= \frac{1}{45}(64 - 34\mu) = \beta_3.
\end{aligned}$$

Si $D_1 = 0$, alors

$$\beta_2 = \frac{1}{15}(8 - 38\mu).$$

La constante du terme principal de l'erreur est:

$$\begin{aligned} D_7 &= \frac{1}{7!}\left(-2^7\alpha_0 - \alpha_1 + \alpha_3 + 2^7\alpha_4\right) - \frac{1}{6!}\left(2^6\beta_0 + \beta_1 + \beta_3 + 2^6\beta_4\right) \\ &= \frac{2}{7!}\left(2^7 - 2\mu\right) - \frac{2}{6!}\left(2^6\beta_0 + \beta_1\right) \\ &= -\frac{16 + 5\mu}{1890}. \end{aligned}$$

Puisque $\mu = \cos\varphi$, $0 < \varphi < \pi$, alors $-1 < \mu < 1$ et $D_7 \neq 0$; donc l'odre est ≤ 6. De plus, $\beta_4 \neq 0$, donc la méthode est implicite. D_7 atteint son minimum lorsque $\mu \to -1$, mais alors -1 devient une racine triple, c'est-à-dire la méthode devient instable.

Puisque D_7 varie seulement par un facteur de 2, la position de μ n'est pas critique. On remarque les deux cas suivants:

$$\mu = 0 \Rightarrow \alpha_1 = 0, \alpha_3 = 0,$$

et

$$\mu = \frac{4}{19} \Rightarrow \beta_2 = 0, \qquad \text{méthode de Quade.}$$

3.10 Spécification des méthodes multipas

Au temps des calculatrices on considérait le développement selon les différences

$$y_{n+1} - y_n = h(1 - \frac{1}{2}\nabla - \frac{1}{12}\nabla^2 - \frac{1}{24}\nabla^3 - \cdots)f_{n+1}. \tag{3.31}$$

Si l'on tronque après 2 termes, on obtient la méthode des trapèzes (3.10):

$$y_{n+1} - y_n = \frac{1}{2}h(f_{n+1} + f_n),$$

et après 3 termes, on obtient la méthode d'Adams–Moulton à 2 pas (3.13):

$$y_{n+1} - y_n = \frac{1}{12}h(5f_{n+1} + 8f_n - f_{n-1}).$$

On utilise (3.31) pour inclure des différences de f plus élevées si elles deviennent importantes à un certain point du calcul. Cette façon de varier l'ordre d'une formule zéro stable comporte un danger: $\sigma(\zeta)$ peut causer une autre espèce d'instabilité comme on verra au chapitre suivant. Aujourd'hui sur ordinateurs, on varie l'ordre de la méthode et le pas h. Les diverses formes de $\rho(\zeta)$ donnent lieu à diverses familles de méthodes.

(A) Méthode d'Adams:

$$\rho(\zeta) = \zeta^k - \zeta^{k-1},$$

Toutes les racines parasites sont à $\zeta = 0$, donc ces méthodes sont 0 stables. On nomme Adams–Bashforth les méthodes explicites et Adams–Moulton les méthodes implicites.

(B) Méthodes de Nyström:

$$\rho(\zeta) = \zeta^k - \zeta^{k-2} = \zeta^{k-2}(\zeta^2 - 1)$$

On nomme Nyström les méthodes explicites et Milne-Simpson généralisées les méthodes implicites. Elles sont zéro stables: en effet il y a une racine parasite simple à $\zeta = -1$ et les autres sont à $\zeta = 0$.

Méthodes explicites

$k = 1$:

$$\begin{aligned} &\alpha_1 = 1, \\ &\alpha_0 = -1, \quad \beta_0 = 1, \\ &p = 1; \qquad C_{p+1} = \tfrac{1}{2}. \end{aligned}$$

$k = 2$:

$$\begin{aligned} &\alpha_2 = 1, \\ &\alpha_1 = -1 - a, \quad \beta_1 = \tfrac{1}{2}(3 - a), \\ &\alpha_0 = a, \qquad \beta_0 = \tfrac{1}{2}(-1 + a), \\ &p = 2; \qquad C_{p+1} = \tfrac{1}{12}(5 + a). \end{aligned}$$

La zéro stabilité limite l'ordre à 2.

$k = 3$:

$$\begin{aligned} &\alpha_3 = 1, \\ &\alpha_2 = -1 - a, \quad \beta_2 = \tfrac{1}{12}(23 - 5a - b), \\ &\alpha_1 = a + b, \qquad \beta_1 = \tfrac{1}{3}(-4 - 2a + 2b), \\ &\alpha_0 = -b, \qquad \beta_0 = \tfrac{1}{12}(5 + a + 5b), \\ &p = 3; \qquad C_{p+1} = \tfrac{1}{24}(9 + a + b). \end{aligned}$$

La zéro stabilité limite l'ordre à 3.

$k = 4$:

$$\begin{aligned}
&\alpha_4 = 1, \\
&\alpha_3 = -1 - a, \quad \beta_3 = \tfrac{1}{24}(55 - 9a - b - c), \\
&\alpha_2 = a + b, \quad \beta_2 = \tfrac{1}{24}(-59 - 19a + 13b - 19c), \\
&\alpha_1 = -b - c, \quad \beta_1 = \tfrac{1}{24}(37 + 5a + 13b - 19c), \\
&\alpha_0 = c, \quad \beta_0 = \tfrac{1}{24}(-9 - a - b - 9c), \\
&p = 4; \quad C_{p+1} = \tfrac{1}{720}(251 + 19a + 11b + 19c).
\end{aligned}$$

La zéro stabilité limite l'ordre à 4.

Méthodes implicites

$k = 1$:

$$\begin{aligned}
&\alpha_1 = 1, \quad \beta_1 = \tfrac{1}{2}, \\
&\alpha_0 = -1, \quad \beta_0 = \tfrac{1}{2}, \\
&p = 2; \quad C_{p+1} = -\tfrac{1}{12}.
\end{aligned}$$

$k = 2$:

$$\begin{aligned}
&\alpha_2 = 1, \quad \beta_2 = \tfrac{1}{12}(5 + a), \\
&\alpha_1 = -1 - a, \quad \beta_1 = \tfrac{2}{3}(1 - a), \\
&\alpha_0 = a, \quad \beta_0 = \tfrac{1}{12}(-1 - 5a), \\
&\text{Si } a \neq -1, \quad p = 3; \quad C_{p+1} = -\tfrac{1}{24}(1 + a), \\
&\text{Si } a = -1, \quad p = 4; \quad C_{p+1} = -\tfrac{1}{90}.
\end{aligned}$$

$k = 3$:

$$\begin{aligned}
&\alpha_3 = 1, \quad \beta_3 = \tfrac{1}{24}(9 + a + b), \\
&\alpha_2 = -1 - a, \quad \beta_2 = \tfrac{1}{24}(19 - 13a - 5b), \\
&\alpha_1 = a + b, \quad \beta_1 = \tfrac{1}{24}(-5 - 13a + 19b), \\
&\alpha_0 = -b, \quad \beta_0 = \tfrac{1}{24}(1 + a + 9b), \\
&p = 4; \quad C_{p+1} = -\tfrac{1}{720}(19 + 11a + 19b).
\end{aligned}$$

La zéro stabilité limite l'ordre à 4.

$k = 4$:

$$\begin{aligned}
\alpha_4 &= 1, & \beta_4 &= \tfrac{1}{720}(251 + 19a + 11b + 19c),\\
\alpha_3 &= -1 - a, & \beta_3 &= \tfrac{1}{360}(323 - 173a - 37b - 53c),\\
\alpha_2 &= a + b, & \beta_2 &= \tfrac{1}{30}(-11 - 19a + 19b + 11c),\\
\alpha_1 &= -b - c, & \beta_1 &= \tfrac{1}{360}(53 + 37a + 173b - 323c),\\
\alpha_0 &= c, & \beta_0 &= \tfrac{1}{720}(-19 - 11a - 19b - 251c).
\end{aligned}$$

Si $27 + 11a + 11b + 27c \neq 0$, alors

$$p = 5; \qquad C_{p+1} = -\frac{1}{1440}(27 + 11a + 11b + 27c).$$

Si $27 + 11a + 11b + 27c = 0$, alors

$$p = 6; \qquad C_{p+1} = -\frac{1}{15\,120}(74 + 10a - 10b - 74c).$$

La zéro stabilité limite l'ordre à 6.

Chapitre 4

MISE EN ŒUVRE DES MÉTHODES MULTIPAS

4.1 Quelques difficultés

On fait d'abord les quatre considérations qui suivent.

(i) Il est assez facile de trouver les valeurs initiales additionnelles nécessaires y_μ, $\mu = 1, 2, \cdots, k-1$, $k > 1$.

(ii) Il est difficile de choisir une valeur appropriée pour le pas h.

(iii) Si la méthode est implicite, comment résoudre l'équation non-linéaire pour y_{n+k}?

(iv) Enfin qu'elle est la précision de la solution obtenue?

Dans ce chapitre, on essaiera de répondre à ces questions de mise en œuvre.

4.2 Les valeurs initiales

Soient le problème

$$y' = f(x, y), \qquad y(a) = \eta,$$

et la méthode d'ordre p:

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f_{n+j}, \qquad \alpha_k = +1.$$

Posons $y_0 = \eta$. On cherche des valeurs initiales à l'ordre de la méthode près:

$$y_\mu - y(x_\mu) = O\left(h^{p+1}\right), \qquad \mu = 1, \ldots, k-1. \tag{4.1}$$

On peut recourir à l'**algorithme de Taylor d'ordre** p. C'est une méthode explicite à un pas:

$$y_1 = y(x_0) + hy^{(1)}(x_0) + \ldots + \frac{h^p}{p!}y^{(p)}(x_0). \tag{4.2}$$

On satisfait (4.1) puisque

$$y_1 - y(x_1) = \frac{h^{p+1}}{(p+1)!}y^{(p)}(\xi), \qquad x_0 < \xi < x_1.$$

On obtient les dérivées apparaissant dans (4.2) en dérivant l'équation différentielle:

$$\begin{aligned} y(x_0) &= y_0, \\ y^{(1)}(x_0) &= f(x_0, y_0), \\ y^{(2)}(x_0) &= [\partial_x f + f\partial_y f]\big|_{x=x_0, y=y_0}, \\ y^{(3)}(x_0) &= [\partial_x^2 f + 2f\partial_{xy}^2 f + f^2\partial_y^2 f + \partial_x f \partial_y f + f(\partial_y f)^2]\big| x = x_0, y = y_0. \\ &\vdots \end{aligned} \tag{4.3}$$

De cette façon, on obtient les autres valeurs intiales:

$$y_\mu = y(x_{\mu-1}) + hy^{(1)}(x_{\mu-1}) + \cdots + \frac{h^p}{p!}h^{(p)}(x_{\mu-1}), \qquad \mu = 2, 3, \ldots, k-1.$$

La méthode de Taylor n'est pas sans inconvénients: certaines dérivées de f peuvent ne pas exister et la dérivation peut être un travail fastidieux.

Il y a en deuxième lieu les **méthodes d'Obrechkoff à un pas**: Ce sont, en général, des méthodes implicites multipas qui emploient des dérivées totales élevées de y:

$$\alpha_1 y_{n+1} + \alpha_0 y_n = \sum_{s=1}^{l} h^s(\beta_{s1}y_{n+1}^{(s)} + \beta_{s0}y_n^{(s)}), \qquad l \geq 2. \tag{4.4}$$

Avec $l = 2$, on a

$$\begin{aligned} y_{n+1} - y_n &= \frac{1}{2}h(y_{n+1}^{(1)} - y_n^{(1)}) - \frac{1}{12}h^2(y_{n+1}^{(2)} - y_n^{(2)}), \\ T_{n+1} &= \frac{1}{720}h^5 y^{(5)}(x_n) + O\left(h^6\right). \end{aligned} \tag{4.5}$$

Exemple 4.1 Soit $y' = x^{\frac{3}{2}} + y$, $y(0) = 1$. Trouver la valeur de départ y_1 la plus précise avec (4.2) et (4.5).

Enfin il y a les **méthodes de Runge–Kutta** dont on traite dans le chapitre 2.

4.3 Borne de l'erreur locale

Soit

$$\mathcal{L}[y(x_n); h] = C_{p+1} h^{p+1} y^{(p+1)}(x_n) + O\left(h^{p+2}\right). \tag{4.6}$$

Peut-on absorber le terme $O\left(h^{p+2}\right)$ dans le premier terme en écrivant

$$\mathcal{L}[y(x_n); h] = C_{p+1} h^{p+1} y^{(p+1)}(x_n + \theta h), \qquad 0 < \theta < k? \tag{4.7}$$

On considère une forme plus générale du reste R_{p+1} du développement

$$F(a+h) = F(a) + hF^{(1)}(a) + \cdots + \frac{h^p}{p!} F^{(p)}(a) + R_{p+1},$$

soit

$$\begin{aligned} R_{p+1} &= \frac{1}{p!} \int_0^h (h-t)^p F^{(p+1)}(a+t)\, dt \\ &= \frac{F^{(p+1)}(a+\theta h)}{p!} \int_0^h (h-t)^p\, dt \\ &= \frac{h^{p+1}}{(p+1)!} F^{(p+1)}(a+\theta h), \qquad 0 < \theta < 1. \end{aligned}$$

On a donc

$$\begin{aligned} y(x_n + jh) &= y(x_n) + jhy^{(1)}(x_n) + \cdots + \frac{1}{p!} j^p h^p y^{(p)}(x_n) \\ &\quad + \frac{1}{p!} \int_0^{jh} (jh-t)^p y^{(p+1)}(x_n + t)\, dt. \end{aligned}$$

Posons $t = hs$ dans le dernier terme:

$$\frac{1}{p!} h^{p+1} \int_0^j (j-s)^p y^{(p+1)}(x_n + sh)\, ds.$$

De même:

$$\begin{aligned} y^{(1)}(x_n + jh) &= y^{(1)}(x_n) + \cdots + \frac{1}{(p-1)!} j^{p-1} h^{p-1} y^{(p)}(x_n) \\ &\quad + \frac{1}{(p-1)!} h^p \int_0^j (j-s)^{(p-1)} y^{(p+1)}(x_n + sh)\, ds. \end{aligned}$$

On substitue ces expressions dans (3.17)

$$\begin{aligned}\mathcal{L}[y(x_n);h] &= \sum_{j=0}^{k}[\alpha_j y(x+jh) - h\beta_j y'(x+jh)] \\ &= \frac{1}{p!}h^{p+1}\sum_{j=0}^{k}\alpha_j\int_0^j (j-s)^p y^{(p+1)}(x_n+sh)\,ds \\ &\quad -\frac{1}{(p-1)!}h^{p+1}\sum_{j=0}^{k}\beta_j\int_0^j (j-s)^p y^{(p+1)}(x_n+sh)\,ds.\end{aligned}$$

On note

$$z_+ = \begin{cases} z & \text{si} \quad z \geq 0, \\ 0 & \text{si} \quad z < 0. \end{cases}$$

Puisque $\mathcal{L}$ est d'ordre p, on obtient

$$\mathcal{L}[y(x_n);h] = \frac{1}{p!}\int_0^k \sum_{j=0}^{k}[\alpha_j(j-s)_+^p - p\beta_j(j-s)_+^{p-1}]y^{(p+1)}(x_n+sh)\,ds.$$

On définit la *fonction d'influence*:

$$G(s) = \sum_{j=0}^{k}\left[\alpha_j(j-s)_+^p - p\beta_j(j-s)_+^{p-1}\right]. \tag{4.8}$$

On voit que G ne dépend que des coefficients de la méthode: $\{\alpha_j, \beta_j\}$, et puisque, par la consistance, $p \geq 1$, $G(s)$ est un polynôme dans chacun des intervalles $[0,1]$, $[1,2]$, ..., $[k-1,k]$. Alors

$$\mathcal{L}[y(x_n);h] = \frac{1}{p1!}h^{p+1}\Big[\int_0^k G(s)\,ds\Big]y^{(p+1)}(x_n+\theta h), \qquad 0<\theta<k. \tag{4.9}$$

Si $G(s)$ ne change pas de signe sur l'intervalle $[0,k]$, on a, par le théorème des accroissements continus pour les intégrales:

$$\mathcal{L}[y(x_n);h] = \frac{h^{p+1}}{(p+1)!}[\int_0^k G(s)\,ds]y^{(p+1)}(x_n+\theta h), \qquad 0<\theta<k. \tag{4.10}$$

Avec $y(x) = x^{p+1}$,

$$\mathcal{L}[x_n^{p+1};h] = \frac{1}{p!}h^{p+1}\left[\int_0^k G(s)\,ds\right](p+1)!$$

et

$$\mathcal{L}[x_n^{p+1}; h] = C_{p+1} h^{p+1} (p+1)!$$

puisque les dérivées d'ordre supérieur à $p+1$ s'annulent. Enfin,

$$\frac{1}{p!} \int_0^k G(s)\, ds = C_{p+1}. \tag{4.11}$$

Alors la borne pour l'erreur locale à un pas quelconque dans $[a, b]$ est bornée par l'inégalité:

$$\left|\mathcal{L}[y(x_n); h]\right| \leq h^{p+1} G Y \tag{4.12}$$

où

$$Y = \max_{x \in [a,b]} [y^{(p+1)}(x)]$$

et

$$G = |C_{p+1}| = \frac{1}{p!} \left| \int_0^k G(s)\, ds \right|.$$

Si $G(s)$ change de signe dans $[0, k]$ on obtient par (4.9):

$$\begin{aligned} |\mathcal{L}[y(x_n); h]| &\leq \frac{1}{p!} h^{p+1} \int_0^k |G(s)|\, |y^{(p+1)}(x_n + sh)|\, ds \\ &\leq \frac{1}{p!} h^{p+1} Y \int_0^k |G(s)|\, ds. \end{aligned}$$

De nouveau on obtient la borne (4.12) où Y est comme ci-dessus et G est de la forme

$$G = \frac{1}{p!} \int_0^k |G(S)|\, ds.$$

4.4 Borne de l'erreur globale

Nous allons considérer l'erreur globale

$$e_n = y(x_n) - y_n$$

dans des cas particuliers. Nous disons qu'une méthode linéaire à pas multiples:

$$\sum_{j=0}^{k} \alpha_j y_{n+1} = h \sum_{j=0}^{k} \beta_j f_{n+j},$$

est de la *classe A* (nécessairement zéro stable) si

$$\alpha_k = 1; \qquad \alpha_j \leq 0, \quad j = 0, \ldots, k-1; \qquad \sum \alpha_j = 0.$$

La classe A inclut les méthodes d'Adams–Bashforth, d'Adams–Moulton, de Nyström et de Milne–Simpson. On cherche une borne pour une méthode explicite quelconque de la classe A. La solution théorique satisfait

$$\sum_{j=0}^{k-1} \alpha_j y(x_{n+j}) - h \sum_{j=0}^{k-1} \beta_j y'(x_{n_j}) = \mathcal{L}[y(x_n); h],$$

c'est-à-dire

$$y(x_{n+k}) = \sum_{j=0}^{k-1} [-\alpha_j y(x_{n+j}) + h\beta_j f(x_{n+j}, y(x_{n+j}))] + \mathcal{L}[y(x_n); h].$$

S'il n'y a pas d'erreurs d'arrondi:

$$y_{n+k} = \sum_{j=0}^{k-1} [-\alpha_j y_{n+j} + h\beta_j f(x_{n+j}, y_{n+j})]. \tag{4.13}$$

On soustrait:

$$e_{n+k} = \sum_{j=0}^{k-1} \Big\{ -\alpha_j e_{n+j} + h\beta_j [f(x_{n+j}, y(x_{n+j})) - f(x_{n+j}, y_{n+j})] \Big\} + \mathcal{L}[y(x_n); h].$$

Alors

$$|e_{n+k}| \le \sum_{j=0}^{k-1} \Big[|-\alpha_j| \, |e_{n+j}| + h|\beta_j| \, L \, |e_{n+j}| + |\mathcal{L}[y(x_n); h]| \Big], \tag{4.14}$$

où par (4.12)

$$|\mathcal{L}[y(x_n); h]| \le h^{p+1} GY.$$

On substitue maintenant dans (4.14):

$$|e_{n+k}| \le \sum_{j=0}^{k-1} (-\alpha_j + hL|\beta_j|)|e_{n+j}| + h^{p+1} GY, \qquad n = 0, 1, 2, \ldots. \tag{4.15}$$

Pous simplifier on pose

$$P_j = -\alpha_j + hL|\beta_j|, \qquad P = \sum_{j=0}^{k-1} P_j, \qquad Q = h^{p+1} GY. \tag{4.16}$$

Alors $\alpha_k = 1$ et $\sum_{j=0}^{k} \alpha_j = 0$ donnent:

$$P = 1 + hLB, \tag{4.17}$$

et

$$B = \sum_{j=0}^{k-1} |\beta_j|. \tag{4.18}$$

Soit

$$\delta = \max_{\mu=0,1,\ldots,k-1} |e_\mu| \tag{4.19}$$

De (4.16) à (4.19) on obtient quatre inégalités utiles pour la suite:

$$P_j \geq 0, \qquad P \geq 1, \qquad Q \geq 0, \qquad \delta \geq 0. \tag{4.20}$$

De (4.15) on déduit

$$|e_{n+k}| \leq \sum_{j=0}^{k-1} P_j |e_{n+j}| + Q, \qquad n = 0, 1, 2, \ldots. \tag{4.21}$$

Si l'on pose $n = 0$ dans (4.21), il vient de (4.19) et (4.20):

$$|e_k| \leq \sum_{j=0}^{k-1} P_j \delta + Q = P\delta + Q. \tag{4.22}$$

Si l'on pose $n = 1$ dans (4.21), il vient

$$|e_{k+1}| \leq \sum_{j=0}^{k-1} P_j |e_{j+1}| + Q. \tag{4.23}$$

Maintenant, de (4.19) il vient

$$|e_{j+1}| \leq \delta, \qquad j = 0, \ldots, k-2,$$

de (4.22) il vient

$$|e_k| \leq P\delta + Q$$

et de (4.20) il vient

$$\delta \leq P\delta + Q.$$

Alors

$$|e_{j+1}| \leq P\delta + Q, \qquad j = 0, 1, \ldots, k-1.$$

Substituant cette inégalité dans (4.23) il vient

$$|e_{k+1}| \leq \sum_{j=0}^{k-1} P_j (P\delta + Q) + Q = P^2\delta + (P+1)Q. \tag{4.24}$$

Par l'hypothèse d'induction

$$|e_{l+k}| \leq P^{l+1}\delta + Q\sum_{s=0}^{l} P^S, \qquad l = 0, \ldots, m-1. \tag{4.25}$$

On voit que (4.22) et (4.24) vérifie (4.25) pour $m = 1, 2$. Par (4.21)

$$|e_{m+k}| \leq \sum_{j=0}^{k-1} P_j |e_{m+j}| + Q. \tag{4.26}$$

Par (4.25)

$$\begin{aligned} |e_{m+j}| &\leq P^{m-k+j+1}\delta + Q\sum_{s=0}^{m-k+j} P^s, \qquad j = 0, 1, \ldots, k-1, \\ &\leq P^m\delta + Q\sum_{s=0}^{m-1} P^s. \end{aligned}$$

Donc par (4.26)

$$\left(P^m\delta + Q\sum_{s=0}^{m-1} P^s\right) + Q$$

$$= P^{m+1}\delta + Q\sum_{s=0}^{m} P^s.$$

Donc (4.25) est vraie pour $l = m$.

Conclusion.

$$\begin{aligned} |e_{n+k}| &\leq P^{n+1}\delta + Q\sum_{s=0}^{n} P^s, \qquad n = 0, 1, 3, \ldots, \\ &\leq P^{n+k}\delta + Q\sum_{j=0}^{n+k-1} P^s, \qquad n = 0, 1, 2, \ldots, \end{aligned}$$

puisque $k \geq 1$, ou bien

$$\begin{aligned} |e_n| &\leq P^n\delta + Q\sum_{s=0}^{n-1} P^s \\ &= P^n\delta + Q\frac{P^n - 1}{P - 1}, \qquad n = k, k+1, \ldots. \end{aligned}$$

La même borne est valide pour $n = 0, 1, \ldots, k-1$, puisque par (4.20)

$$|e_n| \leq \delta \leq P^n\delta + Q\frac{P^n - 1}{P - 1}, \qquad n = 0, 1, \ldots, k-1.$$

Donc

$$|e_n| \leq \delta(1 + hLB)^n + \frac{h^pGY}{LB}[(1 + hLB)^n - 1]. \tag{4.27}$$

et d'après $nh = x_n - a$:

$$(1 + hLB)^n \leq e^{nhLB} = e^{LB(x_n - a)}.$$

Alors

$$|e_n| \leq \delta\, e^{LB(x_n - a)} + \frac{h^pGY}{LB}\left[e^{LB(x_n - a)} - 1\right], \tag{4.28}$$

et par $e^z - 1 < z\, e^z$,

$$|e_n| \leq [\delta + (x_n - a)h^pGY]\, e^{LB(x_n - a)}. \tag{4.29}$$

Si les erreurs d'arrondi à chaque pas sont d'ordre h^{q+1}, alors on remplace (4.13) par

$$\tilde{y}_{n+k} = \sum_{j=0}^{k-1}[-\alpha_j\tilde{y}_{n+j} + h\beta_j f(x_{n+j}, \tilde{y}_{n+j}] + \theta K_1 h^{q+1},$$

où $|\theta| \leq 1$ et $k_1 > 0$. Et avec $\tilde{e}_n = y(x_n) - \tilde{y}_n$, on a la borne de l'erreur globale de méthode et d'arrondi:

$$|\tilde{e}_n| \leq [\delta + (x_n - a)(h^pGY + h^qK_1)]\, e^{LB(x_n - a)}. \tag{4.30}$$

Dans le cas général

$$|\tilde{e}_n| \leq \Gamma^*[Ak\delta + (x_n - a)(h^pPGY + h^qK_1)]\, e^{\Gamma^* LB(x_n - a)} \tag{4.31}$$

si $|h\beta_k\alpha_k^{-1}|L < 1$.

4.5 Commentaires sur les bornes d'erreurs

Si l'erreur initiale est de l'ordre $O\left(h^{p+1}\right)$, les erreurs dans les valeurs de démarrage ne domineront pas l'erreur globale quand $h \to 0$. Si l'erreur locale de méthode est $h^{p+1}GY$, l'erreur globale devient h^pGY. Si l'erreur locale d'arrondi est $h^{q+1}K_1$, l'erreur globale d'arrondi devient h^qK_1, et si $q > p$ il n'y a pas de problème. En pratique on a

$$|\text{erreur locale d'arrondi}| < \varepsilon;$$

alors dans (4.30) $h^pGY + \varepsilon$ et dans (4.31) $h^pGY + \varepsilon/h$. Egalement d'après (4.30) et (4.31), si $p \geq 1, q \geq 1$ et $\delta \to 0$ lorsque $h \to 0$, alors $\tilde{e}_n \to 0$ lorsque $h \to 0$ et $nh = x_n - a$.

Remarque 4.1 Si $K_1 = 0$ (on ignore les erreurs d'arrondi),

$$\tilde{e}_n = e_n = y(x_n) - y_n;$$

alors on a démontré que la consistance $(p \geq 1) \Rightarrow$ convergence pour les méthodes explicites de la classe A, (la zéro stabilité est utilisée pour établir (4.31)).

Applications. Si l'on peut estimer:

$$y = \max_{x \in [a,b]} \left| y^{(p+1)}(x) \right|,$$

on a une borne a priori de l'erreur. Mais cette borne est souvent très pessimiste.

Exemple 4.2 Considérons: $y' = \lambda y$, $y(0) = 1$, $\lambda < 0$, et appliquons la méthode d'Euler:

$$y_{n+1} - y_n = h f_n = h\lambda y_n.$$

On néglige les erreurs d'arrondi, $K_1 = 0$, et les erreurs initiales, $\delta = 0$, et l'on prend $y_0 = 1$. Alors $L = |\lambda|$, $\beta = 1$, $p = 1$ et

$$\begin{aligned} \mathcal{L}[y(x_n); h] &= \frac{1}{p!} \int_0^k G(s)\, ds \\ &= \frac{1}{1!} \int_0^1 \left[-1(0-s)_+^0 + 1(1-s)_+^1 - 1 \times 0 \right] ds \\ &= \int_0^1 (1-s)\, ds = s - \frac{s^2}{2} \Big|_0^1 = 1 - \frac{1}{2} = \frac{1}{2}. \end{aligned}$$

Donc de (4.30) il vient

$$|e_n| < \frac{1}{2} x_n h Y\, e^{|\lambda| x_n}. \tag{4.32}$$

On obtient une borne plus précise de la façon suivante. Posons:

$$y(x_{n+1}) = y(x_n) + h\lambda y(x_n) + T_{n+1},$$

où $T_{n+1} = \frac{1}{2} h^2 y^{(2)}(x_n + \theta h)$, $0 < \theta < 1$. Puisque $y^{(2)}(x) = \lambda^2 y(x)$, $y(0) > 0$ et $y(x) = 0$ seulement lorsque $y'(x) = 0$, alors $y^{(2)}(x) \geq 0$ pour tout $x \geq 0$. Donc $0 \leq T_{n+1} \leq \frac{1}{2} y^2 Y$, et

$$e_{n+1} = (1 + h\lambda) e_n + T_{n+1}.$$

Par conséquent,

$$\begin{aligned} e_0 &= 0, \\ e_1 &= T_1, \\ e_2 &= (1 + h\lambda) e_1 + T_2 = (1 + h\lambda) T_1 + T_2 \\ &\vdots \\ e_n &= (1 + h\lambda)^{n-1} T_1 + (1 + h\lambda)^{n-2} T_2 + \cdots + T_n. \end{aligned}$$

Si $h < -1/\lambda$, alors

$$\begin{aligned} e_n &\leq \frac{1}{2}h^2 y\left[(1+h\lambda)^{n-1} + \cdots + 1\right] \\ &\leq \frac{1}{2}nh^2 y = \frac{1}{2}x_n h y, \end{aligned}$$

c'est-à-dire

$$0 \leq e_n \leq \frac{1}{2}x_n h Y, \tag{4.33}$$

alors que (4.32) donne la borne

$$\frac{1}{2}x_n h Y\, e^{|\lambda| x_n}$$

qui peut être beaucoup plus grande.

Conclusion. Si l'erreur locale est $O\left(h^{p+1}\right)$, l'erreur globale est $O\left(h^p\right)$.

4.6 Théorie de la stabilité faible

Soit la méthode consistante et 0 stable

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f_{n+j} \tag{4.34}$$

La solution exacte du problème aux valeurs initiales satisfait

$$\sum_{j=0}^{k} \alpha_j y(x_{n+j}) = h \sum_{j=0}^{k} \beta_j f(x_{n+j}, y(x_{n+j})) + T_{n+k}, \tag{4.35}$$

où $T_{n+k} = \mathcal{L}[y(x_n); h]$ est erreur locale de la méthode. On note $\{\tilde{y}_n\}$ la solution de (4.34) où une *erreur d'arrondi* R_{n+k} est commise à la $n^{ième}$ application de (4.34):

$$\sum_{j=0}^{k} \alpha_j \tilde{y}_{n+j} = h \sum_{j=0}^{k} \beta_j f\left(x_{n+j}, \tilde{y}_{n+j}\right) + R_{n+k}. \tag{4.36}$$

On note $\tilde{e}_n = y(x_n) - \tilde{y}_n$, l'*erreur globale de méthode*; alors de (4.35) moins (4.36) il vient:

$$\sum_{j=0}^{k} \alpha_j \tilde{e}_{n+j} = h \sum_{j=0}^{k} \beta_j \left[f(x_{n+j}, y(x_{n+j})) - f(x_{n+j}, \tilde{y}_{n+j})\right] + \phi_{n+k},$$

où $\phi_{n+k} = T_{n+k} - R_{n+k}$. Si $f \in C'(y)$,

$$\sum_{j=0}^{k} \alpha_j \tilde{e}_{n+j} = h \sum_{j=0}^{k} \beta_j \frac{\partial f}{\partial y}(x_{n+j}, \xi_{n+j}) \tilde{e}_{n+j} + \phi_{n+k},$$

où ξ_{n+j} se trouve dans l'intervalle borné par $y(x_{n+j})$ et $\tilde{y}_{n+j}$.
On introduit deux hypothèses de simplification:

$$\begin{aligned} \frac{\partial f}{\partial y} &= \lambda, \qquad \text{une constante,} \\ \phi_n &= \phi, \qquad \text{une constante.} \end{aligned} \tag{4.37}$$

Sous ces hypothèses, l'équation aux différences pour l'erreur $\tilde{e}_n$ se linéarise:

$$\sum_{j=0}^{k} (\alpha_j - h\lambda\beta_j) \tilde{e}_{n+j} = \phi. \tag{4.38}$$

Par la section 1.2, la solution de (4.38) est

$$\tilde{e}_n = \sum_{s=1}^{k} d_s r_s^n - \frac{\phi}{h\lambda \sum_{j=0}^{k} \beta_j} \tag{4.39}$$

où les d_s sont des constantes arbitraires et les r_s sont les racines, supposées distinctes, du polynôme

$$\sum_{j=0}^{k} (\alpha_j - h\lambda\beta_j) r^j = 0, \tag{4.40}$$

c'est-à-dire

$$\pi\left(r, \tilde{h}\right) = \rho(r) - \tilde{h}\sigma(r), \tag{4.41}$$

où

$$\tilde{h} = h\lambda. \tag{4.42}$$

On appelle $\pi\left(r, \tilde{h}\right)$ le *polynôme de stabilité* de la méthode. L'intérêt de cette analyse n'est pas d'estimer $\tilde{e}_n$ mais de décider si oui ou non $\tilde{e}_n$ va croître avec n.

Définition 4.1 La méthode (4.34) est *absolument stable*, pour un $\tilde{h}$ donné, si $|r_s| < 1, s = 1, \ldots, k$, et *absolument instable* autrement.
(α, β) est l'*intervalle de stabilité absolue* réel si (4.34) est absolument stable pour tout $\tilde{h} \in (\alpha, \beta)$.
Si (4.34) est absolument instable pour tout $\tilde{h}$, on dit qu'elle n'a pas d'intervalle de stabilité absolue.

Remarque 4.2 La condition $|r_s| < 1$ reste suffisante pour que $\tilde{e}_n$ décroisse dans le cas de racines multiples.

Remarque 4.3 L'intervalle de stabilité est uniquement déterminé par les coefficients de la méthode. Toutefois la borne supérieure de h dépend de λ et donc de l'équation différentielle particulière. Si celle-ci n'est pas linéaire, $\partial_y f$ n'est pas constante; on choisit alors une borne pour λ ou une valeur typique pour $\partial_y f$ sur un sous-intervalle.

Remarque 4.4 Une méthode consistante et zéro stable (4.34) est absolument instable pour $\tilde{h}$ petit > 0.

Démonstration. Lorsque $\tilde{h} = 0$, les r_s sont les racines ζ_s de $\rho(\zeta)$ au bord ou à l'intérieur du disque unité. Mais $\zeta_1 = 1$. Soit r_1 la racine de (4.41) qui converge vers ζ_1 avec $\tilde{h} \to 0$. On montre que

$$r_1 = e^{\tilde{h}} + O\left(\tilde{h}^{p+1}\right) \qquad \text{lorsque } \tilde{h} \to 0. \tag{4.43}$$

Par la définition de l'ordre,

$$\mathcal{L}\left[e^{\lambda x}; h\right] = O\left(\tilde{h}^{p+1}\right).$$

Donc avec $y(x) = e^{\lambda x}$ on obtient:

$$\mathcal{L}\left[e^{\lambda x}; h\right] = \sum_{j=0}^{k} \left[\alpha_j \, e^{\lambda(x_n + jh)} - h\beta_j \lambda \, e^{\lambda(x_n+jh)}\right] = O\left(\tilde{h}^{p+1}\right),$$

c'est-à-dire

$$e^{\lambda x_n} \sum_{j=0}^{k} \left[\alpha_j \left(e^{\tilde{h}}\right)^j - \tilde{h}\beta_j \left(e^{\tilde{h}}\right)^j\right] = O\left(\tilde{h}^{p+1}\right).$$

On divise par $e^{\lambda x_n}$:

$$\pi\left(e^{\tilde{h}}; \tilde{h}\right) := \rho\left(e^{\tilde{h}}\right) - \tilde{h}\sigma\left(e^{\tilde{h}}\right) = O\left(\tilde{h}^{p+1}\right).$$

Puisque les racines de (4.41) sont $r_1, r_2, \ldots, r_k$, on peut écrire

$$\pi\left(r; \tilde{h}\right) := (\alpha_k - \tilde{h}\beta_k)(r - r_1)(r - r_2)\cdots(r - r_k);$$

on pose $r = e^{\tilde{h}}$:

$$\left(e^{\tilde{h}} - r_1\right)\left(e^{\tilde{h}} - r_2\right)\cdots\left(e^{\tilde{h}} - r_k\right) = O\left(\tilde{h}^{p+1}\right).$$

Lorsque $\tilde{h} \to 0$, $e^{\tilde{h}} \to 1$ et $r_s \to \zeta_s$. Alors le premier facteur du premier membre $\to 0$ quand $\tilde{h} \to 0$. Aucun autre facteur ne peut faire de même puisque par la 0 stabilité, $\zeta_1 = +1$ est une racine simple de $\rho(\zeta)$. Donc (4.43) suit. On a aussi démontré que

$$r_1 = e^{\tilde{h}} + O\left(\tilde{h}^{p+1}\right) \geq 1$$

lorsque $\tilde{h} \to 0+$. □

Remarque 4.5 Certaines méthodes n'ont pas d'intervalle de stabilité absolue; pour celles-ci il est inévitable que l'erreur croisse.

Les méthodes optimales (k pair d'ordre $k+2$), n'ont pas d'intervalle de stabilité absolue. Intuitivement si toutes les racines parasites de $\rho(\zeta)$ sont strictement à l'intérieur du cercle unité, la méthode admet un intervalle de stabilité absolue non nul. Or, les méthodes d'Adams ont toutes les racines parasites à l'origine; donc l'intervalle de stabilité absolue est assez grand.

Remarque 4.6 Si $\lambda \approx \partial_y f > 0$ pour un $\tilde{h}$ petit, l'erreur augmente au rythme des calculs. En effet, l'hypothèse $\partial_y f = \lambda$ n'est valide que pour les équations de la form $y' = \lambda y + g(x)$, λ constante, dont la solution contient un terme de la forme $A\,e^{\lambda x}$. Si la solution est dominée par des termes en r_1^n, puisque $r_1^n = \left(e^{\tilde{h}}\right)^n + O\left(\tilde{h}^{p+1}\right)$, l'erreur augmente au même taux que la solution, ce qui est acceptable.

Définition 4.2 La méthode (4.34) est *relativement stable*, si pour un $\tilde{h}$ donné, les racines r_s de (4.41) satisfont $|r_s| < |r_1|$, $s = 2, \ldots, k$; elle est relativement instable autrement. L'intervalle $(\alpha, \beta) \subset \mathbb{R}$ est un *intervalle de stabilité relative* si (4.34) est relativement stable pour tout $\tilde{h} \in (\alpha, \beta)$.

Remarque 4.7 Une méthode (4.34) dont l'intervalle de stabilité absolue est vide, peut avoir un intervalle de stabilité relative non nul.

Exemple 4.3 La méthode de Simpson, étant optimale, ne possède pas d'intervalle de stabilité absolue non nul. En effet, $\rho(r) = r^2 - 1$ et $\sigma(r) = \frac{1}{3}(r^2 + 4r + 1)$. Le polynôme de stabilité (4.41) devient

$$\pi\left(r; \tilde{h}\right) = \rho(r) - \tilde{h}\sigma(r) = \left(1 - \frac{1}{3}\tilde{h}\right) r^2 - \frac{4}{3}\tilde{h}r - \left(1 + \frac{1}{3}\tilde{h}\right) = 0. \qquad (4.44)$$

Les racines sont réelles pour tout $\tilde{h}$:

$$r = \frac{\frac{2}{3}\tilde{h} \pm \sqrt{\frac{4}{9}\tilde{h}^2 + 1 - \frac{1}{9}\tilde{h}^2}}{1 - \frac{1}{3}\tilde{h}}.$$

Si $k \geq 3$, en général il est impossible d'exprimer r_s explicitement comme fonction de $\tilde{h}$. Dans ce cas on peut approcher les racines r_s à $\tilde{h}$ près. De (4.43) il vient:

$$r_1 = 1 + \tilde{h} + O\left(\tilde{h}^2\right).$$

Considérons de nouveau la méthode de Simpson. Puisque $\zeta_2 = -1$ est l'unique racine parasite de $\rho(\zeta) = 0$, on peut écrire

$$r_2 = -1 + \gamma\tilde{h} + O\left(\tilde{h}^2\right).$$

Si l'on substitue cette valeur dans (4.44) on obtient

$$\gamma = \frac{1}{3}.$$

Voici le détail des calculs:

$$\left(1 - \frac{1}{3}\tilde{h}\right)\left(-1 + \gamma\tilde{h}\right)^2 - \frac{4}{3}\tilde{h}\left(-1 + \gamma\tilde{h}\right) - \left(1 + \frac{1}{3}\tilde{h}\right) + O\left(\tilde{h}^2\right) = 0,$$

$$\left(1 - \frac{1}{3}\tilde{h}\right)\left(1 - 2\gamma\tilde{h} + \cdots\right) - \frac{4}{3}\tilde{h}(-1 + \cdots) - \left(1 + \frac{1}{3}\tilde{h}\right) + \cdots = 0,$$

$$\left(1 - 2\gamma\tilde{h} - \frac{1}{3}\tilde{h} + \cdots\right) + \frac{4}{3}\tilde{h} - 1 - \frac{1}{3}\tilde{h} + \cdots = 0,$$

$$-2\gamma - \frac{1}{3} + \frac{4}{3} - \frac{1}{3} = 0,$$

$$-2\gamma + \frac{2}{3} = 0 \Rightarrow \gamma = \frac{1}{3}. \quad \square$$

On voit que l'intervalle de stabilité absolue est nul. L'intervalle de stabilité relative est de la forme $(0, \beta)$, $\beta > 0$. De (4.44) il vient que $\beta = +\infty$ pour la méthode de Simpson. Cette méthode est donc à déconseiller si $\partial_y f < 0$; mais si $\partial_y f > 0$ l'erreur n'augmentera pas relativement à la solution. On peut dire la même chose de n'importe quelle méthode optimale.

Exemple 4.4 Voici un exemple simple dû à Stetter et qui illustre l'utilité de la stabilité relative quand $\partial_y f < 0$. Soit la méthode consistante et 0 stable:

$$y_{n+2} - y_n = \frac{1}{2}h(f_{n+1} + 3f_n). \tag{4.45}$$

De (4.41) il vient

$$r^2 - \frac{1}{2}\tilde{h}r - \left(1 + \frac{3}{2}\tilde{h}\right) = 0.$$

En appliquant une analyse à $O\left(\tilde{h}^2\right)$, on obtient pour $\tilde{h}$ suffisamment petit:

$$r_1 = 1 + \tilde{h}, \quad r_2 = -1 - \frac{1}{2}\tilde{h}.$$

On voit que l'intervalle de la stabilité relative est de la forme $(0, \beta)$, $\beta > 0$ et l'intervalle de la stabilité absolue de la forme $(\alpha, 0)$, $\alpha < 0$. Si l'on adopte la définition de la stabilité absolue, on utilise la méthode seulement quand $\partial_y f < 0$; mais dans ce cas, le module de l'erreur, dominé par $|r_2|^n$, décroît moins lentement que la solution qui, elle, est dominée par $|r_1|^n$. Si d'autre part, on adopte la définition de la stabilité relative, on utilise la méthode seulement quand $\partial_y f > 0$; dans ce cas on accepte une erreur qui augmente en module mais plus lentement que ne le fait la solution. Cet exemple nous fait préférer la stabilité relative.

Exemple 4.5 Considérons une autre méthode consistante et 0 stable

$$y_{n+3} - y_{n+2} + y_{n+1} - y_n = \frac{h}{5}(f_{n+2} - 4f_{n+1} + 13f_n). \tag{4.46}$$

De (4.41) il vient

$$r^3 - \left(1 + \frac{1}{5}\tilde{h}\right) r^2 + \left(1 + \frac{4}{5}\tilde{h}\right) r - \left(1 + \frac{13}{5}\tilde{h}\right) = 0. \tag{4.47}$$

Lorsque $\tilde{h} = 0$,

$$r_1 = 1, \qquad r_2 = i, \qquad r_3 = -i.$$

Quand $\tilde{h} \neq 0$, $r_1 = 1 + \tilde{h} + O\left(\tilde{h}^2\right)$. Donc aux termes en $\tilde{h}$ près,

$$r_2 = \left(1 + \gamma\tilde{h}\right) e^{i[(\pi/2)+\theta\tilde{h}]}, \qquad r_3 = \left(1 + \gamma\tilde{h}\right) e^{-i[(\pi/2)+\theta\tilde{h}]},$$

puisque pour $\tilde{h}$ petit, r_2 et r_3 restent complexes et de ce fait conjugées. Alors aux termes en $\tilde{h}$ près, (4.47) devient

$$\left(r - 1 - \tilde{h}\right)\left[r - \left(1 + \gamma\tilde{h}\right) e^{i[(\pi/2)+\theta\tilde{h}]}\right]\left[r - \left(1 + \gamma\tilde{h}\right) e^{-i[(\pi/2)+\theta\tilde{h}]}\right] = 0.$$

Les termes indépendants de r de l'équation précédente et de (4.47) donnent

$$-\left(1 + \tilde{h}\right)\left(1 + \gamma\tilde{h}\right)^2 = -\left(1 + \frac{13}{5}\tilde{h}\right).$$

De nouveau négligeant les termes en $\tilde{h}^2$ et $\tilde{h}^3$, on obtient

$$\gamma = \frac{4}{5}.$$

Ainsi

$$|r_2| = |r_3| = 1 + \frac{4}{5}\tilde{h} + O\left(\tilde{h}^2\right)$$

et

$$|r_1| = 1 + \tilde{h} + O\left(\tilde{h}^2\right).$$

Donc l'intervalle de stabilité relative est de la forme $(0, \beta), \beta > 0$. Cependant l'intervalle exact de la stabilité relative est la réunion de trois intervalles:

$$\left(-\infty, -\frac{5+15\sqrt{3}}{13}\right) \bigcup \left(0, -\frac{5-15\sqrt{3}}{13}\right) \bigcup (10, +\infty).$$

Cet exemple nous rend excessivement pessimistes devant la stabilité relative. D'autre part, l'exemple de Stetter nous rend excessivement pessimistes devant la stabilité absolue.

Conclusions:

i) Pour un petit $\tilde{h} > 0$, toutes les méthodes consistantes et 0 stables sont instables dans le sens absolu.

ii) Une méthode relativement stable pour un petit $\tilde{h} < 0$ est absolument stable puisque $|r_1| < 1$.

iii) L'analyse $O\left(\tilde{h}\right)$ donne de l'information utile sur la nature de l'intervalle de stabilité mais non sur sa longueur.
Par exemple, considérons les méthodes à un pas respectivement d'Euler et des trapèzes. Pour Euler, $r_1 = 1 + \tilde{h}$ et l'intervalle de stabilité absolue est $(-2, 0)$. Pour les trapèzes, $r_1 = (1 + \frac{1}{2}\tilde{h})/(1 - \frac{1}{2}\tilde{h})$ et l'intervalle de stabilité absolue est $(-\infty, 0)$. Mais l'analyse $O\left(\tilde{h}\right)$ donne $r_1 = \left(1 + \tilde{h}\right)$ pour les deux méthodes.
La définition de la stabilité relative n'est pas applicable aux méthodes à un pas.

iv) La théorie de la stabilité faible dépend fortement de l'hypothèse (4.37): $\partial_y f = \lambda$ une constante.

Exemple 4.6 Considérons

$$y_{n+2} - (1+a)y_{n+1} + ay_n = \frac{h}{12}[(5+a)f_{n+2} + 8(1-a)f_{n+1} - (1+5a)f_n], \ -1 \leq a < 1.$$

i) Prouver que l'intervalle de stabilité absolue est

$$\left(6\frac{a+1}{a-1}, 0\right)$$

et l'intervalle de stabilité relative est

$$\left(\frac{3}{2}\frac{a+1}{a-1}, \infty\right).$$

ii) Illustrer le cas $a = -0.9$ en résolvant $y' = -20y$, $y(0) = 1$.

Solution. i) $-1 \leq a < 1 \Rightarrow 0$ stabilité, d'ordre 3 si $a \neq -1$, d'ordre 4 si $a = -1$ (méthode de Simpson). Soient le polynôme de stabilité en la variable r:

$$\pi\left(r, \tilde{h}\right) = \left[1 - \frac{\tilde{h}}{12}(5+a)\right] r^2 - \left[(1+a) + \frac{2}{3}\tilde{h}(1-a)\right] r + \left[a + \frac{\tilde{h}}{12}(1+5a)\right] = 0 \tag{4.48}$$

et son discriminant:

$$\Delta = (1-a)^2 + \tilde{h}\left(1-a^2\right) + \frac{1}{12}\tilde{h}^2\left(7 - 2a + 7a^2\right).$$

Considérant Δ comme un polynôme du second degré en $\tilde{h}$, on trouve que son discriminant est

$$-\frac{4}{3}(1-a)^4 < 0.$$

Donc $\Delta > 0$ pour tout $\tilde{h} \Rightarrow$ (4.48) admet des racines réelles distinctes r_1, r_2 tout $\tilde{h}$ et tout a.

Stabilité absolue. On sait que $\tilde{h} = 0 \Rightarrow r_1 = 1$. Alors $r = -1$ dans (4.48) $\Rightarrow \tilde{h} = 6(a+1)/(a-1) < 0$.

Remarque: L'intervalle $(6(a+1)/(a-1), 0) \to (0,0)$ lorsque $a \to -1+$ et $\to (-\infty, 0)$ lorsque $a \to +1-$.

Stabilité relative. Puisque r_1 et r_2 sont réelles, les bornes de l'intervalle de stabilité relative sont données par $|r_1| = |r_2|$, c'est-à-dire:

$$r_1 = r_2, \quad r_1 = -r_2.$$

Or $r_1 = r_2$ ne se produit jamais; donc l'intervalle se prolonge à $+\infty$.
Mais $r_1 = -r_2$ se présente lorsque le coefficient de r dans (4.48) s'annule c'est-à-dire lorsque

$$\tilde{h} = \frac{3}{2}\frac{a+1}{a-1}.$$

Ceci termine la démonstration de la partie (i).

ii) On pose $a = -0.9$ dans (i). Alors on a:
pour la stabilité absolue: $-0.316 < -20h < 0$, c'est-à-dire si $h < 0.016$.
pour la stabilité relative: $-0.079 < -20h < +\infty$, c'est-à-dire si $h < 0.00395$
Si l'on prend: $h = 0.01$, l'erreur décroît (stabilité absolue), si $h = 0.02$, l'erreur augmente, et si $h = 0.04$, l'erreur augmente.

Exemple 4.7 Démontrer l'instabilité absolue et relative de la méthode de Simpson lorsque $\partial f_y < 0$ en l'appliquant à l'équation

$$y' = -10(y-1)^2, \qquad y(0) = 2.$$

Solution On a $\partial_y f = -20(y-1) < 0$ près de $x = 0$. La solution théorique est $y = 1/(1+10x) \leq 0$ pour tout $x \geq 0$ (voir la table 7 du manuel).

Dans l'avant-dernier exemple, on avait une équation linéaire et on restait près des bornes de l'intervalle de stabilité absolue; les instabilités étaient par conséquent faibles.
Dans le dernier exemple, l'accroissement de l'erreur dû à l'instabilité faible est rapide. On voit aussi qu'on ne peut employer une méthode optimale qu'avec discernement.
Remarque. La méthode reste convergente quand

$$h \to 0, \qquad nh = x_n - a = \text{constante},$$

mais pour un h fixé, $x_n \to \infty$ quand $n \to \infty$.

4.7 Détermination des intervalles de stabilité

Méthode du tracé du module des racines pour une méthode à k pas. On résout (4.41):

$$\pi(r, \tilde{h}) = \rho(r) - \tilde{h}\sigma(r) = 0$$

pour plusieurs valeurs de $\tilde{h}$ près de 0, en utilisant, par exemple, Newton-Raphson et l'on trace la courbe

$$|r_s(\tilde{h})|, \qquad s = 1, 2, \ldots, k,$$

en fonction de $\tilde{h}$.

Exemple 4.8 Appliquer cette méthode à l'exemple de Stetter (4.45):

$$y_{n+2} - y_n = \frac{1}{2}h(f_{n+1} + 3f_n).$$

On obtient le polynôme de stabilité:

$$\pi\left(r_1\tilde{h}\right) = r^2 - \frac{1}{2}\tilde{h}r - \left(1 + \frac{3}{2}\tilde{h}\right) = 0,$$

qu'on résout pour les valeurs de $\tilde{h} = -2.0(0.2)(1.0)$. On trouve que l'intervalle de stabilité absolue est $(-1.33, 0)$ et l'intervalle de stabilité relative est $(0, \beta)$, $\beta > 1$, $(\beta = \infty)$.

Critère de Schur.

Définition 4.3 Le polynôme

$$\phi(r) = c_k r^k + c_{k-1} r^{k-1} + \ldots + c_1 r + c_0, \qquad c_j \in \mathbb{C}, c_k \neq 0, c_0 \neq 0, \tag{4.49}$$

est un *polynôme de Schur* si les racines sont toutes strictement à l'intérieur du disque unité:

$$|r_s| < 1, \qquad s = 1, \ldots, k.$$

Notons le conjugué complexe de $c = a + ib \in \mathbb{C}$: $c^* = a - ib$, et considérons le polynôme adjoint de ϕ:

$$\hat{\phi}(r) = c_0^* r^k + \ldots + c_k^*. \tag{4.50}$$

Maintenant formons le polynôme

$$\phi_1(r) = \frac{1}{r}\left[\hat{\phi}(0)\phi(r) - \phi(0)\hat{\phi}(r)\right]. \tag{4.51}$$

Il est évident que $\deg \phi_1 \leq k - 1$.

Théorème 4.1 (Schur)*: $\phi(r)$ est un polynôme de Schur $\Leftrightarrow$ $|\hat{\phi}(0)| > |\phi(0)|$ et $\phi_1(r)$ est un polynôme de Schur.*

Une première application, à la stabilité absolue. L'intervalle (α, β) est un intervalle de stabilité absolue si pour tout $\tilde{h} \in (\alpha, \beta)$, le polynôme $\pi\left(r, \tilde{h}\right)$ (4.41) est un polynôme de Schur réel. Par récurrence, on obtient:

$$\hat{\pi}\left(0, \tilde{h}\right) > \pi\left(0, \tilde{h}\right), \qquad \deg \pi_1 \leq k - 1,$$

$$\hat{\pi}_1\left(r, \tilde{h}\right) > \hat{\pi}_1\left(0, \tilde{h}\right), \qquad \deg \pi_2 \leq k - 2,$$

etc. jusqu'au degré 1.

Une seconde application, à la stabilité relative. On adopte la définition: $|r_s| < e^{\tilde{h}}$, $s = 1, \ldots, k$, et l'on substitue $r = Re^{\tilde{h}}$ dans (4.41); alors

$$\rho\left(R\, e^{\tilde{h}}\right) - \tilde{h}\sigma\left(R e^{\tilde{h}}\right) = 0.$$

Notons les racines de cette équation: R_s, $s = 1, 2, \ldots, k$; alors le critère de Schur donne des conditions nécessaires et suffisantes pour que $|R_s| < 1, s = 1, 2, \ldots, k$, c'est-à-dire

$$|r_s| < e^{\tilde{h}};$$

ici les inégalités transcendantes en $\tilde{h}$ peuvent être difficiles à manipuler.

Exemple 4.9 Appliquer le critère de Schur à l'exemple de Stetter (4.45). On a

$$\begin{aligned}
\pi\left(r,\tilde{h}\right) &= r^2 - \frac{1}{2}\tilde{h}r - \left(1+\frac{3}{2}\tilde{h}\right), \\
\hat{\pi}\left(r,\tilde{h}\right) &= -\left(1+\frac{3}{2}\tilde{h}\right)r^2 - \frac{1}{2}\tilde{h}r + 1, \\
\left|\hat{\pi}\left(0,\tilde{h}\right)\right| &> \left|\pi\left(0,\tilde{h}\right)\right| \qquad \text{si} \quad \left|\left(1+\frac{3}{2}\tilde{h}\right)\right| < 1,
\end{aligned}$$

c'est-à-dire si

$$\tilde{h} \in \left(-\frac{4}{3}, 0\right).$$

De (4.51) on obtient le polynôme du premier degré

$$\pi_1\left(r,\tilde{h}\right) = -\frac{1}{2}\tilde{h}\left(2+\frac{3}{2}\tilde{h}\right)(3r+1);$$

ce dernier s'annule à $r = -1/3$; on a donc un polynôme de Schur et l'intervalle de stabilité absolue est $(-4/3, 0)$.

Critère de Routh–Hurwitz. La transformation de Möbius

$$r = \frac{1+z}{1-z}$$

applique le demi-plan $\{\Re z \leq 0\}$ sur le disque disque unité $\{|r| \leq 1\}$ et $r = 1 \Rightarrow z = 0$. De (4.41) il vient

$$\rho\left(\frac{1+z}{1-z}\right) - \tilde{h}\sigma\left(\frac{1+z}{1-z}\right) = 0.$$

On chasse les dénominateurs en multipliant par $(1-z)^k$ et on obtient un polynôme de degré k:

$$a_0 z^k + a_1 z^{k-1} + \ldots + a_k = 0, \tag{4.52}$$

où sans perte de généralité $a_0 > 0$. Les racines de (4.52) satisfont $\{\Re z < 0\}$, c'est-à-dire les racines de $\pi\left(r,\tilde{h}\right) = 0$ satisfont $\{|r| < 1\} \Leftrightarrow$ les mineurs principaux de la matrice Q d'ordre k sont positifs:

$$Q = \begin{bmatrix}
a_1 & a_3 & a_5 & \cdots & a_{2k-1} \\
a_0 & a_2 & a_4 & \cdots & a_{2k-2} \\
0 & a_1 & a_3 & \cdots & a_{2k-3} \\
0 & a_0 & a_2 & \cdots & a_{2k-4} \\
\vdots & \vdots & \vdots & \cdots & \vdots \\
0 & 0 & 0 & \cdots & a_k
\end{bmatrix},$$

où $a_j = 0$ si $j > k$.
Il est possible de démontrer que cette condition implique que

$$a_j > 0, \qquad j = 0, 1, \ldots, k.$$

Donc la positivité des coefficients de (4.52) est une condition nécessaire mais non suffisante pour la stabilité absolue. Pour $k = 2, 3$ et 4 les conditions suivantes sont nécessaires et suffisantes pour la stabilité absolue.

$k = 2$: $\quad a_0 > 0, \quad a_1 > 0, \quad a_2 > 0, \quad a_0 z^2 + a_1 z + a_2,$

$$\begin{bmatrix} a_1 & 0 \\ a_0 & a_2 \end{bmatrix}.$$

$k = 3$: $\quad a_0 > 0, \quad a_1 > 0, \quad a_3 > 0, \quad a_1 a_2 - a_3 a_0 > 0,$

$$a_0 z^3 + a_1 z^2 + a_2 z + a_3,$$

$$\begin{bmatrix} a_1 & a_3 & 0 \\ a_0 & a_2 & 0 \\ 0 & a_1 & a_3 \end{bmatrix}.$$

$k = 4$: $\quad a_0 > 0, \quad a_1 > 0, \quad a_2 > 0, \quad a_3 > 0, \quad a_4 > 0,$

$$a_1 a_2 a_3 - a_0 a_3^2 - a_4 a_1^2 > 0,$$

$$\begin{bmatrix} a_1 & a_3 & 0 & 0 \\ a_0 & a_2 & a_4 & 0 \\ 0 & a_1 & a_3 & 0 \\ 0 & a_0 & a_2 & a_4 \end{bmatrix}.$$

Méthode de la position du bord. On considère r_s et $\tilde{h} \in \mathbb{C}$ et on définit une *région* R de *stabilité absolue* pour $\tilde{h} \in \mathbb{C}$ telle que les racines de $\pi\left(r, \tilde{h}\right)$ sont à l'intérieur du cercle unité: $|r_s| < 1$ si et seulement si $\tilde{h} \in R$. Le bord de cette région est noté ∂R. Puisque les racines sont des fonctions continues de $\tilde{h}$, une racine r_s va satisfaire $|r_s| = 1$ lorsque $\tilde{h} \in \partial R$, c'est-à-dire

$$\pi\left(e^{i\theta}, \tilde{h}\right) := \rho\left(e^{i\theta}\right) - \tilde{h}\sigma\left(e^{i\theta}\right) = 0;$$

alors

$$\tilde{h}(\theta) = \frac{\rho\left(e^{i\theta}\right)}{\sigma\left(e^{i\theta}\right)}.$$

Pour $\tilde{h}$ réel, les bornes de l'intervalle de stabilité absolue vont être des points où ∂R coupe l'axe des réels. On doit vérifier une valeur de $\tilde{h}$ pour s'assurer que les racines sont bien à l'intérieur et non à l'extérieur du disque unité.

Exemple 4.10 Utiliser la méthode de la position du bord pour établir l'intervalle de stabilité absolue pour la méthode (4.45)

$$y_{n+2} - y_n = \frac{1}{2}h(f_{n+1} + 3f_n).$$

Ici

$$\rho(r) = r^2 - 1, \quad \sigma(r) = \frac{1}{2}(r+3)$$

et ∂R est donné par

$$\begin{aligned}
\tilde{h}(\theta) &= \frac{\rho\left(e^{i\theta}\right)}{\sigma\left(e^{i\theta}\right)} \\
&= 2\frac{e^{2i\theta}-1}{e^{i\theta}+3} \\
&= \frac{3(\cos 2\theta - 1) + i(3\sin 2\theta + 2\sin\theta)}{5 + 3\cos\theta} \\
&= \frac{3(\cos 2\theta - 1) + i2\sin\theta(3\cos\theta + 1)}{5 + 3\cos\theta}.
\end{aligned}$$

Le bord coupe l'axe des réels si $\sin\theta = 0$ ou $3\cos\theta = -1$ c'est-à-dire quand $\theta = 0, \pi, \pi \pm \arccos(1/3)$. On a employé l'identité $\cos(\theta - \pi) = -\cos\theta$.
Quand $\theta = 0$ et π, $\tilde{h}(\theta) = 0$.
Quand $\theta = \pi \pm \arccos(1/3), \tilde{h}(\theta) = -\frac{4}{3}$.
Donc les bornes de l'intervalle de stabilité absolue sont $-\frac{4}{3}$ et 0.
On vérifie que les racines sont à l'intérieur du cercle unité: quand $\tilde{h} = -\frac{2}{3}$ l'équation

$$\rho(r) - \tilde{h}\sigma(r) = 0$$

devient

$$r^2 + \frac{1}{3}r = 0.$$

On voit donc que les racines $r_1 = 0$ et $r_2 = -1/3$ de cette dernière sont à l'intérieur du cercle unité. Donc l'intervalle de stabilité absolue est $\left(-\frac{4}{3}, 0\right)$.

4.8 Comparaison entre les méthodes explicites et implicites

La supériorité des méthodes implicites sur les méthodes explicites provient du plus grand intervalle de stabilité absolue des premières, comme on peut le voir aux tableaux suivantes pour les méthodes d'Adams–Bashforth et d'Adams–Moulton.

TABLEAU. Méthodes d'Adams–Bashforth — explicites.

Nombre de pas	k	1	2	3	4
Ordre	p	1	2	3	4
Const. de l'erreur	C_{p+1}	$\frac{1}{2}$	$\frac{5}{12}$	$\frac{3}{8}$	$\frac{251}{720}$
Borne gauche de l'i.s.a.	α	-2	-1	$-\frac{6}{11}$	$-\frac{3}{10}$

TABLEAU. Méthodes d'Adams–Moulton — implicites.

Nombre de pas	k	1	2	3	4
Ordre	p	2	3	4	5
Const. de l'erreur	C_{p+1}	$-\frac{1}{12}$	$-\frac{1}{24}$	$-\frac{19}{720}$	$-\frac{3}{160}$
Borne gauche de l'i.s.a.	α	$-\infty$	-6	-3	$-\frac{90}{40}$

4.9 Méthodes prédicteurs-correcteurs

On aborde maintenant la question (iii) soulevée à la section 4.1: Quelle est la meilleure façon de résoudre l'équation implicite

$$y_{n+k} + \sum_{j=0}^{k-1} \alpha_j y_{n+j} = h\beta_k f(x_{n+k}, y_{n+k}) + h\sum_{j=0}^{k-1} \beta_j f_{n+j}. \tag{4.53}$$

pour y_{n+k}? La récurrence

$$y_{n=k}^{[s+1]} + \sum_{j=0}^{k-1} \alpha_j y_{n+j} = h\beta_k f\left(x_{n+k}, y_{n+k}^{[s]}\right) + h\sum_{j=0}^{k-1} \beta_j f_{n+j}, \quad s = 0, 1, \ldots, \tag{4.54}$$

où $y_{n+k}^{[0]}$ est arbitraire, converge si

$$h < \frac{1}{L|\beta_k|}. \tag{4.55}$$

On emploie une méthode explicite, appelée *prédicteur* pour obtenir la valeur approchée $y_{n+k}^{[0]}$ qu'on corrige au moyen du *correcteur* (4.53).

On dit qu'on corrige à la limite si l'on itère (4.54) jusqu'à ce que $\left|y_{n+k}^{[s+1]} - y_{n+k}^{[s]}\right| < \varepsilon$, où ε est fixé. Dans ce cas, l'erreur locale et la stabilité faible sont celles du correcteur.

Si l'on applique (4.54) m fois, m fixé, l'erreur locale et la stabilité faible peuvent dépendre également du prédicteur.

Notation 4.1 On emploie la notation précisée par Hull et Creemer:

P une application du prédicteur

C une application du correcteur

E une évaluation de f à des valeurs connues de ses arguments.

Alors on écrit PEC si l'on a la suite des opérations

$P \to y_{n+k}^{[0]}$

$E \to f_{n+k}^{[0]} := f\left(x_{n+k}, y_{n+k}^{[0]}\right)$

$C \to y_{n+k}^{[1]}$

On écrit $PECEC$ ou $P(EC)^2$. On écrit $P(EC)^m$ si l'on accepte

$$f_{n+k}^{[m-1]} = f\left(x_{n+k}, y_{n+k}^{[m-1]}\right)$$

pour f_{n+k}; de même on écrit $P(EC)^m E$ si l'on accepte

$$f_{n+k}^{[m]} = f\left(x_{n+k}, y_{n+k}^{[m]}\right)$$

pour f_{n+k}.
On verra que les caractéristiques de la stabilité faible de $P(EC)^m$ et de $P(EC)^m E$ sont très différentes.

En pratique on choisit P et C du même ordre. Pour simplifier l'écriture des formules, il sera plus simple de supposer artificiellement que $k^* = k$, quitte à accepter que, pour le correcteur, α_0 et β_0 puissent tous les deux s'annuler. Soient les polynômes caractéristiques, respectivement, du prédicteur:

$$\rho^*(\zeta) = \sum_{j=0}^{k} \alpha_j^* \zeta^j, \qquad \alpha_k^* = 1, \qquad \sigma^*(\zeta) = \sum_{j=0}^{k} \beta_j^* \zeta^j, \tag{4.56}$$

et du correcteur:

$$\rho(\zeta) = \sum_{j=0}^{k} \alpha_j \zeta^j, \qquad \alpha_j = 1, \qquad \sigma(\zeta) = \sum_{j=0}^{k} \beta_j \zeta^j. \tag{4.57}$$

On peut alors définir formellement $P(EC)^m E$:

$$\begin{aligned}
y_{n+k}^{[0]} + \sum_{j=0}^{k-1} \alpha_j^* y_{n+j}^{[m]} &= h \sum_{j=0}^{k-1} \beta_j^* f_{n+j}^{[m]}, \\
f_{n+k}^{[s]} &= f\left(x_{n+k}, y_{n+k}^{[s]}\right), \\
y_{n+k}^{[s+1]} + \sum_{j=0}^{k-1} \alpha_j y_{n+j}^{[m]} &= h\beta_k f_{n+k}^{[s]} + h \sum_{j=0}^{k-1} \beta_j f_{n+j}^{[m]}, \qquad s = 0, 1, \ldots, m-1, \\
f_{n+k}^{[m]} &= f\left(x_{n+k}, y_{n+k}^{[m]}\right).
\end{aligned} \tag{4.58}$$

et $P(EC)^m$:

$$\begin{aligned} y_{n+k}^{[0]} + \sum_{j=0}^{k-1} \alpha_j^* y_{n+j}^{[m]} &= h \sum_{j=0}^{k-1} \beta_j^* f_{n+j}^{[m-1]}, \\ f_{n+k}^{[s]} &= f\left(x_{n+k}, y_{n+k}^{[s]}\right), \\ y_{n+k}^{[s+1]} + \sum_{j=0}^{k-1} \alpha_j y_{n+j}^{[m]} &= h\beta_k f_{n+k}^{[s]} + h \sum_{j=0}^{k-1} \beta_J f_{n+j}^{[m-1]}, \qquad s = 0, 1, \ldots, m-1. \end{aligned} \tag{4.59}$$

4.10 Erreur locale

Sous l'hypothèse de localisation, on suppose que y_{m+j}, $j = 0, \ldots, k-1$, sont exactes. Si

$$y(x) \in C^{p^*+2} \bigcup C^{p+2},$$

les erreurs du prédicteur et du correcteurs sont respectivement de la forme

$$\begin{aligned} \mathcal{L}^*[y(x); h] &= C_{p^*+1}^* h^{p^*+1} y^{(p^*+1)}(x) + O\left(h^{p^*+2}\right) \\ \mathcal{L}[y(x); h] &= C_{p+1} h^{p+1} y^{(p+1)}(x) + O\left(h^{p+2}\right). \end{aligned} \tag{4.60}$$

Si on applique (3.25) au prédicteur P on obtient

$$y(x_{n+k}) - y_{n+k}^{[0]} = C_{p^*+1}^* h^{p^*+1} y^{(p^*+1)}(x_n) + O\left(h^{p^*+2}\right) \tag{4.61}$$

On redérive (3.25) pour le correcteur C.

$$\begin{aligned} \sum_{j=0}^{k} \alpha_j y(x_{n+j}) &= h \sum_{j=0}^{k} \beta_j f(x_{n+j}, y(x_{n+j})) + \mathcal{L}[y(x_n); h], \\ y_{n+k}^{[s+1]} + \sum_{j=0}^{k-1} \alpha_j y_{n+j}^{[m]} &= h\beta_k f\left(x_{n+k}, y_{n+k}^{[s]}\right) \\ &\quad + h \sum_{j=0}^{k-1} \beta_j f\left(x_{n+j}, y_{n+j}^{[m-t]}\right), \qquad s = 0, \ldots, m-1, \end{aligned}$$

où $t = 0$ dans le mode $P(EC)^m E$ et $t = 1$ dans le mode $P(EC)^m$.
On soustrait et on utilise l'hypothèse de localisation:

$$y(x_{n+k}) - y_{n+k}^{[s+1]} = h\beta_k \left[f(x_{n+k}, y(x_{n+k})) - f\left(x_{n+k}, y_{n+k}^{[s]}\right)\right] + \mathcal{L}[y(x_n); h]$$

$$= h\beta_k \frac{\partial f}{\partial y}(x_{n+k}, \eta_{n+k,s}) \left[y(x_{n+k}) - y_{n+k}^{[s]} \right] + \mathcal{L}[y(x_n); h], \quad s = 0, 1, \ldots, m-1, \tag{4.62}$$

Le cas: $p^* \geq p$ ($h^* \approx h$). Si l'on substitue (4.61) dans le second membre de (4.62) avec $s = 0$, alors de (4.60) il vient

$$y(x_{n+k}) - y_{n+k}^{[1]} = C_{p+1} h^{p+1} y^{(p+1)}(x_n) + O\left(h^{p+2}\right).$$

Par substitutions successives dans (4.62), avec $s = 2, 3, \ldots, m-1$, on obtient

$$y(x_{n+k}) - y_{n+k}^{[m]} = C_{p+1} h^{p+1} y^{(p+1)}(x_n) + O\left(h^{p+2}\right), \qquad m = 1, 2, \ldots.$$

Conclusion. L'erreur locale principale du PC est uniquement celle de C pour les modes $P(EC)^m E$ et $P(EC)^m$ si $p^* \geq p$.

Le cas: $p^* = p - 1$. Si l'on substitue (4.61) dans (4.62) avec $s = 0$, on obtient

$$y(x_{n+k}) - y_{n+k}^{[1]} = \left[\beta_k \frac{\partial f}{\partial y} C_p^* y^{(p)}(x_n) + C_{p+1} y^{(p+1)}(x_n) \right] h^{p+1} + O\left(h^{p+2}\right).$$

Pour $m = 1$ on voit que l'erreur locale principale est du même ordre que celle du correcteur, mais diffère de celle-ci; mais par substitutions successives dans (4.62) avec $m \geq 2$, on voit qu'elles sont identiques.

Le cas: $p^* = p - 2$. (4.61) dans (4.62) donne:

$$y(x_{n+k}) - y_{n+k}^{[1]} = \beta_k \frac{\partial f}{\partial y} C_{p-1}^* y^{(p-1)}(x_n) + O\left(h^{p+1}\right).$$

Alors pour $m = 1$, l'ordre de l'erreur locale principale est un de moins que celui du correcteur.
Une seconde substitution dans (4.62) donne:

$$y(x_{n+k}) - y_{n+k}^{[2]} = \left[(\beta_k \partial f_y)^2 C_{p-1}^* y^{(p-1)}(x_n) + C_{p+1} y^{(p+1)}(x_n) \right] h^{p+1} + O\left(h^{p+2}\right)$$

qui est du même ordre que l'erreur de C. Pour $m \geq 3$ les erreurs principales sont identiques.

Le cas général: $p^* \geq 0, p \geq 1, m \geq 1$, voir le manuel, pp. 90-91.

L'estimation de Milne. Soit $p^* = p$. Alors on peut estimer l'erreur locale principale de la méthode PC sans estimer de dérivées supérieures de $y(x)$. Du fait que

$$C_{p+1} h^{p+1} y^{(p+1)}(x_n) = y(x_{n+k}) - y_{n+k}^{[m]} + O\left(h^{p+2}\right),$$

et

$$C^*_{p+1}h^{p+1}y^{(p+1)}(x_n) = y(x_{n+k}) - y^{[0]}_{n+k} + O\left(h^{p+2}\right),$$

après soustraction, on obtient l'estimation

$$C_{p+1}h^{p+1}y^{(p+1)}(x_n) = \frac{C_{p+1}}{C^*_{p+1} - C_{p+1}}\left(y^{[m]}_{n+k} - y^{[0]}_{n+k}\right). \tag{4.63}$$

Remarque: Comme (4.63) dépend de l'hypothèse de localisation, cette estimation donne seulement une approximation de l'erreur locale principale. Aujourd'hui on emploie (4.63) surtout pour estimer le pas h approprié. Hamming a employé (4.63) pour estimer l'erreur locale principale soit de P, soit de C.
Pour P:

$$C^*_{p+1}h^{p+1}y^{(p+1)}(x_n) = \frac{C^*_{p+1}}{C^*_{p+1} - C_{p+1}}\left(y^{[m]}_{n+k} - y^{[0]}_{n+k}\right).$$

On ne peut utiliser cette estimation pour améliorer la valeur prédite puisque $y^{[m]}_{n+k}$ n'est pas encore connue. Toutefois, dans l'expression précédente,

$$\begin{aligned} \text{le } 1^{er}\text{M} &= C^*_{p+1}h^{p+1}y^{(p+1)}(x_{n+k-1}) + O\left(h^{p+2}\right) \\ &= \frac{C^*_{p+1}}{C^*_{p+1} - C_{p+1}}\left(y^{[m]}_{n+k-1} - y^{[0]}_{n+k-1}\right) + O\left(h^{p+2}\right). \end{aligned}$$

On remplace $y^{[0]}_{n+k}$ par la *valeur modifiée*:

$$\hat{y}^{[0]}_{n+k} = y^{[0]}_{n+k} + \frac{C^*_{p+1}}{C^*_{p+1} - C_{p+1}}\left(y^{[m]}_{n+k-1} - y^{[0]}_{n+k-1}\right) \tag{4.64}$$

Cette étape s'appelle *modificateur* et se note M.
Pour C: (4.63) donne:

$$\hat{y}^{[m]}_{n+k} = y^{[m]}_{n+k} + \frac{C_{p+1}}{C^*_{p+1} - C_{p+1}}\left(y^{[m]}_{n+k} - y^{[0]}_{n+k}\right). \tag{4.65}$$

On a donc les modes $PM(EC)^mME$ et $PM(EC)^mM$.

Remarque 4.8 L'utilisation de M après la correction finale enlève la possibilité d'utiliser (4.63) pour contrôler les pas. Il est probablement préférable d'utiliser des méthodes d'ordre plus élevé et ensuite utiliser M. Toutefois, les algorithmes modernes contrôlent le pas avec (4.63) et modifient $y^{[s]}_{n+k}$ par (4.65) soit avec $s = 1, \ldots, m$, soit avec $s = m$ seulement.

Exemple 4.11 Soient les méthodes d'ordre 4:

$$\begin{array}{lll} P: & \rho^*(\zeta) = \zeta^4 - 1; & \sigma^*(\zeta) = \frac{4}{3}(2\zeta^3 - \zeta^2 + 2\zeta), \\ C^{(1)}: & \rho_1(\zeta) = \zeta^2 - 1; & \sigma_1(\zeta) = \frac{1}{3}(\zeta^2 + 4\zeta + 1), \\ C^{(2)}: & \rho_2(\zeta) = \zeta^3 - \frac{9}{8}\zeta^2 + \frac{1}{8} & \sigma_2(\zeta) = \frac{3}{8}(\zeta^3 + 2\zeta^2 - \zeta). \end{array} \tag{4.66}$$

La méthode de Milne est donnée par $PEC^{(1)}E$ et celle de Hamming par $PMEC^{(2)}ME$.

Exemple 4.12 Comparer P et $C^{(2)}$ dans

i) le mode de correction jusqu'à convergence,
ii) le mode $PECE$,
iii) le mode $PMECE$,

pour résoudre le problème à valeur initiale

$$y' = -10(y-1)^2, \qquad y(0) = 2, \qquad 0 \le x \le 0.2, \qquad h = 0.01.$$

La solution $y = 1 + 1/(1+10x)$ dans (i) donne $\left| y_{n+k}^{[s+1]} - y_{n+k}^{[s]} \right| < 10^{-9}$.

4.11 Stabilité faible

La stabilité faible des méthodes prédicteurs-correcteurs dépend à la fois de P et de C, si l'on ne corrige pas jusqu'à la limite. Comme à la section 4.6, notons les polynômes de stabilité absolue du prédicteur et du correcteur:

$$\pi_P(r, \tilde{h}) = \rho^*(r) - \tilde{h}\sigma^*(r), \quad \pi_C(r, \tilde{h}) = \rho(r) - \tilde{h}\sigma(r).$$

On montre que le polynôme de stabilité absolue du PC en mode $PECE$ est

$$\pi_{PECE}(r, \tilde{h}) = \rho(r) - \tilde{h}\sigma(r) + \tilde{h}\beta_k[\rho^*(r) - \tilde{h}\sigma^*(r)].$$

Pour ce faire, soient les valeurs approchées prédites et corrigées

$$\tilde{y}_{n+k}^{[0]}, \tilde{y}_{n+k}^{[1]} \approx y(x_{n+k})$$

donnés par le prédicteur et le correcteur:

$$\tilde{y}_{n+k}^{[0]} + \sum_{j=0}^{k-1} \alpha_j^* \tilde{y}_{n+j}^{[1]} = h \sum_{j=0}^{k-1} \beta_j^* f\left(x_{n+j}, \tilde{y}_{n+j}^{[1]}\right) + R_n^*,$$

$$\sum_{j=0}^{k} \alpha_j \tilde{y}_{n+j}^{[1]} = h\beta_k f(x_{n+k}, \tilde{y}_{n+k}^{[0]}) + h \sum_{j=0}^{k-1} \beta_j f(x_{n+j}, \tilde{y}_{n+j}^{[1]}) + R_n,$$

où R_n^* et R_n sont les erreurs locales d'arrondi. La solution théorique $y(x)$ du problème à valeur initiale satisfait

$$\sum_{j=0}^{k} \alpha_j^* y(x_{n+j}) = h \sum_{j=0}^{k-1} \beta_j^* f(x_{n+j}, y(x_{n+j})) + T_n^*$$

et

$$\sum_{j=0}^{k} \alpha_j y(x_{n+j}) = h \sum_{j=0}^{k} \beta_j f(x_{n+j}, y(x_{n+j})) + T_n,$$

où T_n^* et T_n sont les erreurs locales de P et C respectivement. On définit les erreurs globales:

$$\tilde{e}_n^{[0]} = y(x_n) - \tilde{y}_n^{[0]}, \quad \tilde{e}_n^{[1]} = y(x_n) - \tilde{y}_n^{[1]}.$$

Si les expressions suivantes sont des constantes:

$$\partial_y f = \lambda, \qquad \tilde{h} = h\lambda, \qquad T_N^* - R_n^*, \qquad T_n - R_n,$$

on obtient les équations des erreurs linéarisées:

$$\tilde{e}_{n+k}^{[0]} + \sum_{j=0}^{k-1} \alpha_j^* \tilde{e}_{n+j}^{[1]} = \tilde{h} \sum_{j=0}^{k-1} \beta_j^* \tilde{e}_{n+j}^{[1]} + \text{constante},$$

$$\sum_{j=0}^{k} \alpha_j \tilde{e}_{n+j}^{[1]} = \tilde{h}\beta_k \tilde{e}_{n+k}^{[0]} + \tilde{h} \sum_{j=0}^{k-1} \beta_j \tilde{e}_{n+j}^{[1]} + \text{constante}.$$

Si l'on substitue la valeur de $\tilde{e}_{n+k}^{[0]}$ de la première expression dans la seconde, on obtient

$$\sum_{j=0}^{k} \alpha_j \tilde{e}_{n+j}^{[1]} - \tilde{h} \sum_{j=0}^{k-1} \beta_j \tilde{e}_{n+1}^{[1]} = -\tilde{h}\beta_k \left[\sum_{j=0}^{k-1} \alpha_j^* \tilde{e}_{n+j}^{[1]} - \tilde{h} \sum_{j=0}^{k-1} \beta_j^* \tilde{e}_{n+j}^{[1]} \right] + \text{constante}.$$

On additionne $-\tilde{h}\beta_k \tilde{e}_{n+k}^{[1]}$ aux deux membres et on emploie: $\alpha_k^* = 1$, $\beta_k^* = 0$, pour obtenir

$$\sum_{j=0}^{k} \left(\alpha_j - \tilde{h}\beta_j \right) \tilde{e}_{n+j}^{[1]} = -\tilde{h}\beta_k \sum_{j=0}^{k} \left(\alpha_j^* - \tilde{h}\beta_j^* \right) \tilde{e}_{n+j}^{[1]} + \text{constante}.$$

La solution $\tilde{e}_n^{[1]}$, dans le mode $PECE$, est de la forme

$$\tilde{e}_n^{[1]} = \sum_{s=1}^{k} d_s r_s^n + \text{constante},$$

où les d_s sont des constantes arbitraires et les r_s sont les racines (supposées distinctes) du polynôme de stabilité

$$\pi_{PECE}\left(r,\tilde{h}\right) := \rho(r) - \tilde{h}\sigma(r) + \tilde{h}\beta_k\left[\rho^*(r) - \tilde{h}\sigma^*(r)\right] = 0.$$

De la même façon, on a les polynômes de stabilité suivants dans les modes supérieurs indiqués:

$$\pi_{P(EC)^m E}\left(r,\tilde{h}\right) = \rho(r) - \tilde{h}\sigma(r) + M_m\left(\tilde{h}\right)\left[\rho^*(r) - \tilde{h}\sigma^*(r)\right], \tag{4.67}$$

où

$$M_m\left(\tilde{h}\right) = \left(\tilde{h}\beta_k\right)^m \frac{1-\tilde{h}\beta_k}{\left(1-\tilde{h}\beta_k\right)^m} \qquad m = 1,2,\ldots; \tag{4.68}$$

$$\pi_{P(EC)^m}\left(r,\tilde{h}\right) = \beta_k r^k\left[\rho(r) - \tilde{h}\sigma(r)\right] + M_m\left(\tilde{h}\right)\left[\rho^*(r)\sigma(r) - \rho(r)\sigma^*(r)\right]. \tag{4.69}$$

Si $L = |\partial_y f|$, la condition (4.55) devient

$$|\tilde{h}\beta_k| < 1. \tag{4.70}$$

Alors $M_m(\tilde{h}) \to 0$ quand $m \to \infty$. Donc $\pi_{P(EC)^m} \to \pi$ et $\pi_{P(EC)^m} \to \pi$, c'est-à-dire l'erreur dépend uniquement de C si l'on corrige jusqu'à la convergence. Les polynômes de stabilité (4.67) et (4.69) ont toujours une racine de la forme

$$r_1 = e^h + O\left(\tilde{h}^{p+1}\right);$$

alors les méthodes correspondantes sont absolument instables pour $\tilde{h}$ petit positif. On peut appliquer les méthodes de la position des racines, le critère de Schur et le critère de Routh–Hurwitz aux nouveaux polynômes de stabilité sans modification. On doit cependant modifier la méthode de la position du bord puisque le polynôme de stabilité π n'est plus linéaire en $\tilde{h}$.

Exemple 4.13 Chase a montré par la méthode de la position des racines, que l'intervalle de stabilité absolue de la paire prédicteur-correcteur de Milne dans le mode $PECE$ est $(-0.8, -0.3)$. Vérifier ce résultat pour $y' = y$, $y(0) = 1$, $0 \le x \le 100$. Ici $\partial_y f = -1$, $\tilde{h} = -h$.

Remarque: Si $h = 0.8$ ou 0.3, l'erreur est persistante. La méthode est absolument stable si $0.8 < h < 0.3$.

Exemple 4.14 Répéter l'exemple précédent avec

$$y' = -5xy^2 + \frac{5}{x} - \frac{1}{x^2}, \qquad y(1) = 1.$$

Solution. La solution analytique est $y(x) = 1/x$. On voit que

$$\partial_y f = -10xy = -10 \quad \Rightarrow \quad \tilde{h} = -10h.$$

Ici encore l'erreur est persistante lorsque $h = 0.8$ ou 0.3.

4.12 Contrôle du pas

On s'attaque maintenant à la plus difficile des quatre questions de la section 4.1.

Les bornes de l'erreur globale de la section 4.5 n'assurent pas une base adéquate pour choisir h. La stabilité faible, à la section 4.6, donne un intervalle pour h tel que l'erreur globale ne "croît" pas. L'erreur locale des méthodes PC et le critère de Milne à la section 4.10 peuvent servir au contrôle des pas.

On a les trois contrôles suivants:

(i) contrôle de l'erreur locale principale (4.63): $|\,\text{e.l.p.}\,| < \epsilon$ à chaque pas,

(ii) contrôle de la stabilité: $\tilde{h} \in$ intervalle de stabilité absolue ou relative,

(iii) contrôle de la convergence (4.55): $\tilde{h} < 1/(L|\beta_k|)$ est satisfait.

Dans le cas d'une seule équation, on peut employer le truc de Nordsieck pour estimer $\partial_y f$ dans $\tilde{h} = h\partial_y f$. Après deux applications successives de C en mode $P(EC)^m E$ ou bien $P(EC)^m$, $m \geq 2$ on a:

$$t = 0, \qquad P(EC)^m E: \quad y_{n+k}^{[m-1]} + \sum_{j=0}^{k-1} \alpha_j y_{n+j}^{[m]} = h\beta_k f_{n+k}^{[m-2]} + h\sum_{j=0}^{k-1} \beta_j f_{n+j}^{[m-t]},$$

$$t = 1, \qquad P(EC)^m: \quad y_{n+k}^{[m]} + \sum_{j=0}^{k-1} \alpha_j y_{n+j}^{[m]} = h\beta_k f_{n+k}^{[m-1]} + h\sum_{j=0}^{k-1} \beta_j f_{n+j}^{[m-t]}.$$

Si l'on soustrait, on obtient

$$y_{n+k}^{[m]} - y_{n+k}^{[m-1]} = h\beta_k \frac{\partial f}{\partial y}\left(x_{n+k}, \eta_{n+k}^{[m-1]}\right)\left(y_{n+k}^{[m-1]} - y_{n+k}^{[m-2]}\right).$$

Alors

$$\begin{aligned} \tilde{h} &= h\frac{\partial f}{\partial y} \approx \frac{y_{n+k}^{[m]} - y_{n+k}^{[m-1]}}{\beta_k\left(y_{n+k}^{[m-1]} - y_{n+k}^{[m-2]}\right)}, \qquad m \geq 2, \\ \tilde{h} &= h\frac{\partial f}{\partial y} \approx h\frac{f\left(x_{n+k}, y_{n+k}^{[1]}\right) - f\left(x_{n+k}, y_{n+k}^{[0]}\right)}{y_{n+k}^{[1]} - y_{n+k}^{[0]}}, \qquad m = 1. \end{aligned} \tag{4.71}$$

Remarque 4.9 Les expressions (4.71) ne s'appliquent pas à un système d'équations.

4.13 Choix des méthodes

La plupart des algorithmes font appel au plus à 2 évaluations par pas et par conséquent on limite cette discussion aux modes $PEC, PECE$ et $P(EC)^2$. De plus on suppose que $p^* = p$ afin d'utiliser le truc de Milne pour contrôler le pas. Ces modes ont la même erreur locale principale. On donne les intervalles de stabilité absolue pour les paires d'Adams–Bashforth–Moulton d'ordre 4:

$$\begin{array}{ll} P(EC)^\infty & (-3.00, 0) \\ PEC & (-0.16, 0) \\ PECE & (-1.25, 0) \\ P(EC)^2 & (-0.30, 0) \\ PM(ECM)^2 & (-0.66, 0) \\ P(ECM)^2 & (-0.95, 0) \end{array} \tag{4.72}$$

On constate la supériorité du mode $PECE$. Dans le mode $P(EC)^2$ on peut estimer $\tilde{h}$ au moyen de la première estimation de (4.71). L'utilisation du modificateur dans les deux derniers modes donne en fait des résultats d'ordre 5. On considère maintenant le choix des paires PC. On remarque d'abord que la méthode de Milne est en générale inadéquate à cause de sa pauvre stabilité faible. La paire d'Adams–Bashforth–Moulton d'ordre 4:

$$\begin{aligned} P: \quad y_{n+4} - y_{n+3} &= \frac{h}{24}(55f_{n+3} - 59f_{n+2} + 37f_{n+1} - 9f_n), \\ C: \quad y_{n+4} - y_{n+3} &= \frac{h}{24}(9f_{n+4} + 19f_{n+3} - 5f_{n+2} + f_{n+1}), \end{aligned} \tag{4.73}$$

avec l'estimation de l'erreur locale:

$$C_5 h^5 y^{(5)}(x_n) \approx -\frac{19}{270}\left(y_{n+4}^{[1]} - y_{n+4}^{[0]}\right),$$

est très populaire. On obtient la constante de l'erreur principale, $-19/270$ de la façon suivante:

$$C_{p+1}^* = \frac{251}{720}, \qquad C_{p+1} = -\frac{19}{720}\frac{C_{p+1}}{C_{p+1}^* - C_{p+1}} = -\frac{19}{251+19} = -\frac{19}{270}.$$

L'intervalle de stabilité absolue en mode $PECE$ est $(-1.25, 0)$. Crane-Klopfenstein ont proposé le prédicteur suivant

$$\begin{aligned} &y_{n+4} - 1.547652y_{n+3} + 1.867503y_{n+2} - 2.017204y_{n+1} + 0.697353y_n \\ &\quad = h(2.002247f_{n+3} - 2.031690f_{n+2} + 1.818609f_{n+1} - 0.714320f_n). \end{aligned} \tag{4.74}$$

avec le correcteur de (4.73). On a alors l'estimation de l'erreur:

$$c_5 h^5 y^{(5)}(x_n) = -\frac{1}{16.219\,66}\left(y_{n+4}^{[1]]} - y_{n+4}^{[0]}\right)$$

et l'intervalle de stabilité absolue: $(-2.48, 0)$.
A la suite de beaucoup d'expériences numériques pour une efficacité plus grande, on devrait utiliser une méthode d'ordre plus élevé en mode $PECE$. Cependant en mode PEC, on conseille le prédicteur de Klopfenstein-Millman

$$\begin{aligned} &y_{n+4} + 0.29y_{n+3} + 15.39y_{n+2} - 12.13y_{n+1} - 4.55y_n \\ &\quad = h(2.24f_{n+3} + 6.65f_{n+2} + 13.91f_{n+1} + 0.69f_n) \end{aligned} \tag{4.75}$$

avec le correcteur de (4.73). Alors l'estimation d'erreur donne

$$C_5 h^5 y^{(5)}(x_n) \approx -\frac{1}{18.0274}\left(y_{n+4}^{[1]} - y_{n+4}^{[0]}\right)$$

et l'intervalle de stabilité absolue est $(-0.78, 0)$.

4.14 Mise en œuvre

On passe enfin au contrôle du pas. Par exemple, pour une méthode à 4 pas, si l'on double h on n'a qu'à employer y et f à $x_{n+3}, x_{n+1}, x_{n-1}$ et x_{n-3}, qu'on aurait eu soin de conserver en mémoire, pour calculer y_{n+1}. Mais si l'on diminue le pas de moitié, on aura besoin de y et f à $x_{n+3/2}$ et $x_{n+5/2}$. On peut recourrir à un des procédés suivants.

(i) On peut employer Runge–Kutta (plus facile à programmer), l'algorithme de Taylor ou la méthode d'Obrechkoff à un pas.

(ii) On peut pour obtenir les valeurs de y manquantes interpoler y à $x_{n+5/2}$ et $x_{n+3/2}$ à l'ordre du PC près.

(iii) On peut recourrir aux formules de Ceschino: Ceschino a construit 3 formules explicites et 2 formules implicites d'ordre 4 à utiliser lorsque le pas change de h à ωh.

(iv) Enfin, Nordsieck a suggéré d'emmagasiner les dérivées de l'interpolant polynomial local, qui représente la solution, évaluées à un seul point, au lieu d'emmagasiner des valeurs rétrogrades de y et de f. On utilise cette méthode lorsque les évaluations sont dispendieuses. Gear a mis en œuvre la méthode de Nordsieck: il emploie les PC d'Adams–Bashforth–Moulton d'ordre 1 à 7 en mode $P(EC)^m$. Le programme s'autodémarre et ajuste le pas automatiquement. De (4.58) on a le PC:

$$P: y_{n+k}^{[0]} \quad = \quad y_{n+k-1}^{[m]} + h\sum_{j=0}^{k-1} b_j^* f_{n+j}^{[m-1]} \tag{4.76}$$

$$C : y_{n+k}^{[s+1]} = y_{n+k-1}^{[m]} + h\beta_k f_{n+k}^{[s]} + h\sum_{j=0}^{k-1} \beta_j f_{n+j}^{[m-1]} \qquad s = 0, \ldots, n-1,$$

avec $\beta_0 = 0$ en mode $P(EC)^m$. Ainsi

$$y_{n+k}^{[s+1]} - y_{n+k}^{[s]} = h\beta_k \left(f_{n+k}^{[s]} - f_{n+k}^{[s-1]} \right), \qquad s = 1, \ldots, m-1, \tag{4.77}$$

et

$$y_{n+k}^{[1]} - y_{n+k}^{[0]} = h\beta_k \left(f_{n+k}^{[0]} - \sum_{j=0}^{k+1} \frac{\beta_j^* - \beta_j}{\beta_k} f_{n+j}^{[m-1]} \right). \tag{4.78}$$

On pose

$$\delta_j^* = (\beta_j^* - \beta_j)/\beta_k, \qquad j = 0, 1, \ldots, k-1,$$

et

$$d_{n+k} = \sum_{j=0}^{k-1} \delta_j^* f_{n+j}^{[m-1]}.$$

Alors (4.78) devient

$$y_{n+k}^{[1]} - y_{n+k}^{[0]} = h\beta_k \left(f_{n+k}^{[0]} - d_{n+k} \right); \tag{4.79}$$

ainsi d_{n+k} joue le rôle de $f_{n+k}^{[-1]}$. On définit le vecteur d'ordre $(k+1)$:

$$\begin{aligned} \mathbf{y}_{n+k}^{[s]} &= \left[y_{n+k}^{[0]}, h d_{n+k}, h f_{n+k-1}^{[m-1]}, \ldots, h f_{n+1}^{[m-1]} \right]^T, \qquad s = 0, \\ &= \left[y_{n+k}^{[s]}, h f_{n+k}^{[s-1]}, h f_{n+k-1}^{[m-1]}, \ldots, h f_{n+1}^{[m-1]} \right]^T, \qquad s = 1, 2, \ldots, m. \end{aligned}$$

Alors le PC de (4.76) devient

$$\begin{aligned} P : \mathbf{y}_{n+k}^{[0]} &= B\mathbf{y}_{n+k-1}^{[m]}, \\ C : \mathbf{y}_{n+k}^{[s+1]} &= \mathbf{y}_{n+k}^{[s]} + F(\mathbf{y}_{n+k}^{[s]})\mathbf{c}, \qquad s = 0, 1, \ldots, m-1. \end{aligned} \tag{4.80}$$

où $B \in \mathbb{R}^{(k+1)\times(k+1)}$, $\mathbf{c} \in \mathbb{R}^{k+1}$ et $F \in \mathbb{R}$ sont donnés par:

$$B = \begin{bmatrix} 1 & \beta_{k-1}^* & \beta_{k-2}^* & \cdots & \beta_1^* & \beta_0^* \\ 0 & \delta_{k-1}^* & \delta_{k-2}^* & \cdots & \delta_1^* & \delta_0^* \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \qquad \mathbf{c} = \begin{bmatrix} \beta_k \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

$$\begin{aligned} F(y_{n+k}^{[s]}) &= h\left(f_{n+k}^{[0]} - d_{n+k}\right) := h\left[f\left(x_{n+k}, y_{n+k}^{[0]}\right) - d_{n+k}\right], \qquad s = 0, \\ &= h\left(f_{n+k}^{[s]} - f_{n+k}^{[s-1]}\right) := h\left[f\left(x_{n+k}, y_{n+k}^{[s]}\right) - f_{n+k}^{[s-1]}\right], \qquad (4.81) \\ &\qquad s = 1, 2, \ldots, m-1. \end{aligned}$$

On remarque que la matrice B et et le vecteur $\mathbf{c}$ dépendent seulement des coefficients dans (4.76) et sont indépendants de h. Dans (4.80), on a une reformulation à un pas de la méthode PC (4.76). Toutefois, si l'on prévoit changer h, les difficultés antérieures sont encore présentes puisque le vecteur de valeurs rétrogrades $y_{n+k-1}^{[m]}$ contient des données calculées à différents points.

A la section 3.4 on a obtenu les méthodes linéaires à pas multiples en éliminant les coefficients d'un polynôme d'interpolation $I(x)$ qui représentait la solution locale. Considérons, par exemple, le cas $k = 3$ pour la méthode d'Adams–Bashforth d'ordre $p = 3$; on a le polynôme

$$I(x) = ax^3 + bx^2 + cx + d.$$

On pose

$$\begin{aligned} I(x_{n+3}) &= y_{n+3}^{[0]}, \qquad I(x_{n+2}) = y_{n+2}^{[m]}, \\ I'(x_{n+2}) &= f_{n+2}^{[m-1]}, \qquad I'(x_{n+1}) = f_{n+1}^{[m-1]}, \qquad (4.82) \\ I'(x_n) &= f_n^{[m-1]}. \end{aligned}$$

L'élimination des coefficients a, b, c et d entre ces 5 équations donne:

$$y_{n+3}^{[0]} - y_{n+2}^{[m]} = h\left[\frac{23}{12} f_{n+2}^{[m-1]} - \frac{4}{3} f_{n+1}^{[m-1]} + \frac{5}{12} f_n^{[m-1]}\right].$$

C'est le P prédicteur d'Adams–Bashforth d'ordre 3, pour lequel $C_4 = \frac{3}{24}$ et le terme principal de l'erreur est $C_4 h^4 f^{(4)}(\zeta_n)$. On voit que le vecteur rétrograde

$$\mathbf{y}_{n+2}^{[m]} = \left[y_{n+2}^{[m]}, h f_{n+2}^{[m-1]}, h f_{n+1}^{[m-1]}, h f_n^{[m-1]}\right]^T$$

détermine $I(x)$ uniquement. Donc toute l'information se retrouve en un seul point:

$$\mathbf{z}_{n+2}^{[m]} = \left[I(x_{n+2}), hI'(x_{n+2}), \frac{h^2}{2!} I^{(2)}(x_{n+2}), \frac{h^3}{3!} I^{(3)}(x_{n+2})\right]^T.$$

En fait

$$\mathbf{z}_{n+2}^{[m]} = Q\mathbf{y}_{n+2}^{[m]}, \qquad Q = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 3/4 & -1 & 1/4 \\ 0 & 1/6 & -1/3 & 1/6 \end{bmatrix}. \qquad (4.83)$$

On remarque que la présence des puissances de h dans les composantes de $\mathbf{z}_{n+2}^{[m]}$ rend Q indépendante de h. On vérifie (4.83):

$$I(x_{n+k}) \approx y_{n+k}^{[m]}.$$

En effet,

$$Q\mathbf{y}_{n+2}^{[m]} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 3/4 & -1 & 1/4 \\ 0 & 1/6 & -1/3 & 1/6 \end{bmatrix} \begin{bmatrix} y_{n+2}^{[m]} \\ hf_{n+2}^{[m-1]} \\ hf_{n+1}^{[m-1]} \\ hf_{n}^{[m-1]} \end{bmatrix} = \begin{bmatrix} I(x_{n+2}) \\ hI'(x_{n+2}) \\ h^2/2I''(x_{n+2}) \\ h^3/6I'''(n_{n+1}) \end{bmatrix} + O\left(h^4\right).$$

On voit tout de suite que les deux premières équations sont satisfaites:

$$\begin{aligned} y_{n+2}^{[m]} &= I(x_{n+2}), \\ hf_{n+2}^{[m-1]} &= hI'(x_{n+2}). \end{aligned}$$

On vérifie la $3^{\text{ième}}$ équation:

$$\frac{3}{4}hf_{n+2} - hf_{n+1} + \frac{1}{4}hf_n = \frac{h^2}{2}I''(x_{n+2}) + O\left(h^4\right).$$

On a:

$$\begin{aligned} \text{le } 1^{\text{er}} \text{ M} &= h\left[\frac{3}{4}y'_{n+2} - y'_{n+1} + \frac{1}{4}y'_n\right] \\ &\approx h\left[\frac{3}{4}y'_{n+2} - \left(y'_{n+2} - y''_{n+2}h + y'''_{n+2}\frac{h^2}{2} + O\left(h^3\right)\right)\right. \\ &\quad \left. + \frac{1}{4}\left(y'_{n+2} - y''_{n+2}2h + y'''_{n+2}\frac{4h^2}{2} + O\left(h^3\right)\right)\right] \\ &= h\left[\left(1 - \frac{1}{2}\right)h''_{n+2}h + O\left(h^3\right)\right] \\ &= \frac{1}{2}h^2y''_{n+2} + O\left(h^4\right) \\ &= \frac{h^2}{2}I''(x_{n+2}) + O\left(h^4\right) \\ &= \text{le } 2^{\text{ième}} \text{ M}. \end{aligned}$$

On vérifie maintenant la $4^{\text{ième}}$ équation:

$$\frac{1}{6}hf_{n+2} - \frac{1}{3}hf_{n+1} + \frac{1}{6}hf_n = \frac{h^3}{6}I'''(x_{n+2}) + O\left(h^4\right).$$

On a:

$$\begin{aligned}
\text{le } 1^{\text{er}}\text{ M} &= h\left\{\frac{1}{6}y'_{n+2} - \frac{1}{3}\left[y'_{n+2} - y''_{n+2}h + y'''_{n+2}\frac{4h^2}{2} + O\left(h^3\right)\right]\right. \\
&\quad \left. + \frac{1}{6}\left[y'_{n+2} - y''_{n+2}h + y'''_{n+2}\frac{h^2}{2} + O\left(h^3\right)\right]\right\} \\
&= h\left[\left(-\frac{1}{6} + \frac{1}{3}\right)h^2 y'''_{n+2} + O\left(h^3\right)\right] \\
&= \frac{h^3}{6}y'''_{n+2} + O\left(h^4\right) \\
&= \frac{h^3}{6}I'''(x_{n+2}) + O\left(h^4\right) \\
&= \text{le } 2^{\text{ième}}\text{ M.}
\end{aligned}$$

On peut donc écrire le prédicteur et le correcteur (4.76) en mode $P(EC)^m$ sous la forme:

$$\begin{aligned}
P: \qquad \mathbf{y}_{n+k}^{[0]} &= B\mathbf{y}_{n+k-1}^{[m]} \\
C: \qquad \mathbf{y}_{n+k}^{[s+1]} &= \mathbf{y}y_{n+k}^{[s]} + F\left(\mathbf{y}_{n+k}^{[s]}\right)\mathbf{c}, \qquad s = 0, \ldots, m-1.
\end{aligned} \tag{4.84}$$

pour le PC:

$$\begin{aligned}
P: \qquad y_{n+k}^{[0]} &= -\sum_{j=0}^{k-1}\alpha_j^* y_{n+j} + h\sum_{j=0}^{k-1}\beta_j^* f_{n+j}, \\
C: \qquad y_{n+k}^{[1]} &= \sum_{j=0}^{k-1}\alpha_j y_{n+j} + h\sum_{j=0}^{k-1}\beta_j f_{n+1} + h\beta_k f_{n+k}^{[0]},
\end{aligned}$$

et si l'on corrige de nouveau:

$$\begin{aligned}
y_{n+k}^{[s+1]} &= y_{n+k}^{[s]} + \beta_k\left[hf_{n+k}^{[s]} - hf_{n+k}^{[s-1]}\right], \qquad s = 1, 2, \ldots, m-1, \\
y_{n+k} &= y_{n+k}^{[m]}, \qquad f_{n+k} = f\left(x_{n+k}, y_{n+k}^{[m-1]}\right.
\end{aligned}$$

On récrit le correcteur

$$y_{n+k}^{[1]} = y_{n+k}^{[0]} - \beta_k\left[\sum_{j=0}^{k-1}\left(\frac{\alpha_j^* - \alpha_j}{\beta_k}\right)y_{n+j} + \sum_{j=0}^{k-1}\left(\frac{\beta_j^* - \beta_j}{\beta_k}\right)hf_{n+j} - hf_{n+k}^{[0]}\right].$$

On note

$$\begin{aligned}\mathbf{y}_{n+k} &:= \left[y_{n+k}, y_{n+k-1}, \ldots, y_{n+1}, hy'_{n+k}, \ldots, hy'_{n+k}\right]^T, \\ \mathbf{y}^{[s]}_{n+k} &:= \left[y^{[s]}_{n+k}, y_{n+k-1}, \ldots, y_{n+1}, hf^{[m-1]}_{n+k}, hy'_{n+k-1}, \ldots, hy'_{n+1}\right]^T, \\ & \qquad s = 1, 2, \ldots, m,\end{aligned}$$

et

$$\mathbf{y}^{[0]}_{n+k} := \left[y^{[0]}_{n+k}, y_{n+k-1}, \ldots, y_{n+1}, d_{n+k}, hy'_{n+k-1}, \ldots, hy'_{n+1}\right]^T,$$

où

$$d_{n+k} = -\sum_{j=0}^{k-1} \left(\frac{\alpha^*_j - \alpha_j}{\beta_k}\right) y_{n+j} + \left(\frac{\beta^*_j - \beta_j}{\beta_k}\right) hy'_{n+j}.$$

Soit la matrice $2k \times 2k$

$$B = \left[\begin{array}{cccc|cccc} -\alpha^*_{k-1} & -\alpha^*_{k-2} & \cdots & -\alpha^*_0 & \beta^*_{k-1} & \beta^*_{k-2} & \cdots & \beta^*_0 \\ 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots & \vdots & \vdots & & \vdots \\ \vdots & & & \vdots & \vdots & \vdots & & \vdots \\ 0 & & 1 & 0 & 0 & 0 & \cdots & 0 \\ \hline \gamma_{k-1} & \gamma_{k-2} & \cdots & \gamma_0 & \delta_{k-1} & \delta_{k-2} & \cdots & \delta_0 \\ 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & 0 & 1 & & \vdots \\ \vdots & \vdots & & \vdots & \vdots & & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & & 1 & 0 \end{array}\right]$$

où

$$\gamma_j = -\frac{\alpha^*_j - \alpha_j}{\beta_k}, \qquad \delta_j = \frac{\beta^*_j - \beta_j}{\beta_k},$$

et

$$\mathbf{c} = [\beta_k, 0, \ldots, 0_k, 1, 0, \ldots, 0_{2k}]^T \in \mathbb{R}^{2k}.$$

Alors le PC devient:

$$\begin{aligned} P: \quad \mathbf{y}^{[0]}_{n+k} &= B\mathbf{y}_{n+k-1} \\ C: \quad \mathbf{y}^{[1]}_{n+k} &= \mathbf{y}^{[0]}_{n+k} + F\left(\mathbf{y}^{[0]}_{n+k}\right)\mathbf{c}, \end{aligned}$$

où

$$F(\mathbf{y}^{[s]}_{n+k}) = \begin{cases} h\left(f^{[0]}_{n+k} - d_{n+k}\right), & \text{si} \quad s = 0, \\ h\left(f^{[s]}_{n+k} - f^{[s-1]}_{n+k}\right), & \text{si} \quad s = 1, 2, \ldots, m-1, \end{cases}$$

et on corrige de nouveau:

$$\begin{aligned} \mathbf{y}_{n+k}^{[s+1]} &= \mathbf{y}_{n+k}^{[s]} + F\left(\mathbf{y}_{n+k}^{[s]}\right)\mathbf{c}, \\ \mathbf{y}_{n+k} &= \mathbf{y}_{n+k}^{[m]}. \end{aligned}$$

Remarque. Puisque les première et deuxième composantes de $\mathbf{y}_{n+k-1}^{[m]}$ et de $\mathbf{y}_{n+k-1}^{[m]}$ sont identiques, les première et deuxième lignes de Q sont respectivement $\mathbf{e}_1^T$ et $\mathbf{e}_2^T$. Toutefois $F(\mathbf{v}) = F(Q\mathbf{v})$. Alors le prédicteur et le correcteur s'écrivent sous la forme d'une méthode à un pas:

$$\begin{aligned} \mathbf{z}_{n+k}^{[0]} &= QBQ^{-1}\mathbf{z}_{n+k-1}^{[m]} \\ \mathbf{z}_{n+k}^{[s+1]} &= \mathbf{z}_{n+k}^{[s]} + F\left(\mathbf{z}_{n+k}^{[s]}\right)\mathbf{l}, \qquad s = 0, 1, \ldots, m-1, \end{aligned} \tag{4.85}$$

où $\mathbf{l} = Q\mathbf{c}$. On voit l'avantage de cette formulation: le vecteur rétrograde $\mathbf{z}_{n+k-1}^{[m]}$ contient toute l'information au seul point x_{n+k-1}. Pour changer le pas, on n'a qu'à multiplier la $i^{\text{ième}}$ composante par $\alpha^i, i = 0, \ldots, k$. C'est la méthode de Gear.

4.15 Comparaison entre les méthodes PC et RK

On considère les critères de comparaison suivants:

(i) la précision locale et l'ordre;

(ii) la stabilité faible;

(iii) le nombre d'évaluations de la fonction f: 2 évaluations pour PC;

(iv) la facilité de programmation.

(i) On choisit P et C du même ordre p; alors l'erreur locale principale du PC est

$$C_{p+1}h^{p+1}y^{(p+1)}(x_n), \tag{4.86}$$

et l'erreur locale principale de RK est

$$\psi(x_n, y(x_n))h^{p+1}. \tag{4.87}$$

Pour les méthodes d'ordre 4, RK est plus précise que PC. Puisque RK exige p évaluations de la fonction f par pas, alors que PC n'exige que 2 évaluations de la fonction par pas, on peut avoir le pas h avec RK et le pas $h = 2/p$ avec PC.

Il s'ensuit qu'on a le même nombre d'évaluations de la fonction; alors l'erreur du PC est

$$C_{p+1}\left(\frac{2h}{p}\right)^{p+1} y^{(p+1)}(x_n) \tag{4.88}$$

et sur cette base le PC est fréquemment plus précis que RK.

Exemple 4.15 Soit l'équation $y' = \lambda y$, $y(0) = 1$. Considérons les méthodes RK et PC d'ordre 4 où le prédicteur P est à 4 pas et le correcteur C est à 3 pas. Montrer que $|T_{RK}| > |T^*_{PC}|$, toute méthodes PC et $|T_{PC}| > |T|_{RK}$ pour certaines méthodes PC.

On a

$$T_{RK} = \frac{h^5\lambda^5}{120}, \qquad T_{PC} = C_5 h^5 \lambda^5, \qquad T^*_{PC} = \frac{C_5 h^5 \lambda^5}{32}.$$

Or, de la section 3.10, on a pour $k = 3$, $p = 4$,

$$C_5 = \frac{-19 + 11a + 19b}{720}$$

Pour avoir la zéro stabilité, d'après le critère d'Hurwitz, les paramètres a et b doivent satisfaire:

$$1 + a + b > 0, \qquad 1 - b > 0, \qquad 1 - a + b > 0.$$

Donc

$$c_5 < 1/2$$

et

$$|T^*_{PC}| < \frac{1}{12} h^5 \frac{|\lambda^5|}{32} < |T_{RK}|.$$

D'autre part, avec $a = b = 0$, on a

$$|T_{PC}| = \frac{19}{720} h^5 |\lambda^5| > |T_{RK}|.$$

Exemple 4.16 Comparer la précision globale de RK et $PECE$ d'ordre 4 pour

$$y' = -y, \qquad y(0) = 1, \qquad 0 \le x \le 10.$$

(ii) Voici une table de l'intervalle de stabilité absolue pour certaines méthodes d'ordre 4:

Runge–Kutta	$(-2.78, 0)$
A-B-M $(PECE)$	$(-1.25, 0)$
A-B-M $(P(EC)^2)$	$(-0.90, 0)$
Crane-Klopfenstein $(PECE)$	$(-2.48, 0)$

(iii) Pour le même prix on a le même nombre d'évaluation de la fonction sur un intervalle donné avec PC et $h/2$ et RK et h.

(iv) Quant à la facilité de programmation, RK est plus facile, mais le contrôle du pas avec PC est plus facile.

Chapitre 5

MÉTHODES DE RUNGE–KUTTA–NYSTRÖM

5.1 Introduction

On peut obtenir la solution numérique de systèmes d'équations différentielles du second ordre de deux façons différentes. La première consiste à transformer le système en un système d'équations différentielles du premier ordre et à appliquer les méthodes de Runge-Kutta bien connues. La deuxième consiste à appliquer une méthode directe inventée par Nyström [10], qu'on nomme méthode Runge-Kutta-Nyström et qu'on note RKN. Les méthodes explicites RKN forment une classe d'algorithmes numériques intéressante pour résoudre les problèmes aux valeurs initiales de systèmes non raides d'équations différentielles du second ordre de la forme:

$$y'' = f(x, y, y'), \quad y(x_0) = y_0, \quad y'(x_0) = y'_0,$$

où $f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$.

Le présent travail traite le cas spécial où f est indépendante de y', c'est-à-dire on considère le système:

$$y'' = f(x, y), \quad y(x_0) = y_0, \quad y'(x_0) = y'_0, \tag{5.1}$$

où $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$.

Comme c'est le cas pour les systèmes d'équations différentielles du premier ordre, la majorité des algorithmes de Runge-Kutta-Nyström actuels utilisent des paires de formules emboîtées, c'est-à-dire une formule d'ordre inférieur $p - 1$ et

une formule d'ordre supérieur p . La technique des méthodes de Runge-Kutta-Nyström emboîtées est généralement reconnue comme étant une technique efficace pour la résolution numérique des problèmes aux valeurs initiales du second ordre.

Les méthodes RKN explicites utilisent les approximations y_0 pour $y(x_0)$ et y'_0 pour $y'(x_0)$ et un pas de longueur h_0 pour évaluer f à plusieurs points entre x_0 et $x_1 = x_0 + h_0$ et ainsi approximer $y(x_1)$ à ces points.

Lorsque la formule d'ordre supérieur est utilisée pour avancer la solution numérique d'un pas, nous parlons de mode d'extrapolation locale. Par contre, lorsque la formule d'ordre inférieur est utilisée pour avancer la solution numérique, alors nous parlons de mode standard.

A la section 2, on étudie la théorie des méthodes de Nyström d'une manière detaillée et on présente les arbres de Nyström.

A la section 3, on présente, dans une forme convenable, les conditions d'ordre jusqu'à l'ordre six. Puis à la section 4, on donne les preuves d'inexistence de méthodes à $s-1$ stages et on donne des méthodes à s stages pour les types I à V. Finalement, à la section 5, on résume les résultats obtenus et on indique des extensions possibles à notre travail.

On ne discutera pas des applications pratiques des résultats numériques obtenus.

Tous les calculs ont été effectués avec MAPLE sur l'ordinateur Amdahl installé à l'Université d'Ottawa.

Ce chapitre est essentiellement la thèse de Maîtrise en informatique de M. Fadi MALEK [9].

5.2 Théorie des méthodes de Nyström

5.2.1 Paires de formules

On considère le système d'équations différentielles du second ordre:

$$y'' = f(x,y), \quad y(x_0) = y_0, \quad y'(x_0) = y'_0,$$

où $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$.

Une paire de formules emboîtées de Runge-Kutta-Nyström (RKN) à s stages se définit par les formules de récurrence suivantes:

$$y_{i+1} = u_i + h_i u'_i + h_i^2 \sum_{j=1}^{s} b_j f_j, \qquad y'_{i+1} = u'_i + h_i \sum_{j=1}^{s} b'_j f_j, \tag{5.2}$$

$$\hat{y}_{i+1} = u_i + h_i u'_i + h_i^2 \sum_{j=1}^{s} \hat{b}_j f_j, \qquad \hat{y}'_{i+1} = u'_i + h_i \sum_{j=1}^{s} \hat{b}'_j f_j, \tag{5.3}$$

où $h_i = x_{i+1} - x_i$ et

$$f_1 = f(x_i, u_i), \tag{5.4}$$

$$f_j = f\big(x_i + h_i c_j, u_i + h_i c_j u_i' + h_i^2 \sum_{k=1}^{j-1} a_{jk} f_k\big), \quad j = 2, \ldots, s. \tag{5.5}$$

Les deux formules de (5.2) sont d'ordre p et celles de (5.3) sont d'ordre $p-1$. Si on avance les approximations numériques par les formules d'ordre $p-1$, alors $u_i = \hat{y}_i$, $u_i' = \hat{y}_i'$. Si on avance les approximations numériques par les formules d'ordre p, alors $u_i = y_i$, $u_i' = y_i'$.

Pour les méthodes de Runge-Kutta-Nyström, $\hat{b}_j$ et b_j sont les poids des formules de la solution, $\hat{b}_j'$ et b_j' les poids des formules de la dérivée, c_j les nœuds, a_{jk} les coefficients de couplage de la paire de formules et h_i le pas d'intégration. Les y_i, y_{i+1}, y_i' et y_{i+1}' représentent les approximations de la solution y et de la dérivée y' aux points x_i et x_{i+1}. Le choix des paramètres mentionés ci-haut a pour but de maximiser le nombre de termes des développements de Taylor respectifs de y_{i+1} et y_{i+1}' qui sont identiques à ceux des développoments de $y(x_i + h_i)$ et $y'(x_i + h_i)$ lorsque l'on suppose que $y_i = y(x_i)$ et $y_i' = y'(x_i)$.

On présente, au tableau 1, dit de Butcher, les coefficients d'une paire de formules de Runge-Kutta-Nyström emboîtées.

Tableau 1. Coefficients d'une paire de formules de Runge-Kutta-Nyström emboîtées à s stages.

c_1	0				
c_2	a_{21}	0			
c_3	a_{31}	a_{32}	0		
$\vdots$	$\vdots$	$\vdots$		$\ddots$	
c_s	a_{s1}	a_{s2}	$\cdots$	$a_{s\,s-1}$	0
$\hat{b}$	$\hat{b}_1$	$\hat{b}_2$	$\cdots$	$\cdots$	$\hat{b}_s$
b	b_1	b_2			b_s
$\hat{b}'$	$\hat{b}_1'$	$\hat{b}_2'$			$\hat{b}_s'$
b'	b_1'	b_2'	$\cdots$	$\cdots$	b_s'

5.3 Théorie des arbres

Pour mieux comprendre et mieux apprécier l'analyse des paires de formules pour les méthodes RKN, que l'on étudiera dans ce travail, il est important d'expliquer quelques notions sur les arbres de Nyström et la dérivation des conditions d'ordre qui ont été établies par Butcher et Hairer:

Définition 5.1 (Butcher [3], pp. 80 et 88, v. Verner [12]) *Un arbre enraciné t, est un graphe connexe sans boucle où l'on choisit pour racine un nœud à l'une des extrémités de l'arbre. L'ordre $r(t)$ est égal au nombre de nœuds de l'arbre. La hauteur $h(t)$ est le nombre d'arcs du plus long chemin relié à la racine.*

Définition 5.2 (Hairer [8], p. 145. Arbres enracinés étiquetés) *Soit A une chaîne d'indices ordonnés: $A = \{j < k < l < m < \ldots\}$ et A_q le sous-ensemble des q premiers indices.*

Un arbre enraciné étiqueté d'ordre q, $q \geq 1$, est une application:

$$t : A_q - \{j\} \to A_q$$

telle que $t(z) < z$ pour tout $z \in A_q - \{j\}$. L'ensemble de tous les arbres étiquetés est noté LT_q ("Labelled Trees of order q"). On appelle z l'enfant de $t(z)$, $t(z)$ le parent de z et le nœud j la racine de l'arbre.

Afin de distinguer la première dérivée de la seconde dans les arbres associés aux méthodes RKN, il est nécessaire d'introduire deux sortes de nœuds: les *nœuds épais* et les *nœuds minces.*

Ceci nous conduit à définir certains termes relatifs aux arbres de Nyström.

Définition 5.3 (Hairer [8], p. 263) *Un arbre de Nyström (N-arbre) étiqueté d'ordre q est un arbre étiqueté (v. la définition précédente) donné par l'application:*

$$t : A_q - \{j\} \to A_q,$$

et une second application:

$$t' : A_q \to \{mince,\ épais\},$$

qui satisfait les règles suivantes:

a) La racine de l'arbre t est toujours épaisse, c'est-à-dire $t'(j) = épais$.

b) Un nœud mince admet au plus un enfant, et cet enfant doit être épais. Les arbres de Nyström étiquetés d'ordre q sont notés LNT_q ("Labelled Nyström Trees of order q").

Deux arbres de Nyström étiquetés t et u sont équivalents (v. [8], p. 264) s'ils sont du même ordre q et se distinguent seulement par une permutation de leurs indices. L'ensemble des arbres de Nyström équivalents forment une classe d'équivalence qu'on note NT_q. De plus, on note $\alpha(t)$ le nombre d'éléments de la classe d'équivalence.

La solution exacte de l'équation différentielle (5.1) est donnée par le développement de Butcher:

$$\begin{aligned} y^{(q)} &= \sum_{t \in LNT_{q-1}} F(t)(y) \\ &= \sum_{t \in NT_{q-1}} \alpha(t) F(t)(y) \end{aligned}$$

où la différentielle $F(t)(y)$ est une somme sur les indices de tous les nœuds épais de t, (sauf la racine j), et sur les indices de tous les nœuds minces extrémaux ("end-vertex").

Afin de calculer la solution numérique, on a recours à quelques définitions formulées par Butcher et par Hairer:

Définition 5.4 (Butcher [3], p. 132, v. Verner [12]) *Pour chaque arbre enraciné, t, d'ordre $r(t)$, la fonction $\gamma(t)$ est l'entier positif calculé de la manière suivante: on assigne un entier à chaque nœud; les nœuds extrémaux prennent la valeur 1; les autres nœuds prennent la valeur 1 plus la somme des valeurs de tous leurs successeurs immédiats. En particulier, la valeur assignée à la racine est $r(t)$. La valeur de $\gamma(t)$ est le produit de tous les entiers assignés aux nœuds de t.*

Définition 5.5 (Hairer [8], p. 267) *Un N-arbre s'appelle arbre spécial ou SN-arbre si les nœuds épais ont seulement des enfants minces.*

Les SN-arbres sont conçus spécifiquement pour les équations différentielles spéciales (5.1).

Définition 5.6 (Hairer [8], p. 266) *Pour chaque SN-arbre de Nyström t, la fonction $\varphi_j(t)$ est définie comme la sommation sur les indices de tous les nœuds épais de t, sauf la racine j.*

Le terme général de $\varphi_j(t)$ est la somme du produit de:

(a) a_{kl} si le nœud épais k est relié à un enfant mince l,

(b) c_k^m si le nœud épais k est relié à m nœuds minces qui sont situés au-dessus et ne sont pas reliés à d'autres nœuds.

La solution numérique de l'équation différentielle (5.1) est donnée par le développement de Butcher:

$$\begin{aligned} y_{i+1}^{(q)} &= q \sum_{t \in LNT_{q-1}} \gamma(t) \sum_{j=1}^{s} b_j \varphi_j(t) F(t)(y_i), \\ y'^{(q-1)}_{i+1} &= \sum_{t \in LNT_{q-1}} \gamma(t) \sum_{j=1}^{s} b'_j \varphi_j(t) F(t)(y_i). \end{aligned}$$

Une méthode de Nyström pour le système d'équations différentielles (5.1) est d'ordre p si et seulement si (v. [8], p. 267)

$$\sum_j b_j \varphi_j(t) = \frac{1}{(r(t)+1)\gamma(t)}$$

pour les SN-arbres t d'ordre $r(t) \leq p-1$, et

$$\sum_j b'_j \varphi_j(t) = \frac{1}{\gamma(t)}$$

pour les SN-arbres t d'ordre $r(t) \leq p$.

5.4 Les cinq types de formules

On peut distinguer cinq types de paires différents, selon les formules de (5.2) et (5.3) qu'on utilise. Ces cinq types sont representés au tableau 2 où 1 indique que la formule correspondante est utilisée et 0 indique qu'elle ne l'est pas. On note ces paires de type I, II, III, IV et V respectivement.

Tableau 2. *Représentation schématique des cinq types.*

I	$p-1$	p
y	1	1
y'	1	1

II	$p-1$	p
y	1	1
y'	1	0

III	$p-1$	p
y	1	1
y'	0	1

IV	$p-1$	p
y	1	0
y'	1	1

V	$p-1$	p
y	0	1
y'	1	1

Les paires du type I contrôlent l'erreur locale en y et y'. Elles peuvent être utilisées en mode standard (où le pas se fait au moyen des formules d'ordre $p-1$), ou en mode d'extrapolation locale (où le pas se fait au moyen des formules d'ordre p). Les paires du type I sont l'œuvre de plusieurs chercheurs, entre autres, Bettis (1973) [1], Bettis–Horn (1978) [2], et Dormand–El-Mikkway–Prince (1987) [4]–[5].

Les paires du type II et III n'ont qu'une formule pour la dérivée; elles contrôlent donc l'erreur locale en y seulement. Les paires du type II (respectivement du type III) n'ont pas la formule de la dérivée d'ordre supérieur (respectivement d'ordre inférieur); on ne les utilise donc qu'en mode standard (respectivement en mode d'extrapolation locale). Quelques exemples de paires des types II et III ont été construites respectivement par Fehlberg–Filippi–Gräf (1986) [6], et Filippi–Gräf (1986) [7]. Voir aussi [11].

Les paires du type IV (respectivement V) n'ont qu'une formule pour la solution; elles contrôlent donc l'erreur locale en y' seulement. Puisque les paires du

type IV (respectivement V) n'ont que la formule de la solution d'ordre inférieur (respectivement d'ordre supérieur), alors elle ne peuvent être utilisés qu'en mode standard (respectivement en mode d'extrapolation locale).

Si on échange les formules de la solution et de la dérivée, les paires des types II et III deviennent semblables à celles des types IV et V. Cependant, il y a une différence importante: presque toutes les paires des types IV et V deviennent des paires du type I par les transformations suivantes:

$$\hat{b}_i = (1 - c_i)\hat{b}'_i, \quad b_i = (1 - c_i)b'_i, \quad i = 1, \ldots, s.$$

Il est évident que s'il existe une paire de méthodes à s-stages du type I d'ordre $(p-1, p)$, alors il existe des paires à s-stages des types II à V d'ordre $(p-1, p)$. On peut poser la question suivante: "Existe-t-il des paires des types II à V qui ont moins de stages que celles du même ordre du type I?"

Comme c'est le cas pour les paires de Runge-Kutta, on utilise ordinairement le nombre minimum de stages possible pour les paires RKN. Dans ce travail, on détermine le nombre minimum de stages pour les paires des types I à V d'ordre $(p-1, p)$ où $p = 2, \ldots, 6$. On a inclus les preuves pour les paires d'ordre $p = 2$ et $p = 3$ bien qu'elles soient simples. Toutes les preuves d'inexistence sont par contradiction: pour chaque type de paires et chaque valeur de p, on suppose l'existence d'une paire qui admet un stage de moins que la valeur minimum. On montre alors, pour les paires du type I, soit que les conditions d'ordre pour les formules de la dérivée ne peuvent pas être satisfaites ou que les formules d'ordre $(p-1)$ et p de la dérivée sont identiques. Puisque les résultats pour les paires du type I sont basés exclusivement sur les formules de la dérivée, ils s'appliquent immédiatement aux paires des types IV et V. Pour les paires des types II et III, les démonstrations utilisent les conditions d'ordre des formules de la dérivée et de la solution, pour déduire que celles-ci ne peuvent pas être satisfaites ou que les deux formules de la solution sont identiques.

5.5 Conditions d'ordre

Les conditions d'ordre pour la formule de la dérivée d'ordre $(p-1)$ comprennent les équations de quadrature:

$$\sum_{i=1}^{s} \hat{b}'_i c_i^k = \frac{1}{k+1}, \quad k = 0, 1, \ldots, p-2, \tag{5.6}$$

et les équations de non quadrature:

$$\sum_{i=1}^{s} \hat{b}'_i S^q_{i,k}(a, c) = 0, \quad k = 1, \ldots, N_q - 1, \quad q = 1, \ldots, p-1, \tag{5.7}$$

où N_q est le nombre d'équations de conditions d'ordre q.

Les conditions d'ordre pour la formule de la solution d'ordre $p-1$ sont

$$\sum_{i=1}^{s} \hat{b}_i c_i^k = \frac{1}{(k+1)(k+2)}, \quad k = 0, 1, \ldots, p-3, \tag{5.8}$$

et

$$\sum_{i=1}^{s} \hat{b}_i S_{i,k}^q(a,c) = 0, \quad k = 1, \ldots, N_q - 1, \quad q = 1, \ldots, p-1. \tag{5.9}$$

Les conditions d'ordre pour la formule de la dérivée d'ordre p sont

$$\sum_{i=1}^{s} b_i' c_i^k = \frac{1}{k+1}, \quad k = 0, 1, \ldots, p-1, \tag{5.10}$$

et

$$\sum_{i=1}^{s} b_i' S_{i,k}^q(a,c) = 0, \quad k = 1, \ldots, N_q - 1, \quad q = 1, \ldots, p. \tag{5.11}$$

Les conditions d'ordre pour la formule de la solution d'ordre p sont

$$\sum_{i=1}^{s} b_i c_i^k = \frac{1}{(k+1)(k+2)}, \quad k = 0, 1, \ldots, p-2, \tag{5.12}$$

et

$$\sum_{i=1}^{s} b_i S_{i,k}^q(a,c) = 0, \quad k = 1, \ldots, N_q - 1, \quad q = 1, \ldots, p. \tag{5.13}$$

Aux colonnes 1 et 3 du tableau 3, on présente les $S_{i,k}^q$ pour les conditions d'ordre sur y' et $\hat{y}'$ de l'ordre 1 à 6; les indices j répétés impliquent une sommation et

$$Q_{i,k} := \frac{c_i^{k+2}}{(k+1)(k+2)} - \sum_{j=1}^{i-1} a_{ij} c_j^k, \quad i = 2, \ldots, s. \tag{5.14}$$

Aux colonnes 2 et 3 du tableau 3, on présente les $S_{i,k}^q$ pour les conditions d'ordre sur y et $\hat{y}$ de l'ordre 1 à 6.

Tableau 3. *Les $S_{i,k}^q(a,c)$ d'ordre q, de un à six, pour les formules respectivement de la dérivée et de la solution.*

q pour y' et $\hat{y}'$	q pour y et $\hat{y}$	$S_{i,k}^q$			
1, 2, 3	1, 2, 3, 4	aucune			
4	5	Q_{i1}			
5	6	$c_i Q_{i1}$	Q_{i2}		
6		$c_i^2 Q_{i1}$	$c_i Q_{i2}$	Q_{i3}	$a_{ij} Q_{j1}$

Au tableau 3, on a supposé que l'hypothèse simplificatrice suivante est satisfaite:

$$\frac{c_i^2}{2} = \sum_{j=1}^{i-1} a_{ij}, \qquad i = 1, \ldots, s. \tag{5.15}$$

Cette hypothèse exprime le fait que la valeur approchée de y à chaque point $x_i + h_i c_j$,

$$u_i + h_i c_j u_i' + h_i^2 \sum_{k=1}^{j-1} a_{jk} f_k,$$

où l'on évalue f_j dans (5.5), approxime la solution locale $y(x_i + h_i c_j)$ à l'ordre $O(h_i^3)$ près. Sous cette hypothèse, certaines des conditions de non quadratures deviennent identiques, d'où une réduction du nombre de conditions distinctes. Par la suite, on supposera toujours l'hypothèse (5.15).

Les hypothèses simplificatrices suivantes:

$$b_i = b_i'(1 = c_i), \quad \hat{b}_i = \hat{b}_i'(1 - c_i), \qquad i = 1, \ldots, t, \tag{5.16}$$

impliquent l'existence d'une formule pour la solution y s'il existe une formule pour la dérivée y'. On supposera l'une ou l'autre ou les deux de ces hypothèses selon les circonstances.

Pour réduire de un le nombre d'évaluations de f par pas, on ré-utilise le dernier stage du pas actuel comme premier stage du pas suivant. On appelle FSAL ("First Same As Last") les paires qui ré-utilisent le dernier stage et dans le cas contraire, non FSAL.

Pour les paires FSAL, on doit avoir $c_s = 1$ et on doit aussi satisfaire les conditions suivantes:

$$\hat{b}_s = 0, \quad a_{s,j} = \hat{b}_j, \quad j = 1, \ldots, s-1, \quad Q_{s1} = 0. \tag{5.17}$$

en mode standard, et les conditions semblables:

$$b_s = 0, \quad a_{s,j} = b_j, \quad j = 1, \ldots, s-1, \quad Q_{s1} = 0, \tag{5.18}$$

en mode d'extrapolation locale.

On notera $c_i^0 = 1$ même si $c_i = 0$.

5.6 Nombre minimum de stages

5.6.1 Paires du type I

Comme on l'a déjà mentionné, les paires du type I contrôlent l'erreur locale en y et y'. On doit donc considérer les équations de conditions en y, y', $\hat{y}$ et $\hat{y}'$ afin de

trouver les méthodes pour les paires RKN de l'ordre $(1,2)$ jusqu'à l'ordre $(5,6)$.

Paires d'ordre $(1,2)$ du type I.

Il n'existe pas de méthode à un seul stage ($s=1$) pour les paires de formules du type I. En effet, les équations de conditions pour les ordres 1 et 2 sont respectivement:

$$\begin{array}{ll} \text{Ordre 1 en } \hat{y}: & \text{aucune condition d'ordre,} \\ \text{Ordre 2 en } y: & b_1 = \frac{1}{2}, \\ \text{Ordre 1 en } \hat{y}': & \hat{b}'_1 = 1, \\ \text{Ordre 2 en } y': & b'_1 = 1, \\ & b'_1 c_1 = \frac{1}{2}. \end{array}$$

On constate qu'on ne peut pas satisfaire l'équation de quadrature d'ordre 2 en y' puisque $c_1 = 0$. Il s'ensuit que quelque soit le mode d'avancement, il n'existe pas de méthode à un stage pour les paires de formules d'ordre $(1,2)$.

Il existe des méthodes FSAL à deux stages pour les paires de formules d'ordre $(1,2)$ du type I. On donne ci-dessous les équations de conditions pour ces paires:

$$\begin{array}{ll} \text{Ordre 1 en } \hat{y}: & \text{aucune condition,} \\ \text{Ordre 2 en } y: & b_1 + b_2 = \frac{1}{2}, \\ \text{Ordre 1 en } \hat{y}': & \hat{b}'_1 + \hat{b}'_2 = 1, \\ \text{Ordre 2 en } y': & b'_1 + b'_2 = 1, \\ & b'_2 c_2 = \frac{1}{2}. \end{array}$$

Pour les paires FSAL on a:

$$c_2 = 1, \begin{cases} \hat{b}_2 = 0, & a_{21} = \hat{b}_1, \quad \text{en mode standard,} \\ b_2 = 0, & a_{21} = b_1, \quad \text{en mode d'extrapolation locale.} \end{cases}$$

Le système ci-dessus admet la solution:

$$b'_2 = \frac{1}{2}, \quad b'_1 = \frac{1}{2},$$

où b_1, $\hat{b}_1$, $\hat{b}'_1$ sont arbitraires en mode standard et $\hat{b}_1$, $\hat{b}_2$ et $\hat{b}'_1$ sont arbitraires en mode d'extrapolation locale.

Donc il existe des méthodes FSAL d'ordre $(1,2)$ dans les deux modes.

Paires d'ordre $(2,3)$ du type I.

Il n'existe pas de méthode à deux stages pour les paires d'ordre $(2,3)$ du type I. Les équations de conditions pour ces paires sont les suivantes:

$$\begin{array}{ll}
\text{Ordre 2 en } \hat{y}: & \hat{b}_1 + \hat{b}_2 = \frac{1}{2}, \\
\text{Ordre 3 en } y: & b_1 + b_2 = \frac{1}{2}, \\
& b_2 c_2 = \frac{1}{6}, \\
\text{Ordre 2 en } \hat{y}': & \hat{b}'_1 + \hat{b}'_2 = 1, \\
& \hat{b}'_2 c_2 = \frac{1}{2}, \\
\text{Ordre 3 en } y': & b'_1 + b'_2 = 1, \\
& b'_2 c_2 = \frac{1}{2}, \\
& b'_2 c_2^2 = \frac{1}{3}.
\end{array}$$

Les équations en y' nous donnent:

$$c_2 = \frac{2}{3}, \qquad b'_2 = \frac{3}{4}, \qquad b'_1 = \frac{1}{4},$$

et les équations en $\hat{y}'$ nous donnent:

$$\hat{b}'_2 = \frac{3}{4}, \qquad \hat{b}'_1 = \frac{1}{4}.$$

Donc

$$b'_i = \hat{b}'_i, \qquad i = 1, 2,$$

et on déduit que quelque soit le mode d'avancement, il n'existe pas de méthode à deux stages pour les paires de formules $(2,3)$.

Il existe des méthodes FSAL à trois stages pour les paires de formules d'ordre $(2,3)$. Les équations de conditions pour ces paires sont les suivantes:

$$\begin{array}{ll}
\text{Ordre 2 en } \hat{y}: & \hat{b}_1 + \hat{b}_2 + \hat{b}_3 = \frac{1}{2}, \\
\text{Ordre 3 en } y: & b_1 + b_2 + b_3 = \frac{1}{2}, \\
& b_2 c_2 + b_3 c_3 = \frac{1}{6}, \\
\text{Ordre 2 en } \hat{y}': & \hat{b}'_1 + \hat{b}'_2 + \hat{b}'_3 = 1, \\
& \hat{b}'_2 c_2 + \hat{b}'_3 c_3 = \frac{1}{2}, \\
\text{Ordre 3 en } y': & b'_1 + b'_2 + b'_3 = 1, \\
& b'_2 c_2 + b'_3 c_3 = \frac{1}{2}, \\
& b'_2 c_2^2 + b'_3 c_3^2 = \frac{1}{3}.
\end{array}$$

Pour les paires FSAL on a:

$$c_3 = 1 \text{ et } \left\{ \begin{array}{lll}
\hat{b}_3 = 0, & a_{31} = \hat{b}_1, a_{32} = \hat{b}_2 & \text{en mode standard,} \\
b_3 = 0, & a_{31} = b_1, a_{32} = b_2 & \text{en mode d'extrapolation locale.}
\end{array} \right.$$

On peut facilement trouver une solution pour le système donné ci-haut et par conséquent il existe des méthodes FSAL à 3 stages pour les paires de formules d'ordre $(2,3)$.

Paires d'ordre $(3,4)$ du type I.

Il n'existe pas de méthode à trois stages pour les paires d'ordre $(3,4)$ du type I. Les équations de conditions pour ces paires sont les suivantes:

Ordre 3 en $\hat{y}$:

$$\begin{aligned} \hat{b}_1 + \hat{b}_2 + \hat{b}_3 &= \frac{1}{2}, \\ \hat{b}_2 c_2 + \hat{b}_3 c_3 &= \frac{1}{6}, \end{aligned}$$

Ordre 4 en y:

$$\begin{aligned} b_1 + b_2 + b_3 &= \frac{1}{2}, \\ b_2 c_2^k + b_3 c_3^k &= \frac{1}{(k+1)(k+2)}, \qquad k = 1, 2, \end{aligned}$$

Ordre 3 en $\hat{y}'$:

$$\hat{b}'_1 + \hat{b}'_2 + \hat{b}'_3 = 1, \tag{5.19}$$

$$\hat{b}'_2 c_2^k + \hat{b}'_3 c_3^k = \frac{1}{k+1}, \qquad k = 1, 2, \tag{5.20}$$

Ordre 4 en y':

$$b'_1 + b'_2 + b'_3 = 1, \tag{5.21}$$

$$b'_2 c_2^k + b'_3 c_3^k = \frac{1}{k+1}, \qquad k = 1, 2, 3, \tag{5.22}$$

$$b'_2 Q_{21} + b'_3 Q_{31} = 0.$$

On démontre le théorème suivant.

Théorème 5.1 *Il n'existe pas de méthode à trois stages pour les paires de formules d'ordre* $(3,4)$.

Démonstration. La démonstration fait appel uniquement aux équations de quadrature des formules des dérivées $\hat{y}'$ et y'. On suppose l'existence d'une méthode à trois stages et l'on considère la matrice et les vecteurs suivants:

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & c_2 & c_3 \\ 0 & c_2^2 & c_3^2 \end{bmatrix}, \quad \widehat{\mathbf{b}}' = \begin{bmatrix} \hat{b}'_1 \\ \hat{b}'_2 \\ \hat{b}'_3 \end{bmatrix}, \quad \mathbf{b}' = \begin{bmatrix} b'_1 \\ b'_2 \\ b'_3 \end{bmatrix}, \quad \mathbf{r} = \begin{bmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{3} \end{bmatrix}.$$

On écrit l'équation (5.19) et les deux équations (5.20), avec $k = 1, 2$, sous forme matricielle:

$$A\widehat{\mathbf{b}}' = \mathbf{r},$$

et, de la même façon, l'équation (5.21) et les deux premières équations de (5.22), avec $k = 1, 2$:

$$A\mathbf{b}' = \mathbf{r}.$$

On a donc

$$A\widehat{\mathbf{b}}' = A\mathbf{b}'.$$

Si A est régulière, la solution est unique: $\mathbf{b}' = \widehat{\mathbf{b}}'$, ce qui contredit l'hypothèse d'existence. Donc A est singulière.

La matrice A est singulière si et seulement si $c_2 = c_3$ ou l'un des c_2 ou c_3 est zéro. Dans les trois cas, les trois équations de (5.22), pour $k = 1, 2, 3$, peuvent s'écrire de la façon suivante:

$$x\alpha = \frac{1}{2}, \quad x\alpha^2 = \frac{1}{3}, \quad x\alpha^3 = \frac{1}{4}. \tag{5.23}$$

La solution des deux premières équations est

$$\alpha = \frac{2}{3}, \qquad x = \frac{3}{4}.$$

On remarque alors que la troisième équation n'est pas satisfaite. Par conséquent, même si A est singulière, la paire n'existe pas parce que la formule d'ordre supérieure ne peut être que d'ordre trois. □

Il existe des méthodes FSAL à quatre stages pour les paires de formules d'ordre $(3, 4)$. On présente une telle méthode au tableau 4.

Tableau 4. Une paire FSAL d'ordre $(3, 4)$ à quatre stages du type I

$\frac{1}{3}$	$\frac{1}{18}$			
$\frac{1}{2}$	$\frac{1}{8}$	0		
1	$\frac{1}{6}$	0	$\frac{1}{3}$	0
b	$\frac{1}{6}$	0	$\frac{1}{3}$	0
$\hat{b}$	$\frac{1}{2}$	-1	1	0
$\hat{b}'$	$\frac{1}{2}$	$-\frac{3}{2}$	2	0
b'	$\frac{1}{6}$	0	$\frac{2}{3}$	$\frac{1}{6}$

Paires d'ordre $(4, 5)$ du type I.

On montre qu'il n'existe pas de méthode d'ordre $(4, 5)$ à 4 stages pour les paires non FSAL ni à 5 stages pour les paires FSAL. On utilise la même démonstration pour les paires non FSAL et FSAL.

Théorème 5.2 *Il n'existe pas de méthode RKN d'ordre* $(4,5)$ *du type I pour les paires non FSAL à 4 stages ni pour les paires FSAL à 5 stages.*

Démonstration. La démonstration est par contradiction, en supposant l'existence d'une telle paire.

On considère les équations de non quadrature:

$$\left.\begin{array}{rcl} b'_2Q_{21} + b'_3Q_{31} + b'_4Q_{41} &=& 0, \\ b'_2c_2Q_{21} + b'_3c_3Q_{31} + b'_4c_4Q_{41} &=& 0, \\ \hat{b}'_2Q_{21} + \hat{b}'_3Q_{31} + \hat{b}'_4Q_{41} &=& 0. \end{array}\right\} \tag{5.24}$$

L'existence d'une paire implique que ces équations sont satisfaites. On remarque l'absence du terme en Q_{51} puisque $\hat{b}'_5 = b'_5 = 0$ pour les paires non FSAL (méthode à 4 stages), et $Q_{51} = 0$ pour les paires FSAL.

Si les trois équations (5.24) sont indépendantes, elles admettent l'unique solution nulle: $Q_{21} = Q_{31} = Q_{41} = 0$. Mais Q_{21} ne peut s'annuler parce que la méthode se réduirait à une méthode non FSAL d'ordre $(4,5)$ à 3 stages qui n'existe pas d'après le théorème 5.1.1 qu'on vient d'établir, ou à une paire FSAL d'ordre $(4,5)$ à 4 stages, que l'on considère dans la première partie du présent théorème.

Alors, les trois équations (5.24) sont dépendantes. Donc, il existe trois nombres α, β et γ, non tous nuls, tels que:

$$\left.\begin{array}{rcl} \alpha\hat{b}'_2 + \beta b'_2 + \gamma b'_2c_2 &=& 0, \\ \alpha\hat{b}'_3 + \beta b'_3 + \gamma b'_3c_3 &=& 0, \\ \alpha\hat{b}'_4 + \beta b'_4 + \gamma b'_4c_4 &=& 0. \end{array}\right\} \tag{5.25}$$

Pour chacune des trois valeurs de k, $k = 1,2,3$, après avoir multiplié les trois équations de (5.25) par c_2^k, c_3^k et c_4^k, on additionne les équations et on utilise les équations de quadrature (5.6) et (5.10); on obtient alors le système linéaire suivant:

$$\left.\begin{array}{rcl} \left(\frac{1}{2} - \hat{b}'_5\right)\alpha + \left(\frac{1}{2} - b'_5\right)\beta + \left(\frac{1}{3} - b'_5\right)\gamma &=& 0, \\ \left(\frac{1}{3} - \hat{b}'_5\right)\alpha + \left(\frac{1}{3} - b'_5\right)\beta + \left(\frac{1}{4} - b'_5\right)\gamma &=& 0, \\ \left(\frac{1}{4} - \hat{b}'_5\right)\alpha + \left(\frac{1}{4} - b'_5\right)\beta + \left(\frac{1}{5} - b'_5\right)\gamma &=& 0, \end{array}\right\} \tag{5.26}$$

qu'on récrit de la façon suivante:

$$\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ 1 & \frac{1}{3} & \frac{1}{4} \\ 1 & \frac{1}{4} & \frac{1}{5} \end{bmatrix} \begin{bmatrix} -\hat{b}'_5\alpha - b'_5(\beta+\gamma) \\ \alpha+\beta \\ \gamma \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{5.27}$$

Puisque la matrice des coefficients est régulière, le système admet l'unique solution nulle:

$$-\hat{b}'_5\alpha - b'_5(\beta+\gamma) = 0, \qquad \alpha+\beta = 0, \qquad \gamma = 0.$$

On a donc $\beta = -\alpha$, ce qui, combiné avec le premier élément du vecteur nul, donne $b'_5 = \hat{b}'_5$.

Si on substitue ces valeurs de α, β et γ dans le système (5.25), on voit que $b'_i = \hat{b}'_i$, $i = 2, 3, 4$. Ce résultat, combiné avec les deux premières équations de quadrature de (5.6) et (5.10), implique que $b'_1 = \hat{b}'_1$, et par conséquent, la paire n'existe pas. □

Dormand, El-Mikkawy et Prince ont construit en [4] une méthode à 6 stages pour les paires FSAL d'ordre $(4, 5)$. On présente, au tableau 5, une nouvelle méthode à 5 stages pour les paires non FSAL d'ordre $(4, 5)$.

Tableau 5. Une paire non FSAL d'ordre $(4, 5)$ à cinq stages du type I

$\frac{1}{6}$	$\frac{1}{72}$				
$\frac{3}{10}$	$\frac{459}{23\,000}$	$\frac{72}{2\,875}$			
$\frac{2}{5}$	$\frac{41}{2\,875}$	$\frac{189}{2\,875}$	0		
$\frac{2}{3}$	$\frac{1\,145}{29\,187}$	$\frac{203}{6\,486}$	$\frac{385}{2\,538}$	0	
$\hat{b}$	$-\frac{19}{36}$	$\frac{155}{84}$	$-\frac{35}{36}$	$-\frac{5}{28}$	$\frac{1}{3}$
b	$\frac{17}{18}$	$-\frac{125}{28}$	$\frac{3\,325}{396}$	$-\frac{275}{56}$	$\frac{47}{88}$
$\hat{b}'$	$-\frac{19}{36}$	$\frac{31}{14}$	$-\frac{25}{18}$	$-\frac{25}{84}$	1
b'	$\frac{17}{18}$	$-\frac{75}{14}$	$\frac{2\,375}{198}$	$-\frac{1\,375}{168}$	$\frac{141}{88}$

Paires d'ordre $(5, 6)$ du type I.

On démontre l'inexistence de méthode à cinq stages par contradiction.

Théorème 5.3 *Il n'existe pas de méthode à 5 stages pour les paires RKN d'ordre $(5, 6)$ du type I.*

Démonstration. On suppose qu'une paire existe et on démontre que les équations de conditions suivantes sont incompatibles pour les paires à 5 stages.

Ordre 5 en $\hat{y}'$:

$$\hat{b}'_1 + \hat{b}'_2 + \hat{b}'_3 + \hat{b}'_4 + \hat{b}'_5 = \frac{1}{2}, \tag{5.28}$$

$$\hat{b}'_2 c_2^k + \hat{b}'_3 c_3^k + \hat{b}'_4 c_4^k + \hat{b}'_5 c_5^k = \frac{1}{k+1}, \quad k = 1, 2, 3, 4, \tag{5.29}$$

$$\begin{aligned} \hat{b}'_2 c_2^k Q_{21} + \hat{b}'_3 c_3^k Q_{31} + \hat{b}'_4 c_4^k Q_{41} + \hat{b}'_5 c_5^k Q_{51} &= 0, \quad k = 0, 1, \\ \hat{b}'_2 Q_{22} + \hat{b}'_3 Q_{32} + \hat{b}'_4 Q_{42} + \hat{b}'_5 Q_{52} &= 0, \end{aligned} \tag{5.30}$$

Ordre 6 en y':

$$
\begin{aligned}
b'_1 + b'_2 + b'_3 + b'_4 + b'_5 &= 1, && (5.31)\\
b'_2 c_2^k + b'_3 c_3^k + b'_4 c_4^k + b'_5 c_5^k &= \frac{1}{k+1}, \quad k = 1, \ldots, 5, && (5.32)\\
b'_2 c_2^k Q_{21} + b'_3 c_3^k Q_{31} + b'_4 c_4^k Q_{31} + b'_5 c_5^k Q_{51} &= 0, \quad k = 0, 1, 2, && (5.33)\\
b'_2 c_2^k Q_{22} + b'_3 c_3^k Q_{32} + b'_4 c_4^k Q_{32} + b'_5 c_5^k Q_{52} &= 0, \quad k = 0, 1,\\
b'_2 Q_{22} + b'_3 Q_{32} + b'_4 Q_{42} + b'_5 Q_{52} &= 0,\\
b'_3 a_{32} Q_{21} + b'_4 \left(a_{42} Q_{21} + a_{43} Q_{31}\right)&\\
+ b'_5 \left(a_{52} Q_{21} + a_{53} Q_{31} + a_{54} Q_{41}\right) &= 0,
\end{aligned}
$$

Ordre 5 en $\hat{y}$:

$$
\begin{aligned}
\hat{b}_1 + \hat{b}_2 + \hat{b}_3 + \hat{b}_4 + \hat{b}_5 &= \frac{1}{2},\\
\hat{b}_2 c_2^k + \hat{b}_3 c_3^k + \hat{b}_4 c_4^k + \hat{b}_5 c_5^k &= \frac{1}{(k+1)(k+2)}, \quad k = 1, 2, 3,\\
\hat{b}_2 Q_{21} + \hat{b}_3 Q_{31} + \hat{b}_4 Q_{41} + \hat{b}_5 Q_{51} &= 0,
\end{aligned}
$$

Ordre 6 en y:

$$
\begin{aligned}
b_1 + b_2 + b_3 + b_4 + b_5 &= \frac{1}{2},\\
b_2 c_2^k + b_3 c_3^k + b_4 c_4^k + b_5 c_5^k &= \frac{1}{k+1}, \quad k = 1, 2, 3, 4,\\
b_2 c_2^k Q_{21} + b_3 c_3^k Q_{31} + b_4 c_4^k Q_{41} + b_5 c_5^k Q_{51} &= 0, \quad k = 0, 1,\\
b_2 Q_{22} + b_3 Q_{32} + b_4 Q_{42} + b_5 Q_{52} &= 0.
\end{aligned}
$$

On procède comme au théorème 4.1.2. Quatre équations quelconques parmi les cinq équations de non quadrature (5.30), $k = 0, 1$, et (5.33), $k = 0, 1, 2$, doivent être dépendantes, sinon $Q_{21} = 0$ et par suite $c_2 = 0$. Ceci réduirait le nombre de stages de la méthode à 4; donc celle-ci n'existerait pas par le théorème 4.1.2. Alors il existe quatre scalaires, α, β, γ et δ, non tous nuls, tels que

$$
\alpha \hat{b}'_i + \beta b'_i + \gamma b'_i c_i + \delta b'_i c_i^2 = 0, \quad i = 2, \ldots, 5. \qquad (5.34)
$$

On multiplie la $i$ème équation par c_i^k, $k = 1, 2, 3$ et on somme sur les i. Si on utilise les équations de quadrature (5.29) et (5.32), on obtient le système suivant:

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \end{bmatrix} \begin{bmatrix} \alpha + \beta \\ \gamma \\ \delta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

La matrice des coefficients est régulière. Donc le système admet l'unique solution

$$(\alpha + \beta, \gamma, \delta) = (0, 0, 0).$$

Il s'ensuit que $\alpha = -\beta \neq 0$, sinon la paire n'existe pas.

On remplace α, β, γ et δ dans (5.34) par leurs valeurs respectives, et on obtient $b'_i = \hat{b}'_i$, $i = 2, \ldots, 5$. Enfin, de (5.28) et de (5.31) on peut tirer que $b'_1 = \hat{b}'_1$. Donc la paire n'existe pas. □

On complète l'étude des paires d'ordre $(5, 6)$ en donnant au tableau 6 une paire FSAL à six stages. Un élément du tableau de la forme (a, b) représente le nombre algébrique $a + b\sqrt{5}$.

Tableau 6. Une paire FSAL d'ordre $(5, 6)$ à six stages du type I

$\frac{1}{2}$	$\frac{1}{8}$					
1	$\frac{1}{6}$	$\frac{1}{3}$				
$(\frac{1}{2}, -\frac{1}{10})$	$(\frac{11}{150}, -\frac{1}{50})$	$(\frac{13}{150}, -\frac{1}{30})$	$(-\frac{1}{100}, \frac{1}{300})$			
$(\frac{1}{2}, \frac{1}{10})$	$(\frac{13}{300}, \frac{1}{100})$	$(\frac{7}{150}, -\frac{1}{150})$	$(\frac{1}{100}, -\frac{1}{300})$	$(\frac{1}{20}, \frac{1}{20})$		
1	$\frac{1}{12}$	0	0	$(\frac{5}{24}, \frac{1}{24})$	$(\frac{5}{24}, -\frac{1}{24})$	
$\hat{b}$	$\frac{1}{12}$	0	-2	$(\frac{5}{24}, \frac{1}{24})$	$(\frac{5}{24}, -\frac{1}{24})$	2
b	$\frac{1}{12}$	0	0	$(\frac{5}{24}, \frac{1}{24})$	$(\frac{5}{24}, -\frac{1}{24})$	0
$\hat{b}'$	$\frac{1}{12}$	0	$-\frac{11}{12}$	$\frac{5}{12}$	$\frac{5}{12}$	1
b'	$\frac{1}{12}$	0	$(-\frac{1}{6}, \frac{1}{20})$	$\frac{5}{12}$	$\frac{5}{12}$	$(\frac{1}{4}, -\frac{1}{20})$

5.6.2 Paires du type II

Comme on l'a déjà mentionné, les paires du type II contrôlent l'erreur locale en $\hat{y}$ et on avance le pas avec la méthode d'ordre inférieur.

Paires d'ordre $(1, 2)$ du type II.

On peut montrer facilement qu'il existe des méthodes à un seul stage pour les paires non FSAL d'ordre $(1, 2)$, mais qu'il n'en existe pas pour les paires FSAL.

Les équations de conditions pour les méthodes à un stage, $s = 1$, sont:

Ordre 1 en $\hat{y}'$: $\hat{b}'_1 = 1$,
Ordre 1 en $\hat{y}$: aucune condition d'ordre,
Ordre 2 en y: $b_1 = \frac{1}{2}$.

On a donc

$$\hat{b}'_1 = 1, \qquad b_1 = \frac{1}{2}, \qquad \hat{b}_1 \text{ arbitraire.}$$

Dans le cas FSAL, on devrait avoir $c_1 = 1$, ce qui n'est pas possible puisque $c_1 = 0$. On déduit donc qu'il n'y a pas de paire FSAL d'ordre $(1, 2)$.

Il existe des méthodes à deux stages pour les paires FSAL d'ordre $(1, 2)$ du type II puisqu'il en existe du type I.

Paires d'ordre $(2, 3)$ du type II.

Il n'y a pas de méthode à un stage pour les paires non FSAL d'ordre $(2, 3)$.

(a) *Inexistence de méthode non FSAL à un stage.*
Les équations de conditions pour les méthodes à un stage, $s = 1$, sont:

Ordre 2 en $\hat{y}'$: $\hat{b}'_1 = 1$,
$\hat{b}'_1 c_1 = \frac{1}{2}$,
Ordre 2 en $\hat{y}$: $\hat{b}_1 = \frac{1}{2}$,
Ordre 3 en y: $b_1 = \frac{1}{2}$,
$b_1 c_1 = \frac{1}{6}$.

Il est évident que les équations:

$$b'_1 c_1 = \frac{1}{2}, \qquad b_1 c_1 = \frac{1}{2},$$

ne peuvent pas être satisfaites parce que $c_1 = 0$.

(b) *Existence de méthodes FSAL à deux stages.*
Il existe des méthodes à deux stages pour les paires FSAL d'ordre $(2, 3)$. On en présente une au tableau 7.

Tableau 7. Une paire non FSAL d'ordre $(2, 3)$ à 2 stages du type II

0		
1	$\frac{1}{2}$	
b	$\frac{1}{3}$	$\frac{1}{6}$
$\hat{b}$	$\frac{1}{2}$	0
$\hat{b}'$	$\frac{1}{2}$	$\frac{1}{2}$

Paires d'ordre $(3,4)$ **du type II.**

On montre qu'il n'existe pas de méthode à 2 stages pour les paires non FSAL d'ordre $(3,4)$.

(a) *Inexistence de méthode à deux stages.*
On énumère les équations de conditions pour les paires à deux stages:

Ordre 3 en $\hat{y}'$:

$$\begin{aligned} \hat{b}'_1 + \hat{b}'_2 &= 1, \\ \hat{b}'_2 c_2^k &= \frac{1}{k+1} \qquad k = 1,2, \end{aligned} \tag{5.35}$$

Ordre 3 en $\hat{y}$:

$$\begin{aligned} \hat{b}_1 + \hat{b}_2 &= \frac{1}{2}, \\ \hat{b}_2 c_2 &= \frac{1}{6}, \end{aligned}$$

Ordre 4 en y:

$$\begin{aligned} b_1 + b_2 &= \frac{1}{2}, \\ b_2 c_2^k &= \frac{1}{(k+1)(k+2)}, \qquad k = 1,2. \end{aligned} \tag{5.36}$$

Les deux équations en (5.35) impliquent que $c_2 = 2/3$; d'autre part, les deux équations en (5.36) impliquent que $c_2 = 1/2$. Elles sont donc incompatibles et il n'y a pas de méthode à 2 stages pour les paires non FSAL $(3,4)$.

(b) *Existence de méthodes FSAL à trois stages.*
On présente, au tableau 8, une méthode à 3 stages pour les paires FSAL d'ordre $(3,4)$.

Tableau 8. Une paire FSAL d'ordre $(3,4)$ à trois stages du type II

$\frac{2}{3}$	$\frac{2}{9}$		
1	$\frac{1}{4}$	$\frac{1}{4}$	0
$\hat{b}$	$\frac{1}{4}$	$\frac{1}{4}$	0
b	$\frac{5}{24}$	$\frac{3}{8}$	$-\frac{1}{12}$
$\hat{b}'$	$\frac{1}{4}$	$\frac{3}{4}$	0

Paires d'ordre $(4,5)$ **du type II.**

On montre qu'il n'existe pas de méthode du type II à 3 stages pour les paires non FSAL d'ordre $(4,5)$ ni à 4 stages pour les paires FSAL d'ordre $(4,5)$. On présente deux nouvelles méthodes, une non FSAL à 4 stages et l'autre FSAL à 5 stages.

(a) *Inexistence de méthode non FSAL à trois stages.*

On énumère les équations de conditions pour les méthodes à 3 stages, $s = 3$, pour les paires non FSAL:

Ordre 4 en $\hat{y}'$:

$$\begin{aligned} \hat{b}'_1 + \hat{b}'_2 + \hat{b}'_3 &= 1, \\ \hat{b}'_2 c_2^k + \hat{b}'_3 c_3^k &= \frac{1}{k+1} \qquad k = 1,2,3, \\ \hat{b}'_2 Q_{21} + \hat{b}'_3 Q_{31} &= 0, \end{aligned} \tag{5.37}$$

Ordre 4 en $\hat{y}$:

$$\hat{b}_1 + \hat{b}_2 + \hat{b}_3 = \frac{1}{2}, \tag{5.38}$$

$$\hat{b}_2 c_2^k + \hat{b}_3 c_3^k = \frac{1}{(k+1)(k+2)}, \qquad k = 1,2, \tag{5.39}$$

Ordre 5 en y:

$$b_1 + b_2 + b_3 = \frac{1}{2}, \tag{5.40}$$

$$b_2 c_2^k + b_3 c_3^k = \frac{1}{(k+1)(k+2)}, \qquad k = 1,2,3, \tag{5.41}$$

$$b_2 Q_{21} + b_3 Q_{31} = 0.$$

On montre que les conditions de quadratures pour y et $\hat{y}$ sont incompatibles. Soit la matrice et les vecteurs

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & c_2 & c_3 \\ 0 & c_2^2 & c_3^2 \end{bmatrix}, \quad \widehat{\mathbf{b}} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}, \quad \mathbf{r} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{6} \\ \frac{1}{12} \end{bmatrix}.$$

Alors les trois équations (5.38) et (5.39), $k = 1,2$, d'une part et les trois équations (5.40) et (5.41), $k = 1,2$, d'autre part, s'écrivent respectivement de la façon suivante:

$$A\widehat{\mathbf{b}} = \mathbf{r}, \qquad A\mathbf{b} = \mathbf{r}.$$

Donc

$$A\widehat{\mathbf{b}} = A\mathbf{b}. \tag{5.42}$$

Si A est régulière, le système (5.42) admet l'unique solution $\mathbf{b} = \widehat{\mathbf{b}}$, ce qui contredit l'existence d'une paire. Il faut donc que A soit singulière.

Dans ce cas, $c_2 = c_3$ ou $c_2 c_3 = 0$. Dans ces trois cas, les trois équations (5.37) se réduisent à la forme suivante:

$$x\alpha = \frac{1}{2}, \quad x\alpha^2 = \frac{1}{3}, \quad x\alpha^3 = \frac{1}{4}.$$

On voit immédiatement que ce système est incompatible.

On conclut à l'inexistence de méthode à 3 stages pour les paires non FSAL d'ordre $(4,5)$.

(b) *Inexistence de méthode FSAL à quatre stages.*

Les équations de conditions pour les méthodes FSAL à 4 stages sont les suivantes:

Ordre 4 en $\hat{y}'$:

$$\begin{aligned} \hat{b}'_1 + \hat{b}'_2 + \hat{b}'_3 + \hat{b}'_4 &= 1, \\ \hat{b}'_2 c_2^k + \hat{b}'_3 c_3^k + \hat{b}'_4 c_4^k &= \frac{1}{k+1}, \qquad k = 1,2,3, \\ \hat{b}'_2 Q_{21} + \hat{b}'_3 Q_{31} &= 0, \end{aligned} \tag{5.43}$$

Ordre 4 en $\hat{y}$:

$$\begin{aligned} \hat{b}_1 + \hat{b}_2 + \hat{b}_3 + \hat{b}_4 &= \frac{1}{2}, \\ \hat{b}_2 c_2^k + \hat{b}_3 c_3^k + \hat{b}_4 c_4^k &= \frac{1}{(k+1)(k+2)}, \qquad k = 1,2, \end{aligned}$$

Ordre 5 en y:

$$\begin{aligned} b_1 + b_2 + b_3 + b_4 &= \frac{1}{2}, \\ b_2 c_2^k + b_3 c_3^k + b_4 c_4^k &= \frac{1}{(k+1)(k+2)}, \qquad k = 1,2,3, \\ b_2 Q_{21} + b_3 Q_{31} &= 0. \end{aligned} \tag{5.44}$$

On rappelle que $Q_{41} = 0$ puisqu'on est dans le cas FSAL. Les équations (5.43) et (5.44) doivent être dépendantes, sinon on aurait $Q_{21} = 0$, une impossibilité. Alors, il existe deux nombres, α et β, non tous nuls tels que

$$\alpha \hat{b}'_i + \beta b_i = 0, \qquad i = 2,3.$$

On multiplie cette équation par c_i^k, $k = 1, 2, 3$, et on somme sur les i. Puis on utilise les équations de quadrature en $\hat{y}'$ et y pour obtenir le système suivant:

$$\begin{aligned}
\left(\frac{1}{2} - \hat{b}'_4\right)\alpha + \left(\frac{1}{6} - b_4\right)\beta &= 0, \\
\left(\frac{1}{3} - \hat{b}'_4\right)\alpha + \left(\frac{1}{12} - b_4\right)\beta &= 0, \\
\left(\frac{1}{4} - \hat{b}'_4\right)\alpha + \left(\frac{1}{20} - b_4\right)\beta &= 0.
\end{aligned}$$

Si on soustrait la deuxième équation de la première et la troisième de la deuxième, on aura

$$\frac{1}{2}\alpha + \frac{1}{12}\beta = 0, \qquad \frac{1}{12}\alpha + \frac{1}{30}\beta = 0.$$

Alors $\alpha = \beta = 0$ et par conséquent, la paire n'existe pas.

(c) *Existence de méthodes non FSAL à quatre stages.*

On présente au tableau 9 une méthode à 4 stages pour les paires non FSAL d'ordre $(4, 5)$.

Tableau 9. Une paire non FSAL d'ordre $(4, 5)$ à quatre stages du type II

$\frac{1}{12}$	$\frac{1}{288}$			
1	$-\frac{445}{298}$	$\frac{297}{149}$		
$\frac{4}{5}$	$-\frac{13\,132}{18\,625}$	$\frac{19\,092}{18\,625}$	0	
$\hat{b}$	$-\frac{2}{3}$	$\frac{12}{11}$	$\frac{5}{66}$	0
b	$-\frac{13}{48}$	$\frac{288}{473}$	$-\frac{3}{44}$	$\frac{475}{2\,064}$
$\hat{b}'$	$-\frac{11}{24}$	$\frac{432}{473}$	$-\frac{2}{33}$	$\frac{625}{1\,032}$

(d) *Existence de méthodes FSAL à cinq stages.*

Enfin, on présente au tableau 10 une méthode à 5 stages pour les paires FSAL d'ordre $(4, 5)$.

Tableau 10. Une paire FSAL d'ordre $(4, 5)$ à cinq stages du type II

$\frac{1}{12}$	$\frac{1}{288}$				
1	$-\frac{6\,583}{4\,390}$	$\frac{4\,389}{2\,195}$			
$\frac{9}{10}$	$-\frac{462\,537}{439\,000}$	$\frac{160\,083}{109\,750}$	0		
1	$-\frac{2}{3}$	$\frac{12}{11}$	$\frac{5}{66}$	0	
$\hat{b}$	$-\frac{2}{3}$	$\frac{12}{11}$	$\frac{5}{66}$	0	0
b	$\frac{17}{54}$	$\frac{360}{539}$	$-\frac{40}{33}$	$\frac{475}{1\,323}$	1
$\hat{b}'$	$-\frac{31}{54}$	$\frac{576}{539}$	$-\frac{95}{66}$	$\frac{1\,250}{1\,323}$	1

Paires d'ordre $(5,6)$ du type II.

Il n'existe pas de méthode à 5 stages pour les paires non FSAL d'ordre $(5,6)$. La démonstration qui est par contradiction suppose qu'une paire existe.

(a) *Inexistence de méthode non FSAL à cinq stages.*

Les équations de conditions pour les méthodes à 5 stages du type II sont les suivantes:

Ordre 5 en $\hat{y}'$:

$$\begin{aligned}
\hat{b}'_1 + \hat{b}'_2 + \hat{b}'_3 + \hat{b}'_4 + \hat{b}'_5 &= 1, && (5.45)\\
\hat{b}'_2 c_2^k + \hat{b}'_3 c_3^k + \hat{b}'_4 c_4^k + \hat{b}'_5 c_5^k &= \frac{1}{k+1}, \qquad k = 1,2,3,4, && (5.46)\\
\hat{b}'_2 c_2^k Q_{21} + \hat{b}'_3 c_3^k Q_{31} + \hat{b}'_4 c_4^k Q_{41} + \hat{b}'_5 c_5^k Q_{51} &= 0, \qquad k = 0,1, && (5.47)\\
\hat{b}'_2 Q_{22} + \hat{b}'_3 Q_{32} + \hat{b}'_4 Q_{42} + \hat{b}'_5 Q_{52} &= 0,
\end{aligned}$$

Ordre 5 en $\hat{y}$:

$$\begin{aligned}
\hat{b}_1 + \hat{b}_2 + \hat{b}_3 + \hat{b}_4 + \hat{b}_5 &= \frac{1}{2}, && (5.48)\\
\hat{b}_2 c_2^k + \hat{b}_3 c_3^k + \hat{b}_4 c_4^k + \hat{b}_5 c_5^k &= \frac{1}{(k+1)(k+2)}, \qquad k = 1,2,3, && (5.49)\\
\hat{b}_2 Q_{21} + \hat{b}_3 Q_{31} + \hat{b}_4 Q_{41} + \hat{b}_5 Q_{51} &= 0, && (5.50)
\end{aligned}$$

Ordre 6 en y:

$$\begin{aligned}
b_1 + b_2 + b_3 + b_4 + b_5 &= \frac{1}{2}, && (5.51)\\
b_2 c_2^k + b_3 c_3^k + b_4 c_4^k + b_5 c_5^k &= \frac{1}{(k+1)(k+2)}, \quad k = 1,2,3, && (5.52)\\
b_2 c_2^k Q_{21} + b_3 c_3^k Q_{31} + b_4 c_4^k Q_{41} + b_5 c_5^k Q_{51} &= 0, \qquad k = 0,1, && (5.53)\\
b_2 Q_{22} + b_3 Q_{32} + b_4 Q_{42} + b_5 Q_{52} &= 0,
\end{aligned}$$

On considère les conditions de non quadrature en Q_{i1}, c'est-à-dire les cinq équations (5.47), (5.50) et (5.53). Quatre quelconque de ces équations linéaires en les Q_{i1} doivent être dépendantes, sinon on aura $Q_{21} = 0$ et, par une seconde application de ce procédé, le nombre de stages de la méthode se réduirait à 3; alors on aurait l'inexistence par le cas précédent.

En premier lieu, on considère les quatre équations (5.47) et (5.53) respectivement pour $k = 0, 1$. Il existe alors quatre nombres, α, β, γ et δ, non tous nuls tels que

$$\alpha \hat{b}_i' + \beta \hat{b}_i' c_i + \gamma b_i + \delta b_i c_i = 0, \qquad i = 2, \ldots, 5. \tag{5.54}$$

On multiplie ces équations par c_i^l, $l = 1, 2, 3$, on somme sur les i et on utilise les équations de quadrature (5.46) et (5.52); on obtient alors le système linéaire:

$$\begin{aligned} \frac{1}{2}\alpha + \frac{1}{3}\beta + \frac{1}{6}\gamma + \frac{1}{12}\delta &= 0, \\ \frac{1}{3}\alpha + \frac{1}{4}\beta + \frac{1}{12}\gamma + \frac{1}{20}\delta &= 0, \\ \frac{1}{4}\alpha + \frac{1}{5}\beta + \frac{1}{20}\gamma + \frac{1}{30}\delta &= 0. \end{aligned}$$

La solution générale de ce système est:

$$\alpha = a, \quad \beta = -a, \quad \gamma = -a, \quad \delta = 0,$$

où $\alpha \neq 0$, sinon la méthode n'existerait pas. Si on substitue ces valeurs de $\alpha, \beta, \gamma, \delta$ dans (5.54) on a:

$$b_i = \hat{b}_i' - \hat{b}_i' c_i, \qquad i = 2, \ldots, 5. \tag{5.55}$$

De même, si l'on considère les quatre conditions de non quadrature suivantes: (5.47) avec $k = 0, 1$, (5.50) et (5.53) avec $k = 1$, on a, comme avant:

$$\alpha \hat{b}_i' + \beta \hat{b}_i' c_i + \gamma \hat{b}_i + \delta b_i c_i = 0, \qquad i = 2, \ldots, 5. \tag{5.56}$$

On multiplie ces équations par c_i^l, $l = 1, 2, 3$, on somme sur les i et on utilise les équations de quadrature (5.46), (5.49) et (5.52); on obtient alors le système linéaire suivant:

$$\begin{aligned} \frac{1}{2}\alpha + \frac{1}{3}\beta + \frac{1}{6}\gamma + \frac{1}{12}\delta &= 0, \\ \frac{1}{3}\alpha + \frac{1}{4}\beta + \frac{1}{12}\gamma + \frac{1}{20}\delta &= 0, \\ \frac{1}{4}\alpha + \frac{1}{5}\beta + \frac{1}{20}\gamma + \frac{1}{30}\delta &= 0. \end{aligned}$$

La solution génrérale de ce système est:

$$\alpha = a, \quad \beta = -a, \quad \gamma = -a, \quad \delta = 0,$$

où $\alpha \neq 0$, sinon la méthode n'existerait pas. Si on substitue les valeurs de $\alpha, \beta, \gamma, \delta$ obtenues dans (5.56) on a:

$$\hat{b}_i = \hat{b}_i' - \hat{b}_i' c_i, \qquad i = 2, \ldots, 5. \tag{5.57}$$

Il est évident que (5.55) et (5.57) donnent:

$$b_i = \hat{b}_i, \qquad i = 2, \ldots, 5,$$

et par les équations de quadrature (5.48) et (5.51), on a $b_1 = \hat{b}_1$. Par conséquent, la paire n'existe pas.

(b) *Existence de méthodes FSAL à six stages.*

Il existe une méthode à 6 stages pour les paires FSAL d'ordre $(5, 6)$ du type II puisqu'une telle paire existe pour le type I.

5.6.3 Paires du type III

Les paires du type III contrôlent l'erreur locale en y et on avance le pas avec la méthode d'ordre supérieur.

Paires d'ordre $(1, 2)$ du type III.

On peut facilement montrer qu'il n'existe pas de méthode à un seul stage pour les paires non FSAL d'ordre $(1, 2)$.

Les équations de conditions pour les méthodes à un stage, c'est-à-dire pour $s = 1$, sont:

Ordre 2 en y':	$b'_1 = 1$,
	$b'_1 c_1 = \frac{1}{2}$,
Ordre 1 en $\hat{y}$:	aucune condition d'ordre,
Ordre 2 en y:	$b_1 = \frac{1}{2}$.

Il est évident que la deuxième équation de condition en y' ne peut pas être satisfaite puisque $c_1 = 0$. On déduit donc qu'il n'y a pas de paire non FSAL d'ordre $(1, 2)$.

Il existe des méthodes à deux stages pour les paires FSAL d'ordre $(1, 2)$ du type III puisqu'il en existe du type I.

Paires d'ordre $(2, 3)$ du type III.

Il n'y a pas de méthode à deux stages pour les paires FSAL d'ordre $(2, 3)$, mais il y a des paires non FSAL.

(a) *Inexistence de méthode FSAL à deux stages.*

Les équations de conditions pour les méthodes à deux stages, $s = 2$, sont:

$$
\begin{array}{ll}
\text{Ordre 3 en } y': & b'_1 + b'_2 = 1, \\
& b'_2 c_2^k = \frac{1}{k+1} \quad k = 1, 2, \\
\text{Ordre 2 en } \hat{y}: & \hat{b}_1 + \hat{b}_2 = \frac{1}{2}, \\
\text{Ordre 3 en } y: & b_1 + b_2 = \frac{1}{2}, \\
& b_2 c_2 = \frac{1}{6}.
\end{array}
$$

Dans le cas FSAL on a $c_2 = 1$, $b_2 = 0$. Alors on voit clairement que la deuxième équation de quadrature en y ne peut pas être satisfaite.

(b) *Existence de méthodes non FSAL à deux stages.*

Il y a des méthodes à deux stages pour les paires non FSAL d'ordre $(2, 3)$. On en présente une au tableau 11.

Tableau 11. *Une paire non FSAL d'ordre* $(2, 3)$ *à 2 stages du type III*

$$
\begin{array}{c|cc}
0 & & \\
\frac{2}{3} & 1 & \\
\hline
b & \frac{1}{4} & \frac{1}{4} \\
\hat{b} & 0 & \frac{1}{2} \\
b' & \frac{1}{4} & \frac{3}{4}
\end{array}
$$

(c) *Existence de méthodes FSAL à trois stages.*

Il existe des méthodes à trois stages pour les paires FSAL d'ordre $(2, 3)$ du type III puisqu'il en existe du type I.

Paires d'ordre $(3, 4)$ du type III.

Il n'y a pas de méthode à deux stages non FSAL ni à trois stages FSAL, mais il y a des paires non FSAL à trois stages.

(a) *Inexistence de méthode non FSAL à deux stages.*

On montre qu'il n'existe pas de méthode à 2 stages pour les paires non FSAL d'ordre $(3, 4)$. On donne ci-dessous les équations de conditions pour les paires à deux stages.

Ordre 4 en y':

$$
\begin{aligned}
b'_1 + b'_2 &= 1, && (5.58) \\
b'_2 c_2^k &= \frac{1}{k+1}, \qquad k = 1, 2, 3, && (5.59) \\
b'_2 Q_{21} &= 0,
\end{aligned}
$$

Ordre 3 en $\hat{y}$:

$$\hat{b}_1 + \hat{b}_2 = \frac{1}{2}, \tag{5.60}$$

$$\hat{b}_2 c_2 = \frac{1}{6}, \tag{5.61}$$

Ordre 4 en y:

$$b_1 + b_2 = \frac{1}{2}, \tag{5.62}$$

$$b_2 c_2^k = \frac{1}{(k+1)(k+2)}, \qquad k = 1, 2. \tag{5.63}$$

Il est évident qu'on ne peut pas satisfaire la deuxième équation de quadrature en y' pour la même raison qu'en (5.23). Donc il n'y a pas de méthode à 2 stages pour les paires non FSAL d'ordre $(3, 4)$.

(b) *Inexistence de méthode FSAL à trois stages.*

Il n'existe pas de méthode à 3 stages pour les paires FSAL d'ordre $(3, 4)$. On donne ci-dessous les équations de conditions pour les paires à trois stages.

Ordre 4 en y':

$$b'_1 + b'_2 + b'_3 = 1,$$

$$b'_2 c_2^k + b'_3 c_3^k = \frac{1}{k+1}, \qquad k = 1, 2, 3, \tag{5.64}$$

$$b'_2 Q_{21} + b'_3 Q_{31} = 0, \tag{5.65}$$

Ordre 3 en $\hat{y}$:

$$\hat{b}_1 + \hat{b}_2 + \hat{b}_3 = \frac{1}{2},$$

$$\hat{b}_2 c_2 + \hat{b}_3 c_3 = \frac{1}{6},$$

Ordre 4 en y:

$$b_1 + b_2 + b_3 = \frac{1}{2},$$

$$b_2 c_2^k + b_3 c_3^k = \frac{1}{(k+1)(k+2)}, \qquad k = 1, 2. \tag{5.66}$$

Dans le cas FSAL, on a $c_3 = 1$, $b_3 = 0$ et $Q_{31} = 0$ puisqu'on est dans la mode d'extrapolation locale. Les deux équations en (5.66) donnent $c_2 = \frac{1}{2}$, et

avec ces valeurs de c_2 et c_3, les trois équations en (5.64) et l'équation (5.65) sont incompatibles puisque Q_{21} ne peut pas être nul.

(c) *Existence de méthode non FSAL à trois stages.*

On présente, au tableau 12, une méthode à 3 stages pour les paires non FSAL d'ordre $(3,4)$ du type III.

Tableau 12. Une paire non FSAL d'ordre $(3,4)$ à trois stages du type III

$\frac{1}{2}$	$\frac{1}{8}$		
1	0	$\frac{1}{2}$	
$\hat{b}$	$\frac{13}{6}$	$-\frac{11}{3}$	2
b	$\frac{1}{6}$	$\frac{1}{3}$	0
b'	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

(d) *Existence de méthode FSAL à quatre stages.*

Il existe des méthodes à quatre stages pour les paires FSAL d'ordre $(3,4)$ du type III puisqu'il en existe du type I.

Paires d'ordre $(4,5)$ du type III.

On montre qu'il n'existe pas de méthode du type III à 3 stages pour les paires d'ordre $(4,5)$ non FSAL ni à 4 stages pour les paires FSAL d'ordre $(4,5)$.

(a) *Inexistence de méthode non FSAL à trois stages.*

Les équations de conditions pour les méthodes à 3 stages, $s = 3$, pour les paires non FSAL sont les suivantes:

Ordre 5 en y':

$$\begin{aligned} b'_1 + b'_2 + b'_3 &= 1, \\ b'_2 c_2^k + b'_3 c_3^k &= \frac{1}{k+1} \qquad k = 1,2,3,4, \\ b'_2 c_2^k Q_{21} + b'_3 c_3^k Q_{31} &= 0 \qquad k = 0,1, \\ b'_2 Q_{22} + b'_3 Q_{32} &= 0, \end{aligned} \tag{5.67}$$

Ordre 4 en $\hat{y}$:

$$\hat{b}_1 + \hat{b}_2 + \hat{b}_3 = \frac{1}{2}, \tag{5.68}$$

$$\hat{b}_2 c_2^k + \hat{b}_3 c_3^k = \frac{1}{(k+1)(k+2)}, \qquad k = 1,2, \tag{5.69}$$

Ordre 5 en y:

$$b_1 + b_2 + b_3 = \frac{1}{2}, \tag{5.70}$$

$$\begin{aligned} b_2c_2^k + b_3c_3^k &= \frac{1}{(k+1)(k+2)}, \qquad k = 1,2,3, \qquad (5.71)\\ b_2Q_{21} + b_3Q_{31} &= 0. \end{aligned}$$

On montre que les conditions de quadratures pour y et $\hat{y}$ sont incompatibles. Soient la matrice et les vecteurs:

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & c_2 & c_3 \\ 0 & c_2^2 & c_3^2 \end{bmatrix}, \quad \widehat{\mathbf{b}} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}, \quad \mathbf{r} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{6} \\ \frac{1}{12} \end{bmatrix}.$$

Alors les trois équations (5.68), (5.69), $k = 1,2$, et les trois équations (5.70), (5.71), $k = 1,2$, s'écrivent respectivement de la façon suivante:

$$A\widehat{\mathbf{b}} = \mathbf{r}, \qquad A\mathbf{b} = \mathbf{r}.$$

Donc

$$A\widehat{\mathbf{b}} = A\mathbf{b}. \qquad (5.72)$$

Si A est régulière, le système (5.72) admet l'unique solution $\mathbf{b} = \widehat{\mathbf{b}}$, ce qui contredit l'existence d'une paire. Donc il faut que A soit singulière.

Dans ce cas, $c_2 = c_3$ ou $c_2c_3 = 0$. Dans ces trois cas, les trois équations (5.67) se réduisent à la forme suivante:

$$x\alpha = \frac{1}{2}, \quad x\alpha^2 = \frac{1}{3}, \quad x\alpha^3 = \frac{1}{4}. \qquad (5.73)$$

On voit immédiatement que ce système est incompatible.

On conclut à l'inexistence de méthode à 3 stages pour les paires non FSAL d'ordre $(4,5)$.

(b) *Inexistence de méthode FSAL à quatre stages.*

Les équations de conditions pour les méthodes FSAL à 4 stages sont les suivantes:

Ordre 5 en y':

$$\begin{aligned} b_1' + b_2' + b_3' + b_4' &= 1, \\ b_2'c_2^k + b_3'c_3^k + b_4'c_4^k &= \frac{1}{k+1}, \qquad k = 1,2,3,4, \\ b_2'c_2^kQ_{21} + b_3'c_3^kQ_{31} &= 0, \qquad k = 0,1, \qquad (5.74)\\ b_2'Q_{22} + b_3'Q_{32} &= 0, \end{aligned}$$

Ordre 4 en $\hat{y}$:

$$\hat{b}_1 + \hat{b}_2 + \hat{b}_3 + \hat{b}_4 = \frac{1}{2}, \qquad (5.75)$$

$$\hat{b}_2c_2^k + \hat{b}_3c_3^k + \hat{b}_4c_4^k = \frac{1}{(k+1)(k+2)}, \qquad k = 1,2, \qquad (5.76)$$

Ordre 5 en y:

$$b_1 + b_2 + b_3 + b_4 = \frac{1}{2}, \tag{5.77}$$

$$b_2 c_2^k + b_3 c_3^k + b_4 c_4^k = \frac{1}{(k+1)(k+2)}, \qquad k = 1, 2, 3, \tag{5.78}$$

$$b_2 Q_{21} + b_3 Q_{31} = 0. \tag{5.79}$$

On rappelle que $Q_{41} = 0$ puisqu'on est dans le cas FSAL. Les équations (5.74), pour $k = 0$, et (5.79) doivent être dépendantes, sinon on aurait $Q_{21} = 0$, une impossibilité. Alors, il existe deux nombres, α et β, non tous nuls tels que

$$\alpha b_i' + \beta b_i = 0, \qquad i = 2, 3.$$

On multiplie cets équations par c_i^k, $k = 1, 2, 3$, et on somme sur les i. Puis on utilise les équations de quadrature en y' et y pour obtenir le système suivant:

$$\begin{aligned}
\left(\frac{1}{2} - b_4'\right)\alpha + \left(\frac{1}{6} - b_4\right)\beta &= 0, \\
\left(\frac{1}{3} - b_4'\right)\alpha + \left(\frac{1}{12} - b_4\right)\beta &= 0, \\
\left(\frac{1}{4} - b_4'\right)\alpha + \left(\frac{1}{20} - b_4\right)\beta &= 0.
\end{aligned}$$

Si on soustrait la deuxième équation de la première et la troisième de la deuxième, on aura

$$\frac{1}{2}\alpha + \frac{1}{12}\beta = 0, \qquad \frac{1}{12}\alpha + \frac{1}{30}\beta = 0.$$

Alors $\alpha = \beta = 0$ et par conséquent, la paire n'existe pas.

(b) *Existence de méthodes non FSAL à quatre stages.*

On présente au tableau 13 une méthode à 4 stages pour les paires non FSAL d'ordre $(4, 5)$.

Tableau 13. *Une paire non FSAL d'ordre* $(4, 5)$ *à quatre stages du type III*

$\frac{17}{20}$	$\frac{289}{800}$			
$\frac{1}{2}$	$-\frac{179}{136}$	$\frac{49}{34}$		
$\frac{1}{3}$	$\frac{18\,275\,573}{66\,096\,000}$	$-\frac{8\,162\,827}{46\,267\,200}$	$-\frac{1\,211\,573}{27\,216\,000}$	
$\frac{1}{2}$	$\frac{1}{6}$	0	$\frac{1}{3}$	0
b	$\frac{11}{102}$	$\frac{200}{3\,689}$	$\frac{1}{21}$	$\frac{9}{31}$
b'	$\frac{11}{102}$	$\frac{4\,000}{11\,067}$	$\frac{2}{21}$	$\frac{27}{62}$

(c) *Existence de méthodes FSAL à cinq stages.*

Enfin, on présente au tableau 14 une méthode à 5 stages pour les paires FSAL d'ordre $(4,5)$.

Tableau 14. *Une paire d'ordre FSAL* $(4,5)$ *à cinq stages du type III*

$\frac{1}{5}$	$\frac{1}{50}$				
1	$-\frac{3}{10}$	$\frac{4}{5}$			
$\frac{2}{3}$	$-\frac{14\,839}{121\,500}$	$\frac{35\,539}{97\,200}$	$-\frac{10\,339}{486\,000}$		
1	$\frac{1}{24}$	$\frac{25}{84}$	0	$\frac{9}{56}$	0
$\hat{b}$	$-\frac{1}{12}$	$\frac{25}{48}$	$-\frac{15}{16}$	0	1
b	$\frac{1}{24}$	$\frac{25}{84}$	0	$\frac{9}{56}$	0
b'	$\frac{1}{24}$	$\frac{125}{336}$	$\frac{5}{48}$	$\frac{27}{56}$	0

Paires d'ordre $(5,6)$ **du type III.**

Il n'existe pas de méthode à 5 stages pour les paires non FSAL d'ordre $(5,6)$. La démonstration est par contradiction en supposant qu'une paire existe.

(a) *Inexistence de méthode non FSAL à cinq stages.*

Les équations de conditions pour une méthode à 5 stages du type III sont les suivantes:

Ordre 6 en y':

$$b'_1 + b'_2 + b'_3 + b'_4 + b'_5 = 1, \tag{5.80}$$

$$b'_2 c_2^k + b'_3 c_3^k + b'_4 c_4^k + b'_5 c_5^k = \frac{1}{k+1}, \quad k = 1, \ldots, 5, \tag{5.81}$$

$$b'_2 c_2^k Q_{21} + b'_3 c_3^k Q_{31} + b'_4 c_4^k Q_{41} + b'_5 c_5^k Q_{51} = 0, \quad k = 0, 1, 2, \tag{5.82}$$

$$b'_2 c_2^k Q_{22} + b'_3 c_3^k Q_{32} + b'_4 c_4^k Q_{42} + b'_5 c_5^k Q_{52} = 0, \quad k = 0, 1,$$

$$b'_2 Q_{23} + b'_3 Q_{33} + b'_4 Q_{43} + b'_5 Q_{53} = 0,$$

$$b'_3 a_{32} Q_{21} + b'_4 (a_{42} Q_{21} + a_{43} Q_{31}) + b'_5 (a_{52} Q_{21} + a_{53} Q_{31} + a_{54} Q_{51}) = 0,$$

Ordre 5 en $\hat{y}$:

$$\hat{b}_1 + \hat{b}_2 + \hat{b}_3 + \hat{b}_4 + \hat{b}_5 = \frac{1}{2}, \tag{5.83}$$

$$\hat{b}_2 c_2^k + \hat{b}_3 c_3^k + \hat{b}_4 c_4^k + \hat{b}_5 c_5^k = \frac{1}{(k+1)(k+2)}, \quad k = 1, 2, 3, \tag{5.84}$$

$$\hat{b}_2 Q_{21} + \hat{b}_3 Q_{31} + \hat{b}_4 Q_{41} + \hat{b}_5 Q_{51} = 0, \tag{5.85}$$

Ordre 6 en y:

$$\begin{aligned} b_1 + b_2 + b_3 + b_4 + b_5 &= \frac{1}{2}, && (5.86)\\ b_2c_2^k + b_3c_3^k + b_4c_4^k + b_5c_5^k &= \frac{1}{(k+1)(k+2)},\ k = 1,2,3, && (5.87)\\ b_2c_2^kQ_{21} + b_3c_3^kQ_{31} + b_4c_4^kQ_{41} + b_5c_5^kQ_{41} &= 0, \qquad k = 0,1, && (5.88)\\ b_2Q_{22} + b_3Q_{32} + b_4Q_{42} + b_5Q_{52} &= 0, \end{aligned}$$

On considère les cinq conditions de non quadrature en Q_{i1}, c'est-à-dire les équations (5.82), (5.85) et (5.88). Quatre quelconques de ces équations linéaires en Q_{i1} doivent être dépendantes, sinon on aura $Q_{21} = 0$ et le nombre de stages de la méthode se réduirait à 4; alors l'inexistence s'ensuivrait par le cas précédent.

En premier lieu, on considère les quatre équations (5.82) et (5.88) pour $k = 0, 1$. Il existe alors quatre nombres, α, β, γ et δ, non tous nuls, tels que

$$\alpha\hat{b}_i' + \beta\hat{b}_i'c_i + \gamma b_i + \delta b_ic_i = 0, \qquad i = 2,\ldots,5. \tag{5.89}$$

On multiplie ces équations par c_i^l, $l = 1,2,3$, on somme sur les i et on utilise les équations de quadrature (5.81) et (5.87); on obtient alors le système linéaire suivant:

$$\begin{aligned} \frac{1}{2}\alpha + \frac{1}{3}\beta + \frac{1}{6}\gamma + \frac{1}{12}\delta &= 0,\\ \frac{1}{3}\alpha + \frac{1}{4}\beta + \frac{1}{12}\gamma + \frac{1}{20}\delta &= 0,\\ \frac{1}{4}\alpha + \frac{1}{5}\beta + \frac{1}{20}\gamma + \frac{1}{30}\delta &= 0. \end{aligned}$$

La solution générale de ce système est:

$$\alpha = a, \quad \beta = -a, \quad \gamma = -a, \quad \delta = 0,$$

où $\alpha \neq 0$, sinon la méthode n'existerait pas. Si on substitue dans (5.89) les valeurs de $\alpha, \beta, \gamma, \delta$ obtenues, on a:

$$b_i = b_i' - b_i'c_i, \qquad i = 2,\ldots,5. \tag{5.90}$$

De même, si l'on considère les quatre conditions de non quadrature (5.82) avec $k = 0, 1$, (5.85) et (5.88) avec $k = 1$, on a, comme avant:

$$\alpha b_i' + \beta b_i'c_i + \gamma\hat{b}_i + \delta b_ic_i = 0, \qquad i = 2,\ldots,5. \tag{5.91}$$

On multiplie ces équations par c_i^l, $l = 1, 2, 3$, on somme sur les i et on utilise les équations de quadrature (5.81), (5.84) et (5.87); on obtient alors le système linéaire suivant:

$$\begin{aligned} \frac{1}{2}\alpha + \frac{1}{3}\beta + \frac{1}{6}\gamma + \frac{1}{12}\delta &= 0, \\ \frac{1}{3}\alpha + \frac{1}{4}\beta + \frac{1}{12}\gamma + \frac{1}{20}\delta &= 0, \\ \frac{1}{4}\alpha + \frac{1}{5}\beta + \frac{1}{20}\gamma + \frac{1}{30}\delta &= 0. \end{aligned}$$

La solution générale de ce système est:

$$\alpha = a, \quad \beta = -a, \quad \gamma = -a, \quad \delta = 0,$$

où $\alpha \neq 0$, sinon la méthode n'existerait pas. Si on substitue dans (5.91) les valeurs de $\alpha, \beta, \gamma, \delta$ obtenues, on a:

$$\hat{b}_i = b'_i - b'_i c_i, \qquad i = 2, \ldots, 5. \tag{5.92}$$

Il est évident que (5.90) et (5.92) donnent:

$$b_i = \hat{b}_i, \qquad i = 2, \ldots, 5,$$

et par les équations de quadrature (5.83) et (5.86), on a $b_1 = \hat{b}_1$. Par conséquent, la paire n'existe pas.

(b) *Existence de méthodes FSAL à six stages.*

Il existe une méthode à 6 stages pour les paires FSAL d'ordre $(5, 6)$ du type III puisqu'il en existe du type I.

5.6.4 Paires des types IV et V

Les résultats de non-existence sont pareils à ceux des paires du type I, puisqu'on a utilisé uniquement les formules des dérivées dans les preuves du type I. Ainsi, l'existence d'une méthode pour une paire du type I implique l'existence d'une méthode semblable des types IV et V.

5.7 Conclusion

Dans ce travail, on a déterminé le nombre minimum de stages requis pour l'existence de paires de formules de Runge–Kutta–Nyström des ordres $(1, 2)$ à $(5, 6)$. On a considéré cinq types de paires selon qu'on emploie une ou deux formules pour y et y'; par conséquent, ces types se distinguent par le mode

d'avancement soit selon les formules d'ordre inférieur (mode standard), soit celles d'ordre supérieur (mode d'extrapolation locale), ainsi que par le contrôle de l'erreur locale sur y, sur y', ou sur y et y'. La recherche nous a conduits à trouver des nouvelles paires que nous n'avons pas essayé d'optimaliser pour une classe donnée. Le but de ces paires est de confirmer l'existence d'au moins une paire dont le nombre de stages est minimum.

On résume au tableau 15 les résultats obtenus pour chaque type de paires. Le tableau donne le nombre minimum de stages pour les paires jusqu'à l'ordre $(5, 6)$.

Tableau 15. *Résumé des résultats. Nombre mininum de stages pour l'existence de chacune des paires d'un type donné.*

Ordres	Type I	Type II	Type III	Type IV	Type V
$(1, 2)$	2F	1NF 2F	2F	voir I	voir I
$(2, 3)$	3F	2F	2NF 3F	voir I	voir I
$(3, 4)$	4F	3F	3NF 4F	voir I	voir I
$(4, 5)$	5NF 6F	4NF 5F	4NF 5F	voir I	voir I
$(5, 6)$	6F	6F	6F	voir I	voir I

Il est important de remarquer que, si une méthode FSAL existe pour une paire, alors une méthode non-FSAL existe aussi, puisque cette dernière classe inclut la première. Il est également important de remarquer que si une méthode existe pour une paire du type I, alors une méthode doit exister pour les paires semblables des autres types. De plus, si une méthode existe pour une paire du type III, alors il en existe une semblable du type II. Les divers types satisfont donc les relations d'ordre partiel suivantes: I $\subset$ III $\subset$ II, I $\subset$ V $\subset$ IV et FSAL $\subset$ non FSAL.

On pourrait prolonger ce travail en considérant des paires d'ordres plus élevés, et essayer d'établir le nombre minimum de stages pour chacun des types de paires. Il est important de remarquer qu'au fur et à mesure que l'ordre augmente, les démonstrations se compliquent.

Une autre extension à ce travail, du point de vue pratique, serait de tester les nouvelles paires obtenues. Ces paires utilisent le nombre minimum de stages, mais plusieurs paires admettent des nœuds égaux. Deux questions se posent. Premièrement, même si le coût par pas est minimisé, l'est-il par pas unitaire? Deuxièmement, l'égalité des nœuds réduit-elle la fiabilité de l'estimation de l'erreur locale?

Bibliography

[1] D. G. Bettis, A Runge–Kutta–Nyström algorithm, *Celestial Mechanics*, **8** (1973) 229-233.

[2] D. G. Bettis, M. K. Horn, *Embedded Pairs of Runge-Kutta-Nyström Algorithms of Order Two through Six*, manuscript, Texas Institute of Computational Mechanics, Univ. of Texas at Austin 1978.

[3] J. C. Butcher, *The Numerical Analysis of Ordinary Differential Equations: Runge–Kutta and General Linear Methods*, Wiley, Chichester, 1987.

[4] J. R. Dormand, M. E. A. El-Mikkawy and P. J. Prince, Families of Runge–Kutta–Nyström formulae, *IMA J. Num. Anal.* **8** (1987) 235–250.

[5] J. R. Dormand, M. E. A. El-Mikkawy and P. J. Prince, Higher order embedded Runge–Kutta–Nyström formulae, *IMA J. Num. Anal.* **8** (1987) 423–430.

[6] E. Fehlberg, S. Filippi, and J. Gräf, Ein Runge–Kutta–Nyström-Formelpaar der Ordnung 10(11) für Differentialgleichungen der Form $y'' = f(x, y)$, *ZAMM*, **66**(7) (1987) 265-270.

[7] S. Filippi and J. Gräf, New Runge–Kutta–Nyström formula-pairs of order 8(7), 9(8), 10(9) and 11(10) for differential equations of the form $y'' = f(x, y)$, *J. Comp. Appl. Math.* **14** (1986) 361–370.

[8] E. Hairer, S. P. Nørsett and G. Wanner, *Solving Ordinary Differential Equations I: Nonstiff Problems*, Springer-Verlag, Berlin, 1987.

[9] Fadi Malek, *Formules de type Runge–Kutta–Nyström*, Thèse de M.Inf., Université d'Ottawa, Ottawa, Ontario, Canada K1N 6N5, janvier 1992, 45 pp.

[10] E. J. Nyström, Über die numerische Integration von Differentalgleichungen, *Acta Societatis Scientiarum Ferricae*, **50**, no 13, (1925) 1–55.

[11] P. W. Sharp, Fadi Malek et Rémi Vaillancourt, The minimum number of stages for explicit Runge–Kutta–Nyström pairs, *CRM–1791*, Centre de recherches mathématiques, Université de Montréal, C. P. 6128–A, Montréal, Qc, Canada H3C 3J7, octobre 1991.

[12] J. H, Verner, *On Deriving High Order Interpolants*, Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada K7L 3N6, manuscript, 23 February 1990, 23 pp.