Numer. Math. 32, 75–82 (1979)

Numerische Mathematik

Sur la *B*-stabilité des méthodes de Runge-Kutta

Michel Crouzeix

Département de Mathématiques et Informatique, Université de Rennes, Rennes-Beaulieu, F-35042 Rennes, Cédex, France

On *B*-Stability of the Methods of Runge Kutta

Summary. In this paper, we slightly modify the definition of *B*-stability of Butcher [1], so as to cover a wider class of differential equations, and we give simple characterizations of this property.

Subject Classifications. AMS(MOS): 65L05; CR: 5.17.

Résumé. Dans cet article, nous modifions légèrement la définition de la *B*-stabilité donnée par J.C. Butcher [1] afin qu'elle s'applique à une plus large classe d'équations différentielles et nous donnons des caractérisations simples de cette propriété.

1. Introduction

Considérons l'équation différentielle

$$y'(t)=f(t,y(t)) \qquad t_0<t<t_0+a \tag{1}$$

munie de la condition initiale

$$y(t_0)=\eta \tag{2}$$

où η est un vecteur donné de $\mathbb{C}^N$ et $f(.,.)$ une fonction définie sur $[t_0, t_0+a]\times\mathbb{C}^N$, à valeurs dans $\mathbb{C}^N$. Une méthode de Runge-Kutta implicite est définie par la donnée de q nombres réels positifs ou nuls $\tau_1, \tau_2, \ldots, \tau_q$, d'une matrice carrée $A(q\times q)$ d'élément générique $a_{ij}\in\mathbb{R}$, d'un vecteur colonne $b=(b_1, b_2, \ldots, b_q)^T\in\mathbb{R}^q$ et d'un pas de discrétisation en temps $\Delta t>0$.

Posons $t_n=t_0+n\,\Delta t$ et $t_{n,i}=t_n+\tau_i\,\Delta t$. Nous obtenons une approximation y_{n+1} de la solution $y(t_{n+1})$ de (1) à partir de l'approximation y_n de $y(t_n)$ par les équations

$$\begin{aligned} y_{n,i}&=y_n+\Delta t\sum_{j=1}^{q} a_{ij}\,f(t_{n,j},y_{n,j}) \qquad 1\leqq i\leqq q \\ y_{n+1}&=y_n+\Delta t\sum_{j=1}^{q} b_j\,f(t_{n,j},y_{n,j}). \end{aligned} \tag{3}$$

0029-599X/79/0032/0075/$01.60

Si nous partons d'une autre valeur z_n, nous obtenons z_{n+1} vérifiant

$$\begin{aligned} z_{n,i} &= z_n + \Delta t \sum_{j=1}^{q} a_{ij} f(t_{n,j}, z_{n,j}) \qquad 1 \leqq i \leqq q \\ z_{n+1} &= z_n + \Delta t \sum_{j=1}^{q} b_j f(t_{n,j}, z_{n,j}). \end{aligned} \tag{4}$$

Il est bien connu que si f vérifie la condition

$$\forall\, y, z \in \mathbb{C}^N, \quad \forall\, t, \quad \mathrm{Re}\,(f(t,y)-f(t,z), y-z) \leqq 0 \tag{5}$$

alors les solutions $y(.)$ et $z(.)$ de l'équation différentielle (1) vérifient

$$\forall\, t \geqq t', \quad |y(t)-z(t)| \leqq |y(t')-z(t')| \tag{6}$$

où nous avons noté respectivement par $(.,.)$ et par $|.|$ le produit scalaire et la norme euclidienne dans $\mathbb{C}^N$.

Nous dirons que la méthode de Runge-Kutta définie par (3) *est B-stable si pour toute fonction* $f(t, y)$ *vérifiant* (5) *et pour toute solution* y_{n+1} *et* z_{n+1} *de* (3) *et* (4), *nous avons*

$$|y_{n+1} - z_{n+1}| \leqq |y_n - z_n|. \tag{7}$$

En prenant $f(t, y)$ de la forme $f(t, y) = -\lambda y$ avec $\mathrm{Re}\,\lambda \geqq 0$, on voit immédiatement que toute méthode B-stable est obligatoirement A-stable.

Remarque. Il est habituel de ramener le système différentiel (1) en un système autonome en posant

$$Y(t) = \begin{bmatrix} y(t) \\ t \end{bmatrix} \quad \text{et} \quad F(Y(t)) = \begin{bmatrix} f(t, y(t)) \\ 1 \end{bmatrix}.$$

L'équation (1) équivaut alors à

$$Y'(t) = F(Y(t)). \tag{1.bis}$$

Si on pose aussi

$$Y_n = \begin{bmatrix} y_n \\ t_n \end{bmatrix}, \qquad Y_{n,i} = \begin{bmatrix} y_{n,i} \\ t_{n,i} \end{bmatrix}$$

le schéma (3) devient alors équivalent à

$$\begin{aligned} Y_{n,i} &= Y_n + \Delta t \sum_{j=1}^{q} a_{ij} F(Y_{n,j}) \\ Y_{n+1} &= Y_n + \Delta t \sum_{j=1}^{q} b_j F(Y_{n,j}) \end{aligned} \tag{3.bis}$$

pourvu que

$$\tau_i = \sum_{j=1}^{q} a_{ij} \qquad 1 \leqq i \leqq q. \tag{8}$$

(Il est à noter que l'hypothèse (8) est vérifiée par toutes les méthodes de Runge-Kutta qui présentent un intérêt pratique, cependant on peut construire des méthodes ne vérifiant pas cette hypothèse.) On définirait de même Z_{n+1} à partir de Z_n.

La définition originelle de la B-stabilité est due à Butcher [1]; d'après cette définition, une méthode de Runge-Kutta est dite B-stable si pour toute fonction F de $\mathbb{R}^p$ dans $\mathbb{R}^p$ vérifiant

$$\forall\, Y, Z \in \mathbb{R}^p \quad (F(Y)-F(Z), Y-Z) \leqq 0 \tag{5. bis}$$

on a

$$|Y_{n+1}-Z_{n+1}| \leqq |Y_n - Z_n|.$$

Le fait de considérer des fonctions à valeurs réelles au lieu de fonctions à valeurs complexes ne change rien car on se ramène aisément de $\mathbb{C}^N$ à $\mathbb{R}^{2N}$ et on plonge facilement $\mathbb{R}^p$ dans $\mathbb{C}^p$. Par contre, le fait de se limiter à considérer les systèmes autonomes, restreint l'intérêt de la définition. En effet (5. bis) équivaut à

$$\forall y, z \in \mathbb{R}^N, \quad \forall t, t' \in \mathbb{R} \quad (f(t, y)-f(t', z), y-z) \leqq 0 \tag{9}$$

ce qui est une condition beaucoup plus restrictive que (5). (En fait, si f est continue, (9) entraîne que $f(t, y)$ est indépendant de t.)

Il est clair que, si une méthode est B-stable au sens où nous l'entendons dans cet article, elle est aussi B-stable au sens de Butcher. Nous ne connaissons pas d'exemples de méthodes qui soient B-stables au sens de Butcher et qui ne soient pas B-stables suivant notre définition.

2. Caractérisation de la B-stabilité

Introduisons maintenant la matrice diagonale $B=\operatorname{diag}(b_i)$ d'éléments diagonaux $b_{ii}=b_i$, $1 \leqq i \leqq q$ ($b_{ij}=0$ si $i \neq j$). Nous noterons par A^T la matrice transposée de A et par b^T le vecteur transposé de b. Nous considérons aussi la matrice symétrique M définie par

$$M = BA + A^T B - b\, b^T. \tag{10}$$

Théorème. *On suppose que les nombres $\tau_1, \tau_2, \ldots, \tau_q$ sont tous distincts. Alors une condition nécessaire et suffisante pour que le méthode de Runge-Kutta* (3) *soit B-stable est que les matrices M et B soient semi-définies positives. Si les τ_i ne sont pas tous distincts, la condition est seulement suffisante.*

Démonstration.

a) *La condition est suffisante.* Posons

$$\varphi_i = \Delta t[f(t_{n,i}, y_{n,i})-f(t_{n,i}, z_{n,i})] \quad \text{et} \quad \Phi=(\varphi_1, \varphi_2, \ldots, \varphi_q)^T. \tag{11}$$

On a alors

$$\begin{aligned} y_{n,i} - z_{n,i} &= y_n - z_n + \sum_{j=1}^{q} a_{ij}\, \varphi_j \\ y_{n+1} - z_{n+1} &= y_n - z_n + \sum_{i=1}^{q} b_i\, \varphi_i \end{aligned} \tag{12}$$

d'où

$$|y_{n+1} - z_{n+1}|^2 = |y_n - z_n|^2 + 2\,\mathrm{Re} \sum_{i=1}^{q} b_i(\varphi_i, y_n - z_n) + |b^T\, \Phi|^2. \tag{13}$$

On déduit de (12) que l'on a

$$\sum_{i=1}^{q} b_i(\varphi_i, y_n - z_n) = \sum_{i=1}^{q} b_i(\varphi_i, y_{n,i} - z_{n,i}) - \sum_{i=1}^{q} (b_i\, \varphi_i,\ \sum_j a_{ij}\, \varphi_j). \tag{14}$$

Etant donnés deux vecteurs $Y = (y_1, \ldots, y_q)^T$ et $Z = (z_1, \ldots, z_q)^T$ de $(\mathbb{R}^N)^q$, nous noterons

$$((Y, Z)) = \sum_{i=1}^{q} (y_i, z_i);$$

l'équation (14) s'écrit alors:

$$\sum_{i=1}^{q} b_i(\varphi_i, y_n - z_n) = \sum_{i=1}^{q} b_i(\varphi_i, y_{n,i} - z_{n,i}) - ((B\, \Phi, A\, \Phi))$$

ce qui nous donne, en reportant dans (13)

$$|y_{n+1} - z_{n+1}|^2 = |y_n - z_n|^2 + 2\,\mathrm{Re} \sum_{i=1}^{q} b_i(\varphi_i, y_{n,i} - z_{n,i}) - ((M\, \Phi, \Phi)).$$

L'hypothèse (5) entraîne $\mathrm{Re}(\varphi_i, y_{n,i} - z_{n-i}) \leqq 0$; la matrice B étant semidéfinie positive, on a $b_i \geqq 0$ pour $1 \leqq i \leqq q$. On en déduit que

$$|y_{n+1} - z_{n+1}|^2 \leqq |y_n - z_n|^2.$$

b) *La condition est nécessaire.* Pour montrer que la condition est nécessaire, nous nous plaçons dans le cas où $N = 1$, $y_n = 1$, $z_n' = 0$ et où $f(t, y)$ est de la forme

$$f(t, y) = -\delta(t)\, y$$

où δ est une fonction arbitraire, définie sur $[t_0, t_0 + a]$, à valeurs dans $\mathbb{C}$ et vérifiant pour tout $t \in [t_0, t_0 + a]$ $\mathrm{Re}\, \delta(t) \geqq 0$.

Nous avons alors $z_{n+1} = 0$. Pour que la méthode (3) soit B-stable il est donc nécessaire que l'on ait $|y_{n+1}| \leqq 1$.

Posons $Y=(y_{n,1}, \ldots, y_{n,q})^T$, $e=(1,1,\ldots,1)^T$ et notons par D la matrice diagonale d'éléments diagonaux $d_{ii}=\Delta t\, \delta(t_{n,i})$, $1\leqq i\leqq q$ ($d_{ij}=0$ si $i\neq j$). Les équations (3) deviennent alors

$$Y=e-ADY \tag{15}$$

et

$$y_{n+1}=1-b^T DY.$$

On en déduit

$$|y_{n+1}|^2=1-2\,\mathrm{Re}\, b^T DY+|b^T DY|^2$$

ce qui s'écrit aussi

$$|y_{n+1}|^2=1-2\,\mathrm{Re}(DY, Be)+(bb^T DY, Y).$$

En utilisant (15), on obtient

$$|y_{n+1}|^2=1-2\,\mathrm{Re}(DY, BY)-(MDY, DY). \tag{16}$$

La fonction δ pouvant prendre n'importe quelle valeur dans le demiplan $\{z\in\mathbb{C};\ \mathrm{Re}\, z\geqq 0\}$, la matrice D peut être n'importe quelle matrice diagonale semi-définie positive. Pour que la méthode de Runge-Kutta soit B-stable, il faudra donc d'après (15) que, pour toute matrice diagonale semi-définie positive D, et pour tout vecteur Y vérifiant

$$(I+AD)\,Y=e \tag{15}$$

on ait

$$2\,\mathrm{Re}(DY, BY)+(MDY, DY)\geqq 0.$$

Soit D_0 une matrice diagonale, à coefficients diagonaux réels positifs; prenons $D=\varepsilon D_0$ avec $\varepsilon>0$. Lorsque ε tend vers zéro, on a $Y=e+O(\varepsilon)$ et

$$2\,\mathrm{Re}(DY, BY)+(MDY, DY)=2\varepsilon(D_0 e, Be)+O(\varepsilon^2).$$

Il faut donc que l'on ait $(D_0 e, Be)\geqq 0$ pour toute matrice diagonale D_0 à coefficients diagonaux positifs, ce qui entraîne $b_i\geqq 0$, $1\leqq i\leqq q$. La matrice B doit donc être semi-définie positive.

Prenons maintenant D de la forme $D=\varepsilon\, i D_1$ où D_1 est une matrice diagonale quelconque à coefficients réels. On a encore $Y=e+O(\varepsilon)$, mais maintenant

$$2\,\mathrm{Re}(DY, BY)+(MDY, DY)=\varepsilon^2(MD_1 e, D_1 e)+O(\varepsilon^3).$$

Il faut donc que l'on ait $(MD_1 e, D_1 e)\geqq 0$ pour toute matrice diagonale D_1 à coefficients réels, ce qui entraîne que M est semi-définie positive. •

Remarque. Si les τ_i ne sont pas tous distincts, les conditions M et B semi-définies positives ne sont pas nécessaires. Considérons par exemple la méthode de Runge-

Kutta définie par

$$\tau_1 = \tau_2 = \tfrac{1}{2}$$

$$y_{n,1} = y_n + \frac{\Delta t}{2} f(t_{n,1}, y_{n,1})$$

$$y_{n,2} = y_n + \frac{\Delta t}{2} f(t_{n,2}, y_{n,2})$$

$$y_{n+1} = y_n + 2\Delta t f(t_{n,1}, y_{n,1}) - \Delta t f(t_{n,2}, y_{n,2}).$$

Ce schéma correspond à

$$A = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \quad B = \begin{pmatrix} 2 & 0 \\ 0 & -1 \end{pmatrix} \quad M = \begin{pmatrix} -2 & 2 \\ 2 & -2 \end{pmatrix}.$$

Il est clair que l'on a $y_{n,1} = y_{n,2}$, de sorte que ce schéma se réduit en fait à la méthode du point milieu

$$y_{n,1} = y_n + \frac{\Delta t}{2} f(t_{n,1}, y_{n,1})$$

$$y_{n+1} = y_n + \Delta t f(t_{n,1}, y_{n,1}).$$

Pour cette deuxième méthode, on a

$$A = (\tfrac{1}{2}), \quad B = (1), \quad M = (0).$$

Les deux méthodes sont donc B-stables, mais, pour la première méthode, les matrices M et B ne sont pas semi-définies positives.

Proposition 2. *Si la méthode de Runge-Kutta est d'ordre p, le rang de la matrice M est inférieur ou égal à $2q-p$.*

Démonstration. Considérons les quantités $(MT^{k'}e, T^{k'}e)$ où T désigne la matrice diagonale telle que $t_{ii} = \tau_i$, $1 \leqq i \leqq q$. On a

$$(MT^k e, T^{k'} e) = b^T T^{k'} A T^k e + b^T T^k A T^{k'} e - (b^T T^k e)(b^T T^{k'} e).$$

La méthode étant d'ordre p, on a (cf. Butcher [2] ou Crouzeix [4])

$$\forall k \leqq p-1, \quad b^T T^k e = 1/(k+1),$$

$$\forall k \text{ et } k' \text{ avec } k+k' \leqq p-2, \quad b^T T^k A T^{k'} e = 1/(k'+1)(k+k'+2),$$

ce qui nous donne

$$\forall k \text{ et } k' \text{ avec } k+k' \leqq p-2, \quad (MT^k e, T^{k'} e) = 0. \tag{17}$$

Notons par E_m (resp. F_m) le sous-espace vectoriel de $\mathbb{R}^q$ engendré par les vecteurs $T^k e$ (resp. $MT^k e$), $0 \leqq k \leqq m-1$, et notons par r la partie entière de $(p+1)/2$.

D'après les relations (17), les espaces F_{p-r} et E_r sont orthogonaux. On en déduit

$$\dim F_{p-r} + \dim E_r \leqq q$$

et par suite

$$\dim E_{p-r} + \dim E_r \leqq q + \dim \operatorname{Ker} M.$$

La méthode de Runge-Kutta étant d'ordre p, il existe au moins r nombres distincts parmi $\tau_1, \tau_2, \ldots, \tau_q$; il en résulte que $\dim E_r = r$ et $\dim E_{p-r} = p-r$, d'où $\dim \operatorname{Ker} M \geqq p-q$ et par suite $\operatorname{rang}(M) \leqq 2q-p$.

3. Applications

Théorème 3. *La méthode de Runge-Kutta d'ordre $2q$ est B-stable.*

Démonstration. On sait d'après Butcher [2], qu'il existe une et une seule méthode de Runge-Kutta d'ordre $2q$ et que ses coefficients b_i sont tous strictement positifs. D'après la proposition 2, on a $M=0$. Cette méthode est donc B-stable. ●

Théorème 4. *Pour qu'une méthode de Runge-Kutta d'ordre $\geqq 2q-1$ soit B-stable, il faut et il suffit que les coefficients b_j soient positifs ou nuls et que* $\sum_{i=1}^{q} b_i(2a_{ii}-b_i) \geqq 0$.

Démonstration. La matrice M étant de rang 0 ou 1 d'après la proposition 2, pour qu'elle soit semi-définie positive il faut et il suffit que sa trace soit positive ou nulle, ce qui nous donne $\sum_{i=0}^{q} b_i(2a_{ii}-b_i) \geqq 0$. ●

Dans le cas où $f(t, y) = -\lambda y$, le schéma (3) nous donne

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

avec

$$r(z) = 1 - z\, b^T (I + z A)^{-1} e. \tag{18}$$

La fraction rationnelle $r(z)$ est l'approximation rationnelle de e^{-z} associée à la méthode de Runge-Kutta.

Théorème. 5. *Pour qu'une méthode de Runge-Kutta d'ordre $\geqq 2q-1$ soit B-stable, il suffit que les coefficients b_i soient positifs ou nuls, que la matrice A soit inversible et que $|r(\infty)| < 1$.*

Démonstration. Remarquons d'abord que $r(\infty) = 1 - b^T A^{-1} e$. On a aussi

$$(MA^{-1} e, A^{-1} e) = 2 b^T A^{-1} e - (b^T A^{-1} e)^2 = 1 - [r(\infty)]^2 > 0.$$

D'après la proposition 2, la matrice M est de rang 1; on déduit donc de l'inégalité précédente qu'elle est semi-définie positive. ●

On retrouve ainsi que les méthodes d'ordre $2q-1$ qui correspondent aux approximations sous-diagonales de Padé décrites dans B.L. Ehle [5] sont B-stables ainsi que les méthodes de la classe III_c de F.H. Chipman [3]. Donnons pour terminer deux exemples de méthodes semi-implicites B-stables.

Exemple 1. *Méthode semi-implicite d'ordre* 3.

Elle est définie par

$$\tau=\tfrac{1}{2}+\frac{1}{2\sqrt{3}}, \quad q=2$$

$$\tau_1=\tau \qquad b_1=\tfrac{1}{2} \qquad A=\begin{pmatrix} \tau & 0 \\ \frac{-1}{\sqrt{3}} & \tau \end{pmatrix}.$$

$$\tau_2=1-\tau \qquad b_2=\tfrac{1}{2}$$

Exemple 2. *Méthode semi-implicite d'ordre* 4.

Elle est définie par

$$\alpha=\frac{2}{\sqrt{3}}\cos\frac{\pi}{18}, \quad q=3$$

$$\tau_1=\frac{1+\alpha}{2} \qquad b_1=\frac{1}{6\alpha^2}$$

$$\tau_2=\tfrac{1}{2} \qquad b_2=1-\frac{1}{3\alpha^2} \qquad A=\begin{pmatrix} \frac{1+\alpha}{2} & 0 & 0 \\ -\frac{\alpha}{2} & \frac{1+\alpha}{2} & 0 \\ 1+\alpha & -(1+2\alpha) & \frac{1+\alpha}{2} \end{pmatrix}.$$

$$\tau_3=\frac{1-\alpha}{2} \qquad b_3=\frac{1}{6\alpha^2}$$

Remerciements. L'auteur tient à remercier le professeur G. Wanner qui lui a signalé une erreur commise dans une première rédaction de cet article.

Bibliographie

1. Butcher, J.C.: A stability property of implicit Runge-Kutta methods. Nordisk Tidskr. Informationsbehandling (BIT) **15**, 358–361 (1975)
2. Butcher, J.C.: Implicit Runge-Kutta processes. Math. Comput. **18**, 50–64 (1964)
3. Chipman, F.H.: A-stable Runge-Kutta processes. Nordisk Tidskr. Informationsbehandling (BIT) **11**, 384–388 (1971)
4. Crouzeix, M.: Sur l'approximation des equations differentielles opérationnelles linéaires par des méthodes de Runge-Kutta. Thèse, Paris, 1975
5. Ehle, B.L.: On pade approximation to the exponential function and A-stable methods for the numerical solution of initial value problems. Research Report CSRR 2010, Dept. AACS, University of Waterloo, 1969

Reçu le 2 mai 1978