RAIRO
ANALYSE NUMÉRIQUE

CATHERINE BOLLEY
MICHEL CROUZEIX

Conservation de la positivité lors de la discrétisation des problèmes d'évolution paraboliques

RAIRO – Analyse numérique, tome 12, n° 3 (1978), p. 237-245.
<http://www.numdam.org/item?id=M2AN_1978__12_3_237_0>

R.A.I.R.O. Analyse numérique/Numerical Analysis
(vol. 12, n° 3, 1978, p. 237 à 245).

CONSERVATION DE LA POSITIVITÉ LORS DE LA DISCRÉTISATION DES PROBLÈMES D'ÉVOLUTION PARABOLIQUES (*)

par Catherine BOLLEY (1) et Michel CROUZEIX (2)

Communiqué par P-A RAVIART

Résumé. — *Il est connu que la solution de l'équation de la chaleur associée à des données positives est elle-même positive. Dans cet article, nous montrons que pour conserver cette propriété par discrétisation, il faut soit utiliser des schémas en temps d'ordre au plus un, soit imposer des conditions de type stabilité liant le pas de temps à la précision de la discrétisation en espace.*

1. NOTATIONS ET POSITION DU PROBLÈME

Soit $(\Omega, \mathscr{F}, \mu)$ un espace mesuré et $H = L^2_{\mathbb{R}}(\Omega)$ l'espace des fonctions de carré sommable sur Ω. Soit $f \in H$; on dira que f est positif si pour tout $x \in \Omega$, on a $f(x) \geqq 0$; on écrira alors $f \geqq 0$. Soit B un opérateur linéaire continu de H dans H; on dira que B est positif si pour tout $f \geqq 0$, $f \in H$, on a $Bf \geqq 0$; on notera alors $B \geqq 0$.

Soit A un opérateur maximal accrétif sur H (*cf.* Kato [5]) de domaine dense dans H; on dira que A conserve la positivité si

$$\forall \alpha > 0, \quad (\alpha I + A)^{-1} \geqq 0.$$

Considérons maintenant le problème d'évolution

$$\left.\begin{aligned} u'(t) + A u(t) &= f(t), \qquad t > 0, \\ u(0) &= u_0 \end{aligned}\right\} \tag{1}$$

On sait que

$$f \geqq 0 \quad \text{et} \quad u_0 \geqq 0 \quad \Rightarrow \quad \forall t \geqq 0, \quad u(t) \geqq 0. \tag{2}$$

Le but de cet article est de rechercher les méthodes de discrétisation en temps qui conservent cette dernière propriété. Ces méthodes peuvent être utiles dans

(*) Reçu septembre 1977
(1) Laboratoire d'Analyse numérique, Institut National des Sciences appliquées, Rennes.
(2) Département de Mathématiques et Informatique, Université de Rennes, Rennes.

l'approximation de certains problèmes non linéaires (*cf.* C. Bolley [2]). Pour des propriétés analogues dans le cadre des problèmes hyperboliques, *cf.* Lax [6], A. Harten, J. M. Hyman et P. D. Lax [4].

2. MÉTHODES DE DISCRÉTISATION A UN PAS

Donnons-nous un pas de discrétisation $\Delta t > 0$. Les méthodes usuelles à un pas nous donnent une approximation u_n de la solution u du problème (1) à l'instant $t_n = n\Delta t$, par un schéma de la forme

$$u_{n+1} = r(\Delta t\, A)\, u_n + \Delta t \sum_{i=1}^{q} r_i(\Delta t\, A)\, f(t_n + \tau_i \Delta t), \tag{3}$$

dans lequel $r(z)$ est une approximation rationnelle de e^{-z} au voisinage de $z=0$, $r_i(z)$, $1 \leqq i \leqq q$, des fractions rationnelles vérifiant $\sum_{i=1}^{q} r_i(0) = 1$ et τ_i, $1 \leqq i \leqq q$, q nombres réels donnés.

Lorsque la méthode à un pas est d'ordre p, on a

$$r(z) = e^{-z} + \mathcal{O}(z^{p+1}) \quad \text{au voisinage de } z=0. \tag{4}$$

Nous recherchons ici les schémas qui conservent l'analogue discret de la propriété (2), c'est-à-dire les schémas tels que

$$f \geqq 0 \quad \text{et} \quad u_0 \geqq 0 \quad \Rightarrow \quad \forall n, \quad u_n \geqq 0. \tag{5}$$

Nous dirons alors que le schéma conserve la positivité. Pour qu'il en soit ainsi, il est nécessaire et suffisant que l'on ait

$$r(\Delta t\, A) \geqq 0 \qquad \text{et} \qquad \forall i = 1, \ldots, q, \quad r_i(\Delta t\, A) \geqq 0.$$

Rappelons maintenant la :

DÉFINITION : *On dit qu'une fonction f est totalement monotone sur un intervalle I de* $\mathbf{R}$, *si elle est indéfiniment dérivable sur I et si :*

$$\forall x \in I, \quad \forall k \in N \qquad (-1)^k f^{(k)}(x) \geqq 0.$$

On a la caractérisation suivante due à S. Bernstein [1].

THÉORÈME: *Pour que f soit totalement monotone sur* $[0, +\infty[$, *il faut et il suffit qu'il existe une mesure $d\alpha$, positive et bornée sur* $[0, +\infty[$ *telle que :*

$$f(x) = \int_0^{\infty} e^{-xt}\, d\alpha(t).$$

LEMME 1 : *Soit r une fraction rationnelle; une condition nécessaire et suffisante pour que l'on ait* $r(A) \geqq 0$ *pour tout opérateur maximal accrétif A conservant la positivité, est que r soit totalement monotone sur* $[0, +\infty[$.

Démonstration : Remarquons tout d'abord que r ne peut pas avoir de pôle réel positif car sinon $r(A)$ ne serait pas défini pour certains opérateurs de la forme $A = xI$ avec $x \geqq 0$, opérateurs qui conservent la positivité.

Condition nécessaire

Nous nous plaçons dans le cas où $H = \mathbf{R}^n$.

Soit $x > 0$; prenons $A = xI - \varepsilon B$ où B est la matrice $(n \times n)$ définie par

$$b_{ij} = 0 \quad \text{si} \quad |i-j| \neq 1,$$
$$b_{ij} = 1 \quad \text{si} \quad |i-j| = 1.$$

Pour $0 < \varepsilon < x/2$, la matrice A est une M-matrice, elle conserve donc la positivité. Pour ε suffisamment petit, on a

$$r(A) = \sum_{k=0}^{\infty} (-1)^k \frac{\varepsilon^k r^{(k)}(x)}{k\ !} B^k.$$

L'élément de la $(k+1)$-ième ligne, 1re colonne de $r(A)$ est égal à

$$(r(A))_{k+1,1} = (-1)^k \frac{\varepsilon^k r^{(k)}(x)}{k\ !} + \mathcal{O}(\varepsilon^{k+1}).$$

Pour que $r(A)$ soit positif, il est donc nécessaire que

$$\forall k = 0, \ldots, n-1, \quad (-1)^k r^{(k)}(x) \geqq 0.$$

Condition suffisante

D'après le théorème de Bernstein, si r est totalement monotone sur $[0, +\infty[$, alors r est holomorphe dans le demi-plan $\operatorname{Re} z > 0$ et on a

$$\forall z \quad \text{avec} \quad \operatorname{Re} z > 0, \qquad |r(z)| \leqq r(0).$$

On en déduit que le rayon de convergence de la série

$$r(x) + \ldots + (-1)^k \frac{r^{(k)}(x)}{k\ !} z^k + \ldots$$

est strictement supérieur à x.

Soit maintenant $\varepsilon > 0$; posons $A_\varepsilon = A(I + \varepsilon A)^{-1}$. Pour $x \geqq 1/\varepsilon$, on a

$$xI - A_\varepsilon = \left(x - \frac{1}{\varepsilon}\right) I + \frac{1}{\varepsilon}(I + \varepsilon A)^{-1},$$

d'où

$$xI - A_\varepsilon \geqq 0$$

et

$$\| xI - A_\varepsilon \|_{\mathcal{L}(H,H)} \leqq x - \frac{1}{\varepsilon} + \frac{1}{\varepsilon} = x.$$

On a donc

$$r(A_\varepsilon) = r(x)I + \ldots + (-1)^k \frac{r^{(k)}(x)}{k\,!}(xI - A_\varepsilon)^k + \ldots$$

d'où

$$r(A_\varepsilon) \geqq 0, \qquad \forall\, 0 < \varepsilon \leqq \frac{1}{x}.$$

Pour achever la démonstration du lemme, il suffit de montrer que $r(A_\varepsilon)$ tend vers $r(A)$ lorsque ε tend vers zéro.

D'après une variante d'un théorème de J. von Neumann (*cf.* Crouzeix, [3], p. 34) :

$$\| r(A_\varepsilon) - r(A) \|_{\mathcal{L}(H,H)}$$
$$= \| r(A(I + \varepsilon A)^{-1}) - r(A) \|_{\mathcal{L}(H,H)} \leqq \sup_{\mathrm{Re}\, z \geqq 0} \left| r\left(\frac{z}{1+\varepsilon z}\right) - r(z) \right|,$$

$$\| r(A_\varepsilon) - r(A) \|_{\mathcal{L}(H,H)}$$
$$\leqq \sup_{\substack{\mathrm{Re}\, z \geqq 0 \\ |z| \leqq M}} \left| r\left(\frac{z}{1+\varepsilon z}\right) - r(z) \right| + \sup_{|z| \geqq M} \left| r\left(\frac{z}{1+\varepsilon z}\right) - r(z) \right|.$$

Choisissons ε et M tels que $\varepsilon \leqq 1/2M$; alors :

$$|z| \geqq M \quad \Rightarrow \quad \left| \frac{z}{1+\varepsilon z} \right| \geqq \frac{M}{1+\varepsilon M} \geqq \frac{2M}{3}$$

et donc

$$\| r(A_\varepsilon) - r(A) \|_{\mathcal{L}(H,H)}$$
$$\leqq \sup_{\substack{\mathrm{Re}\, z \geqq 0 \\ |z| \leqq M}} \left| r\left(\frac{z}{1+\varepsilon z}\right) - r(z) \right| + 2 \sup_{|z| \geqq 2M/3} |r(z) - r(\infty)|,$$

ce qui montre que

$$\| r(A_\varepsilon) - r(A) \|_{\mathcal{L}(H,H)}$$

tend vers zéro lorsque ε tend vers zéro.

LEMME 2 : *Si f est totalement monotone sur* $[0, +\infty[$, *et si* $f(z) = e^{-z} + \mathcal{O}(z^3)$ *au voisinage de* $z = 0$, *alors* $f(z) = e^{-z}$.

Démonstration : D'après le théorème de Bernstein :

$$f(0) = \int_0^{+\infty} d\alpha(t) = 1,$$

$$f'(0) = -\int_0^{+\infty} t\, d\alpha(t) = -1,$$

$$f''(0) = \int_0^{+\infty} t^2\, d\alpha(t) = 1.$$

On en déduit que

$$\int_0^{+\infty} (1-t)^2\, d\alpha(t) = 0,$$

$d\alpha$ étant une mesure positive, $d\alpha$ a son support réduit au point $t=1$; comme de plus $\int_0^{\infty} d\alpha(t) = 1$, $d\alpha$ est la mesure de Dirac au point $t=1$. Donc

$$f(x) = \int_0^{+\infty} e^{-xt}\, d\alpha(t) = e^{-x}.$$

Il résulte des lemmes 1 et 2 et de la condition (4) que l'on a le :

THÉORÈME 1 : *Il n'existe pas de méthode à un pas du type* (3) *conservant inconditionnellement la positivité, qui soit d'ordre supérieur ou égal à* 2.

REMARQUE : Il existe des méthodes à un pas d'ordre 1, conservant inconditionnellement la positivité, par exemple le schéma implicite classique

$$U_{n+1} = (I + \Delta t\, A)^{-1} (U_n + \Delta t\, f(t_{n+1})).$$

En fait, dans la pratique, les méthodes de discrétisation en temps sont toujours utilisées dans le cas où $H = \mathbf{R}^n$, cas obtenu par discrétisation en espace des problèmes d'évolution.

Dans ce cas A est une matrice carrée $n \times n$ semi-définie positive. Une condition nécessaire et suffisante pour que A conserve la positivité est alors que, pour tout $\alpha > 0$, $\alpha I + A$ soit une M-matrice (*cf.* Varga [8]). En effet, pour $\alpha > 0$ tendant vers l'infini

$$(\alpha I + A)^{-1} = \frac{1}{\alpha} I - \frac{1}{\alpha^2} A + \mathcal{O}\left(\frac{1}{\alpha^3}\right).$$

La condition $(\alpha I + A)^{-1} \geqq 0$ entraîne donc que les coefficients non diagonaux de A sont négatifs.

Dans ce qui suit, nous dirons que A est une $\overline{M}$-matrice si pour tout $\alpha > 0$,

$\alpha I + A$ est une M-matrice. Nous noterons $\sigma(A) = \max_{1 \leqq i \leqq n} a_{ii}$ où a_{ii}, $1 \leqq i \leqq n$, désignent les coefficients diagonaux de A.

LEMME 3 : *Soit r une fraction rationnelle; une condition nécessaire et suffisante pour que l'on ait $r(A) \geqq 0$ pour toute $\overline{M}$-matrice A vérifiant $\sigma(A) \leqq M$ est que r soit totalement monotone sur* $[0, M]$.

Démonstration : La démonstration de la condition nécessaire se fait comme dans le lemme 1. Pour la condition suffisante, remarquons tout d'abord que $\sigma(A) \leqq M$ entraîne $MI - A \geqq 0$. Notons par ρ le rayon spectral de la matrice $MI - A$; d'après le théorème de Perron-Frobenius, il existe un vecteur $x \geqq 0$, $x \neq 0$, tel que $(MI - A)\, x = \rho x$, d'où

$$\forall \alpha > 0, \quad (\alpha I + A)\, x = (\alpha + M - \rho)\, x$$

$(\alpha I + A)^{-1}$ étant positif, on en déduit que $\forall \alpha > 0$ $(\alpha + M - \rho) \geqq 0$ d'où $\rho \leqq M$.

Notons par R le rayon de convergence de la série

$$r(M) + \ldots + (-1)^k \frac{r^{(k)}(M)}{k\,!} z^k + \ldots$$

Les coefficients $(-1)^k (r^{(k)}(M)/k\,!)$ étant positifs, pour $z = R$ cette série est divergente; on en déduit que $M - R$ est un pôle de r, d'où, puisque $r \in C^\infty([0, M])$, $R > M$.

On a donc

$$r(A) = r(M)\, I + \ldots + (-1)^k \frac{r^{(k)}(M)}{k\,!} (MI - A)^k + \ldots$$

ce qui montre que $r(A) \geqq 0$.

On obtient ainsi le théorème de conservation conditionnelle de la positivité.

THÉORÈME 2 : *On suppose que les fractions rationnelles r et r_i, $1 \leqq i \leqq q$, du schéma* (3) *sont totalement monotones sur* $[0, M]$, *alors ce schéma conserve la positivité pour toute $\overline{M}$-matrice A et pour tout Δt vérifiant :*

$$\Delta t\, \sigma(A) \leqq M. \tag{6}$$

Exemples :

1° Le schéma explicite

$$U_{n+1} = (I - \Delta t\, A)\, U_n + \Delta t\, f(U_n),$$

est d'ordre 1 et conserve la positivité dès que $\Delta t\, \sigma(A) \leqq 1$.

2° Le schéma de Crank-Nicolson :

$$U_{n+1}=\left(I-\frac{\Delta t}{2}A\right)\left(I+\frac{\Delta t}{2}A\right)^{-1}U_n+\Delta t\left(I+\frac{\Delta t}{2}A\right)^{-1}f(t_{n+1/2}),$$

est d'ordre 2 et conserve la positivité dès que $\Delta t\,\sigma(A)\leqq 2$.

3° Le schéma de Hammer-Hollingsworth, d'ordre 4 correspond au cas où

$$r(z)=\frac{1-(1/2)z+(1/12)z^2}{1+(1/2)z+(1/12)z^2},$$

est l'approximation de Padé (2/2) de e^{-z}.

Des calculs simples montrent que

$$r^{(n)}(0)=(-1)^n\frac{n\,!}{(2\sqrt{3})^{n-1}}2\sin n\frac{\pi}{6}\qquad(n>0).$$

Il n'existe donc pas de réel $M>0$ tel que r soit totalement monotone sur $[0, M]$. Ce schéma ne conservera jamais la positivité.

3. MÉTHODES DE DISCRÉTISATION A PAS MULTIPLES

Étant donné $\Delta t>0$ un pas de discrétisation, et $U_0, U_1, \ldots, U_{q-1}$ q valeurs initiales, un schéma linéaire à q pas s'écrit sous la forme suivante :

$$\sum_{i=0}^{q}\alpha_i U_{n+i}=\Delta t\sum_{i=0}^{q}\beta_i(-A\,U_{n+i}+f(t_{n+i})),\tag{7}$$

où α_i et β_i, $0\leqq i\leqq q$, sont $2q+2$ réels donnés, avec $\alpha_q\neq 0$ et $t_{n+i}=(n+i)\,\Delta t$, pour $i=0, \ldots, q$.

On supposera que $\alpha_q>0$.

Lorsque la méthode est d'ordre p, on a

$$\left.\begin{aligned}&\sum_{i=0}^{q}\alpha_i=0\\ \text{et}\qquad&\sum_{i=0}^{q}i^k\alpha_i=k\sum_{i=0}^{q}i^{k-1}\beta_i,\qquad 1\leqq k\leqq p.\end{aligned}\right\}$$

Lorsque $(\alpha_q I+\beta_q\Delta t\,A)$ est inversible, le schéma s'écrit sous la forme équivalente suivante :

$$U_{n+q}=\sum_{i=0}^{q-1}r_i(\Delta t\,A)\,U_{n+i}+\Delta t\sum_{i=0}^{q}s_i(\Delta t\,A)\,f(t_{n+i}),$$

où

$$r_i(z) = -\frac{\alpha_i + \beta_i z}{\alpha_q + \beta_q z}, \qquad 0 \leqq i \leqq q-1$$

et

$$s_i(z) = \frac{\beta_i}{\alpha_q + \beta_q z}, \qquad 0 \leqq i \leqq q$$

On dira que le schéma (7) conserve la positivité si

$$f \geqq 0 \quad \text{et} \quad U_i \geqq 0, \qquad 0 \leqq i \leqq q-1 \quad \Rightarrow \quad \forall n, \quad U_n \geqq 0.$$

Une condition nécessaire et suffisante pour qu'il en soit ainsi est que l'on ait

$$r_i(\Delta t\, A) \geqq 0, \quad 0 \leqq i \leqq q-1 \qquad \text{et} \qquad s_i(\Delta t\, A) \geqq 0, \quad 0 \leqq i \leqq q.$$

THÉORÈME 3 : *Il n'existe pas de schémas à pas multiples conservant inconditionnellement la positivité qui soient d'ordre supérieur ou égal à 2.*

Démonstration : Supposons l'existence d'un tel schéma; alors β_q est nécessairement positif ou nul car sinon $r_i(\Delta t\, A)$ et $s_i(\Delta t\, A)$ ne seraient pas définis pour certains opérateurs A de la forme λI avec $\lambda \geqq 0$.

Soit $i \in \{1, \ldots, q-1\}$; pour que l'on ait $\forall \lambda \geqq 0$, $r_i(\lambda I) \geqq 0$, il est nécessaire et suffisant que $\alpha_i \leqq 0$ et $\beta_i \leqq 0$.

Une méthode d'ordre 2 vérifie :

$$\left.\begin{aligned} \sum_{i=0}^{q} \alpha_i &= 0, \\ \sum_{i=0}^{q} i\,\alpha_i &= \sum_{i=0}^{q} \beta_i, \\ \sum_{i=0}^{q} i^2 \alpha_i &= 2 \sum_{i=0}^{q} i\,\beta_i, \end{aligned}\right\} \tag{8}$$

d'où

$$\sum_{i=0}^{q} (q-i)^2 \alpha_i = -2 \sum_{i=0}^{q} (q-i)\,\beta_i.$$

Comme on doit avoir $\alpha_i \leqq 0$ et $\beta_i \leqq 0$ pour $i = 0, \ldots, q-1$, on en déduit que $\forall i = 0, \ldots, q-1$, $\alpha_i = \beta_i = 0$. En reportant dans les équations (8) on obtient aussi $\alpha_q = \beta_q = 0$, ce qui est contraire à l'hypothèse $\alpha_q > 0$.

REMARQUE : Ce théorème reste valable lorsque nous nous limitons au cas où f est identique à 0. De plus, nous n'avons utilisé que des opérateurs de la forme λI avec $\lambda \geqq 0$; le théorème reste encore vrai si on se limite aux équations différentielles

$$u'(t) + \alpha\, u(t) = 0 \qquad \text{avec} \quad \alpha \geqq 0.$$

Remarquons maintenant que pour avoir $r_i(z)$ et $s_i(z)$ totalement monotones sur l'intervalle $[0, M]$ avec $M>0$, il faut et il suffit que l'on ait

$$\left.\begin{aligned} \beta_i &\geqq 0, && i=0, \ldots, q, \\ \alpha_i &\leqq 0, && i=0, \ldots, q-1, \\ \alpha_i + M\beta_i &\leqq 0, && i=0, \ldots, q-1. \end{aligned}\right\} \tag{9}$$

On en déduit le :

THÉORÈME 4 : *Une condition nécessaire et suffisante pour que le schéma* (7) *conserve la positivité pour tout* $\Delta t>0$ *et toute* $\overline{M}$*-matrice vérifiant* $\Delta t\,\sigma(A)\leqq M$, *est que* α_i, β_i *et* M *satisfassent aux conditions* (9).

Exemple : Le schéma multipas défini par (*cf.* Zlámal [9] :

$$\left(\frac{1}{2}+\eta\right)U_{n+2}-2\eta\,U_{n+1}-\left(\frac{1}{2}-\eta\right)U_n$$

$$=\Delta t\left[\frac{1}{4}(1+\eta)^2(-A\,U_{n+2}+f(t_{n+2}))\right.$$

$$\left.+\frac{1}{2}(1-\eta^2)(-A\,U_{n+1}+f(t_{n+1}))+\frac{1}{4}(1-\eta)^2(-AU_n+f(t_n))\right]$$

est d'ordre 2 dès que $\eta>0$ et conserve la positivité dès que $0<\eta\leqq 1/2$ et :

$$\Delta t\,\sigma(A)\leqq \mathrm{Inf}\left\{\frac{4\eta}{1-\eta^2}, 4\frac{(1/2)-\eta}{(1-\eta)^2}\right\}=M,$$

M est maximal pour $\eta=1/3$; on a alors $M=3/2$.

BIBLIOGRAPHIE

1. S. BERNSTEIN, *Sur les fonctions absolument monotones*, Acta Mathematica, vol. 51, 1928, p. 1-66.
2. C. BOLLEY, *Thèse de 3e Cycle*, Rennes, 1977.
3. M. CROUZEIX, *Thèse*, Paris, 1975.
4. A. HARTEN, J. M. HYMAN et P. D. LAX, *On Finite Difference Approximations and Entropy Conditions for Shocks*, Comm. Pure Appl. Math., vol. 29, 1976, p. 297-322.
5. T. KATO, *Perturbation Theory for Linear Operators*, Springer Verlag, Berlin-Heidelberg-New York.
6. P. D. LAX, *On the Stability of Difference Approximations to solutions of Hyperbolic Equations with Variable Coefficients*, Comm. Pure Appl. Math., vol. 14, 1961, p. 497-520.
7. J. VON NEUMANN, *Eine Spektraltheorie für allgemeine Operatoren eines unitären Raumes*, Math. Nachrichten, vol. 4, 1951, p. 258-281.
8. R. S. VARGA, *Matric Iterative Analysis*, Prentice Hall, 1962.
9. M. ZLÁMAL, *Finite Element Methods in Heat Conduction Problems*, Proceedings of the Brunel Conference of Finite Elements, 1975.