Institutional Repository - Research Portal
Dépôt Institutionnel - Portail de la Recherche
researchportal.unamur.be
University of Namur

THESIS / THÈSE

**MASTER EN SCIENCES MATHÉMATIQUES**

# Méthodes d'intégration numérique à pas et dérivées multiples

Crowet, Françoise; Debleser, Myriam

*Award date:*
1979

*Awarding institution:*
Universite de Namur

Link to publication

**General rights**
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

**Take down policy**
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 15. May. 2024

FACULTÉS UNIVERSITAIRES NOTRE-DAME DE LA PAIX

NAMUR

# METHODES D'INTEGRATION NUMERIQUE A PAS ET DERIVEES MULTIPLES

PROMOTEUR : J.P. THIRAN

Françoise CROWET

Myriam DEBLESER

Mémoire présenté pour l'obtention du grade de Licencié en Sciences Mathématiques.

1979

19695

A Jean Compère
et
Jean-Paul Godefroid

Nous tenons tout d'abord à exprimer nos très sincères remerciements à Monsieur J.P. Thiran qui a accepté de diriger nos recherches et nous a aimablement apporté son dévouement et sa compétence.

Nous avons aussi beaucoup apprécié l'aide et les encouragements de Monsieur Ph. Dontaine. Nous l'en remercions vivement.

Nous témoignons également notre reconnaissance à Madame D. Antoine et Madame C. Noël qui ont veillé avec soin à la dactylographie, ainsi qu'à toute l'équipe qui s'est chargée de la reproduction de ce mémoire.

Durant ces quatre années d'études, nos parents ont contribué à notre réussite, par leur compréhension et leurs encouragements répétés. Pour cela, nous les remercions profondément.

Françoise Myriam

# PREFACE

Depuis plusieurs années, des méthodes d'intégration numérique à pas multiples, telles les méthodes d'Adams, sont utilisées pour la résolution de problèmes différentiels à valeurs initiales. Cependant, ces méthodes posent des problèmes de stabilité numérique pour l'intégration de systèmes particuliers : les systèmes Stiff. En effet, si nous sélectionnons parmi ces méthodes, celles qui sont A-stables et conviennent ainsi à l'intégration de tels systèmes, l'ordre d'erreur se réduit à 2 [11], alors que de manière générale, il dépend linéairement du nombre de pas.
Il y a, par conséquent, incompatibilité entre la A-stabilité et la précision.

Deux façons d'augmenter la précision, tout en gardant la A-stabilité, ont été proposées : l'utilisation des méthodes composites [37] et celle des méthodes à pas et dérivées multiples. C'est à ces dernières que nous nous intéresserons dans ce mémoire. En effet, même si les évaluations des dérivées successives sont coûteuses au point de vue calcul, ces méthodes ont l'avantage d'offrir un ordre d'erreur maximum proportionnel à l'ordre des dérivées et ce, en gardant la forte hypothèse de A-stabilité.

Outre l'avantage d'allier la précision à la stabilité, l'étude des méthodes à pas et dérivées multiples constitue un préliminaire à celle des méthodes composites, car elle fournit pour ces dernières une borne supérieure de l'ordre d'erreur.

L'étude des méthodes à pas et dérivées multiples, réalisée dans ce mémoire, se scinde en deux parties.
La première, plus générale, introduit les définitions relatives à ces méthodes (chapitre I), et recherche l'ordre de consistance maximum qu'elles peuvent atteindre (chapitre II).
Dans le troisième chapitre, nous introduisons la stabilité asymptotique et nous étudions l'ordre d'erreur maximum.

La seconde partie s'intéresse à la A-stabilité qui apporte une solution au problème posé par les systèmes Stiff, introduits au chapitre IV.
Une caractérisation algébrique de cette A-stabilité est proposée au chapitre V alors que le sixième chapitre étudie l'ordre d'erreur maximum des méthodes A-stables et convergentes.

Les notes de Jeltsch [ 25 ] sont à l'origine de notre travail. Nous devons aussi mentionner le rapport de Rubin [ 37 ] qui nous fut utile pour l'étude algébrique de la A-stabilité.

Les chapitres I et IV sont le fruit d'un travail commun tandis que les chapitres II et VI ont été rédigés par Myriam Debleser et les chapitres III et V par Françoise Crowet.

°

° °

TABLE DES MATIERES

°

° °

# CHAPITRE I

## GÉNÉRALITÉS SUR LES MÉTHODES D'INTÉGRATION NUMÉRIQUE À PAS ET DÉRIVÉES MULTIPLES

## 1.1. DEFINITION

Considérons l'équation différentielle

$$y' = f(x,y) \tag{1.1}$$

où ' désigne la dérivée par rapport à la variable x considérée dans l'intervalle $[a,b]$, $-\infty < a < b \leq +\infty$.

satisfaisant à la condition initiale.

$$y(a) = n \tag{1.2}$$

Nous savons que si $f(x,y)$ est continue sur $[a,b] \times \mathbb{R}$ et Lipschitzienne en la seconde variable, alors l'équation différentielle (1.1) admet une et une seule solution.

Une méthode d'intégration numérique cherche à approcher la solution exacte $y(x)$, en les points $x_n = a + n\,h$, où $n \in \mathbb{N}$, $x_n \in [a,b]$ et $h \in \mathbb{R}_0^+$, par $y_n$ déterminés par la récurrence suivante :

$$\sum_{i=0}^{k} \alpha_i\, y_{n+i} = \sum_{i=0}^{k} \sum_{j=1}^{\ell_i} h^j\, \beta_{ij}\, f_{n+i}^{(j-1)} \qquad n=0,1,2,\ldots \tag{1.3}$$

Cette formule (1.3) met en jeu les éléments suivants :

- k : le nombre de pas, qui est entier
- $\ell_i$ : l'ordre maximum de la dérivée de la solution exacte en l'abscisse $x_{n+i}$
- $\alpha_i$, $\beta_{ij}$ : constantes réelles

  Nous supposerons toujours $\alpha_k \neq 0$, ce qui définit une méthode à itération directe.
- $f_{n+i}^{(j-1)} = f^{(j-1)}(x_{n+i}, y_{n+i})$

  où $i = 0,1,\ldots,k$

  $j = 1,\ldots,\ell_i$

  Ces fonctions sont définies par les relations :

  $f^{(0)}(x,y) = f(x,y)$

  $f^{(j)}(x,y) = \frac{\partial}{\partial x} f^{(j-1)}(x,y) + f(x,y) \frac{\partial}{\partial y} f^{(j-1)}(x,y)$

  pour $j = 1,\ldots, \ell-1$ où l'on définit $\ell$ comme étant $\max \{\ell_i : i = 0,1,\ldots,k\}$.

Nous dirons que (1.3) définit *une méthode* $(k,\ell)$.

Si $\sum_{j=1}^{\ell_k} | \beta_{kj} | = 0$ , la méthode est dite *explicite*.

Dans le cas contraire, elle est *implicite*.

Voyons sous quelles conditions la relation (1.3) admet une et une seule solution.

Théorème 1.1 [ 25 ]

*Supposons que* $f^{(j)}(x,y)$ *existent pour* $j=0,1,\ldots,\ell-1$ *et satisfassent à la condition de Lipschitz.*

*Alors on peut trouver un* $h^* > 0$ *tel que, quels que soient les réels* $\eta_0, \eta_1, \ldots, \eta_{k-1}$ *et* $h \in [0,h^*]$ , *il existe une et une seule suite* $\{y_n\}$ *solution de (1.3) avec :*

$$y_i = \eta_i \qquad i = 0,1,\ldots,k-1.$$

La démonstration utilise le théorème du point fixe et ne présente aucun intérêt particulier dans le cadre de ce mémoire.

## 1.2. POLYNOMES CARACTERISTIQUES - ORDRES DE CONSISTANCE ET D'ERREUR

Notons $\alpha_{io} = \alpha_i$ pour $i = 0,1,\ldots, k$

$$\alpha_{ij} = \begin{cases} -\beta_{ij} & \text{pour } j = 1,2,\ldots,\ell_i \\ 0 & \text{pour } j = \ell_i+1, \ell_i+2, \ldots, \ell \end{cases} \qquad i=0,1,\ldots, k$$

La relation (1.3) peut alors s'écrire :

$$\sum_{i=0}^{k} \sum_{j=1}^{\ell} \alpha_{ij} h^j f_{n+i}^{(j-1)} = 0 \qquad (1.4)$$

Dans la suite, nous utiliserons *les polynômes caractéristiques* qui sont définis de la manière suivante :

$$\rho_j(\xi) = \sum_{i=0}^{k} \alpha_{ij} \xi^i \qquad j=0,1,\ldots,\ell \qquad (1.5)$$

Rappel

L'opérateur de déplacement E est défini par la relation :

$$E\, y_n = y_{n+1}$$

tandis que la formule

$$D\, y(x) = \frac{d}{dx}\, y(x)$$

définit l'opérateur de dérivation D.

Dès lors, (1.4) prend la forme

$$\sum_{j=o}^{\ell} h^j \rho_j(E)\, y_n^{(j)} = 0$$

ou encore

$$\left( \sum_{j=o}^{\ell} h^j \rho_j(E)\, D^j \right) y_n = 0 \qquad (1.6)$$

Cette dernière relation définit un nouvel opérateur

$$L\,[y(x),h] = \left( \sum_{j=o}^{\ell} h^j \rho_j(E)\, D^j \right) y(x)$$

qui par (1.5) s'écrit encore :

$$L\,[y(x),h] = \sum_{i=o}^{k} \sum_{j=o}^{\ell} \alpha_{ij}\, h^j\, y^{(j)}(x+ih) \qquad (1.7)$$

Définitions

1° Une méthode $(k,\ell)$ a *l'ordre de consistance* $q$ si et seulement si pour tout $y \in \mathcal{C}^{q+1}$, on a :

$$L\,[y(x),h] = C_{q+1}\, h^{q+1}\, y^{(q+1)}(x) + O(h^{q+2}) \qquad (1.8)$$

où $C_{q+1}$ est une constante non nulle.
Une méthode $(k,\ell)$ est *consistante* si et seulement si $q \geq 1$.

2° Si $\xi = 1$ est racine de multiplicité m du polynôme $\rho_o(\xi)$, nous définirons *l'ordre d'erreur* $p$ de la méthode comme étant :

$$p = q - m + 1 \qquad (1.9)$$

Si m = 1,

$$c_{p+1} = \frac{C_{p+1}}{\rho_1(1)} \quad \text{est appelé } \textit{la constante} \qquad (1.10)$$

*d'erreur.*

Remarque

Si m = 1, alors l'ordre d'erreur et l'ordre de consistance coïncident.

## 1.3. METHODES ASYMPTOTIQUEMENT STABLES

Soient $\xi_i$ les racines de $\rho_0(\xi)$.
Une méthode $(k,\ell)$ est *asymptotiquement stable* si et seulement si

$$|\xi_i| \leq 1 \qquad (1.11)$$

et si $|\xi_i| = 1$ , alors la multiplicité de cette racine vaut 1.

Note : souvent nous employerons le mot "stable" à la place "asymptotiquement stable".

## 1.4. CONVERGENCE DES METHODES A PAS ET DERIVEES MULTIPLES

Définition

Une méthode $(k,\ell)$ est *convergente* si , pour toute équation différentielle du type (1.1), (1.2) où y(x) est la solution exacte et où $f^{(j)}(x,y)$, pour $j = 0,1,\ldots,\ell-1$, satisfait à la condition de Lipschitz en la seconde variable, on a

$$\lim_{\substack{h\to 0 \\ nh=x-a}} y_n = y(x) \qquad \text{pour tout } x \in [a,b\,].$$

En outre, il faut que toutes les solutions $\{y_n\}$ de l'équation aux différences (1.3), avec les conditions initiales $y_i = \eta_i(h)$, $i=0,1,\ldots,k-1$, satisfassent aux conditions suivantes :

$$\lim_{h\to o} \eta_i(h) = \eta \qquad \text{pour } i=0,1,\ldots,k-1 \text{ et } h \in [\,o,h^*]$$

Théorème 1.2

*Une condition nécessaire et suffisante pour qu'une méthode (k-ℓ) soit convergente est qu'elle soit consistante et asymptotiquement stable.*

## 1.5. DIAGRAMME DE PUISEUX

Définition

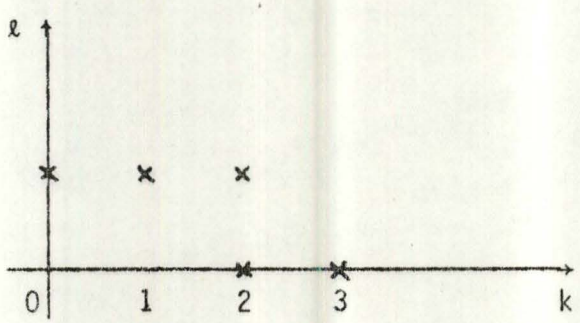
Soit une méthode $(k,\ell)$ donnée par (1.4).
Considérons l'ensemble $P = \{(i,j) \mid \alpha_{ij} \neq 0\}$.
On appelle *diagramme de Puiseux* de la méthode, le graphe de P dans $\mathbb{R}^2$.

Exemple

Soit la méthode d'Adams-Bashforth donnée par la relation :

$$y_{n+1} - y_n = \frac{h}{12} (23\, f_n - 16\, f_{n-1} + 5\, f_{n-2}).$$

Le diagramme de Puiseux de cette méthode a la forme suivante :




## 1.6. GENERALISATION

Des problèmes physiques demandent parfois d'intégrer sur un intervalle $[a,b]$, une équation différentielle d'ordre r :

$$y^{(r)} = f(x,y) \qquad \text{où} \quad r \in \mathbb{N} \tag{1.12}$$

satisfaisant à r conditions initiales :

$$y^{(s)}(a) = n_s \qquad \text{pour } s = 0,1,\ldots,r-1$$

La méthode $(k,\ell)$ que nous utiliserons pour résoudre un tel problème sera fournie par l'opérateur suivant :

$$L[y(x),h] = \sum_{i=0}^{k} \alpha_{i0}\, y(x+ih) + h^r \sum_{i=0}^{k} \sum_{j=1}^{\ell} h^{j-1} \alpha_{ij}\, y^{(r+j-1)}_{(x+ih)} \tag{1.13}$$

La *stabilité asymptotique* d'une méthode $(k,\ell)$ définie en (1.13) se généralise de la manière suivante :

soient $\xi_i$ les racines de $\rho_0(\xi)$

$$|\xi_i| \leq 1 \qquad (1.14)$$

et si $|\xi_i| = 1$ , alors la multiplicité de cette racine est au plus r.

Notons qu'en posant r=1, (1.13) se ramène à (1.1) et (1.14) à (1.11).

o

o o

# CHAPITRE II

## ORDRE DE CONSISTANCE MAXIMUM D'UNE MÉTHODE À PAS ET DÉRIVÉES MULTIPLES

## 2.1. INTRODUCTION

Ce chapitre est consacré à l'étude de l'ordre de consistance maximum des méthodes à pas et dérivées multiples et à l'étude des méthodes atteignant cet ordre, sans tenir compte des contraintes de stabilité.

Soit $J = \{(i,j) \mid 0 \leq i \leq K \quad \text{et} \quad 0 \leq j \leq r-1\}$.

On se donne une méthode contenant au plus r coefficients non nuls, $\alpha_{ij}$ où $(i,j) \in I$, un ensemble de couples ordonnés, inclus dans J, et on recherche l'ordre de consistance maximum que peut atteindre une telle méthode.

Le lien entre ce problème et le problème d'interpolation H-B, signalé à la section 2.2, est exploité à la section 2.3, afin de déterminer les conditions sous lesquelles on peut effectivement trouver l'ordre maximum.

Dans la section 2.4, nous chercherons des critères d'existence et d'unicité d'une méthode atteignant l'ordre de consistance maximum.

Enfin, la section 2.5 comporte l'étude des méthodes dites d'interpolation d'Hermite.

## 2.2. POSITION DU PROBLEME

Considérons l'opérateur

$$L[y(x), h] = \sum_{(i,j)\in I} \alpha_{ij} h^j y^{(j)}(x + ih) \qquad (2.0)$$

où I est l'ensemble de couples ordonnés fixé et
où $y(x)$ est de classe $C^\infty$.

En développant $y^{(j)}(x + ih)$ au voisinage de x, il prend la forme suivante :

$$L[y(x), h] = \sum_{m\in IN} h^m y^{(m)}(x) C_m \qquad (2.1)$$

où $C_m = \sum_{(i,j)\in I} \alpha_{ij} \frac{i^{m-j}}{(m-j)!}$

Par définition, la méthode à pas et dérivées multiples associée à cet opérateur, a l'ordre de consistance p si et seulement si

$$\sum_{(i,j)\in I} \alpha_{ij} \frac{i^{m-j}}{(m-j)!} = 0 \qquad \forall\, m \in \bar{p}\ (+) \tag{2.2}$$

L'ordre p peut être atteint si ce système homogène admet une solution non triviale.
L'ordre de consistance maximum sera donc déterminé par les conditions d'existence d'une solution non triviale.
Les équations (2.2) peuvent s'écrire :

$$\sum_{(i,j)\in I} \alpha_{ij} \frac{m!}{(m-j)!}\, i^{m-j} = 0 \qquad \forall\, m \in \bar{p}$$

Considérons le système transposé.
Celui-ci s'écrit :

$$\sum_{m=0}^{p} a_m \frac{m!}{(m-j)!}\, i^{m-j} = 0 \qquad \forall\, (i,j) \in I$$

ou encore

$$q^{(j)}(i) = 0 \qquad \forall\, (i,j) \in I \tag{2.3}$$

$$\text{où } q(x) = \sum_{m=0}^{p} a_m x^m$$

Le problème se situe donc dans la recherche d'un polynôme q(x) non nul, de degré au plus p, vérifiant les égalités (2.3). Ce problème a été étudié par Karlin-Karon [ 26 ], Ferguson [ 17 ], Atkinson-Sharma [ 2 ] et Lorentz [ 30 ].
Il s'agit de l'interpolation d'Hermite-Birkhoff.
Notre intérêt se situant dans la recherche d'une solution non triviale, la section 2.3 contient quelques résultats qui nous mèneront à un critère d'existence et d'unicité d'une méthode d'ordre p.

---

(+) $\bar{p} = \{0,1,2,\ldots,p\}$.

## 2.3. RECHERCHE DE L'ORDRE DE CONSISTANCE MAXIMUM

Le problème d'interpolation d'Hermite-Birkhoff s'énonce de manière générale, de la façon suivante.

On se donne $\{x_i\}_{i=1}^{k} \subset \mathbb{R}$, tel que $x_1 < \ldots < x_k$

et n nombres $f_i^{(j)}$ où $(i,j) \in I \subseteq \{(i,j) \mid 1 \leqslant i \leqslant k \text{ et } 0 \leqslant j \leqslant n-1\}$

Sous quelle(s) condition(s) existe-t-il un polynôme unique $p(x) \in \pi_{n-1}$ (+) satisfaisant aux conditions :

$$p^{(j)}(x_i) = f_i^{(j)} \qquad \forall\ (i,j) \in I. \tag{2.4}$$

Le polynôme d'interpolation est unique ssi seul, le polynôme trivial appartenant à $\pi_{n-1}$, satisfait aux équations (2.4) avec $f_i^{(j)} = 0$, $\forall\ (i,j) \in I$.

On peut caractériser cette unicité, en introduisant une matrice associée au problème.

A I, on peut faire correspondre une et une seule matrice

$$E = (e_{ij})_{i=1,j=o}^{k \quad n-1} \tag{2.4bis}$$

où $e_{ij} = 1$ si $(i,j) \in I$

$= 0$ si $(i,j) \notin I$

Remarquons que cette matrice a exactement n colonnes et n éléments non nuls. On l'appelle une matrice n-incidente.

Définition

La matrice E est dite "order poised" si seul le polynôme trivial $p(x) \equiv 0$ dans $\pi_{n-1}$, satisfait à (2.4) avec $f_i^{(j)} = 0$ pour tout $(i,j) \in I$ et cela pour tout $\{x_i\}_{i=1}^{k} \subseteq \mathbb{R}$ tel que $x_1 < x_2 < \ldots < x_k$.

Reprenons le problème d'interpolation de la section précédente (2.3). Nous pouvons lui associer une matrice r-incidente : A à K+1 lignes (car $0 \leqslant i \leqslant k$) et r colonnes.

---

(+) $\pi_{n-1}$ est l'ensemble des polynômes à coefficients réels, de degré au plus n-1.

Il est évident que si la matrice A est order poised, alors l'ordre p = r-1 ne pourra être atteint car seul le polynôme trivial sera solution du problème d'interpolation (2.3) et, par suite, seule la méthode triviale sera solution du système homogène de départ.
Il semble donc naturel de rechercher l'ordre maximum d'une méthdde, I étant fixé, en caractérisant la matrice incidente associée à I.
Le théorème suivant se déduit facilement des remarques précédentes.

Théorème 2.1 [25]

*Soient k et r deux entiers positifs*
*I un ensemble de r+2 couples ordonnés*
*où* $0 \leq j \leq r$
$0 \leq i \leq k$

*Si la matrice (r+2)-incidente P associée à l'ensemble I est une matrice order poised, alors il n'existe pas de méthodes à pas et dérivées multiples de la forme (2.1) non triviales avec un ordre de consistance* $q > r$.

En effet, supposons qu'il existe une telle méthode.
Si $q \geq r+1$, alors

$$C_0 = C_1 = \ldots = C_{r+1} = 0.$$

Le système comporte donc r+2 équations à (r+2) inconnues :

$$\alpha_{ij} \quad \text{où } (i,j) \in I.$$

En le transposant, le problème d'interpolation H-B obtenu s'écrit :

$$q^{(j)}(i) = 0 \qquad \forall\ (i,j) \in I$$

où q(x) est un polynôme de degré au plus r+1.
La matrice (r+2)-incidente associée à I étant order poised, seul le polynôme trivial répond à ce problème d'interpolation et donc seule la méthode triviale peut atteindre l'ordre r+1. ∎

Le critère d'"order poised" tel qu'il a été défini précédemment est relativement lourd à mettre en oeuvre. Cependant, d'autres critères, plus simples à utiliser, ont été trouvés par les auteurs déjà cités préalablement.
Ils nécessitent l'introduction de quelques définitions.

Définition d'un bloc supporté [ 25 ] , [ 30 ]

Soit E une matrice n-incidente.

- Un bloc de E est une suite maximale de 1 dans une ligne de E, ne commençant pas en j = 0.

  Exemple : $e_{i,j-1} = 0$ , $e_{ij} = e_{i,j+1} = \dots = e_{i,j+l-1} = 1$.

- Si l est impair, on dira que le bloc est impair.
- Si l est pair, on dira que le bloc est pair.
- Un bloc est supporté si

$$\exists\ (i_1,j_1),\ (i_2,j_2) \qquad \text{tels que} \quad \begin{matrix} i_1 < i < i_2 \\ j_1 < j \\ j_2 < j \end{matrix}$$

$$\text{et } e_{i_1 j_1} = e_{i_2 j_2} = 1$$

Exemples

1. Si $E_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$

$E_1$ a un bloc impair supporté. Il est formé du seul élément $e_{11}$.

2. Si

$$E_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$E_2$ a deux blocs pairs. L'un est supporté (à la 3ème ligne); l'autre pas.
$E_2$ a un seul bloc impair non supporté.

Définition des constantes de Polya [26 ]

Soit E une matrice n-incidente : $(e_{ij})_{i=1,j=o}^{k \quad n-1}$

Posons

$$m_j = \sum_{i=1}^{k} e_{ij}$$

On a que

$$M_j = \sum_{\nu=0}^{j} m_\nu$$

$M_j$ (j = 0, ..., n-1) sont les constantes de Polya.

Signification

Si p(x) est un polynôme vérifiant les relations (2.4)

- $m_j$ est le nombre de conditions requises sur la $j^e$ dérivée du polynôme
- $M_j$ est le nombre de conditions requises sur les $(j+1)^{es}$ premières dérivées du polynôme.

Définition des conditions de Polya [ 26 ]

Une matrice n-incidente E vérifie

- les conditions faibles de Polya si $M_j \geq j+1$ $\quad j=0,1,\ldots,n-2$
- les conditions fortes de Polya si $M_j \geq j+2$ $\quad j=0,1,\ldots,n-2$

Exemples

1. Reprenons les deux exemples précédents :

   $E_1$ et $E_2$ vérifient les conditions de Polya.

2. Soit

$$E = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

E ne vérifie pas les conditions de Polya.

Le théorème suivant est une condition suffisante d'"order poised" :

Théorème 2.2 [26] , [2]

*Une matrice n-incidente E satisfaisant aux conditions faibles de Polya est "order poised" si elle ne contient pas de blocs supportés impairs.*

Théorème 2.3 [ 30 ] , [ 26 ]

*Soit E une matrice n-incidente satisfaisant aux conditions fortes de Polya.*

*Si une ligne de E contient exactement un bloc supporté impair, alors E n'est pas "order poised".*

Remarques

1. Si on ajoute ou si on retire une ligne de zéros à une matrice incidente E, on ne change pas le problème d'interpolation.
   La matrice obtenue est équivalente à E. [ 17 ]
   Remarquons que les conditions de caractérisation d'"order poised" ne sont affectées d'aucun changement si l'on retire ou non une ligne de zéros à la matrice incidente.

2. Dans le théorème 2.1, on considère $0 \leq j \leq r$.
   Par contre, la matrice (r+2)-incidente associée à I, a exactement r+2 colonnes. La dernière colonne de cette matrice est donc composée uniquement de zéros et une colonne au moins de cette matrice comportera deux 1.
   La méthode d'ordre de consistance maximum, associée à I, aura donc bien au minimum un pas.
   D'autre part, si on admettait une dérivée $(r+1)^e$ dans la formule de récurrence, le coefficient $\alpha_{i',r+1}$ y correspondant ne contribuerait en aucune façon à atteindre l'ordre r,
   car
   $$L[y(x),h] = \sum_{(i,j)\in I\setminus\{(i',r+1)\}} \alpha_{ij}\, h^j\, y^{(j)}(x + ih) + \alpha_{i',r+1}\, h^{r+1}\, y^{(r+1)}(x + i'h)$$
   $$= C_{r+1}\, h^{r+1}\, y^{(r+1)}(x) + O(h^{r+2})$$
   Si le système homogène
   $$C_0 = C_1 = \ldots = C_r = 0$$
   admettait une solution non triviale, celle-ci ne serait donc pas unique (car seul un coefficient peut être normalisé).

Applications :

1. Supposons que la méthode ait le diagramme de Puiseux représenté à la figure 2.1

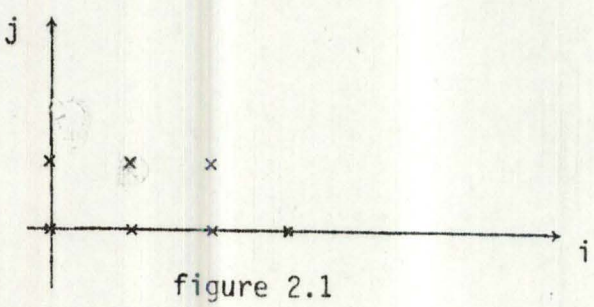



figure 2.1

La matrice P associée à cette méthode est

$$P = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Le théorème 2.2 permet d'affirmer que P est "order poised".
Par le théorème 2.1, on sait donc que l'ordre de consistance maximal d'une méthode ayant ce diagramme de Puiseux est 5.
Si on résout le système

$$C_0 = C_1 = \ldots = C_5 = 0$$

en fixant, par exemple, $\alpha_{30} = 1$, on obtient la méthode suivante :

$$y_{n+3} + 18\, y_{n+2} - 9\, y_{n+1} - 10\, y_n - h(9\, f_{n+2} + 18\, f_{n+1} + 3\, f_n) = 0$$

Remarquons que si on pose $\alpha_{30} = \lambda \neq 0$, on obtient une méthode identique.

2. Il se peut qu'un couple (i,j) figure dans l'ensemble I et que le coefficient $\alpha_{ij}$ correspondant à ce couple, s'annule.
Le cas se présente dans l'exemple suivant.
Considérons la matrice incidente

$$P = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Cette matrice étant "order poised", l'ordre de consistance de la méthode associée à P, ne peut pas dépasser 4.
On construit la méthode, en résolvant le système

$$C_0 = C_1 = \ldots = C_4 = 0$$

en fixant $\alpha_{20} = 1$.

On peut remarquer alors que $\alpha_{10} = 0$ et que la méthode obtenue est celle de Milne :

$$y_{n+2} - y_n = \frac{1}{3} h(f_{n+2} + 4\, f_{n+1} + f_n)$$

Si par contre, nous avions considéré

$$P' = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

cette matrice n'étant pas order poised, le théorème 2.1 n'est pas applicable.

Rien ne dit que l'ordre de consistance de la méthode associée à P' ne dépasse pas 3. En effet, la résolution ci-dessus nous indique que l'ordre de cette méthode vaut 4.

3. Pour la même raison que celle évoquée dans l'exemple précédent, il se peut que si $\tau = \max \{j \mid (i,j) \in I\}$

$\ell = \max \{j \mid \alpha_{ij} \neq 0\}$

où $\alpha_{ij}$ sont les coefficients de la méthode d'ordre de consistance maximal, on ait $\tau \neq \ell$.

Soit

$$P = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

La matrice P est order poised [29].

L'ordre de consistance maximum est donc 4.

La résolution du système indique qu'il s'agit également de la méthode de Milne.

Le coefficient $\alpha_{14} = 0$ et donc $\ell \neq \tau$

En conclusion, si I et k sont fixés et si la matrice incidente associée à I est order poised, le théorème 2.1 nous donne la borne maximale de l'ordre de consistance que l'on peut atteindre. La question qui se pose immédiatement est de savoir sous quelle(s) condition(s), il existe une et une seule méthode atteignant cet ordre maximum.

Ce problème est abordé dans la section 2.4.

## 2.4. EXISTENCE ET UNICITE D'UNE METHODE D'ORDRE DE CONSISTANCE MAXIMUM

Notation : Soit I un ensemble de r+2 couples ordonnés.

Si P est la matrice (r+2)-incidente associée à I, on note $P^{(\mu,\nu)}$, la matrice (r+1)-incidente associée à $I\setminus\{(\mu,\nu)\}$.

Remarquons que la dernière colonne de $P^{(\mu,\nu)}$ ne sera pas nécessairement une colonne de zéros.

Par un raisonnement analogue à celui du paragraphe 2.3, nous pouvons démontrer le critère d'existence et d'unicité.

Théorème 2.4 [ 25 ]

*Soient $k$ et $r$ deux entiers positifs*

*$I$ un ensemble de $(r+2)$ paires ordonnées $(i,j)$*

*où $0 \leq i \leq k$*

*$0 \leq j \leq r$*

*On suppose que $P^{(\mu,\nu)}$ est order poised, où $(\mu,\nu) \in I$*

*alors (i) il existe une et une seule méthode non triviale avec $q \geq r$ et*

$$\alpha_{\mu\nu} \neq 0$$

*(ii) si $P^{(\bar{\mu},\bar{\nu})}$ est order poised, $(\bar{\mu},\bar{\nu}) \in I$, alors $\alpha_{\bar{\mu}\bar{\nu}} \neq 0$*

*(iii) si $P$ est order poised, alors il existe une et une seule méthode non triviale avec $q = r$.*

Démonstration

(i) Si on impose un ordre de consistance supérieur ou égal à r, les relations suivantes sont vérifiées

$$C_0 = C_1 = \ldots = C_r = 0 \qquad (2.5)$$

D'autre part, si on fixe $\alpha_{\mu\nu}$ à 1, le système (2.5) s'écrit

$$\sum_{(i,j)\in I\setminus\{(\mu,\nu)\}} \alpha_{ij} \frac{i^{\tau-j}}{(\tau-j)!} = -\frac{\mu^{\tau-\nu}}{(\tau-\nu)!} \qquad (2.6)$$

où $\tau \in \bar{F}$.

Ce système non homogène admet une et une seule solution si son déterminant est non nul ou de manière équivalente, si le déterminant du système homogène transposé est non nul.

Ce dernier s'écrit :

$$q^{(j)}(i) = 0 \quad \forall\ (i,j) \in I\setminus\{(\mu,\nu)\} \quad \text{où } q(x) \in \pi_r. \qquad (2.7)$$

La matrice incidente correspondant à ce problème d'interpolation, $P^{(\mu,\nu)}$ est order poised.

Donc, seul le polynôme trivial est solution de (2.7) et le déterminant du système est non nul, ce qui démontre l'assertion (i).

(ii) Si on note par $C^{(\mu\nu)}$, la matrice des coefficients de $\alpha_{ij}$ $(i,j) \neq (\mu,\nu)$

$\underline{u}$, le vecteur formé des $\alpha_{ij}$ $(i,j) \neq (\mu,\nu)$

et $\underline{d}$, le vecteur formé des coefficients de $\alpha_{\mu\nu}$, changés de signe,

le système (2.6) peut s'écrire sous la forme plus compacte

$$C^{(\mu\nu)} . \underline{u} = \underline{d} \qquad (2.8)$$

Nous avons démontré précédemment que si $P^{(\mu,\nu)}$ est order poised, alors $\det(C^{(\mu\nu)})$ est non nul.

Il en est de même du $\det(C^{(\bar{\mu}\bar{\nu})}$ si $P^{(\bar{\mu}\bar{\nu})}$ est order poised.

La règle de Cramer appliquée à (2.8) permet de trouver $\alpha_{\bar{\mu}\bar{\nu}}$.

$$\alpha_{\bar{\mu}\bar{\nu}} = \pm \frac{\det(C^{(\bar{\mu},\bar{\nu})})}{\det(C^{(\mu,\nu)})}$$

Ces déterminants étant non nuls, il en est de même de $\alpha_{\bar{\mu}\bar{\nu}}$ .

(iii) Cette assertion découle immédiatement du théorème 2.1 et de l'assertion (i).

■

Corollaire

*Soient $k$ et $r$ deux entiers positifs*

*$I$ un ensemble de $(r+2)$ paires ordonnées $(i,j)$*

*où $0 \leq i \leq k$*

*$0 \leq j \leq r$*

*Une condition suffisante pour qu'il existe une et une seule méthode non triviale, d'ordre de consistance maximum est que*

*1° $P$ soit order poised, où $P$ est la matrice $(r+2)$-incidente associée à $I$*

*2° $\exists$ $(\mu,\nu)$ tel que $P^{(\mu,\nu)}$ soit order poised.*

Ce corollaire découle immédiatement du théorème précédent.

Sous les conditions décrites ci-dessus, l'existence et l'unicité d'une méthode atteignant l'ordre maximum r, sont assurées. L'opérateur associé à cette méthode aura donc la forme :

$$L[y(x), h] = C_{r+1} h^{r+1} y^{(r+1)}(x) + O(h^{r+2}) \qquad (2.9)$$

$$\forall\, y(x) \in \mathcal{C}^{r+1}$$

où $C_{r+1}$ est une constante.

Il est parfois utile de connaître la constante d'erreur d'une méthode (1.10), si on désire comparer deux méthodes, par exemple. Pour cette raison, nous allons rechercher l'expression de la constante $C_{r+1}$ dans (2.9).

Auparavant, introduisons un nouvel élément du problème d'interpolation H-B, qui simplifiera l'expression de cette constante.

Considérons le problème d'interpolation H-B :

$$p^{(j)}(x_i) = 0 \qquad \forall\, (i,j) \in I \qquad (2.10)$$

$$\text{où } p(x) = \sum_{m=0}^{n-1} a_m x^m$$

Soit E, la matrice n-incidente associée à I.
Les conditions d'interpolation (2.10) peuvent s'écrire :

$$\sum_{m=0}^{n-1} a_m \frac{x_i^{m-j}}{(m-j)!} = 0 \qquad \forall\, (i,j) \in I$$

soit

$$D(E) \cdot \underline{a} = 0$$

où $\underline{a} = [a_0 \;\; a_1 \;\; \dots \;\; a_{n-1}]^T$

et où la ligne de D(E) correspondant au couple (i,j) de I est donnée par

$$[\underbrace{0 \dots 0}_{j \text{ fois}} \quad 1 \quad x_i \quad \frac{x_i^2}{2!} \cdots \frac{x_i^{n-1-j}}{(n-1-j)!}]$$

Les lignes de D(E) sont rangées suivant l'ordre lexicographique : la ligne correspondant à (i,j) précède celle correspondant à (i',j') si

$$i < i'$$

ou

$$i = i' \quad \text{et} \quad j < j'$$

Si on note par K(E), le déterminant de la matrice D(E), on a le théorème suivant :

Théorème 2.5 [ 25 ], [ 26 ]

*Soit E une matrice incidente, satisfaisant les conditions faibles de Polya.*
*On suppose que E n'a pas de blocs impairs dans ses lignes intérieures.*
*Soient $j_0, j_1, \ldots, j_\ell$, les indices pour lesquels $\alpha_{0 j_s} = 1$.*

*Alors*

$$K(E) \quad (-1)^{\sum_{s=0}^{\ell} (j_s - s)} > 0$$

L'expression de la constante $C_{r+1}$ dans (2.9) est donnée par le théorème suivant.

Théorème 2.6 [ 25 ]

*Soient k et r deux entiers positifs.*
*I un ensemble de (r+2) paires ordonnées (i,j)*
*où $0 \leq j \leq r$*
*$0 \leq i \leq k$.*

*On suppose que P, matrice incidente associée à I, et $P^{(\mu,\nu)}$ sont order poised où $(\mu,\nu) \in I$.*
*Alors, l'opérateur associé à la méthode unique d'ordre de consistance r, construite sur I, s'écrit :*

$$L[y(x),h] = C_{r+1}\, h^{r+1}\, y^{(r+1)}(x) + o(h^{r+2}) \quad \forall\, y(x) \in \mathcal{C}^{r+1}$$

$$\text{où } C_{r+1} = (-1)^{\tau}\, \frac{1}{(r+1)!}\, \frac{K(P)}{K(P^{(\mu,\nu)})}.$$

*$\tau$ est le nombre de couples dans I, strictement supérieurs lexicographiquement à $(\mu,\nu)$.*

Démonstration

Par (2.1), on a

$$C_{r+1} = \sum_{(i,j) \in I} \alpha_{ij}\, \frac{i^{r+1-j}}{(r+1-j)!}$$

Donc

$$(r+1)!\, C_{r+1} = \sum_{(i,j) \in I} \alpha_{ij}\, \frac{(r+1)!}{(r+1-j)!}\, i^{r+1-j} \qquad (2.11)$$

Normalisons le coefficient $\alpha_{\mu\nu}$ : $\alpha_{\mu\nu} = 1$

(2.11) s'écrit alors

$$\sum_{(i,j)\in \Lambda\setminus\{(\mu,\nu)\}} \alpha_{ij} \frac{(r+1)!}{(r+1-j)!}\, i^{r+1-j} - C_{r+1}(r+1)! = -\frac{(r+1)!}{(r+1-\nu)!}\, u^{r+1-\nu}$$

On ajoute cette équation au système (2.8), avec $C_{r+1}$ comme inconnue. Matriciellement, on obtient le système :

$$\left(\begin{array}{ccc|c} & C^{(\mu,\nu)} & & \begin{matrix}0\\ \vdots \\ 0\end{matrix} \\ \hline e_1 & e_2 \ \cdots & e_{r+1} & 1 \end{array}\right) \left(\begin{matrix} u_1 \\ \vdots \\ u_{r+1} \\ -(r+1)!\ C_{r+1}\end{matrix}\right) = \left(\begin{matrix} d_1 \\ \vdots \\ d_{r+1} \\ d_{r+2}\end{matrix}\right)$$

où $\begin{cases} e_s = \dfrac{(r+1)!}{(r+1-j)!}\, i^{r+1-j} & (i,j) \neq (\mu,\nu) \\ d_{r+2} = -\dfrac{(r+1)!}{(r+1-\nu)!}\, u^{r+1-\nu} \end{cases}$

On recherche ensuite $C_{r+1}$ par la règle de Cramer.

$$-(r+1)!\ C_{r+1} = \frac{\det \left(\begin{array}{ccc|c} & C^{(\mu,\nu)} & & \begin{matrix}d_1\\ \vdots \\ d_{r+1}\end{matrix} \\ \hline e_1 & \cdots\cdots & e_{r+1} & d_{r+2} \end{array}\right)}{\det\left(C^{(\mu,\nu)}\right)}$$

soit

$$C_{r+1} = \frac{1}{(r+1)!} \cdot \frac{\det \left(\begin{array}{ccc|c} & C^{(\mu,\nu)} & & \begin{matrix}-d_1\\ \vdots \\ -d_{r+1}\end{matrix} \\ \hline e_1 & \cdots\cdots & e_{r+1} & -d_{r+2} \end{array}\right)}{\det\left({}^{\tau}C^{(\mu,\nu)}\right)}$$

Si on permute la dernière colonne de la matrice apparaissant au numérateur avec les $\tau$ colonnes qui la précèdent, la matrice obtenue transposée sera celle du problème d'interpolation H-B.

$$q^{(j)}(i) = 0 \qquad \forall\ (i,j) \in I$$

$$\text{où } q(x) \in \pi_{r+1}$$

La matrice incidente associée à ce problème est P.

Le déterminant apparaissant au numérateur sera donc

$$K(P)\ (-1)^{\tau}$$

Celui du dénominateur sera $K(P^{(\mu,\nu)})$ car ${}^{\tau}C^{(\mu,\nu)}$ est la matrice du problème d'interpolation H-B correspondant à $P^{(\mu,\nu)}$.

On obtient donc l'expression de $C_{r+1}$ :

$$\boxed{C_{r+1} = \frac{1}{(r+1)!}\ (-1)^{\tau}\ \frac{K(P)}{K(P^{(\mu,\nu)})}}$$

Remarque

On a posé $\alpha_{\mu\nu} = 1$.

Si ce n'est pas le cas, on obtient alors

$$\boxed{C_{r+1} = \frac{1}{(r+1)!}\ (-1)^{\tau}\ \frac{K(P)}{K(P^{(\mu,\nu)})} \cdot \alpha_{\mu\nu}}$$

En effet, le facteur $\alpha_{\mu\nu}$ intervient également dans le vecteur $\underline{d}$.

Applications

Reprenons les exemples cités à la section 2.3..

1. Soit

$$P = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

P et $P^{(o,o)}$ étant order poised, le corollaire du théorème 2.4 affirme l'existence et l'unicité d'une méthode d'ordre de consistance q = 5.

2. Soit

$$P = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

P et $P^{(2,0)}$ sont order poised. Il n'existe donc qu'une seule méthode non triviale d'ordre de consistance maximum :

$$q = 4.$$

Le théorème 2.4 assure la non-nullité de $\alpha_{20}$.
Remarquons que $P^{(1,0)}$ n'est pas order poised.
On ne peut donc rien conclure quant à $\alpha_{10}$.

3. Considérons la matrice 9-incidente

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

P est order poised.
D'autre part, $P^{(\mu,\nu)}$ est order poised $\quad \forall\ (\mu,\nu) \in I \setminus \{(1,1),(3,0)\}$.
Par le théorème 2.4, on a donc

$$\alpha_{\mu\nu} \neq 0 \qquad \forall\ (\mu,\nu) \in I \setminus \{(1,1),(3,0)\}$$

et l'ordre de consistance de la méthode est 7.
Le diagramme de Puiseux associé à cette méthode unique aura donc la forme, représentée à la figure 2.2.

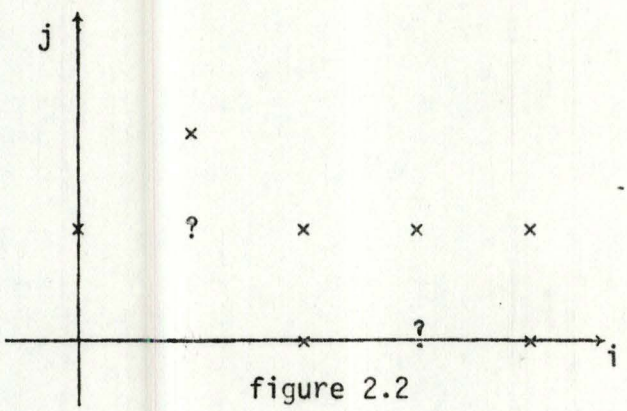



figure 2.2

Par ailleurs, le théorème 2.6 nous assure que

$$C_8 = (-1)^{\tau} \frac{1}{8!} \frac{K(P)}{K(P^{(4,0)})} \cdot \alpha_{40}$$

$\tau$ prend ici la valeur 1.

En appliquant le théorème 2.5, on déduit que

$$\text{sign}\left(\frac{K(P)}{K(P^{(4,0)})}\right) = +1.$$

Nous pouvons donc conclure que

$$\text{sign } C_8 = -\text{ sign } \alpha_{40}.$$

## 2.5. METHODES D'INTERPOLATION D'HERMITE

Dans cette section, nous considérerons un ensemble de couples particulier :

$$I = \{(i,j) \mid 0 \leq i \leq k \text{ et } 0 \leq j \leq \ell_i \quad \forall\, i \in \bar{k}\}$$

où les $\ell_i$ sont des entiers non négatifs $\forall\, i \in \bar{k}$.

Soit $\ell = \max\{\ell_i \mid i \in \bar{k}\}$

La formule de récurrence associée à ces méthodes a la forme générale définie au chapitre I. (1-3).

### 2.5.1. Définition

Le théorème 2.1 nous permet d'affirmer immédiatement :

Théorème 2.7

*La méthode $(k,\ell)$ de la forme*

$$\sum_{i=0}^{k} \alpha_{i0}\, y_{n+i} + \sum_{i=0}^{k} \sum_{j=1}^{\ell_i} h^j \alpha_{ij}\, f_{n+i}^{(j-1)} = 0 \qquad (2.12)$$

*a un ordre de consistance* $q \leq \sum_{i=0}^{k} \ell_i + k - 1$

En effet, l'ensemble I correspondant à cette méthode contient $\left(\sum_{i=0}^{k} \ell_i + k + 1\right)$ couples et la matrice incidente associée à cet ensemble I est order poised.

Remarque

Si $\ell_i = 1 \quad \forall\, i \in \bar{k}$, il s'agit alors de méthodes à k pas et une dérivée. L'ordre de consistance maximum pouvant être atteint est 2 k [21].

Le théorème 2.7 affirme également que l'ordre de consistance maximum est $\sum_{i=0}^{k} \ell_i + k-1 = 2k$.

Définition

La méthode $(k,\ell)$ de la forme (2.12) est appelée interpolation d'Hermite si son ordre de consistance est exactement $\sum_{i=0}^{k} \ell_i + k-1$.

Si les $\ell_i$ sont donnés ($\forall\, i \in \bar{k}$), une telle méthode existe et est unique, comme le prouve le théorème suivant :

Théorème 2.8 [ 25 ]

*Soient* $k$ *un entier positif*

$\ell_0, \ell_1, \ldots, \ell_k$ *des entiers positifs, tels que*

$$\max \{\ell_i \mid i \in \bar{k}\} = \ell \geq 1.$$

*Alors* (i) *il existe une et une seule méthode* $(k,\ell)$ *d'interpolation d'Hermite avec les* $\ell_i$ *donnés*

(ii) $\operatorname{sign} C_{q+1} = (-1)^{\ell_k} \operatorname{sign} \alpha_{k_0}$

Démonstration

(i) Soit P, la matrice incidente associée à l'ensemble I.
Le théorème 2.2 de la section 2.3, nous assure que P et $P^{(k,0)}$ sont order poised. Par le corollaire du théorème 2.4, ces conditions sont suffisantes pour assurer l'existence et l'unicité d'une méthode non triviale d'ordre de consistance maximum.

(ii) Le théorème 2.6 affirme que

$$C_{q+1} = (-1)^{\ell_k} \frac{K(P)}{K(P^{(k,0)})} \frac{1}{(q+1)!} \alpha_{k_0}$$

On en déduit que

$$\operatorname{sign} C_{q+1} = (-1)^{\ell_k} \operatorname{sign} \left( \frac{K(P)}{K(P^{(k,0)})} \right) \operatorname{sign} \alpha_{k_0} \qquad (2.13)$$

Or, par le théorème 2.5, on a que

$$K(P) > 0$$

$$K(P^{(k,0)}) > 0$$

L'égalité (2.13) s'écrit alors

$$\text{sign } C_{q+1} = (-1)^{\ell_k} \text{ sign } \alpha_{k_o} \quad .$$

La méthode optimale peut s'exprimer sous une forme particulière, appelée multigradient car elle généralise la notion de bigradient et trigradient [ 6 ]. Cette idée est développée à l'appendice A.

## 2.5.2. Propriété des coefficients

Utilisant les résultats précédents, il est possible de déterminer, sans les calculer, quels sont les coefficients non nuls d'une méthode d'interpolation d'Hermite, ainsi que leurs signes en fonction de celui de $\alpha_{k_o}$.
Le théorème 2.9 rassemble ces propriétés.

Théorème 2.9 [ 25 ]

*Soient* $\alpha_{ij}$, $\begin{cases} 0 \leq i \leq k \\ 0 \leq j \leq \ell_i \end{cases}$ *les coefficients d'une méthode d'interpolation d'Hermite.*

*On a* (i) $\left.\begin{array}{l} \alpha_{kj}\ ,\ j \in \bar{\ell}_k \\ \alpha_{oj}\ ,\ j \in \bar{\ell}_o \\ \alpha_{i,\ell_i-2t}\ ,\ \begin{cases} t = 0,1,\ldots \left[\dfrac{\ell_i}{2}\right]^{(+)} \\ i = 1,\ldots,k-1 \end{cases} \end{array}\right\}$ *sont non nuls.*

(ii) *si* $\nu_i = \sum_{s=i+1}^{k-1} (\ell_s + 1)$

*alors*
$$\text{sign } \alpha_{kj} = (-1)^j \text{ sign } \alpha_{ko}\ ,\ j \in \bar{\ell}_k$$
$$\text{sign } \alpha_{oj} = (-1)^{\nu_o+1} \text{ sign } \alpha_{ko},\ j \in \bar{\ell}_o$$
$$\text{sign } \alpha_{i,\ell_i-2t} = (-1)^{\nu_i+1} \text{ sign } \alpha_{ko},\ t = 0,1,\ldots \left[\frac{\ell_i}{2}\right],\ i = 1,\ldots,k-1$$

---

(+) [ a ] est le plus grand entier inférieur ou égal à a.

Démonstration

Soit P, la matrice incidente associée à la méthode; P est order poised.

Recherchons les couples $(\mu,\nu)$ pour lesquels la matrice $P^{(\mu,\nu)}$ reste order poised.

- si $\mu = 0$, $P^{(0,\nu)}$ est order poised pour $\nu = 0,1,\ldots,\ell_0$.
- si $\mu = k$, $P^{(k,\nu)}$ est order poised pour $\nu = 0,1,\ldots,\ell_k$.
- si $\mu = i$ et $0 < i < k$, pour certains $j$ $(0 \leq j \leq \ell_i)$, la matrice $P^{(i,j)}$ contient un bloc supporté impair et n'est donc pas order poised (par le théorème 2.3).

  Cela se produira pour $j = \ell_i-1, \ell_i-3, \ldots, 1$ si $\ell_i$ est pair
  $\ldots, 0$ si $\ell_i$ est impair.

  Donc, $P^{(i,j)}$ ne sera order poised que si

  $$j = \ell_i - 2t \quad \text{où} \quad t = 0,1,2,\ldots\left[\frac{\ell_i}{2}\right]$$

  L'assertion (i) découle alors immédiatement de ces résultats et du théorème 2.4.

(ii) Par le théorème 2.6, nous savons que si $P^{(\mu,\nu)}$ est order poised, alors

$$C_{r+1} = \frac{(-1)^{\tau}}{(r+1)!} \frac{K(P)}{K(P^{(\mu,\nu)})} \alpha_{\mu\nu} \tag{2.14}$$

où $\tau$ est le nombre de couples $(i,j)$ strictement supérieurs lexicographiquement à $(\mu,\nu)$ dans I.

Comme $P^{(k,o)}$ est order poised, on a

$$C_{r+1} = \frac{(-1)^{\ell_k}}{(r+1)!} \frac{K(P)}{K(P^{(k,o)})} \alpha_{ko} \tag{2.15}$$

Le quotient de (2.15) et (2.14) donne l'expression de $\alpha_{\mu\nu}$ :

$$\alpha_{\mu\nu} = (-1)^{\ell_k - \tau} \frac{K(P^{(\mu,\nu)})}{K(P^{(k,o)})} \alpha_{ko}$$

On a donc

$$\operatorname{sign} \alpha_{\mu\nu} = (-1)^{\ell_k - \tau} \operatorname{sign} \frac{K(P^{(\mu,\nu)})}{K(P^{(k,o)})} \operatorname{sign} \alpha_{ko} \tag{2.16}$$

Le théorème 2.5 nous assure que

$$K(P^{(\mu,\nu)}) > 0 \quad \text{pour } \mu = k \ , \ \nu \in \bar{\ell}_k$$
$$\text{et } 0 < \mu < k \ , \ \nu = \ell_\mu - 2t$$
$$\text{où } t = 0,1,\ldots \left[\frac{\ell_\mu}{2}\right]$$

car
$$\sum_{s=0}^{\ell} (j_s - s) = \sum_{s=0}^{\ell_0} (s-s) = 0.$$

Par contre, si $\mu = 0$ et $\nu \in \bar{\ell}_0$,

$$\sum_{s=0}^{\ell} (j_s - s) = \sum_{s=0}^{\nu-1} (s-s) + \sum_{s=\nu}^{\ell_0 - 1} (s+1) - s$$

$$= \ell_0 - \nu$$

On obtient donc

$$K(P^{(0,\nu)}) \ (-1)^{\ell_0 - \nu} > 0.$$

Dès lors,

- si $\mu = k$ et $\nu \in \bar{\ell}_k$,

$$\text{sign } \alpha_{\mu\nu} = (-1)^{\ell_k - \tau} \ \text{sign } \alpha_{k0}.$$

où $\tau = \ell_k - \nu$

Il en résulte donc que

$$\boxed{\text{sign } \alpha_{k\nu} = (-1)^{\nu} \ \text{sign } \alpha_{k0}}$$

- si $0 < \mu < k$ et $\nu = \ell_\mu - 2t$
où $t = 0,1,\ldots \left[\frac{\ell_\mu}{2}\right]$

$$\text{sign } \alpha_{\mu\nu} = (-1)^{\ell_k - \tau} \ \text{sign } \alpha_{k0}$$

où $\tau = \ell_\mu - \nu + \sum_{s=\mu+1}^{k} (\ell_s + 1)$

$$= 2t + \sum_{s=\mu+1}^{k} (\ell_s + 1)$$

Il en résulte que

$$\boxed{\text{sign } \alpha_{\mu\nu} = (-1)^{\nu_\mu + 1} \ \text{sign } \alpha_{k0}}$$

- si $\mu = 0$ et $\nu \in \bar{\ell}_0$,

$$\text{sign } \alpha_{o\nu} = (-1)^{\ell_k - \tau} \; (-1)^{\ell_o - \nu} \; \text{sign } \alpha_{ko}$$

$$\text{où } \tau = \ell_o - \nu + \sum_{s=1}^{k} (\ell_s + 1)$$

Il en résulte que

$$\boxed{\text{sign } \alpha_{o\nu} = (-1)^{\nu_o + 1} \; \text{sign } \alpha_{ko}}$$

■

Remarques

1. Si $0 \leqslant i < k$, $\nu_i$ ne dépend que de i et est indépendant de l'indice de la colonne.
   Tous les coefficients que l'on sait non nuls, correspondant à une même ligne, sont donc tous du même signe, sauf pour i=k.
2. Les coefficients correspondant à l'avant-dernière ligne (i = k-1) et que l'on sait non nuls, seront du signe contraire de $\alpha_{ko}$ car $\nu_{k-1} = 0$.
3. Les coefficients de la dernière ligne, que l'on sait non nuls, ont des signes alternés.

Exemple

Si $P = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$, la matrice des coefficients se présentera comme suit :

$$\begin{pmatrix} - & - & 0 & \dots \\ ? & - & 0 & \dots \\ ? & - & 0 & \dots \\ + & 0 & 0 & \dots \end{pmatrix}$$

La méthode s'avère être

$$Y_{n+3} + 18\, y_{n+2} - 9\, y_{n+1} - 10\, y_n - h(9\, f_{n+2} + 18\, f_{n+1} + 3\, f_n) = 0$$

(cf. application 1, section 2.3).

## 2.5.3. Définition de méthodes particulières

Méthodes de Brown [25], [9]

Les méthodes de Brown sont des méthodes du type interpolation d'Hermite, avec $\ell_0 = \ell_1 = \ldots = \ell_{k-1} = 0$ et $\ell_k = \ell$.
L'ordre de consistance d'une telle méthode est donc $\ell + k-1$.
Par le théorème 2.8, on a

$$\text{sign } C_{q+1} = (-1)^{\ell} \text{ sign } \alpha_{ko}$$

et par le théorème 2.9, on peut connaître le signe des coefficients de la méthode, par rapport au signe de $\alpha_{ko}$

$$\text{sign } \alpha_{kj} = (-1)^{j} \text{ sign } \alpha_{ko}$$

$$\text{sign } \alpha_{io} = (-1)^{k-i} \text{ sign } \alpha_{ko}$$

Si $P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}$, la matrice des signes des coefficients aura la forme

$$\begin{pmatrix} - & 0 & \ldots & & \\ + & 0 & \ldots & & \\ - & 0 & \ldots & & \\ + & - & + & 0 & \ldots \end{pmatrix}$$

Le diagramme de Puiseux d'une telle méthode est représenté à la figure 2.3.

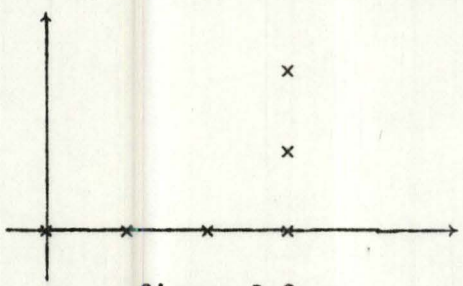


figure 2.3

Méthodes de type Adams

Une méthode $(k,\ell)$ de la forme

$$y_{n+k} - y_{n+k-1} + \sum_{i=0}^{k} \sum_{j=1}^{\ell_i} h^j \alpha_{ij} f^{(j-1)}(x_{n+i}, y_{n+i}) = 0$$

est appelée méthode du type Adams si son ordre d'erreur p est au moins $\sum_{i=0}^{k} \ell_i$.

Remarques

1. Pour une telle méthode, $\rho(\xi) = \xi^k - \xi^{k-1}$
   $\xi = 1$ est donc une racine simple de $\rho(\xi)$ et dès lors, l'ordre de consistance et l'ordre d'erreur sont identiques.

2. On sait que l'ordre de consistance d'une méthode $(k,\ell)$ ne peut pas dépasser $\sum_{i=0}^{k} \ell_i + k-1$.
   Une méthode est donc de type Adams si elle vérifie les deux conditions
   $$(i)\ \rho(\xi) = \xi^k - \xi^{k-1}$$
   $$(ii)\ \sum_{i=0}^{k} \ell_i \leq q \leq \sum_{i=0}^{k} \ell_i + k-1$$

3. Dans le cas particulier où $\ell_i = 1$ pour $0 \leq i \leq k$, on obtient les formules classiques de Adams-Bashforth et Adams-Moulton [21].
   Les méthodes d'Adams-Bashforth s'écrivent de façon générale :
   $$y_{n+1} - y_n = h \sum_{j=0}^{k} \nabla^j f_n \gamma_j$$
   où $\gamma_j$ sont des constantes.
   Cette méthode à k+1 pas est d'ordre de consistance $k+1 = \sum_{i=0}^{k+1} \ell_i$
   Les méthodes d'Adams-Moulton s'écrivent
   $$y_n - y_{n-1} = h \sum_{j=0}^{k} \nabla^j f_n \gamma_j$$
   où $\gamma_j$ sont des constantes.
   Cette méthode à k pas est d'ordre de consistance $k+1 = \sum_{i=0}^{k} \ell_i$.

Théorème 2.10 [25]

*Soient $\ell_0, \ell_1, \ldots, \ell_k$ des entiers positifs donnés et $\ell = \max_{i=0,\ldots k} \ell_i > 0$*

*Alors (i) il existe une et une seule méthode de type Adams avec les $\ell_i$ donnés*

*(ii) $p = q = \sum_{i=0}^{k} \ell_i$.*

En effet, si P est la matrice $(\sum_{i=o}^{k} \ell_i+2)$-incidente, associée à une telle méthode, le théorème 2.2 affirme qu'elle est order poised. Par le théorème 2.1, il n'existe donc pas de méthode non triviale avec $q > \sum_{i=o}^{k} \ell_i$. Par ailleurs, $P^{(k,o)}$ étant également order poised, le théorème 2.4 nous assure qu'il existe une et une seule méthode non triviale avec $q = \sum_{i=o}^{k} \ell_i$.

Remarquons qu'on ne peut pas prouver que

$$\text{sign } C_{q+1} = (-1)^{\ell_k} \text{ sign } \alpha_{ko}.$$

En effet, il se pourrait que P et $P^{(k,o)}$ aient des blocs impairs dans leurs lignes intérieures; le théorème 2.5 ne serait pas applicable.

Théorème 2.11 [25]

*Si $\alpha_{ij}$ où $j = 1, \ldots, \ell_i$ et $i = o, \ldots, k$ sont les coefficients d'une méthode de type Adams, alors les coefficients suivants sont non nuls :*

$$\alpha_{kj} \quad , \quad j = 1,\ldots,\ell_k$$

$$\alpha_{i,\ell_i-2t} \quad , \quad t = 0,1,\ldots \left[\frac{(\ell_i-1)}{2}\right] \quad i = 0,1,\ldots,k$$

$$\alpha_{oj} \quad , \quad j = 1,2,\ldots,\ell_o$$

La démonstration est similaire à celle du théorème 2.9.

Méthodes d'Enright [25], [16]

Une méthode d'Enright est une méthode du type Adams avec $\ell_o = \ell_1 = \ldots = \ell_{k-1} = 1$ et $\ell_k = 2$.

L'ordre d'erreur, identique à l'ordre de consistance, prend la valeur k+2. Un exemple du diagramme de Puiseux de telles méthodes, est donné à la figure 2.4.

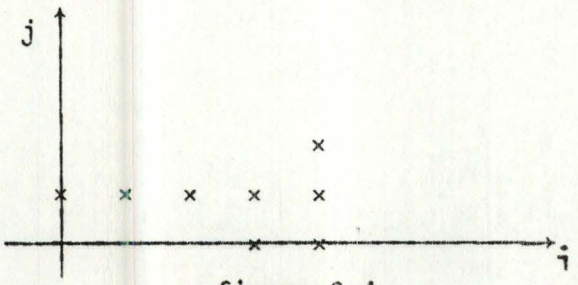



figure 2.4

## 2.6. CONCLUSION

Nous nous sommes efforcés, tout au long de ce chapitre, de trouver une borne supérieure pour l'ordre de consistance d'une méthode à pas et dérivées multiples, et de discerner les conditions suffisantes pour atteindre cette borne. Nous n'avons imposé aucune condition de stabilité pour construire ces méthodes.

Les chapitres suivants poursuivent le même but, mais avec des contraintes de stabilité de plus en plus exigeantes.

# CHAPITRE III

## ORDRE D'ERREUR MAXIMUM D'UNE MÉTHODE (K-L) ASYMPTOTIQUEMENT STABLE

## 3.1. INTRODUCTION

Pour des méthodes à pas multiples et à une ou deux dérivées, Dahlquist a obtenu l'ordre maximum que pouvait atteindre une méthode asymptotiquement stable [ 11 ].

Dans ce chapitre, nous étudierons la généralisation de ces résultats au cas des méthodes à nombre de dérivées arbitraires.

Le cas particulier de l'intégration de l'équation (1-12) où il faut distinguer entre ordre de consistance et ordre d'erreur est également abordé.

Le chapitre se terminera par l'étude des méthodes "offstep" où le coefficient $\alpha_{k,0}$ de (1-13) n'est plus différent de zéro.

Nous énonçons d'abord quelques propriétés relatives aux polynômes, qui seront utilisées dans les démonstrations qui suivent.

## 3.2. RESULTATS UTILES CONCERNANT LES POLYNOMES

### 3.2.1. Définitions

$U_n = \{p \mid p \in \pi_n$, degré = n, si $p(\alpha) = 0$ alors $|\alpha| < 1\}$

$\bar{U}_n = \{p \mid p \in \pi_n$, degré = n, si $p(\alpha) = 0$ alors $|\alpha| \leqslant 1\}$

### 3.2.2. Opération $*$ sur un polynôme

Soit $u \in \pi_n$

On définit $u^* \in \pi_n$ de la façon suivante :

$$u^*(s) = s^n\, u(s^{-1}) \qquad (3.1)$$

### 3.2.3. Opérateur D

Nous définirons un opérateur D :

$$D = s\,\frac{d}{ds}$$

Cet opérateur jouit des propriétés suivantes :

a) $D\, s^\nu = \nu\, s^\nu$ (3-2)

b) $D^j s^\nu = \nu^j s^\nu$ (3-3)

où $D^j$ indique que l'on applique j fois l'opérateur.

c) si $u(s) = \sum_{\nu=0}^{n} a_\nu \, s^\nu$

$v(s) = \sum_{\nu=0}^{m} b_\nu \, s^\nu$

$f(s) = \sum_{\nu=0}^{k} c_\nu \, s^\nu$

alors on a que :

$$f(D)\, u(s) = \sum_{\nu=0}^{n} a_\nu \, s^\nu \, f(\nu) \qquad (3\text{-}4)$$

$$f(D)\,(u\,v) = \sum_{\nu=0}^{n} a_\nu \, s^\nu \, f(\nu+D)\, v(s) \qquad (3\text{-}5)$$

$$[\, f(D)\, u \,]^* = f(n-D)\, u^*$$

d) si $f \equiv v \bmod \omega$ (c'est-à-dire $f(s)-v(s)$ peut être mis sous une forme où $w(s)$ apparaît en facteur)

avec $\omega(s) = s(s-1) \ldots (s-k)$

alors $f(D)\, u = v(D)\, u$ (3-6 bis)

e) $[\, f(D)\, (s-1)^\lambda \,]_{|s=1} = [\Delta^\lambda f(s)]_{|s=0}$

où $\Delta$ est l'opérateur de différence progressive

$$\Delta f(s) = f(s+1)-f(s) \qquad (3\text{-}7)$$

## 3.3. NOUVELLE CARACTERISATION POUR LA STABILITE ASYMPTOTIQUE - L'ORDRE D'ERREUR ET L'ORDRE DE CONSISTANCE D'UNE METHODE A PAS ET DERIVEES MULTIPLES

Nous cherchons à résoudre l'équation (1-1, 1-2)

### 3.3.1. La stabilité asymptotique

Rappelons qu'une méthode (k-ℓ) est asymptotiquement stable ssi elle vérifie la condition suivante :
les racines du polynôme $\rho_0(s)$ sont à l'intérieur du cercle unité et s'il y en a sur la frontière, elles sont de multiplicité 1.

Théorème de Lucas [31, p. 21]

Ce théorème décrit la position relative des racines et des points critiques d'un polynôme à variables complexes.
Les points critiques sont les racines de la dérivée première du polynôme.

Lucas assure que :

*Tous les points critiques d'un polynôme p non constant, se trouvent dans l'enveloppe convexe H de l'ensemble des racines de p.*
*Si les racines de p ne sont pas colinéaires, aucun des points critiques ne se trouve sur la frontière de H à moins que ce ne soit une racine multiple de p.*

Caractérisation de la stabilité symptotique

En vertu du théorème de Lucas, nous déduisons immédiatement qu'une méthode (k-$\ell$) est asymptotiquement stable ssi :

$$\rho_\bullet(s) \in U_k$$

$$\text{et } \rho'_\bullet(s) \in U_{k-1}$$

3.3.2. Théorème 3-1 sur l'ordre de consistance [25]

*L'opérateur L défini en (1-7), ou la méthode qui lui est associée, à l'ordre de consistance q ssi :*

$$\left[\sum_{j=0}^{\ell} y^{(j)}(D)\,\rho_j(s)\right]_{|s=1} \begin{cases} =0 & \forall\, y \in \pi_q \\ \neq 0 & \text{pour un } y \in \pi_{q+1} \end{cases} \qquad (3.8)$$

Démonstration

Pas 1 - Lemme

Une méthode a un ordre de consistance q ssi :

$$L[x^j, h] = 0 \qquad \text{pour } j=0,1,\ldots,q$$

$$L[x^{q+1}, h] = (q+1)!\,h^{q+1}\,C_{q+1}$$

où $C_{q+1}$ est une constante non nulle.

Démonstration du lemme

Supposons que la méthode ait l'ordre de consistance q.
La définition (1-8) assure alors que :

$$L[y(x), h] = C_{q+1}\,h^{q+1}\,y^{(q+1)}(x) + O(h^{q+2})$$

et ceci quel que soit $y(x) \in \mathcal{C}^{q+1}$
En particulier, prenons $y(x) = x^s$ où $s \leq q$
Dès lors, $y^{(q+1)}(x) = 0$ et $L[x^s, h] = O(h^{q+2})$
Par ailleurs $h^j\,y^{(j)}(x+ih) = h^j \left.\dfrac{d^j}{dx^j}(x^s)\right|_{x+ih}$
et donc $L[x^s, h]$ est un polynôme en h de degré au plus s ($s \leq q$).

Par conséquent $L[x^s,h] = O(h^{q+2}) = 0 \qquad s = 0,1,\ldots,q$

D'autre part, si $y(x) = x^{q+1}$, nous avons que $L[x^{q+1},h]$ est un polynôme en h de degré au plus q+1.

Dès lors, nous pouvons écrire que :

$$L[x^{q+1},h] = C_{q+1}\, h^{q+1}\, \frac{d^{q+1}}{dx^{q+1}}\,(x^{q+1})$$

ou encore

$$L[x^{q+1},h] = C_{q+1}\, h^{q+1}\, (q+1)\ !$$

La nécessité du lemme est ainsi prouvée.

Montrons la suffisance.

Nous introduisons un nouvel opérateur:

$$R[y(x),h] = L[y(x),h] - C_{q+1}\, h^{q+1}\, y^{(q+1)}\,(x)$$

$R[y(x),h]$ ainsi défini est visiblement une fonction linéaire qui s'annule pour tout $y(x) \in \Pi_{q+1}$.

En appliquant alors le théorème de Peano [14]
on a que :

$$R[y(x),h] = 0\,(h^{q+2}) \qquad \text{pour tout } y(x)\ \Pi_{q+1}$$

Ceci assure que l'ordre de consistance de L est q.

Notons que, puisque l'opérateur L est linéaire, l'énoncé du lemme peut prendre la forme suivante :

Une méthode (k-ℓ) associée à l'opérateur L défini en (1-7) a un ordre de consistance q ssi :

$$L[y(x),h] = 0 \qquad \text{pour tout } y \in \pi_q$$

$$\neq 0 \qquad \text{pour un } y \in \pi_{q+1}$$

Pas 2 - Démonstration du théorème 3-1

Soient : $z(x)$ un polynôme quelconque à coefficients réels

$$y(i) = z(h\,i)$$

$$l(y) = L[z(x+ih),h]_{|x=0} = L[z(ih),h]$$

En utilisant la définition (1-7), nous écrirons que :

$$L[z(ih),h] = \sum_{i=0}^{k} \sum_{j=0}^{\ell} \alpha_{ij}\, h^j\, z^{(j)}(ih)$$

Or : $y(i) = z(hi)$

et donc : $y^{(j)}(i) = h^j\, z^{(j)}\,(hi)$

Par conséquent, nous trouvons que :

$$L[z(ih),h] = \sum_{i=0}^{k} \sum_{j=0}^{\ell} \alpha_{ij}\, y^{(j)}(i)$$

k et $\ell$ étant des entiers finis, les signes de sommation peuvent être intervertis et nous avons :

$$L[z(ih),h] = \sum_{j=0}^{\ell} \sum_{i=0}^{k} \alpha_{ij}(1)^i\, y^{(j)}(i)$$

ce qui, par (3-4) donne :

$$L[z(ih)\ h] = l(y) = \left[\sum_{j=0}^{\ell} y^{(j)}(D)\, \varrho_j{}^{(s)}\right]_{|s=1} \qquad (3\text{-}9)$$

Nous pouvons donc affirmer que :

$$l(y) = 0 \Leftrightarrow L[z(ih),h] = 0.$$

Le lemme assure alors que l'opérateur L a un ordre de consistance q ssi :

$$L[z(ih),h] = 0 \quad \text{pour tout } z(ih) = y(i) \in \pi_q$$

$$\neq 0 \quad \text{pour un } y(ih) = y(i) \in \pi_{q+1}$$

Etant donné la forme (3-9) de l(y), il est alors bien évident que l'opérateur $L[y(x),h]$ a l'ordre de consistance q ssi :

$$\left[\sum_{j=0}^{\ell} y^{(j)}(D)\, \varrho_j(s)\right]_{|s=1} = 0 \quad \forall\, y \in \pi_q$$

$$\neq 0 \quad \text{pour un } y \in \pi_{q+1}$$

■

3.3.3. Théorème 3-2

*Si $q \geq 0$, alors $\varrho_0(1) = 0$.*

Démonstration

Prenons q=0

La formule (3-8) assure alors

$$\left[\sum_{j=0}^{\ell} y^{(j)}(D)\, \varrho_j(s)\right]_{|s=1} = 0 \qquad \forall\, y \in \pi_0$$

c'est-à-dire ∀ y = constante notée a.

Si y = a, alors $y^{(j)}=0$ ∀ j≠0 et par conséquent la somme ci-dessus se ramène à

$$a\ \varrho_0(s)_{|s=1} = 0 \qquad \forall\, a$$

et par conséquent : $\varrho_0(1) = 0$.

Supposons $q \geq 0$

L'opérateur de la formule (3-8) doit alors s'annuler pour tout $y \in \pi_q$ c'est-à-dire pour tout polynôme de degré plus petit ou égal à q.
En particulier, il doit donc s'annuler $\forall\ y \in \pi_0$ ce qui nous permet de nous ramener au cas précédent et de conclure : $\rho_0(1) = 0$. ■

3.3.4. Lien entre les ordres d'erreur et de consistance pour des méthodes asymptotiquement stables.

L'ordre de consistance égant non négatif, le théorème précédent nous permet de dire que s=1 est racine de $\rho_0(s)$.
La stabilité asymptotique impose alors que s=1 soit de multiplicité 1 et par conséquent p=q c'est-à-dire que l'ordre d'erreur est égal à l'ordre de consistance.
Par conséquent, dans le cas des méthodes stables, les théorèmes 3-1 et 3-2 peuvent s'énoncer en remplaçant q par p.

3.3.5. Généralisation dans le cas d'intégration d'une équation différentielle d'un ordre $r > 1$.

A présent, nous cherchons à résoudre (1-12).

a) Rappelons que dans ce cas, la méthode (k-$\ell$) utilisée est celle définie par l'opérateur (1-13).
Cette méthode est asymptotiquement stable si et seulement si elle satisfait à l'exigence (1-14).
En utilisant le théorème de Lucas, cette condition se ramène à :

$$\rho_0^{(j)}(s) \in U_{k-j} \quad j=0,1,\ldots,r-1$$
$$\text{et } \rho_0^{(r)}(s) \in U_{k-r}$$

b) L'opérateur L défini par (1-13), ou la méthode qui lui est associée, a l'ordre d'erreur p ssi :

$$\left[ y(D)\ \rho_0(s) + \sum_{j=1}^{\ell} y^{(r+j-1)}(D)\ \rho_j(s) \right]_{|s=1} = 0 \quad \forall\ y \in \pi_{p+r-1} \qquad (3.10)$$
$$\neq 0 \quad \text{pour un } y \in \pi_{p+r}$$

c) A partir de (3-10) on peut alors montrer [ 36 ] que :

$$\text{si } p \geq 0, \text{ alors } \rho_0(1) = \rho_0'(1) = \ldots\ldots = \rho_0^{(r-1)}(1) = 0 \qquad (3\text{-}11)$$

d) Un ordre d'erreur non négatif impose donc que s=1 soit une racine de multiplicité au moins r, du polynôme $\rho_0(s)$.
Si nous imposons la condition de stabilité asymptotique, la multiplicité de s=1

ne peut dépasser r et donc, si $p \geqslant 0$, alors s=1 est de multiplicité égale à r. Nous constatons par conséquent que même dans l'hypothèse de la stabilité asymptotique, les ordres d'erreur et de convergence sont distincts. Ils sont liés par la relation : $p = q-r+1$.

## 3.4. BORNES DE L'ORDRE D'ERREUR DES METHODES ASYMPTOTIQUEMENT STABLES

Nous allons considérer des méthodes du type (1-13).
Nous donnerons des théorèmes tout à fait généraux (r et $\ell$ quelconques) concernant les bornes d'erreur p, dans le cas où l'on impose la stabilité asymptotique. Les résultats de Dahlquist en seront une déduction immédiate.

Remarque préliminaire

Il est immédiat qu'un opérateur $L[y(x),h]$ du type (1-13) est d'ordre $\geqslant 0$ ssi

$$1 \leqslant r \leqslant k$$
$$\text{et } \rho_0(s) = (s-1)^r \ \varphi(s) \text{ où } \varphi(s) \in \pi_{k-r} \qquad (3.12)$$

Ceci découle de (3-11) et puisque, par sa définition, $\rho_0$ est un polynôme de degré $\leqslant k$, il est logique que $\varphi(s) \in \pi_{k-r}$ et que $1 \leqslant r \leqslant k$.
Si nous imposons en outre, la condition de stabilité asymptotique, nous pouvons en plus préciser que $\varphi(1) \neq 0$.

### 3.4.1. Théorème 3-3 [34]

*Pour tout $\rho_0(s)$ vérifiant les relations (3-12), il existe un et un seul $L[y(x),h]$ du type (1-13) avec $p \geqslant \ell(k+1)$.*

Note : la démonstration pour un r quelconque n'entraînant pas de complications trop grandes, nous établirons le théorème dans ce cadre général .

Démonstration

Notons : $v(y) = \left[ y(D)\, \rho_0(s) + \sum_{j=1}^{\ell} y^{(r+j-1)}(D)\, \rho_j(s) \right]_{s=1}$

Si nous appliquons (3-10), pour amener p au niveau $\ell(k+1)$ il faut et il suffit que : $v(y) = 0 \quad \forall y \in \pi_{\ell(k+1)+r-1}$

Cela revient à dire que v doit s'annuler sur toutes les fonctions de base qui engendrent les polynômes de degré $\leqslant \ell(k+1)+r-1$.
Nous allons construire cette base comme une extension d'une base de $\pi_{r-1}$ par exemple $\{1,x,\ldots,x^{r-1}\}$
Pour cela, nous définirons un ensemble d'indices :

$$F = \{(i,j) \mid i,j \in N,\ 0 \leqslant i \leqslant k,\ 0 \leqslant j \leqslant \ell-1 \}$$

Pour un couple (i,j) quelconque de F, nous pouvons alors définir un polynôme $q_{ij}$ vérifiant les 3 conditions suivantes :

a. $q_{ij} \in \pi_{\ell(k+1)+(r-1)}$

b. $q_{ij}^{(r+\mu)}(\nu) = \begin{cases} 0 \text{ si } (\nu,\mu) \neq (i,j) \text{ et } (\nu,\mu) \in F \\ 1 \text{ si } (\nu,\mu) = (i,j) \end{cases}$ (3-13)

c. $q_{ij}^{(\lambda)}(\frac{k}{2}) = 0$ pour $\lambda = 0,1,\ldots,r-1$

On peut montrer qu'avec la base $\{1,x,\ldots,x^{r-1}\}$ de $\pi_{r-1}$, les polynômes $q_{ij}$ qui sont au nombre de $\ell(k+1)$ forment une base de $\pi_{\ell(k+1)+r-1}$

v doit s'annuler sur tous ces éléments de base.

1er cas soit $y(x) \in \{1,x,\ldots,x^{r-1}\}$

v(y) se restreint alors à :

$$v(y) = [y(D)\, \varrho_0(s)]_{|s=1}$$

Par hypothèse (3-12) est vérifiée

Par conséquent $\varrho_0(1) = \varrho_0'(1) = \ldots = \varrho_0^{(r-1}(1) = 0$

Dès lors :

$$v(y) = [D^j\, \varrho_0(s)]_{|s=1} = 0 \quad j=0,\ldots,r-1$$

L'annulationde v est donc assurée sur cette première partie de la base de $\pi_{\ell(k+1)+r-1}$

2ème cas considérons un élément $\in \{q_{\nu\,\mu} | (\nu,\mu) \in F\}$

Il faut que :

$$v(q_{\nu\,\mu}) = [q_{\nu\,\mu}(D)\, \varrho_0(s) + \sum_{j=1}^{\ell} q_{\nu\,\mu}^{(r+j-1)}(D)\, \varrho_j(s)]_{|s=1} = 0 \quad \forall (\nu\,\mu) \in F$$

Par (3-4), on a

$$v(q_{\nu\mu}) = [\sum_{i=0}^{k} q_{\nu\,\mu}(i)\, \alpha_{oi}\, s^i + \sum_{j=1}^{\ell} (\sum_{i=0}^{k} q_{\nu\,\mu}^{(r+j-1)}(i)\, \alpha_{ij}\, s^i)]_{|s=1}$$

or par (3-13) $q_{\nu\,\mu}^{(r+j-1)}(i) = 0$ si $(i,j-1) \neq (\nu,\mu)$

$= 1$ si $(i,j-1) = (\nu,\mu)$

Pour tout $(\nu,\mu)$ fixé, il n'y a qu'un seul couple $(i,j-1) = (\nu,\mu)$.

Par conséquent, nous avons :

$$v(q_{\nu\,\mu}) = [\sum_{i=0}^{k} q_{\nu\,\mu}(i)\, \alpha_{oi}\, s^i + \alpha_{\nu,\mu+1}\, s^\nu]_{|s=1}$$

$$v(q_{\nu\,\mu}) = \sum_{i=0}^{k} q_{\nu\,\mu}(i)\, \alpha_{oi} + \alpha_{\nu\,\mu+1}$$

Nous déduisons donc que v s'annulera pour tout $(\nu\mu) \in F$ ssi

$$\alpha_{\nu\,\mu+1} = -\sum_{i=0}^{k} q_{\nu\,\mu}(i)\,\alpha_{oi} \qquad \forall(\nu\mu) \in F$$

$\alpha_{oi}$ étant connus pour tout i, $0 \leqslant i \leqslant k$, puisque $\varrho_0(s)$ est donné, la relation ci-dessus permet de déterminer de manière unique, tous les $\alpha_{\nu\ \mu+1}$, $0 \leqslant \nu \leqslant k$, $0 \leqslant \nu \leqslant \ell-1$, c'est-à-dire tous les coefficients de $L[y(x),h]$ qui, par la construction, aura un ordre $p \geqslant \ell(k+1)$.

■

3.4.2. Théorème 3 -4 sur l'ordre d'erreur maximal des méthodes asymptotiquement stables [36]

*Si $L[y(x),h]$ donné par (1-13) est asymptotiquement stable, et $\alpha_{k,0} \neq 0$. Si l'une des trois conditions suivantes est vérifiée :*

a. *$r=1$ ou $r=2$*
b. *$\ell$ est pair*
c. *$\ell$ est impair et $\ell > \bar{\ell}(k,r)$ c'est-à-dire $\ell$ est minoré par une quantité dépendant du k et du r.*

*Alors :* $p \leqslant \ell(k+1)+1$ *si $\ell$ est impair et k pair*; $\ell(k+1)$ *sinon* (3.14)

*En outre (3-14) est vérifiée dans l'égalité ssi* $\varrho_0^* = (-1)^r \varrho_0$

Démonstration

Nous nous limiterons ici à démontrer le cas où r=1, les autres cas nécessitant de longs développements de calcul qui peuvent être obtenus dans Reimer [36]

Pas 1 - Lemmes et définitions utiles

• Définition 1

Posons $\omega(x) = x(x-1)\,(x-2)\,\ldots(x-k)$.

Ce polynôme est lié à la fonction bêta par la relation

$$\omega(x) = \alpha\,\beta(x+1,k+1-x)\,\sin\pi x$$

où $\alpha$ est une constante.

L'obtention de cette forme est décrite dans Reimer [35]

Nous noterons $\beta(x+1,k+1-x) = A(x)$

et la définition de la fonction bêta donne :

$$A(x) = \int_0^1 t^x\,(1-t)^{k-x}\,dt \qquad -1 < x < k+1$$

Nous écrirons donc :

$$\omega(x) = \alpha\,A(x)\,\sin\pi x \qquad -1 < x < k+1 \qquad (3\text{-}15a)$$

Rappelons qu'une fonction $\beta$ est liée à la fonction $\Gamma$ par la relation :

$$\beta(x,y) = \frac{\Gamma(x)\ \Gamma(y)}{\Gamma(x+y)}$$

et par conséquent :

$$A(x) = A(k-x) \qquad (3\text{-}15b)$$

- Définition 2

Nous serons amenés à considérer les fonctions suivantes :

$$A_{mj}(x) = (x - \frac{k}{2})^j\ A^m(x) \qquad -1 < x < k+1 \quad (3\text{-}15c)$$
$$j = 0,1,\ldots$$

La relation (3-15b) permet alors d'écrire :

$$A_{mj}^{(\lambda)}(k-x) = (-1)^{j+\lambda}\ A_{mj}^{(\lambda)}(x) \qquad j,\lambda = 0,1,\ldots$$

où $(\lambda)$ indique que l'on dérive $\lambda$ fois.

- Lemme 1

Soient $j$ et $\lambda \in \mathbb{N}$

Supposons $0 \leqslant j \leqslant \lambda$

$\lambda \equiv j \bmod 2$ (c'est-à-dire $\lambda - j$ est multiple de 2)

Alors : $A_{mj}^{(\lambda)}(x) > 0$ pour $-1 < x < k+1$

- Lemme 2

Soit H une fonction réelle en x qui est (k-r) fois continuement différentiable dans [ 0,k-r ]

Supposons que :

$$H(k-r-x) = (-1)^{\nu}\ H(x)$$
$$H^{(\lambda)}(x) > 0 \quad \text{pour } 0 \leqslant x \leqslant k-r$$
$$0 \leqslant \lambda \leqslant k-r$$
$$\lambda \equiv \nu \bmod 2$$
$$\nu \in \mathbb{N}$$

Soit G un polynôme t.q

$$G(\nu) = (-1)^{\mu\nu} H(\nu) \quad \text{pour } \nu = 0,1,\ldots,k-r$$
$$\mu \in \mathbb{N}$$

Supposons enfin que U soit un polynôme réel de degré exactement k-r, n'ayant pas de racine hors du cercle unité.

Alors :

$$[G(D)\ U(s)]_{|s=1} = 0 \text{ ssi } U^* = (-1)^{\mu(k-r)+\nu+1}\ U$$

Pas 2 - Idée de la démonstration - Nouvelle formulation du problème

Le théorème 3-3 nous donne une borne inférieure pour l'ordre d'erreur : $p \geq \ell(k+1)$.

Ici, nous supprserons que $p = \ell(k+1)+d$ où $d \geq 0$.

Notre problème est donc de voir quelles valeurs peuvent être attribuées à d.

Rappelons que nous nous limitons au cas r=1.

Dans la démonstration du théorème 3-3, nous avons obtenu une base pour $\pi_{\ell(k+1)}$.

Nous allons procéder à une extension de cette base.

Pour cela, choisissons des polynômes $g_j$ de degré $\ell(k+1)+1+j$ où $j = 0,1,\ldots,d$.

Nous obtiendrons ainsi une base de $\pi_{\ell(k+1)+d}$.

Puisque $p = \ell(k+1)+d$, par (3-8) où p=q, nous pouvons écrire que :

$$v(g_j) = 0 \text{ pour } j = 0,1,\ldots,d-1$$
$$v(g_j) \neq 0 \text{ pour } j=d \qquad (3\text{-}16)$$

où $v(y) = \left[\sum_{k=0}^{\ell} y^{(k)}(D)\ {}_k(s)\right]_{|s=1}$

Ceci donne une caractérisation du d.

Les $g_j$ peuvent être choisis de façons différentes mais nous les définirons comme suit :

a. $g_j(x) = c(x-\frac{k}{2})^j\ \omega^{\ell}(x)$ où c est une constante arbitraire non nulle (3-16bis)

b. $g_j(\frac{k}{2})$ qui est la valeur initiale est arbitraire.

Justifions le choix de tels $g_j$.

On peut montrer que ceux-ci vérifient :

$$g_j^{(\lambda)} \equiv 0 \bmod \omega \quad \text{pour} \quad \lambda=1,\ldots,\ell \qquad (3\text{-}17)$$

Ceci va nous permettre d'écrire $v(g_j)$ sous une nouvelle forme.

Nous avons :

$$v(g_j) = \left[g_j(D)\,\varrho_0(s) + \sum_{m=1}^{\ell} g_j^{(m)}(D)\,\varrho_m(s)\right]_{|s=1}$$

qui, par (3-17) et (3-6bis) se restreint à :

$$v(g_j) = \left[g_j(D)\,\varrho_0(s)\right]_{|s=1}$$

Le nombre p étant positif et la méthode étant stable, (3-12) nous permet d'écrire :

$$\varrho_0(s) = (s-1)\,\varphi(s)$$
où $\varphi(s) \in \pi_{k-1}$; $\varphi(s) = \sum_{\nu=0}^{k} a_\nu s^\nu$
et $\varphi(1) \neq 0$.

Par conséquent : $v(g_j) = \left[g_j(D)\,(s-1)\,\varphi(s)\right]_{|s=1}$

et par (3-5), on a $v(g_j) = \left[ \sum_{\nu=0}^{k-1} a_\nu s^\nu g_j(D+\nu)\ (s-1) \right]_{|s=1}$

ce qui en vertu de (3-7) donne :

$$v(g_j) = \left[ \sum_{\nu=0}^{k-1} a_\nu \Delta g_j(\nu)\ s^\nu \right]_{|s=0}$$

En appliquant (3-4), nous obtenons finalement que :

$$v(g_j) = \left[ \Delta g_j(D)\ \varphi(s) \right]_{|s=1} \qquad (3\text{-}18)$$

Notons : $G_j(x) = \Delta g_j(x)$ (3-18bis)

Par le théorème de Peano [ 14 ], on a alors :

$$G_j(x) = \int_0^1 g'_j(x+t)\ dt$$

ce qui par (3-16bis) devient :

$$G_j(x) = \int_0^1 c(t+x-\tfrac{k}{2})^j\ \omega^{\ell}(x+t)\ dt$$

ou encore, par (3-15a)

$$G_j(x) = \int_0^1 c(t+x-\tfrac{k}{2})^j\ \alpha^{\ell}\ A^{\ell}(x)\ \sin^{\ell} \pi(x+t)\ dt \qquad (3\text{-}19)$$

où $\alpha^{\ell}$ étant une constante, entre dans c.

Par ailleurs, en normalisant, on peut obtenir c=1.

Etant donné les restrictions sur A(x), la formule (3-19) est valable pour $-1 < x < k+1$

et donc à fortiori pour $-1 < x < k$.

Grâce à (3-15c), (3-19) s'écrit :

$$G_j(x) = \int_0^1 A_{\ell j}(x+t)\ \sin^{\ell} \pi(x+t)\ dt \qquad -1 < x < k \qquad (3\text{-}20)$$

Définissons enfin, en vue d'appliquer le lemme 3, la fonction :

$$H_j(x) = \int_0^1 A_{\ell j}(x+t)\ \sin^{\ell} \pi t\ dt \qquad -1 < x < k \qquad (3\text{-}21)$$

Les propriétés du sinus et de $A_{mj}(x)$ permettent alors d'obtenir les relations qui suivent entre $H_j(x)$ et $G_j(x)$ :

$$G_j(\nu) = (-1)^{\nu\ell} H_j(\nu) \qquad \nu = 0,1,\ldots,k-1 \qquad (3\text{-}22)$$

$$H_j(k-1-x) = (-1)^j H_j(x) \qquad (3\text{-}23)$$

Pas 3 - Discussion proprement dite de l'ordre d'erreur pour r=1

Montrons que, dans ce cas, (3-21) vérifie les hypothèses du lemme 3.

Nous avons que $0 \leqslant t \leqslant 1$

et donc $\sin^{\ell} \pi t \geqslant 0$

Par ailleurs, le lemme 1, assure que :

$$A_{1\,j}^{(\lambda)}(x+t) > 0 \quad \text{pour 1)} \; 0 \leqslant x+t \leqslant k-1$$

ce qui revient à

$-1 \leqslant x \leqslant k$ car $0 \leqslant t \leqslant 1$

2) $\lambda \equiv j \bmod 2$.

L'intégrant de (3-21) étant donc une fonction non négative et non constamment nulle, nous pouvons alors dire que :

$$H_j^{(\lambda)}(x) > 0 \quad \text{pour} \quad -1 < x < k$$

$$\lambda \equiv j \bmod 2.$$

Grâce aux relations (3-22) et (3-23), nous constatons alors que les hypothèses du lemme 2 sont vérifiées par $H_j$ et $G_j$ et le rôle du polynôme U sera tenu par $\varphi(s)$.

Nous avons, par conséquent, que :

$$[G_j(D)\ \varphi(s)]_{|s=1} = 0 \text{ ssi } \varphi^* = (-1)^{\ell(k-1)+j+1}\varphi$$

ou encore par (3-18) et (3-18bis)

$$v(g_j) = 0 \text{ ssi } \varphi^* = (-1)^{\ell(k-1)+j+1}\varphi \qquad (3\text{-}24)$$

1er cas : si $\ell$ est pair

(3-24) s'écrit alors :

$v(g_j) = 0$ ssi $\varphi^* = (-1)^{j+1}\varphi$

et en particulier :

$v(g_0) = 0$ ssi $\varphi^* = -\varphi$

Il est évident que $\varphi^*(1) = \varphi(1)$

et par conséquent $\varphi^* = -\varphi$ ssi $\varphi^*(1) = \varphi(1) = 0$

ce qui ne se peut par la condition de stabilité.

Donc $v(g_0) \neq 0$. Dès lors, par (3-16) et puisque $d \geqslant 0$, nous pouvons conclure que $d = 0$.

Par conséquent, si r=1, et $\ell$ est pair, $p = \ell(k+1)$ dans le cas d'une méthode stable.

2ème cas : si $\ell$ est impair

(3-24) se ramène alors à :

$$v(g_j) = 0 \quad \text{ssi} \quad \varphi^* = (-1)^{k+j}\,\varphi \qquad (3\text{-}25)$$

Par conséquent, si k est pair, la condition (3-25) devient :

$\varphi^* = \varphi$ si j=0

$\varphi^* = -\varphi$ si j=1

si k est impair, elle devient :

$\varphi^* = -\varphi$ si j=0

$\varphi^* = \varphi$ si j=1

Pour obtenir en même temps $\varphi^* = \varphi$ et $\varphi^* = -\varphi$, il faudrait que $\varphi(s) = 0$ $\forall$ s et donc que $\varphi(1) = 0$, ce qui est contraire à la stabilité.

Par conséquent, $g_0$ et $g_1$ ne peuvent annuler v en même temps et donc si $v(g_0) = 0$, alors $v(g_1) \neq 0$.

Par (3-16), nous concluons donc $d \leq 1$.

Supposons maintenant que d=1.

Alors par (3-16) et (3-25) :

$$v(g_0) = 0 \quad \text{ssi} \quad \varphi^* = (-1)^k \varphi$$

$$v(g_1) \neq 0 \quad \text{ssi} \quad \varphi^* \neq (-1)^{k+1}\varphi$$

Si k est impair : $v(g_0) = 0$ ssi $\varphi^* = -\varphi$ ce qui est impossible à cause de l'hypothèse de stabilité.

Par conséquent, si k est impair, nous aurons $d < 1$ et donc $p < \ell(k+1)+1$

ou $p \leq \ell(k+1)$

puisque $p \in \mathbb{N}$.

Si k est pair : $v(g_0) = 0$ ssi $\varphi^* = \varphi$, relation qui est réalisable de plusieurs façons.

Si k est pair, nous avons donc bien $d \leq 1$ et l'égalité se produit ssi $\varphi^* = \varphi$

Par conséquent $p \leq \ell(k+1)+1$ $\ell$ impair et k pair

et $p = \ell(k+1)+1$ ssi $\varphi^* = \varphi$

Pas 4 - Montrons que $\varphi^* = \varphi$ ssi $\rho_0^* = -\rho_0$

Nous savons que : $\rho_0(s) = (s-1)\,\varphi(s)$

ou encore $\varphi(s) = \dfrac{\rho_0(s)}{s-1}$

Par ailleurs $\varphi^*(s) = s^k \dfrac{1}{s} \dfrac{\rho_0(\frac{1}{s})}{\frac{1}{s}-1}$ par (3-1)

$$= \left(\frac{1}{1-s}\right) \rho_0^*(s)$$

Nous aurons donc que :

$$\varphi(s) = \varphi^*(s) \quad \text{ssi} \quad \frac{\rho_0(s)}{s-1} = \frac{\rho_0^*(s)}{1-s}$$

Ce qui revient à écrire :

$$\varphi(s) = \varphi^*(s) \quad \text{ssi} \quad \rho_0(s) = -\rho_0^*(s).$$

Remarque

Si $\ell=1$ et $r=1$, le théorème précédent nous permet de retrouver les résultats établis par Dahlquist puisque $p \leqslant k+2$

et $p = k+2$ si $k$ est pair et $\rho_0^* = -\rho_0$

Dahlquist avait formulé la deuxième partie du résultat de façon un peu différente en disant :

$p = k+2$ si $k$ est pair et si $\rho_0(s) = 1$ alors $|s| = 1$

Montrons que $\rho_0^* = -\rho_0$ entraîne la condition que :

si $\rho_0(s) = 0$ alors $|s| = 1$

Nous savons que la stabilité conduit à la propriété suivante :

si $\rho_0(s) = 0$ alors $|s| \leqslant 1$

Par (3-1) $\rho_0^*(s) = s^k \rho_0(s^{-1})$

Si $s$ est racine de $\rho_0$, puisque $\rho_0^* = -\rho_0$, nous avons :

$$\rho_0^*(s) = \rho_0(s^{-1}) = 0 \qquad (3\text{-}26)$$

Dès lors, si $|s| < 1$, alors $|s^{-1}| > 1$ ce qui est impossible par la stabilité car (3-26) assure que $s^{-1}$ est racine de $\rho_0$.

Par conséquent, nous avons que $|s| = 1$.

## 3.5. METHODES OFFSTEP

Jusqu'ici, nous avons considéré des méthodes pour lesquelles $\alpha_{k,0} \neq 0$

Si nous n'imposons pas cette condition, nous obtiendrons une méthode "offstep" qui est donnée par un opérateur de la forme :

$$L[y(x),h] = \sum_{i=s}^{k-t} \alpha_{i0}\, y[x+ih] + h^r \sum_{i=0}^{k} \sum_{j=1}^{\ell} \alpha_{ij}\, h^{j-1}\, y^{(j+r-1)}_{[x+ih]} \qquad (3\text{-}27)$$

où $\alpha_{k-t,0} \neq 0$ et $\alpha_{s,0} \neq 0$

L'utilisation de telles méthodes permet d'accroître l'ordre d'erreur comme nous allons le voir dans les théorèmes qui suivent.

Nous nous limiterons au cas où $r=\ell=1$.

3.5.1. Théorème 3.-5 [25]

*Soient $\ell=r=1$*

*Une méthode du type (3-27) a un ordre*

$$p \leqslant 2k - t - s$$

Ceci est une borne assez grossière mais qui se comprend par le raisonnement suivant :

p est l'ordre d'erreur et dès lors l'ordre de consistance s'exprime :

$q = p+m-1$ où m est la multiplicité de s=1 racine de $\rho_0(s)$.

La matrice d'incidence (2-4bis) associée à une méthode du type (3-27) où $r=\ell=1$ aura la forme suivante :

| | | | | |
|---|---|---|---|---|
| ligne 0 | 0 | 1 | 0 ..... | 0 |
| | ⋮ | 1 | 0 | |
| ligne (s-1) | 0 | ⋮ | ⋮ | |
| ligne s | 1 | ⋮ | ⋮ | |
| | 1 | ⋮ | ⋮ | |
| | ⋮ | ⋮ | ⋮ | |
| ligne (k-t) | 1 | ⋮ | ⋮ | |
| | 0 | ⋮ | ⋮ | |
| | ⋮ | ⋮ | ⋮ | |
| | 0 | 1 | 0 ..... | 0 |

Cette matrice comporte (k-t-s+1)+(k+1) éléments non nuls et est order poised (Chapitre II - section 3).

Par conséquent, le théorème 2-1 assure que

$$q \leqslant (2k-t-s+2)-2$$

c'est-à-dire

$$q \leqslant 2k-t-s$$

Etant donné le lien entre l'ordre d'erreur et l'ordre de consistance, nous avons dès lors que :

$$p + m-1 \leqslant 2k-t-s$$

Puisque nous considérons toujours des ordres non négatifs, m sera plus grand ou égal à 1 et donc :

$$p \leqslant 2k-t-s$$

Nous affinerons ce résultat dans le théorème suivant :

3.5.2. Théorème 3-6 [25]

*Soit une méthode du type (3-27) asymptotiquement stable, où $r=\ell=1$*

*Soit* $R(z) = (1/2)^k (z-1)^{k-t} (z+1)^{-s} \rho_0\left(\frac{z+1}{z-1}\right)$

*Alors* $p \leqslant 2[k/2] + 2$ *si* $t=s$

*et R est un polynôme impair*

$(k-1) + \max\{t-s, 0\}+2$ *sinon.*

*où* $[a]$ *est le plus grand entier n'excédant pas a.*

Posons v = t-s.

Ce théorème montre que pour augmenter l'ordre d'erreur d'une méthode offstep stable, il faut choisir $v > 0$.

Par ailleurs, comme nous cherchons toujours des ordres non négatifs, le théorème 3.-2 nous assure que $\rho_0(1) = 0$ et par conséquent, $\rho_0$ se factorise comme suit :

$$\rho_0(s) = (s-1)\,\varphi(s) \text{ où } \varphi(s) \text{ est un polynôme.}$$

Il est donc bien évident que $\rho_0(s)$ doit être de degré $\geqslant 1$

Or le polynôme caractéristique $\rho_0(s)$ lié à (3-27) est de degré : k-t-s.

Nous devons donc vérifier la relation suivante :

$$k-t-s \geqslant 1$$

ce qui revient à écrire :

$$v \leqslant k-2s-1$$

Nous avons donc obtenu des bornes sur le v :

$$0 < v \leqslant k-2s-1$$

3.5.3. Théorème 3.-7 [25]

*Soit $\ell=1$, $r=1$*

*et $t > s$*

*Alors pour $0 < v < k-1-2s$, il existe une méthode stable (3-27) avec $p \geqslant k+2$*

*Si $v = k-1-2s$, alors $p \leqslant k+1$*

D'où vient cette limitation du p quand v a sa valeur maximale permise ?

Si v = k-1-2s, la matrice d'incidence liée à la méthode possèdera k+3 éléments non nuls et sera order poised.

Le même raisonnement que pour le théorème 3-5, nous conduira à la conclusion que : $p \leqslant k+1$.

Il n'y a donc pas intérêt à prendre v le plus grand possible comme le suggérait le théorème 3-6, puisque si v a sa valeur maximale, l'ordre d'erreur admet une borne supérieure qui n'offre pas d'avantages par rapport à celle donnée au théorème 3-4.

3.6. CONCLUSIONS

Nous retiendrons de ce chapitre les faits suivants :

1. Lors de l'intégration d'une équation différentielle du premier ordre, une méthode à k pas et dérivées multiples, stable, fournit un ordre d'erreur maximal valant :

   $\ell(k+1)+1$ si $\ell$ est impair et k pair et $\rho_0^* = -\rho_0$

   $\ell(k+1)$ sinon.

2. Cet ordre peut être augmenté par l'utilisation de méthodes offstep.
Notons cependant que ces méthodes présentent aussi certains inconvénients :
- les méthodes offstep ne sont pas à itérations directes puisque $\alpha_{k0} \neq 0$
- par ailleurs, les méthodes offstep intéressantes c'est-à-dire celles pour lesquelles $0 < v < k-1-2s$ ne sont pas A stables.
En effet, nous verrons au chapitre VI que l'ordre maximum d'une méthode (k-1), A.stable vaut 2 (Dahlquist [12]).
Par conséquent, puisque le théorème 3-7 fournit, pour une méthode offstep (k-1) une borne inférieure de l'ordre d'erreur p, qui vaut au moins 3, nous concluons immédiatement que la méthode n'est pas A.stable.
Dès lors, les méthodes offstep présentent peu d'intérêt pour l'intégration des systèmes Stiff que nous introduirons au chapitre IV.

°
° °

# CHAPITRE IV

## PROBLÈME DE STABILITÉ DES SYSTÈMES STIFF

## 4.1. NOTATIONS

Nous définirons les demi-plans suivants, que nous utiliserons dans la suite.

$$R = \{\xi \in \mathbb{C} : \text{Re } \xi > 0 \}$$
$$L = \{\xi \in \mathbb{C} : \text{Re } \xi < 0 \}$$
$$G = \{\xi \in \mathbb{C} : \text{Im } \xi > 0 \}$$
$$H = \{\xi \in \mathbb{C} : \text{Im } \xi < 0 \}$$

Nous noterons $\overline{R}$, respectivement $\overline{L}$, $\overline{G}$, $\overline{H}$, la fermeture de R , respectivement L, G , H .

## 4.2. INTRODUCTION

Jusqu'ici, nous avons envisagé la convergence des méthodes (k,ℓ) c'est-à-dire que, en une abscisse x fixée, la solution numérique vérifie :

$$\lim_{\substack{h \to 0 \\ nh = x-a}} y_n = y(x)$$

A présent, nous allons considérer un h fixé. Il se peut alors que la solution numérique ne tende pas toujours vers la solution théorique lorsque x tend vers l'infini. Ce phénomène est appelé *stabilité faible*. Nous allons tâcher de trouver des conditions pour que

$$\lim_{\substack{n \to \infty \\ x=a+nh \\ h \text{ fixé}}} y_n = y(x) \qquad (4.0)$$

## 4.3. STABILITE FAIBLE. REGION DE STABILITE ABSOLUE

Intégrons l'équation différentielle

$$y' = \lambda y \qquad (4.1)$$

où $y(0) = 1$ et $\lambda \in \mathbb{C}$

Nous justifierons dans la suite le choix d'une telle équation différentielle scalaire avec un $\lambda$ complexe.
Si nous utilisons la méthode $(k,\ell)$ définie par l'opérateur (1.7), nous obtenons la récurrence :

$$\sum_{i=o}^{k} \sum_{j=o}^{\ell} \alpha_{ij} h^j \lambda^j y_{n+i} = 0 \qquad n=0,1,... \tag{4.2}$$

Notons

$$\eta_i(\mu) = \sum_{j=o}^{\ell} \alpha_{ij} \mu^j \tag{4.3}$$

Ceci nous permet d'écrire (4.2) comme suit :

$$\sum_{i=o}^{k} \eta_i(h\lambda)\, y_{n+i} = 0 \qquad n=0,1,... \tag{4.4}$$

Dans ce qui suit, nous utiliserons l'abréviation :

$$\mu = h\lambda.$$

Pour exprimer les conditions que nous cherchons, il est nécessaire d'exprimer la solution numérique de (4.1) par la méthode $(k,\ell)$ (4.4).
Si k valeurs initiales sont fournies, cette relation définit de façon unique une suite $\{y_n\}_{n=0,1,...}$, à condition que $\eta_k(\mu)$ soit non nul.

Dans une première étape, nous fixerons un $\mu$ tel que $\eta_k(\mu)$ soit non nul.

Dans une seconde étape, nous laisserons varier $\mu$ dans le plan complexe afin de déterminer ceux qui permettent de vérifier la relation (4.0).
Un problème se pose si un des $\mu$ est racine de $\eta_k(\mu)$.

Dans ce cas, la récurrence (4.4) devient :

$$\sum_{i=o}^{k-1} \eta_i(\mu)\, y_{n+i} = 0 \qquad n=0,1,...$$

et fournira, avec k-1 valeurs initiales, une suite $\{y_n\}_{n=0,1,...}$ unique à condition que $\eta_{k-1}(\mu)$ soit non nul.
Si $\eta_{k-1}(\mu)$ est nul, on applique le même raisonnement.
Ainsi donc, nous pourrons obtenir une suite $\{y_n\}_{n=0,1,...}$ unique, pour autant que $\mu$ ne soit pas racine de $\eta_i(\mu)$ $\quad i=0,1,...,\ k$.

C'est pourquoi nous ferons l'hypothèse, dans tout ce qui suit, que $\eta_i(\mu)$ $i=0,1,\ldots,k$ n'ont pas de facteur commun non constant.

Première étape : résolution de l'équation (4.4) avec $\mu$ fixé

La relation (4.4) détermine dans ce cas une récurrence linéaire d'ordre k, à coefficients constants.
La forme générale de la solution s'écrira alors [ 25 ]

$$y_n = \sum_{i=1}^{s} \sum_{j=1}^{m_i} C_{ij}\, n^{j-1}\, \zeta_i^n$$

où $\zeta_i$, $i=1,\ldots$ s sont les racines distinctes de multiplicité $m_i$ de l'équation

$$\Phi(\zeta) = \sum_{i=o}^{k} \eta_i(\mu)\, \zeta^i = 0 \qquad (4.5)$$

et $C_{ij}$ $\begin{matrix} i=1,2,\ldots,s \\ j=1,2,\ldots,m_i \end{matrix}$ sont des constantes.

Ce même raisonnement peut être mené quel que soit le $\mu$ fixé, pourvu que $\eta_k(\mu)$ soit différent de 0.
Par conséquent, la formule (4.5) s'écrira :

$$\Phi(\zeta,\mu) = \sum_{i=o}^{k} \eta_i(\mu)\, \zeta^i = \sum_{j=o}^{\ell} \rho_j(\zeta)\, \mu^j = 0 \qquad (4.6)$$

Dans la littérature, l'équation (4.6) porte le nom d'*équation caractéristique* d'une méthode $(k,\ell)$.
Il est clair que $\Phi(\zeta,\mu)$ est un polynôme en deux variables, de degré k en $\zeta$ et $\ell$ en $\mu$. Nous le nommerons *polynôme de caractérisation*.

Deuxième étape : $\mu$ variant dans le plan complexe

Remarque : nous ferons pour la suite l'hypothèse que $\Phi(\zeta,\mu)$ ne peut pas se factoriser sous la forme :

$$\Phi(\zeta,\mu) = P(\mu)\, R(\zeta,\mu)$$

où $P(\mu)$ est un polynôme en $\mu$ de degré au plus $\ell$.
Dès lors, il n'existera pas de $\mu$ tel que $\eta_i(\mu) = 0$ $i=0,1,\ldots,k$.

La résolution de l'équation (4.6) permet de trouver les racines $\zeta_i$ , i=1,2,...,s ($s \leq k$) en fonction de $\mu$. Ces $\zeta_i(\mu)$ sont des branches de la fonction algébrique définie par

$$\Phi(\zeta(\mu),\mu) \equiv 0 \quad \text{pour tout } \mu \text{ complexe.}$$

Supposons que nous travaillons avec une méthode convergente.
Si h est nul, il en est de même pour $\mu$, et $\zeta_i(0)$ est racine de $\rho_0(\zeta)$.
On déduit alors de la stabilité asymptotique (théorème 1.2) que :

$$|\zeta_i(0)| \leq 1$$

si $|\zeta_i(0)| = 1$ alors $m_i = 1$, où $m_i$ est la multiplicité.

Par ailleurs, la consistance (théorème 1.2) assure qu'une racine de $\rho_0(\zeta)$ vaut 1.
Nous la noterons :

$$\zeta_1(0) = 1$$

Puisque $m_1 = 1$, $\zeta_1(\mu)$ détermine une branche de la fonction algébrique $\zeta(\mu)$, analytique dans un voisinage de 0 et prenant la valeur 1 en 0.
Cette branche $\zeta_1(\mu)$ est appelée *racine principale*.
De manière analogue, on définit les *racines essentielles* $\zeta_i(\mu)$ pour i=1,2, ..., t ($t \leq s$) comme étant les branches de $\zeta(\mu)$ telles que $|\zeta_i(0)| = 1$.
Avec ces nouvelles notions, la solution numérique de (4.1) par la méthode (k,ℓ) (4.4) s'écrit comme suit :

$$\begin{aligned}
y_n = {} & C_{11}\,\xi_1^n(\mu) && \text{racine principale} \\
& +C_{21}\,\xi_2^n(\mu) \\
& +C_{31}\,\xi_3^n(\mu) && \text{contribution des racines essentielles} \\
& + \ldots \\
& +C_{t1}\,\xi_t^n(\mu) \\
& +C_{t+1,1}\,\xi_{t+1}^n(\mu) + C_{t+1,2}\,n\,\xi_{t+1}^n(\mu) + \ldots + C_{t+1,m_{t+1}}\,n^{m_{t+1}}\,\xi_{t+1}^n(\mu) && \text{contribution des racines non essentielles} \\
& + \ldots \\
& +C_{s,1}\,\xi_s^n(\mu) + C_{s,2}\,n\,\xi_s^n(\mu) + \ldots + C_{s,m_s}\,n^{m_s}\,\xi_s^n(\mu)
\end{aligned} \qquad (4.7)$$

Par ailleurs, la solution théorique de (4.1) est :

$$y(x) = e^{\lambda x}.$$

Cherchons à présent les $\mu$ tels que (4.0) soit vraie.

Pour cela, fixons un h. Voyons à quelles conditions nous avons que

$$\lim_{n\to\infty} y_n = \lim_{x\to\infty} e^{\lambda x}$$

Il est clair que

$$\lim_{x\to\infty} e^{\lambda x} = 0 \qquad \text{ssi} \qquad \lambda \in L$$

D'autre part, si nous notons $\mathcal{A}$ l'ensemble suivant :

$$\mathcal{A} = \{\mu \in \overline{\mathbb{C}}^{(+)} \text{ t.q. } |\zeta_i(\mu)| < 1,\ i=1,2,\ldots,k\} \tag{4.8}$$

nous pouvons dire que

$$\lim_{n\to\infty} y_n = 0 \qquad \text{ssi} \qquad h\lambda = \mu \in \mathcal{A}$$

$\mathcal{A}$ est appelée *la région de stabilité absolue.*

Nous comparons dans le tableau suivant les limites des solutions numérique et théorique, selon les zones dans lesquelles se trouvent $\lambda$ et $\mu$.

| $\lambda$ | $\mu$ | $\lim_{x\to\infty} \lvert e^{\lambda x}\rvert$ | $\lim_{n\to\infty} \lvert y_n\rvert$ |
|---|---|---|---|
| $\in L$ | $\in \mathcal{A}$ | 0 | 0 |
| $\in L$ | $\notin \mathcal{A}$ | 0 | $\neq 0$, souvent $\infty$ |
| $\notin L$ | $\in \mathcal{A}$ | $\neq 0$, souvent $\infty$ ($\mathrm{Re}\ \lambda \neq 0$) | 0 |
| $\notin L$ | $\notin \mathcal{A}$ | $\neq 0$, souvent $\infty$ ($\mathrm{Re}\ \lambda \neq 0$) | $\neq 0$, souvent $\infty$ |

Nous voyons que le comportement des solutions numériques pose des problèmes dans les deuxième et troisième cas. Ceci est le phénomène de stabilité faible. Pour éviter cela, il est donc naturel de rechercher des méthodes pour lesquelles $\mathcal{A} = L$.

---

(+) $\overline{\mathbb{C}}$ est le plan complexe augmenté des points à l'infini.

De telles méthodes offrent un intérêt tout particulier car elles conviennent à l'intégration des systèmes stiff, comme nous allons le voir dans la section qui suit.

## 4.4. SYSTEMES STIFF

### 4.4.1. Intégration d'un système d'équations différentielles ordinaires

Considérons le système différentiel

$$Y' = A\,Y \tag{4.9}$$

satisfaisant à la condition initiale

$$Y(o) = Y_o$$

où Y est le vecteur $(y^{(1)}(x), y^{(2)}(x), \ldots, y^{(m)}(x))^T$

et A, une matrice réelle de dimension m x m.

Utilisons pour résoudre ce système, la méthode $(k,\ell)$ définie par l'opérateur (1.7).

Nous obtenons dès lors la récurrence

$$\sum_{i=o}^{k} \sum_{j=o}^{\ell} \alpha_{ij}\, h^j\, A^j\, Y_{n+i} = 0 \qquad n=0,1,\ldots \tag{4.10}$$

La relation (4.10) s'écrit en vertu de (4.3)

$$\eta_k(hA)\, Y_{n+k} = - \sum_{i=o}^{k-1} \eta_i(hA)\, Y_{n+i} \tag{4.11}$$

Remarquons que

$\eta_k(o) = \alpha_{ko}$ qui a été supposé non nul au chapitre I.

Sans perdre de généralité, posons -le égal à 1.

Par conséquent,

$$\frac{1}{\eta_k(o)} = 1$$

et nous pouvons conclure que la série

$$\frac{1}{\eta_k(\mu)} = \sum_{i=o}^{\infty} c_i\, \mu^i$$

est convergente dans un disque ouvert de centre 0 et de rayon $r_o$.
Si $\rho(A)$ est le rayon spectral(+) de la matrice A, les séries

$$\frac{1}{\eta_k(hA)} = \sum_{i=0}^{\infty} c_i (hA)^i \qquad (4.12)$$

convergent uniformément pour tout h dans $[o, \frac{r_o}{\rho(A)}[$ .
Puisque $\eta_k(o) = 1$, on a que $c_o = 1$ et (4.12) s'écrit :

$$\frac{1}{\eta_k(hA)} = I + h\, G(hA)$$

où I est la matrice unité m x m
et G(hA) est une matrice dont tous les éléments sont bornés uniformément pour tout h dans $[o, \frac{r_o}{\rho(A)}[$ .

De la relation (4.11), nous déduisons que

$$Y_{n+k} = - \sum_{i=o}^{k-1} B_i \; Y_{n+i} \qquad (4.13)$$

où $B_i = \frac{\eta_i(hA)}{\eta_k(hA)}$ $\qquad i=0,1,\ldots,k-1$

Soit

$$Z_n = (Y_{n+k-1}, Y_{n+k-2}, \ldots, Y_n)^\tau \in \mathbb{R}^N$$

où $N = k.m$

Dès lors, nous pouvons écrire de façon matricielle :

$$Z_{n+1} = C_N(h)\, Z_n$$

où $C_N(h) = \begin{pmatrix} -B_{k-1} & -B_{k-2} & \cdots & -B_1 & -B_o \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & I & 0 \end{pmatrix}$

---

(+) Le rayon spectral d'une matrice est la plus grande de ses valeurs propres en module.

De proche en proche, on a donc que

$$Z_n = C_N^n(h)\, Z_o \qquad (4.14)$$

puisque $C_N(h)$ est indépendant de n.
Le comportement de $Z_n$ dépend donc du rayon spectral de la matrice $C_N(h)$.
On peut montrer [ 25 ] que

*$\zeta$ est une valeur propre de $C_N(h)$ si et seulement si $\lambda$ est une valeur propre de A et que*

$$\sum_{i=0}^{k} \eta_i(h\lambda)\, \zeta^i = 0 \qquad (4.15)$$

Notons que, même si A est réelle, $\lambda$ peut être complexe.

Remarque

Observons que l'équation (4.15) s'identifie à la relation (4.6). Dès lors, pour étudier les solutions d'un système du type (4.9), il est suffisant de considérer des équations tests scalaires du type (4.1) où $\lambda$ est une valeur propre de A, $\lambda \in \mathbb{C}$.

4.4.2. Définition d'un système Stiff

Considérons le système différentiel (4.9).
Soient $\lambda_1, \lambda_2, \ldots, \lambda_r$, les valeurs propres de la matrice A.
Le système est appelé *système Stiff* si

$$Re\ \lambda_r \leq Re\ \lambda_{r-1} \leq \ldots\ldots \leq Re\ \lambda_2 \leq Re\ \lambda_1 \leq 0$$

$$et\ Re\ \lambda_r \ll Re\ \lambda_1$$

Remarque

Si le système est non linéaire, le rôle de la matrice A est joué par le jacobien du système.

4.4.3. Problème stiff

Nous intégrons un système différentiel stiff du type (4.9) par une méthode dont la région de stabilité absolue a la configuration représentée à la figure 4.1.

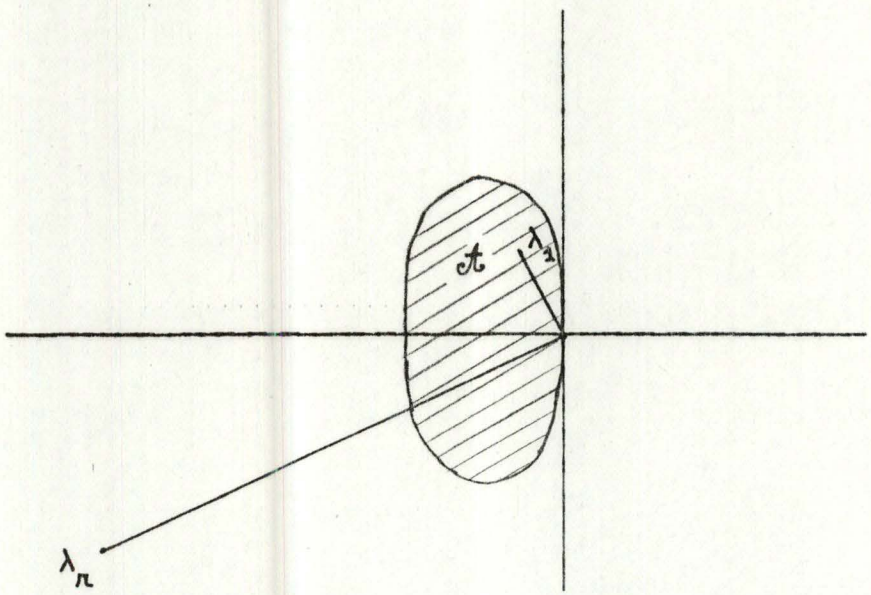



figure 4.1

Rappelons que la solution théorique générale du système (4.9) a la forme :

$$Y(x) = \sum_{i=1}^{r} k_i e^{\lambda_i x} v_i \qquad (4.16)$$

où $v_i$ est le vecteur propre associé à la valeur propre $\lambda_i$ $i=1,2,\ldots,r$
et $k_i$ est une constante $i=1,2,\ldots,r$.
Si le système est stiff, la forme (4.16) assure que la solution théorique tend vers 0 pour x tendant vers l'infini.
Par ailleurs, dans cette solution, le terme dominant est celui qui correspond à $\lambda_1$, tandis que la valeur propre $\lambda_m$ fournit un terme négligeable. C'est pourtant cette dernière qui posera des problèmes pour l'intégration des systèmes stiff.

Posons h=1. Par conséquent, au vu de la figure (4.1), $\mu_r$ n'appartient pas à $\mathcal{A}$. Dès lors, la matrice $C_N(h)$ a une valeur propre de module strictement supérieur à 1 et donc, par (4.14), la solution numérique $Z_n$ explose à l'infini.

Pour que cette situation ne se présente pas, il faut choisir h tel que $h\lambda_r = \mu_r$ soit dans $\mathcal{A}$.

Ceci amène souvent des restrictions très sévères sur le pas, comme nous allons le constater dans l'exemple qui suit.

Exemple [ 28 ]

Soit

$$A = \begin{pmatrix} -21 & 19 & -20 \\ 19 & -21 & 20 \\ 40 & -40 & 40 \end{pmatrix}$$

la matrice du système (4.0), avec les conditions initiales

$$Y(o) = Y_o = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$$

Les valeurs propres de A sont :

$$\lambda_1 = -2$$
$$\lambda_2 = -40 + i\ 40$$
$$\lambda_3 = -\ 40 - i\ 40$$

Il est clair que nous avons affaire à un système stiff.
La solution théorique de ce système s'exprime comme suit :

$$Y(x) = \begin{pmatrix} y_1(x) \\ y_2(x) \\ y_3(x) \end{pmatrix} = \begin{pmatrix} 1/2\ e^{-2x} + 1/2\ e^{-40x}(\cos 40\ x + \sin 40\ x) \\ 1/2\ e^{-2x} - 1/2\ e^{-40x}(\cos 40\ x + \sin 40\ x) \\ -e^{-40x}\ (\cos 40\ x - \sin 40\ x) \end{pmatrix}$$

Cette solution est représentée à la figure 4.2.

Supposons à présent que l'on veuille intégrer ce système par la méthode d'Euler directe, dont la relation de récurrence s'écrit :

$$Y_{n+1} - Y_n = h\ A\ Y_n$$

c'est-à-dire $$Y_{n+1} = [\ I + hA\ ]\ Y_n$$

Par conséquent, nous pouvons exprimer la solution numérique de la façon suivante :

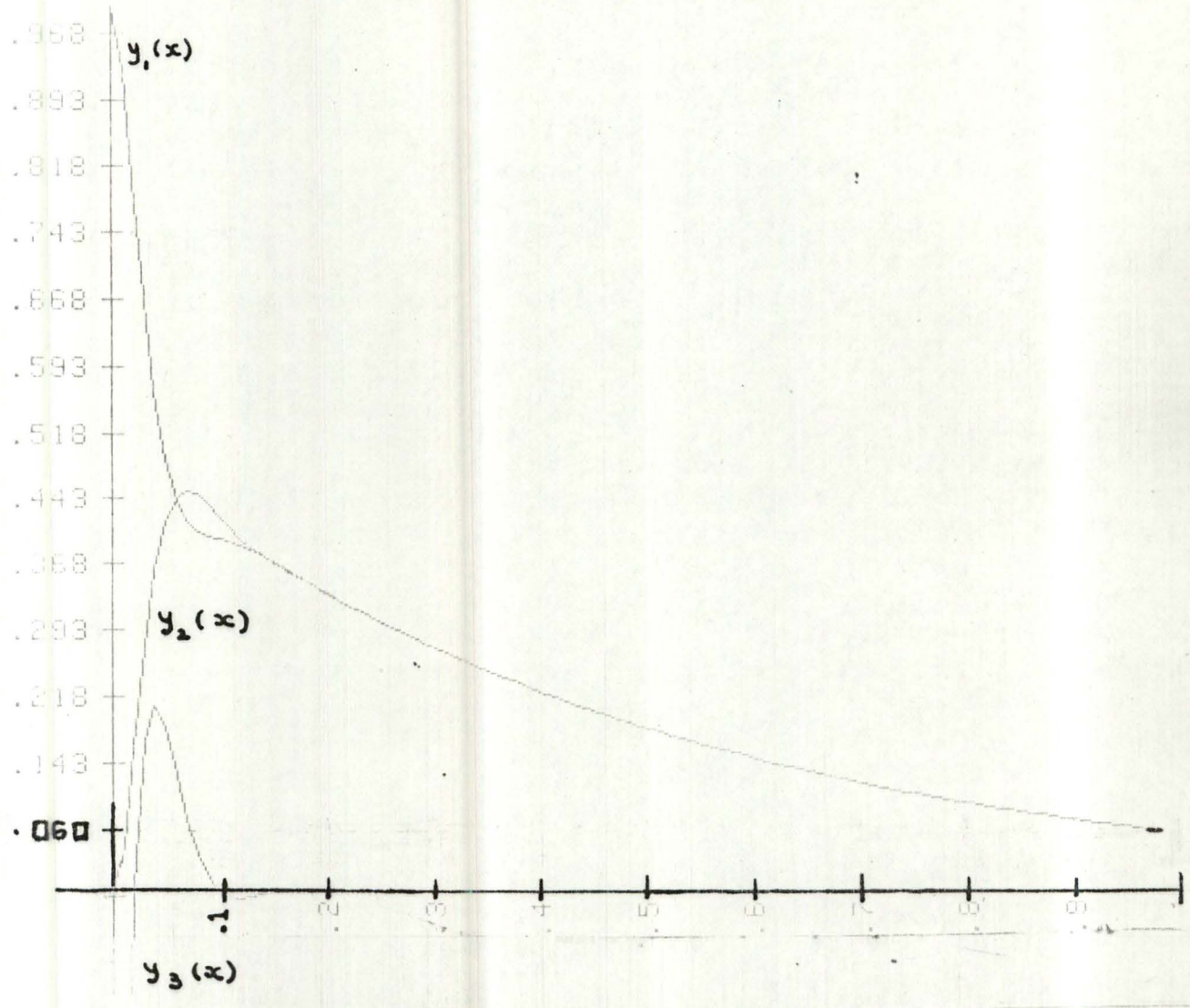



FIGURE 4.2.

$$Y_n = [\, I + hA \,]^n \, Y_o$$

La convergence de cette solution numérique vers 0 dépend du rayon spectral de I + hA, qui lui-même est gouverné par la plus grande des valeurs propres en module de A.

En effet, $Y_n$ tend vers 0 lorsque n tend vers l'infini ssi

$$|1 + h( -40 + i40)|^2 < 1$$

c'est-à-dire

$$(1 - h\,40)^2 + (40\,h)^2 < 1$$

De là, on déduit que

$$h < \frac{1}{40} \, . \qquad (4.17)$$

Si nous intégrons le système entre 0 et 100, un pas aussi petit amène des calculs longs et coûteux.
Notons que l'écart entre les parties réelles de $\lambda_1$ et $\lambda_r$ est souvent beaucoup plus prononcé (par exemple, $Re\ \lambda_1 = -1$ et $Re\ \lambda_r = -10^6$). Les restrictions sur le pas sont alors beaucoup plus sévères.

Si nous examinons la figure 4.2, la restriction (4.17) semble acceptable au voisinage de l'origine, car la variation des solutions théoriques est brutale. Par contre, lorsque l'on s'éloigne de l'abscisse 0.1, cette variation est de moins en moins marquée et on serait tenté d'augmenter le pas. Or, cela n'est pas permis car $h\lambda_r$ serait alors hors de $\mathcal{S}$ et la solution numérique du système exploserait à l'infini.

En conclusion, pour intégrer des systèmes stiff, il est souhaitable d'utiliser des méthodes dont le domaine de stabilité absolue contient $\mathbb{L}$, car alors tout pas est permis.
Les méthodes qui vérifient une telle propriété sont appelées les méthodes *A-stables*.

En vertu de la remarque de la section (4.4.1), la A-stabilité d'une méthode sera étudiée par le biais des équations tests de type (4.1) et par le polynôme de caractérisation $\Phi(\zeta,\mu)$ (4.6) associé.

## 4.5. METHODES A-STABLES

*Une méthode numérique est A-stable, si appliquée à l'équation* $y' = \lambda y$ *avec* $Re\ \lambda < 0$, *elle engendre des valeurs* $y_n$ *telles que*

$$\lim_{n\to\infty} y_n = 0 \qquad \text{pour tout } h > 0.$$

Par la formule (4.7), cela revient à dire que pour tout $\mu$ de partie réelle strictement négative, on a

$$|\zeta_i(\mu)| < 1 \qquad i=1, \ldots, s$$

où $\zeta_i(\mu)$ sont les racines du polynôme $\Phi(\zeta,\mu)$ (4.6).

° ° °

# CHAPITRE V

## CARACTÉRISATION ALGÉBRIQUE DE LA A-STABILITÉ

## 5.1. INTRODUCTION

La A-stabilité, telle que nous l'avons définie au chapitre IV nécessite le traitement de polynômes en 2 variables.
Dans ce chapitre, nous nous efforcerons d'abord de créer un lien entre la A-stabilité et les propriétés de tels polynômes.
Ensuite, nous nous poserons la question de savoir s'il y a moyen de vérifier la propriété de A-stabilité en un nombre fini d'opérations arithmétiques.

## 5.2. POLYNOME CANONIQUE D'UNE METHODE $(k,\ell)$

### 5.2.1. Changement de variable

Considérons l'équation caractéristique :

$$\Phi(\xi,\mu) = \sum_{i=0}^{k} \eta_i(\mu)\ \xi^i = \sum_{j=0}^{\ell} \rho_j(\xi)\ \mu^j = 0 \qquad (5\text{-}1)$$

Pour caractériser la A-stabilité de la méthode à laquelle est associée (5-1), il est utile de considérer le changement de variable suivant :

$$z = \frac{\xi+1}{\xi-1} \qquad (5\text{-}2)$$

$$\xi = \frac{z+1}{z-1}$$

Ceci a pour effet de transformer l'intérieur du disque unité du plan des $\xi$ en le demi-plan gauche: $\mathbb{L}$ et le cercle unité en l'axe imaginaire. En particulier, l'image de $\xi = 1$ vaut l'infini.
Lorsque l'on travaille avec la variable $z$ , on utilise à la place des polynômes caractéristiques $\rho_j(\xi)$, les polynômes $r_j(z)$ qui vérifient :

$$r_j(z) = (z-1)^k\ \rho_j\ (\frac{z+1}{z-1}) \qquad j = 0,1,\ldots,\ell \qquad (5\text{-}3)$$

$$= \sum_{i=0}^{k} a_{ij}\ z^i$$

On introduit aussi des polynômes $e_i(\mu)$ $i = 0,1,\ldots, k$, qui sont tels que

$$e_i(\mu) = \sum_{j=0}^{\ell} a_{ij}\ \mu^j \qquad i = 0,1,\ldots,k \qquad (5\text{-}4)$$

Par ailleurs, la transformation (5-2) nous conduit à utiliser, à la place de $\Phi(\xi,\mu)$, le polynôme $H(z,\mu)$ défini par la relation suivante :

$$H(z,\mu) = (z-1)^k\ \Phi(\frac{z+1}{z-1},\mu) \qquad (5\text{-}5)$$

qui, par (5-1) peut encore s'écrire

$$H(z,\mu) = \sum_{j=0}^{\ell} (z-1)^k\ \rho_j\ (\frac{z+1}{z-1})\ \mu^j \qquad (5\text{-}6)$$

Afin de pouvoir rattacher la A-stabilité à des propriétés de polynômes, il nous est utile d'introduire des polynômes $R_j(z)$ :

$$R_j(z) = (-1)^j r_j(z) = \sum_{i=0}^{k} a_{ij} (-1)^j z^i = \sum_{i=0}^{k} A_{ij} z^i \quad j = 0,1,\ldots,\ell \qquad (5\text{-}7)$$

Nous introduisons ces polynômes dans (5-6) de la manière suivante :

$$H(z,\mu) = \sum_{j=0}^{\ell} (-1)^j (z-1)^k \rho_j \left(\frac{z+1}{z-1}\right) (-\mu)^j$$

Par (5-3), cela revient à noter

$$H(z,\mu) = \sum_{j=0}^{\ell} (-1)^j r_j(z) (-\mu)^j$$

Si nous posons $q = -\mu$ , nous obtenons par (5-7) que

$$H(z,q) = \sum_{j=0}^{\ell} R_j(z) q^j$$

Le polynôme $H(z,q)$ peut aussi s'écrire en faisant apparaître les puissances successives de z.
En effet, si nous portons (5-7) dans (5-8) nous avons que

$$H(z,q) = \sum_{j=0}^{\ell} \sum_{i=0}^{k} A_{ij} z^i q^j$$

$$= \sum_{i=0}^{k} \sum_{j=0}^{\ell} A_{ij} q^j z^i$$

Posons $E_i(q) = \sum_{j=0}^{\ell} A_{ij} q^j \quad i=0,\ldots,k$

Dès lors, nous obtenons que

$$H(z,q) = \sum_{i=0}^{k} E_i(q) z^i$$

Notons le lien entre les polynômes $E_i$ et $e_i$.
Il est clair que

$$E_i(q) = e_i(\mu)$$

c'est-à-dire, puisque $q = -\mu$

$$E_i(-\mu) = e_i(\mu)$$

Au chapitre IV, nous avions fait l'hypothèse que $\Phi(\xi,\mu)$ ne pouvait pas être factorisé selon la forme :

$$\Phi(\xi,\mu) = P(\mu) \quad R(\xi,\mu)$$

où $P(\mu)$ est un polynôme en $\mu$ de degré au plus $\ell$.

Après le changement de variable (5-2), cette hypothèse se traduit de la manière suivante :

H(z,q) ne peut se factoriser selon la forme

$$H(z,q) = v(q)\, H^{*}(z,q) \qquad (5\text{-}9bis)$$

où v(q) est un polynôme en q de degré au plus $\ell$.

5.2.2. Définitions

a. H(z,q) défini par les égalités (5-8) et (5-9) est appelé *le polynôme canonique* de la méthode (k,$\ell$) (1-3).
H(z,q) est donc un polynôme en 2 variables, à coefficients réels, de degré k en z et de degré $\ell$ en q.

b. La fonction algérgique z(q) obtenue en résolvant

$$H(z(q),q)) = 0$$

est appelée la *z-fonction canonique* de la méthode(k-$\ell$).

c. De manière analogue, nous appellerons *q-fonction canonique*, la fonction algébrique qui vérifie :

$$H(z,q(z)) = 0.$$

5.2.3. Pôles d'un polynôme en 2 variables [37]

Soit P(x,y) un polynôme en deux variables x et y.
Les pôles du polynôme P(x,y) sont les $x \in \mathbb{C}$ tels que :

$$P(x,\infty) = 0.$$

Si nous écrivons P(x,y) en faisant apparaître les puissances successives de y, cela signifie que les pôles de P(x,y) sont les racines du polynôme en x, qui est le coefficient du terme en y de plus haut degré.

Notons que cette notion de pôle peut aussi être définie en considérant les $y \in C$ tels que :

$$P(\infty,y) = 0.$$

5.3. CARACTERISATION DE LA A-STABILITE

Rappellons qu'une (k-$\ell$) méthode est A-stable ssi les racines de $\Phi(\xi,\mu)$ sont à l'intérieur du disque unité, pour tout $\mu$ t.q. Re $\mu < 0$.
En termes du polynôme canonique, cela revient à dire que les racines de H(z,q) doivent se trouver dans $\mathbb{L}$ pour tout q tel que Re $q > 0$.
Nous traduirons cela dans la formule suivante :

Une (k-$\ell$) méthode est A-stable

ssi

H(z,q), son polynôme canonique, est tel que

$$\text{Re } q > 0 \Rightarrow \text{Re } z(q) < 0 \qquad (5\text{-}10)$$

pour q et z(q) vérifiant H(z,q(z)) = 0

La formule (5-10) est logiquement équivalente à

$$\text{Re } z \geqslant 0 \Rightarrow \text{Re } q(z) \leqslant 0 \qquad (5\text{-}11)$$

5.3.1. Théorème 5-1 [19]

*Soient z et q(z) vérifiant*

$H(z,q(z)) = 0.$

*La condition (5-11) est équivalente à*

$$\text{Re } z > 0 \Rightarrow \text{Re } q(z) \leqslant 0 \qquad (5\text{-}12)$$

*ou encore à*

$$\begin{aligned} &\text{Re } z > 0 \Rightarrow \text{Re } q(z) < 0 \\ &\text{Re } z = 0 \Rightarrow \text{Re } q(z) \leqslant 0 \end{aligned} \qquad (5\text{-}13)$$

Démonstration

Montrons d'abord l'équivalence de (5-11) et (5-12)

L'implication [(5-11) ⇒ (5-12)] est triviale.

Pour montrer [(5-12) ⇒ (5-11)], il nous suffit de voir que (5-12) entraîne :

$$\text{Re } z = 0 \Rightarrow \text{Re } q(z) \leqslant 0 \qquad (5\text{-}14)$$

Si z est un pôle de H(z,q), alors q vaut l'infini et le signe de sa partie réelle est donc indéterminé. Par conséquent la relation (5-14) sera toujours vérifiée dans ce cas.

Sans perdre de généralité, nous pouvons donc considérer que z n'est pas un pôle.

Supposons alors, par l'absurde, que (5-14) ne soit pas vraie, c'est-à-dire qu'il existe une paire $(z_0, q(z_0))$ telle que

$$H(z_0 , q(z_0)) = 0$$

avec $\text{Re } z_0 = 0$ et $\text{Re } q(z_0) \geqslant 0$.

Puisque z n'est pas un pôle, nous pouvons affirmer que les racines du polynôme H(z,q(z)) sont des fonctions continues des coefficients [31].

Par conséquent q(z) est une fonction continue de z.

Nous pourrons donc trouver une paire $(z_1, q(z_1))$ telle que

$$H(z_1, q(z_1)) = 0$$

$$\text{avec Re } z_1 > 0 \text{ et Re } q(z_1) > 0 \qquad (5\text{-}15)$$

Cette relation (5-15) contredit (5-12).
Notre hypothèse de l'absurde est donc fausse et nous concluons que (5-14) est vraie.

Montrons l'équivalence de (5-11) et (5-13)

Considérons les relations (5-13). Il est immédiat que celles-ci entraînent (5-11).
Il nous reste donc à prouver que [ (5-11) ⇒ (5-13) ]
Grâce à l'équivalence entre (5-11) et (5-12), pour obtenir que
[ Re $z > 0 \Rightarrow$ Re $q(z) < 0$ ], il nous suffit de montrer que

$$\text{Re } z > 0 \Rightarrow \text{Re } q(z) \neq 0 \qquad (5\text{-}16)$$

ce qui est logiquement équivalent à

$$\text{Re } q = 0 \Rightarrow \text{Re } z(q) \leqslant 0 \qquad (5\text{-}17)$$

Si q est un pôle de $H(z(q),q)$, z devient l'infini et (5-17) est vérifiée.
Plaçons nous donc dans le cas où q n'est pas un pôle et, par l'absurde, supposons que (5-17) ne soit pas vraie c'est-à-dire qu'il existe une paire $(z(q_0),q_0)$ telle que :

$$H(z(q_0),q_0) = 0$$
$$\text{avec Re } q_0 = 0 \text{ et Re } z(q_0) > 0$$

La continuité des racines de $H(z(q),q)$ par rapport aux coefficients, nous permet alors de trouver une paire $(z(q_1),q_1)$ telle que

$$H(z(q_1),q_1) = 0$$
$$\text{avec Re } q_1 > 0 \text{ et Re } z(q_1) > 0 \qquad (5\text{-}18)$$

Cette dernière relation (5-18) contredit (5-10) qui est logiquement équivalente à (5-11).
Par conséquent, notre hypothèse de l'absurde est fausse et nous avons (5-17).
La dernière chose que nous devons voir est que (5-11) ou (5-12) qui lui est équivalent, conduit à

[ Re $z=0 \Rightarrow$ Re $q \leqslant 0$ ]

Ceci a été démontré dans la preuve de l'implication [ (5-12) ⇒ (5-11)]

■

Remarque

L'argument de continuité des racines q du polynôme $H(z,q)$ par rapport à ses coefficients est valable pour autant que $H(z,q)$ ne puisse pas se factoriser sous la forme

$$H(z,q) = V(q)\, H^*(z,q)$$

L'hypothèse (5-9bis) que nous avons faite sur le polynôme canonique dans la section 5-2 permet donc l'emploi de cet argument.

5.3.2. Définition

f(z) est appelée fonction positive si et seulement si

$$\text{Re } z \geqslant 0 \Rightarrow \text{Re } f(z) \geqslant 0 \qquad (5\text{-}19)$$

Si, en outre, les coefficients de f(z) sont réels, alors f(z) est une fonction positive réelle.

Ces définitions restent valables dans le cas où f(z) a la forme d'une fraction :

$$f(z) = \frac{n(z)}{d(z)}$$

où n(z) et d(z) sont des polynômes et d(z) non constamment nul.

5.3.3. Théorème 5-2 [19]

*Une méthode (k,ℓ) est A-stable si et seulement si l'opposé de sa q-fonction canonique, ou l'opposé de sa z-fonction canonique, est une fonction algébrique positive.*

Ce théorème est évident si l'on interprète (5-11) et (5-13) au vu de (5-19).

5.3.4. Définitions

Un polynôme p(z), non nul, est Hurwitzien s'il n'a pas de racines dans R. Il est strictement Hurwitzien si toutes ses racines sont dans L.

Ansell [1] a généralisé cette notion d'Hurwitzité pour un polynôme à deux variables, de la manière suivante :

H(z,q(z)), polynôme réel non nul, à deux variables est Hurwitzien au sens strict, si et seulement si il n'a pas de racines

$$\text{ni dans Re } z > 0,\ \text{Re } q(z) > 0 \qquad (5\text{-}20a)$$

$$\text{ni dans Re } z > 0,\ \text{Re } q(z) = 0 \qquad (5\text{-}20b)$$

$$\text{ni dans Re } z = 0,\ \text{Re } q(z) > 0 \qquad (5\text{-}20c)$$

5.3.5. Théorème 5-3 [19]

*Une méthode (k,ℓ) est A-stable si et seulement si son polynôme canonique H(z,q(z)) est un polynôme en 2 variables, Hurwitzien au sens strict.*

Démonstration

L'Hurwitzité d'un polynôme peut encore s'exprimer de la façon suivante :

Soit (z,q(z)) tel que H(z,q(z)) = 0

Les relations (5-20a) et (5-20b) sont équivalentes à

$$\text{Re } z > 0 \Rightarrow \text{Re } q(z) < 0$$

et (5-20c) est équivalente à

$$\text{Re } z = 0 \Rightarrow \text{Re } q(z) \leqslant 0$$

Par conséquent, en interprétant les relations (5-13) nous obtenons immédiatement le théorème 5-3.

■

Notre problème de A-stabilité s'est ainsi ramené à la vérification de l'Hurwitzité stricte d'un polynôme en deux variables. Il est donc naturel de rechercher un critère d'Hurwitzité stricte.

5.3.6. Théorème 5-4 [1]

*Soit $H(z,q)$ un polynôme réel, non constamment nul.*
*$H(z,q)$ est un polynôme en 2 variables, Hurwitzien au sens strict si et seulement si :*
*(i) le polynôme réel en z, $H(z,1)$ est strictement Hurwitzien.*
*(ii) pour tout w réel, fini, le polynôme complexe en q, $H(iw,q)$ est Hurwitzien.*
*(iii) $H(z,q)$ n'a pas de facteur $(q-q_0)$ avec Re $q_0 = 0$.*

Démonstration

1. Montrons la nécessité

Soit $(z,q)$ tel que $H(z,q) = 0$

L'Hurwitzité stricte de $H(z,q)$ nous permet de dire, par (5-20a) et (5-20c) que si Re $q > 0$ alors Re $z < 0$.

Par conséquent $H(z,1)$ a toutes ses racines dans $\mathbb{L}$ et est donc strictement Hurwitzien.

D'autre part, par (5-20c) nous avons aussi que si Re $z = o$ alors Re $q \leqslant 0$ ce qui assure l'Hurwitzité de $H(iw,q)$.

L'Hurwitzité conduit également à (iii). En effet, supposons qu'il existe un facteur $(q-q_0)$ avec Re $q_0 = 0$

Dès lors, on peut écrire

$$H(z,q) = (q-q_0)\, H'(z,q)$$

ce qui assure que

$$H(z,q_0) = 0 \quad \text{pour tout } z \in \mathbb{C} .$$

Par conséquent, on pourrait avoir des racines de $H(z,q_0)$ telles que

$$\text{Re } z \leqslant 0 \text{ et Re } q_0(z) = 0$$

Ce qui serait contraire à l'Hurwitzité stricte (5-20b).

2. Montrons la suffisance

Si $H(z,q)$ est constant, comme par hypothèse $H(z,q)$ est non nul, l'Hurwitzité stricte est immédiate.

Supposons donc $H(z,q)$ non constant.

Dans ce cas, $H(z,q)$ peut se décomposer selon un produit de polynômes irréductibles (c'est-à-dire de polynômes qui ne peuvent plus être exprimés comme produit de polynômes non constants).

Dans cette factorisation, nous noterons :

$h(z)$ le produit des polynômes irréductibles qui sont indépendants de q.

$h(q)$ le produit des polynômes irréductibles qui sont indépendants de z.

Le produit de tous les autres polynômes irréductibles s'écrira comme le produit de $H_s(z,q)$ et $H_r(z,q)$ où $H_r(z,q)$ regroupe seulement les polynômes qui sont, à une constante multiplicative près, égaux à certains facteurs de $H_s(z,q)$. Par conséquent, les polynômes de $H_s(z,q)$ ne sont pas multiples constants les uns des autres.

Nous écrirons donc $H(z,q)$ sous la forme

$$H(z,q) = h(z)\ h(q)\ H_r(z\,q)\ H_s(z\,q) \qquad (5\text{-}21)$$

La condition i) assure $h(z)$ n'a pas de racine dans $\overline{R}$.

D'autre part ii) et iii) permettent de dire qu'il n'existe pas de racine de $h(q)$ telle que $\mathrm{Re}\ q \geqslant 0$.

Par conséquent, étant donné (5-21) et grâce aux définitions de $H_s(z,q)$ et $H_r(z,q)$, nous aurons établi l'Hurwitzité au sens trict de $H(z,q)$ si nous parvenons à montrer que $H_s(z,q)$ est Hurwitzien au sens strict.

La résolution de l'équation $H_s(z,q) = 0$ nous permet d'obtenir q comme fonction algébrique de z : $q(z)$

Considérons alors la transformation bilinéaire :

$$S = \frac{q+1}{q-1} \qquad (5\text{-}22)$$

En portant l'expression de $q(z)$ dans (5-22), on obtient S, fonction algébrique de z. Nous la noterons $S(z)$.

Pas 1 - S(z) a tous ses pôles dans L

La condition ii) assure que $q(iw)$ sera tel que

$$\mathrm{Re}\ q(iw) \leqslant 0$$

Par conséquent, nous aurons que

$$|\mathrm{Re}\ (q(iw)+1)| < |\mathrm{Re}\ (q(iw)-1)|$$

et donc $|S(z)| < 1$ pour $\mathrm{Re}\ z=0$ (5-23)

La fonction $S(z)$ étant ainsi bornée sur l'axe imaginaire, elle ne peut y avoir de pôles.

Ceci nous permet d'affirmer que $S(z)$ n'a pas de pôles à l'infini. (5-24)

Par ailleurs, de i) nous tirons que $H_s(z,1)$ a toutes ses racines dans L

Remarquons que les racines de $H_s(z,1)$ sont les pôles de $S(z)$.

En tenant compte de (5-24) nous concluons donc que tous les pôles de $S(z)$ sont finis et situés dans L .

Pas 2 - Représentation conforme de S(z)

La fonction algébrique $S(z)$ détermine une surface de Riemann, comportant un nombre fini de régions ou feuilles [8]. Si nous considérons la surface de Riemann correspondant à $\overline{R}$, nous aurons plusieurs feuilles $F_i$ pour $i = 1,\ldots,n$ chacune contenant l'axe imaginaire, endroit où elles se joignent.

Par la définition des feuilles d'une surface de Riemann et puisque S(z) n'a pas de pôles sur l'axe imaginaire, nous pouvons dire que sur chaque feuille $F_i$, $|S(z)|$ est injective et continue à l'intérieur et sur la frontière de $F_i$.
Par conséquent, sur chaque $F_i$, $|S(z)|$ admet une valeur maximum et puisque les $F_i$ sont en nombre fini, $|S(z)|$ possède un maximum sur $\bigcup_{i=1}^{n} F_i$.
Par ailleurs, choisissons arbitrairement $z_0$ dans l'intérieur de R.
Les valeurs de S(z), pour z dans un voisinage de $z_0$, peuvent être données de façon paramétrique par plusieurs fonctions :

$$s_i(t_j) \quad i=1,\ldots,n$$

chacune étant analytique dans le voisinage considéré (ceci par le pas 1).
Les $t_j$ sont donc des puissances fractionnelles positives de $(z-z_0)$.
Par la factorisation (5-21) nous avons éliminé de S(z) les branches constantes.
Dès lors, le théorème du maximum du module assure que chacune des fonctions $s_i(t_j)$ i = 1, ...,n atteint son module maximum sur l'axe imaginaire. Par conséquent $|S(z)|$ atteint aussi son maximum sur l'axe imaginaire.
Par la relation (5-23) nous avons dès lors que

$$|S(z)| < 1 \text{ pour tout } z \text{ tel que } \mathrm{Re}\ z > 0 \qquad (5\text{-}25)$$

Etant donné la forme (5-22) nous constatons que

- si $\mathrm{Re}\ q(z) > 0$ alors $|S(z)| > 1$ pour tout z et donc en particulier pour z tel que $\mathrm{Re}\ z \geqslant 0$
- si $\mathrm{Re}\ q(z) = 0$ alors $|S(z)| = 1$ pour tout z et donc en particulier pour z tel que $\mathrm{Re}\ z > 0$.

En conséquence, (5-25) sera vérifiée si et seulement si

$H_s(p,q)$ n'a de racines

ni dans $\mathrm{Re}\ z > 0$ et $\mathrm{Re}\ q > 0$
ni dans $\mathrm{Re}\ z = 0$ et $\mathrm{Re}\ q > 0$
ni dans $\mathrm{Re}\ z > 0$ et $\mathrm{Re}\ q = 0$

ce qui est une traduction de l'Hurwitzité stricte du polynôme $H_s(p,q)$.

■

Remarques

1. Si nous prenons en considération l'hypothèse (5-9bis), la condition iii) du théorème 5-4 est immédiatement vérifiée et il n'y a pas lieu de la prendre en considération dans l'énoncé.
2. La propriété de A-stabilité est symétrique en les variables z et q du polynôme canonique H(z,q). Par conséquent, le théorème 5-4 peut s'énoncer en permutant les rôles de z et de q.

5.3.7. Théorème 5-5 [ 19 ]

Avant de clore cette section énonçant les théorèmes fondamentaux de la caractérisation algébrique de la A-stabilité, nous énonçons une dernière condition nécessaire, qui nous sera utile dans la suite pour établir l'ordre maximum que peut atteindre une méthode A-stable.

*Soit une méthode* $(k-\ell)$ *et son polynôme canonique* $H(z,q)$ *vérifiant l'hypothèse (5-9bis).*

*Si la méthode est A-stable, alors :*

*i)* $R_j(z)$ $j=0,1,\ldots,\ell$ ;

$E_i(q)$ $i=1,\ldots,k$, *sont réels, non nuls et Hurwitziens.*

*ii) les fonctions* $\dfrac{R_{j-1}(z)}{R_j(z)}$ $j=1,2,\ldots,\ell$

*et* $\dfrac{E_{i-1}(q)}{E_i(q)}$ $i=1,2,\ldots,k$

*sont des fonctions positives réelles.*

*iii) tous les coefficients* $A_{ij}$ *ont le même signe.*

Nous omettons la démonstration de ce théorème car elle utilise un lemme énoncé par Jeltsch [ 25 ] et dont la démonstration offre peu d'intérêt dans le cadre de ce chapitre.

## 5.4. VERIFICATION DE LA A-STABILITE EN UN NOMBRE FINI D'ETAPES

### 5.4.1. Introduction

Le problème que nous nous posons à présent est comment vérifier la A-stabilité en un nombre fini d'opérations arithmétiques ?
Le théorème 5-4 suggère qu'il y a moyen de résoudre ce problème en utilisant l'algorithme de Routh-Hurwitz [ 18 ] pour vérifier i). La condition ii) revient à dénombrer les racines d'un polynôme à coefficients complexes dans un demi-plan. Ce problème peut être résolu grâce au formalisme de Bézout [ 23 ]
C'est cette idée qu'a suivie Ansell [ 1 ] dans son théorème II.
Ansell considère un polynôme particulier, vérifiant l'hypothèse (5-9bis) où V(q) serait de la forme $(q-q^*)$ avec Re $q^* = 0$

Sans utiliser le vocabulaire de l'Hurwitzité, Rubin [37] a, de son côté écrit un algorithme semblable.
Son mérite est cependant d'avoir considéré des polynômes quelconques, sans émettre l'hypothèse (5-9bis) concernant la factorisation.
La technique est donc applicable à la vérification de la A-stabilité pour des méthodes d'intégration, dont le polynôme de caractérisation n'a pas une forme

simple, telles les méthodes composites.
Par ailleurs, Rubin a su tirer profit de certaines propriétés des mineurs principaux emboîtés de la matrice de Bézout, de façon à obtenir un critère moins coûteux du point de vue calcul et dont l'énoncé montre que la A-stabilité peut être vérifiée en un nombre fini d'opérations.

Puisqu'il paraît de portée plus générale, nous adopterons le procédé employé par Rubin [ 37 ], même si souvent nous poserons une hypothèse simplificatrice.

5.4.2. Utilisation de la matrice de Bézout pour la localisation des racines d'un polynôme dans un demi-plan

Le but de cette section est de justifier le procédé employé par Rubin [37] . Nous ne ferons donc aucune démonstration. Celles-ci peuvent être obtenues dans Househölder [23] et Gantmacher [18]

Soit le polynôme

$$f(z) = a_0 + a_1 z + a_2 z^2 + \dots a_n z^n$$

où $a_i \in C$ pour $i = 0,1,\dots,n$.

Nous nous proposons de compter le nombre de racines de ce polynôme dans un demi-plan. Notons qu'il est toujours possible de passer d'un demi-plan à un autre, ou d'un demi-plan à un disque unité, par une transformation fractionnelle linéaire Sans perdre de généralité, nous pouvons donc considérer un demi-plan quelconque. Tout comme Hermite (+) l'a fait, nous compterons le nombre de racines d'un polynôme au-dessus de l'axe réel.

Première étape

Construisons la matrice $B(f,\bar{f})$ qui est la matrice de Bézout associée à $f(z)$ et $\bar{f}(z)$ où $\bar{f}(z)$ est le polynôme $f(z)$ dans lequel les coefficients ont été conjugués [23]

$$B(f,\bar{f}) = \begin{pmatrix} |a_0\,\bar{a}_1| & |a_0\,\bar{a}_2| & \dots\dots\dots & |a_0\,\bar{a}_n| \\ |a_0\,\bar{a}_2| & |a_0\,\bar{a}_3| + |a_1\,\bar{a}_2| & \dots & |a_0\,\bar{a}_{n+1}| + |a_1\,\bar{a}_n| \\ \vdots & & & \\ |a_0\,\bar{a}_n| & |a_0\,\bar{a}_{n+1}| + |a_1\,\bar{a}_n| & \dots\dots\dots & \end{pmatrix}$$

(+) C. HERMITE - Sur le nombre des racines d'une équation algébrique comprises entre des limites données.
J. Reine Angew. Math. 52 (1856), pp. 39-51.

où • $|a_j.\bar{a}_k| = a_j \bar{a}_k - a_k \bar{a}_j$

• les coefficients d'indice supérieur à n sont nuls.

Définissons alors une nouvelle matrice

$$A(f) = i\ B(f,\bar{f}) \qquad (5\text{-}26)$$

Comme $B(f,\bar{f})$, A(f) est symétrique, carrée, de dimension n.
Notons que, puisque f(z) est un polynôme à coefficients complexes, il peut se mettre sous la forme suivante :

$$f(z) = f_R(z) + i\ f_I(z)$$

où $f_R(z)$ et $f_I(z)$ sont des polynômes réels correspondant respectivement aux parties réelle et imaginaire de f(z).

En développant les éléments de A(f), on voit alors très facilement que

$$A(f) = 2\ B(f_R,f_I) \qquad (5\text{-}27)$$

C'est cette forme de A(f) que nous utiliserons par la suite comme Rubin [ 37 ] le fait.

Deuxième étape : théorème d'Hermite [ 23 ]

*La signature de A(f) est égale à la différence entre le nombre de racines de f(z) se situant strictement au-dessus de l'axe réel et strictement en-dessous de l'axe réel.*

Troisième étape

Le théorème d'Hermite nécessite la détermination de la signature de la matrice A(f). Ceci peut se faire par la méthode de Jacobi.
Pour définir la signature de A(f), nous considérons la forme quadratique qui lui est associée :

$$Q(x,x) = x\ A(f)\ x^{\tau}$$

où x est un vecteur ligne à n éléments.

La signature de A(f), notée $\sigma$ est, par définition, la différence entre le nombre $\pi$ de carrés positifs et le nombre $\nu$ de carrés négatifs dans la forme quadratique Q(x,x). Nous avons donc

$$\sigma = \pi - \nu$$

et nous savons que le rang r de A(f) est tel que

$$r = \pi + \nu$$

Le théorème de Jacobi nous permet alors de déterminer la signature de A(f) de la manière suivante [18]

*Si pour la forme quadratique*

$$Q(x,x) = \sum_{i,k=1}^{n} a_{ik}\ x_i\ x_k$$

*de rang r, l'inégalité*

$$\nabla_k = A\begin{pmatrix} 1 & 2 & \ldots & k \\ 1 & 2 & \ldots & k \end{pmatrix} \neq 0 \qquad \text{pour } k = 1, 2, \ldots r \quad (+)$$

---

(+) $A\begin{pmatrix} i_1, & i_2, & \ldots, & i_k \\ j_1, & j_2, & \ldots, & j_k \end{pmatrix}$ est une notation pour le déterminant de la sous-matrice (kxk) de A, formée des lignes $i_1, i_2, \ldots, i_k$ et des colonnes $j_1, j_2, \ldots, j_k$

*est satisfaite, le nombre $\pi$ de carrés positifs et le nombre $\nu$ de carrés négatifs de $Q(x,x)$, coïncident respectivement avec le nombre P de permanences de signe et le nombre V de variations de signe dans la suite :*

$$1,\ \nabla_1,\ \nabla_2,\ \ldots,\ \nabla_n$$

*La signature de A est alors*

$$\sigma = r - 2V$$

Par conséquent, si tous les $\nabla_i$ (qui sont aussi appelés mineurs principaux emboîtés) sont strictement positifs, nous aurons que

$$\sigma = r$$

et donc, par le théorème d'Hermite, toutes les racines de f(z) seront situées strictement au-dessus de l'axe réel.

Remarques

1. L'application du théorème de Jacobi nécessite que les $\nabla_k$ soient non nuls pour $k = 1, 2, \ldots, r$.
   Il se pose donc un problème pour le nombre de variations de signe dans la suite $(1, \nabla_1, \ldots, \nabla_r)$ si certains de ses éléments sont nuls.
   Gantmacher [18, p. 248] a résolu ce problème en disant que

   si $\nabla_l \neq 0$
   $\nabla_{l+1} = 0 \ldots\ldots = \nabla_{l+p} = 0$
   $\nabla_{l+p+1} \neq 0$
   alors, le nombre de variations de signe de la suite $(1, \nabla_1, \ldots \nabla_r)$ vaut

   si p est impair : $\frac{p+1}{2}$

   si p est pair : $\frac{p+1-\varepsilon}{2}$ où $\varepsilon = (-1)^{P/2}$ signe $\frac{\nabla_{l+p+1}}{\nabla_l}$

   Par cette dernière formule, il est clair que la nullité d'un terme dans la suite $(1, \nabla_1, \ldots, \nabla_r)$ entraîne au moins un changement de signe.

2. La matrice de Bézout associée à deux polynômes à coefficients réels a un autre usage que celui de compter les racines dans un demi-plan.
   En effet, ce formalisme permet de dire si deux polynômes à coefficients réels admettent un diviseur commun ou pas.
   Householder [23] démontre que deux polynômes réels de degré n, f(z) et g(z) ont un diviseur commun si et seulement si le déterminant de la matrice de Bézout qui leur est associée est nul.

## 5.4.3. Définitions du polynôme transformé et de la A-stabilité transformée

Considérons de manière générale

$$\Phi(\mu,\xi) = \sum_{i=0}^{m} \sum_{j=0}^{n} \alpha_{ij}\, \xi^i\, \mu^j \qquad (5\text{-}28)$$

un polynôme réel en deux variables, qui est le polynôme de caractérisation d'une méthode d'intégration.

Ce polynôme est de degré m en $\xi$ et nous ne faisons aucune hypothèse quant à son irréductibilité ou ses factorisations possibles.

Nous effectuons alors deux transformations successives :

- la première change le disque unité du plan des $\xi$ en le demi-plan gauche $\mathbb{L}$ des y (5-2)

  On considère ainsi le nouveau polynôme

$$P(\mu,y) = (y-1)^m \; \Phi(\mu, \frac{y+1}{y-1}) \qquad (5\text{-}29)$$

- la seconde opère une rotation de $\pi/2$, dans le sens anti-horlogique, sur la variable $\mu$ de P, et une rotation de $\pi/2$, dans le sens horlogique, sur la variable y de P.

  On définit dès lors

$$P(w,z) = P(iw, -iz) \qquad (5\text{-}30)$$

En combinant (5-29) et (5-30), et en négligeant un facteur $(i)^{-m}$, on peut écrire P'(w,z), appelé polynôme transformé associé à $\Phi(\mu,\xi)$ de la façon suivante :

$$P'(w,z) = (z-i)^m \; \Phi(iw, \frac{z+i}{z-i}) \qquad (5\text{-}31)$$

Définition

Soit P'(w,z) un polynôme complexe.

P'(w,z) satisfait le critère de A-stabilité transformée si et seulement si, pour tout w dans le demi-plan supérieur $\mathbb{G}$, les racines de P'(w,.) sont situées dans le demi-plan inférieur $\mathbb{H}$.

Théorème 5-6 [37]

*Soit $\Phi(\mu,\xi)$ un polynôme réel de degré m en $\xi$ et P'(w,z) son polynôme transformé, de degré m' en z.*

*$\Phi(\mu,\xi)$ satisfait la A-stabilité si et seulement si*

*i) m' = m*

*ii) P' satisfait la A-stabilité transformée.*

Démonstration

Pas 1 - lemme

Soient $\Phi(\mu,\xi)$ comme donné en (5-28) et P'(w,z) comme en (5-31)

Supposons que $\Phi(\mu,\xi)$ soit factorisé de la façon suivante :

$$\Phi(\mu,\xi) = \varphi(\mu) \;\; \Psi(\xi) \;\; \bar{\Phi}(\mu,\xi) \qquad (5\text{-}32)$$

où $\varphi(\mu)$ est le P.G.C.D. des coefficients des puissances successives de $\xi$ dans $\Phi(\mu,\xi)$ et $\Psi(\xi)$ le P.G.C.D. des coefficients des puissances de $\mu$ .

La multiplicité de 1 comme racine de $\Psi$ et m-m'.

En particulier m = m' si et seulement si $\Psi(1) \neq 0$.

Démonstration du lemme

Faisons subir à (5-32) la transformation (5-31)

Nous obtenons :

$$P'(w,z) = \varphi'(w)\ \Psi'(z)\ \overline{P}'(w,z) \qquad (5\text{-}33)$$

où $\varphi'(w) = \varphi(iw)$

$$\Psi'(z) = (z-i)^a\ \Psi\left(\frac{z+i}{z-i}\right) \qquad (5\text{-}34)$$

où a est le degré de $\Psi$

$$\overline{P}'(w,z) = (z-i)^{m-a}\ \overline{\Phi}\left(iw,\frac{z+i}{z-i}\right) \qquad (5\text{-}35)$$

Notons j la multiplicité de 1 comme racine de $\Psi(\xi)$; $\Psi(\xi)$ peut dès lors se factoriser :

$$\Psi(\xi) = (\xi-1)^j\ \overline{\Psi}(\xi)$$

où $\overline{\Psi}(\xi)$ est de degré a-j.

La formule (5-34) nous donne l'effet de la transformation (5-31) sur $\Psi(\xi)$.

Nous pouvons dès lors écrire :

$$\Psi'(z) = (z-i)^a\ \left(\frac{z+i}{z-i} - 1\right)^j\ \Psi\left(\frac{z+i}{z-i}\right)$$

ou encore

$$\Psi'(z) = (2\ i)^j\ (z-i)^{a-j}\ \overline{\Psi}\left(\frac{z+i}{z-i}\right)$$

Par conséquent $\Psi'$ est de degré a-j.

La relation (5-32) assure par ailleurs que $\overline{\Phi}(\mu,\xi)$ est de degré m-a en $\xi$.

Le facteur $\Psi(\xi)$ étant un P.G.C.D., $\overline{\Phi}(\mu,\xi)$ ne peut plus se factoriser avec un terme ne dépendant que de $\xi$ . Par conséquent, $\overline{P}'(w,z)$ défini par (5-35) sera de degré m-a en z.

La forme (5-33) nous permet alors de conclure que m', le degré en z de P'(w,z) vérifie :

$$m' = a-j+m-a$$

Nous avons donc que j, la multiplicité de 1, comme racine de $\Psi(\xi)$ est m-m'.

De là, nous déduisons immédiatement que si m=m', cela entraîne et réciproquement que $\Psi(1) \neq 0$.

■

Pas 2 - Démonstration du théorème 5-6

Supposons que $\Phi(\mu,\xi)$ soit A-stable et par l'absurde, posons $m \neq m$. Le lemme nous assure alors que $\Psi(1) = 0$ et donc $\Phi(\mu,1) = 0$ pour tout $\mu \in C$. Ceci contredit l'hypothèse de la A-stabilité de $\Phi(\mu,\xi)$. Dès lors, notre hypothèse de l'absurde est fausse.
La condition : $m = m'$ est donc une condition nécessaire.
Par ailleurs, si $m = m'$, la transformation (5-31) a un inverse qui peut s'écrire :

$$\Phi(\mu,\xi) = (\xi-1)^{m'} \quad P'(-i\,\mu,\ i\,\frac{\xi+1}{\xi-1}) \qquad (5\text{-}36)$$

Par conséquent, si $-i\,\mu \in G$, $-\mu \in R$ et donc $\mu \in L$
D'autre part, si $i\,\frac{\xi+1}{\xi-1} \in H$, $\frac{\xi+1}{\xi-1} \in L$ et donc $\xi$ appartient à l'intérieur du disque unité.

Il est donc bien clair que si $m=m'$, la condition ii) du théorème 5-6 est équivalente à la condition de A-stabilité.

■■

Théorème 5-7 [ 37 ]

*Soit $P'(w,z)$ un polynôme complexe.*
*$P'(w,z)$ satisfait la A-stabilité transformée si et seulement si :*
*i) toutes les racines de $P'(.,i)$ sont dans $\overline{H}$*
*ii) $P'(.,z)$ n'est pas la fonction constante nulle, pour tout $z \in \mathbb{R}$.*
*iii) Pour tout $w \in \mathbb{R}$, si $P'(w,.)$ n'est pas la fonction constante nulle, toutes les racines de $P'(w,.)$ sont dans $\overline{H}$*

La démonstration de ces assertions se base sur un théorème de caractérisation analytique de la A-stabilité que Rubin énonce dans son rapport [ 37 ] et qu'il s'agit de transformer pour pouvoir l'utiliser.
Ces transformations sont de peu d'intérêt dans l'optique que nous avons choisie et nous accepterons le théorème 5-7 sans démonstration.

## 5.4.4. Application du formalisme de Bézout pour la vérification de la A-stabilité transformée

Le théorème 5-6 ramène le problème de la A-stabilité à celui de la A-stabilité transformée. Ce concept est utile dans la mesure où nous voulons utiliser le théorème d'Hermite.

Considérons le polynôme transformé $P'(w,z)$ qui est complexe, de degré $m'$ en $z$. Il peut, par conséquent s'écrire sous la forme

$$P'(w,z) = [1 \quad i] \quad \tau(w) \quad [1 \quad z \quad z^2 \quad \ldots \quad z^{m'}]^{\tau} \qquad (5\text{-}37)$$

où $\tau(w)$ est une matrice $2x(m'+1)$ dont les éléments sont des polynômes réels en $w$ ou la fonction constante nulle.

Si P'(w,z) est de degré n en w, les éléments de $\tau(w)$ sont de degré au plus n.

Nous définissons alors la matrice symétrique T, de dimension m'xm' dont les éléments ont la forme suivante ;

$$t_{ij} = \sum_{s=\max(0,m'+1-i-j)}^{m'-\max(i,j)} \tau\begin{pmatrix} 1 & , & 2 \\ 2\,m'+1-i-j-s, & & s \end{pmatrix} \qquad (5\text{-}38)\ (+)$$

pour i,j = 1,2, ...,m'.

La matrice T ainsi formée est, à deux permutations de lignes et de colonnes près, la matrice de Bézout $B(P'_R, P'_I)$ précédée du signe moins.

Puisque deux permutations ne changent pas la signature d'une matrice, nous pouvons utiliser la matrice T dans la formule (5-27) et lui appliquer le théorème d'Hermite pour déterminer, lorsque w réel est fixé, où se situent les racines de P'(w,.)

Nous aurons donc besoin des mineurs principaux emboîtés

$$\nabla'_i(w) = T\begin{pmatrix} 1, & 2, & \ldots, & i \\ 1, & 2, & \ldots, & i \end{pmatrix} \text{ pour } i = 0,1,2,\ldots,m' \qquad (5.39)$$

avec $\nabla'_0(w) = 1$ pour tout w.

Ces mineurs sont des polynômes réels en w de degré au plus égal à 2ni .

Remarque

Rubin utilise le formalisme de Bézout de manière détournée. Ceci lui est utile lorsqu'il recherche des formules explicites de factorisation du polynôme P'(w,z) sous la forme :

$$P'(w,z) = \hat{D}(w,z)\,\hat{P}'(w,z)$$

où D(w,z) est le plus grand diviseur réel au sens où il le définit [37] Householder [23] a indiqué que l'on pouvait obtenir une telle factorisation par la matrice de Bézout mais il ne donnait pas de telles formules explicites.

Théorème 5-8 [37]

*Soit P'(w,z) un polynôme complexe de degré m' en z.*
*Soient τ, T, $\nabla'_i$, i = 0,1,...,m' définis respectivement en (5-37) (5-38), (5-39).*
*Soient w ∈ ℝ, w fixé et supposons que $\nabla'_{m'}(w) \neq 0$*
*Alors P'(w,.) n'a pas de racines réelles.*
*En outre, P'(w,.) a toutes ses racines dans H si et seulement si $\nabla'_i(w) > 0$ pour i = 1,2,...,m'.*

---

(+) la notation $\tau\begin{pmatrix} 1 & & 2 \\ 2m'+1-i-j-s & & s \end{pmatrix}$ a la même signification que celle des $\nabla_k$ dans le théorème de Jacobi.

Démonstration

Si $\nabla'_{m'}(w) \neq 0$, la remarque 2 de la section 5.4.2. permet de conclure que les parties réelle et imaginaire $P'_R$ et $P'_I$ du polynôme $P'(w,.)$ n'ont pas de diviseur commun.

Or toute racine réelle de $P'(w,.)$ annule simultanément $P'_R(w,.)$ et $P'_I(w,.)$

Par conséquent $P'_R(w,.)$ et $P'_I(w,.)$ n'ont pas de racines réelles, et il en est de même pour $P'(w,.)$.

Par la section 5.4.2, nous savons que le nombre de racines dans un demi-plan est déterminé par le nombre de changements de signe de la suite

$$\nabla'_i(w) \quad i = 0,1,\ldots, m'$$

S'il existe des indices i tels que $\nabla'_i(w) = 0$, la remarque 1 de la section 5.4.2. assure que l'on aura au moins un changement de signe.

Par conséquent, puisque nous avons posé $\nabla'_0(w) = 1$, la suite $(1,\nabla'_1(w), \nabla'_2(w) \ldots,\nabla'_{m'}(w))$ ne présentera pas de changements de signe si et seulement si

$$\nabla'_i(w) > 0 \quad i = 0,1,\ldots,m' \qquad (5\text{-}40)$$

Si (5-40) est vérifiée, cela entraîne et réciproquement, par le théorème de Jacobi que

$\sigma(T) = -m'$ puisque T peut être considérée comme la matrice de Bézout affectée d'un signe -

Par le théorème d'Hermite, nous concluons alors que toutes les racines de $P'(w,.)$ sont strictement sous l'axe réel c'est-à-dire dans H .

■

Les théorèmes 5-7 et 5-8 vont à présent nous permettre d'énoncer un critère de A-stabilité transformée.

Théorème 5-9 [ 37 ]

*Soit $P'(w,z)$ un polynôme complexe de degré $m'$ en z.*

*Soient $\tau$, $T$, $\nabla'_i$, $i = 0,1,\ldots,m'$ définis respectivement en (5-37), (5-38), (5-39).*

*Supposons que $\nabla'_{m'}$, ne soit pas la fonction constante nulle.*

*Alors $P'(.,z)$ n'est pas la fonction constante nulle pour tout $z \in \mathbb{R}$.*

*En outre, $P'(w,z)$ satisfait le critère de A-stabilité transformée si et seulement si :*

*i) toutes les racines de $P'(.,i)$ sont dans $\overline{H}$*

*ii) $\nabla'_i$ n'est pas la fonction constante nulle pour $i = 1, 2, \ldots, m'-1$*

*iii) $\nabla'_i(w) \geq 0$ pour $i = 1, 2, \ldots, m'$*

*et pour tout $w \in \mathbb{R}$.*

Démonstration

L'idée centrale de cette démonstration est d'appliquer le théorème 5-8 pour toutes les valeurs réelles de w.

Pas 1 - P'(.,w) n'est pas constamment nul pour tout $z \in \mathbb{R}$

Si $\nabla'_{m'}$ n'est pas la fonction constante nulle, c'est un polynôme en la variable w.
Par conséquent, il existe un w réel tel que

$$\nabla'_{m'}(w) \neq 0$$

Le théorème 5-8 assure alors que $P'(w,z) \neq 0$ pour tout $z \in \mathbb{R}$.
Nous avons ainsi prouvé la première conclusion.

Pas 2 - Lien entre les théorèmes 5-7 et 5-9

Les conditions i) sont identiques dans l'un et l'autre.
L'assertion ii) du théorème 5-7 est immédiatement vérifiée grâce au pas 1.
Par conséquent, la démonstration du théorème 5-9 revient à établir l'équivalence entre la condition iii) du théorème 5-7 et les conditions ii) et iii) du théorème 5-9.
Soit w un réel.
Supposons que P'(w,.) soit la fonction constante nulle.
Alors (5-37) montre que $\tau(w) = 0$ et par conséquent $T(w) = 0$ et donc $\nabla'_{m'}(w) = 0$.
Ceci montre que si $w \in Y$, où $Y = \{w \in \mathbb{R}; \nabla'_{m'}(w) \neq 0\}$
alors P'(w,.) n'est pas la fonction constante nulle (5-41).
Or, par hypothèse $\nabla'_{m'}$ n'est pas constamment nulle.
Par conséquent Y est l'axe réel à l'exception peut être d'un nombre fini de points qui seraient des racines éventuelles de $\nabla'_{m'}$.

Pas 3 - Prenons pour hypothèse : iii) du théorème 5-7

Soit w fixé dans Y.
L'affirmation (5-41) et iii) du théorème 5-7 assurent alors que P'(w,.) a toutes ses racines dans $\overline{H}$
Par ailleurs, puisque $w \in Y$, P'(w,.) n'a pas de racines réelles par le théorème 5-8.
Par conséquent, toutes les racines de P'(w,.) sont dans H.
De nouveau, par le théorème 5-8, pour le w choisi, nous avons que $\nabla'_i(w) > 0$ pour $i = 1, 2, \ldots, m'$ et donc l'assertion ii) du théorème 5-9 est vérifiée.
En outre, les $\nabla'_i$ étant des polynômes sont des fonctions continues et nous pouvons aussi remarquer que la fermeture de Y est $\mathbb{R}$.
Dès lors, nous pouvons conclure que

$$\nabla'_i(w) \geqslant 0 \qquad \text{pour tout } w \text{ dans } \mathbb{R} \text{ et } i = 1, 2, \ldots, m'-1$$

ce qui est la condition iii) du théorème 5-9.

Pas 4 - Prenons pour hypothèse les conditions ii) et iii) du théorème 5-9

Désignons par $\hat{y} = \{w \in \mathbb{R} : \nabla'_i(w) \neq 0 \text{ pour } i = 1, 2, \ldots, m'\}$

La condition ii) du théorème 5-9 assure que $\hat{y}$ est l'axe réel à l'exception d'un nombre fini de points.

$\mathbb{R}$ sera donc la fermeture de $\hat{y}$.

Si nous prenons w dans $\hat{y}$, par iii) du théorème 5-9, nous avons que $\nabla'_i(w) > 0$ pour $i = 1, 2, \ldots, m'$.

Dès lors, en vertu du théorème 5-8, P'(w,.) a toutes ses racines dans $\mathbb{H}$ quel que soit w dans $\hat{y}$. (5-42)

Considérons à présent w dans $\mathbb{R}$, la fermeture de $\hat{y}$.

Si w est un pôle de P'(w,z), z est infini et sa partie imaginaire n'ayant dès lors pas de signe déterminé, nous pouvons affirmer que P'(w,.) a ses racines dans $\overline{\mathbb{H}}$. (5-43)

Supposons à présent que w ne soit pas un pôle. Les racines de P'(w,.) sont alors des fonctions continues des coefficients et donc de w.

Dès lors, s'il existait un w dans $\mathbb{R}$ tel que Im $z > 0$ où z est une racine de P'(w,.), nous aurions dans $\hat{y}$, un w tel que Im $z \geq 0$; ce qui contredit l'affirmation (5-42).

Par conséquent, nous concluons en considérant l'affirmation (5-43) que P'(w,.) a toutes ses racines dans $\overline{\mathbb{H}}$ quel que soit le w réel. L'assertion iii) du théorème 5-7 est ainsi vérifiée.

■

Remarques

1. L'argument de continuité des racines z du polynôme P'(w,z) par rapport à ses coefficients est valable pour autant que P'(w,z) ne puisse se factoriser sous la forme

   $$P'(w,z) = \Psi'(z)\ \overline{P}'(w,z)$$

   Cette condition est assurée par le fait que $\nabla'_{m'}$ ne soit pas une fonction constamment nulle car dans ce cas P'(w,z) admet une factorisation triviale au sens où Rubin le définit[37, proposition 3.6].

   Cette hypothèse simplifie notre raisonnement mais il est à noter que Rubin établit aussi un critère de A-stabilité sans cette hypothèse [37 - théorème 3-16].

2. C'est ce théorème 5-9 qui peut être comparé à celui de Ansell [1 - théorème II]. La condition que m'=m (voir théorème 5-6) est immédiatement vérifiée car Ansell considère au départ un polynôme qui correspond à celui défini en (5-29) et la seule transformation qu'il opère est celle répondant à (5-30).

   Par conséquent, ce que Ansell vérifie est, en fait, ce que Rubin appelle la A-stabilité transformée.

   Par ailleurs, la forme du polynôme utilisé par Ansell assure que les parties réelle et imaginaire n'ont pas de facteur commun. La remarque 2 de la section

5.4.2. nous assure dès lors que $\nabla'_m$, ne soit pas constamment nul.

Tout ceci étant noté, le critère de A-stabilité diffère pourtant légèrement d'un auteur à l'autre.

Ansell renforce i) en disant que toutes les racines de P'(.,i) sont dans **H**.

Par ailleurs, il ne mentionne pas la condition ii)

A ce stade, il resterait donc à établir quel est le lien entre les résultats des 2 auteurs.

5.4.5. Propriété des $\nabla'_i$. Nouveau critère de A-stabilité transformée

Théorème 5-10 [ 37 ]

*Soit P'(w,z) le polynôme complexe transformé de degré m' en z associé au polynôme réel* $\Phi(\mu,\xi)$

*Soient* $\tau$, $T$ *et* $\nabla'_i$ *pour* $i = 0,1,2, \ldots,m'$ *donnés en (5-37), (5-38), (5-39).*

*Alors* $\nabla'_i$ *est pair pour* $i = 0, 1, 2, \ldots, m'$.

Démonstration

Cette preuve repose sur un lemme qui donne une condition nécessaire et suffisante pour qu'un polynôme complexe P'(w,z) soit le polynôme associé à un polynôme réel par le biais de la formule (5-31)

Pas 1 - lemme

P'(w,z), polynôme complexe, est le polynôme transformé associé au polynôme réel $\Phi(\mu,\xi)$ si et seulement si

$$P'_{RM} = P'_R$$

$$P'_{IM} = -P'_I$$

où $P'_M(w,z) = P'(-w,-z)$ pour tout $w,z \in C$

et les lettres R et I indiquent respectivement les parties réelle et imaginaire des polynômes considérés.

Démonstration du lemme

Le polynôme transformé associé à $\Phi(\mu,\xi)$ résultait des deux transformations successives (5-29) et (5-30). $\Phi(\mu,\xi)$ est réel.

Dès lors P(μ,y) obtenu par (5-29) est également réel.

Par conséquent :

$$[P(iw,-iz)]^* = P(-iw,iz) \text{ pour tout } w \text{ et } z \text{ dans } \mathbb{R}.$$

où $*$ indique le polynôme conjugué (5-44)

Nous allons développer chacun des membres de cette égalité.

La relation (5-30) nous permet d'écrire

$$[P(iw,-iz)]^* = [P'(w,z)]^*$$

$$= [P'_R(w,z) + i\, P'_I(w,z)]^*$$

Puisque $P'_R$ et $P'_I$ sont des polynômes réels, cela revient encore à

$$[P(iw,-iz)]^* = P'_R(w,z) - i\, P'_I(w,z) \quad \text{pour tout } w \text{ et } z \text{ dans } \mathbb{R}.$$

Par ailleurs (5-30) donne au second membre de (5-44) la forme suivante :

$$P(-iw,iz) = P'(-w,-z)$$

$$= P'_M(w,z)$$

$$= P'_{RM}(w,z) + i\, P'_{IM}(w,z) \quad \text{pour tout } w \text{ et } z \text{ dans } \mathbb{R}.$$

La relation (5-44) peut donc s'écrire

$$P'_R(w,z) - i\, P'_I(w,z) = P'_{RM}(w,z) + i\, P'_{IM}(w,z) \qquad (5\text{-}45)$$

pour tout $w$ et $z$ dans $\mathbb{R}$.

Nous concluons dès lors que (5-45) est vérifiée si et seulement si

$$P'_R(w,z) = P'_{RM}(w,z)$$

$$\text{et } -P'_I(w,z) = P'_{IM}(w,z)$$

■

Pas 2 - Mise en évidence d'une propriété de la structure de T.

Notons

$$P'_R(w,z) = \sum_{j=0}^{m'} \Theta_j(w)\, z^j$$

P'(w,z) étant par hypothèse un polynôme transformé, nous avons, grâce au lemme, que

$$P'_{RM} = P'_R$$

c'est-à-dire

$$\sum_{j=0}^{m'} \Theta_j(-w)\, (-1)^j\, z^j = \sum_{j=0}^{m'} \Theta_j(w)\, z^j$$

Par conséquent, si j est pair, on a :

$$\Theta_j(-w) = \Theta_j(w)$$

et si j est impair

$$\Theta_j(-w) = -\Theta_j(w)$$

c'est-à-dire que les coefficients des puissances paires (respectivement impaires) de z dans $P'_R$ sont des polynômes pairs (respectivement impairs) en w.

Par contre, puisque nous avons aussi que

$$P'_{IM} = -P'_I$$

le même raisonnement nous mène à la conclusion que les coefficients des puissances paires (respectivement impaires) de z dans $P'_I$ sont des polynômes impairs (respectivement pairs) en w.

Les coefficients de $P'_R$ et $P'_I$ étant les éléments respectivement de la première et de la second ligne de $\tau(w)$ dans (5-37), les éléments de cette dernière matrice forment un damier de polynômes pairs et impairs de la façon suivante :

$$\begin{pmatrix} \text{Pair} & \text{Impair} & \text{Pair} & \cdots\cdots \\ \text{Impair} & \text{Pair} & \text{Impair} & \end{pmatrix}$$

Dès lors, les éléments de T (5-38) forment aussi un damier de ce genre.

En effet :

1. $t_{11}$ est pair car par (5-38), nous avons

   $$t_{11} = \tau \begin{pmatrix} 1 & 2 \\ m' & m'+1 \end{pmatrix}$$

   qui est un polynôme pair par la structure de $\tau(w)$ que nous venons de mettre en évidence.

2. Supposons que i+j soit pair

   Si s est pair : 2m'+1-i-j-s = 2m'+1-(i+j+s)

   est impair et $t_{ij}$ sera une somme de déterminants de matrices du type

   $$\begin{bmatrix} \text{Impair} & \text{Pair} \\ \text{Pair} & \text{Impair} \end{bmatrix}$$

   Si s est impair : 2m'+1-i-j-s est pair et $t_{ij}$ sommera les déterminants de matrices du genre

   $$\begin{bmatrix} \text{Pair} & \text{Impair} \\ \text{Impair} & \text{Pair} \end{bmatrix}$$

   Puisque le produit de 2 polynômes pairs ou de 2 polynômes impairs est un polynôme pair et que la somme des 2 polynômes pairs est un polynôme pair, nous concluons que $t_{ij}$ est pair si i+j est pair.

3. Supposons que i+j soit impair

   Un raisonnement analogue à celui mené ci-dessus conduit à la conclusion que $t_{ij}$ est impair si i+j l'est aussi.

Pas 3 - $\nabla'_i$ est pair pour $i = 0,1,2,\ldots,m'$

$\nabla'_i$ est un déterminant d'ordre i.

Le développement de Laplace (+) permet donc de l'exprimer comme une somme algébrique (avec les signes appropriés) de i! termes de la forme suivante :

$$\prod_{s=1}^{i} t_{s,r_s}$$

où $r_1, r_2, r_3 \ldots r_i$ est une permutation de 1, 2, .., i

Le pas 2 nous permet de dire que $t_{s,r_s}$ est impair si et seulement si $s + r_s$ est impair.

Par conséquent, il est clair que $\prod_{s=1}^{i} t_{s,r_s}$ est pair si et seulement si

$$\sum_{s=1}^{i} s + r_s \text{ est pair.}$$

Mais $\sum_{s=1}^{i} s + r_s = \sum_{s=1}^{i} s + \sum_{s=1}^{i} r_s$

et puisque l'addition est commutative dans $\mathbb{R}$, nous pouvons aussi écrire que

$$\sum_{s=1}^{i} s + r_s = \sum_{s=1}^{i} s + \sum_{s=1}^{i} s = 2 \sum_{s=1}^{i} s$$

Donc $\prod_{s=1}^{i} t_{s,r_s}$ est pair, pour toute permutation $r_1, r_2, \ldots, r_i$.

Puisque $\nabla'_i$ est une somme de tels termes, nous concluons que $\nabla'_i$ est pair pour $i = 0,1,2, \ldots,m'$.

■■

(+) A.C. Aitken, *Determinants and Matrices*", Edinburgh, London, Oliver and Boyd (1967).

Transformation des $\nabla_i'$

Le théorème 5-10 suggère d'utiliser le changement de variable suivant

$$\Omega = w^2$$

Dès lors, on définit les fonctions $\nabla_i''(\Omega)$ dans la formule (5-46)

$$\nabla_i''(\Omega) = \nabla_i'(\Omega^{1/2}) \qquad (5\text{-}46)$$

pour $i = 0,1, \ldots,m'$

Notons que $\nabla_i''(\Omega)$ est, ou la fonction constante nulle, ou un polynôme réel de degré plus petit ou égal à n i.
Il est clair que si $\nabla_i'$ est constamment nul, il en est de même pour $\nabla_i''$.

Appelons $k_i$ la multiplicité de 0 comme racine de $\nabla_i''$ pour $i = 0,1,\ldots,m'$. Si $\nabla_i''$ est la fonction constante nulle, posons $k_i = 0$.

Nous définissons alors de nouvelles fonctions $\nabla_i(\Omega)$ de la façon suivante :

$$\nabla_i(\Omega) = \nabla_i''(\Omega) / \Omega^{k_i} \qquad i = 0, 1,\ldots, m' \qquad (5\text{-}47)$$

Cette dernière formule nous permet de dire que, si $\nabla_i''(\Omega)$ est la fonction constante nulle, il en est de même de $\nabla_i(\Omega)$; sinon, $\nabla_i(\Omega)$ est un polynôme réel de degré $n_i - k_i$.

Notons en outre que la forme (5-47) assure que

$$\nabla_i(0) \neq 0 \qquad i = 0,1,\ldots,m' \qquad (5\text{-}48)$$

Grâce à ces nouveaux polynômes $\nabla_i(\Omega)$, nous allons pouvoir formuler un nouveau critère de A-stabilité.
L'avantage des $\nabla_i(\Omega)$ est qu'ils sont de degré nettement moins élevé que les $\nabla_i'$.

Théorème 5-11 [37]

*Soit $\Phi(\mu,\xi)$ un polynôme réel de degré m en $\xi$*

*Soit $P'(w,z)$ le polynôme complexe qui lui est associé, de degré m' en z.*

*Soit $\nabla_i(\Omega)$ $i = 0,1,\ldots,m'$ définis en (5-39).*

*Supposons que $\nabla_{m'}(\Omega)$ n'est pas la fonction constante nulle.*

*Alors, $\Phi(\mu,\xi)$ satisfait la propriété de A-stabilité si et seulement si*

*i) tous les pôles de $\Phi(\mu,\xi)$ sont dans $\overline{R}$*

*ii) $m' = m$*

*iii) $\nabla_i(0) > 0$ pour tout $i = 1,2,\ldots,m'$*

*iV) $\nabla_i$ n'a pas de racines réelles positives de multiplicité impaire pour $i = 1,2,\ldots,m'$.*

Démonstration

Par le théorème 5-6, il nous suffit de démontrer que les conditions i) iii) et iV) sont équivalentes à la A-stabilité transformée du polynôme complexe P'(w,z). Nous utiliserons, pour cela, le théorème 5-9.

Pas 1 - Les assertions i) des théorèmes 5-9 et 5-11 sont équivalentes

Cette équivalence devient évidente lorsque l'on considère la transformation (5-31). Si z = i, $\xi$ vaut l'infini et donc examiner les racines de P'(.,i) revient à examiner les pôles de $\Phi(\mu,\xi)$.
Or (5-31) multiplie la première variable par i.
Nous avons donc : $i\mu = w$
et inversément : $-\mu = iw$ ou encore $\mu = -iw$ (5-49).
La transformation (5-49) envoie $\overline{R}$ dans $\overline{H}$
Par conséquent $\Phi(\mu,\xi)$ a tous ses pôles dans $\overline{R}$ si et seulement si P'(.,i) a toutes ses racines dans $\overline{H}$

Pas 2 - Nouvelle formulation de ce qu'il faut démontrer

Notons que si ii) du théorème 5-9 n'est pas vérifiée, la condition iii) du théorème 5-11 tombe en défaut puisque $\nabla_i$ et $\nabla'_i$ vérifient la relation

$$\nabla_i(\Omega) = \frac{\nabla'_i(\Omega^{1/2})}{\Omega^{k_i}}$$

où $\Omega = w^2$

et donc si $\nabla'_i$ est la fonction nulle, il en est de même pour $\nabla_i$.
Il est donc suffisant de voir que, si on a ii) du théorème 5-9, alors il y a équivalence entre les conditions iii) du théorème 5-9 et iii) et iV) du théorème 5-11.

Pas 3 - Ecrivons les conditions iii) du théorème 5-9 d'une nouvelle façon

Par (5-46) et le théorème 5-10, nous avons que

$$\nabla_i''(w^2) = \nabla_i'(w) = \nabla_i'(-w) \qquad \text{pour tout } w \in \mathbb{R}$$

La condition iii) du théorème 5-9 revient donc à :

$$\nabla_i''(\Omega) \geq 0 \quad \text{pour } i = 1,2,\ldots,m'$$
$$\text{et } \Omega \in \mathbb{R} \text{ et } \Omega \geq 0 \quad (\text{car } \Omega = w^2)$$

Par ailleurs $\Omega^{k_i} > 0$ pour tout $\Omega > 0$

La formule (5-47) assure que $\nabla_i(\Omega)$ est continue en $\Omega = 0$ et donc nous pouvons écrire la condition iii) du théorème 5-9 sous la forme équivalente suivante :

$$\nabla_i(\Omega) \geq 0 \quad \text{pour } i = 1,2,\ldots,m'$$
$$\text{et } \Omega \in \mathbb{R},\ \Omega \geq 0 \qquad (5\text{-}50).$$

Pas 4 - Equivalence de (5-50) et des assertions iii) et iV) du théorème 5-11

Nous utilisons le lemme suivant [ 39 ]

Soit g(u) un polynôme réel.

Alors : g(u) ≥ 0 pour tout u ≥ 0, réel

si et seulement si

a. lorsque g(u) est une fonction constante

$$g(u) \equiv g_0 \text{ où } g_0 \geq 0 \text{ et réel}$$

b. lorsque g(u) n'est pas une fonction constante

- g(u) n'a pas de racines réelles, positives de multiplicité impaire
- il existe un $u_1 \geq 0$ réel tel que $g(u_1) > 0$

Démonstration du lemme

Supposons d'abord g(u) ≥ 0 pour tout u ≥ 0, réel.

Si g(u) est la fonction constante, la condition a) est immédiate.

Si g(u) n'est pas constante, elle n'est pas nulle partout, et donc, il existe $u_1$ réel et $u_1 \geq 0$, tel que

$$g(u_1) > 0.$$

Par ailleurs, nous savons qu'une fonction d'une variable réelle change de signe lorsqu'elle a une racine de multiplicité impaire.

Par conséquent, puisque, par hypothèse, g est de signe constant sur tout l'axe réel positif, g(u) n'a pas de racines réelles, positives de multiplicité impaire.

La nécessité du lemme est ainsi prouvée.

La suffisance est aussi évidente en utilisant la même propriété du changement de signe d'une fonction à variable réelle.

■

Appliquons ce lemme pour obtenir la thèse du théorème 5-11.
Au pas 2, nous avons fait l'hypothèse que $\nabla_i'$ n'est pas constamment nulle pour $i = 1, \ldots, m'$.
Par (5-47), il en est de même pour $\nabla_i(\Omega)$ $\quad i = 1, \ldots, m'$

Nous savons aussi de $\nabla_i(0) \neq 0 \quad i = 1, \ldots, m' \quad$ par (5-48).
Dès lors, grâce au lemme, la condition iii) du théorème 5-9, formulée d'une nouvelle façon en (5-50) est réalisée si et seulement si :

- $\nabla_i(\Omega)$ n'a pas de racines réelles positives de multiplicité impaire pour $i = 1, 2, \ldots, m'$
- $\nabla_i(0) > 0$ pour $i = 1, 2, \ldots, m'$.

■■

5.4.6. Conclusions

La première condition du théorème 5-11 peut se formuler d'une façon un peu différente. En effet, nous savons que les pôles $\Phi(\mu,\xi)$ sont les racines du polynôme, coefficient du terme en $\xi$ du plus haut degré.

Par conséquent, i) revient à dire que les racines de :

$\sum_{j=0}^{n} \alpha_{m_j} \mu^j$ doivent être dans $\overline{R}$ et que $\alpha_{m,n}$ soit non nul.

Supposons en effet que $\alpha_{m,n}$ soit nul. Dès lors, $\Phi(\mu,\xi)$ aurait un pôle infini qui ne serait donc pas forcément situé dans $\overline{R}$.
Cette nouvelle formulation nous montre que la condition i) du théorème 5-11 peut être vérifée par l'algorithme de Routh-Hurwitz.

La seconde condition est triviale à contrôler.
Il en est de même de iii) puisque $\nabla_i(0)$ est la valeur du terme indépendant de ce polynôme.

Le dernier point peut être vérifié par l'utilisation d'un algorithme basé sur les suites de Sturm [4].

o

o o

# CHAPITRE VI

## ETUDE DE L'ORDRE D'ERREUR MAXIMUM DES MÉTHODES À PAS ET DÉRIVÉES MULTIPLES, CONVERGENTES ET A-STABLES

Ce chapitre rassemble les différents résultats obtenus dans la recherche de l'ordre d'erreur maximum des méthodes $(k,\ell)$ convergentes et A-stables.

Après avoir caractérisé une méthode A-stable en fonction des parties paire et impaire de son polynôme canonique (section 6-1), nous démontrerons une propriété intéressante de ces polynômes, à la section 6-2.

La section 6-3 contient plusieurs caractérisations de l'ordre d'erreur d'une méthode convergente, que nous utiliserons pour démontrer les théorèmes généraux de la section 6-4.

Enfin, les conjectures de Daniel-Moore sont énonçées à la section 6-5 et démontrées dans certains cas particuliers, notamment dans le cas important où k vaut 1 et $\ell$ est quelconque.

## 6.1. CARACTERISATION D'UNE METHODE A-STABLE, EN FONCTION DES PARTIES PAIRE ET IMPAIRE DE SON POLYNOME CANONIQUE

Soit $H(z,q)$ le polynôme canonique d'une méthode A-stable.
Définissons sa partie paire :

$$H_e(z,q) = \frac{1}{2}\,[\,H(z,q) + H(-z,-q)\,]$$

et sa partie impaire

$$H_0(z,q) = \frac{1}{2}\,[\,H(z,q) - H(-z,-q)\,]$$

Supposons que :

$$H_e(z,q) \not\equiv 0 \text{ et } H_0(z,q) \not\equiv 0$$

Posons

$$Z(z,q) = \frac{H_e(z,q)}{H_0(z,q)}$$

On vérifie immédiatement que $Z(z,q)$ est une fonction réelle impaire.
La caractérisation que nous nous proposons d'établir, portera non pas sur le polynôme canonique lui-même (cfr. chapitre V), mais sur la fonction réelle $Z(z,q)$ ainsi définie.
La démonstration de cette caractérisation nécessite l'introduction de deux lemmes, que nous allons examiner préalablement.

Lemme 6.1.[1]

*Supposons que la fonction rationnelle à deux variables, $S(p,q)$, soit analytique dans le domaine* $\mathrm{Re}\ p \geq 0$, $\mathrm{Re}\ q > 0$.
*Supposons également que pour tout $p_0$ de partie réelle nulle, $S(p_0,q)$ soit une fonction rationnelle de $q$ telle que*

$$|S(p_0,q)| \leq 1 \qquad \text{pour tout } q \text{ tel que } \mathrm{Re}\ q = 0$$

*Alors, on peut affirmer que :*

$$|S(p,q)| \leq 1 \qquad \text{dans le domaine } \mathrm{Re}\ p > 0,\ \mathrm{Re}\ q > 0$$

Démonstration

Cette démonstration comprend deux parties, au courant desquelles nous appliquerons le théorème du maximum du module.

1ère partie

Fixons $p_0$ tel que sa partie réelle soit nulle, et considérons $S(p_0,q)$ fonction rationnelle en la variable $q$.
On sait que cette fonction rationnelle est analytique dans le domaine $\mathrm{Re}\ q > 0$.
D'autre part, par hypothèse, nous pouvons affirmer que :

$$|S(p_0,q)| \leq +1 \quad \text{si } \mathrm{Re}\ q = 0.$$

Donc, $S(p_0,q)$ est analytique dans tout le demi-plan de droite ($\text{Re } q \geqslant 0$).
Par le théorème du maximum du module, le module de $S(p_0,q)$ atteint son maximum en un point de l'axe $\text{Re } q = 0$.
Donc

$$|S(p_0,q)| \leqslant +1 \quad \text{pour tout } q \text{ tel que } \text{Re } q \geqslant 0.$$

2ème partie

Fixons, par ailleurs, $q_0$ de partie réelle strictement positive et considérons $S(p,q_0)$.
Par hypothèse, $S(p,q_0)$ est analytique dans le demi-plan droit ($\text{Re } p \geqslant 0$).
Le maximum du module de $S(p,q_0)$ est donc atteint en un point $p_0$ de partie réelle nulle.
Pour tout p de partie réelle strictement positive, on a donc

$$|S(p,q_0)| \leqslant |S(p_0,q_0)|$$

Or, dans la première partie, nous avons montré que

$$|S(p_0,q_0)| \leqslant +1$$

On en déduit que

$$|S(p,q_0)| \leqslant +1 \quad \text{pour tout } p \text{ tel que } \text{Re } p > 0.$$

La thèse résulte du fait que ce raisonnement est valable quel que soit $q_0$ de partie réelle strictement positive. ∎

Lemme 6.2. [5]

*Une fonction positive non nulle $f(\lambda)$ n'a pas de pôle dans $\text{Re } \lambda > 0$. Si elle a un pôle fini sur $\text{Re } \lambda = 0$, ce pôle est simple et à résidu réel et positif.*

Démonstration

L'inverse d'une fonction positive est une fonction positive car

$$\text{Re } \frac{1}{f(\lambda)} = \frac{\text{Re } f(\lambda)}{|f(\lambda)|^2}$$

Prouver la thèse revient donc à prouver qu'une fonction positive n'a pas de zéro dans $\text{Re } \lambda > 0$ et que si elle a un zéro fini sur $\text{Re } \lambda = 0$, alors ce zéro est simple et la dérivée en ce point est réelle et positive.
Soit $\lambda_0$, un zéro d'ordre k de $f(\lambda)$, tel que $\text{Re } \lambda_0 > 0$.
On a alors que

$$f(\lambda) \simeq \frac{1}{k!} \frac{d^k f(\lambda)}{d\lambda^k}\Big|_{\lambda=\lambda_0} (\lambda-\lambda_0)^k$$

Posons

$$\lambda - \lambda_0 = \rho e^{i\theta}$$

et

$$\frac{1}{k!} \frac{d^k f(\lambda)}{d\lambda^k}\Big|_{\lambda=\lambda_0} = A e^{i\varphi} \quad \text{où } A > 0$$

On a alors

$$\text{Re } f(\lambda) \simeq A\, \rho^k \cos(k\Theta + \varphi)$$

Pour que la fonction soit positive, il faut que pour tout $\Theta$ tel que

$$-\pi \leqslant \Theta \leqslant \pi \quad , \text{ on ait}$$

$$\text{Re } f(\lambda) \geqslant 0$$

Or, quel que soit $\varphi$ , il existe des valeurs de $\Theta$ pour lesquelles on a que

$$\text{Re } f(\lambda) < 0.$$

Il faut donc que k=0, autrement dit $f(\lambda)$ n'a pas de zéro dans Re $\lambda > 0$.

D'autre part, si Re $\lambda_0 = 0$, une condition nécessaire pour que la fonction $f(\lambda)$ soit positive, est que pour tout $\Theta$ tel que

$$-\frac{\pi}{2} \leqslant \Theta \leqslant \frac{\pi}{2} \quad ,$$

$$\text{Re } f(\lambda) \geqslant 0$$

Cela est possible en imposant k = 1 et $\varphi = 0$.

Donc, $\lambda_0$ est un zéro imaginaire simple de $f(\lambda)$ et la dérivée en ce point est un nombre réel positif.

■

Théorème 6.1. [1]

*Soient $N(z,q)$ et $D(z,q)$, deux polynômes en les variables z et q et à coefficients réels tels que,*

$$Z(z,q) = \frac{N(z,q)}{D(z,q)}$$

*soit une fraction rationnelle impaire.*

*Supposons que $N(z,q)$ et $D(z,q)$ n'ont pas de facteur commun non constant.*

*Alors, les deux assertions suivantes sont équivalentes :*

*(i) Si* $\text{Re } z > 0$, $\text{Re } q > 0$ *alors* $D(z,q) \neq 0$, $\text{Re } Z(z,q) > 0$

*(ii) Le polynôme*

$$H(z,q) = N(z,q) + D(z,q)$$

*est un polynôme à deux variables, Hurwitz au sens strict.*

Démonstration

1. Démontrons que la condition est nécessaire

Considérons la transformation suivante :

$$S(z,q) = \frac{Z(z,q)-1}{Z(z,q)+1} = \frac{N(z,q)-D(z,q)}{N(z,q)+D(z,q)} \qquad (6.1.)$$

Par hypothèse, on peut affirmer que

$$|S(z,q)| < 1 \quad \text{dans le domaine} \quad \begin{matrix} \text{Re } z > 0 \\ \text{Re } q > 0 \end{matrix}$$

Par continuité, on a que

$$|S(z,q)| \leqslant 1 \quad \text{dans le domaine} \quad \begin{matrix} \text{Re } z \geqslant 0 \text{ et Re } z > 0 \\ \text{Re } q > 0 \quad \text{Re } q \geqslant 0 \end{matrix} \qquad (6.2.)$$

Pour simplifier les écritures, désignons ce domaine par D.
Or, par hypothèse, N(z,q) et D(z,q) n'ont pas de facteur commun non constant. Donc, ces deux polynômes ne peuvent s'annuler simultanément qu'en un nombre fini de paires de points isolés. En ces points, S(z,q) est indéterminé. Outre ces cas litigieux, l'inégalité (6.2) nous permet d'affirmer que N(z,q) + D(z,q) est non nul dans le domaine D. Autrement dit, H(z,q) est un polynôme à deux variables Hurwitz au sens strict.
Traitons les cas particuliers où N(z,q) et D(z,q) s'annulent simultanément.
Supposons que $H(z,q) = R(z).V(q).U(z,q)$.
On a que

$$R(z) \neq 0 \quad \text{si Re } z \geqslant 0,$$

car sinon N(z,q) et D(z,q) s'annuleraient simultanément en une infinité de paires de points.

De même

$$V(q) \neq 0 \quad \text{si Re } q \geqslant 0.$$

Supposons que U(z,q) = 0, dans le domaine D.
On peut alors exprimer z comme fonction algébrique de q et par continuité de la fonction algébrique, il existe une infinité de (z,q) dans le domaine D tels que H(z,q) s'annule, ce qui contredit le fait que N(z,q) et D(z,q) ne s'annulent simultanément qu'en un nombre fini de paires de points isolées.

2. Montrons que la condition est suffisante

Considérons la transformation (6.1)

Si Re z = 0 = Re q, on a que

$$|S(z,q)| \leqslant 1.$$

D'autre part, H(z,q) ne s'annule pas si $\operatorname{Re} z \geqslant 0$, $\operatorname{Re} q > 0$.

Donc, S(z,q) est analytique dans ce domaine.

On peut donc appliquer le lemme 6.1. :

$|S(z,q)| \leqslant 1$ dans le domaine $\operatorname{Re} z > 0$, $\operatorname{Re} q > 0$.

Désignons ce domaine par R.

On a donc que

$$\operatorname{Re} Z(z,q) \geqslant 0 \quad \text{dans R.} \qquad (6.3.)$$

Il reste à prouver que

si $\operatorname{Re} Z(z,q) \geqslant 0$ dans R

alors $D(z,q) \neq 0$

et $\operatorname{Re} Z(z,q) > 0$ dans R

c'est-à-dire que Z(z,q) n'a pas de pôle dans R et qu'elle ne prend pas de valeurs purement imaginaires dans R.

a. Montrons que Z(z,q) n'a pas de pôle dans R

Il y a lieu de distinguer trois cas :

- 1er cas : $Z(a,q) = \infty$ pour tout q et où $\operatorname{Re} a > 0$.
  Considérons alors d tel que $\operatorname{Re} d > 0$ et la fonction à une variable Z(z,d).
  Si $\operatorname{Re} z > 0$, alors $\operatorname{Re} Z(z,d) \geqslant 0$.
  Cette fonction est donc positive en la variable z.
  Par le lemme 6.2., elle ne peut donc pas avoir de pôle dans $\operatorname{Re} z > 0$.

- 2ème cas : $Z(z,b) = \infty$ pour tout z et où $\operatorname{Re} b > 0$.
  La démonstration est semblable à celle du premier cas.

- 3ème cas : $Z(a,b) = \infty$ où $\operatorname{Re} a > 0$ et $\operatorname{Re} b > 0$.
  Fixons a et considérons Z(a,q) comme fonction de q. Cette fonction est positive en q. Elle n'a donc pas de pôle dans $\operatorname{Re} q > 0$.

b. Montrons que $Z(z,q)$ ne prend pas de valeurs purement imaginaires dans R

Supposons qu'il existe (a,b) dans R, tel que Z(a,b) soit purement imaginaire.

Considérons la fraction rationnelle

$$P(z,q) = \frac{1}{Z(z,q)-Z(a,b)}$$

Si Re $Z(z,q) \geqslant 0$ dans R, il en est de même pour Re Z(p,q). On démontre de la même façon que précédemment que P(z,q) n'a pas de pôle dans R, ce qui contredit l'existence de (a,b). ∎

## 6.2. PROPRIETE DES PARTIES PAIRE ET IMPAIRE DU POLYNOME CANONIQUE D'UNE METHODE A-STABLE

Supposons que $H_e(z,q)$ et $H_0(z,q)$, les parties paire, respectivement impaire, de H(z,q), n'ont pas de facteur commun non constant. Dès lors, les hypothèses du théorème 6.1. étant satisfaites, la méthode de polynôme canonique H(z,q) est A-stable si et seulement si $H_0(z,q)$ est non nul et la partie réelle de Z(z,q) est strictement positive, lorsque z et q sont à parties réelles strictement positives.

Ce critère est le point de départ de la propriété suivante.

Théorème 6.2. [19]

*Considérons une méthode à pas et dérivées multiples A-stable, de polynôme canonique H(z,q).*
*Supposons que* $H_e(z,q) \not\equiv 0$, $H_0(z,q) \not\equiv 0$ *et que ces deux polynômes n'ont pas de facteur commun non constant.*
*Alors* $H_e(z,q)$ *et* $H_0(z,q)$ *ont éventuellement des zéros imaginaires simples, indépendants de l'autre variable et sont Hurwitziens au sens strict lorsqu'on leur a extrait ces zéros éventuels.*

Démonstration

Etudions, tout d'abord, les zéros imaginaires éventuels de $H_e(z,q)$. L'étude de ceux de $H_0(z,q)$ sera tout-à-fait similaire.
Nous envisagerons dans un premier cas, un zéro imaginaire de $H_e(z,q)$, indépendant de la variable z et dans un second cas, de manière similaire, un zéro imaginaire de $H_e(z,q)$, indépendant de la variable q.

1er cas : supposons que $H_e(z^*,q^*) = 0$ où Re $z^* > 0$, et Re $q^* = 0$

Il est bien évident que $H_0(z^*,q^*)$ est non nul car sinon le polynôme $H(z,q)$ ne serait pas strictement Hurwitzien.

D'autre part, l'hypothèse de A-stabilité nous permet d'affirmer, par le théorème 6.1, que

si

$$\text{Re } z > 0$$
$$\text{Re } q > 0$$

alors

$$\text{Re } Z(z,q) > 0.$$

Par continuité, on a immédiatement que

si

$$\text{Re } z > 0$$

alors

$$\text{Re } Z(z,q^*) \geqslant 0,$$

ce qui signifie que $Z(z,q^*)$ est une fonction positive en la variable z.

Le lemme 6.2. démontre qu'une telle fonction ne peut avoir de pôles dans R.

Il est donc évident que $Z(z,q^*)$ est analytique dans R.

Par ailleurs,

$$\text{Re } Z(z^*,q^*) = 0 \quad \text{où Re } z^* > 0.$$

La partie réelle de la fonction $Z(z,q^*)$ atteint donc son minimum dans R.

Appliquons alors le théorème suivant [ 32 ]

> si la partie réelle d'une fonction analytique dans un contour R, atteint sa valeur minimale en un point intérieur de R, alors cette fonction est constante.

On en déduit que

$$Z(z,q^*) = \text{cste} = Z(z^*,q^*)$$

quel que soit z de partie réelle strictement positive

ou encore

$$H_e(z,q^*) = 0 \quad \text{pour tout z dans R.}$$

Nous venons de prouver que $H_e(z,q)$ peut avoir un zéro imaginaire, indépendant de la variable z : $q=q^*$.

Montrons ensuite que la multiplicité de cette racine vaut 1.

Considérons, pour cela, la fonction $Z(z,q)$ où z est fixé de partie réelle strictement positive.

On a que

$$\text{Re } Z(z,q) \geqslant 0 \quad \text{pour tout q dans } \overline{R}.$$

La fonction $Z(z,q)$ est donc positive en la variable q. Par conséquent, elle ne peut avoir qu'un zéro simple sur l'axe imaginaire.

La fonction $H_e(z,q)$ possède donc éventuellement un zéro simple imaginaire, indépendant de la variable z.

2ème cas : supposons que $H_e(\tilde{z},\tilde{q}) = 0$ où Re $\tilde{z} = 0$ et Re $\tilde{q} > 0$

On applique le même raisonnement que dans le premier cas, en intervertissant les rôles de z et de q.
La fonction $H_e(z,q)$ admet donc éventuellement un zéro simple imaginaire, indépendant de q.

En conclusion, si on extrait tous ces zéros imaginaires de $H_e(z,q)$, ce polynôme prend la forme :

$$H_e(z,q) = \prod_{j=1}^{s} (q-q_j) \prod_{i=1}^{t} (z-z_i)\, H_{e,r}(z,q)$$

où $H_{e,r}(z,q)$ ne s'annule ni dans Re $z > 0$
Re $q > 0$

ni dans Re $z = 0$
Re $q > 0$

ni dans Re $z > 0$
Re $q = 0$

et où Re $q_j = 0 \qquad \forall j \in \overline{s}$

Re $z_i = 0 \qquad \forall i \in \overline{t}$

Le polynôme $H_{e,r}(z,q)$ est donc bien un polynôme à deux variables, Hurwitz au sens strict.

On peut traiter $H_0(z,q)$ de la même façon, en raisonnant avec $\frac{1}{Z(z,q)}$ au lieu de $Z(z,q)$, qui vérifie également les hypothèses du théorème 6.1.
Le critère de A-stabilité est alors le suivant :

si
Re $z > 0$
Re $q > 0$

alors
$H_e(z,q) \neq 0$
Re $\frac{1}{Z(z,q)} > 0$

Le polynôme $H_0(z,q)$ peut alors s'écrire sous la forme :

$$H_0(z,q) = \prod_{j=1}^{s'} (q-q'_j) \prod_{i=1}^{t'} (z-z'_i)\, H_{0,r}(z,q)$$

où $H_{0,r}(z,q)$ est un polynôme à deux variables, Hurwitz au sens strict
et Re $q'_j = 0 \qquad \forall j \in \overline{s'}$
Re $z'_i = 0 \qquad \forall i \in \overline{t'}$

Abordons maintenantles caractérisations de l'ordre d'erreur de méthodes convergentes, que nous appliquerons ensuite aux méthodes A-stables.

## 6.3. CARACTETISATIONS DE L'ORDRE D'ERREUR D'UNE METHODE CONVERGENTE

Remarquons tout d'abord que puisque les méthodes envisagées sont convergentes et que nous n'intégrons que des équations différentielles du premier ordre, l'ordre d'erreur et l'ordre de consistance sont identiques. Il suit donc de la définition de l'ordre de consistance qu'une méthode $(k,\ell)$ a l'ordre d'erreur p ssi

$$L[e^{\lambda x},h] = C_{p+1}(h\lambda)^{p+1} e^{\lambda x} + O((h\lambda)^{p+1}) \qquad (6.4)$$

avec $C_{p+1} \neq 0$

et pour tout $\lambda \in C$.

En se rappelant la définition du polynôme de caractérisation et de l'opérateur $L[y(x),h]$, on prouve aisément que

$$\Phi(e^{\lambda h},\mu) = L[e^{\lambda x},h]\ e^{-\lambda x} \qquad (6.5.)$$

On déduit de (6.4) et (6.5) que la méthode a l'ordre d'erreur p ssi

$$\Phi(e^{\mu},\mu) = C_{p+1}\ \mu^{p+1} + O(\mu^{p+2}) \quad \text{si } \mu \to 0 \qquad (6.6)$$

En effectuant le changement de variable

$$\xi = e^{\mu},$$

l'égalité (6.6) devient

$$\Phi(\xi,\log\xi) = C_{p+1}(\xi-1)^{p+1} + O((\xi-1)^{p+2}) \quad \text{si } \xi \to 1 \qquad (6.7)$$

Nous avons donc obtenu une condition nécessaire et suffisante pour que la méthode soit d'ordre p, en fonction de son polynôme de caractérisation, comme fonction de $\xi$ (6.7) et comme fonction de $\mu$ (6.6).
Nous souhaitons maintenant trouver l'équivalent de ces conditions en fonction du polynôme canonique.
Le théorème 6.3 donne la caractérisation équivalente en fonction de la variable z, tandis que le théorème 6.4 la donne en fonction de q.

Théorème 6.3 [ 19 ]

*Une méthode à pas et dérivées multiples, convergente a l'ordre d'erreur p ssi*

$$\frac{1}{z^k} H(z, -\log\frac{z+1}{z-1}) \simeq C_{p+1}\ (\frac{2}{z})^{p+1} \quad si\ z \to \infty \qquad (6.8)$$

Démonstration

Par la transformation (5.2), on obtient que

$$z-1 = \frac{2}{\xi-1}$$

et

$$\xi - 1 = \frac{2}{z} + O(z^{-2}) \quad \text{si } z \to \infty \tag{6.9}$$

Ces égalités, ainsi que (5.5) nous assurent que

$$H(z, -\log \frac{z+1}{z-1}) = (\frac{2}{\xi - 1})^k \quad \Phi(\xi, \log \xi).$$

En vertu de (6.7) et (6.9), on obtient que

$$H(z, -\log \frac{z+1}{z-1}) = 2^{p+1} C_{p+1} z^{-p-1+k} + O(z^{-p-2+k}) \quad \text{si } z \to \infty$$

De là

$$\frac{1}{z^k} H(z, -\log \frac{z+1}{z-1}) \simeq C_{p+1} (\frac{2}{z})^{p+1} \quad \text{si } z \to \infty$$

■

Théorème 6.4. [19]

*Une méthode à pas et dérivées multiples, convergente à l'ordre d'erreur p ssi*

$$(\frac{q}{2})^k H(-\coth \frac{q}{2}, q) = (-1)^{p+1-k} C_{p+1} q^{p+1} + O(q^{p+2}) \tag{6.10}$$

Démonstration

Ce théorème se déduit directement du précédent, en considérant

$$q = -\log \frac{z+1}{z-1} = -\frac{2}{z} + O(z^{-2}) \quad \text{si } z \to \infty$$

Exprimons z en fonction de q :

$$z = -\frac{e^q+1}{e^q-1} = -\coth \frac{q}{2}$$

On déduit de ces égalités et du théorème précédent que

$$(-1)^k q^k \frac{1}{2^k} H(-\coth \frac{q}{2}, q) \simeq C_{p+1} (-1)^{p+1} q^{p+1} \quad \text{si } q \to 0.$$

ou encore

$$(\frac{q}{2})^k H(-\coth \frac{q}{2}, q) \simeq C_{p+1} (-1)^{p+1-k} q^{p+1} \quad \text{si } q \to 0.$$

■

Nous possédons maintenant les outils nécessaires pour aborder l'étude de l'ordre d'erreur maximal de méthodes $(k,\ell)$ A-stables.

## 6.4. THEOREMES GENERAUX CONCERNANT L'ORDRE D'ERREUR MAXIMAL D'UNE METHODE A PAS ET DERIVEES MULTIPLES A-STABLE ET CONVERGENTE

Une méthode convergente est telle que $\rho_0(\xi)$ a une racine simple en $\xi = 1$. Or, le changement de variable (5.2) transforme $\xi = 1$ en $z = \infty$. Dès lors, si la méthode $(k,\ell)$ est convergente, le polynôme $R_0(z)$ est exactement de degré k-1.
Il suit donc que

$$A_{k-1,0} \neq 0$$

et

$$A_{k,0} = 0.$$

Par ailleurs, une condition nécessaire de convergence est que l'ordre de consistance de la méthode soit au moins égal à 1.
En remplaçant p par 1 dans (6.8), on obtient la condition nécessaire et suffisante suivante :
la méthode est consistance ssi $A_{k,0} = 0$

$$A_{k,1} = 2\, A_{k-1,0}$$

Une condition nécessaire de convergence est donc que

$$A_{k,1} \neq 0 \qquad (6.11)$$

et

$$A_{k-1,0} \neq 0.$$

Théorème 6.5. [ 19 ], [ 25 ]

*Pour trouver le plus grand ordre d'erreur possible des méthodes $(k,\ell)$ convergentes et A-stables, il est suffisant de considérer les méthodes $(k',\ell')$ A-stables et convergentes, avec un polynôme canonique*
*pair si k est impair*
*impair si k est pair*
*et où* $1 \leq k' \leq k$ *et* $1 \leq \ell' \leq \ell$

Démonstration

Soit une méthode $(k,\ell)$ convergente, A-stable et d'ordre d'erreur p. Distinguons deux cas, suivant la parité de k.

1er cas : k est pair.

Il est évident que le polynôme $H_0(z,q)$ est non identiquement nul (car $A_{k-1,0}$ est non nul).
La méthode est d'ordre d'erreur p si et seulement si l'égalité (6.8) est vérifiée. Scindons cette égalité en deux parties, en développant en série au voisinage de $z = \infty$.

$$z^{-k} H_e\left(z, -\log \frac{z+1}{z-1}\right) = \sum_{s=t}^{\infty} \gamma_{2s+2} z^{-2s} \qquad (6.12)$$

$$z^{-k} H_0\left(z, -\log \frac{z+1}{z-1}\right) = \sum_{s=u}^{\infty} \gamma_{2s+1} z^{-2s-1} \qquad (6.13)$$

Si p est pair, on a que

$$u = \frac{p}{2}$$

et

$$t = \frac{p+2}{2}$$

Si p est impair, on a que

$$u = \frac{p+1}{2}$$

et

$$t = \frac{p+1}{2}$$

Considérons la méthode de polynôme canonique

$$\tilde{H}(z,q) = H_0(z,q)$$

et d'ordre d'erreur $\tilde{p}$ .
Remarquons que la condition nécessaire de convergence (6.11) est encore satisfaite. Par contre, elle ne le serait plus, si on considérait :

$$\tilde{H}(z,q) = H_e(z,q).$$

Voyons ce que devient l'ordre d'erreur $\tilde{p}$ ce cette méthode.
Si p est pair, la formule (6.13) s'écrit :

$$z^{-k} \tilde{H}(z,q) = \gamma_{p+1} \frac{1}{z^{p+1}} + 0\left(\frac{1}{z^{p+3}}\right).$$

Dans ce cas, $\tilde{p} = p$.
Par contre, si p est impair, (6.13) s'écrit

$$z^{-k} \tilde{H}(z,q) = \gamma_{p+2} \frac{1}{z^{p+2}} + 0\left(\frac{1}{z^{p+4}}\right).$$

L'ordre d'erreur $\tilde{p}$ est dans ce cas, $p+1+2\nu$ , où $\nu$ est un entier positif ou nul.

2ème cas : k est impair.

Le polynôme $H_e(z,q)$ est non identiquement nul.
On peut également scinder (6.8) en deux parties :

$$z^{-k} H_e(z, -\log \frac{z+1}{z-1}) = \sum_{s=t}^{\infty} \gamma_{2s+1} z^{-2s-1} \qquad (6.14)$$

$$z^{-k} H_0(z, -\log \frac{z+1}{z-1}) = \sum_{s=u}^{\infty} \gamma_{2s} z^{-2s} \qquad (6.15)$$

Si p est pair, on a

$$t = \frac{p}{2}$$

et

$$u = \frac{p+2}{2}.$$

Si p est impair, on a

$$t = \frac{p+1}{2}$$

et

$$u = \frac{p+1}{2}.$$

Considérons la méthode de polynôme canonique

$$\tilde{H}(z,q) = H_e(z,q)$$

et d'ordre d'erreur $\tilde{p}$.
Cette méthode satisfait encore à la condition nécessaire de convergence (6.11).
Examinons son ordre d'erreur.
Si p est pair, (6.14) s'écrit :

$$z^{-k} \tilde{H}(z,q) = \gamma_{p+1} \frac{1}{z^{p+1}} + 0\left(\frac{1}{z^{p+3}}\right)$$

On en déduit que $\tilde{p} = p$.
Par contre, si p est impair, (6.14) devient :

$$z^{-k} \tilde{H}(z,q) = \gamma_{p+2} \frac{1}{z^{p+2}} + 0\left(\frac{1}{z^{p+4}}\right).$$

Dans ce cas, $\tilde{p}$ vaut $p+1+2\nu$ , où $\nu$ est un nombre entier positif ou nul.

■

Rassemblons les résultats obtenus concernant l'ordre d'erreur $\tilde{p}$.

Si k est pair, $\tilde{p} =$ p si p est pair
$p+1+2\nu$ si p est impair.

Si k est impair, $\tilde{p} =$ p si p est pair
$p+1+2\nu$ si p est impair.

Par conséquent, si p est impair, on peut trouver une méthode $(k,\ell)$ avec un ordre d'erreur plus élevé, en négligeant soit $H_e(z,q)$ (si k est pair), soit $H_0(z,q)$ (si k est impair) et cet ordre d'erreur plus élevé est pair. Par contre, si p est pair, en négligeant $H_e(z,q)$ ou $H_0(z,q)$, on ne peut augmenter l'ordre d'erreur.

On déduit aisément le corollaire suivant :

Corollaire 6.1. [19], [25]

*Le plus grand ordre d'erreur possible d'une méthode $(k,\ell)$ A-stable et convergente est pair.*

## 6.5. LES CONJECTURES DE DANIEL-MOORE [13] : ORDRE D'ERREUR MAXIMUM ET METHODES OPTIMALES

En 1970, Daniel-Moore a abordé les méthodes à pas et dérivées multiples et a émis une hypothèse assez intéressante à leur sujet : l'ordre d'erreur d'une méthode $(k,\ell)$ A-stable et convergente, ne peut pas dépasser $2\ell$.

Il a également supposé que parmi toutes les méthodes A-stables et convergentes, d'ordre $2\ell$, celles qui ont la plus petite constante d'erreur sont les méthodes de la forme :

$$y_{i+1} = y_i + h \sum_{j=0}^{\ell-1} \alpha_j \left(f_i^{(j)} + (-1)^j f_{i+1}^{(j)}\right).$$

Ces conjectures ont pu être démontrées dans plusieurs cas. Toutefois, il n'existe pas encore de démonstration dans le cas tout à fait général.

La première conjecture nous incite à porter plus d'attention aux méthodes à un pas et à dérivées multiples.
En effet, si on augmente le nombre de pas, l'ordre d'erreur maximum restera inchangé, si le nombre de dérivées, $\ell$, reste inchangé.

### 6.5.1. Etude du cas : k=1, $\ell$ qcq

On a alors

$$H(z,q) = \sum_{j=0}^{\ell} A_{0j}\, q^j + z \sum_{j=0}^{\ell} A_{1j}\, q^j = 0.$$

De là

$$-z(q) = \frac{\sum_{j=0}^{\ell} A_{0j}\, q^j}{\sum_{j=0}^{\ell} A_{1j}\, q^j}$$

6.5.1.1. Ordre d'erreur maximum

Avant d'aborder la question de l'ordre d'erreur maximum de ces méthodes, démontrons un résultat qui sera utile dans la suite.

Lemme 6.3. [5]

*Soit une fraction rationnelle admettant la représentation en fraction continue*

$$f(z) = b_1 (z+i\,\omega_1)^{\alpha_1} + \cfrac{1}{b_2 (z+i\,\omega_2)^{\alpha_2} + \cfrac{1}{b_3 (z+i\,\omega_3)^{\alpha_3} + \ldots}}$$

où $\alpha_i \in \{-1,0,+1\}$

$\omega_i \in \mathbb{R}$

*si* $\alpha_i = 0$, $\mathrm{Re}\ b_i \geq 0$

*si* $\alpha_i = \pm 1$, $b_i$ *est réel et positif.*

*alors* $f(z)$ *est une fonction positive.*
*Inversément, si* $f(z)$ *est une fonction positive, alors elle admet le développement en fraction continue ci-dessus.*

Démonstration

Considérons f(z), une fonction positive et i $\omega_0$, un des points où Re f(z) atteint son minimum, noté $R_0$.
Soit

$$f(i\,\omega_0) = R_0 + i\,X_0 = Z_0.$$

La fonction $f(z) - Z_0$ est encore positive à condition que $X_0$ soit fini, autrement dit que i $\omega_0$ ne soit pas un pôle de f(z).
Il est bien évident que $f(z) - Z_0$ s'annule en $z = i\,\omega_0$.
Donc, la fonction $\dfrac{1}{f(z) - Z_0}$ a un pôle en $z = i\,\omega_0$.
On extait ce pôle :

$$\frac{1}{f(z) - Z_0} = \frac{h_0}{z - i\,\omega_0} + \frac{1}{Z_1(z)} \qquad (6.16)$$

où $h_0$ est le résidu au pôle i $\omega_0$ ($h_0 > 0$).

La fraction $Z_1(z)$ est encore positive et est d'un degré inférieur à celui de f(z).
L'égalité (6.16) s'écrit encore

$$f(z) - Z_0 = \frac{1}{\dfrac{h_0}{z - i\,\omega_0} + \dfrac{1}{Z_1(z)}}$$

Or,

$$Z_0 = R_0 + i\,X_0.$$

Donc,

$$f(z) = R_0 + i\,X_0 + \cfrac{1}{\cfrac{h_0}{z - i\,\omega_0} + \cfrac{1}{Z_1(z)}}$$

On procède de la même façon pour la fonction $Z_1(z)$ et ainsi de suite.
Dans le cas où $X_0$ est infini, on extrait tout d'abord ce pôle infini de $f(z)$ et on applique ensuite le procédé ci-dessus au reste.

Théorème 6.6. [7]

*L'ordre d'erreur maximum d'une méthode $(1,\ell)$ A-stable et convergente est $2\ell$.*
*De plus la méthode optimale est unique.*

Démonstration

Par le théorème 6.4. et l'égalité suivante :

$$H(-\coth\tfrac{q}{2},\, q) = \sum_{j=0}^{\ell} A_{0j}\, q^j - \coth\tfrac{q}{2} \sum_{j=0}^{\ell} A_{1j}\, q^j,$$

on a que

$$(\tfrac{q}{2})^k \Big[\sum_{j=0}^{\ell} A_{0j}\, q^j - \coth\tfrac{q}{2} \cdot \sum_{j=0}^{\ell} A_{1j}\, q^j\Big] = (-1)^{p+1-k}\, C_{p+1}\, q^{p+1} + O(q^{p+2}).$$

En divisant les deux membres de l'égalité par $\sum_{j=0}^{\ell} A_{1j}\, q^j$,
on obtient :

$$-z(q) = \coth\tfrac{q}{2} + (\tfrac{2}{q})^k\, (-1)^{p+1-k}\, C_{p+1}\, \frac{q^{p+1}}{\sum_{j=0}^{\ell} A_{1j}\, q^j} + O(q^{p+2-k}).$$

Or

$$k = 1 \text{ et } A_{10} = 0$$

Donc

$$-z(q) = \coth\tfrac{q}{2} + (-1)^p\, \frac{2\,C_{p+1}}{A_{11}}\, q^{p-1} + O(q^p).$$

La méthode ne peut être A-stable que si $-z(q)$ est une fonction positive (théorème 5.2).
Considérons le développement de $\coth\frac{q}{2}$ en fraction continue autour de $q=0$ :

$$\coth\tfrac{q}{2} = \frac{2}{q} + \cfrac{1}{6/q + \cfrac{1}{10/q + \cfrac{1}{14/q + \cdots \cfrac{1}{2(2n-1)/q + \cdots}}}}$$

Les hypothèses du lemme 6.3. étant satisfaites , on peut en déduire que $\coth \frac{q}{2}$ est une fonction positive.
Dès lors, l'ordre d'erreur maximum sera l'ordre de consistance maximum que peut atteindre une méthode $(1,\ell)$ convergente, à savoir (cfr. chapitre 2)

$$\sum_{i=0}^{1} \ell_i + k-1 = 2\ell$$

Pour un $\ell$ fixé, on obtient donc la méthode optimale en tronquant l'expansion en fraction continue de $\coth \frac{q}{2}$, après $\ell$ termes.
La fonction caractéristique de cette méthode s'écrit donc :

$$-z(q) = \frac{2}{q} + \cfrac{1}{\frac{6}{q} + \cfrac{1}{\frac{10}{q} + \cfrac{1}{\ldots + \cfrac{1}{\frac{2(2\ell-1)}{q}}}}} \qquad (6.17)$$

■

Le polynôme caractéristique de la méthode optimale prend une forme bien particulière, que nous allons étudier maintenant.

6.5.1.2. Méthodes optimales

RAPPEL : Les approximants de Padé

Définition

Soit une fonction A(x)
On note l'approximant de Padé L,M de A(s) par

$$[L/M] = \frac{P_L(x)}{Q_M(x)}$$

où $P_L(x)$ est un polynôme de degré au plus L

et $Q_M(x)$ est un polynôme de degré au plus M.

On détermine les coefficients de $P_L(x)$ et $Q_M(x)$ grâce au développement en série de

$$A(x) = \sum_{j=0}^{\infty} a_i \, x^j$$

et à l'équation

$$A(x) - \frac{P_L(x)}{Q_M(x)} = O(x^{L+M+1}).$$

Théorème [3]

*Lorsqu'il existe, l'approximant de Padé [L/M] d'une série en puissances de x, A(x), est unique.*

La table de Padé

On peut dresser la table de Padé de A(x) de la façon suivante :

[0/0] [0/1] [0/2] [0/3] ...

[1/0] [1/1] [1/2] [1/3] ...

[2/0] [2/1] [2/2] [2/3] ...

⋮ ⋮ ⋮ ⋮

La première colonne est donc formée des sommes partielles de la série de Taylor de A(x).

Lien entre le développement en fraction continue de A(x) et ses approximants de Padé :

Théorème [3]

*Soit le développement en fraction continue de A(x) :*

$$A(x) = a_0 \cfrac{a_1 x}{1+a_2 x - \cfrac{a_3 x}{1+a_4 x - \cfrac{a_5 x}{1+a_6 x - \cfrac{a_7 x}{\ddots}}}}$$

*où* $a_0, a_1, a_2, \ldots \in \mathbb{R}$.

*Les troncatures successives de ce développement donnent la suite [0/0], [1/1], [2/2], ... de la table de Padé de A(x).*

Reprenons la caractérisation de l'ordre d'erreur d'une méthode convergente, en fonction de son polynôme de caractérisation (6.6).
Une méthode $(1,\ell)$ convergente est d'ordre d'erreur p ssi

$$\sum_{j=0}^{\ell} \alpha_{0j} \mu^j + e^{\mu} \sum_{j=0}^{\ell} \alpha_{1j} \mu^j = C_{p+1} \mu^{p+1} + O(\mu^{p+2})$$

si $\mu \to 0$.
Or, $\xi(\mu)$ est défini par (5.1.).
Donc,

$$\xi(\mu) = - \frac{\sum_{j=0}^{\ell} \alpha_{0j} \mu^j}{\sum_{j=0}^{\ell} \alpha_{1j} \mu^j} = - \frac{\eta_0(\mu)}{\eta_1(\mu)}$$

La méthode est donc d'ordre p ssi

$$- \xi(\mu)\ \eta_1(\mu) + e^{\mu}\ \eta_1(\mu) = C_{p+1}\ \mu^{p+1} + O(\mu^{p+2})$$

ou encore

$$\xi(\mu) = e^{\mu} - \frac{C_{p+1}}{\alpha_{10}}\ \mu^{p+1} + O(\mu^{p+2}).$$

Chaque fois qu'on donne une approximation rationnelle de $e^{\mu}$ , avec $p \geq 1$, on donne une méthode $(1,\ell)$ et inversément.
Or, donner une approximation rationnelle de $e^{\mu}$, c'est donner un approximant de Padé de $e^{\mu}$ :

$$[L/M] \underset{\text{noté}}{=} E_{L,M}\ .$$

On vient donc de démontrer le théorème suivant :

Théorème 6-7 [7], [20], [25]

*A chaque approximant de Padé de $e^{\mu}$ , correspond une et une seule méthode $(1,\ell)$ convergente.*

A l'approximant de Padé $E_{L,M}$ de $e^{\mu}$ , correspond une méthode d'ordre L+M, car

$$e^{\mu} - \frac{P_L(\mu)}{Q_M(\mu)} = O(\mu^{L+M+1}).$$

Le théorème 6.7 nous suggère de dresser la table des méthodes $(1,\ell)$ convergentes (voir fig. 6.1.).

L'observation de la table de Padé nous porte à croire que pour un $\ell$ fixé, la méthode d'ordre d'erreur maximum est celle basée sur $E_{\ell,\ell}$.
C'est ce que nous allons prouver à présent.

La méthode $(1,\ell)$ optimale est donnée par la relation (6.17) ou encore par

$$- z(q) = \frac{u_\ell(q)}{v_\ell(q)}$$

où $u_\ell(q)$ et $v_\ell(q)$ vérifient les relations de récurrence suivantes [19]

$$u_\ell(q) = (2\ell-1)\ u_{\ell-1}(q) + \frac{1}{4}\ q^2\ u_{\ell-2}(q) \qquad \text{si } \ell > 2$$
$$v_\ell(q) = (2\ell-1)\ v_{\ell-1}(q) + \frac{1}{4}\ q^2\ v_{\ell-2}(q)$$

initialisées par

$$u_1(q) = 1$$
$$v_1(q) = \frac{q}{2}$$

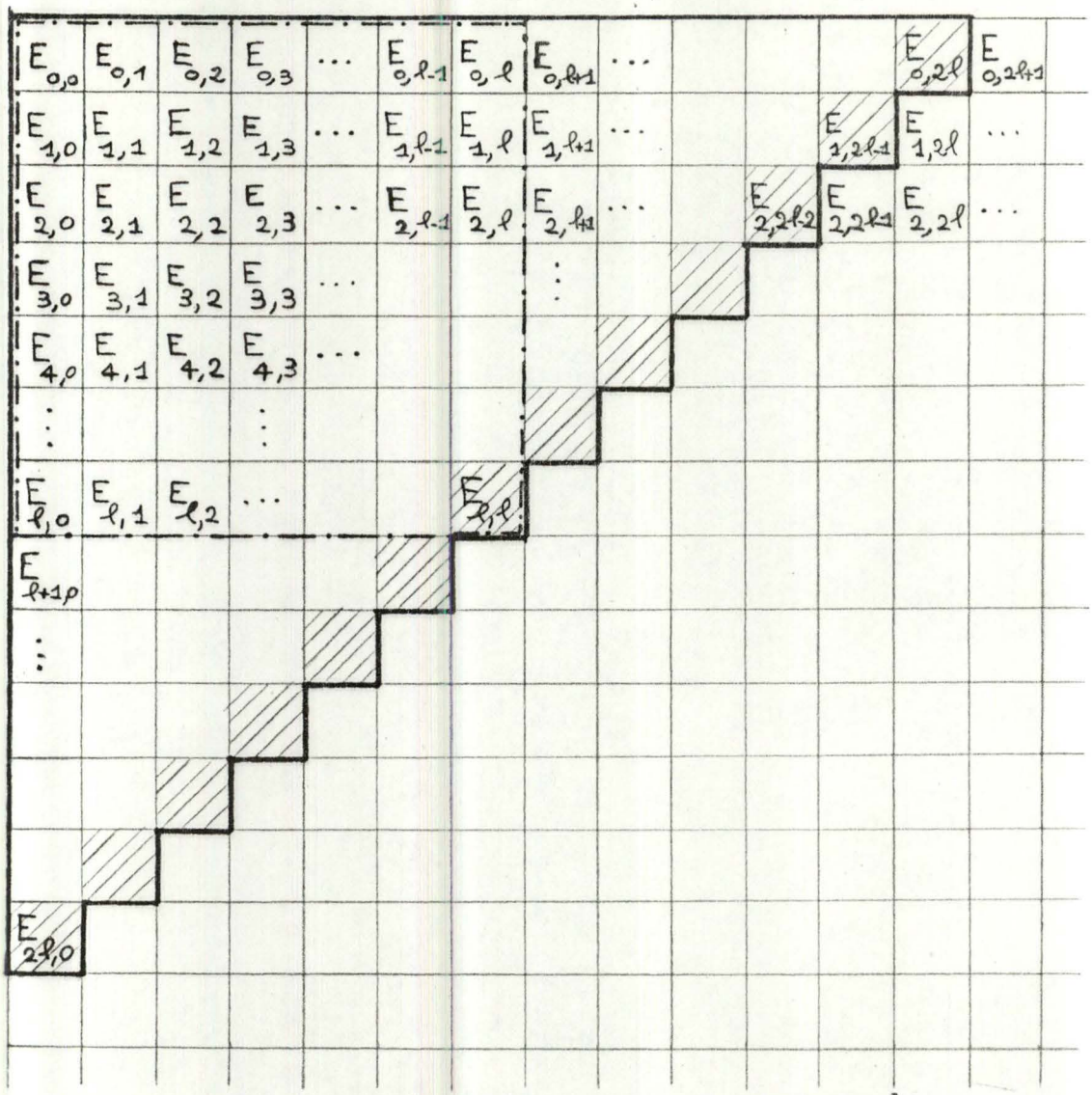



FIGURE 6.1. : TABLE DE PADE DE $e^{\mu}$

Légende :

——— : méthodes d'ordre d'erreur $\leqslant 2\ell$

—.—.— : méthodes $(1,\ell)$ où $\ell$ est fixé

/// : méthodes d'ordre $2\ell$

et $u_2(q) = 3\, u_1(q) + \frac{1}{4} q^2$

$v_2(q) = 3\, v_1(q)$.

On applique la transformation inverse de (5.2), à savoir

$$\xi = \frac{z+1}{z-1}$$

et $\mu = -q$

Donc,

$$\xi = \frac{u_\ell - v_\ell}{u_\ell + v_\ell} = \frac{n_\ell(\mu)}{m_\ell(\mu)}$$

Ces polynômes $n_\ell(\mu)$ et $m_\ell(\mu)$ obéissent aux relations de récurrence :

$$\begin{aligned} n_\ell(\mu) &= (2\ell-1)\, n_{\ell-1}(\mu) + \tfrac{1}{4}\mu^2\, n_{\ell-2}(\mu) \\ m_\ell(\mu) &= (2\ell-1) m_{\ell-1}(\mu) + \tfrac{1}{4}\mu^2\, m_{\ell-2}(\mu) \end{aligned} \qquad (6.18)$$

initialiséespar

$$\begin{aligned} n_1(\mu) &= 1 - \tfrac{\mu}{2} \\ m_1(\mu) &= 1 + \tfrac{\mu}{2} \end{aligned} \qquad (6.19)$$

et

$$\begin{aligned} n_2(\mu) &= 3\, n_1(\mu) + \tfrac{1}{4}\mu^2 \\ m_2(\mu) &= 3\, m_1(\mu) + \tfrac{1}{4}\mu^2 \end{aligned} \qquad (6.19)$$

On a immédiatement que

$$n_\ell(-\mu) = m_\ell(\mu)$$

et dès lors

$$\xi_{opt} = \frac{m_\ell(-\mu)}{m_\ell(\mu)} \qquad (6.20)$$

Si on considère l'ensemble des approximants de Padé de $e^{\mu}$, $E_{L,M}$ tels que $L+M = 2\ell$, on va montrer que $E_{\ell,\ell}$ a la forme (6.20).
On aura prouvé le théorème suivant :

Théorème 6.8 [ 20 ]

*La méthode $(1,\ell)$ convergente et A-stable, d'ordre d'erreur maximum est celle basée sur l'entrée diagonale de la table de Padé de $e^{\mu}$ et est telle que*

$$-n_0(\mu) = \sum_{j=0}^{\ell} \frac{(2\ell-j)!}{(2\ell)!} \binom{\ell}{j} \mu^j = n_1(-\mu)$$

Démonstration

Considérons le développement en fraction continue de $e^{\mu}$ :

$$e^{\mu} = 1 + \frac{2\mu}{2-\mu+} \quad \frac{\mu^2}{6+} \quad \frac{\mu^2}{10+} \cdots \frac{\mu^2}{4m-2+} \cdots$$

Les troncatures successives de cette expression donnent la suite

$$E_{0,0}, E_{1,1}, E_{2,2}, \cdots$$

de la table de Padé de $e^{\mu}$ (cfr. rappel).

Donc,

$$E_{\ell,\ell} = 1 + \frac{2\mu}{2-\mu+} \quad \frac{\mu^2}{6+} \quad \frac{\mu^2}{10+} \cdots \frac{\mu^2}{4\ell-2} \tag{6.21}$$

On vérifie aisément que ces troncatures, soit $\dfrac{u_\ell(\mu)}{v_\ell(\mu)}$ vérifient les relations de récurrence :

$$\begin{aligned} u_\ell(\mu) &= (4\ell-2)\, u_{\ell-1}(\mu) + \mu^2\, u_{\ell-2}(\mu) \\ v_\ell(\mu) &= (4\ell-2)\, v_{\ell-1}(\mu) + \mu^2\, v_{\ell-2}(\mu) \end{aligned} \tag{6.22}$$

initialisées par

$$\begin{aligned} u_0(\mu) &= 1 \\ v_0(\mu) &= 1 \end{aligned} \tag{6.23}$$

et

$$\begin{aligned} u_1(\mu) &= 2 + \mu \\ v_1(\mu) &= 2 - \mu \end{aligned} \tag{6.23}$$

qui s'identifient à (6.18) et (6.19).

La méthode optimale, pour un $\ell$ fixé, est donc bien la méthode basée sur $E_{\ell,\ell}$.

Il reste à montrer que $E_{\ell,\ell}$ s'identifie avec

$$q_\ell(\mu) = \frac{\sum_{j=0}^{\ell} \dfrac{(2\ell-j)!}{(2\ell)!} \binom{\ell}{j} \mu^j}{\sum_{j=0}^{\ell} \dfrac{(2\ell-j)!}{(2\ell)!} \binom{\ell}{j} (-\mu)^j}$$

$$= \frac{P_\ell(\mu)}{P_\ell(-\mu)}$$

où $P_\ell(\mu) = \sum_{j=0}^{\ell} \dfrac{(2\ell-j)!}{j!(\ell-j)!} \mu^j$

On peut montrer par récurrence que $P_\ell(\mu)$ vérifie la relation (6.22), initialisée par (6.23), [20], ce qui termine la démonstration du théorème 6.8.

Remarques

1. Ehle [ 73 ] a démontré que pour tout $n \geq 0$, les approximants de Padé $E_{n,n+1}$ et $E_{n,n+2}$ de $e^{\mu}$ correspondent à des méthodes A-stables.

2. Pour un $\ell$ fixé, Genin [ 74 ] a démontré que la constante d'erreur de la méthode $(1,\ell)$ convergente et A-stable optimale est

$$c_{2\ell+1} = (-1)^{\ell} \frac{\ell!\ \ell!}{(2\ell+1)!\ (2\ell)!}$$

Remarquons que si $\ell$ augmente, la constante d'erreur en valeur absolue diminue.

3. Le plus grand ordre d'erreur des méthodes $(1,\ell)$ convergentes et A-stables est pair. Ce résultat concorde bien avec la corollaire 6.1.
Par ailleurs, $K$ étant impair, on montre que les polynômes canoniques correspondant aux méthodes optimales sont pairs.
En effet,

où $H(z,q) = u_{\ell}(q) + z\, v_{\ell}(q)$

$u_{\ell}(q)$ et $v_{\ell}(q)$ vérifient (6.22) et (6.23).

On voit immédiatement que $u_{\ell}(q)$ est pair pour tout $\ell \geq 1$

$v_{\ell}(q)$ est impair pour tout $\ell \geq 1$.

Il suit que $u_{\ell}(q) + z\, v_{\ell}(q)$ est pair pour tout $\ell \geq 1$.

Abordons maintenant un second cas particulier : les méthodes à pas multiples, mais à une seule dérivée.

6.5.2. Etude du cas : h qcq, $\ell = 1$

En 1963, Dahlquist a déjà étudié longuement ces méthodes, ainsi que leur ordre d'erreur. Il a ainsi prouvé des résultats connus tels que : "L'ordre d'erreur maximum des méthodes à pas multiples, convergentes et A-stables est 2".

Néanmoins, Genin [74] a publié une démonstration analogue à celle du premier cas étudié.

Théorème 6.9. [12], [19]

*L'ordre d'erreur d'une méthode (k,1) convergente et A-stable, ne peut pas dépasser 2.*

Démonstration

On a que

$$H(z,q) = \sum_{i=0}^{k-1} A_{i0} z^i + q \sum_{i=0}^{k} A_{i1} z^i = 0 \qquad (6.24)$$

et donc

$$-q(z) = \frac{\sum_{i=0}^{k-1} A_{i0} z^i}{\sum_{i=0}^{k} A_{i1} z^i} \qquad (6.25)$$

Par le théorème 5.2, la méthode est A-stable si et seulement si -q(z) est une fonction positive.
D'autre part, par (6.24) et le théorème 6.3, on sait que la méthode est d'ordre d'erreur p ssi

$$\frac{1}{z^k}\left[\sum_{i=0}^{k-1} A_{i0} z^i - \log\frac{z+1}{z-1} \sum_{i=0}^{k} A_{i1} z^i\right] = C_{p+1}\left(\frac{2}{z}\right)^{p+1} + O(z^{-p-2})$$

ou encore, en divisant par $\sum_{i=0}^{k} A_{i1} z^i$,

$$-q(z) = \log\frac{z+1}{z-1} + \frac{C_{p+1}}{A_{k1}}\left(\frac{2}{z}\right)^{p+1} + O(z^{-p-2})$$

Sans perdre de généralité, posons

$$A_{k1} = 1.$$

Supposons que p vaut 3.

On a alors

$$-q(z) = \log\frac{z+1}{z-1} + C_{p+1}\left(\frac{2}{z}\right)^4 + O(z^{-5})$$

Or,

$$\log\frac{z+1}{z-1} = \frac{2}{z} + \frac{2}{3z^3} + O(z^{-5})$$

$$= \frac{1}{\frac{z}{2} - \frac{1}{6z}} + O(z^{-5}).$$

Comme $-\frac{1}{6} < 0$, le lemme 6.3. nous permet d'affirmer que -q(z) n'est pas une fonction positive.

En conclusion, l'ordre d'erreur 3 ne peut donc pas être atteint.
Si p vaut 2, alors on a l'égalité suivante :

$$-q(z) = \frac{2}{z} + \frac{2}{3z^3} + C_3\left(\frac{2}{z}\right)^3 + O(z^{-4})$$

Or, par (6.25), on a

$$-q(z) = \frac{\sum_{i=0}^{k-1} A_{i0}\, z^{i-k}}{1 + \sum_{i=0}^{k-1} A_{i1}\, z^{i-k}}$$

Donc, si on pose

$$d = \frac{2}{3} + 8\, C_3,$$

on obtient l'égalité :

$$\frac{2}{z} + d\,\frac{1}{z^3} + O(z^{-4}) = \frac{\sum_{i=0}^{k-1} A_{i0}\, z^{i-k}}{1 + \sum_{i=0}^{k-1} A_{i1}\, z^{i-k}}$$

ou encore

$$\frac{2}{z} + d\,\frac{1}{z^3} + O(z^{-4}) = \left(A_{k-1,0}\,\frac{1}{z} + A_{k-2,0}\,\frac{1}{z^2} + A_{k-3,0}\,\frac{1}{z^3} + \ldots\right)$$

$$\cdot\left[1 - \left(A_{k-1,1}\,\frac{1}{z} + A_{k-2,1}\,\frac{1}{z^2} + \ldots\right) + \left(A_{k-1,1}\,\frac{1}{z} + \ldots\right)^2 + \ldots\right].$$

Egalons les coefficients de $\frac{1}{z}$, $\frac{1}{z^2}$ et $\frac{1}{z^3}$

$$A_{k-1,0} = 2$$

$$A_{k-2,0} = 2\, A_{k-1,1}$$

$$d = A_{k-3,0} + 2(-A_{k-2,1} + A_{k-1,1}^2) - A_{k-2,0}\, A_{k-1,1}$$

$$= A_{k-3,0} - 2\, A_{k-2,1}.$$

On obtient dès lors, la fonction caractéristique suivante :

$$-q(z) = \frac{2z^{k-1} + 2\, A_{k-1,1}\, z^{k-2} + (d+2\, A_{k-2,1})\, z^{k-3} + \ldots}{z^k + A_{k-1,1}\, z^{k-1} + A_{k-2,1}\, z^{k-2} + \ldots}$$

ou encore

$$-q(z) = \cfrac{1}{\frac{z}{2} + \cfrac{1}{\frac{-4z}{d} + Q(z)}}$$

où Q(z) est une fraction rationnelle positive de degré au maximum k-2 et où $-\frac{4}{d}$ est strictement positif, pour que -q(z) soit positive.
Il faut donc que $d < 0$ pour que la méthode soit A-stable.
L'ordre d'erreur 2 peut donc être atteint.

■

Dahlquist a également recherché la méthode (k,1) A-stable et convergente, d'ordre 2, possédant la plus petite constante d'erreur en valeur absolue. Il s'agit de la règle trapézoïdale, donnée par la relation suivante :

$$y_n - y_{n-1} = \frac{h}{2}\left[f_n + f_{n-1}\right]$$

de constante d'erreur, $-\frac{1}{12}$.

En effet, puisque

$$d = \frac{2}{3} + 8\,C_3,$$

on a que

$$C_3 = \frac{d - \frac{2}{3}}{8}$$

D'autre part,

$$c_3 = -\frac{C_3}{\rho_1(1)}$$

et

$$\rho_1(1) = -A_{k,1} = -1$$

On a donc

$$c_3 = \frac{d - \frac{2}{3}}{8}.$$

Or, pour que la méthode soit A-stable, il faut que

$$d \leqslant 0$$

c'est-à-dire

$$c_3 \leqslant -\frac{1}{12}$$

Le théorème suivant est alors immédiat.

Théorème 6.10 [ 12 ], [ 19 ]

*Parmi toutes les méthodes (k,1) convergentes et A-stables, d'ordre 2, la règle trapézoïdale est celle qui a la plus petite constante d'erreur en valeur absolue.*

### 6.5.3. Etude du cas général : $k > 1$, $\ell > 1$

L'étude de l'ordre maximum des méthodes a plusieurs pas et plusieurs dérivées, a été abordé par Genin [19]. Toutefois, ce dernier n'a démontré la conjecture de Daniel-Moore que dans certains cas particuliers : k=2 et $\ell$ = 2,3 ou 4, k=3 et $\ell$=2.

Nous examinerons de plus près deux de ces cas particuliers dans le théorème suivant .

Théorème 6.11 [19]

*Le plus grand ordre d'erreur des méthodes (2,2) respectivement (2,3), (2,4), (3,2), convergentes et A-stables est 4, respectivement 6,8,4.*

Démonstration

Les preuves de ces différentes assertions présentant le même type d'arguments, nous nous limiterons aux cas k=2, $\ell$=2 et k=2, $\ell$=3.

1er cas : k=2, $\ell$=2

Le théorème 3.4. du chapitre III nous indique que l'ordre d'erreur de ces méthodes ne peut pas dépasser 6.
Par ailleurs, le polynôme canonique d'une méthode (2,2) d'ordre de consistance 6, s'écrit de manière générale [19] :

$$H(z,q) = [30z + (15\,z^2-1)\,q + 2z\,q^2]\,A_{12} + [48 + 18z\,q - (3\,z^2 - 5)\,q^2]\,A_{02}$$

Or, une condition nécessaire de A-stabilité est que tous les coefficients de H(z,q) aient le même signe (théorème 5.5). Il n'existe donc pas de méthode (2,2) A-stable d'ordre de consistance 6.
Si l'ordre de consistance ne dépasse pas 5, le polynôme s'écrit alors :

$$H(z,q) = A_{22}\,(2 + z\,q)^2 + \frac{A_{12}}{2}\,(30 + (15\,z^2-1)\,q + 2\,z\,q^2) + A_{02}\,(12 + 6\,z\,q + q^2).$$

Les méthodes correspondant à ce polynôme canonique ne peuvent être A-stables que si $A_{12} = 0$. Le polynôme se réduit alors à :

$$H(z,q) = A_{22}\,(2 + z\,q)^2 + A_{02}\,(12 + 6\,z\,q + q^2) \qquad (6.26).$$

On peut vérifier, en utilisant le théorème 5.4, que ce polynôme est Hurwitz au sens strict pour toutes valeurs positives de $A_{22}$ et $A_{02}$.

Nous venons donc de prouver que l'ordre de consistance maximum d'une méthode (2,2) A-stable est 5 et que la forme la plus générale de son polynôme canonique est donnée par (6.26).
Or, l'ordre d'erreur d'une telle méthode vaut 4.
En effet, si on applique la transformation inverse de (5.2), on remarque que 1 est une racine double du polynôme $\rho_0(\xi)$.
Remarquons, par ailleurs, que la méthode basée sur l'approximant de Padé $E_{2,2}$ de $e^q$, atteint cet ordre d'erreur, ce qui prouve la première assertion du théorème.

2ème cas : k=2, $\ell$=3

Supposons que l'ordre de consistance de la méthode soit 8. La forme générale de son polynôme canonique s'écrit alors [ 19 ] :

$$H(z,q) = H_e(z,q) + H_0(z,q)$$

où

$$H_e(z,q) = [\,384 + 174\, z\, q - (9\, z^2 - 39)\, q^2 + 23\, q^3]\, \frac{A_{13}}{2}$$

et

$$H_0(z,q) = [\,1050\, z + (525\, z^2 - 3)\, q + 102\, z\, q^2 + 8\, z^2\, q^3]\, \frac{A_{23}}{8}$$
$$+ [\,630\, z + (315\, z^2 + 75)\, q + 90\, z\, q^2 + 8\, q^3\,]\, \frac{A_{03}}{8} \qquad (6.27).$$

Si le polynôme H(z,q) est un polynôme à deux variables, Hurwitz au sens strict, en vertu du théorème 6.2, ses parties paire et impaire sont des polynômes Hurwitz.

Pour que la méthode soit A-stable, il faut donc que

$$A_{13} = 0$$

et

$$75\, A_{03} - 3\, A_{23} \geqslant 0$$

où

$A_{03}$ et $A_{23}$ sont des réels positifs.

On peut montrer par des calculs élémentaires mais relativement longs que sous ces contraintes, le minimum de la partie réelle de -z(q) est négatif sur l'axe Re q = 0.

La fonction -z(q) n'est donc pas positive et il n'existe pas de méthode (2,3) A-stable atteignant l'ordre de consistance 8.

Par contre, si l'ordre de consistance vaut 7, le polynôme canonique s'écrit

$$H(z,q) = [\,48 + 18\, z\, q + (-3\, z^2 + 5)\, q^2\,]\, \frac{A_{02}}{5}$$
$$+ [\,24 + 84\, z\, q + 36\, z^2\, q^2 + 5\, z\, q^3]\, \frac{A_{13}}{5}$$
$$+ [\,1050\, z + (525\, z^2 - 3)\, q + 102\, z\, q^2 + 8\, z^2\, q^3]\, \frac{A_{23}}{8}$$
$$+ [\,630\, z + (315\, z^2 + 75)\, q + 90\, z\, q^2 + 8\, q^3\,]\, \frac{A_{03}}{8} \qquad (6.28)$$

La partie impaire de (6.28) s'identifie avec (6.27). Par un raisonnement similaire au précédent, nous pouvons conclure que la condition nécessaire pour avoir la A-stabilité est la suivante :

$$A_{23} = A_{03} = 0.$$

Dès lors, le polynôme canonique peut s'écrire sous la forme

$$H(z,q) = x_1 (2 + z\,q)(12 + 6\,z\,q + q^2) + x_2 (120 + 60\,z\,q + 12\,q^2 + z\,q^3) \qquad (6.29)$$

où $x_1 = (12\,A_{13} - A_{02}) / 10$

et $x_2 = (A_{02} - 2\,A_{13}) / 10$

Le réel $x_1$ est positif pour que tous les coefficients du polynôme soient non négatifs, tandis que le réel $x_2$ est non négatif pour que le minimum de la partie réelle de $-z(q)$ soit positif ou nul sur l'axe Re q = 0.
Avec ces conditions et en appliquant le théorème 5.4., on montre que H(z,q) est un polynôme à deux variables Hurwitz au sens strict.
L'ordre de consistance 7 peut donc être atteint par des méthodes (2,3) A-stables).
La thèse résulte du fait que l'ordre d'erreur des méthodes décrites par un polynôme canonique de la forme (6.29) vaut 6 et que la méthode basée sur $E_{3,3}$ atteint cet ordre d'erreur.

■

L'étude de l'ordre d'erreur des méthodes à pas et dérivées multiples est encore l'objet d'une recherche intensive aujourd'hui. En effet, les conjectures de Daniel-Moore restent encore des hypothèses non démontrées de manière générale. Bien qu'elles semblent se vérifier dans plusieurs cas particuliers, aucune généralisation n'a pu être dégagée des démonstrations existantes.

## 6.6. CONCLUSION

Nous nous sommes bornés, dans ce chapitre, à étudier l'ordre d'erreur des méthodes $(k,\ell)$ A-stables. L'hypothèse de convergence des méthodes est nécessaire dans ce cas, afin de pouvoir considérer les égalités (6.8) et (6.10) comme caractérisations de l'ordre d'erreur.

Une étude simulaire a été réalisée par Genin [19], concernant l'ordre de consistance des méthodes $(k,\ell)$ A-stables.
L'hypothèse de convergence des méthodes n'a alors plus de raison d'être imposée. Remarquons que les méthodes d'ordre de consistance maximum, obtenues par Genin, ne sont pas toujours convergentes mais que leur ordre d'erreur prend la valeur $2\ell$.

Il est donc bien évident que les méthodes $(k,\ell)$ A-stables d'ordre d'erreur maximum ne sont pas nécessairement convergentes.

°

° °

# A P P E N D I C E  A

## EXPRESSION GENERALE DU POLYNOME CANONIQUE DE METHODES $(k,\ell)$ D'ORDRE DE CONSISTANCE MAXIMUM

Nous avons vu au chapitre II que l'ordre de consistance maximum d'une méthode $(k,\ell)$ vaut

$$(k+1)(\ell+1) - 2$$

et que la méthode optimale est unique pour un k et un $\ell$ fixés.
Posons

$$\nu = (k+1)(\ell+1) - 2.$$

Notre but est de rechercher l'expression générale du polynôme canonique de la méthode $(k,\ell)$ d'ordre de consistance $\nu$.

Considérons l'égalité (6.8) où $H(z,q)$ est le polynôme canonique de la méthode optimale. Cette égalité peut également s'écrire :

$$\frac{1}{z^k} \sum_{i=0}^{l} R_i(z) \left(-\log \frac{z+1}{z-1}\right)^i \sim C_{\nu+1} \left(\frac{2}{z}\right)^{\nu+1} \qquad (A.1.)$$

Développons $\left(\log \frac{z+1}{z-1}\right)^i$ au voisinage de $z = \infty$

$$\left(\log \frac{z+1}{z-1}\right)^i = c_i^{(i)} z^{-i} + c_{i+2}^{(i)} z^{-i-2} + c_{i+4}^{(i)} z^{-i-4} + \ldots$$

où les coefficients $c_j^{(i)}$ satisfont aux relations suivantes :

$$c_j^{(i)} = 0 \quad \text{si } j < i$$
$$c_j^{(i)} = 2^i \quad \text{si } j = i$$
$$c_j^{(i)} = 0 \quad \text{si } i+j \text{ est impair.}$$

Les coefficients $c_{i+2k}^{(i)}$ sont repris dans la table (A.1)

TABLE (A.1)

| i \ k | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 2 | 2/3 | 2/5 | 2/7 | 2/9 | 2/11 |
| 2 | 4 | 8/3 | 92/45 | 176/105 | 2252/1575 | 13016/10395 |
| 3 | 8 | 24/3 | 112/15 | 6544/945 | 9152/1575 | - |
| 4 | 16 | 64/3 | 352/15 | 22976/945 | - | - |
| 5 | 32 | 160/3 | 608/9 | 14624/189 | - | - |
| 6 | 64 | 384/3 | - | - | - | - |

Si on tronque le polynôme $(\log \frac{z+1}{z-1})^i$ après le terme $c_\nu^{(i)} z^{-\nu}$, on obtient un polynôme de degré $\nu$ en $z^{-1}$.

On désignera le produit de ce polynôme par $(-1)^i$, par $a_i(z)$.

Si on fait de même pour tout $i \in \overline{\ell}$, on déduit alors de l'égalité (A.1) que

$$\frac{1}{z^k} \sum_{i=0}^{\ell} R_i(z)\ a_i(z) = 0 \qquad (A.2)$$

où $R_i(z)$, $i = 0,1,\ldots,\ell$ sont les polynômes inconnus.

Introduisons les matrices triangulaires suivantes :

$$\pi_i = (-1)^i \begin{pmatrix} c_0^{(i)} & & & 0 \\ c_1^{(i)} & c_0^{(i)} & & \\ c_2^{(i)} & c_1^{(i)} & c_0^{(i)} & \\ \vdots & \vdots & & c_0^{(i)} \\ & & & \vdots \\ c_\nu^{(i)} & c_{\nu-1}^{(i)} & \cdots & c_{\nu-k}^{(i)} \end{pmatrix} \qquad i = 0,1,\ldots,\ell$$

où $\pi_i$ a $(\nu+1)$ lignes et $(k+1)$ colonnes.

Par ailleurs, les coefficients $A_{ij}$ du polynôme $R_i(z)$ forment le vecteur $\overline{a}_i$ défini par

$$\overline{a}_i = (A_{i0}\ A_{i1}\ A_{i2} \cdots A_{ik})^\tau$$

où $i = 0,1,2,\ldots,\ell$.

Avec cette notation, on montre aisément, en égalant les coefficients des puissances de $\frac{1}{z}$ à zéro, que l'égalité (A.2) est équivalente au système suivant :

$$\sum_{i=0}^{\ell} \pi_i\ \overline{a}_i = 0 \qquad (A.3)$$

Les polynômes $R_i(z)$, $i = 0,1,\ldots,\ell$, de la méthode optimale peuvent alors s'exprimer de la façon suivante :

$$R_i(z) = \left|\begin{array}{ccc:c:ccc} \pi_0 & \pi_1 \cdots & \pi_{i-1} & \pi_i & \pi_{i+1} & \cdots & \pi_\ell \\ \hline 0 & 0 \ldots & 0 & 1\ z\ z^2 \ldots z^{k-1}\ z^k & 0 & \ldots & 0 \end{array}\right| \qquad (A.4)$$

où $i = 0,1,\ldots,\ell$

En effet, l'égalité (A.3) peut encore s'écrire :

$$\sum_{i=0}^{\ell} R_i(z)\, a_i(z) = 0 \qquad (A.5)$$

Notons

$$(1 \; z \; z^2 \; \ldots \; z^{k-1} \; z^k) = P.$$

Montrons que si on remplace $R_i(z)$ par (A.4) dans le membre de gauche de (A.5), alors celui-ci s'annule.
Par ce remplacement, le membre de gauche devient

$$\begin{vmatrix} \pi_0 & \pi_1 & \pi_2 & & \pi_\ell \\ a_0(z)\,P & a_1(z)\,P & a_2(z)\,P & \ldots\ldots & a_\ell(z)\,P \end{vmatrix}$$

Ce déterminant est nul, car le système

$$\begin{pmatrix} \pi_0 & \pi_1 & \pi_2 & \ldots\ldots & \pi_\ell \\ a_0(z)\,P & a_1(z)\,P & a_2(z)\,P & \ldots\ldots & a_\ell(z)\,P \end{pmatrix} \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_\ell \end{pmatrix} = 0$$

admet une solution non triviale

$$u_i = \bar{a}_i \qquad \text{où } i = 0,1,\ldots,\ell.$$

■

Remarques :

1. Les déterminants de la forme (A.4) sont appelés multigradients car ils apparaissent comme une simple généralisation des bigradients et trigradients [ 6 ].
2. Suivant la parité de k et de $\ell$, le second membre de (A.1) sera soit pair, soit impair (si k est pair et $\ell$ impair, il sera impair, sinon, il sera pair).

   Puisque $\log \frac{z+1}{z-1}$ est une fonction impaire, il faut donc que $R_i(z)$ soit pair ou impair suivant la parité de i.

°

° °

# BIBLIOGRAPHIE


[1] ANSELL, H.G. : "*On certain two-variable generalizations of circuit theory with applications to networks of transmission lines and lumped reactances*", IEEE Trans. on C.T. 11, pp. 214-223, 1964.

[2] ATKINSON K. et A. SHARMA : "*A partial characterisation of poised Hermite-Birkloff interpolation problems*", SIAM J. Numer Anal. 6, pp. 230-235, 1969.

[3] BAKER George A.,Jr., "*Essantials of Padē Approximants*", Academic Press, New York, San Francisco, London, 1975.

[4] BARNETT, "*A new formulation of the theorems of Hurwitz, Routh and Sturm*", J. Inst. Math. Appl., 8, pp. 240-250, 1971.

[5] BELEVITCH, "*Classical Network Theory*", Holdenday - San Francisco, 1968.

[6] BELEVITCH-GENIN, "*Implicit interpolation, trigradients and continued fractions*", Philips Res. Reports 26, pp. 453-470, 1971.

[7] BIRKHOFF-VARGA, "*Discretisation errors for well-set Cauchy problems*", J.J. Math. and Phys., 44, pp. 1-23, 1965.

[8] KREYSZIG, E., "*Advanced Engineering Mathematics*", J. Wiley and Sons, New York.

[9] BROWN, "*Multi-derivative numerical methods for the solution of stiff ordinary differential equations*", Department of Computer Science, University of Illinois, Report UIUCDCS-R-74-672, 1974.

[10] CALAHAN, D.A., "*Numerical solution of linear systems with widely separated time constants*", Proc. IEEE, Nov., 2016-2017, 1967.

[11] DAHLQUIST, "*Stability and error bounds in the numerical integration of ordinary differential equations*", Kungl. Tekn Högsk, Stockholm, n° 130, 1959.

[12] DAHLQUIST, "*A special stability problem for linear multistep methods*", BIT 3, 27-43, 1963.

[13] DANIEL-MOORE, "*Computation and theory in ordinary differential equations*", San Francisco, Freeman and Co., 1970.

[ 14 ] DAVIS, P.J., "*Interpolation and Approximation*", New York, Blaisdell Publ. Co., 1965.

[ 15 ] EHLE, B.L., "*A-stable methods and Padē approximations to the exponential*" SIAM J. Math. Anal. November 1973.

[ 16 ] ENRIGHT,"*Second derivative multistep methods for Stiff ordinary differential equations*", SINUM, n, pp. 321-331, 1974

[ 17 ] FERGUSON, "*The question of uniqueness for G.D. Birkhoff interpolation problems*", J. Approximation Theory, 2, pp. 1-28,1969.

[ 18 ] GANTMACHER, F.R., "*Theory of Matrices*", Vols. 1 and 2, Chelsea Pub. Co, New York, 1959.

[ 19 ] GENIN Y., "*An Algebraic Approach to A-stable Linear Multistep - multiderivative Integration formulas*", BIT 14, pp. 382-406, 1974.

[ 20 ] GRIEPENTROG, "*Mehrschrittverfahren zur numerischen Integration von gewöhnlichen Differentialgleichungssystemen und asymptotische Exaktheit*", Wiss. Z. Humboldt, Univ. Berlin Math., Natur. Reihe, vol. 19, pp. 637-653, 1970.

[ 21 ] HENRICI, "*Discrete Variable Methods in Ordinary Differential Equations*", Wiley, New York, 1962.

[ 22 ] HILLE, "*Analytic Function Theory*", vol. II, Ginn and Co.,Boston, 1962.

[ 23 ] HOUSEHÖLDER , "*Bezoutiants, Elimination and Localization*", SIAM Review, vol. 12, n° 1, January 1970.

[ 24 ] JELTSCH, "*Integration of iterated integrals by multistep methods*", Numer. Math. 21, pp. 303-316, 1973.

[ 25 ] JELTSCH, "*Multistep Multiderivative Methods for the Numerical Solution of Initial Value Problems of Ordinary Differential Equations*", Sem. Notes 1975-1976, University of Kentucky.

[ 26 ] KARLIN-KARON, "*On Hermite-Birkhoff Interpolation*", J. Approximation Theory, 6, pp. 90-114, 1972.a .

[ 27 ] KARLIN-KARON, "*Poised and Non-poised Hermite Birkhoff Interpolation*", Indiana University Mathematics Journal, vol. 21, n° 12, pp. 1131-1170, 1972.b.

[ 28 ] LAMBERT,"*Computational Methods in Ordinary Differential Equations*", Wiley, London, 1973.

[ 29 ] LORENTZ - ZELLER,"*Birkhoff Interpolation*", SIAM J. Numer. Anal., 8, pp. 43-48, 1971.

[ 30 ] LORENTZ, "*Birkhoff interpolation and the problem of free matrices*", J. Approx. Theory, vol. 6, pp. 283-290, 1972.

[ 31 ] MARDEN, M., "*Geometry of Polynomials*", American Mathematical Society, Providence, Rhode Island, 1966

[ 32 ] OZAKI, H., KASAMI, T., "*Positive Real Functions of Several Variables and their Applications to Variable Networks*", IRE Transactions on Circuit Theory , 1960

[ 33 ] OSGOOD, W.F., "*Functions of real and complex Variables*", Chelsea Publ. Co., New York, NY.

[ 34 ] RALSTON, A. "*A first course in Numerical Analysis*", New York, London; Mc Graw Hill; Tokyo, Kogakusha, 1965.

[ 35 ] REIMER, "*Zur Theorie der linearen Differenzenformeln*", Math. Zeitschr. 95, pp. 373-402, 1967.

[ 36 ] REIMER, "*Finite difference forms containing derivatives of higher order,*" SIAM J. Numer. Anal., vol. 5, pp. 725-738, 1968.

[ 37 ] RUBIN, "*A-stability and composite multistep methods*", Ph. D Thesis, EE Dept, Syracuse University, New York, 1973.

[ 38 ] SHARMA A. et J. PRASAD, "*On Abel-Hermite Birkhoff interpolation*", SIAM J. Num. Anal. 5, pp. 864-881, 1968.

[ 39 ] SILJAK, "*New Algebraic Criteria for Positive Realness*", J. Franklin Inst. 291, pp. 109-120, 1971.

°

° °