Étude de procédés d'extrapolation en analyse numérique

Pierre-Jean Laurent

N· D'ORDRE
408

THÈSES

PRÉSENTÉES

A LA FACULTÉ DES SCIENCES
DE L'UNIVERSITÉ DE GRENOBLE

POUR OBTENIR

LE GRADE DE DOCTEUR ÈS SCIENCES MATHÉMATIQUES

PAR

Pierre-Jean LAURENT
Ingénieur des Arts et Métiers
Docteur du 3ème cycle

PREMIÈRE THÈSE

Étude de procédés d'extrapolation en analyse numérique

DEUXIÈME THÈSE

Propositions données par la Faculté

Le théorème du graphe fermé dans les espaces de Banach

Soutenues le 15 juin 1964, devant la Commission d'examen :

MM. J. FAVARD	Président
J. KUNTZMANN	Rapporteur
N. GASTINEL	
J. BARRA	

N· D'ORDRE
408

THÈSES

PRÉSENTÉES

A LA FACULTÉ DES SCIENCES
DE L'UNIVERSITÉ DE GRENOBLE

POUR OBTENIR

LE GRADE DE DOCTEUR ÈS SCIENCES MATHÉMATIQUES

PAR

Pierre-Jean LAURENT
Ingénieur des Arts et Métiers
Docteur du 3ème cycle

PREMIÈRE THÈSE

Étude de procédés d'extrapolation en analyse numérique

DEUXIÈME THÈSE

Propositions données par la Faculté

Le théorème du graphe fermé dans les espaces de Banach

Soutenues le 15 juin 1964, devant la Commission d'examen :

MM.	J. FAVARD	Président
	J. KUNTZMANN	Rapporteur
	N. GASTINEL	
	J. BARRA	

FACULTÉ DES SCIENCES – UNIVERSITÉ DE GRENOBLE

Doyens honoraires

M. FORTRAT P.
M. MORET L. Membre de l'Institut

Doyen

M. WEIL L.

Professeurs honoraires

M.M. FORTIER A. ; FAVARD J. ; BRELOT M. ; FORTRAT R.

Professeurs

M. NEEL L. Membre de l'Institut — Magnétisme et physique du solide

MM.	DORIER A.	Zoologie	
	HEILMANN R.	Chimie organique	
	KRAVTCHENKO J.	Mécanique rationnelle	
	CHABAUTY C.	Calcul différentiel et intégral	
	PARDE M.	Potamologie	
	BENOIT J.	Radioélectricité	
	CHENE M.	Chimie papetière	
	BESSON J.	Electrochimie	
	WEIL L.	Thermodynamique	
	FELICI N.	Electrostatique	
	KUNTZMANN J.	Mathématiques appliquées	
	BARBIER R.	Géologie appliquée	
	SANTON L.	Mécanique des fluides	
	OZENDA P.	Botanique	
	FALLOT M.	Physique industrielle	
	GALVANI O.	Mathématiques	
	MOUSSA A.	Chimie nucléaire	
	TRAYNARD P.	Chimie	
	SOUTIF M.	Physique	
	CRAYA A.	Hydrodynamique	
MM.	REULOS R.	Théorie des champs	
	AYANT Y.	Physique approfondie	
	GALLISSOT F.	Mathématiques appliquées	
Mle.	LUTZ E.	Mathématiques	
MM.	BLAMBERT M.	Mathématiques	
	BOUCHEZ R.	Physique nucléaire	
	LLIBOUTRY L.	Géophysique	
	MICHEL R.	Géologie et minéralogie	
	BONNIER E.	Electrochimie	
	DESSAUX G.	Physiologie animale	
	PILLET E.	Electrotechnique	
	DEBELMAS J.	Géologie	
	GERBER R.	Mathématiques	
	PAUTHENET R.	Electrotechnique	
	VAUQUOIS B.	Mathématiques appliquées	
	SILBER R.	Mécanique des fluides	
	MOUSSIEGT J.	Electronique	
	BARBIER J.C.	Physique	
	KOSZUL J.L.	Mathématiques	
	BUYLE-BODIN M.	Electronique	

Professeurs sans chaire

Mme.	KOFLER L.	Botanique	
MM.	DREYFUS B.	Thermodynamique	
	VAILLANT F.	Zoologie et hydrobiologie	
	GIRAUD P.	Géologie	
	GIDON P.	Géologie et minéralogie	
	ARNAUD P.	Chimie	
	PERRET R.	Servomécanismes	
Mme.	LUMER L.	Mathématiques	
Mme.	BARBIER M.J.	Electrochimie	
Mme.	SOUTIF J.	Physique	
MM.	BRISSONNEAU P.	Physique	
	COHEN J.	Electrotechnique	
	DEPASSEL R.	Mécanique	

Professeurs associés

MM.	LUMER G.	Mathématiques
	HIGUCHI	Biosynthèse de la cellulose
	WAGNER	Botanique

Maîtres de conférences

MM.	ROBERT A.	Chimie papetière	
	ANGLES D'AURIAC P.	Mécanique des fluides	
	BIAREZ J.P.	Mécanique physique	
	COUMES A.	Electronique	
	DODU J.	Mécanique des fluides	
	DUCROS P.	Minéralogie et cristallographie	
	GLENAT R.	Chimie	
	HACQUES G.	Calcul numérique	
	LANCIA R.	Physique automatique	
	PEBAY-PEROULA J.C.	Physique	
	GASTINEL A.	Mathématiques appliquées	
	LACAZE A.	Thermodynamique	
Mme.	KAHANE J.	Physique	
MM.	DEGRANGE C.	Zoologie	
	GAGNAIRE D.	Chimie papetière	
	RASSAT A.	Chimie systématique	
MM.	KLEIN J.	Mathématiques	
	BETHOUX P.	Mathématiques appliquées	
	POLOUJADOFF M.	Electrotechnique	
	DEPOMMIER P.	Physique nucléaire	
	DEPORTES C.	Chimie	
	BARRA J.	Mathématiques appliquées	
Mme.	BOUCHE L.	Mathématiques	
MM.	PERRIAUX J.	Géologie	
	SARROT-REYNAULD J.	Géologie	
	CAUQUIS G.	Chimie générale	
	LABBE A.	Botanique	
	BONNET G.	Physique générale	
	BARNOUD F.	Biosynthèse de la cellulose	
Mme.	BONNIER M.J.	Chimie	
MM.	KAHANE	Physique générale	
	DOLIQUE	Electronique	

Maîtres de conférences associés

MM.	ISHIKAWA Y.	Magnétisme
	QUATTROPANI	Thermodynamique

Ce travail a été effectué sous la direction de Monsieur le Professeur KUNTZMANN, Directeur de l'Institut de Mathématiques Appliquées de Grenoble. Je tiens à lui exprimer ma plus profonde reconnaissance pour l'intérêt permanent qu'il a manifesté pour mon travail et pour les nombreux conseils qu'il m'a donnés depuis mon arrivée dans son service en 1958.

J'adresse le témoignage de ma respectueuse gratitude à Monsieur FAVARD, Professeur à la Sorbonne et à l'Ecole Polytechnique, qui a accepté de présider le jury. Les indications qu'il m'a données en cours de thèse ont été déterminantes pour la suite de ces recherches.

Je remercie vivement Monsieur GASTINEL qui a constamment suivi la progression de mon travail. Ses conseils, donnés au cours d'innombrables discussions et ses encouragements ont constitué pour moi une aide et un soutien inestimables.

Je remercie également Monsieur BARRA qui a bien voulu accepter de faire partie du jury.

Je voudrais signaler la contribution d'ingénieurs et de programmeurs du Laboratoire de Mathématiques Appliquées pour l'exécution de nombreux calculs. Je les assure de toute ma reconnaissance.

Enfin, que l'imprimerie LOUIS-JEAN trouve ici le témoignage de mon admiration pour la qualité et la rapidité de son travail.

CHAPITRE 1

LE PROCÉDÉ D'EXTRAPOLATION DE RICHARDSON

§ 1.1 - INTRODUCTION

Pour résoudre *numériquement* des équations dans lesquelles interviennent des opérateurs différentiels ou intégraux on ne considère souvent que les valeurs de fonctions en un nombre fini de points de leur domaine de définition : les opérateurs sont remplacés par des approximations qui ne font intervenir que ces points. Aux équations primitives on substitue ainsi un ensemble fini de relations entre les valeurs pontuelles de certaines fonctions. Cette méthode est appelée la discrétisation du problème. Elle est en général caractérisée par un certain pas h : dans le cas simple où le domaine est un intervalle de la droite numérique et où l'on a choisi des abcisses équidistantes, h représente l'écart entre deux points successifs : le pas peut, plus généralement, être un paramètre convenable dont dépend l'ensemble des points utilisés. Le pas h est souvent de la forme $\frac{F}{n}$ ou F est imposé par la géométrie du problème. Si $\Phi(n)$ désigne la solution obtenue avec $h = \frac{F}{n}$ (cela peut-être la valeur d'une fonction en un point, l'intégrale d'une fonction, une valeur propre etc.) on exige :

$$\lim_{n\to\infty} \Phi(n) = \Phi_\infty$$

où Φ_∞ est la solution exacte du problème. En pratique $\Phi(n)$ admet souvent un développement de la forme :

$$\Phi(n) = \Phi_\infty + \frac{A_1}{n} + \frac{C(n)}{n^2}$$

où C est une fonction bornée de n. Et même, en prenant quelques précautions, on peut s'arranger pour que les premiers termes impairs du développement disparaissent :

$$\Phi(n) = \Phi_\infty + \frac{B_2}{n^2} + \frac{D(n)}{n^4}$$

(D, fonction bornée de n).

Pour n assez grand, $\Phi(n) - \Phi_\infty$ se comporte donc comme $\frac{B_2}{n^2}$. Connaissant $\Phi(n_1)$ et $\Phi(n_2)$ on pourra donc former $\Phi^* = a_1 \Phi(n_1) + a_2 \Phi(n_2)$ qui sera en général plus proche de Φ_∞ que $\Phi(n_1)$ ou $\Phi(n_2)$:

$$\Phi^* = \frac{n_1^2}{n_1^2 - n_2^2} \times \Phi(n_1) + \frac{n_2^2}{n_2^2 - n_1^2} \times \Phi(n_2)$$

Si l'on porte $\Phi(n_1)$ à l'abcisse $\left(\frac{1}{n_1}\right)^2$ et $\Phi(n_2)$ en $\left(\frac{1}{n_2}\right)^2$, Φ^* est obtenue par extrapolation linéaire en zéro. En passant $n_2 = k \times n_1$, l'erreur s'écrit :

$$\Phi^* - \Phi_\infty = \frac{1}{n_1^4}\left(\frac{1}{1 - k^2}\right)\left(D(n_1) - \frac{D(n_2)}{k^2}\right)$$

Ce procédé simple a été suggéré par L. F. Richardson [44] et étudié surtout dans l'article qu'il a intitulé : "The deferred approach to the limit", [43] ; voir aussi N. Bogolouboff et N. Kryloff [3].

Comme l'erreur due au terme en $\frac{1}{n^2}$, (donc en h^2), a été éliminée, Richardson a appelé cette méthode une "h^2 - extrapolation". Si $\Phi(n)$ admet un développement de la forme

$$\Phi(n) = \Phi_\infty + \frac{B_2}{n^2} + \frac{B_4}{n^4} + \frac{D(n)}{n^6},$$

en considérant des combinaisons de 3 termes.

$$\Phi^{**} = a\,\Phi(n_1) + b\,\Phi(n_2) + c\,\Phi(n_3)$$

on peut éliminer dans un certain sens la contribution de $\frac{B_2}{n^2}$ et $\frac{B_4}{n^4}$: on aura alors une h^4 - extrapolation. Plus récemment le procédé a été utilisé par Salvadori [47]. Osborne [41] a étudié l'"h^2 - extrapolation" pour des problèmes de valeurs propres. Signalons enfin que la méthode du doublement du pas pour évaluer l'erreur dans l'intégration approchée d'une équation différentielle est basée sur le même principe [25].

Le but de ce travail est de généraliser l'idée de Richardson et de l'utiliser systématiquement. Nous pensons que pour certains problèmes, la règle de calcul rudimentaire que consitue l'"h^2 - extrapolation" peut ainsi devenir une véritable méthode donnant d'excellents résultats.

Le chapitre 1 est consacré à l'étude du procédé en tant qu'extrapolation d'une fonction.

L'existence d'un développement en puissances de $\frac{1}{n}$ pouvant être difficile à prouver en pratique, il est bon d'assurer la convergence du procédé lorsque l'on a seulement :

$$\lim_{n\to\infty} \Phi(n) = \Phi_\infty$$

On donne au chapitre 2 une condition portant sur le choix des pas successifs, nécessaire et suffisante pour que ceci soit réalisé. La stabilité est également étudiée.

Le chapitre 3 étudie l'application du principe de Richardson à l'intégration approchée. On considère l'application de l'extrapolation au résultat de la formule des trapèzes, fonction du pas d'intégration.

Comme cas particulier on trouve au passage la méthode de Romberg [45] que celui-ci a introduite d'une tout autre manière. L'application de l'extrapolation à l'intégration double (ou multiple) nous semble particulièrement intéressante : on obtient de façon simple des formules dont la silhouette de validité, de forme triangulaire est de plus en plus grande.

Le chapitre 4 étudie de manière analogue l'application de l'extrapolation à des formules de dérivation approchée.

Les chapitres 5 et 6 sont consacrés aux équations différentielles (problèmes de conditions initiales ou de conditions aux limites), aux équations aux dérivées partielles (en se limitant à des exemples simples) et aux équations intégrales. On obtient des conditions suffisantes d'existence des premiers termes des développements en puissances de $\frac{1}{n}$ pour ces différents problèmes. Des exemples numériques montrent que pour certains problèmes l'extrapolation est très efficace.

Enfin les chapitres 7 et 8 traitent de l'évaluation d'intégrales par la méthode de Monte-Carlo. De nombreuses techniques de réduction de la variance sont déjà connues : presque toutes nécessitent une étude préalable de la fonction à intégrer. Nous développons une méthode qui n'a pas cet inconvénient mais qui, en revanche, suppose l'existence des premières dérivées : on applique l'échantillonnage systématique sur une transformée de la fonction à intégrer. Le paragraphe 7. 4 fait le lien avec le procédé de Richardson. La méthode de Monte-Carlo ainsi modifiée est utilisée pour résoudre des équations intégrales de Fredholm de 2ème espèce.

Les principaux algorithmes de calcul ont été donnés sous forme de procédures Algol [4]. De nombreuses expériences numériques ont été effectuées afin d'apprécier la valeur pratique des méthodes.

§ 1.2 - EXTRAPOLATION D'UNE FONCTION. ERREUR

Au lieu d'étudier la fonction $\Phi(n)$ quand n tend vers l'infini nous considérons, ce qui revient au même l'extrapolation de la fonction v au point zéro.

$$v\left(\frac{C}{n}\right) = v(h) = \Phi(n)$$

Supposons donc que $v(x)$, $x \in [0, \Omega]$ possède n dérivées sur $[0, \Omega]$; (dérivées à droite en zéro).

On a donc :

$$v(x) = v(0) + x\, v'(0) + \frac{x^2}{2!} v''(0) + \ldots + \frac{x^{n-1}}{(n-1)!} v^{(n-1)}(0) + x^n S(x) \qquad (1.2.1)$$

où S est une fonction bornée de $x \in [0, \Omega]$.

Soit $\{h_k\}$ une suite d'abcisses positives strictement décroissantes et tendant vers 0 quand k tend vers l'infini. Nous supposerons dès maintenant $h_k = \frac{h}{m_k}$ où m_k est un entier. Formons le procédé d'extrapolation destiné à fournir une valeur approchée de $v(0)$:

$$L_n(v) = \sum_{k=1}^{n} A_k^n \, v(h_k) \qquad (1.2.2)$$

Les coefficients A_k^n sont choisis pour que le procédé L_n soit exact pour tout polynôme de degré $n - 1$:

$$L_n(P_{n-1}) = P_{n-1}(0)$$

On a donc :

$$A_k^n = \prod_{\substack{j=1 \\ j \neq k}}^{n} \frac{1}{1 - \frac{h_k}{h_j}} \quad (k \leqslant n) \qquad (1.2.3)$$

Progression des calculs :

On prend en fait la valeur en zéro : $L_n(v)$ d'un polynôme de degré $\leqslant n - 1$ qui coïncide avec v aux abcisses $h_1, h_2, \ldots, h_n$. Mais on ne connaît pas a priori la valeur de n.

On s'intéresse à la suite des valeurs $\{L_n(v)\}$ pour $n = 1, 2, 3$ etc. (l'arrêt est en général déclenché par un test convenable sur les valeurs obtenues). Nous allons comparer entre elles les trois méthodes suivantes :

a) Méthode de Lagrange
b) Méthode de Neville
c) Méthode de Newton.

Nous évaluerons le coût du passage de $L_n(v)$ à $L_{n+1}(v)$ en nombre d'additions-soustractions (A. S.) et de multiplications-divisions (M. D.). (On suppose pour simplifier que ces deux dernières opérations sont d'un coût égal). Pour passer commodément d'un degré au degré suivant on est amené à conserver en mémoires (dans le cas d'un traitement sur calculateur électronique) un certain nombre de quantités : on notera selon les méthodes le nombre de mémoires nécessaires et la nature du contenu.

a) Méthode de Lagrange

On utilise directement les formules (1.2.2) et (1.2.3) et la relation suivante qui s'en déduit :

$$A_i^{n+1} = \frac{h_{n+1}}{h_{n+1} - h_i} \times A_i^n \qquad (1.2.4)$$

Pour passer de $L_n(v)$ à $L_{n+1}(v)$ on garde en mémoires les n quantités :

$$A_1^n\ v(h_1)\ ,\ A_2^n\ v(h_2)\ \ldots\ldots,\ A_n^n\ v(h_n)$$

On les multiplie respectivement par :

$$\frac{h_{n+1}}{h_{n+1} - h_1}\ ,\ \frac{h_{n+1}}{h_{n+1} - h_2}\ \ldots\ldots,\ \frac{h_{n+1}}{h_{n+1} - h_n}$$

On calcule :

$$A_{n+1}^{n+1}\ v(h_{n+1}) = \left(\prod_{j=1}^{n} \frac{h_j}{h_j - h_{n+1}}\right) \times v(h_{n+1})$$

(Noter que les dénominateurs de ce produit ont déjà été calculés pour les facteurs multiplicatifs précédents). On effectuera ensuite la somme des n + 1 quantités $A_i^{n+1}\ v(h_i)$ qui sont gardées en mémoires pour le degré suivant.

Le passage de L_n à L_{n+1} coûte 4n (M. D.) + 2n (A. S.).

b) Méthode de Neville

Il s'agit en fait d'une variante du procédé d'Aitken [26]. Appelons $P_n(x_1, x_2, \ldots, x_n, x)$ le polynôme de degré $\leqslant n - 1$ qui coïncide avec v aux abcisses $x_1, x_2, \ldots, x_n$.

On a la relation :

$$P_{n+1}(x_1, \ldots, x_n, x_{n+1}, x) = \frac{\begin{vmatrix} P_n(x_1, x_2, \ldots, x_{n-1}, x_n, x)\ (x - x_1) \\ P_n(x_2, x_3, \ldots, x_n, x_{n+1}, x)\ (x - x_{n+1}) \end{vmatrix}}{(x_1 - x_{n+1})} \qquad (1.2.5)$$

Pour alléger l'écriture, notons $L_m^n(v)$ la valeur en zéro du polynôme de degré n - 1 qui coïncide avec v aux abcisses

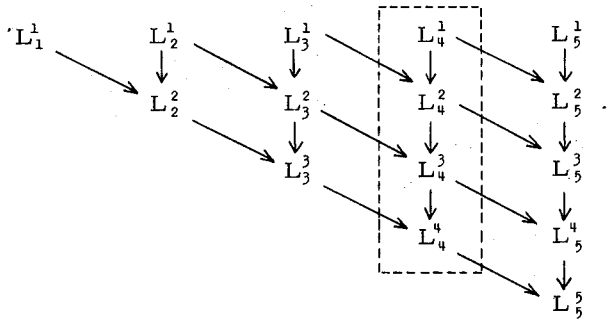
$$h_m,\ h_{m-1},\ \ldots,\ h_{m-n+1} \qquad (m \geqslant n)$$

($L_n^n(v)$ représente ici le $L_n(v)$ précédent)

La formule 1.2.5 s'écrit alors : (x = 0)

$$L_{m+1}^{n+1}(v) = \frac{(h_{m-n+1}) \times L_{m+1}^n(v) - (h_{m+1}) \times L_m^n(v)}{(h_{m-n+1} - h_{m+1})} \qquad (1.2.6)$$

En posant $L_m^1 = v(h_m)$, on a la disposition en triangle suivante :



(1.2.7)

Pour passer de $L_n = L_n^n$ à $L_{n+1} = L_{n+1}^{n+1}$ on garde en mémoire les *n quantités* $L_n^i (i \leqslant n)$ et on ajoute au tableau triangulaire la colonne des L_{n+1}^i en utilisant la formule (1.2.6).

Ce passage coûte 3n (M. D.) + 2n (A. S.).

c) Méthode de Newton

Nous employerons la notation condensée suivante pour les différences divisées :

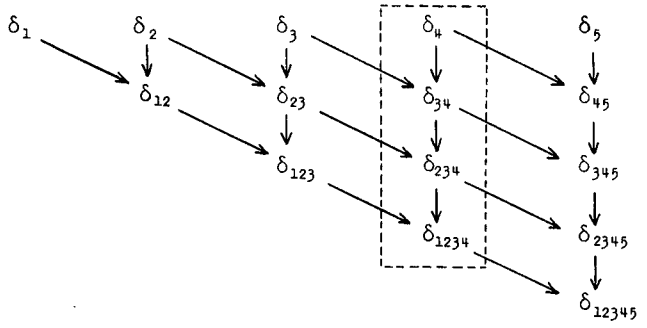
$$\delta(h_i,\ h_j) = \delta_{ij} \qquad \delta(h_i,\ h_j,\ h_k) = \delta_{ijk} \quad \text{etc.}$$

(On pose $v(h_i) = \delta_i$)

La formule de Newton s'écrit avec ces notations [26] :

$$P_{n-1}(x) = \delta_1 + \delta_{12}(x - h_1) + \delta_{123}(x - h_1)(x - h_2) + \ldots + \delta_{12\ldots n}(x - h_1)(x - h_2)\ldots(x - h_{n-1}) \quad (1.2.8)$$

On construit un tableau triangulaire analogue au tableau (1.2.7) pour obtenir les différences divisées :



(1.2.9)

en employant les formules

$$\delta_{i,i+1,i+2,\ldots,k} = \frac{\delta_{i,i+1\ldots,k-1} - \delta_{i+1,\ldots,k}}{h_i - h_k}$$

Pour passer de $L_n(v)$ à $L_{n+1}(v)$ on garde donc en mémoire les n quantités $\delta_n,\ \delta_{n-1,n},\ \ldots,\ \delta_{1,2\ldots n}$.

Le calcul de la n + 1ème colonne coûte n (M. D.) + 2 n (A. S).

On forme ensuite $L_{n+1}(v)$ par la formule suivante :

$$L_{n+1}(v) = L_n(v) + \delta_{1,2\ldots n,n+1} \times (-h_n) \times Q$$

où Q représente la quantité $(-h_1)(-h_2)\ldots(-h_{n-1})$ calculée au tour précédent. Dans ce cas-ci où l'on veut obtenir les valeurs des polynômes pour tous les différents degrés successifs on n'utilise pas le schéma de Hörner. Le passage de $L_n(v)$ à $L_{n+1}(v)$, coûte donc en tout :

$$n + 2 \text{ (M. D.) et } 2n + 1 \text{ (A. S.)}$$

Conclusions :

La méthode de Lagrange est nettement la moins économique des 3 méthodes. La méthode de Newton est moins coûteuse que celle de Neville mais cette dernière présente toutefois l'avantage de manipuler dans le tableau triangulaire des quantités qui sont du même ordre de grandeur (les L_m^n représentent toujours la valeur en zéro d'un certain polynôme).

Calcul de l'erreur : (hypothèse : v dérivable n fois sur $[0,\ \Omega]$)

α) Appelons $Q_n(z)$ le polynôme de degré n - 1 qui coïncide avec v aux abcisses $h_1,\ \ldots,\ h_n$ (on a $Q_n(0) = L_n(v)$) et posons $\omega(z) = (z - h_1)(z - h_2)\ldots(z - h_n)$.

Etudions la fonction g définie par :

$$g(z) = v(z) - Q_n(z) - k\,\omega(z) \qquad \text{avec} \qquad k = \frac{v(0) - Q(0)}{\omega(0)}$$

g s'annule pour $z = h_1,\ h_2,\ \ldots,\ h_n,\ 0$.

g' admet donc au moins n zéros distincts dans $[0,\ \Omega]$ et finalement $g^{(n)}(z) = v^{(n)}(z) - k \times (n!)$ s'annule au moins une fois en $\xi \in [0,\ \Omega]$, d'où

$$k = \frac{v^n(\xi)}{n!}$$

L'erreur est donc de la forme :

$$L_n(v) - v(0) = -\frac{\omega(0)}{n!} \times v^{(n)}(\xi) = (-1)^{n+1} \left(\prod_{i=1}^{n} h_i\right) \times \frac{v^{(n)}(\xi)}{n!} = (-1)^{n+1} \frac{H^n\, v^{(n)}(\xi)}{\left(\prod_{k=1}^{n} m_k\right) n!} \qquad (1.2.10)$$

β) On peut également mettre l'erreur sous forme intégrale [26] :

$$L_n(v) - v(0) = \int_0^{\Omega} \mathcal{E}_n(t)\, v^{(n)}(t)\, dt \qquad (1.2.11)$$

Pour exprimer $\mathcal{E}_n$, écrivons le développement de Taylor de v avec reste sous forme intégrale :

$$v(z) = v(0) + \frac{z\, v'(0)}{1!} = \ldots + \frac{z^{n-1}\, v^{(n-1)}(0)}{(n-1)!} + \int_0^z \frac{(z-t)^{n-1}}{(n-1)!} \times v^{(n)}(t)\, dt \qquad (1.2.12)$$

En tenant compte du fait que L_n est exact pour tout polynôme de degré n - 1, on obtient :

$$L_n(v) - v(0) = \sum_{k=1}^{n} A_k^n \int_0^{h_k} \frac{(h_k - t)^{n-1}}{(n-1)!}\, v^{(n)}(t)\, dt \qquad (1.2.13)$$

Si $\varepsilon_k(t)$ est une fonction valant 1 dans l'intervalle $[0, h_k]$ et 0 ailleurs :

$$\mathcal{E}_n(t) = \sum_{k=1}^{n} \frac{A_k^n\, \varepsilon_k(t)\, (h_k - t)^{n-1}}{(n-1)!} \qquad (1.2.14)$$

Pour étudier $\mathcal{E}_n$, posons

$$\Phi_j(t) = \sum_{k=1}^{n} \frac{A_k^n\, \varepsilon_k(t)\, (h_k - t)^{j-1}}{(j-1)!}$$

Φ_1 est une fonction constante par morceaux, nulle à gauche de 0 et à droite de h : elle a au plus n - 1 changements de signe à l'intérieur de l'intervalle $[0, h_1]$.

On a la propriété :

$$\Phi_j'(t) = -\Phi_{j-1}(t) \quad (j = 2, \ldots, n)$$

(Sauf pour Φ_2 aux points de discontinuité de Φ_1)

Les fonctions Φ_j sont continues en 0 et h_1 et valent zéro en ces points pour j = 2, ..., n : Pour h_1 c'est évident ; pour l'abcisse 0, cela est dû au fait que :

$$\sum_{k=1}^{n} A_k^n (h_k)^{j-1} = 0 \quad \text{pour } j = 2, \ldots, n$$

(Ce sont les équations mêmes qui définissent les A_k^n).

On voit alors facilement que Φ_2 admet au plus n - 2 zéros à l'intérieur de $[0, h_1]$; en continuant Φ_{n-1} admet 1 zéro au plus et $\Phi_n = \mathcal{E}_n$ ne change pas de signe dans l'intervalle $[0, h_1]$.

$$\int_0^{h_1} \mathcal{E}_n(t)\, dt = \sum_{k=1}^{n} \frac{A_k^n}{(n-1)!} \int_0^{h_k} (h_k - t)^{n-1}\, dt = \frac{\sum_{k=1}^{n} A_k^n\, (h_k)^n}{n!}$$

D'après l'expression des A_k^n (1.2.3), on obtient :

$$\int_0^{h_1} \mathcal{E}_n(t)\, dt = \frac{\sum_{k=1}^{n} A_k^n\, (h_k)^n}{n!} = \frac{(-1)^{n+1}}{n!} \prod_{k=1}^{n} (h_k) \qquad (1.2.15)$$

(Notons que (1.2.10) et (1.2.11) appliqués avec $v^{(n)}(t) \equiv 1$ donnent (1.2.15).

Ainsi les noyaux $\mathcal{E}_n$ sont alternativement positifs et négatifs (positifs si n est impair) et admettent un seul extremum entre 0 et h_1. $\mathcal{E}_n(0) = \mathcal{E}_n(h_1) = 0$. $(\mathcal{E}_n)^{(n-2)}$ est encore continue partout mais n'est plus dérivable aux abcisses h_k et zéro. (1.2.16)

On peut démontrer (Kuntzmann [26]) que le noyau garde un signe constant en utilisant (1.2.5) pour certaines fonctions v.

§ 1.3 - ABCISSES EN PROGRESSION GEOMETRIQUE DE RAISON $\frac{1}{2}$

Il est particulièrement intéressant de choisir les abcisses (c'est-à-dire en pratique les pas successifs dans une méthode de résolution approchée) en progression géométrique de raison $\frac{1}{2}$: cela revient à choisir $m_1 = 1$; $m_2 = 2$; ... ; $m_k = 2^{k-1}$. Appelons encore L_m^n le procédé d'extrapolation basé sur les points

$$\frac{h}{2^{m-1}}, \frac{h}{2^{m-2}}, \ldots, \frac{h}{2^{m-n}}$$

valable pour tout polynôme de degré n - 1.

On remarque que L_{m+1}^n représente en fait le procédé L_m^n appliqué avec des abcisses moitié. Ces deux formules étant exactes pour des polynômes de degré n - 1, le procédé a L_{m+1}^n + b L_m^n avec a + b = 1 le sera également. On peut choisir a et b pour que ce nouveau procédé soit exact pour les polynômes de degré n.

L_{m+1}^{n+1} et (a L_{m+1}^n + b L_m^n) utilisent les mêmes abcisses : ils sont donc identiques.

Ecrivons l'erreur sous forme intégrale :

$$L_m^n(v) - v(0) = \int_0^{\frac{h}{2^{m-n}}} \mathcal{E}_m^n(t)\ v^{(n)}(t)\ dt \qquad (1.3.1)$$

On montre que :

$$\mathcal{E}_{m+1}^n(t) = \begin{cases} \dfrac{\mathcal{E}_m^n(2t)}{2^{n-1}} & \text{pour } t \in \left[0,\ \dfrac{h}{2^{m+1-n}}\right] \\ 0 \text{ ailleurs} \end{cases}$$

La formule a L_{m+1}^n + b L_m^n (avec a + b = 1) a une erreur :

$$\int_0^{\frac{h}{2^{m-n}}} \mathcal{E}_m^{n*}(t)\ v^n(t)\ dt$$

avec

$$\mathcal{E}_m^{n*}(t) = \begin{cases} \dfrac{a\ \mathcal{E}_m^n(2t)}{2^{n-1}} + b\ \mathcal{E}_m^n(t) & \text{pour } t \in \left[0,\ \dfrac{h}{2^{m+1-n}}\right] \\ b\ \mathcal{E}_m^n(t) & \text{pour } t \in \left[\dfrac{h}{2^{m+1-n}},\ \dfrac{h}{2^{m-n}}\right] \end{cases} \qquad (1.3.2)$$

Posons :

$$\mathcal{E}_{m+1}^{n+1}(t) = -\int_0^t \mathcal{E}_m^{n*}(x)\ dx \qquad (1.3.3)$$

L'erreur (1.3.2) devient :

$$[-\mathcal{E}_{m+1}^{n+1}(t)\ v^{(n)}(t)]_0^{\frac{h}{2^{m-n}}} + \int_0^{\frac{h}{2^{m-n}}} \mathcal{E}_{m+1}^{n+1}(t)\ v^{(n+1)}(t)\ dt$$

On a déjà $\mathcal{E}_{m+1}^{n+1}(0) = 0$; On choisit a et b pour que

$$\mathcal{E}_{m+1}^{n+1}\left(\frac{h}{2^{m-n}}\right) = 0 \qquad (1.3.4)$$

Cela donne :

$$a = \frac{2^n}{2^n - 1} \qquad b = \frac{-1}{2^n - 1}$$

qui ne dépendent que de n.

On obtient donc L_{m+1}^{n+1} par la formule commode suivante qui est un cas particulier de la formule de Neville (1.2.6) :

$$L_{m+1}^{n+1} = \frac{2^n}{2^n - 1} L_{m+1}^n - \frac{1}{2^n - 1} L_m^n \tag{1.3.4}$$

Le noyau correspondant $\mathcal{E}_{m+1}^{n+1}$ est défini par la formule :

$$\mathcal{E}_{m+1}^{n+1}(t) = \begin{cases} \dfrac{-1}{2^n - 1} \displaystyle\int_t^{2t} \mathcal{E}_m^n(x)\, dx & \text{pour } t \in \left[o, \dfrac{h}{2^{m-n+1}} \right] \\ \dfrac{-1}{2^n - 1} \displaystyle\int_t^{\frac{h}{2^{m-n}}} \mathcal{E}_m^n(x)\, dx & \text{pour } t \in \left[\dfrac{h}{2^{m-n+1}}, \dfrac{h}{2^{m-n}} \right] \end{cases} \tag{1.3.5}$$

Pour faire progresser les calculs on pourra adopter le schéma triangulaire (1.2.7) et la formule (1.3.4).

§ 1.4 - EXTRAPOLATION BASEE SUR DES POLYNOMES PAIRS

Comme nous l'avons signalé dans l'introduction (§ 1.1), on peut presque toujours faire "disparaître" les termes impairs du développement de $\Phi(n)$. Nous allons donc reprendre rapidement l'étude précédente pour une fonction w dérivable jusqu'à l'ordre 2n sur [0, Ω] et dont les dérivées d'ordre impair sont nulles en zéro :

$$w^{(2p+1)}(0) = 0 \quad . \quad (p = 0, 1, \ldots, n - 1)$$

On a donc :

$$w(h) = w(0) + \frac{h^2}{2!} w''(0) + \ldots + \frac{h^{2n-2}}{(2n-2)!} w^{(2n-2)}(0) + h^{2n} S(h) \tag{1.4.1}$$

où S est une fonction bornée de h. On forme le procédé M_n sur les mêmes abcisses h_k que précédemment :

$$M_n(w) = \sum_{k=1}^{n} B_k^n\, w(h_k) \tag{1.4.2}$$

tel que :

$$M_n(Q_{2n-2}) = Q_{2n-2}(0) \text{ pour tout polynome } Q_{2n-2} \text{ pair et de degré } 2n - 2$$

Les relations (1.2.3) et (1.2.4) deviennent :

$$\begin{aligned} B_k^n &= \prod_{\substack{j=1 \\ j \neq k}}^{n} \frac{1}{1 - \left(\frac{h_k}{h_j}\right)^2} \qquad k \leqslant n \\ B_i^{n+1} &= \frac{h_{n+1}^2}{h_{n+1}^2 - h_i^2} \times B_i^n \end{aligned} \tag{1.4.3}$$

et les méthodes de Lagrange, Neville et Newton s'adaptent sans difficulté.

L'erreur est de la forme :

$$M_n(w) - w(0) = \frac{h^{2n}\, w^{(2n)}(\xi)}{\left(\prod_{k=1}^{n} m_k \right)^2 (2n)!} \qquad \text{avec } \xi \in [0, \Omega] \tag{1.4.4}$$

$$M_n(w) - w(0) = \int_o^{\Omega} \mathfrak{M}_{2n}(t)\, w^{(2n)}(t)\, dt$$

avec

$$\mathfrak{M}_{2n}(t) = \sum_{k=1}^{n} \frac{B_k^n\, \varepsilon_k(t)\, (h_k - t)^{2n-1}}{(2n-1)!} \tag{1.4.5}$$

Pour étudier $\mathfrak{M}_{2n}$, en posant

$$\psi_j(t) = \sum_{k=1}^{n} \frac{B_k^n \, \varepsilon_k(t) \, (h_k - t)^{j-1}}{(j-1)!}$$

on peut procéder comme § 1.2.

On remarque cependant que $\psi_j(0^+)$ est différent de 0 pour j pair.

Ainsi, ψ_1 a au plus n - 1 zéros à l'intérieur de $[0, h_1]$ et de même ψ_2. ψ_{2i-1} et ψ_{2i} ont au plus n - i zéros dans $[0, h_1]$. ψ_{2n-1} est donc de signe constant et $\psi_{2n} = \mathfrak{M}_{2n}$ est monotone.

$$\int_0^\Omega \mathfrak{M}_{2n}(t) \, dt = \frac{(-1)^{n+1}}{(2n)!} \times \prod_{i=1}^{n} (h_i)^2 \qquad (1.4.6)$$

Les noyaux $\mathfrak{M}_{2n}$ sont donc alternativement positifs décroissants (n impair) et négatifs croissants.

Abcisses en progression géométrique de raison $\frac{1}{2}$

Soit M_m^n le procédé d'extrapolation exact pour tout polynome pair de degré 2n - 2, utilisant les mêmes abcisses qu'au § 1.3. On a une relation analogue à (1.3.4) qui permet une progression des calculs selon un schéma triangulaire comme en (1.2.7) :

$$M_{m+1}^{n+1} = \frac{2^{2n}}{2^{2n}-1} M_{m+1}^n - \frac{1}{2^{2n}-1} M_m^n \qquad (1.4.7)$$

Si $\mathfrak{M}_m^{2n}$ désigne le noyeu de M_m^n, on a la relation de récurrence suivante :

$$\mathfrak{M}_{m+1}^{2n+1}(t) = \begin{cases} \dfrac{1}{2^{2n}-1} \displaystyle\int_t^{2t} \mathfrak{M}_m^{2n}(x) \, dx & \text{pour } t \in \left[0, \dfrac{h}{2^{m+1-n}}\right] \\ \dfrac{1}{2^{2n}-1} \displaystyle\int_0^{\frac{h}{2^{m-n}}} \mathfrak{M}_m^{2n}(x) \, dx & \text{pour } t \in \left[\dfrac{h}{2^{m+1-n}}, \dfrac{h}{2^{m-n}}\right] \end{cases} \qquad (1.4.8)$$

$$\mathfrak{M}_{m+1}^{2n+2}(t) = \int_{\frac{h}{2^{m-n}}}^{t} \mathfrak{M}_{m+1}^{2n+1}(t) \, dt$$

§ 1.5 - PROCEDURE ALGOL POUR LE PROCEDE D'EXTRAPOLATION DE RICHARDSON

La fonction v sur laquelle porte l'extrapolation n'est pas en général une fonction simple connue explicitement : v(h) est le résultat d'un calcul approché fait avec un pas h. On suppose que v est donnée sous forme d'une procédure Algol. Des exemples d'utilisation seront examinés dans les chapitres suivants.

a) Par la méthode de Lagrange

```
procédure RILAGRANGE (V) vitesse : (ALPHA) maximum : (L)
précision : (EPS) pas : (HD) type : (T) résultats : (R, S) ;
valeur ALPHA, L, EPS, HD ; réel procédure V ; réel ALPHA, EPS, HD ;
entier L, T ; tableau R, S ;
commentaire. Cette procédure effectue une extrapolation à zéro sur la fonction V. On calcule V(H)
pour des pas H successifs qui sont approximativement dans un rapport ALPHA (ALPHA ⩾ 1.25) et les
résultats sont rangés dans le tableau R. Le tableau S reçoit les valeurs extrapolées successives.
On arrête le procédé quand on a calculé L + 1 valeurs de V ou quand l'écart relatif entre 2 valeurs
successives de S est inférieur à EPS. L'extrapolation est basée sur les polynomes ordinaires ou sur
les polynomes pairs selon que T vaut 1 ou 2. Les bornes de Ret S doivent être [0 : L] ;
début tableau W[0 : L] ; entier tableau K[0 : L] ;
entier N, M ; réel HE, U, A, WS, SS ;
      N := 0 ; K[0] := 1 ; K[1] := 2 ;
```

```
RET : HE := HD/K [N] ; R[N] := V(HE) ;
     si N ≠ 0 alors aller à EXTRAPOLATION ;
     S[0] := W[0] := R[0] ; N := N + 1 ; aller à RET ;
EXTRAPOLATION : WS := 1.0 ; SS := 0 ;
pour M := 0 pas 1 jusqua N - 1 faire
       début U := (K[N]/K[M]) ↑ T ; A := 1.0 - U ;
             W[M] := W[M]/A ;
             WS := WS × U/A ; SS := SS + W[M]
       fin ;
W[N] := ABS(WS) × R[N] ; S[N] := SS + W[N] ;
si ABS ((S[N] - S[N - 1]) / S[N]) ⩾ EPS ∧ N < L alors
                                 début  N := N + 1 ;
                                        K[N] := ALPHA × K[N - 1] ;
                                        aller à RET
                                 fin
fin
```

b) Par la méthode de Neville

Nous appellerons cette procédure RINEVILLE ; le reste de la tête de procédure est inchangé par rapport à la procédure RILAGRANGE (y compris le commentaire).

Nous n'écrivons que le corps de procédure :

```
début tableau W[0 : L] ; entier tableau K[0 : L] ;
      entier N, I, KN ; réel HE;
      N := 0 ; K[0] := 1 ; K[1] := 2 ;
   RET : HE := HD/K [N] ; W[N] := R[N] := V(HE) ;

      si N ≠ 0 alors allera EXTRAPOLATION ;
                     S[0] := R[0] ; N := N + 1 ; allerà RET ;
EXTRAPOLATION : KN := K[N] ; pour I := N - 1 pas - 1 jusqu'à 0 faire
      W[I] := W[I + 1] + (W[I + 1] - W[I])/ ((KN /K[I]) ↑ T - 1.0) ;
      S[N] := W[0] ;
si ABS ((S[N] - S[N - 1]) / S[N]) ⩾ EPS ∧ N < L alors
                           début N : = N + 1 ;
                                 K[N] := ALPHA × K[N - 1] ;
                                 allerà RET
                           fin
fin
```

c) Par la méthode Newton

Appelons cette procédure RINEWTON. Nous ne donnons que le corps de la procédure (tête inchangée).

```
début tableau W, HE [0 : L] ; entier N, I ; réel HC, P, DP, HEN ;
N := 0 ; HE[0] := - HD ↑ T ; HC := HD ; DP := 1.0 ;
RET : W[N] := R[N] := V (HC) ;
si N ≠ 0 alors allerà EXTRAPOLATION ;
P := S[0] := R[0] ; N := 1 ; HC := HD/2.0 ;
HE[1] := - HC ↑ T ; allerà RET ;
EXTRAPOLATION : HEN := HE[N] ;
pour I := N - 1  pas - 1  jusqu'à 0 faire
      W[I] := (W[I + 1] - W[I])/(HE[I] - HEN) ;
DP := DP × HE[N - 1] ; S[N] := P := P + DP × W[0] ;
si ABS ((P - S [N - 1])/ P) ⩾ EPS ∧ N < L alors
    début N := N + 1 ; HC := HD / ENTIER (HD × ALPHA/HC + 0.5) ;
          HE[N] := - HC ↑ T ; allerà RET
    fin
fin
```

CHAPITRE 2

CONVERGENCE ET STABILITÉ

§ 2.1 - UN THEOREME DE CONVERGENCE

Dans la plupart des cas, le coût d'un calcul de v(h) ou w(h) n'est pas indépendant de h ; au contraire, ce coût augmente très vite quand h tend vers 0 ; il suffit de penser à la discrétisation d'un problème aux dérivées partielles elliptique sur le carré unitaire : on aboutit à un système linéaire de $(m - 1)^2$ équations avec $m = \frac{1}{h}$. On semble donc avoir intérêt à faire diminuer aussi lentement que possible les abcisses h_i. Dans le cas du problème évoqué, le choix le plus économique serait $h_i = \frac{1}{i}$. Les théorèmes de convergence que nous allons établir montrent que ce choix est impossible et qu'une progression trop lente peut conduire à une divergence. L'application des procédés L_n ou M_n se justifie quand les fonctions v ou w admettent un développement limité à droite au voisinage de l'origine. C'est souvent le cas en pratique, mais comme il est en général difficile de le prouver, il est raisonnable d'exiger que les procédés L_n et M_n convergent encore dans le cas minimum, c'est-à-dire lorsque v ou w sont seulement continues à droite en x = 0. Le théorème suivant permet de choisir les abcisses h_i pour qu'il en soit ainsi :

Théorème 1 : Soit $\{x_k\}$ une suite d'abcisses positives strictement décroissantes et tendant vers 0 quand k tend vers l'infini et L_n le procédé d'extrapolation défini au § 1.2 basé sur les n premières abcisses.

Une condition nécessaire et suffisante pour que l'on ait :

$$\lim_{n\to\infty} L_n(G) = G(0)$$

pour toute fonction G continue à droite en x = 0 est qu'il existe un nombre $\alpha > 1$ tel que pour tout h on ait : $\frac{x_h}{x_{h+1}} \geqslant \alpha$. C'est ce que nous appellerons la condition (α).

Plus généralement, on peut énoncer un théorème relatif à un procédé d'extrapolation basé sur les puissances successives d'une fonction strictement croissante.

Théorème 2 : Soit φ une fonction numérique définie sur $[0, \Omega]$ strictement croissante et continue à droite en zéro. On note φ^n la fonction telle que :

$$\varphi^n(x) = [\varphi(x)]^n, \quad n = 0, 1, 2, \ldots, ;$$

Λ_n le procédé d'extrapolation

$$\Lambda_n(G) = \sum_{k=1}^{n} \alpha_k^n \, G(x_k)$$

tel que

$$\Lambda_n (\Pi_{n-1}) = \Pi_{n-1}(0)$$

pour tout "polynome" de la forme

$$a_{n-1} \varphi^{n-1} + a_{n-2} \varphi^{n-2} + \ldots + a_o \varphi^o$$

§ 2.2

Une condition nécessaire et suffisante pour que l'on ait :

$$\lim_{n\to\infty} \Lambda_n(G) = G(0)$$

pour toute fonction G continue à droite en x = 0 est qu'il existe un nombre $\alpha > 1$ tel que pour tout h on ait :

$$\frac{\varphi(0) - \varphi(x_h)}{\varphi(0) - \varphi(x_{h+1})} \geqslant \alpha$$

Exemples :

a) $\varphi(x) \equiv x$ on obtient le procédé L_n et le th. 1

b) $\varphi(x) \equiv x^2$ on obtient le procédé M_n

c) $\varphi(x) \equiv e^x$

d) $\varphi(x) \equiv \sin x \qquad \left(x \in \left[0, \frac{\pi}{2}\right]\right)$

§ 2.2 - RAPPEL D'UN THEOREME DE BASE

Pour démontrer le théorème 2 nous utiliserons le théorème général suivant (Favard [10] tome 2 pages 40 - 41 ; Krylov [23] page 61 ; Kantorovitch et Akylov [22] pages 229 - 234) :

Théorème 3 : Soit L une application linéaire continue et $\{L_n\}$ une suite d'applications linéaires continues d'un espace de Banach S dans un espace de Banach T. Une condition nécessaire et suffisante pour que l'on ait :

$$\lim_{n\to\infty} L_n(s) = L(s)$$

pour tout $s \in S$ est que :

1/ $\lim_{n\to\infty} L_n(a) = L(a)$ pour tout $a \in A$ ou A est une partie dense dans S, $(\bar{A} = S)$.

2/ Il existe une constante M tel que pour tout n :

$$\| L_n \|_{ST} \leqslant M$$

$\| x \|_S$ désigne la norme d'un élément $x \in S$, $\| y \|_T$ la norme de $y \in T$ et $\| f \|_{ST}$ la norme habituelle d'une application linéaire continue de S dans T.

Condition suffisante : Supposons les 2 conditions vérifiées et prenons un s quelconque de S. Il faut montrer que $\| L_n(s) - L(s) \|_T$ peut être rendu arbitrairement petit pour n > N. On a l'inégalité :

$$\| L_n(s) - L(s) \|_T \leqslant \| L_n(s) - L_n(a) \|_T + \| L_n(a) - L(a) \|_T + \| L(a) - L(s) \|_T$$

Prenons un nombre $\varepsilon > 0$ arbitraire : on peut choisir $a \in A$ tel que $\| a - s \|_S \leqslant \varepsilon$. Comme la convergence a lieu pour $a \in A$, il existe N tel que pour n > N :

$$\| L_n(a) - L(a) \|_T \leqslant \varepsilon$$

On a donc :

$$\| L_n(s) - L(s) \|_T \leqslant \| L_n \|_{ST} \times \| s - a \|_S + \varepsilon + \| L \|_{ST} \times \varepsilon \leqslant (M + \| L \|_{ST} + 1) \times \varepsilon = K \times \varepsilon$$

Condition nécessaire : Supposons que la convergence ait lieu pour tout $s \in S$. La première condition est évidemment vérifiée. Choisissons un nombre $\varepsilon > 0$ arbitraire.

Pour tout $s \in S$ il existe N_s tel que pour tout $n > N_s$ on ait :

$$\| L_n(s) - L(s) \|_T \leqslant \varepsilon$$

Pour tout $s \in S$ on peut donc trouver M_s tel que pour tout n on ait :

$$\| L_n(s) \|_T \leqslant M_s$$

D'après le théorème de la borne uniforme de Banach-Steinhaus, ([58] page 135) il existe M tel que pour tout n et tout s on ait :

$$\| L_n(s) \|_T \leqslant M \| s \|_s$$

donc :

$$\| L_n \|_{ST} \leqslant M$$

La condition nécessaire est donc démontrée.

§ 2.3 - DEMONSTRATION DU THEOREME DE CONVERGENCE (th. 2)

On peut raisonner, en fait, sur la suite convergente $g = \{G(x_k)\}$. Soit S l'espace de Banach des suites convergentes avec la norme $\| g \|_s = \sup_k |g_k|$. La droite numérique jouera le rôle de l'espace de Banach T du théorème 3. Appelons ψ^i l'élément de S tel que $\psi^i_j = \psi^i(x_j)$ et A l'ensemble des éléments qui sont combinaisons linéaires finies de ψ^i.

A est dense dans S : $\bar{A} = S$. On démontre facilement cette propriété à l'aide du théorème de Stone-Weierstrass. En effet, $\{0, x_k\}$ est une partie compacte de R ; $\bar{A}$ est une algèbre de fonctions continues à valeurs réelles qui séparent les points de $\{0, x_k\}$: si $x_i \neq x_j$ il existe toujours $a \in A$ tel que $a(x_i) \neq a(x_j)$ (prendre par exemple $a = \varphi$ qui est strictement croissante). Puisque A contient les fonctions constantes, de la forme $\lambda \times \psi^o$, la fermeture uniforme $\bar{A}$ de A (c'est-à-dire au sens de la norme introduite) est l'ensemble des fonctions continues sur $\{0, x_k\}$ c'est-à-dire S. L'application Λ_n de S dans R :

$$\Lambda_n(g) = \sum_{k=1}^{n} \alpha_k^n g_k$$

est une fonctionnelle linéaire continue de norme

$$\| \Lambda_n \|_{ST} = \sum_{k=1}^{n} |\alpha_k^n|$$

La fonctionnelle Λ définie par $\Lambda(g) = \lim_{k \to \infty} g_k$ est aussi linéaire continue et sa norme vaut 1. Nous sommes donc dans les conditions d'application du théorème 3. Une condition nécessaire et suffisante pour que l'on ait :

$$\lim_{n \to \infty} \Lambda_n(g) = \Lambda(g) = G(0)$$

pour tout $g \in S$ est qu'il existe M tel que pour tout n on ait :

$$\sum_{k=1}^{n} |\alpha_k^n| \leqslant M \qquad (2.3.1)$$

Montrons que la condition (α) du th. 2 est équivalente à (2.3.1).

Condition suffisante :

Supposons la propriété (α) vérifiée :

$$\frac{\varphi(0) - \varphi(x_h)}{\varphi(0) - \varphi(x_{h+1})} \geqslant \alpha \quad \text{avec } \alpha > 1$$

a) Les coefficients α_n^n sont bornés :

$$\alpha_n^n = \prod_{i=1}^{n-1} \frac{1}{1 - \dfrac{\varphi(0) - \varphi(x_n)}{\varphi(0) - \varphi(x_i)}}$$

§ 2.3

Posons :

$$\beta = \frac{1}{\alpha} \quad , \quad (\beta < 1)$$

On a alors :

$$\alpha_n^n \leqslant \prod_{k=1}^{n-1} \frac{1}{1 - \beta^{k'}} = P_n$$

P_n converge en croissant vers p. On a donc pour tout n : $\alpha_n^n < p$

b) Les coefficients α_i^n sont bornés :

D'après une formule analogue à (1.2.4) on a :

$$\alpha_n^{n+j} = \frac{1}{1 - \dfrac{\varphi(0) - \varphi(x_n)}{\varphi(0) - \varphi(x_{n+j})}} \times \ldots \times \frac{1}{1 - \dfrac{\varphi(0) - \varphi(x_n)}{\varphi(0) - \varphi(x_{n+2})}} \times \frac{1}{1 - \dfrac{\varphi(0) - \varphi(x_n)}{\varphi(0) - \varphi(x_{n+1})}} \times \alpha_n^n$$

Les facteurs multiplicatifs sont bornés par $s = \dfrac{1}{\alpha - 1}$. Si δ est un nombre arbitraire inférieur à 1, il y en a au plus un nombre l (dépendant de α et δ seulement) qui sont $\geqslant \delta$: prendre l tel que :

$$\alpha^l > \frac{1}{\delta} + 1$$

On a donc :

$$|\alpha_n^{n+j}| \leqslant s^l \times p = E$$

c) Etant donné un nombre $r < 1$ on peut lui associer χ entier tel que, pour tout i et tout m on ait :

$$|\alpha_i^{m+\chi}| < r\ |\alpha_i^m|$$

On a la formule :

$$\alpha_i^{m+\chi} = \frac{1}{1 - \dfrac{\varphi(0) - \varphi(x_i)}{\varphi(0) - \varphi(x_{m+\chi})}} \times \ldots \times \frac{1}{1 - \dfrac{\varphi(0) - \varphi(x_i)}{\varphi(0) - \varphi(x_{m+1})}} \times \alpha_i^m$$

L'entier l étant choisi comme en b), le produit des l premiers facteurs (à partir de la droite) sera inférieur à s^l. Les facteurs suivants sont tous inférieurs à $\delta < 1$. Il en faudra un nombre fini fixé σ pour rendre le produit total inférieur à r :

$$\delta^\sigma \times s^l < r$$

On prend pour χ l'entier immédiatement supérieur à :

$$\sigma + l = \frac{\log r}{\log \delta} + l \times \left(1 - \frac{\log s}{\log \delta}\right)$$

qui ne dépend pas de i et m.

d) Il existe M tel que pour tout n : $\sum_{k=1}^{n} |\alpha_k^n| \leqslant M$.

En effet :

$$\sum_{k=1}^{n} |\alpha_k^n| = \sum_{k=1}^{\text{n-entier } (n/\chi) \times \chi} + \ldots + \sum_{k=n-2\chi+1}^{n-\chi} + \sum_{k=n-\chi+1}^{n}$$

$$\leqslant \chi\ E\ \left\{r^{\text{entier } (n/\chi)} + \ldots + r^2 + r + 1\right\}$$

donc :

$$\| A_n \|_{ST} = \sum_{k=1}^{n} | \alpha_k^n | < \frac{\chi E}{1 - r} = M$$

indépendant de n.

La condition suffisante de th. 3 entraîne alors celle du th. 2.

Condition nécessaire :

Montrons que si la condition (α) n'est pas vérifiée, les quantités

$$\sum_{k=1}^{n} | \alpha_k^n |$$

ne sont pas bornées par un nombre M indépendant de n.

Tous les facteurs dans l'expression de α_n^n sont > 1. Il suffit que l'un d'eux ne soit pas borné pour que α_n^n et par suite

$$\sum_{k=1}^{n} | \alpha_n^n |$$

ne le soient pas.

Or, pour tout W > 1 on peut trouver n tel que :

$$\frac{\varphi(0) - \varphi(x_n)}{\varphi(0) - \varphi(x_{n-1})} > 1 - \frac{1}{W}$$

(c'est la négation de la propriété (α))
donc :

$$\frac{1}{1 - \dfrac{\varphi(0) - \varphi(x_n)}{\varphi(0) - \varphi(x_{n-1})}} > W$$

La condition nécessaire du théorème 3 permet alors d'affirmer qu'il existe un $g \in S$ soit encore une fonction G continue à droite en zéro pour laquelle on a divergence. Cela termine la démonstration du théorème 2.

Remarque : Un choix d'abcisses tel que $x_i = \frac{1}{i}$ est donc à éviter : il ne vérifie pas la condition (α).

§ 2.4 - STABILITE DU PROCEDE D'EXTRAPOLATION

Le coefficient $\alpha > 1$ étant choisi, on prendra les abcisses successives par la formule $h_{i+1} = \frac{h_i}{\alpha}$. Si les abcisses sont de la forme $\frac{C}{k_i}$ (k_i entier) on aura :

$$k_{i+1} = \text{Entier } (\alpha \times k_i) + 1$$

En pratique on ne pourra pas prendre un coefficient α trop voisin de 1. Si l'on commet une erreur inférieure à ε en valeur absolue (arrondi ou autre) sur chaque $v(h_i)$, l'erreur sur $L_n(v)$ sera inférieure à :

$$\varepsilon \times \sum_{k=1}^{n} | A_k^n |$$

La quantité

$$\sum_{k=1}^{n} | A_k^n |$$

tend en croissant vers une limite quand n tend vers l'infini. Il est intéressant d'étudier numériquement cette limite en fonction de α. Ce sera une borne (indépendante de n) du coefficient par lequel on multiplie l'erreur ε.

$$\sum_1(\alpha) = \lim_{n\to\infty} \sum_{k=1}^{n} |A_k^n|$$

$$\sum_2(\alpha) = \lim_{n\to\infty} \sum_{k=1}^{n} |B_k^n|$$

On a évidemment $\sum_2(\alpha) = \sum_1(\alpha^2)$

$\alpha \rightarrow$	2	1,9	1,8	1,7	1,6	1,5	1,4	1,3
$\sum_1(\alpha)$	8,25	10	14	21	37	79	250	1680
$\sum_2(\alpha)$	1,97	2,2	2,5	3,0	3,8	5,3	9,05	22

Si l'on veut un facteur < 10 il faut prendre $\alpha \sim 2$ pour L_n et $\alpha \sim 1,4$ pour M_n.

Par ailleurs, pour évaluer v(0), il semble raisonnable d'exiger que les abcisses interviennent avec un poids d'autant plus grand en valeur absolue qu'elles sont plus proches de zéro.

On démontre que ceci est réalisé :

pour L_n si $\alpha \geqslant 2$

pour M_n si $\alpha \geqslant \sqrt{2}$

Ces deux valeurs semblent être les plus intéressantes. Les expériences numériques le confirment (voir chapitre 6).

CHAPITRE 3

APPLICATION DU PROCÉDÉ D'EXTRAPOLATION DE RICHARDSON A L'INTÉGRATION NUMÉRIQUE

§ 3.1 - EXTRAPOLATION SUR LA FORMULE DES TRAPEZES

Considérons la formule des trapèzes avec un pas $h = \frac{1}{m}$ pour le calcul de l'intégrale définie $\int_0^1 F(x)\,dx$. On supposera l'existence des premières dérivées de F jusqu'à l'ordre utilisé dans les formules :

$$T_m\ F = \frac{1}{m} \sum_{j=0}^{m-1} \frac{F(j/m) + F((j+1)/m)}{2} \qquad (3.1.1)$$

L'erreur se met sous la forme :

$$T_m\ F = \int_0^1 F(X)\,dX - \frac{h^2}{2} \int_0^1 \left(\bar{B}_2\left(\frac{t}{h}\right) - B_2\right) F^{(2)}(t)\,dt \qquad (3.1.2)$$

avec $B_2(t)$ polynôme de Bernoulli de degré 2 : $B_2(t) = t^2 - t + \frac{1}{6}$

B_2 nombre de Bernoulli : $B_2 = B_2(0)$

$\bar{B}_2(t)$ fonction périodique, de période 1, coïncidant avec $B_2(t)$ sur l'intervalle $[0,1]$.

On introduit les polynômes de Bernoulli de degré supérieur [9] définis par les relations :

$$\frac{d\,B_n(x)}{dx} = n\,B_{n-1}(x) \quad ; \quad \int_0^1 B_n(x)\,dx = 0 \qquad (3.1.3)$$

et $\bar{B}_n(x)$ la fonction pénodique, de période 1, qui coïncide avec $B_n(x)$ sur $[0,1]$. On pose $B_n = B_n(0)$, nombre de Bernoulli. (Pour les propriétés de ces polynômes voir [10], [13], [26]).

En intégrant par parties la formule (3.1.2) on obtient :

$$T_m\ F = \int_0^1 F(x)\,dx + \sum_{p=1}^{n-1} \frac{h^{2p}\ B_{2p}}{(2p)!} \left[F^{(2p-1)}(1) - F^{(2p-1)}(0)\right] - \frac{h^{2n}}{(2n)!} \int_0^1 \left(\bar{B}_{2n}\left(\frac{t}{h}\right) - B_{2n}\right) F^{(2n)}(t)\,dt \qquad (3.1.4)$$

Posons $T_m\ F = w(h)$. Le développement limité de w au voisinage de l'origine n'a pas de termes impairs. w(h) se comporte au voisinage de zéro comme un polynôme pair de h. Pour que ce développement limité existe il suffit que les premières dérivées de F existent.

Pour trouver $w(0) = \int_0^1 F(x)\,dx$ on peut appliquer le procédé d'extrapolation M_n du § 1.4 :

$$M_n(T_m\ F) = M_n(w) = \sum_{k=1}^{n} B_k^n(Tm_k\ F) \qquad (3.1.5)$$

(On a choisi des pas $h_k = \frac{h}{m_k}$, h étant un pas de départ et les m_k des entiers).

La formule d'intégration approchée (3.1.5) donne le résultat exact si F est un polynôme de degré 2n - 1 au plus.

Evaluons l'erreur :

$$M_n(T_m F) - \int_0^1 F(x)\,dx = M_n\left\{\int_0^1 F(x) + \sum_{p=1}^{n-1} \frac{h^{2p} B_{2p}}{2p!} [F^{2p-1}(1) - F^{2p-1}(0)]\right\} - \int_0^1 F(x)\,dx$$

$$- M_n\left\{\frac{h^{2n}}{2n!} \int_0^1 \left(\bar{B}_{2n}\left(\frac{t}{h}\right) - B_{2n}\right) F^{(2n)}(t)\,dt\right\}$$

$$M_n(T_m F) - \int_0^1 F(x)\,dx = \frac{-h^{2n}}{(2n)!} \int_0^1 \left[\sum_{k=1}^{n} \frac{B_k^n}{(m_k)^{2n}} \left(\bar{B}_{2n}\left(\frac{t}{h_k}\right) - B_{2n}\right)\right] F^{(2n)}(t)\,dt \qquad (3.1.6)$$

L'erreur s'exprime sous forme d'une intégrale portant sur la dérivée 2n ème de F, le noyau étant l'expression entre crochets.

Essayons de majorer l'expression (3.1.6) :

$$|M_n(T_m F) - \int_0^1 F(x)\,dx| \leq \frac{h^{2n}}{2n!} \int_0^1 \left|\sum_{k=1}^{n} \frac{B_k^n}{(m_k)^{2n}} \left(\bar{B}_{2n}\left(\frac{t}{h_k}\right) - B_{2n}\right)\right| dt \times \left(\operatorname{Max}_{t\in[0,1]} |F^{(2n)}(t)|\right) \qquad (3.1.7)$$

$$\int_0^1 \left|\sum_{k=1}^{n} \frac{B_k^n}{(m_k)^{2n}} \left(\bar{B}_{2n}\left(\frac{t}{h_k}\right) - B_{2n}\right)\right| dt \leq \sum_{k=1}^{n} \left(\frac{|B_k^n|}{(m_k)^{2n}} \int_0^1 |\bar{B}_{2n}\left(\frac{t}{h_k}\right) - B_{2n}|\,dt\right)$$

Comme $\bar{B}_{2n}\left(\frac{t}{h_k}\right) - B_{2n}$ garde un signe constant (voir propriétés des polynômes de Bernoulli [13]) on a d'après (3.1.3) :

$$\int_0^1 |\bar{B}_{2n}\left(\frac{t}{h_k}\right) - B_{2n}|\,dt = |B_{2n}|$$

Et finalement :

$$|M_n(T_m F) - \int_0^1 F(x)\,dx| \leq \frac{h^{2n}}{(2n)!} |B_{2n}| \left(\sum_{k=1}^{n} \frac{|B_k^n|}{(m_k)^{2n}}\right) \left(\operatorname{Max}_{t\in[0,1]} |F^{(2n)}(t)|\right) \qquad (3.1.8)$$

La borne est donc de la forme :

$$k_n h^{2n} \left(\operatorname{Max}_{t\in[0,1]} |F^{(2n)}(t)|\right)$$

D'après le théorème 2 du chapitre 2 on peut choisir les m_k de la façon suivante :

$$m_{k+1} = \text{entier } (q \times m_k) + 1 \qquad \text{avec } q > 1 \qquad (3.1.9)$$

On a vu au § 2.4 qu'il fallait de préférence $q \geq \sqrt{2}$.

$M_n(T_m F)$ converge alors vers $\int_0^1 F(x)\,dx$ quand n tend vers l'infini si $T_m F$ converge lui-même quand m tend vers l'infini.

Comme la formule des trapèzes converge pour tout F intégrable Riemann, il en sera de même pour le procédé $M_n T_m$.

Dans le paragraphe suivant, nous étudierons plus en détail le cas où q = 2.

3.2 - CHOIX DES PAS EN PROGRESSION GEOMETRIQUE DE RAISON $\frac{1}{2}$; METHODE DE ROMBERG

Si l'on applique le procédé du paragraphe précédent avec les pas successifs $h_k = \frac{h}{2^{k-1}}$ on obtient une méthode particulièrement intéressante qui a été utilisée pour la première fois par Romberg [45]. Celui-ci l'a d'ailleurs introduite d'une toute autre façon qui ne montre pas qu'il s'agit en fait d'une extrapolation sur la formule des trapèzes considérée comme fonction du pas. D'autres auteurs se sont intéressés à cette méthode : Stiefel [52], [53] et Rutishauser [46], [52] ont étudié les coeffi-

cients de la formule de quadrature, l'erreur de la méthode et le mode de convergence quand la fonction est holomorphe sur [0,1] dans le plan complexe : Bauer [2] a étudié le noyau d'erreur.

Introduisons d'abord quelques notations :

Appelons T_m^n la formule d'intégration qui résulte d'une extrapolation basée sur les pas :

$$\frac{h}{2^{m-1}}, \frac{h}{2^{m-2}}, \ldots\ldots, \frac{h}{2^{m-n}} \tag{3.2.1}$$

T_m^n donne le résultat exact quand F est un polynôme de degré 2n - 2 quel que soit $m \geqslant n$.

On a donc :

$$T_m^1 = T_{\frac{h}{2^{m-1}}}$$

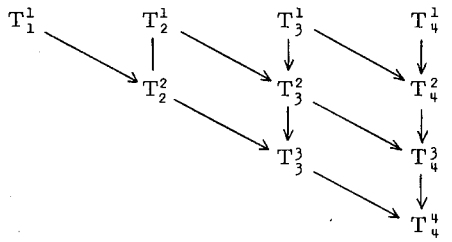
On procède comme aux paragraphes 1.3 et 1.4 : T_{m+1}^n représente en fait la formule T_m^n appliquée avec un pas moitié. Ces deux formules étant exactes pour des polynômes de degré 2n - 1 on peut choisir a et b (avec a + b = 1) tel que la formule :

$$a\ T_{m+1}^n + b\ T_m^n$$

soit exacte pour des polynômes de degré 2n + 1. Cette nouvelle formule utilise en fait les mêmes abcisses que T_{m+1}^n et elle est une combinaison linéaire de T_{m-n+1}^1, T_{m-n+2}^1, ..., T_{m+1}^1 : il s'agit donc de la formule T_{m+1}^{n+1}. Les coefficients a et b sont bien entendu les mêmes que ceux trouvés en (1.4.7) ; il s'agit en fait d'un cas particulier de la méthode de Neville pour polynômes pairs :

$$T_{m+1}^{n+1} = \frac{2^{2n}}{2^{2n} - 1}\ T_{m+1}^n - \frac{1}{2^{2n} - 1}\ T_m^n \tag{3.2.2}$$

La progression des calculs se fera donc suivant le tableau triangulaire suivant :



exactes pour polynômes de degré 1.
" " " 3.
" " " 5.
" " " 7.

(3.2.3)

Si la précision de T_4^4 ne suffit pas, on ajoute une 5ème colonne qui fournit T_5^5 et ainsi de suite.

Examinons les premières formules :

$$T_2^2 = \frac{4}{3}\ T_2^1 - \frac{1}{3}\ T_1^1$$

$$= \frac{1}{6}\ [F(0) + 4\ F(0,5) + F(1)] \qquad (\text{Si } h = 1) \tag{3.2.4}$$

On trouve donc la formule de Simpson.

Désignons par $\mathfrak{C}_2^2(t)$ le noyau d'erreur de T_1^1 :

$$\mathfrak{C}_2^2(t) = -\frac{h^2}{2!}\left(\overline{B}_2\left(\frac{t}{h}\right) - B_2\right)$$

$$T_1^1\ F - \int_0^1 F(x)\ dx = \int_0^1 \mathfrak{C}_2^2(t)\ F^{(2)}(t)\ dt$$

Posons :

$$\mathcal{C}_4^2(t) = M_2 \left\{ -\frac{h^2}{2!} \left(\bar{B}_2\left(\frac{t}{h}\right) - B_2 \right) \right\}$$

$$= -\frac{h^2}{2!3} \left(\left[\bar{B}_2\left(\frac{2t}{h}\right) - B_2 \right] - \left[\bar{B}_2\left(\frac{t}{h}\right) - B_2 \right] \right)$$

L'erreur de T_2^2 F s'écrit :

$$T_2^2 F - \int_0^1 F(x)\, dx = \int_0^1 \mathcal{C}_4^2(t)\, F^{(2)}(t)\, dt$$

et en intégrant 2 fois par parties :

$$T_2^2 F - \int_0^1 F(x)\, dx = -\int_0^1 \mathcal{C}_4^3(t)\, F^{(3)}(t)\, dt = \int_0^1 \mathcal{C}_4^4(t)\, F^{(4)}(t)\, dt$$

avec

$$\begin{aligned} \mathcal{C}_4^3(t) &= -\frac{h^3}{(3!)\times 3} \left[\frac{1}{2} \bar{B}_3\left(\frac{2t}{h}\right) - \bar{B}_3\left(\frac{t}{h}\right) \right] \\ \mathcal{C}_4^4(t) &= -\frac{h^4}{(4!)\times 3} \left[\frac{1}{4} \left(\bar{B}_4\left(\frac{2t}{h}\right) - B_4 \right) - \left(\bar{B}_4\left(\frac{t}{h}\right) - B_4 \right) \right] \end{aligned} \qquad (3.2.5)$$

Si l'on désigne par $\hat{\mathcal{C}}_2^2$ la fonction périodique, de période 1 qui coïncide avec $\mathcal{C}_2^2$ sur [0,1], on passe de $\mathcal{C}_2^2$ à $\mathcal{C}_4^4$ par les formules suivantes :

$$\begin{cases} \mathcal{C}_4^2(t) = \frac{1}{3} \left[\hat{\mathcal{C}}_2^2(2t) - \mathcal{C}_2^2(t) \right] \\ \mathcal{C}_4^3(t) = \int_0^t \mathcal{C}_4^2(t)\, dt \\ \mathcal{C}_4^4(t) = \int_0^t \mathcal{C}_4^3(t)\, dt \end{cases} \qquad (3.2.6)$$

La formule suivante, T_3^3 s'écrit :

$$\begin{aligned} T_3^3 &= \frac{16}{15} T_3^2 - \frac{1}{15} T_2^2 \\ &= \frac{64}{45} T_3^1 - \frac{20}{45} T_2^1 + \frac{1}{45} T_1^1 \\ &= \frac{1}{90} (7\, F(0) + 32\, F(0,25) + 12\, F(0,5) + 32\, F(0,75) + 7\, F(1)) \end{aligned} \qquad (3.2.7)$$

(cette dernière ligne suppose h = 1).

On obtient la formule du type interpolation de Newton-Cotes à 5 abcisses.

Pour n > 3, T_n^n ne coïncide plus avec les formules de Newton-Cotes ou avec des formules connues par ailleurs. On peut dès maintenant dégager les avantages des formules T_n^n par rapport aux formules de Newton-Cotes par exemple :

α) Si l'on veut passer d'une formule de Newton-Cotes d'un certain degré à une autre de degré plus élevé il faut pratiquement refaire tous les calculs et de plus la formation des coefficients n'est pas simple ; au contraire la formule (3.2.2) permet d'atteindre de façon simple et itérative des formules d'ordre aussi élevé qu'on le désire. Chaque fois qu'on ajoute les abcisses d'une formule de trapèzes de pas plus faible, on utilise les résultats de toutes les formules de trapèzes précédentes. On tient compte au maximum des résultats déjà acquis.

β) Kuzmin [27] a montré que les coefficients de Newton-Cotes augmentaient indéfiniment en valeur absolue quand l'ordre n tend vers l'infini. Ainsi une petite erreur sur un calcul de la fonction F peut donc produire une grande erreur pour la valeur approchée de l'intégrale. Si l'on appelle NC(n) la somme des valeurs absolues des coefficients pour la formule de Newton-Cotes d'ordre n et SR(n) la même quantité pour T_n^n, une erreur de ± ε sur chaque valeur de F produira

sur la valeur finale les erreurs NC(n) × ε et SR(n) × ε. Or, NC(n) tend vers l'infini avec n. Au contraire, d'après le tableau du § 2.4, comme T_n^n provient d'une extrapolation pour polynômes pairs avec $\alpha = 2$ on peut déjà affirmer que $SR(n) \leqslant 2$. (On obtiendra un résultat plus précis au § 3.4).

γ) Rappelons le théorème de convergence de Polya-Steckloff [42], [50] : Pour que le procédé de quadrature approchée $Q_n(f) = \sum_{k=1}^{n} q_k^n f(x_k^n)$, (avec $x_k^n \in [0,1]$) converge pour toute fonction *continue* sur [0,1] il faut et il suffit :

1/ qu'il converge pour les polynômes ;

2/ qu'il existe une constante M tel que pour tout n on ait

$$\sum_{k=1}^{n} |q_k^n| \leqslant M$$

(La condition suffisante a été montrée par Steckloff en 1916 ; la difficile condition nécessaire a été montrée directement par Polya en 1933 sans le secours du théorème de Banach-Steinhaus. Il s'agit bien en fait d'une application directe du théorème 3 du § 2.2). Puisque NC(n) n'est pas borné, il existe des fonctions continues pour lesquelles le procédé de Newton-Cotes diverge. En revanche T_n^n converge pour toute fonction continue (et même pour toute fonction intégrable Riemann, voir § 2.1 et § 3.1).

§ 3.3 - ETUDE DU NOYAU D'ERREUR

Appelons $\mathcal{C}_{2n}^{2n}(t)$ le noyau d'erreur de la formule T_n^n

$$T_n^n F - \int_0^1 F(X)\, dX = \int_0^1 \mathcal{C}_{2n}^{2n}(t)\, F^{(2n)}(t)\, dt \qquad (3.3.1)$$

On passe de $\mathcal{C}_{2n}^{2n}$ à $\mathcal{C}_{2n+2}^{2n+2}$ par les formules suivantes qui généralisent (3.2.6) :

$$\begin{cases} \mathcal{C}_{2n+2}^{2n}(t) = \dfrac{1}{2^{2n} - 1} \left(\hat{\mathcal{C}}_{2n}^{2n}(2t) - \mathcal{C}_{2n}^{2n}(t)\right) \\ \mathcal{C}_{2n+2}^{2n+1}(t) = \displaystyle\int_0^t \mathcal{C}_{2n+2}^{2n}(t)\, dt \\ \mathcal{C}_{2n+2}^{2n+2}(t) = \displaystyle\int_0^t \mathcal{C}_{2n+2}^{2n+1}(t)\, dt \end{cases} \qquad (3.3.2)$$

Nous étudierons le noyau pour h = 1. Pour $h = \frac{1}{N}$ on a répétition de la même forme de noyau dans chaque intervalle de longueur $\frac{1}{N}$.

$$\mathcal{C}_2^2(t) = \frac{1}{2}\left(\bar{B}_2(t) - B_2\right) = \frac{x(1-x)}{2}$$

est une *fonction symétrique par rapport à la droite* $x = \frac{1}{2}$ *et strictement croissante positive pour* $x \in \left[0, \frac{1}{2}\right]$.

Supposons que $\mathcal{C}_{2n}^{2n}$ possède ces propriétés et montrons qu'il en sera alors de même pour $\mathcal{C}_{2n+2}^{2n+2}$:

Etudions géométriquement la fonction :

$$\hat{\mathcal{C}}_{2n}^{2n}(2t) - \mathcal{C}_{2n}^{2n}(t) \quad \text{pour} \quad t \in \left[0, \frac{1}{2}\right]$$

Pour $t \in \left[0, \frac{1}{4}\right]$, $\hat{\mathcal{C}}_{2n}^{2n}(2t) = \mathcal{C}_{2n}^{2n}(2t)$ et l'on a :

$$\mathcal{C}_{2n}^{2n}(2t) > \mathcal{C}_{2n}^{2n}(t)$$

à cause de la propriété de croissance.

§ 3.3

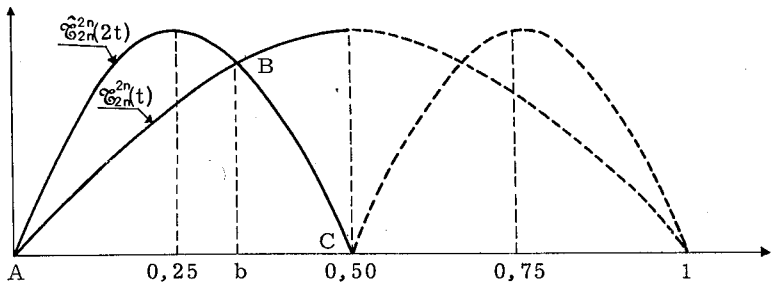
Pour

$$t \in \left[\frac{1}{4}, \frac{1}{2}\right], \ \hat{\mathcal{C}}_{2n}^{2n}(2t)$$

est décroissante alors que $\mathcal{C}_{2n}^{2n}(t)$ est croissante : la fonction

$$\hat{\mathcal{C}}_{2n}^{2n}(2t) - \mathcal{C}_{2n}^{2n}(t)$$

admet donc au plus un zéro dans l'intervalle $\left[0, \frac{1}{2}\right]$.



L'arc AB est la transformé de l'arc CB par une affinité d'axe Bb et de rapport - 2. On a donc ab = 2bC

$\mathcal{C}_{2n+2}^{2n}(t)$ s'annule donc pour $t = \frac{1}{3}$.

$\mathcal{C}_{2n+2}^{2n+1}(t)$ est donc strictement croissante de 0 à $\frac{1}{3}$ et strictement décroissante ensuite de $\frac{1}{3}$ à $\frac{1}{2}$.

Comme $\mathcal{C}_{2n+2}^{2n+1}\left(\frac{1}{2}\right) = 0$, puisque

$$\int_0^{1/2} \hat{\mathcal{C}}_{2n}^{2n}(2t)\, dt = \int_0^{1/2} \mathcal{C}_{2n}^{2n}(t)\, dt,$$

$\mathcal{C}_{2n+2}^{2n+1}$ est strictement positif sur $]0, \frac{1}{2}[$.

On a :

$$\mathcal{C}_{2n+2}^{2n+1}(1 - t) = - \mathcal{C}_{2n+2}^{2n+1}(t)$$

car

$$\mathcal{C}_{2n+2}^{2n+1}(1 - t) = \underbrace{\int_0^1 \mathcal{C}_{2n+2}^{2n}(t)\, dt}_{0} - \int_{1-t}^1 \mathcal{C}_{2n+2}^{2n}(t)\, dt$$
$$= - \int_0^t \mathcal{C}_{2n+2}^{2n}(t)\, dt = - \mathcal{C}_{2n+2}^{2n+1}(t)$$

Donc $\mathcal{C}_{2n+2}^{2n+1}(t)$ est négatif pour $t \in]\frac{1}{2}, 1[$

On montre de même que

$$\mathcal{C}_{2n+2}^{2n+2}(1 - t) = \mathcal{C}_{2n+2}^{2n+2}(t)$$

Finalement on trouve que $\mathcal{C}_{2n+2}^{2n+2}$ *est strictement croissant dans* $[0, \frac{1}{2}]$ *et strictement décroissant dans* $[\frac{1}{2}, 1]$

Comme :

$$\mathfrak{C}_{2n+2}^{2n+2}(1) = \mathfrak{C}_{2n+2}^{2n+2}(0) = 0 \qquad (3.3.3)$$

la fonction $\mathfrak{C}_{2n+2}^{2n+2}$ *est positive.*

Autre évaluation de l'erreur :

Les noyaux d'erreur étant tous positifs, on a :

$$T_n^n F - \int_0^1 F(X)\, dX = \int_0^1 \mathfrak{C}_{2n}^{2n}(t)\, F^{(2n)}(t)\, dt$$
$$= (F^{2n}(\xi)) \times \int_0^1 \mathfrak{C}_{2n}^{2n}(t)\, dt$$

avec $\xi \in [0,1]$

$$\int_0^1 \mathfrak{C}_{2n}^{2n}(t)\, dt = -\frac{h^{2n}}{2n!} \sum_{k=1}^{n} B_k^n \left(\frac{1}{2^{k-1}}\right)^{2n} \int_0^1 \left[\overline{B}_{2n} \left(\frac{t \times 2^{k-1}}{h}\right) - B_{2n} \right] dt$$

Du fait que l'on a :

$$\int_0^1 \left(\overline{B}_{2n} \left(\frac{t \times 2^{k-1}}{h}\right) - \overline{B}_{2n} \right) dt = - B_{2n}$$

on obtient :

$$\int_0^1 \mathfrak{C}_{2n}^{2n}(t)\, dt = \frac{h^{2n}\, B_{2n}}{2n!} \left(\sum_{k=1}^{n} B_k^n \left(\frac{1}{2^{k-1}}\right)^{2n} \right)$$

$$= \frac{h^{2n}\, B_{2n}}{2n!} \prod_{k=1}^{n} \left(\frac{1}{2^{2k-2}}\right)$$

D'où finalement l'expression suivante pour l'erreur :

$$T_n^n F - \int_0^1 F(X)\, dX = \frac{h^{2n}\, B_{2n}}{2n!} \frac{F^{2n}(\xi)}{2^{n(n-1)}} \qquad \text{avec } \xi \in [0,1] \qquad (3.3.4)$$

n	1	2	3	4	5	6
$\frac{\lvert B_{2n}\rvert}{2^{n(n-1)} \times (2n)!}$	$0,833\,333 \times 10^{-1}$	$0,347\,222 \times 10^{-3}$	$0,516\,699 \times 10^{-6}$	$0,201\,836 \times 10^{-9}$	$0,199\,096 \times 10^{-13}$	$0,492\,127 \times 10^{-18}$

§ 3.4 - SIGNE DES COEFFICIENTS

La formule T_n^n est de la forme :

$$T_n^n F = \sum_{k=1}^{n} B_k^n (T_k^1 F) = \sum_i \theta_i F(y_i)$$

Les B_k^n peuvent être négatifs mais $\sum_{k=1}^{n} |B_k^n|$ reste borné par un nombre voisin de 2.

On a :

$$\sum_i |\theta_i| \leqslant \sum_{k=1}^{n} |B_k^n|$$

car plusieurs formules de trapèzes avec pas différents peuvent utiliser une même abcisse y_k de sorte que des coefficients B_k^n de signe contraire peuvent se compenser.

Dans le cas présent, cette compensation est tellement efficace que *tous les* θ_i *sont positifs.* (Ceci n'est plus vrai si les pas successifs des formules de trapèzes utilisées ne sont pas dans le rapport $\frac{1}{2}$). En effet :

§ 3.4

Appelons E_m l'ensemble des points de [0,1] utilisés par la formule des trapèzes T_m^1 de pas $\frac{1}{2^{m-1}}$.

Si $P \in E_m$, il peut aussi éventuellement appartenir à E_{m-1}, E_{m-2}, mais si $P \notin E_k$ alors $P \notin E_i$ pour $i \leqslant k$.

Les E_m sont strictement emboités.

Considérons la formule T_m^m et une abcisse P quelconque qu'elle utilise. Supposons que $P \in E_m$, E_{m-1} , ..., E_{m-r} (r quelconque $\leqslant$ m - 1).

Le poids affecté à P est alors :

$$\frac{B_m^m}{2^{m-1}} + \frac{B_{m-1}^m}{2^{m-2}} + \cdots\cdots + \frac{B_{m-r}^m}{2^{m-r-1}}$$

(Si P se trouve en 0 ou 1 il faut encore multiplier cette quantité par $\frac{1}{2}$).

Montrons que ce poids est positif quels que soient m et r. $\frac{B_m^m}{2^{m-1}}$ est positif ; le deuxième terme $\frac{B_{m-1}^m}{2^{m-2}}$ est négatif ; le 3ème à nouveau positif, et ainsi de suite. Il suffit donc de montrer que ces termes sont toujours décroissants en valeur absolue :

$$\left|\frac{B_{m-j+1}^m}{2^{m-j}}\right| \geqslant \left|\frac{B_{m-j}^m}{2^{m-j-1}}\right|$$

soit :

$$|B_{k+1}^m| \geqslant 2\,|B_k^m|$$

$$\prod_{j=1}^{k} \frac{1}{1 - \left(\frac{1}{4}\right)^{k+1-j}} \prod_{j=k+2}^{m} \frac{1}{4^{j-k-1} - 1} \geqslant 2 \prod_{j=1}^{k-1} \frac{1}{1 - \left(\frac{1}{4}\right)^{k-j}} \times \prod_{j=k+1}^{m} \frac{1}{4^{j-k} - 1}$$

ce qui revient à :

$$\frac{1}{1 - \left(\frac{1}{4}\right)^{k}} \geqslant 2 \times \frac{1}{4^{m-k} - 1}$$

pour tout m et k (k + 1 $\leqslant$ m)

Or ceci est bien vérifié puisque :

$$\frac{2}{4^{m-k} - 1} \leqslant \frac{2}{3} \leqslant 1 \leqslant \frac{1}{1 - \left(\frac{1}{4}\right)^{k}}$$

On a donc bien démontré que les coefficients effectifs θ_i du procédé de quadrature approchée T_m^m étaient tous positifs. On rappelle que $\sum_i \theta_i = 1$.

Supposons que l'on commette sur chaque calcul de F une erreur inférieure à ε en valeur absolue (due aux arrondis successifs) si l'on employait la formule directe $\Sigma\, \theta_i\ F(y_i)$, l'erreur d'arrondi serait inférieure à ε également.

Si on utilise la formule plus commode $\sum_{k=1}^{n} B_k^n\ T_k^1\ F$ (en prenant les formules (1.4.3)) l'erreur d'arrondi sera inférieure à $\left(\sum_{k=1}^{n} |B_k^n|\right)\ \varepsilon$, soit inférieure à 2ε. Il en est de même si l'on utilise comme au paragraphe suivant la progression des calculs (3.2.2) suivant le tableau triangulaire (3.2.3).

§ 3.5 - PROCEDURE ALGOL POUR L'INTEGRATION SIMPLE ET EXEMPLES NUMERIQUES

La procédure ci-dessous utilise la méthode de Romberg (extrapolation sur la méthode des trapèzes avec des pas successifs dans le rapport 1/2). Elle emploie le schéma de calcul (3.2.2) et (3.2.3) et réduit le plus possible le nombre d'évaluations de la fonction. Elle s'arrête automatiquement quand deux résultats successifs sont suffisamment proches l'un de l'autre.

```
réel procédure INSIRO (F, A, B, ORDMAX, PREC, SORT, RES) ;
valeur A, B, ORDMAX, PREC ; réel procédure F ; Booleen SORT ;
réel A, B, PREC ; entier ORDMAX ; tableau RES ;
commentaire : Cette procédure calcule l'intégrale de F sur l'intervalle [A, B]. On trouve une suite
d'évaluations par des formules exactes pour des polynômes de degré de plus en plus élevé. Si l'écart
relatif entre deux résultats successifs est inférieur à PREC, on arrête le calcul en donnant à SORT
la valeur vrai. Si l'on va jusqu'au degré maximum 2 × ORDMAX on arrête le calcul en donnant à
SORT la valeur faux. Le temps de calcul maximum correspondant à 2 ↑ ORDMAX évaluations de F.
Le tableau RES porte plusieurs évaluations de l'intégrale, la plus précise étant en principe RES [1]
qui est affecté à INSIRO. Si PE est le paramètre effectif qui remplace ORDMAX, le tableau effectif qui remplace RES doit avoir comme bornes d'indices [1 : PE + 1] ;
début réel L, T, P, MA ; entier N, J, I, FAC ;
N := 1 ; L := B - A ; SORT := faux ; MA := RES [1] := L × 0.5 × (F(A) + F(B)) ;
pour J := 1 pas 1 jusqu'à ORDMAX faire
début T := 0 ; P := L/N ;
    pour I := 1 pas 1 jusqu'à N faire T := T + F(A + P × (I - 0.5)) ;
    RES [J + 1] := (P × T + RES [J])/2.0 ; FAC := 1 ;
    pour I := J pas - 1 jusqu'à 1 faire
    début FAC := 4 × FAC ;
          RES [I] := RES [I + 1] + (RES [I + 1] - RES [I])/(FAC - 1)
    fin ;
    si ABS ((RES [1] - MA)/RES [1]) ⩽ PREC alors allera TERME ;
    MA := RES [1] ; N := 2 × N
fin ;
allera AFFECT ;
TERME : SORT := vrai ;
AFFECT : INSIRO := RES [1]
fin
```

Remarque : On pourrait aussi employer l'une des 3 procédures d'extrapolation du § 1.5. avec ALPHA = 2 et T = 2 appliquée à la formule des trapèzes de pas H. Pour calculer $\int_0^1 e^{-x^2}\,dx$ on emploierait par exemple le programme suivant :

```
début réel procédure F(X) ; valeur X ; réel X ; F := EXP ( - X ↑ 2) ;
    réel procédure TRAPEZE (H) ; valeur H ; réel H ;
    début entier N, J ; réel U ; N := 1.0/H ; U := 0 ;
          pour J := 1 pas 1 jusqu'à N - 1 faire U := U + F(J/N) ;
          U := (U + (F(0.0) + F(1.0))/2.0)/N ; TRAPEZE := U
    fin ;
          tableau RE, SE [0 : 5] ;
    déclaration de la procédure RINEVILLE, § 1.5 ;
    RINEVILLE (TRAPEZE, 2.0, 5, 10-4, 1.0, 2, RE, SE)
fin
```

Le résultat le plus précis sera en principe porté par la composante du tableau SE d'indice le plus élevé qui ait reçu une valeur.

Exemples numériques :

Exemple 1 : Considérons $I = \int_0^1 \frac{1}{x + 0{,}01}\,dx = 4{,}615\,147$ avec un pas de départ de $h = \frac{1}{3}$. (Noter que l'intégrant varie entre 1 et 100). Nous donnons le tableau complet (3.2.3) jusqu'à T_8^8. (page suivante).

Exemple 2 : $\int_0^1 e^{4x} \sin(2x\pi)\,dx = -6{,}070\,236$ avec un pas de départ h = 1.

§ 3.6

Exemple n° 1 :

18,295 168	10,615 406	7,056 412	5,510 689	4,905 156	4,698 465	4,637 174	4,620 734
	8,055 486	5,870 081	4,995 449	4,703 312	4,629 567	4,616 744	4,615 255
		5,724 387	4,937 140	4,683 837	4,624 651	4,615 889	4,615 155
			4,924 644	4,679 816	4,623 711	4,615 750	4,615 144
				4,678 856	4,623 491	4,615 719	4,615 142
					4,623 438	4,615 711	4,615 141
						4,615 709	4,615 141
							4,615 141

Exemple n° 2 :

$T_1^1 = 0$

$T_2^2 = 0$

$T_3^3 = -\ 6,175\ 024\ 04$

$T_4^4 = -\ 6,079\ 999\ 80$

$T_5^5 = -\ 6,070\ 180\ 88$

$T_6^6 = -\ 6,070\ 236\ 28$

§ 3.6 - AUTRES PROCEDES ANALOGUES

α) Extrapolation sur la formule des rectangles symétriques, qui admet un développement analogue en puissance de h :

$$R_m = \frac{1}{m} \sum_{j=0}^{m-1} F\left(\frac{j + 0,5}{m}\right) \qquad \left(h = \frac{1}{m}\right)$$

$$R_m = \int_0^1 F(X)\ dX - \frac{h^2}{2!} \int_0^1 \left[\bar{B}_2\left(\frac{t}{h} - \frac{1}{2}\right) - \bar{B}_2\left(\frac{1}{2}\right) \right] F''(t)\ dt$$

et en intégrant par parties :

$$R_m = \int_0^1 F(X)\ dX + \sum_{p=1}^{n-1} \frac{h^{2p}\ \bar{B}_{2p}\left(\frac{1}{2}\right)}{2p!}\ [F^{(2p-1)}(1) - F^{(2p-1)}(0)]\ - \frac{h^{2n}}{2n!} \int_0^1 \left(\bar{B}_{2n}\left(\frac{t}{h} - \frac{1}{2}\right) - \bar{B}_{2n}\left(\frac{1}{2}\right)\right) F^{(2n)}(t)\ dt$$

On peut donc encore appliquer l'extrapolation M_n :

$$M_n(R_m) = \sum_{k=1}^{n} B_k^n\ R_{m_k}$$

Dans le cas ou $m_k = \frac{1}{h} \times 2^{k-1}$ on peut obtenir des résultats comparables à ceux obtenus pour T_n^n.

En particulier :

a) R_m^n obéit à la formule de récurrence 3.2.2.

b) $\mathcal{R}_{2n}^{2n}$, noyau de R_n^n, s'obtient par les formules (3.3.2) en prenant soin de changer le noyau de départ :

$$\frac{-h^2}{2} \left[\bar{B}_2\left(\frac{t}{h} - \frac{1}{2}\right) - \bar{B}_2\left(\frac{1}{2}\right) \right]$$

c) $\mathcal{R}_{2n}^{2n}$ est une fonction positive, croissante de 0 à $\frac{1}{2}$, symétrique par rapport à la droite $t = \frac{1}{2}$.

d) $$R_n^n F - \int_0^1 F(X)\, dX = \frac{h^{2n}\, \bar{B}_{2n}\left(\frac{1}{2}\right)}{(2n)!}\, \frac{F^{2n}(\xi)}{2^{n(n-1)}}$$

La formule R_n^n est de la forme :

$$R_n^n F = \sum_{k=1}^{n} B_k^n R_k^1 = \sum_i \eta_i F(z_i)$$

Appelons F_m l'ensemble des points de [0,1] utilisés par la formule :

$$R_m^1 = R_{\frac{2^m-1}{h}}$$

Contrairement au cas de T_n^n (§ 3.4), les ensembles F_m sont tous 2 à 2 disjoints :

En effet, supposons que pour m > n il existe deux entiers i et j ($0 \leqslant j \leqslant 2^n - 1$; $0 \leqslant i \leqslant 2^m - 1$) tels que :

$$\frac{j + 0,5}{2^n} = \frac{i + 0,5}{2^m}$$

soit $2^{m-n}(j + 0,5) = (i + 0,5)$,

on aboutit à une contradiction : le membre de gauche est un entier alors que celui de droite ne peut pas l'être. Puisqu'il y a des B_k^n négatif, il y aura des η_i négatifs ; on a :

$$\sum_{k=1}^{n} |B_k^n| = \sum_i |\eta_i| < ?$$ (voir § 2.4)

On peut, ici aussi, employer des pas successifs h_k en progression géométrique de raison différente de 0,5.

Exemple : $h_k = \frac{h}{m_k}$; m_k = partie entière $(1,5 \times m_{k-1}) + 1$ (voir § 2.4) ; on adopte alors les formules (1.4.3) pour faire progresser les calculs. La procédure du § 1.5 est directement applicable.

β) Extrapolation appliquée à des formules d'ordre plus élevé :

On peut partir d'une formule dont le degré de validité est plus élevé que celui de la formule des trapèzes. Soit par exemple la formule de Gauss à 2 abcisses exactes pour les polynômes de degré 3 :

$$\frac{1}{2}[F(x_1) + F(x_2)]$$

avec

$$x_1 = \frac{1}{2}\left(1 - \frac{\sqrt{3}}{3}\right) \qquad x_2 = \frac{1}{2}\left(1 + \frac{\sqrt{3}}{3}\right)$$

Appelons G_m^2 une telle formule appliquée avec un pas $\frac{h}{2^{m-2}}$;

La formule :

$$G_{m+1}^3 = \frac{16}{15} G_{m+1}^2 - \frac{1}{15} G_m^2$$

est alors exacte pour des polynômes de degré 5. On peut progresser en triangle, pour trouver G_n^n, en utilisant les formules (3.2.2) (3.2.3) (remplacer la lettre T par G). La première ligne du schéma triangulaire disparaît.

De même plus généralement on peut partir d'une formule de Gauss (ou d'un autre type d'ailleurs) exacte pour des polynômes de degré 2q - 1 et symétrique par rapport à $x = \frac{1}{2}$.

Appelons G_m^q cette formule appliquée avec un pas $\frac{h}{2^{m-q}}$: les q - 1 premières lignes du tableau (3.2.3) disparaissent tandis que les suivantes se construisent de manière inchangée. On montre facilement que $\mathcal{G}_{2n}^{2n}$, le noyau de G_n^n possède les propriétés (3.3.3) démontrées pour $\mathcal{C}_{2n}^{2n}$.

Ce procédé à l'avantage d'allier *la précision et l'économie des formules de Gauss* au caractère *simple et itératif du procédé d'extrapolation* qui augmente de 2 à chaque tour le degré de validité de la formule. En pratique, nous conseillons comme formule de départ une formule de Gauss d'ordre modéré (q = 4 ou 5 par exemple) : il n'est pas souhaitable de prendre d'emblée une formule lourde et on limite ainsi le nombre de constantes.

§ 3.7 - FORMULE DES TRAPEZES A 2 DIMENSIONS

Nous nous limitons pour simplifier l'écriture à des calculs d'intégrales doubles. On peut étendre sans difficulté aux intégrales multiples le procédé décrit ci-dessous.

Considérons le calcul de l'intégrale :

$$\int_0^1 \int_0^1 f(x, y)\, dx\, dy$$

On note T_{x_m} la formule des trapèzes appliquée sur la variable x avec un pas $\frac{1}{m}$ et T_{y_m} l'équivalent pour y.

$$T_{x_m} f = \frac{1}{2m} \sum_{j=0}^{m-1} [f(j/m, y) + f((j+1)/m, y)] \qquad (3.7.1)$$

est encore une fonction de y ; on peut lui appliquer T_{y_m} :

$$T_{y_m} (T_{x_m} f) = \frac{1}{4m^2} \sum_{i=0}^{m-1} \sum_{j=0}^{m-1} [f(j/m, i/m) + f((j+1)/m, i/m) + f(j/m, (i+1)/m) + f((j+1)/m, (i+1)/m)] \qquad (3.7.2)$$

on note :

$$f_{s,t} = \frac{\partial^{s+t} f(x, y)}{\partial^s x\, \partial^t y}$$

$$\Delta_y(f_{s,t}) = f_{s,t}(x, 1) - f_{s,t}(x, 0) \quad \text{et de même} \quad \Delta_x(f_{s,t})$$

$$\Delta_x \Delta_y(f_{s,t}) = f_{s,t}(1,1) - f_{s,t}(1,0) - f_{s,t}(0,1) + f_{s,t}(0,0) \qquad (3.7.3)$$

En appliquant la formule (3.1.2) on obtient : $\left(\text{avec } h = \frac{1}{m}\right)$

$$T_{x_m} f = \int_0^1 f(x, y)\, dx - \frac{h^2}{2!} \int_0^1 \left(\bar{B}_2\left(\frac{x}{h}\right) - B_2\right) f_{2,0}\, dx \qquad (3.7.4)$$

Et en appliquant à nouveau (3.1.2) sur (3.7.4) il vient :

$$T_{y_m} T_{x_m} f = T_{y_m}\left\{\int_0^1 f(x, y)\, dx\right\} - T_{y_m}\left\{\frac{h^2}{2} \int_0^1 \left(\bar{B}_2\left(\frac{x}{h}\right) - B_2\right) f_{2,0}\, dx\right\}$$

$$= \int_0^1 \int_0^1 f(x, y)\, dx\, dy - \frac{h^2}{2} \int_0^1 \left(\bar{B}_2\left(\frac{y}{h}\right) - B_2\right) \left(\int_0^1 f_{0,2}\, dx\right) dy$$

$$- \frac{h^2}{2} \int_0^1 \left(\bar{B}_2\left(\frac{x}{h}\right) - B_2\right) T_{y_m}(f_{2,0})\, dx$$

d'où finalement :

$$T_{y_m}\,T_{x_m}\,f = \iint f - \frac{h^2}{2}\iint \left(\bar{B}_2\left(\frac{y}{h}\right) - B_2\right) f_{0,2}$$
$$- \frac{h^2}{2}\iint \left(\bar{B}_2\left(\frac{x}{h}\right) - B_2\right) f_{2,0} \qquad (3.7.5)$$
$$+ \frac{h^4}{4}\iint \left(\bar{B}_2\left(\frac{y}{h}\right) - B_2\right)\left(\bar{B}_2\left(\frac{x}{h}\right) - B_2\right) f_{2,2}$$

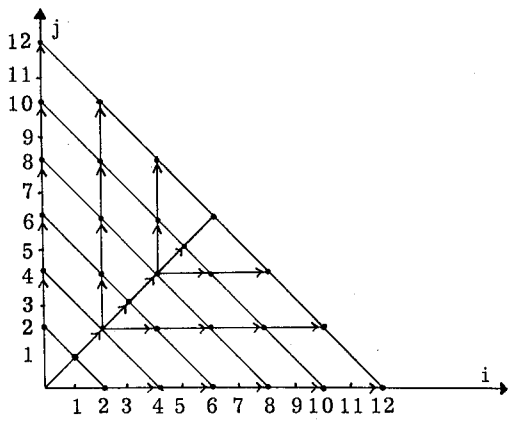
Pour obtenir l'analogue de (3.1.2) on va faire porter les intégrales doubles sur les dérivées $f_{i,j}$ telles que $i + j = 2$: de cette façon le facteur h^2 apparaîtra devant les 3 intégrales doubles :

$$T_{y_m}\,T_{x_m}\,f = \iint f - \frac{h^2}{2}\iint \left(\bar{B}_2\left(\frac{y}{h}\right) - B_2\right) f_{0,2}$$
$$- \frac{h^2}{2}\iint \left(\bar{B}_2\left(\frac{x}{h}\right) - B_2\right) f_{2,0} \qquad (3.7.6)$$
$$+ h^2 \iint \bar{B}_1\left(\frac{x}{h}\right)\bar{B}_1\left(\frac{y}{h}\right) f_{1,1}$$

En intégrant par parties on obtient la formule suivante qui fait intervenir les $f_{i,j}$ avec $i + j = 4$:

$$T_{y_m}\,T_{x_m}\,f = \iint f + \frac{h^2}{2!} B_2 \int_0^1 \Delta_y f_{0,1}\,dx + \frac{h^2}{2!} B_2 \int_0^1 \Delta_x f_{1,0}\,dy$$
$$- \frac{h^4}{4!}\iint \left(\bar{B}_4\left(\frac{y}{h}\right) - B_4\right) f_{0,4}$$
$$- \frac{h^4}{4!}\iint \left(\bar{B}_4\left(\frac{x}{h}\right) - B_4\right) f_{4,0} \qquad (3.7.7)$$
$$+ \frac{h^4}{(2!)^2}\iint \left(\bar{B}_2\left(\frac{y}{h}\right) - B_2\right)\left(\bar{B}_2\left(\frac{x}{h}\right) - B_2\right) f_{2,2}$$

Le graphique ci-dessous montre la génération des termes par intégration par parties : le terme sur la diagonale fournit alternativement 1 terme puis 3 termes :



Ecrivons encore le résultat pour le rang suivant ($i + j = 6$)

$$\begin{aligned}
\underset{y}{T_m}\ \underset{x}{T_m}\ f = \iint f &+ \frac{h^2}{2!} B_2 \int_0^1 \Delta_y\,(f_{0,1})\,dx + \frac{h^2}{2!} B_2 \int_0^1 \Delta_x\,(f_{1,0})\,dy \\
&+ \frac{h^4}{4!} B_4 \int_0^1 \Delta_y\,(f_{0,3})\,dx + \frac{h^4}{4!} B_4 \int_0^1 \Delta_x\,(f_{3,0})\,dy \\
&+ \frac{h^4}{(2!)^2} (B_2)^2\ \Delta_x\,\Delta_y\,(f_{1,1}) \\
&- \frac{h^6}{6!} \iint \left(\overline{B}_6\left(\frac{y}{h}\right) - B_6\right) f_{0,6} - \frac{h^6}{6!} \iint \left(\overline{B}_6\left(\frac{x}{h}\right) - B_6\right) f_{6,0} \\
&- \frac{h^6}{2!}\frac{B_2}{4!} \iint \left(\overline{B}_4\left(\frac{y}{h}\right) - B_4\right) f_{2,4} - \frac{h^6}{2!}\frac{B_2}{4!} \iint \left(\overline{B}_4\left(\frac{x}{h}\right) - B_4\right) f_{4,2} \\
&+ \frac{h^6}{(3!)^2} \iint \overline{B}_3\left(\frac{y}{h}\right) \overline{B}_3\left(\frac{x}{h}\right) f_{3,3}
\end{aligned} \tag{3.7.9}$$

Les intégrales doubles portent sur $f_{0,6}$; $f_{2,4}$; $f_{3,3}$; $f_{4,2}$; $f_{6,0}$. Notons que certaines intégrales peuvent être transformées en intégrales simples de façon évidente. Il y a bien d'autres façons d'exprimer le reste de la formule des trapèzes à deux dimensions ([36], [38], [49]). Celle-ci est parfaitement symétrique en x et y et est particulièrement commode pour le but poursuivi ; elle est de la forme :

$$\underset{y}{T_m}\ \underset{x}{T_m}\ f = Q_4(h) + h^6\ S(h) \tag{3.7.10}$$

où Q_4 est un polynôme pair du 4ème degré et S une fonction bornée.

Pour exprimer facilement la formule générale nous poserons :

$$\begin{aligned}
\Delta f_{s,t} &= \Delta_y\,\Delta_x\,(f_{s,t}) = \Delta_x\,\Delta_y\,(f_{s,t}) \\
f_{-1,t} &= \int \frac{\partial^t f}{\partial y^t}\,dx \qquad \text{(une primitive quelconque)} \\
f_{s,-1} &= \int \frac{\partial^s f}{\partial x^s}\,dy \\
\Delta f_{-1,-1} &= \int_0^1 \int_0^1 f(x, y)\,dx\,dy
\end{aligned} \tag{3.7.11}$$

En supposant n pair, il vient :

$$\begin{aligned}
\underset{y}{T_m}\ \underset{x}{T_m}\ f = &\sum_{p=0}^{n-1} h^{2p} \sum_{\substack{s=0 \\ r+s=p}}^{p} \frac{B_{2r}\ B_{2s}}{(2r)!\,(2s)!}\ \Delta f_{2r-1,2s-1} \\
&+ h^{2n} \left\{ \frac{1}{(n!)^2} \iint \left(\overline{B}_n\left(\frac{y}{h}\right) - B_n\right) \left(\overline{B}_n\left(\frac{x}{h}\right) - B_n\right) f_{n,n} \right. \\
&- \sum_{\substack{i=0 \\ i+j=n}}^{n/2-1} \frac{B_{2i}}{(2i)!\,(2j)!} \iint \left(\overline{B}_{2j}\left(\frac{y}{h}\right) - B_{2j}\right) f_{2i,2j} \\
&\left. - \sum_{\substack{i=n/2+1 \\ i+j=n}}^{n} \frac{B_{2j}}{(2i)!\,(2j)!} \iint \left(\overline{B}_{2i}\left(\frac{x}{h}\right) - B_{2i}\right) f_{2i,2j} \right\}
\end{aligned} \tag{3.7.12}$$

qui est de la forme :

$$\underset{y}{T_m}\ \underset{x}{T_m}\ f = Q_{2n-2}(h) + h^{2n}\ S(h) \tag{3.7.13}$$

où S est une fonction bornée de h.

Quand n est impair, la formule (3.7.12) est légèrement modifiée comme le tableau (3.7.8) le laisse prévoir mais elle reste de la forme (3.7.13).

§ 3.8 - EXTRAPOLATION SUR LA FORMULE DES TRAPEZES A 2 DIMENSIONS ET PROCEDES ANALOGUES

Les formules (3.7.12) et (3.7.13), sous réserve de l'existence des dérivées partielles écrites, montrent que l'on peut appliquer le procédé d'extrapolation M_n sur la fonction $w(h) = T_{y_m}\ T_{x_m}\ f$ avec $h = \frac{1}{m}$ pour trouver $w(0) = \iint f(x,y)\,dx\,dy$.

On choisit $\{m_k\}$ une suite d'entiers rangés par ordre croissant tels que la suite $\left\{\frac{1}{m_k}\right\}$ vérifie la condition (α) du théorème 2, § 2.1.

$$M_n(w) = M_n(T_{y_m}\ T_{x_m}\ f) = \sum_{k=1}^{n} B_k^n\ I_{m_k} \tag{3.8.1}$$

avec $I_m = T_{y_m}\ T_{x_m}\ f$ et B_k^n coefficients d'extrapolation relatifs aux abcisses $\frac{1}{m_k}$ calculés d'après (1.4.3).

Notons que l'on pourrait employer des pas différents pour les variables x et y : $\frac{1}{p_k}$ et $\frac{1}{q_k}$, à condition que $\frac{p_k}{q_k}$ soit constant par rapport à k.

L'extrapolation M_n étant exacte pour un polynôme pair de degré 2n - 2 on a :

$$M_n(I_m) = \iint f + M_n\{h^{2n}\ S(h)\} \tag{3.8.2}$$

Les formules (3.7.9) et (3.8.2) montrent que l'erreur de la formule d'intégration $M_n(I_m)$ peut se mettre sous forme d'une somme d'intégrales doubles portant sur les dérivées partielles $f_{2i,2j}$ avec i + j = n.

Par exemple, pour $i > \frac{n}{2} + 1$ et n pair, en posant j = n - i :

$$-\frac{B_{2j}}{2i!\,2j!} \iint \left[\sum_{k=1}^{n} (B_k^n)\,(h_k)^{2n} \left(\overline{B}_{2i}\left(\frac{x}{h_k}\right) - B_{2i}\right)\right] f_{2i,2j} \tag{3.8.3}$$

Si l'on appelle $F_{2i,2j} = \underset{x,y\in[0,1]}{\text{Max}} |f_{2i,2j}(x,y)|$ on peut écrire des majorations de l'erreur de la forme :

$$|M_n(I_m) - \iint f| \leqslant \sum_{\substack{i=0 \\ i+j=n}}^{n} K_{i,j}\ F_{2i,2j} \tag{3.8.4}$$

On peut prendre par exemple :

$$K_{i,j} = \frac{B_{2i}\ B_{2j}}{(2i)!\,(2j)!} \left(\sum_{k=1}^{n} |B_k^n|\ (h_k)^{2n}\right) \tag{3.8.5}$$

La formule d'intégration $M_n(I_m)$ est exacte pour des monomes de la forme $x^i \times y^j$ avec $i + j \leqslant 2n - 1$. On obtient ainsi de façon simple des formules d'intégration double dont les silhouettes de validité, de forme triangulaire, sont de plus en plus grandes. (3.8.6)

Convergence :

La suite $\left\{\frac{1}{m_k}\right\}$ vérifiant la condition (α) du th. 2, § 2.1, la convergence de I_m vers $\iint f$ quand m tend vers l'infini entraîne la convergence de $M_n(I_m)$ quand n tend vers l'infini. Le procédé d'intégration approché $M_n(I_m)$ convergera donc vers $\iint f$ pour toute fonction f intégrable Riemann.

Par ailleurs, si l'on choisit :

$$m_{k+1} = \text{partie entière } (\alpha \times m_k) + 1 \qquad (3.8.7)$$

et en mettant le procédé sous la forme :

$$M_n(I_m) = \sum_i \theta_i \ f(x_i, y_i)$$

On aura :

$$\sum_i |\theta_i| \leqslant \sum_{k=1}^{n} |B_k^n| \qquad (3.8.8)$$

D'après le § 2.4, $\sum_{k=1}^{n} |B_k^n|$ est borné par un nombre peu différent de $\sum_2(\alpha)$ (voir tableau (2.4.3)). (On aurait exactement $\sum_2(\alpha)$ si les abcisses successives étaient exactement dans le rapport α).

Pour $\alpha = 2$

$$m_{k+1} = 2 \times m_k, \qquad \sum_i |\theta_i| < \sum_{k=1}^{n} |B_k^n| < 2$$

Comme pour l'intégration simple, ce cas présente des propriétés intéressantes. Si I_m^n désigne le procédé provenant d'une extrapolation de la formule des trapèzes sur les abcisses :

$$\frac{h}{2^{m-1}}, \frac{h}{2^{m-2}}, \ldots\ldots, \frac{h}{2^{m-n}}$$

on a la formule suivante, cas particulier de la formule de Neville :

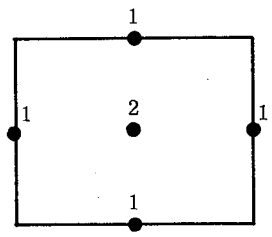
$$I_{m+1}^{n+1} = \frac{2^{2n}}{2^{2n} - 1} I_{m+1}^n - \frac{1}{2^{2n} - 1} I_m^n \qquad (3.8.9)$$

qui permet une progression commode des calculs.

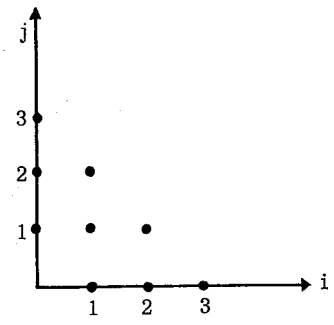
Si l'on reprend dans ce cas l'étude du signe des coefficients θ_i comme au § 3.4, on s'aperçoit que certains coefficients θ_i sont négatifs (contrairement à l'intégration simple). Les ensembles E_k de points des formules de trapèzes successives I_k^1 sont encore emboités. On observe bien une compensation comme au § 3.4 entre les B_i^n positifs et négatifs (de sorte que $\sum_i |\theta_i|$ est strictement inférieur à $\sum_{k=1}^{n} |B_k^n|$) mais cette compensation ne suffit pas pour rendre tous les θ_i positifs. Ceci n'a aucune conséquence numérique puisque $\sum_i |\theta_i| < 2$, ce qui assure une bonne stabilité.

Indiquons enfin explicitement les premières formules obtenues :

Formule I_2^2 (avec h = 1)

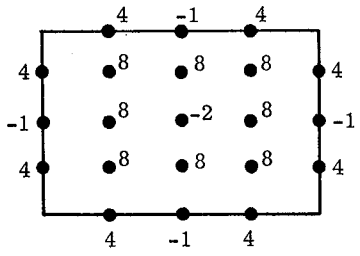


(Dénominateur pour les poids : 6)

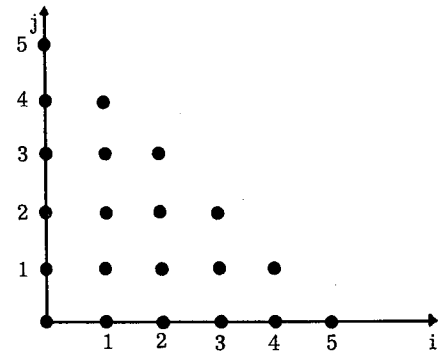


Silhouette de validité

Formule I_3^3 (avec h = 1)



(Dénominateur pour les poids : 90)



Silhouette de validité

Autres procédés analogues :

On peut partir d'une autre formule que celle des trapèzes (comme au § 3.6) : *Par exemple* la formule des rectangles suivantes :

$$R_{y_m}\, R_{x_m} = \frac{1}{m^2} \sum_{j=0}^{m-1} \sum_{i=0}^{m-1} f\left(\frac{i + 0,5}{m}\,,\ \frac{j + 0,5}{m}\right)$$

La formule de récurrence 3.8.9 s'applique sans changement. Plus généralement si $G(F) = \sum_i m_i\, F(x_i)$ est une formule d'intégration simple exacte pour les polynômes de degré 2q - 1 et symétrique par rapport au milieu de l'intervalle, et G_m la même formule appliquée avec un pas $h = \frac{1}{m}$, en appelant

$$\overline{\mathcal{B}}_j(x) = \sum_i m_i\, \overline{B}_j(x - x_i)$$

$$\mathcal{B}_j = \mathcal{B}_j(0)$$

On a : $\mathcal{B}_j = 0$ pour $j = 1, 2, \ldots, 2q - 1$

$\mathcal{B}_0 = 1$

Pour l'intégration double on obtient :

$$G_{y_m}\, G_{x_m}\, f = \iint f + \sum_{p=q}^{n-1} h^{2p} \sum_{\substack{s=0 \\ s=p \\ s=q \text{ à } p-q}}^{r+s=p} \frac{\mathcal{B}_{2r}\, \mathcal{B}_{2s}}{2r!\,2s!}\, \Delta f_{2r-1,2s-1}$$

$$+ h^{2n} \left\{ \frac{1}{(n!)^2} \iint \left(\overline{\mathcal{B}}_n\left(\frac{y}{h}\right) - \mathcal{B}_n\right)\left(\overline{\mathcal{B}}_n\left(\frac{x}{h}\right) - \mathcal{B}_n\right) f_{n,n} \right.$$

$$- \sum_{\substack{i=q \\ i+j=n}}^{n/2-1} \frac{\mathcal{B}_{2i}}{(2i)!\,(2j)!} \iint \left(\overline{\mathcal{B}}_{2j}\left(\frac{y}{h}\right) - \mathcal{B}_{2j}\right) f_{2i,2j}$$

$$\left. - \sum_{\substack{i=n/2+1 \\ i+j=n}}^{n-q} \frac{\mathcal{B}_{2j}}{2i!\,2j!} \iint \left(\overline{\mathcal{B}}_{2i}\left(\frac{x}{h}\right) - \mathcal{B}_{2i}\right) f_{2i,2j} \right\}$$

En appelant G_m^q la formule précédente appliquée avec le pas $h/2^{m-q}$ on peut appliquer la formule de récurrence (3.8.9) avec $n \geqslant q$ (en remplaçant I par G) pour trouver des formules plus puissantes.

§ 3.9 - PROCEDURE ALGOL POUR L'INTEGRATION DOUBLE ET EXEMPLES NUMERIQUES

La procédure ci-dessous calcule $\int_{A_1}^{B_1} dx \int_{A_2}^{B_2} f(x,y)\, dy$ en utilisant la formule de récurrence (3.8.9). Elle donne la possibilité de prendre deux pas de départ HD1 et HD2 différents en x et en y. Elle économise le plus possible les calculs de la fonction f.

```
réel procédure INDOURO (F, A1, B1, A2, B2, N1, N2, ORDMAX, PREC, SORT, RES) ;
valeur A1, B1, A2, B2, N1, N2, ORDMAX, PREC ;
réel A1, B1, A2, B2, PREC ; entier N1, N2, ORDMAX ; Booleen SORT ;
tableau RES ; réel procédure F ;
commentaire: Cette procédure calcule l'intégrale de F(X, Y) sur le rectangle [A1, B1] × [A2, B2].
On choisit des pas de départ (B1 - A1)/N1 et (B2 - A2)/N2 pour X et Y. On trouve une suite d'éva-
luations par des formules dont les silhouettes de validité sont de plus en plus grandes. Les résul-
tats sont portés par le tableau RES, le plus précis étant RES [1] qui est affecté à INDOURO. Si
l'écart relatif entre deux résultats successifs est inférieur à PREC, on arrête le calcul en donnant
à SORT la valeur vrai. Si l'on va jusqu'à la silhouette maximum, on donne à SORT la valeur faux.
Le temps de calcul maximum correspond à N1 × N2 × 4 ↑ ORDMAX évaluations de F. Le tableau
effectif qui remplacera RES doit avoir [1 : ORDMAX + 1] comme bornes d'indices ;
début réel L1, L2, P1, P2, U, MA ;
      entier I, J, K, FAC ;
      L1 := B1 - A1 ; L2 := B2 - A2 ; SORT := faux ; U := 0.0 ;
      P1 := L1/N1 ; P2 := L2/N2 ;
      pour I := 1 pas 1 jusqua N1 - 1 faire
      pour J := 1 pas 1 jusqua N2 - 1 faire
      U := U + F(A1 + P1 × I, A2 + P2 × J) ;
      U := U × 2.0 ;
      pour I := 1 pas 1 jusqua N1 - 1 faire
      U := U + F(A1 + P1 × I, A2) + F(A1 + P1 × I, B2) ;
      pour J := 1 pas 1 jusqua N2 - 1 faire
      U := U + F(A1, A2 + P2 × J) + F(B1, A2 + P2 × J) ;
      U := U × 2.0 ;
      U := U + F(A1, A2) + F(A1, B2) + F(B1, A2) + F(B1, B2) ;
      MA := RES [1] := U × P1 × P2 / 4.0 ;
      pour K := 1 pas 1 jusqua ORDMAX faire
      début U := 0.0 ; P1 := L1/N1 ; P2 := L2/N2 ;
          pour I := 1 pas 1 jusqua N1 faire
          pour J := 1 pas 1 jusqua N2 faire
          U := U + F(A1 + P1 × (I - 0.5), A2 + P2 × (J - 0.5)) ;
          pour I := 1 pas 1 jusqua N1 - 1 faire
          pour J := 1 pas 1 jusqua N2 faire
          U := U + F(A1 + P1 × I, A2 + P2 × (J - 0.5)) ;
          pour I := 1 pas 1 jusqua N1 faire
          pour J := 1 pas 1 jusqua N2 - 1 faire
          U := U + F (A1 + P1 × (I - 0.5) , A2 + P2 × J) ;
          U := U × 2.0 ;
          pour I := 1 pas 1 jusqua N1 faire
          U := U + F(A1 + P1 × (I - 0.5), A2) + F(A2 + P1 × (I - 0.5), B2) ;
          pour J := 1 pas 1 jusqua N2 faire
          U := U + F(A1, A2 + P2 × (J - 0.5)) + F(B1, A2 + P2 × (J - 0.5)) ;
          RES [K + 1] := RES [K]/4.0 + U × P1 × P2/8.0 ;
          FAC := 1 ;
          pour I := K pas - 1 jusqua 1 faire
          début FAC := 4 × FAC ;
                RES [I] := RES [I + 1] + (RES [I + 1] - RES [I])/(FAC - 1)
          fin ;
          si ABS ((RES [1] - MA)/RES [1]) ≤ PREC alors allera TERME ;
          MA := RES [1] ; N1 := 2 × N1 ; N2 := 2 × N2
      fin ;
      allera AFFECT ;
      TERME : SORT := vrai ;
      AFFECT : INDOURO := RES [1]
fin
```

Donnons encore la procédure ALGOL basée sur la formule des rectangles. Nous l'appelons INDOUREC. Le reste de la tête de procédure est inchangé par rapport à la précédente. (Le temps de calcul maximum devient toutefois N1 × N2 × 4 ^ (ORDMAX + 1). Nous n'écrivons que le corps de la procédure INDOUREC :

```
début réel L1, L2, P1, P2, U, MA ; entier I, J, K, FAC ; L1 := B1 - A1 ; L2 := B2 - A2 ;
SORT := faux ; U := 0.0 ; P1 := L1/N1 ; P2 := L2/N2 ;
pour I := 1 pas 1 jusqu'a N1 faire
pour J := 1 pas 1 jusqu'a N2 faire
    U := U + F(A1 + P1 × (I - 0.5), A2 + P2 × (J - 0.5)) ;
RES [1] := MA := U × P1 × P2 ;
pour K := 1 pas 1 jusqu'a ORDMAX faire
début U := 0.0 ; N1 := 2 × N1 ; N2 := 2 × N2 ; P1 := L1/N1 ; P2 := L2/N2 ;
    pour I := 1 pas 1 jusqu'a N1 faire
    pour J := 1 pas 1 jusqu'a N2 faire
    U := U + F(A1 + P1 × (I - 0.5), A2 + P2 × (J - 0.5)) ;
    RES [K + 1] := U × P1 × P2 ; FAC := 1 ;
    pour I := K pas - 1 jusqu'a 1 faire
    début FAC := 4 × FAC ;
        RES [I] := RES [I + 1] + (RES [I + 1] - RES [I])/(FAC - 1)
    fin ;
    si ABS ((RES [1] - MA)/RES [1]) ⩽ PREC alors allera TERME ;
    MA := RES [1]
fin ;
allera AFFECT ;
TERME : SORT := vrai ;
AFFECT : INDOUREC := RES [1]
fin
```

Exemple numérique :

Considérons le calcul de

$$I = \int_{y=0}^{y=2} \int_{x=0}^{x=1} \sin(\pi x)\ \exp(y)\ dx\ dy \ \# \ 4{,}067\ 399\ 44$$

avec un pas de départ de 1 en x comme en y.

	résultats approchés	erreurs
I_6^6 procédure INDOURO (trapèzes)	4,067 380 6	19×10^{-6}
R_5^5 procédure INDOUREC (rectangles)	4,067 391 9	7×10^{-6}

Notons que I_6^6 coûte ici 3 fois plus que R_5^5.

Plus précisément :

$$I_n^n \text{ coûte } (2^{n-1} + 1)^2 \text{ calculs de f}$$

$$R_n^n \text{ coûte } \frac{4^n - 1}{3} \text{ calculs de f}$$

Le coût de I_n^n est donc 3/4 du coût de R_n^n pour n assez grand. (mais on peut espérer avoir une meilleure précision avec R_n^n)

CHAPITRE 4

ÉTUDE DE LA DÉRIVATION APPROCHÉE

Les formules de dérivation étudiées dans ce chapitre permettent l'évaluation de la dérivée d'une fonction dont les valeurs sont calculables en des abcisses quelconques. Ces valeurs ne sont pas entachées d'erreurs de type expérimental mais seulement d'erreurs d'arrondi relativement petites.

§ 4.1 - FORMULES UTILISANT DES POINTS TOUS DU MEME COTE

Soit f une fonction définie sur $[0, \Omega]$. On suppose, sans le dire chaque fois, l'existence des dérivées jusqu'à l'ordre utilisé dans les formules. Nous cherchons des formules approchées utilisant les valeurs de f sur $[0, \Omega]$ pour calculer $f'(0)$ (dérivée à droite éventuellement). Considérons la formule la plus rudimentaire :

$$f'(0) \sim \frac{f(h) - f(0)}{h} = C(h) \tag{4.1.1}$$

L'erreur s'exprime simplement :

$$\begin{aligned} C(h) &= f'(0) + \frac{1}{h}\int_0^h (h - t)\, f^{(2)}(t)\, dt \\ &= f'(0) + \frac{h}{2} f^{(2)}(\xi) \qquad \xi \in [0, h] \end{aligned} \tag{4.1.2}$$

et en intégrant plusieurs fois par parties :

$$C(h) = f'(0) + \frac{h}{2!} f''(0) + \frac{h^2 f^{(3)}(0)}{3!} + \ldots + \frac{h^{n-1} f^{(n)}(0)}{n!} + \frac{1}{h}\int_0^h \frac{(h - t)^n}{n!} f^{(n+1)}(t)\, dt \tag{4.1.3}$$

$$C(h) = P_{n-1}(h) + h^n S(h) \tag{4.1.4}$$

avec S fonction bornée de h.

En effet :

$$\frac{1}{h}\int_0^h \frac{(h - t)^n}{n!} f^{n+1}(t)\, dt = h^n \int_0^1 (1 - T)^n f^{(n+1)}(h T)\, dT = h^n S(h)$$

Appliquons le procédé d'extrapolation L_n (voir § 1.2) à la fonction C pour trouver $C(0) = f'(0)$. Soit $\{h_k\}$ une suite d'abcisses décroissantes tendant vers 0. Le procédé L_n utilise les n premières abcisses h_k.

$$\begin{aligned} L_n(C) &= f'(0) + \sum_{k=1}^{n} A_k^n (h_k)^n \int_0^1 (1 - T)^n f^{(n+1)}(h_k T)\, dT \\ &= f'(0) + \frac{1}{n!}\int_0^\Omega \left[\sum_{k=1}^{n} A_k^n\, \varepsilon_k(t) \frac{(h_k - t)^n}{h_k} \right] f^{(n+1)}(t)\, dt \end{aligned} \tag{4.1.5}$$

($\varepsilon_k(t) = 1$ entre 0 et h_k et 0 ailleurs).

L'erreur de dérivation de la formule $L_n(C)$ est donc mise sous forme intégrale portant sur la dérivée n + 1ème de f. La formule $L_n(C)$ utilise n + 1 points : h_1, h_2, ..., h_n et zéro. Elle est exacte quand f est un polynôme de degré n. $L_n(C)$ est donc simplement la formule classique qui consiste, pour trouver une valeur approchée de f' en un point h_o, à faire coïncider un polynôme de degré n avec f en n + 1 points (dont h_o) et à prendre la dérivée de ce polynôme en h_o. Il faut seulement remarquer que tous les points utilisés sont d'un même côté de h_o.

Progression des calculs

Appelons comme au § 1.2 b) C_m^n la formule de dérivation approchée qui utilise

$$h_m,\ h_{m-1},\ \ldots,\ h_{m-n+1} \text{ et } 0 \qquad (m \geqslant n)$$

$$C_m^1 = C(h_m) = \frac{f(h_m) - f(0)}{h_m} \tag{4.1.6}$$

On a en fait une extrapolation sur la fonction C. On a donc la formule suivante qui permet la progression des calculs :

$$C_{m+1}^{n+1} = \frac{h_{m-n+1} \times C_{m+1}^n - h_{m+1} \times C_m^n}{h_{m-n+1} - h_{m+1}} \tag{4.1.7}$$

Cette relation simple ne semble pas être directement généralisable à l'évaluation des dérivées d'ordre supérieur à 1 ou même à l'évaluation de dérivée première en une abcisse quelconque : elle est due au fait que l'on évalue la dérivée en un point (zéro ici) qui fait partie des abcisses utilisées. Dérivons l'identité d'Aitken (1.2.5) après avoir ajouté le point zéro :

$$P'_{n+1}(x_1,\ x_2,\ \ldots,\ x_n,\ 0,x) = \frac{P'_n(x_1,\ \ldots,\ x_{n-1},\ 0,x) \times (x - x_n) - P'_n(x_2,\ \ldots,\ x_n,\ 0,x) \times (x - x_1)}{x_1 - x_n}$$

$$+ \frac{P_n(x_1,\ \ldots,\ x_{n-1},\ 0,x) - P_n(x_2,\ \ldots,\ x_n,\ 0,x)}{x_1 - x_n}$$

La deuxième partie du terme de droite ne disparaît que lorsque x = 0 et l'on trouve la formule (4.1.7). Pour une abcisse quelconque, il faut construire deux tableaux triangulaires :

Si E_m^n désigne la valeur en x = a du polynôme basé sur h_m, h_{m-1}, ..., h_{m-n+1}

et F_m^n désigne la dérivée en x = a de ce polynôme, on a les 2 relations :

$$\begin{cases} F_{m+1}^{n+1} = \dfrac{F_m^n \times (a - h_{m+1}) - F_{m+1}^n \times (a - h_{m-n+1})}{h_{m-n+1} - h_{m+1}} + \dfrac{E_m^n - E_{m+1}^n}{h_{m-n+1} - h_{m+1}} \\ E_{m+1}^{n+1} = \dfrac{E_m^n \times (a - h_{m+1}) - E_{m+1}^n\ (a - h_{m-n+1})}{h_{m-n+1} - h_{m+1}} \end{cases} \tag{4.1.8}$$

Pour les dérivées d'ordre plus élevé, il faut manipuler plusieurs tableaux, que le point où l'on évalue la dérivée appartienne ou non à l'ensemble des abcisses utilisées.

Etude du noyau d'erreur :

(voir Kuntzmann [26] page 152)

$$L_n(C) = f'(0) + \int_o^{\Omega} \mathcal{C}_n(t)\ f^{(n+1)}(t)\ dt \tag{4.1.9}$$

Pour étudier $\mathcal{C}_n(t)$, donné par (4.1.5), posons :

$$\Phi_j(t) = \sum_{k=1}^{n} \frac{A_k^n\ \varepsilon_k(t)\ (h_k - t)^j}{j!\ h_k}$$

On a :

$$\Phi_n = \mathcal{C}_n$$

Pour $j \geqslant 2$, les Φ_j sont continues pour tout t : en t = 0 cela provient du fait que :

$$\sum_{k=1}^{n} A_k^n \, h_k^{j-1} = 0 \text{ pour } j = 2, 3, \ldots, n \; ; \; \Phi_j(0^+) = 0$$

Pour j = 1 : On a continuité partout sauf au point t = 0 ; $\Phi_1(0^+) = 1$.

Pour j = 0 : Φ_o est constante par morceaux. Φ_o est discontinue en 0, h_n, h_{n-1}, ..., h_1.

On a la relation : $\Phi'_{j+1}(t) = -\Phi_j(t)$ (j = 0, 1, ..., n - 1) sauf au points de discontinuités de Φ_j mentionnés ci-dessus.

Tous les Φ_j sont nuls en dehors de $[0, h_1]$.

Φ_o change de signe au plus n - 1 fois à l'intérieur de $[0, h_1]$ (sans compter les extrémités h_1 et 0).

Φ_1 change de signe également au plus n - 1 fois à l'intérieur de $[0, h_1]$ (du fait que $\Phi_1(0^+) = 1$).

Φ_2 change de signe n - 2 fois au plus à l'intérieur de $[0, h_1]$ (en effet $\Phi_2(0^+) = \Phi_2(h_1) = 0$) et ainsi de suite...

$$\underline{\mathcal{C}_n = \Phi_n \text{ ne change pas de signe dans } [0, h_1]} \qquad (4.1.10)$$

On a donc :

$$L_n(C) - f'(0) = \int_o^{\Omega} \mathcal{C}_n(t) \, f^{n+1}(t) \, dt$$

$$= f^{n+1}(\xi) \times \int_o^{\Omega} \mathcal{C}_n(t) \, dt \text{ avec } \xi \in [0, h_1]$$

Pour trouver $\int_o^{\Omega} \mathcal{C}_n(t) \, dt$ il suffit d'appliquer la formule $L_n(C)$ à la fonction :

$$f(t) \equiv \frac{t^{n+1}}{(n+1)!}$$

on a :

$$f^{(n+1)}(t) \equiv 1$$

et

$$\int_o^{\Omega} \mathcal{C}_n(t) \, dt = \sum_{k=1}^{n} \frac{A_k^n \, (h_k)^n}{(n+1)!}$$

et en tenant compte de l'expression des A_k^n (§ 2.1) :

$$\int_o^{\Omega} \mathcal{C}_n(t) \, dt = \frac{(-1)^{n+1} \prod_{k=1}^{n} (h_k)}{(n+1)!}$$

On a donc finalement :

$$L_n(C) - f'(0) = \frac{(-1)^{n+1}}{(n+1)!} \left(\prod_{k=1}^{n} (h_k) \right) f^{(n+1)}(\xi) \quad \text{avec } \xi \in [0, h_1] \qquad (4.1.11)$$

Convergence :

Pour toute fonction f dérivable à droite en 0, la fonction C est continue à droite en zéro. Le théorème de convergence du § 2.1 permet donc d'affirmer que le procédé $L_n(C)$ converge vers la solution pour toute fonction dérivable à droite en zéro *si et seulement si* les abcisses h vérifient la condition (α) ; on prendra tout simplement des abcisses en progression géométrique de raison $\beta < 1$:

$$h_{k+1} = \beta \times h_k$$

Abcisses en progression géométrique de raison $\frac{1}{2}$

Appelons C_m^n la formule qui correspond à une extrapolation sur la fonction C basée sur les pas $\frac{h}{2^{m-1}}, \ldots, \frac{h}{2^{m-n}}$ et qui est exacte quand f est un polynome de degré n. C_{m+1}^n et C_m^n sont exactes pour des polynômes de degré n : a C_{m+1}^n + b C_m^n utilise les mêmes abcisses que C_{m+1}^{n+1}, et est exacte pour un polynôme de degré n si a + b = 1. En choisissant correctement a et b on peut la rendre exacte pour le degré n + 1 donc identique à C_{m+1}^{n+1}. On a la formule suivante, cas particulier de (4.1.7) :

$$C_{m+1}^{n+1} = \frac{2^n}{2^n - 1} C_{m+1}^n - \frac{1}{2^n - 1} C_m^n$$
$$C_m^1 = C\left(\frac{h}{2^{m-1}}\right) \qquad (4.1.12)$$

Appelons $\mathcal{C}_m^n$ le noyau de la formule C_m^n.

On a la relation :

$$\mathcal{C}_{m+1}^n(t) = \frac{\mathcal{C}_m^n(2t)}{2^{n-1}} \quad \text{pour } t \in \left[0, \frac{h}{2^{m+1-n}}\right] \text{ et 0 ailleurs}$$

Donc :

$$C_{m+1}^{n+1} f - f'(0) = \int_0^{\frac{h}{2^{m-n}}} \mathcal{C}_m^{n*}(t)\, f^{(n+1)}(t)\, dt \qquad (4.1.13)$$

avec

$$\mathcal{C}_m^{n*}(t) = \begin{cases} \frac{1}{2^n - 1}\left[2\,\mathcal{C}_m^n(2t) - \mathcal{C}_m^n(t)\right] & \text{si } t \in \left[0, \frac{h}{2^{m+1-n}}\right] \\ \frac{-1}{2^n - 1}\,\mathcal{C}_m^n(t) & \text{si } t \in \left[\frac{h}{2^{m+1-n}}, \frac{h}{2^{m-n}}\right] \end{cases} \qquad (4.1.14)$$

Posons :

$$\mathcal{C}_{m+1}^{n+1}(t) = -\int_0^t \mathcal{C}_m^{n*}(t)\, dt$$

En intégrant par parties et en tenant compte du fait que :

$$\mathcal{C}_{m+1}^{n+1}(0) = \mathcal{C}_{m+1}^{n+1}\left(\frac{h}{2^{m-n}}\right)$$

on a :

$$C_{m+1}^{n+1} f - f'(0) = \int_0^{\frac{h}{2^{m-n}}} \mathcal{C}_{m+1}^{n+1}(t)\, f^{(n+2)}(t)\, dt$$

Le noyau $\mathcal{C}_{m+1}^{n+1}$ de C_{m+1}^{n+1} s'obtient donc par la formule de passage :

$$\mathcal{C}_{m+1}^{n+1}(t) = \begin{cases} \frac{-1}{2^n - 1}\int_t^{2t} \mathcal{C}_m^n(x)\, dx \text{ pour } t \in \left[0, \frac{h}{2^{m-n+1}}\right] \\ \frac{-1}{2^n - 1}\int_t^{\frac{h}{2^{m-n}}} \mathcal{C}_m^n(x)\, dx \text{ pour } t \in \left[\frac{h}{2^{m-n+1}}, \frac{h}{2^{m-n}}\right] \end{cases}$$

§ 4.2 - FORMULES UTILISANT DES POINTS SYMETRIQUES

Soit f une fonction définie sur $[-\Omega, +\Omega]$.

Appelons D(h) la formule de dérivation approchée :

$$D(h) = \frac{f(h) - f(-h)}{2h} \tag{4.2.1}$$

On a l'expression suivante pour l'erreur :

$$D(h) = f'(0) + \int_{-h}^{0} \frac{(h+t)^2}{4h} f^{(3)}(t)\, dt + \int_{0}^{h} \frac{(h-t)^2}{4h} f^{(3)}(t)\, dt \tag{4.2.2}$$

et en intégrant plusieurs fois par parties :

$$D(h) = f'(0) + \frac{h^2}{3!} f^{(3)}(0) + \frac{h^4}{5!} f^{(5)}(0) + \ldots + \frac{h^{2n-2}}{(2n-1)!} f^{(2n-1)}(0) + \frac{1}{2h}\left[\int_{-h}^{0} \frac{(h+t)^{2n}}{(2n)!} f^{(2n-1)}(t)\, dt + \int_{0}^{h} \frac{(h-t)^{2n}}{(2n)!} f^{(2n+1)}(t)\, dt\right] \tag{4.2.3}$$

$$D(h) = Q_{2n-2}(h) + h^{2n} S(h) \tag{4.2.4}$$

avec S fonction bornée de h. On peut donc appliquer le procédé d'extrapolation M_n (voir § 1.4) sur la fonction D pour obtenir D(0) = f'(0).

$$M_n(D) = f'(0) + \frac{1}{2n!} \int_{-\Omega}^{+\Omega} \left[\sum_{k=1}^{n} B_k^n\, \varepsilon_k(t)\, \frac{(h_k - |t|)^{2n}}{2h_k}\right] f^{(2n+1)}(t)\, dt \tag{4.2.5}$$

où $\varepsilon_k(t)$ vaut 1 entre $-h_k$ et h_k et zéro ailleurs.

La formule $M_n(D)$ est exacte pour les polynômes de degré 2n et utilise en fait 2n abcisses symétriques 2 à 2 par rapport à l'origine. On obtient la même formule en faisant passer un polynôme de degré 2n - 1 par ces 2n points et en prenant sa dérivée en zéro (la validité s'étendra au degré 2n à cause de la symétrie).

Noyau d'erreur :

$$M_n(D) - f'(0) = \int_{-\Omega}^{+\Omega} \varpi_{2n}(t)\, f^{(2n+1)}(t)\, dt \tag{4.2.6}$$

Comme au paragraphe précédent on pose :

$$\psi_j(t) = \sum_{k=1}^{n} \frac{B_k^n\, \varepsilon_k(t)\, (h_k - |t|)^j}{j! \times (2h_k)} \quad ; \quad \text{on a } \psi_{2n} = \varpi_{2n}$$

Les ψ_j sont continues partout pour $j \geqslant 1$.

ψ_o est discontinue aux abcisses $\pm h_k$.

On a $\psi'_{j+1} = -\psi_j$ sauf pour j = 0 aux points $\pm h_k$. Les fonctions ψ_j sont paires.

ψ_o change au plus n - 1 fois de signe à l'intérieur de $[0, h_1]$, (aux points $h_n, h_{n-1}, \ldots, h_2$).

ψ_1 change également au plus n - 1 fois de signe dans $[0, h_1]$, car $\psi_1(0) \neq 0$. De même ψ_2.

Mais ψ_3 change au plus n - 2 fois de signe car $\psi_3(0) = 0$, et de même pour ψ_4. En continuant ainsi, ψ_{2j-1} et ψ_{2j} changent au plus n - j fois de signe. ψ_{2n-1} garde un signe constant et $\psi_{2n} = \varpi_{2n}$ est monotone et de signe constant dans $[0, h_1]$. On a donc pour l'erreur :

$$M_n(D) - f'(0) = \left(\int_{-h_1}^{+h_1} \varpi_{2n}(t)\, dt\right) \times f^{2n+1}(\xi) \text{ avec } \xi \in [-h_1, +h_1]$$

$$= \frac{\sum_{k=1}^{n} B_k^n\, (h_k)^{2n}}{(n+1)!} \times f^{(2n+1)}(\xi) \tag{4.2.7}$$

$$M_n(D) - f'(0) = \frac{(-1)^{n+1} \prod_{k=1}^{n} (h_k)^2}{(n+1)!} \times f^{(2n+1)}(\xi) \qquad (4.2.8)$$

En résumé : Les noyaux ϖ_{2n} sont pairs : dans l'intervalle $[0, h_1]$ ils sont alternativement positifs décroissants (n impair) et négatifs croissants (n pairs). (4.2.9)

Abcisses en progression géométrique de raison $\frac{1}{2}$

Soient D_m^n la formule basée sur les abcisses :

$$\frac{h}{2^{m-1}}, \frac{h}{2^{m-2}}, \ldots\ldots, \frac{h}{2^{m-n}}$$

et leurs symétriques, et ϖ_m^{2n} le noyau d'erreur correspondant. On a les formules de passage :

$$D_{m+1}^{n+1} = \frac{2^{2n}}{2^{2n}-1} D_{m+1}^n - \frac{1}{2^{2n}-1} D_m^n$$

$$\varpi_{m+1}^{2n+1}(t) = \begin{cases} \frac{1}{2^{2n}-1} \int_t^{2t} \varpi_n^{2n}(x)\, dx \text{ pour } t \in \left[\frac{-h}{2^{m+1-n}}, \frac{+h}{2^{m+1-n}} \right] \\ \frac{1}{2^{2n}-1} \int_t^{\frac{h}{2^{m-n}}} \varpi_m^{2n}(x)\, dx \text{ pour } t \in \left[\frac{h}{2^{m+1-n}}, \frac{h}{2^{m-n}} \right] \\ \frac{1}{2^{2n}-1} \int_t^{\frac{-h}{2^{m-n}}} \varpi_m^{2n}(x)\, dx \text{ pour } t \in \left[\frac{-h}{2^{m-n}}, \frac{-h}{2^{m+1-n}} \right] \end{cases} \qquad (4.2.10)$$

$$\varpi_{m+1}^{2n+2}(t) = \int_{\frac{h}{2^{m-n}}}^{t} \varpi_{m+1}^{2n+1}(t)\, dt$$

§ 4.3 - PROGRAMME ALGOL ET EXEMPLES

Nous prendrons comme exemple les formules du § 4.2 avec les abcisses en progression géométrique de raison $\frac{1}{2}$.

```
réel procédure DERIVEE (F, A, H, ORDMAX, PREC, SORT, RES) ;
valeur A, H, ORDMAX, PREC ; réel procédure F ; Booleen SORT ;
réel A, H, PREC ; entier ORDMAX ; tableau RES ;
commentaire : Cette procédure calcule la dérivée de F(X) pour X = A. On calcule F en des abcisses
prises dans l'intervalle [A - H, A + H]. On trouve une suite d'évaluations d'ordre croissant. Si
l'écart relatif entre deux résultats successifs est inférieur à PREC on arrête le calcul en donnant
à SORT la valeur vrai. Si l'on va jusqu'à l'ordre maximum on donne à SORT la valeur faux. Les
résultats sont affectés au tableau RES, le plus précis RES [1] étant aussi affecté à DERIVEE. Le
temps de calcul maximum correspond à 2 × (ORDMAX + 1) calculs de F. Le tableau effectif qui rem-
place RES doit avoir [1 : LE + 1] comme bornes d'indices si LE est la paramètre effectif qui rem-
place ORDMAX ;
début réel MA ; entier J, I, FAC ;
SORT := faux ; MA := RES [1] := (F(A + H) - F(A - H))/2.0/H ;
pour J := 1 pas 1 jusqu'à ORDMAX faire
début RES [J + 1] := (F(A + H/2.0) - F(A - H/2.0))/H ;
    FAC := 1 ;
    pour I := J pas - 1 jusqu'à 1 faire
    début FAC := 4 × FAC ;
          RES [I] := RES [I + 1] + (RES [I + 1] - RES [I])/(FAC - 1)
    fin ;
    si ABS ((RES [1] - MA)/RES [1]) ⩽ PREC alors allera TERME ;
    MA := RES [1] ; H := H/2.0
fin ;
allera AFFECT ;
```

```
. TERME : SORT := vrai ;
AFFECT : DERIVEE := RES [1]
fin
```

Le programme peut s'écrire également en utilisant l'une des procédures définies au § 1.5. Le paramètre ALPHA vaut donc 2 et T est pris égal à 2 puisqu'il s'agit d'une extrapolation pour polynômes pairs. On applique l'extrapolation à la fonction V(H) = (F(H/2) - F(-H/2))/H. On prend H = 1 comme première valeur. On traite la fonction F(X) = 1/(X - 1) dont la dérivée en zéro vaut - 1.

```
début réel procédure F(X) ; valeur X ; réel X ;
{     F := 1.0/(X - 1.0) ;
      réel procédure DER (H) ; valeur H ; réel H ;
      DER := (F(H/2.0) - F(- H/2.0))/H ;
      tableau RE, SE [0 : 10] ;
      procédure RINEVILLE ..... voir déclaration § 1.5 ..... ;
      RINEVILLE (DER, 2.0, 10, 5.0, 10 -5, 1.0, 2, RE, SE)
fin
```

RE [I] sont les valeurs de V(H) pour des pas H successifs
SE [I] sont les valeurs améliorées plus précises. SE [I] est en fait une combinaison linéaire convenable des RE [J] pour J = 1, ..., I.

On trouve les résultats numériques suivants :

H	I	Sans extrapolation RE [I]	Avec extrapolation SE [I]
1	0	- 1,333 333 3	- 1,333 333 3
1/2	1	- 1,066 666 7	- 0,977 777 7
1/4	2	- 1,015 872 9	- 1,000 352 7
1/8	3	- 1,003 921 6	- 0,999 998 62
1/16	4	- 1,000 977 5	- 0,999 999 98

L'erreur sur RE [4] est environ 2×10^{-5} alors que l'erreur sur SE [4] est 2×10^{-8}.

§ 4.4 - QUELQUES EXTENSIONS POSSIBLES

L'application de l'extrapolation à la dérivation approchée étudiée dans les § 4.1 et 4.2 n'a pas conduit à des formules nouvelles. Cela est dû aux formules de départ employées. Il n'en sera plus de même pour les cas traités ci-dessous. Comme pour l'intégration approchée, diverses extensions sont possibles :

α) formules à 1 dimension :

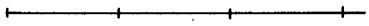
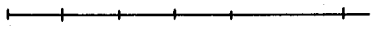
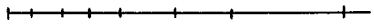
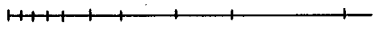
On peut employer la méthode d'extrapolation pour évaluer des dérivées d'ordre plus élevé ou même des expressions différentielles. La formule de départ employé comporte souvent davantage de points que celles utilisées en 4.1 et 4.2. Dans ce cas on n'obtient pas en général des formules classiques : comme pour l'intégration approchée, les formules obtenues ne correspondent pas au degré de validité maximum. Elles s'obtiennent facilement d'une manière itérative.

Exemple : Evaluation de la dérivée seconde de f(x) en zéro. Prenons comme point de départ une formule F à 4 points 0, h, 2h, 3h, valable pour les polynômes de degré 3. On a alors :

$$F(f) - f''(0) = A_2\, h^2 + A_3\, h^3 + A_4\, h^4 + \ldots$$

On employera le procédé L_m^n. La première ligne du schéma triangulaire (1.3.6) est supprimée. Les valeurs de L_m^2 sont remplacées par F(f) calculée avec un pas $\frac{h}{2^{m-2}}$.

Le tableau ci-après donne les nombres de points, degrés de validité et dispositions des premières formules obtenues :

Formule	Nombre de points	Degré de validité	Dispositions
L_m^2	4	3	
L_m^3	6	4	
L_m^4	8	5	
L_m^5	10	6	

β) formules à plusieurs dimensions

L'extrapolation peut-être appliquée au calcul de dérivées partielles ou d'expressions aux dérivées partielles.

Considérons la formule D_{xy} pour le calcul de la dérivée partielle

$$\frac{\partial^2 f}{\partial x\, \partial y}(0,0) = f_{1,1}(0,0)$$

$$D_{x,y}\, f = \frac{f(h,\ h) + f(-h,\ -h) - f(h,\ -h) - f(-h,\ h)}{4h^2}$$

On a alors :

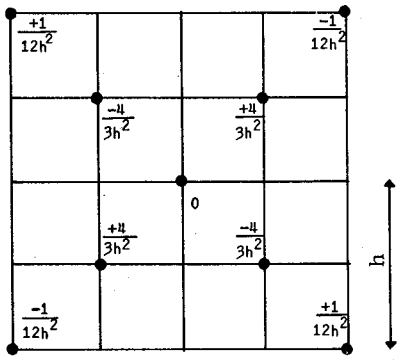
$$D_{x,y}\, f = f_{1,1}(0,0) + h^2 A_2 + h^4 A_4 + h^6 A_6 + \ldots$$

A_2 contient $f_{4,0}(0,0)$; $f_{2,2}(0,0)$; $f_{0,4}(0,0)$
A_4 contient $f_{6,0}(0,0)$; $f_{4,2}(0,0)$; $f_{2,4}(0,0)$; $f_{0,6}(0,0)$.

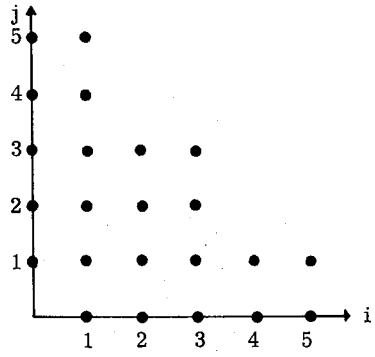
etc.

Appliquons le procédé M_m^n ; Donnons explicitement la première formule obtenue :

$M_2^2\,(D_{x,y}\, f)$:



Silhouette de validité



Exemple :

$$f(x,\ y) = \frac{1}{1 + e^{xy}} \quad ; \quad \frac{\partial^2 f}{\partial x\, \partial y}(0,0) = -0{,}25 \quad ; \quad h_1 = 0{,}5 \quad ; \quad h_2 = 0{,}25 \quad ; \quad h_3 = 0{,}125$$

$D_{x,y}\, f$ pour $h = 0{,}125$	$M_3^3(D_{x,y}\, f)$	erreur sans extrapolation	erreur avec extrapolation
- 0,249 994 75	- 0,249 999 97	525×10^{-8}	3×10^{-8}

CHAPITRE 5

APPLICATION A L'INTÉGRATION DES ÉQUATIONS DIFFÉRENTIELLES

Nous allons étudier dans ce chapitre l'application du procédé d'extrapolation du chapitre 1 à l'intégration approchée des équations différentielles.

Pour les problèmes de conditions initiales il existe déjà de nombreuses méthodes d'intégration dont l'erreur de discrétisation est petite. Par contre il est toujours difficile pour les problèmes de conditions aux limites de discrétiser avec une erreur infinitésimale d'ordre élevé : le procédé de Richardson permet dans certains cas d'augmenter cet ordre et d'obtenir une précision difficile à atteindre par les autres méthodes ; ce phénomène sera encore plus net pour les équations aux dérivées partielles de type elliptique et les équations intégrales de Fredholm traitées au chapitre suivant. Si V(h) désigne la solution approchée en un point fixe obtenue avec un pas h, l'efficacité du procédé de Richardson est liée à l'existence de constantes C_i (i = 0, 1, ..., r - 1) et d'une fonction B bornée de $h \in [0, h_o]$ telle que l'on ait :

$$V(h) = \sum_{i=0}^{r-1} C_i h^i + h^r B(h) \text{ pour tout } h \in [0, h_o]$$

$$(|B(h)| \leqslant K)$$

On appliquera le procédé M_n (§ 1.4) quand les C_i sont nuls pour i impair.

Nous allons examiner l'existence d'un tel développement dans quelques cas simples.

§ 5.1 - ETUDE DU COMPORTEMENT DE LA SOLUTION APPROCHEE EN FONCTION DU PAS h POUR UN PROBLEME DE CONDITIONS INITIALES

Considérons l'équation différentielle $y' = f(y, x)$ avec la condition initiale $y(a) = \alpha$. y(x) désignera la solution exacte de cette équation. Nous employons la méthode d'intégration approchée d'Euler (méthode de la tangente) avec un pas h. La formule de passage de l'abcisse $x_n = a + nh$ à l'abcisse $x_{n+1} = a + (n + 1)h$ est la suivante :

$$y_{n+1} = y_n + hf(y_n, x_n) \tag{5.1.1}$$

on prend $y_o = \alpha$. On intègre dans l'intervalle [a, b]. D désigne le domaine en x, y suivant :

$$a \leqslant x \leqslant b, \quad -\infty < y < +\infty$$

On note l'erreur d'intégration

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

pour l'abcisse x_n. Dans toute la suite on suppose $0 < h \leqslant (b - a)$. Les méthodes de démonstration des théorèmes suivants sont inspirées de P. Henrici [18].

Théorème 1 : Si f(y, x) est continue sur D et possède des dérivées partielles $\frac{\partial f}{\partial y}$ et $\frac{\partial f}{\partial x}$ également continues sur D, si en plus $|\frac{\partial f}{\partial y}|$ est bornée par une constante F_1 sur D, alors l'erreur e_n de la méthode d'Euler à l'abcisse $x_n \in [a, b]$ vérifie :

$$|e_n| < h \times K$$

où K est une constante indépendante de h et n.

Démonstration : Avec ces hypothèses, l'équation différentielle a une seule solution exacte y(x) qui possède une dérivée seconde continue sur [a, b].

Appelons :

$$N(x) = \frac{1}{2} \underset{t\in[a,x]}{\text{Max}} |y''(t)|$$

Par définition on a :

$$y_{m+1} = y_m + hf(y_m, x_m)$$

Et pour la solution exacte, en utilisant le développement de Taylor :

$$y(x_{m+1}) = y(x_m) + hf(y(x_m), x_m) + \frac{h^2}{2} y''(\xi)$$

avec $\xi \in [x_m, x_{m+1}]$.

En soustrayant :

$$e_{m+1} = e_m + h\,[f(y(x_m) + e_m, x_m) - f(y(x_m), x_m)] - \frac{h^2}{2} y''(\xi)$$

On obtient donc :

$$|e_{m+1}| \leqslant |e_m| + h\,F_1\,|e_m| + h^2\,N(x_{m+1})$$

$$m = 0, 1, \ldots, n-1 \quad ; \quad e_o = 0$$

soit encore :

$$|e_{m+1}| \leqslant e_m \times A + B \quad \text{avec} \quad A = 1 + h\,F_1$$
$$B = h^2\,N(x_n)$$

On en déduit :

$$|e_n| \leqslant B\left(\frac{A^n - 1}{A - 1}\right) = h\,N(x_n)\left[\frac{(1 + h\,F_1)^n - 1}{F_1}\right]$$

On voit facilement que :

$$\left(1 + \frac{(x_n - a)\,F_1}{n}\right)^n \leqslant e^{(x_n - a)F_1} \quad (x_n \in [a, b])$$

d'où :

$$|e_n| \leqslant h\,N(x_n) \times \left(\frac{e^{F_1(x_n - a)} - 1}{F_1}\right) \tag{5.1.2}$$

Il suffit donc de prendre :

$$K = N(b) \times \frac{e^{F_1(b-a)} - 1}{F_1} \tag{5.1.3}$$

Théorème 2 : Nous faisons les mêmes hypothèses qu'au théorème 1 mais nous considérons une formule d'intégration de la forme :

$$y_{m+1} = y_m + h\, f(y_m, x_m) + \theta(m, h) \times h^2\, C$$

où $C \geqslant 0$ est une constante indépendante de m et h et où $\theta(m, h)$ vérifie :

$$|\theta(m, h)| \leqslant 1$$

pour tout m entier et tout h $(0 < h \leqslant b - a)$.

On a alors encore la majoration suivante pour l'erreur e_n :

$$|e_n| \leqslant h \times K^*$$

où K^* est une constante indépendante de n et h.

Démonstration : La démonstration est analogue à celle du théorème 1 ; B est remplacé par $h^2(N(x_n) + C)$; il suffit de prendre :

$$K^* = (N(b) + C)\left(\frac{e^{F_1(b-a)} - 1}{F_1}\right)$$

Théorème 3 : Nous supposerons maintenant que f(y, x) est continue sur D et qu'elle admet des dérivées partielles $\frac{\partial f}{\partial y}, \frac{\partial f}{\partial x}, \frac{\partial^2 f}{\partial y^2}, \frac{\partial^2 f}{\partial x^2}, \frac{\partial^2 f}{\partial x\, \partial y}$ continues sur D.

Nous supposons en outre que les quantités $|\frac{\partial f}{\partial y}|$ et $|\frac{\partial^2 f}{\partial y^2}|$ sont bornées sur D par F_1 et F_2 respectivement. Alors l'erreur de la méthode d'Euler à l'abcisse $x_n \in [a, b]$ obtenue avec un pas h vérifie :

$$e_n = h\, e(x_n) + h^2\, B(n, h)$$

où B est une fonction bornée de n et de h, $(|B(n, h)| \leqslant H)$ et e(x) la solution de l'équation différentielle :

$$\frac{de}{dx} = f'_y(y(x), x) \times e - \frac{1}{2} y''(x) \quad ; \quad e(a) = 0$$

Démonstration : On a encore la formule de passage :

$$y_{m+1} = y_m + h\, f(y_m, x_m)$$

D'après les hypothèses, la solution exacte y(x) possède une dérivée troisième continue, et l'on peut écrire :

$$y(x_{m+1}) = y(x_m) + h\, f(y(x_m), x_m) + \frac{h^2}{2} y''(x_m) + \frac{h^3}{3!} y'''(\xi)$$

avec $\xi \in [x_m, x_{m+1}]$.

En soustrayant on obtient :

$$e_{m+1} = e_m + h\,[f(y(x_m) + e_m, x_m) - f(y(x_m), x_m)] - \frac{h^2}{2} y''(x_m) - \frac{h^3}{3!} y'''(\xi)$$

L'expression entre crochets s'écrit :

$$f'_y(y(x_m), x_m)\, e_m + \frac{f''_{y^2}}{2}(y^*, x_m)\, e_m^2$$

(y^* se trouve entre $y(x_m)$ et y_m).

En appelant $\overline{e_m}$ la quantité $\frac{e_m}{h}$, on passe de $\overline{e_m}$ à $\overline{e_{m+1}}$ par :

$$\overline{e_{m+1}} = \overline{e_m} + h\,[f'_y(y(x_m),\ x_m)\ \overline{e_m} - \frac{1}{2}\,y''(x_m)] + h^2\ \rho_m$$

avec

$$|\rho_m| \leqslant |f''_{y2}(y^*,\ x_n)|\ \overline{e_m^2} + \frac{1}{3!}\ |y'''(\xi)|$$

On sait déjà (théorème 1) que :

$$|\overline{e_m}| = |\frac{e_m}{h}| \leqslant C_1 = N(b) \times \left(\frac{e^{F_1(b-a)} - 1}{F_1}\right)$$

Posons :

$$G_3 = \operatorname*{Max}_{t\in[a,b]} |y'''(t)|$$

On a alors :

$$|\rho_m| < F_2 \times C_1^2 + \frac{1}{6}\,G_3 = C_2$$

On est dans les conditions d'application du théorème 2 ; la formule de passage de $\overline{e_m}$ à $\overline{e_{m+1}}$ correspond à une formule d'intégration approchée de l'équation différentielle (y(x) supposé connu):

$$\frac{de}{dx} = f'_y(y(x),\ x) \times e - \frac{1}{2}\,y''(x)\ ;\ e(a) = 0$$

soit :

$$\frac{de}{dx} = g(e,\ x)\ ;\ e(a) = 0$$

Les dérivées partielles $\frac{\partial g}{\partial e}$ et $\frac{\partial g}{\partial x}$ sont continues et $\left|\frac{\partial g}{\partial e}\right|$ est borné sur D. Posons :

$$L_1 = \operatorname*{Max}_{x\in[a,b]} |f'_y(y(x),\ x)|$$

e''(x) existe et est continue sur [a, b].

On a donc d'après le théorème 2 :

$$|\overline{e_n} - e(x_n)| \leqslant h\left[\frac{1}{2}\operatorname*{Max}_{x\in[a,b]}|e''(x)| + C_2\right]\left(\frac{e^{L_1(b-a)} - 1}{L_1}\right)$$

$$\leqslant h\,H$$

où H est indépendant de n et h.

D'où :

$$e_n = h\ e(x_n) + h^2\ B(n,\ h) \quad \text{avec} \quad |B(n,\ h)| \leqslant H$$

Théorème 4 : Supposons que f(y, x) soit continue sur D et admette des dérivées partielles jusqu'à l'ordre 3 continues sur D. Supposons que les quantités $|\frac{\partial f}{\partial y}|$, $|\frac{\partial^2 f}{\partial y^2}|$ et $|\frac{\partial^3 f}{\partial y^3}|$ soient bornées sur D par F_1, F_2 et F_3 respectivement.

Alors l'erreur de la méthode d'Euler à l'abcisse $x_n \in [a, b]$ obtenue avec un pas h vérifie :

$$e_n = h\ e(x_n) + h^2\ \varepsilon(x_n) + h^3\ A(n,\ h)$$

où A est une fonction bornée de n et h

$(|A(n, h)| \leqslant J)$

et $\varepsilon(x)$ la solution de l'équation différentielle :

$$\frac{d\varepsilon}{dx} = f'_y(y(x), x)\,\varepsilon + \frac{f''_{y^2}(y(x), x)}{2} \times e(x)^2 - \frac{1}{6}\,y'''(x) - \frac{e''(x)}{2}$$

$\varepsilon(a) = 0$

et $e(x)$ la même fonction qu'au théorème 3.

Démonstration : Le principe de cette démonstration est analogue à celui de la démonstration précédente. La solution $y(x)$ admet une dérivée quatrième continue et l'on a :

$$e_{m+1} = e_m + h\,[f(y(x_m) + e_m,\ x_m) - f(y(x_m),\ x_m)] - \frac{h^2}{2}\,y''(x_m) - \frac{h^3}{3!}\,y'''(x_m) - \frac{h^4}{4!}\,y^{(4)}(\xi)$$

Posons :

$$e_m = h\,e(x_m) + \varepsilon_m \times h^2$$

(ε_m dépend en fait aussi de h).

On sait déjà que $|\varepsilon_m| \leqslant H$ (théorème 3).

La quantité entre crochets peut s'écrire :

$$f'_y(y(x_m), x_m) \times e_m + \frac{f''_{y^2}(y(x_m), x_m) \times e_m^2}{2} + f'''_{y^3}(y^*, x_m) \times \frac{e_m^3}{6}$$

On obtient alors en simplifiant par h :

$$\left\{e(x_{m+1}) - e(x_m) - h\left[f'_y(y(x_m), x_m)\,e(x_m) - \frac{y'''(x_m)}{2}\right]\right\} + h\,[\varepsilon_{m+1} - \varepsilon_m - \frac{h}{2}\,(f''_{y^2}(y(x_m), x_m)\,e(x_m)^2$$

$$- \frac{y'''(x_m)}{3!} + f'_y(y(x_m), x_m)\,\varepsilon_m)] + \frac{h^3}{4!}\,y^{(4)}(\xi) - f'''_{y^3}(y^*, x_m)\,e_m^3$$

$$- \frac{f''_{y^2}(y(x_m), x_m)}{2}\,[\varepsilon_m\,h^4 + 2h^3\,e(x_m)\,\varepsilon_m] = 0$$

La quantité entre accolades s'écrit :

$$e(x_{m+1}) - e(x_m) - he'(x_m) = \frac{h^2}{2}\,e''(x_m) + \frac{h^3}{3!}\,e'''(\xi^*)$$

(e''' existe et est continue avec les hypothèses faites).

On trouve donc en divisant à nouveau par h :

$$\varepsilon_{m+1} = \varepsilon_m + h\left[f'_y(y(x_m), x_m)\,\varepsilon_m + \frac{f''_{y^2}(y(x_m), x_m)\,e(x_m)^2}{2} - \frac{y'''(x_m)}{3!} - \frac{e''(x_m)}{2}\right] + h^2\,r_m$$

avec :

$$|r_m| \leqslant \frac{G_4}{4!} + \frac{F_3\,C_1^3}{3!} + \frac{F_2\,H(b-a)}{2} + F_2\,H\,E + \frac{E_3}{3!} = C_3$$

On a posé :

$$G_4 = \max_{t\in[a,b]} |y^{(4)}(t)| \quad ; \quad E = \max_{x\in[a,b]} |e(x)| \quad ; \quad E_3 = \max_{x\in[a,b]} |e'''(x)|$$

On obtient donc les ε_m par intégration approchée de l'équation différentielle suivante :

$$\frac{d\varepsilon}{dx} = f'_y(y(x), x) \times \varepsilon + \frac{f''_{y^2}(y(x), x)}{2} e(x)^2 - \frac{y'''(x)}{3!} - \frac{e''(x)}{2} \quad (\text{avec } \varepsilon(a) = 0)$$

au moyen d'une méthode voisine de celle d'Euler : on est dans les conditions d'application du théorème 2 :

$$|\varepsilon_n - \varepsilon(x_n)| \leq h \left[\operatorname*{Max}_{x \in [a,b]} \frac{|\varepsilon''(x)|}{2} + C_3 \right] \times \left(\frac{e^{L_1(b-a)} - 1}{L_1} \right) = h\, J$$

donc $\varepsilon_n = \varepsilon(x_n) + h\, A(n, h)$ où A est une fonction de n et h bornée par J, ce qui achève la démonstration.

Remarques :

1/ En faisant les hypothèses convenables sur les dérivées d'ordre plus élevées de f(y, x) on peut conclure à l'existence d'un développement en puissance de h jusqu'à un ordre quelconque.

2/ On peut aboutir à la même conclusion pour d'autres méthodes que celle d'Euler (Runge Kutta, par exemple). Pour des méthodes plus puissantes, les premiers termes de développement en h sont nuls.

3/ Les hypothèses f'_y, f''_{y^2}, f'''_{y^3} bornées dans D peuvent paraître difficiles à remplir : on peut en fait les alléger : si l'on sait que la solution y(x) et les diverses solutions approchées (x_n, y_n) restent dans un domaine compact et convexe en y : $D_1 \subset D$ pour tout n et tout h, ce qui a lieu dans la plupart des cas pratiques) alors il suffit d'avoir f'_y, f''_{y^2} et f'''_{y^3} bornées dans D_1, ce qui est acquis par la seule continuité de ces quantités (ce sera le cas pour l'exemple n° 1 du § 5.2).

4/ On peut obtenir des résultats analogues pour un système d'équations différentielles d'ordre quelconque.

§ 5.2 - EXTRAPOLATION SUR LA FORMULE D'EULER, PROCEDURE ALGOL ET EXEMPLES NUMERIQUES

Pour un *x fixé*, appelons v(h) la solution approchée obtenue en x avec un pas h (qui est forcément de la forme $\frac{x - a}{N}$, N entier) par la méthode d'Euler. Les théorèmes du § 5.1 montrent qu'avec des hypothèses convenables sur f(y, x), v(h) admet un développement en puissances de h de la forme :

$$v(h) = c_o + c_1 h + c_2 h^2 + \ldots + c_{r-1} h^{r-1} + h^r B(h) \qquad (5.2.1)$$

où les c_i sont des constantes (c_o = solution exacte en x) et B une fonction bornée de h. On peut donc employer le procédé d'extrapolation L_r basé sur les r pas $h_1, h_2, \ldots, h_r$ (voir chapitre 1).

On prendra souvent :

$$h_i = \frac{h}{2^{i-1}} \; ; \quad h = \frac{x - a}{N}$$

Soit $E_m^1(x)$ le résultat en x de l'intégration par la méthode d'Euler avec le pas $\frac{h}{2^{m-1}}$. On forme alors $E_m^n(x)$ de proche en proche par la formule :

$$E_{q+1}^{p+1} = \frac{2^p E_{q+1}^p - E_q^p}{2^p - 1} \qquad (5.2.2)$$

§ 5.2

Formules de passage basées sur une extrapolation :

Au lieu de faire toute l'intégration (jusqu'au bout de l'intervalle considéré avec un pas h, puis $\frac{h}{2}$, puis $\frac{h}{4}$ et ainsi de suite, on peut effectuer le passage de t_i à $t_i + h$ d'abord avec un pas h, ensuite $\frac{h}{2}$, $\frac{h}{4}$ etc. Si l'on appelle encore E_m^1 la valeur numérique obtenue en $t + h$ par la formule de la tangente entre t_i et $t_i + h$ avec un pas $\frac{h}{2^{m-1}}$, on peut appliquer (5.2.2) pour obtenir des formules de passage plus puissantes.

a) *Formule* E_2^2 :

$$E_2^2 = 2\ E_2^1 - E_1^1$$

Cette formule correspond à la méthode de la tangente améliorée [25].

b) *Formule* E_3^3 :

$$E_3^3 = \frac{8}{3} E_3^1 - 2\ E_2^1 + \frac{1}{3} E_1^1$$

Cette formule correspond à la formule de Runge-Kutta suivante :

$$y_{i\alpha} = y_i + h \sum_{\beta=0}^{\alpha-1} A_{\alpha\beta}\ f_{i\beta}$$

$$f_{i\beta} = f(y_{i\beta},\ t_i + \theta_\beta)$$

$$y_{i+1} = y_{iq}$$

avec les constantes suivantes :

$$q = 5 \text{ (méthode de rang 5)}$$

$$\theta_1 = \frac{1}{4} \quad \theta_2 = \frac{1}{2} \quad \theta_3 = \frac{3}{4} \quad \theta_4 = \frac{1}{2} \quad \theta_5 = 1$$

$$A_{10} = A_{21} = A_{20} = A_{32} = A_{31} = A_{30} = \frac{1}{4}$$

$$A_{40} = \frac{1}{2} \ ; \ A_{41} = A_{42} = A_{43} = 0 \ ;$$

$$A_{50} = 0 \ ; \ A_{51} = A_{52} = A_{53} = \frac{2}{3} \ ; \ A_{54} = -1$$

Son développement de Taylor coïncide avec le développement de Taylor de la solution jusqu'en h^3 (voir tableau, [25] page 55) ; l'erreur sur un pas est d'ordre h^4.

Les formules itératives E_n^n sont d'un emploi commode : pour intégrer sur un pas, on calcule les E_n^n successifs jusqu'à stabilisation avant de passer au pas suivant : on dispose ainsi d'un certain contrôle sur l'erreur par pas.

PROCEDURE ALGOL :

```
procédure EQUADIF (F, A, B, YO, H, ORDMAX, EPS, T, Y) ;
valeur A, B, YO, H ; réel procédure F ; réel A, B, YO, H, EPS, T, Y ; entier ORDMAX ;
Commentaire. Cette procédure intègre l'équation différentielle Y' = F(Y, T) avec Y(A) = YO dans
l'intervalle [A, B].
```

On prend un pas de base H. Pour trouver la valeur de la fonction à la fin de chacun de ces intervalles élémentaires on emploie la formule de la tangente avec des pas différents en progression géométrique de raison 1/2 et on fait une extrapolation de Richardson. On passe au pas suivant quand

l'erreur relative est inférieure à EPS (erreur par pas H) ou quand on a pris ORDMAX + 1 pas différents.

Les résultats sont portés par T et Y ;

```
début tableau RES [1 : ORDMAX + 1] ; réel MA, HE, YE, RE ;
entier J, FAC, I ; T := A ; Y := YO ;
INTEGRE : MA := RES [1] := Y + H × F(Y, T) ; HE := H/2.0 ;
          pour J := 1 pas 1 jusqua ORDMAX faire
          début YE := Y ; TE := T ;
          ITER : YE := YE + HE × F(YE, TE) ; TE := TE + HE ;
          si TE < T + H - HE/2.0 alors allera ITER ;
          RES [J + 1] := YE ; FAC := 1 ;
          pour I := J pas - 1 jusqua 1 faire
          début FAC := 2 × FAC ;
               RES [I] := RES [I + 1] + (RES [I + 1] - RES [I])/(FAC - 1)
          fin ;
          si ABS ((RES [1] - MA)/RES [1]) < EPS alors
          allera TERM ; MA := RES [1] ; HE := HE/2.0
          fin ;
TERM : Y := RES [1] ; T := T + H ;
       si T < B - H/2.0 alors allera INTEGRE
fin
```

Exemples numériques :

Nous donnons trois exemples d'utilisation de la procédure précédente avec les valeurs suivantes des paramètres : EPS = 10^{-5} ; ORDMAX = 8 ; H = 0,4.

Exemple n° 1: $y' = -2t\,y^2$; $y(0) = 1$;

t	Solution exacte	solution approchée par EQUADIF	Erreurs par R. K. classique $\times 10^6$	Erreurs EQUADIF $\times 10^6$
0,4	0,862 068 9	0,862 068 6	409	- 0,3
0,8	0,609 756 1	0,609 755 9	297	- 0,2
1,2	0,409 836 0	0,409 835 8	+ 147	- 0,2
1,6	0,280 898 6	0,280 898 4	- 225	- 0,2
2,0	0,200 000 0	0,199 999 8	- 177	- 0,2

La comparaison à égalité de coût avec les méthodes classiques est assez difficile : en effet, la procédure EQUADIF divise chaque intervalle de base (de longueur 0,4 ici) en sous intervalles de plus en plus petits jusqu'à obtention de la précision demandée. Nous donnons à titre indicatif l'erreur pour la méthode de Runge Kutta avec un pas 0,4.

Exemple n° 2 : $y' = y$; $y(0) = 1$;

t	Solution exacte	Solution approchée par EQUADIF	Erreurs R. K. classique $\times 10^6$	Erreurs EQUADIF $\times 10^6$
0,4	1,491 824	1,491 824	91	- 0,6
0,8	2,225 541	2,225 539	272	- 2
1,2	3,320 117	3,320 113	610	- 4
1,6	4,953 032	4,953 025	1213	- 7
2,0	7,389 056	7,389 042	2262	- 14

Exemple n° 3 : $y' = 6y/(1 + t)$; $y(0) = 1$

t	Solution exacte	Solution approchée EQUADIF	Erreurs R. K. classique $\times 10^4$	Erreurs EQUADIF $\times 10^4$
0,4	7,529 53	7,529 52	4152	- 0,1
0,8	34,012 22	34,012 14	24920	- 0,8
1,2	113,379 90	113,379 55	91400	- 3,5
1,6	308,915 7	308,914 6	259620	- 11
2,0	729,000 0	728,996 6	625560	- 34

Remarque :

Si l'on remplace la formule d'Euler par la formule de passage suivante :

$$y_{m+1} = y_m + \frac{h}{2}\,[f(y_m,\ x_m) + f(y_{m+1},\ x_{m+1})] \qquad (5.2.3)$$

les termes impairs du développement en puissances de h de la solution approchée en un point fixe x disparaissent. On pourrait donc employer le procédé M_n pour polynômes pairs (§ 1.4). Cependant, comme la formule (5.2.3) est implicite, y_{n+1} ne peut être obtenue qu'itérativement, et il est à craindre que ceci alourdisse le procédé.

§ 5.3 - ETUDE DU COMPORTEMENT DE LA SOLUTION APPROCHEE EN FONCTION DU PAS h POUR UN PROBLEME DE CONDITIONS AUX LIMITES

Considérons le problème particulier suivant :

$$y'' = f(y, x) \quad ; \quad y(a) = A \quad ; \quad y(b) = B \qquad (5.3.1)$$

Même pour cet exemple simple, la situation est plus compliquée que dans le cas d'un problème de conditions initiales : il peut y avoir selon les cas plusieurs solutions, une solution ou aucune ; de même il peut y avoir plusieurs solutions, une solution ou aucune pour le systéme d'équations obtenu en discrétisant (5.3.1). Pour simplifier nous ferons les hypothèses suivantes *dans tout le paragraphe* 5.3 :

$$\left.\begin{array}{l} f(y, x) \text{ continue sur } D = [a, b] \times]-\infty, +\infty[\\ \dfrac{\partial f}{\partial y} \text{ existe, est continue et vérifie } 0 \leqslant \dfrac{\partial f}{\partial y} \leqslant F_1 \text{ sur } D. \end{array}\right\} \qquad (5.3.2)$$

(F_1 est une constante).

On sait alors qu'il existe une solution unique y(x) *deux fois continûment différentiable vérifiant* (5.3.1) (voir 18).

Théorème 1 : On suppose que f(y, x) possède des dérivées partielles continues jusqu'à l'ordre 2. On pose :

$$G_4 = \max_{t \in [a,b]} |y^{(4)}(t)|$$

Appelons y_i, i = 1, 2, ..., N - 1 une solution du système d'équations :

$$- y_{i-1} + 2y_i - y_{i+1} + h^2 f(y_i, x_i) = 0 \quad ; \quad y_0 = A \quad ; \quad y_N = B$$

avec h = (b - a)/N (on suppose l'existence d'une telle solution) et posons $e_n = y_n - y(x_n)$, erreur à l'abcisse $x_n = a + (b - a) \times \dfrac{n}{N}$

On a alors :

$$|e_n| \leqslant \frac{(b-a)^2}{4} \frac{G_4}{12} h^2$$

Remarques :

1/ On peut parfois établir l'existence et l'unicité de la solution approchée y_i : c'est le cas si $|f''_{y^2}|$ est bornée par L_2 sur D et si h est suffisamment petit ; par exemple :

$$h^2 < \min\left(\frac{1}{F_1}, \frac{32 \times 12}{(b-a)^4 L_2 G_4}\right)$$

2/ Il suffit que les dérivées partielles d'ordre 2 de f(y, x) existent et soient continues pour que y(x) soit 2 fois continûment différentiable.

3/ Nous démontrerons le théorème 2 suivant qui contient le théorème 1 comme cas particulier.

Théorème 2 : Avec les mêmes hypothèses que pour le théorème 1, on suppose que les y_i vérifient :

$$-y_{i-1} + 2y_i - y_{i+1} + h^2 f(y_i, x_i) = h^4 \theta(i, h) K$$

où K est une constante et $|\theta(i, h)| \leqslant 1$ pour tout i et tout h ; on a alors :

$$|e_n| \leqslant \frac{(b-a)^2}{4}\left(\frac{G_4}{12} + K\right) h^2$$

Démonstration : On sait que :

$$-y(x_{i-1}) + 2y(x_i) - y(x_{i+1}) + h^2 f(y(x_i), x_i) = \theta'(i, h) \frac{G_4}{12} h^4$$

avec $|\theta'(i, h)| \leqslant 1$.

En soustrayant avec la relation vérifiée par les y_i on obtient :

$$-e_{i-1} + 2e_i - e_{i+1} + h^2 f'_y(y_i^*, x_i) e_i = \theta''(i, h)\left[\frac{G_4}{12} + K\right] h^4$$

avec y_i^* compris entre y_i et $y(x_i)$ et $|\theta''(i, h)| \leqslant 1$.

- Appelons J la matrice carrée tridiagonale de dimension N - 1 ayant des 2 sur la diagonale principale et - 1 sur les deux sous-diagonales.

- Appelons G la matrice carrée diagonale de dimension N - 1 ayant $f'_y(y_i^*, x_i)$ comme éléments diagonaux. Les éléments de G sont compris entre 0 et F_1.

Le système linéaire qui donne les e_i peut s'écrire :

$$(J + h^2 G)\, e = h^4 \left(\frac{G_4}{12} + K\right) \theta$$

où θ est un vecteur dont les composantes sont inférieures à 1.

En utilisant des propriétés des matrices monotones et irréductibles (voir [18] page 358), on peut démontrer que la matrice $J + h^2 G$ est inversible pour h assez petit $\left(h^2 < \frac{1}{F_1}\right)$ et que les éléments de son inverse $(J + h^2 G)^{-1}$ sont non négatifs et inférieurs aux éléments correspondants de J^{-1} (J est inversible).

Appelons $j_{m,n}$ les éléments de J^{-1} on a alors :

$$|e_m| \leqslant h^4 \left(\frac{G_4}{12} + K\right) \sum_{n=1}^{N-1} \left(j_{m,n}\right)$$

§ 5.3

On peut calculer explicitement les éléments de J^{-1} :

$$j_{m,n} = \begin{cases} \dfrac{(N - m)\, n}{N} & \text{si } n \leqslant m \\ \dfrac{m(N - n)}{N} & \text{si } n > m \end{cases}$$

On a :

$$\sum_{n=1}^{N-1} j_{m,n} = \frac{m(N - m)}{2} = \frac{(x_m - a)\ (b - x_m)}{2h^2} \leqslant \frac{(b - a)^2}{4h^2}$$

d'où :

$$|e_m| \leqslant \frac{(b - a)^2}{4} \left(\frac{G_4}{12} + K\right) h^2$$

pour tout m et tout h $\left(h^2 < \frac{1}{F_1}\right)$.

Théorème 3 : On suppose que f(y, x) possède des dérivées partielles continues jusqu'à l'ordre 4 et que $|\frac{\partial^2 f}{\partial y^2}|$ soit bornée par F_2.

Supposons que les y_i vérifient :

$$- y_{i-1} + 2y_i - y_{i+1} + h^2 f(y_i, x_i) = 0 \qquad i = 1, 2, \ldots, N - 1 ; y_o = A \ ; \ Y_N = B$$

On a alors quel que soit h et n $\left(h^2 < \frac{1}{F_1}\right)$:

$$e_n = h^2\ e(x_n) + h^4\ B(n, h)$$

où B est une fonction bornée de n, h et e(x) la solution (qui existe et est unique) du problème aux limites :

$$\frac{d^2e}{dx^2} = g(x)\ e - \frac{1}{12} y^{(4)}(x) \ ; \ e(a) = 0 \ ; \ e(b) = 0$$

avec $g(x) = f'_y(y(x), x)$

Démonstration : La solution exacte y(x) vérifie :

$$- y(x_{i-1}) + 2y(x_i) - y(x_{i+1}) + h^2\ f(y(x_i), x_i) = \frac{-1}{12} h^4\ y^{(4)}(x_i) + \frac{h^6}{360}\ \theta_i\ G_6$$

avec $G_6 = \underset{t \in [a,b]}{\text{Max}} |y^{(6)}(t)|$ et $|\theta_i| \leqslant 1$

(En fait θ_i dépend aussi de h).

En soustrayant avec la relation que vérifie les y_i et tenant compte du fait que :

$$f(y_i, x_i) - f(y(x_i), x_i) = f'_y(y(x_i), x_i)\ e_i + \frac{f''_{y^2}(y_i^*, x_i)}{2}\ e_i^2$$

on obtient pour $\overline{e_i} = e_i/h^2$

$$- \overline{e_{i-1}} + 2\ \overline{e_i} - \overline{e_{i+1}} + h^2 \left(f'_y(y(x_i), x_i)\ \overline{e_i} - \frac{1}{12} y^{(4)}(x_i)\right) = h^4\ \theta'_i\ C_2$$

avec $|\theta'_i| \leqslant 1$

$$C_2 = \frac{G_6}{360} + \frac{F_2}{2} C_1^2 \qquad C_1 = \frac{(b-a)^2}{4} \frac{G_4}{12}$$

$\overline{e_i}$ est donc obtenue par résolution approchée du problème :

$$\frac{d^2e}{dx^2} = f'_y(y(x), x)\ e - \frac{1}{12} y^{(4)}(x) \qquad e(a) = 0 \qquad e(b) = 0$$

Ce problème admet une solution unique car :

$$0 \leqslant f'_y(y(x), x) \leqslant F_1$$

(il vérifie les hypothèses (5.3.2)).

On est alors dans les conditions d'application du théorème 2 :

$$|\overline{e_n} - e(x_n)| \leqslant \frac{(b-a)^2}{4} \left(\frac{E_4}{12} + C_2\right) h^2 \qquad \text{avec } E_4 = \max_{t\in[a,b]} |e^{(4)}(t)|$$
$$\leqslant H h^2$$

donc :

$$e_n = h^2\ e(x_n) + h^4\ B(n, h) \qquad \text{avec } |B(n, h)| \leqslant H$$

Théorème 4 : On suppose que $f(y, x)$ possède des dérivées partielles continues jusqu'à l'ordre 6 et que $|\frac{\partial^2 f}{\partial y^2}|$ et $|\frac{\partial^3 f}{\partial y^3}|$ sont bornées sur D par F_2 et F_3 respectivement. Les y_i vérifiant les mêmes relations qu'au théorème 3, on a pour $h^2 < \frac{1}{F_1}$:

$$e_n = h^2\ e(x_n) + h^4\ \varepsilon(x_n) + h^6\ A(n, h)$$

où A est une fonction bornée de n et h, et $\varepsilon(x)$ est la solution (qui existe et est unique) du problème aux limites :

$$\frac{d^2\varepsilon}{dx^2} = f'_y(y(x), x)\ \varepsilon + \frac{1}{2} f''_y(y(x), x)\ e(x)^2 - \frac{1}{360} y^{(6)}(x) - \frac{1}{12} e^{(4)}(x)$$

$\varepsilon(a) = 0$; $\varepsilon(b) = 0$

Démonstration : La démonstration est analogue à celle du théorème 4 du § 5.1.

On pose :

$$e_i = h^2\ e(x_i) + h^4\ \varepsilon_i$$

On sait que $|\varepsilon_i| \leqslant H$.

On trouve pour ε_i :

$$-\varepsilon_{i-1} + 2\varepsilon_i - \varepsilon_{i+1} + h^2 \left[f'_y(y(x_i), x_i)\ \varepsilon_i + \frac{1}{2} f''_{y^2}(y(x_i), x_i)\ e(x_i)^2 - \frac{1}{360} y^{(6)}(x_i) - \frac{1}{12} e^{(4)}(x_i) \right] + h^4\ \theta_i\ C_3$$

avec $|\theta_i| \leqslant 1$ et :

$$C_3 = \frac{2}{8!} G_8 + \frac{F_2}{2} (2\ EH + (b-a)^2\ H^2) + \frac{F_3\ C_1^3}{3} + \frac{2}{6!} E_6$$

indépendant de i et h.

Les ε_i proviennent donc de l'intégration approchée du problème aux limites donné dans l'énoncé du théorème pour $\varepsilon(x)$.

Ce problème vérifie les hypothèses (5.3.1) donc a une solution unique.

§ 5.4

On peut appliquer le théorème 2 :

$$|\varepsilon_n - \varepsilon(x_n)| \leqslant \frac{(b-a)^2}{4} \left(\operatorname*{Max}_{t\in[a,b]} \frac{|\varepsilon^{(4)}(t)|}{12} + C_3 \right) h^2 = J\, h^2$$

d'où :

$$e_n = h^2\ e(x_n) + h^4\ \varepsilon(x_n) + h^6\ A(n, h) \quad \text{avec } |A(n, h)| \leqslant J$$

Remarques :

1/ On peut, en faisant des hypothèses sur les dérivées partielles d'ordre plus élevé de f, obtenir l'existence de développements en puissance de h jusqu'à un ordre quelconque.

2/ On peut obtenir des résultats analogues pour des méthodes numériques plus puissantes que celle utilisée ici.

3/ Si l'on peut affirmer que la solution y(x) et les diverses solutions approchées appartiennent à un domaine D_1 compact et convexe en y alors il suffit que les dérivées partielles par rapport à y soient bornées sur D , ce qui est acquis par leur continuité (ce sera le cas pour l'exemple traité au § 5.4).

§ 5.4 - EXEMPLE D'EXTRAPOLATION POUR UN PROBLEME DE CONDITIONS AUX LIMITES

Nous considérons le problème (5.3.1). Si l'on appelle w(h) la solution approchée en un point fixe $x \in [a, b]$ avec un pas $h = \frac{b-a}{N}$, (x et h doivent être pris de façon à ce que l'on ait $x = nh$, n entier), les théorèmes du § 5.3 montrent qu'avec des hypothèses convenables sur f(y, x) on a :

$$w(h) = d_0 + d_2 h^2 + d_4 h^4 + \ldots\ldots + d_{2r-2} h^{2r-2} + h^{2r} B(h)$$

où les d_i sont des constantes et B une fonction bornée de h. On peut donc appliquer le procédé M_n (§ 1.4) basé sur les polynômes pairs.

Exemple :

Considérons le problème aux limites :

$$y'' = \frac{3}{2} y^2 \quad ; \quad y(0) = 4 \quad ; \quad y(1) = 1$$

Notons que les hypothèses (5.3.2) ne sont pas vérifiées. En fait le problème admet deux solutions dont l'une est $y(x) = \frac{4}{(1+x)^2}$. C'est cette solution que nous allons chercher numériquement pour $x = 0,5$ ($y(0,5) = 1,777\,777$). On discrétise avec un pas $h = \frac{1}{2N}$ comme au paragraphe précédent :

$$-y_{i-1} + 2y_i - y_{i+1} + h^2 f(y_i, x_i) = 0 \quad (y_0 = 4 \quad ; \quad y_{2N} = 1)$$

Pour résoudre ce problème on peut se donner une valeur y_1 de départ de la forme :

$$y_1 = y_0 + \mu \times h$$

et corriger μ itérativement afin de trouver $y_{2N} = 1$.

Appelons $y(x, \mu)$ la solution trouvée avec un certain μ : on choisit la correction $\Delta\mu$ telle que :

$$y(1, \mu + \Delta\mu) \# y(1, \mu) + \Delta\mu \times \frac{\partial y}{\partial \mu}(1, \mu) = 4$$

Pour trouver $\frac{\partial y}{\partial \mu}(x, \mu) = \eta(x, \mu)$ on résoud l'équation supplémentaire :

$$\eta'' = f'_y(y(x), x) \times \eta \qquad \text{avec } \eta_0 = \eta(0) = 0$$
$$\eta_1 = \eta(h) = h$$

Selon la valeur de départ donnée à μ on converge vers l'une ou l'autre des deux solutions du problème.

Les résultats sont rassemblés dans le tableau suivant (la troisième colonne du tableau représente le résultat de l'extrapolation M_n portant sur la deuxième colonne) :

h	Solution normale	Solution avec extrapolation	Erreur solution normale $\times 10^5$	Erreur solution avec extrapolation $\times 10^5$
1/2	1,854 88	1,854 88	77 11	77 11
1/4	1,800 79	1,782 76	23 02	4 99
1/8	1,783 88	1,777 95	6 11	18
1/16	1,779 33	1,777 78	1 56	1

Remarques :

1/ En combinant les résultats obtenus avec les pas 1/2, 1/4, 1/6, 1/10, 1/16 (qui sont dans un rapport approximatif 1,5) au lieu de 1/2, 1/4, 1/8, 1/16 (qui sont dans un rapport 2) on obtient une erreur dix fois plus petite.

2/ On peut se donner $y'(0) = \overline{w}$, intégrer par les formules de passage de la tangente pour une équation du deuxième ordre :

$$y_{i+1} = y_i + hy'_i + \frac{h^2}{2} f(y_i, x_i)$$

$$y'_{i+1} = y'_i + h\, f(y_i, x_i)$$

On corrige $\overline{w}$ comme on l'a fait pour μ afin d'atteindre $y_{2N} = 4$. Mais avec ce procédé les puissances impaires du développement ne disparaissent pas ; il faut donc employer le procédé L_n. Les résultats sont moins bons :

Pas	Solution normale	Solution avec extrapolation	Erreur solution normale $\times 10^5$	Erreur solution avec extrapolation $\times 10^5$
1/2	0,920 554	0,920 554	- 857 22	- 857 22
1/4	1,555 880	2,191 206	- 221 89	+ 413 42
1/8	1,685 461	1,689 653	- 92 31	- 88 12
1/16	1,734 840	1,785 988	- 42 93	+ 8 21
1/32	1,756 996	1,777 429	- 20 78	- 34
1/64	1,767 543	1,777 774	- 10 23	- 0,3

§ 5.5 - EXEMPLE D'EXTRAPOLATION POUR UN PROBLEME DE VALEURS PROPRES

Considérons l'équation d'Airy :

$$y'' + \lambda x\, y = 0 \qquad y(0) = y(1) = 0$$

Les valeurs propres sont les racines de $J_{1/3}\left(\frac{2\sqrt{\lambda}}{3}\right) = 0$ et les fonctions propres correspondantes :

$$y = \sqrt{x}\, J_{1/3}\left(\frac{2\sqrt{\lambda}}{3} x^{3/2}\right)$$

Cherchons par exemple les deux premières valeurs propres qui sont voisines des nombres 18 et 80. On résoud le problème avec un pas h :

$$y_{i-1} - (2 - \lambda h^2 x_i)\, y_i + y_{i+1} = 0 \quad \text{avec} \quad \begin{cases} y_o = 0 \\ y_n = 0 \end{cases}$$

$$y_i = y(i \times h)$$

On peut trouver le λ convenable par itérations en résolvant plusieurs fois un problème de conditions initiales. On se donne $y_o = 0$ et $y_1 = h$. Appelons $y(x, \lambda)$ la solution obtenue avec un certain λ.

La quantité $z(x, \lambda) = \frac{\partial y(x, \lambda)}{\partial \lambda}$ obéit à l'équation :

$$z'' + \lambda x z + x y = 0$$

On résoud cette équation en même temps que la précédente :

$$z_{i-1} - (2 - h^2 \lambda x_i)\, z_i + z_{i+1} + h^2 x_i y_i = 0$$

en se donnant $z_o = 0$; $z_1 = 0$.

En partant d'une valeurs approchée $\lambda^{(o)}$ on détermine la correction $\Delta\lambda^{(o)}$ de la façon suivante :

$$y(1, \lambda^{(1)}) \# y(1, \lambda^{(o)}) + \Delta\lambda^{(o)} \times z(1, \lambda^{(o)}) = 0$$

et on recommence l'intégration avec ce nouveau $\lambda^{(1)} = \lambda^{(o)} + \Delta\lambda^{(o)}$ et ainsi de suite.

Appelons $\lambda(h)$ la valeur propre obtenue avec un pas h. La discrétisation étant symétrique en h, les termes impairs du développement en puissance de h de $\lambda(h)$ disparaissent et on peut ainsi employer le procédé M_n (chapitre 1). En appliquant la procédure RINEVILLE pour polynômes pairs (T = 2) sur $\lambda(h)$ avec un pas de départ HD = 0.2 et un rapport ALPHA = 1.5 entre les pas successifs, ont obtient les résultats suivants :

a) 1ère valeur propre : Valeur de départ $\lambda^{(o)} = 18$

Valeur exacte : 18,956.

Pas	solution normale	solution extrapolée	erreur solution normale $\times 10^3$	erreur solution extrapolée $\times 10^3$
1/5	18,251	18,251	- 705	- 705
1/10	18,777	18,953	- 179	- 3
1/15	18,876	18,956	- 80	< 1

b) 2ème valeur propre : Valeur de départ $\lambda^{(o)} = 80$

Valeur exacte : 81,886.

Pas	solution normale	solution extrapolée	erreur solution normale $\times 10^3$	erreur solution extrapolée $\times 10^3$
1/5	69,232	69,232	- 12 654	- 12 654
1/10	78,565	81,676	- 3 321	- 210
1/15	80,402	81,896	- 1 484	- 10
1/25	81,350	81,886	- 536	< 1

CHAPITRE 6

APPLICATION AUX ÉQUATIONS INTÉGRALES ET AUX ÉQUATIONS AUX DÉRIVÉES PARTIELLES

§ 6.1 - ETUDE DU COMPORTEMENT DE LA SOLUTION APPROCHEE POUR UNE EQUATION INTEGRALE DE FREDHOLM DE DEUXIEME ESPECE

Considérons l'équation intégrale suivante :

$$f(x) = \int_0^1 K(x, y)\, f(y)\, dy + g(x) \qquad (6.1.1)$$

Nous supposons que g soit continue et deux fois continûment différentiable sur [0,1], que K(x, y) soit continue et possède des dérivées partielles continues jusqu'à l'ordre 2. Supposons que (6.1.1) admette une solution unique. On sait alors que cette solution admet une dérivée seconde continue.

Pour calculer l'intégrale on emploie la formule des trapèzes avec un pas $h = \frac{1}{N}$:

$$f(x) = \frac{1}{2N} \sum_{j=0}^{N-1} [K(x, x_j)\, f(x_j) + K(x, x_{j+1})\, f(x_{j+1})] + g(x) + \frac{h^2\, B_2\, \theta(x)\, V_2}{2} \qquad (6.1.2)$$

avec :

$$x_j = j/N \quad ; \quad |\theta(x)| \leqslant 1 \quad ; \quad V_2 = \operatorname*{Max}_{x,y} \left| \frac{\partial^2}{\partial y^2} [K(x, y)\, f(y)] \right|$$

En fait on donnera à x les valeurs x_i (i = 0, 1, ..., N) et on obtiendra des valeurs approchées f_i pour les $f(x_i)$ en résolvant le système linéaire de N + 1 équations à N + 1 inconnues :

$$f_i = \frac{1}{2N} \sum_{j=0}^{N-1} [K(x_i, x_j)\, f_j + K(x_i, x_{j+1})\, f_{j+1}] + g(x_i) \qquad (6.1.3)$$

Pour simplifier nous faisons l'hypothèse :

$$|K(x, y)| \leqslant k < 1 \qquad (6.1.4)$$

En soustrayant (6.1.2) et (6.1.3) on obtient pour $e_i = f_i - f(x_i)$:

$$e_i = \frac{1}{2N} \sum_{j=0}^{N-1} K(x_i, x_j)\, e_j + K(x_i, x_{j+1})\, e_{j+1}] + h^2\, \frac{B_2\, V_2}{2}\, \theta_i \quad \text{avec } |\theta_i| \leqslant 1 \qquad (6.1.5)$$

qui est un système linéaire de la forme :

$$(I - hK)\, e = \frac{h^2\, B_2\, V_2}{2}\, \theta$$

En adoptant pour les vecteurs la norme $\|e\| = \max_i |e_i|$ on obtient pour les matrices carrées correspondantes (transformations linéaires) :

$$\|A\| = \max_i \left(\sum_{j=0}^{N} |A_{ij}| \right)$$

θ est un vecteur de norme $\| \theta \| \leqslant 1$,

hK est une matrice carrée de norme $\| hK \| \leqslant k < 1$ (d'après 6.1.4).

D'après un théorème classique, on sait que I - hK admet un inverse et que :

$$\| (I - hK)^{-1} \| \leqslant \frac{1}{1 - k}$$

On en déduit :

$$\| e \| = \max_i |e_i| \leqslant \frac{h^2 \ B_2 \ V_2}{2(1 - k)}$$

Théorème 1 : Si l'on suppose que K et g ont des dérivées secondes continues (dérivées partielles pour K), que la solution de (6.1.1) est unique, que K vérifie (6.1.4), la solution approchée par (6.1.3) existe et est unique et l'erreur vérifie :

$$|f_i - f(x_i)| \leqslant h^2 \ K$$

où K est une constante indépendante de i et h. $\left(K = \frac{B_2 \ V_2}{2(1 - k)}\right)$.

Remarque :

L'hypothèse (6.1.4) a été faite pour simplifier la démonstration. On peut en fait démontrer sans utiliser (6.1.4) que pour h assez petit la matrice I - hK est inversible et que la norme de son inverse est bornée par rapport à h (L. V. Kantorovich, Functional Analysis and Applied Mathematics, édité par G. E. Forsythe, page 66).

Théorème 2 : Avec les mêmes hypothèses qu'au théorème 1 on remplace (6.1.3) par :

$$f_i = \frac{1}{2N} \sum_{j=0}^{N-1} [K(x_i, x_j) \ f_j + K(x_i, x_{j+1}) \ f_{j+1}] + g(x_i) + \rho(i, h) \times Ch^2 \qquad (6.1.6)$$

avec $|\rho(i, h)| \leqslant 1$ et C constante. On a encore :

$$|f_i - f(x_i)| \leqslant h^2 \ K^*$$

(K^* indépendant de i et h) ; (il suffit de prendre $K^* = \frac{1}{(1 - k)} \left| \frac{B_2 \ V_2}{2} + C \right|$).

La démonstration est analogue à celle du théorème 1.

Théorème 3 : En plus des hypothèses du théorème 1, on suppose maintenant l'existence et la continuité des dérivées d'ordre 3 et 4 de K et g (dérivées partielles pour K).

L'erreur vérifie alors :

$$e_i = f_i - f(x_i) = h^2 \ e(x_i) + h^4 \ B(i, h)$$

où B est une fonction bornée de i et h et e(x) la solution de l'équation intégrale :

$$e(x) = \int_0^1 K(x, y) \ e(y) \ dy + \frac{B_2}{2!} \left[\frac{\partial}{y} [K(x, y) \ f(y)] \right]_{y=0}^{y=1} \qquad (6.1.7)$$

Démonstration : La solution exacte f(x) vérifie :

$$f(x_i) = \frac{h}{2} \sum_{j=0}^{N-1} \left[K(x_i, x_j) \ f(x_j) + K(x_i, x_{j+1}) \ f(x_{j+1}) \right] + g(x_i) - \frac{h^2}{2!} \ B_2 \ T_1(x_i) - \frac{h^4 \ \theta_i \ B_4}{4!} \ F_4$$

avec $T_1(x) = \left[\frac{\partial}{\partial y} [K(x, y) \ f(y)] \right]_{y=0}^{y=1}$;

$$F_4 = \max_{x,y} \left| \frac{\partial^4}{\partial y^4} \left(K(x, y)\ f(y) \right) \right|$$

En soustrayant avec (6.1.3) on obtient en posant $\overline{e_i} = e_i / h^2$:

$$\overline{e_i} = \frac{h}{2} \sum_{j=0}^{N-1} \left[K(x_i, x_j)\ \overline{e_j} + K(x_i, x_{j+1})\ \overline{e_{j+1}} \right] + \frac{B_2}{2!} T_1(x_i) + \frac{h^2\ \theta_i}{4!} B_4\ F_4$$

qui correspond à la résolution de (6.1.7) par une méthode (6.1.6). On peut appliquer le théorème 2 et on obtient :

$$\left| \frac{e_i}{h^2} - e(x_i) \right| \leqslant h^2\ K^{**}$$

ce qui démontre le théorème 3.

Théorème 4 : En plus des hypothèses du théorème 3 on suppose l'existence et la continuité des dérivées d'ordre 5 et 6 de K et g. L'erreur vérifie alors :

$$e_i = h^2\ e(x_i) + h^4\ \varepsilon(x_i) + h^6\ A(i, h)$$

où A est une fonction bornée de i et h et $\varepsilon(x)$ la solution de l'équation intégrale :

$$\varepsilon(x) = \int_0^1 K(x, y)\ \varepsilon(y)\ dy + \frac{B_2}{2} \left[\frac{\partial}{\partial y} \left(K(x, y)\ e(x) \right) \right]_{y=0}^{y=1} + \frac{B_4}{4!} \left[\frac{\partial^3}{\partial y^3} \left(K(x, y)\ f(y) \right) \right]_{y=0}^{y=1}$$

La démonstration est comparable à celle du théorème 3.

§ 6.2 - EXEMPLE D'EXTRAPOLATION POUR UNE EQUATION INTEGRALE

Considérons l'équation l'intégrale :

$$f(x) = \int_0^1 \frac{\pi}{4} \cos(x + y)\ f(y)\ dy + \pi x$$

La condition (6.1.4) est vérifiée. En plus toutes les dérivées partielles existent jusqu'à un ordre quelconque. Nous cherchons la solution pour une abcisse fixe : x = 0,5 ; solution exacte : f(0,5) = 1,006 856 727.

La solution approchée w(h) obtenue avec un pas h admet un développement pair jusqu'à un ordre arbitraire :

$$w(h) = f(0,5) + d_2\ h^2 + d_4\ h^4 + \ldots + d_{2r-2}\ h^{2r-2} + h^{2r}\ B(h)$$

On emploie le procédé M_r. Nous essayons des pas successifs dans un rapport 2 puis dans un rapport 1,5.

Première expérience ($\alpha = 2$) :

n = 1/h	Solution normale	Solution avec extrapolation	Erreur solution normale $\times 10^9$	Erreur solution avec extrapolation $\times 10^9$
4	1,036 148 056	1,036 148 056	29 291 329	29 291 329
8	1,014 122 635	1,006 780 828	7 265 908	- 75 899
16	1,008 669 689	1,006 856 786	1 812 962	59
32	1,007 309 750	1,006 856 728	453 023	1

Deuxième expérience ($\alpha = 1,5$)

n = 1/h	Solution normale	Solution avec extrapolation	Erreur solution normale $\times 10^9$	Erreur solution avec extrapolation $\times 10^9$
4	1,036 148 056	1,036 148 056	29 291 329	29 291 329
6	1,019 799 918	1,006 721 407	12 943 191	- 135 320
10	1,011 502 579	1,006 857 023	4 645 852	296
16	1,008 669 689	1,006 856 727	1 812 962	0

On constate que le coefficient $\alpha = 1,5$ conduit à la même précision que $\alpha = 2$ mais avec un coût moindre.

Pour atteindre une précision 10^{-9} sans extrapolation, (solution normale), on serait amené à résoudre un système linéaire de dimension excessive.

§ 6.3 - ETUDE DU COMPORTEMENT DE LA SOLUTION APPROCHEE POUR L'EQUATION AUX DERIVEES PARTIELLES DE POISSON

Considérons le problème suivant :

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y) \quad \text{sur le carré unitaire } [0,1] \times [0,1] \tag{6.3.1}$$

$u(x, y) = \varphi(x, y)$ sur le contour du carré.

Supposons que la solution soit quatre fois continûment différentiable et notons G_4 le maximum sur le carré des dérivées partielles d'ordre 4 (en valeur absolue). Pour trouver une valeur approchée u_{ij} de la solution $u(x_i, x_j)$, $\left(x_k = \frac{k}{N}\right)$ on résoud les équations suivantes :

$$- u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1} + 4u_{ij} + h^2 f(x_i, x_j) = h^4 \theta_{i,j,n} C \tag{6.3.2}$$

où $|\theta_{i,j,n}| \leq 1$ et C constante.

En adaptant un raisonnement fait par Kantorowitsch et Krylow, (Näherungsmethoden der höheren Analysis, (1956) pp 220 - 225) on trouve que l'erreur :

$$e_{n,m} = u_{n,m} - u(x_n, x_m)$$

vérifie pour h assez petit :

$$|e_{n,m}| \leq h^2 K \tag{6.3.3}$$

où K est une constante indépendante de n, m et h.

Exactement on a :

$$|e_{n,m}| \leq h^2 \frac{[G_4/6 + C] N}{6 - h^2 (N_4 + 6C)}$$

Si w est la solution de $\Delta w = -1$ avec $w = 0$ sur le contour, on a posé $N = \max_{x,y} |w(x, y)|$ et N_4 le maximum des dérivées partielles d'ordre 4 sur le carré (en valeur absolue).

Supposons maintenant que la solution u(x, y) soit six fois continûment différentiable.

En posant $\overline{e_{m,n}} = e_{m,n}/h^2$ on obtient :

$$- \overline{e_{i+1,j}} - \overline{e_{i-1,j}} - \overline{e_{i,j+1}} - \overline{e_{i,j-1}} + 4\overline{e_{i,j}} + h^2 \left[- \frac{1}{12} \left(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4} \right) \right] = h^4 \theta_{ij} \frac{G_6}{180} \qquad (6.3.4)$$

donc, si l'on appelle e(x, y) la solution de :

$$\frac{\partial^2 e}{\partial x^2} + \frac{\partial^2 e}{\partial y^2} = - \frac{1}{12} \left[\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4} \right]$$

avec e = 0 sur le contour et en supposant cette solution quatre fois continûment différentiable on a :

$$\left| \frac{e_{m,n}}{h^2} - e(x_m, x_n) \right| \leqslant h^2 H \qquad (6.3.5)$$

pour h assez petit, où H est une constante indépendante de m, n, h.

On a donc :

$$e_{m,n} = h^2 e(x_m, x_n) + h^4 B(m, n, h) \quad \text{avec} \quad |B(m, n, h)| \leqslant H \qquad (6.3.6)$$

On peut obtenir des développements d'ordre plus élevé en faisant les hypothèses correspondantes.

§ 6.4 - EXEMPLE D'EXTRAPOLATION POUR L'EQUATION DE LAPLACE

Exemple n° 1 :

Nous prenons le problème (6.3.1) avec f(x, y) = 0, le procédé (6.3.2) avec C = 0 et u(x, y) = sin (x) e^y + sin (3x) e^{3y} sur le contour. On cherche u (0,5 ; 0,5) = 5,260 90.

Deux valeurs du rapport α entre les pas successifs ont été essayées.

Première expérience : (α = 2)

n = 1/h	Nombre d'équations	Solution normale	Solution avec extrapolation	Erreur solution normale × 10^5	Erreur solution avec extrapolation × 10^5
10	81	5,304 78	5,304 78	4 388	4 388
20	361	5,271 91	5,260 95	1 101	5
40	1 521	5,263 57	5,260 78	267	- 12

Deuxième expérience : (α = 1,4)

n = 1/h	Nombre d'équations	Solution normale	Solution avec extrapolation	Erreur solution normale × 10^5	Erreur solution avec extrapolation × 10^5
10	81	5,304 78	5,304 78	4 388	4 388
14	169	5,283 36	5,261 05	2 246	15
20	361	5,271 91	5,260 86	1 101	- 4
28	729	5,266 49	5,260 82	559	- 8

On constate qu'à partir de la deuxième ou troisième ligne, les résultats avec extrapolation ne s'améliorent plus. Cela provient du fait que l'erreur de résolution du système linéaire affecte déjà le sixième chiffre des résultats de la deuxième colonne. On a employé la méthode de Gauss-Seidel en abaissant le résidu maximum en dessous de 10^{-6} (on ne peut pas faire beaucoup mieux avec la représentation - machine des nombres utilisés ici).

L'extrapolation, qui concerne *l'erreur de discrétisation*, ne peut abaisser *l'erreur totale* en dessous de *l'erreur de résolution du système linéaire*.

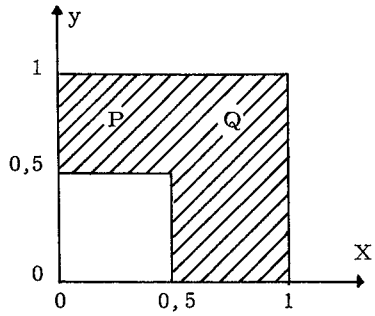
Troisième expérience : ($\alpha = 1,4$)

n = 1/h	Nombre d'équations	Solution normale	Solution avec extrapolation	Erreur solution normale $\times 10^5$	Erreur solution avec extrapolation $\times 10^5$
2	1	6,208 79	6,208 79	94 789	94 789
4	9	5,525 19	5,297 33	26 429	3 643
6	25	5,381 23	5,262 14	12 033	124
8	49	5,329 18	5,260 93	6 828	3
12	121	5,291 44	5,260 88	3 054	- 2

On remarque qu'en partant d'un pas grossier ($h = 0,5$) on a obtenu un excellent résultat en moins de temps. Les coûts des trois expériences sont proportionnels aux nombres 21, 7, 1.

On ne peut atteindre le sixième chiffre significatif à cause de l'erreur de résolution de système linéaire (notons que celle-ci est devenue plus faible, les systèmes étant de dimension moindre). Si l'on voulait atteindre la précision 10^{-6} par résolution normale, sans extrapolation, on serait amené à prendre un pas très petit, c'est-à-dire à résoudre un système linéaire de dimension considérable.

Exemple n° 2 : Equation de Laplace sur un coude.



Conditions sur le contour :

$$u(x, y) = 0,2\ [e^{\pi x} \sin \pi y + e^{\pi x y} \sin \pi x]$$

Point P : $u\ (0,25\ ;\ 0,75) = 1,802\ 274$

Point Q : $u\ (0,75\ ;\ 0,75) = 2,984\ 195$

Point P :

n = 1/h	Nombre d'équations	Solution normale	Solution avec extrapolation	Erreur solution normale $\times 10^6$	Erreur solution avec extrapolation $\times 10^6$
4	5	1,850 307	1,850 307	48 033	48 033
8	33	1,815 111	1,803 377	12 837	1 103
16	161	1,805 553	1,802 300	3 279	26
32	705	1,803 095	1,802 270	821	- 4

§ 6.4

Point Q :

n = 1/h	Nombre d'équations	Solution normale	Solution avec extrapolation	Erreur solution normale $\times 10^6$	Erreur solution avec extrapolation $\times 10^6$
4	5	3,070 424	3,070 424	86 229	86 229
8	33	3,007 103	2,985 995	22 908	1 800
16	161	2,990 037	2,984 237	5 842	43
32	705	2,985 659	2,984 190	1 464	- 5

Remarques :

On a traité dans ce chapitre des problèmes de type "conditions aux limites". Il est souvent difficile pour ces problèmes d'introduire des opérateurs de discrétisation très puissants ; par ailleurs le coût de calcul augmente beaucoup quand le pas diminue (Résolution d'un système linéaire de $\left(\frac{1}{h} - 1\right)^2$ équations pour l'exemple du problème de Laplace).

Il peut donc être très intéressant, si le comportement de la solution le permet, de prendre des pas grossiers et de faire une extrapolation. Notons que la méthode ne s'applique pas facilement si les domaines ne sont pas du "type rectangulaire" (au sens large).

CHAPITRE 7

ÉVALUATION D'INTÉGRALES PAR UNE MÉTHODE DE MONTE-CARLO APPLICATION DU PROCÉDÉ DE RICHARDSON

§ 7.1 - INTRODUCTION

On désigne par "méthode de Monte-Carlo" un certain nombre de procédés de calcul qui peuvent en fait être très différents ; leur seul point commun semble être qu'ils utilisent des nombres aléatoires. Ces procédés sont appelés suivant les cas : simulation, échantillonnage ou méthode de Monte-Carlo. La distinction n'est pas toujours bien établie entre ces trois termes dont les sens se recouvrent en partie. Cela est sans doute dû au fait que Von Neumann, Ulam et leurs successeurs n'ont pas inventé la théorie de l'échantillonnage mais lui ont ouvert, sous le nom de Monte-Carlo, un champ d'application nouveau et très vaste, entraînant ainsi des développements importants de l'ancienne théorie. De nombreux problèmes de la physique (surtout atomique) dont la nature à l'origine, était probabiliste, ont été décrits d'abord à l'aide d'équations mathématiques parfois très compliquées. L'innovation a consisté principalement à inverser cette démarche habituelle qui ramène les problèmes de probabilité à des équations mathématiques. En effet, l'apparition des calculateurs électroniques a permis de *simuler* en machine le processus aléatoire tel que la physique le fournit, et même de résoudre numériquement certains problèmes mathématiques où n'intervenait pas le hasard, au moyen d'une analogie probabiliste. Ces nouvelles perspectives ont entraîné un développement considérable des techniques de réduction de variance : certaines étaient déjà connues dans leur principe mais la théorie de l'échantillonnage n'en faisait pas un emploi aussi systématique (cela ne justifie toutefois pas les noms nouveaux donnés à ces procédés) ; d'autres sont effectivement nouvelles et dûes aux propriétés spéciales des populations considérées.

On appelle "simulation" la simple reproduction d'un phénomène physique dans une machine à calculer. On parle plutôt de "méthode de Monte-Carlo" lorsque le phénomène physique a été fortement modifié pour améliorer la précision des résultats et a fortiori lorsque le processus aléatoire mis en machine n'a d'autre rapport avec le problème proposé que de conduire numériquement à sa solution (ce qui est rare en pratique).

En résumé, ce sont donc le recours à certaines analogies mathématiques, l'utilisation systématique des procédés de réduction de variance et le genre de problèmes traités qui constituent l'originalité des méthodes de Monte-Carlo.

Dans les deux chapitre qui suivent, nous considèrerons le calcul d'intégrales et la résolution d'équations intégrales à l'aide d'une variable aléatoire dont on sait construire des réalisations dans le calculateur et dont la moyenne est égale à la solution cherchée. Pour réduire la variance nous ne ferons pas appel à l'origine physique du problème, que nous ne connaissons pas, mais nous utiliserons plutôt les méthodes habituelles de l'analyse numérique : on appliquera en fait les procédés d'extrapolation du chapitre 1.

§ 7.2 - QUELQUES PROCEDES DE REDUCTION DE VARIANCE

Soit f une fonction numérique, intégrable Riemann sur l'intervalle $[0,1]$. Une variable aléatoire, de variance finie et dont la moyenne est égale à $I = \int_0^1 f(x)\,dx$ est appelée estimateur de I. Dans la suite, ξ ou ξ_i désigneront des variables aléatoires indépendantes uniformément distribuées sur l'intervalle $[0,1]$; si $|f|$ et f^2 sont intégrables Riemann, $f(\xi)$ est un estimateur de I dont la variance vaut $v = \int_0^1 (f(x) - I)^2\,dx$.

La méthode de Monte-Carlo élémentaire consiste à prendre n ξ_i indépendants et à former $\frac{1}{n}\sum_{i=1}^{n} f(\xi_i)$ qui est un estimateur de I de variance $\frac{V}{n}$.

Cette méthode peut être considérée comme un échantillonnage de population : considérons N points équidistants de l'intervalle [0,1] ; N est grand : par exemple 10^p, p étant le nombre de décimales utilisées pour exprimer ξ. On a une population de N éléments, la caractéristique du $j^{ème}$ élément étant $X_j = f\left(\frac{j}{N}\right)$.

On veut estimer :

$$M = \frac{1}{N}\sum_{j=1}^{N} X_j \simeq I$$

à l'aide d'un échantillon de n éléments. Les techniques suivantes, utilisées en théorie de l'échantillonnage peuvent donc être transposées : Cochran [5] Deming [8].

α) Echantillonnage stratifié :

On divise l'intervalle [0,1] en S sous-intervalles ou strates de longueur l_h (h = 1, 2, ..., S). Dans chaque strate on tire n_h abcisses $x_{i,h}$ au hasard (distribution uniforme sur la strate considérée).

$$\sum_{h=1}^{S} n_h = n$$

Appelons :

$$\mathcal{Y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} f(x_{i,h})$$

la moyenne des ordonnées dans la $h^{ème}$ strate. L'estimateur stratifié sera alors :

$$\mathcal{Y}_{ST} = \sum_{h=1}^{S} l_h \mathcal{Y}_h$$

Supposons pour simplifier que l'on ait pris $n_h = l_h \times n$.

On a alors le théorème suivant :

$$\frac{V}{n} - \mathrm{Var}\,(\mathcal{Y}_{ST}) = \frac{1}{n}\sum_{h=1}^{S} l_h\,(Y_h - I)^2 \qquad (7.2.1)$$

où Y_h désigne la moyenne de f dans la hème strate.

Cela signifie que l'échantillonnage stratifié est intéressant quand on peut grouper dans une même strate des éléments ayant une caractéristique comparable. Si les strates constituent une image restreinte mais fidèle de la population totale on gagnera peu. Dans notre cas, si f est continue, on peut prévoir que la stratification sera avantageuse si les strates sont suffisamment petites.

β) Echantillonnage pondéré :

Cette technique a pris le nom d'"Importance Sampling", en anglais, pour la méthode de Monte-Carlo. Soit g une fonction densité pour l'intervalle [0,1] :

$$\int_0^1 g(x)\,dx = 1 \quad ; \quad g(x) > 0$$

et X une variable aléatoire admettant g comme fonction densité :

$Z = \frac{f(X)}{g(X)}$ est une variable aléatoire telle que :

$$E(Z) = I \; ; \; \mathrm{Var}\,(Z) = \int_0^1 \left(\frac{f(x)}{g(x)} - I\right)^2 g(x)\,dx \qquad (7.2.2)$$

On choisit g pour diminuer Var (Z). Le choix optimum, impossible à réaliser en pratique est de

prendre g proportionnel à $|f|$. Quand $f > 0$, $g = \dfrac{f}{\int_0^1 f(x)\,dx}$ conduit à une variance nulle. On essayera de se rapprocher du choix optimum. Cette technique s'étend à presque tous les problèmes traités par la méthode de Monte-Carlo (pour les équations différentielles, les équations aux dérivées partielles et les systèmes d'équations linéaires voir Curtiss (*)). En revanche son emploi est difficile : elle nécessite une étude préalable et peut conduire pour de mauvais choix de g à une augmentation de la variance.

γ) Méthode de la variable de contrôle : (Fieller and Hartley [11])

C'est en fait la méthode de régression de l'échantillonnage. Supposons que l'on puisse trouver une variable aléatoire auxiliaire C de moyenne connue J et telle que le coefficient de corrélation p entre $Y = f(\xi)$ et C soit voisin de + 1 ou - 1 ; alors on pourra utiliser l'estimateur :

$$Y_R = \bar{Y} + b\,(J - \bar{C})$$

avec
$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i \tag{7.2.3}$$

(Y_i et C_i sont des réalisations des variables aléatoires Y et C). On peut choisir b a priori : dans ce cas on a $E(Y_R) = I$ rigoureusement. Si l'on prend b en fonction des Y_i et C_i on introduit un léger biais qui est négligeable quand n est suffisamment grand. En supposant que le calcul de Y_R coûte k fois le calcul de Y, pour que Y_R soit avantageux il faut :

$$\mathrm{Var}\,(Y_R) \leqslant \frac{\mathrm{Var}\,(\bar{Y})}{k}$$

c'est-à-dire :

$$v + b^2 w - 2bp\sqrt{v\,w} \leqslant \frac{v}{k}$$

où w est la variance de C.

Le terme de gauche est minimum quand :

$$b = p\sqrt{\frac{v}{w}} = \frac{E[(C - J)(Y - I)]}{E[(C - J)^2]}$$

Avec ce b optimum, la méthode de régression est avantageuse quand :

$$|p| > \sqrt{1 - \frac{1}{k}}$$

Par ailleurs, pour qu'un certain b laisse un domaine possible en p pour lequel Y_R est avantageux, il faut :

$$\left(1 - \frac{1}{\sqrt{k}}\right)\sqrt{\frac{v}{w}} < b < \left(1 + \frac{1}{\sqrt{k}}\right)\sqrt{\frac{v}{w}}$$

Un mauvais choix de b peut rendre cette méthode moins efficace que la méthode simple.

Pour approcher le b optimum on peut prendre :

$$\tilde{b} = \frac{\sum_j (C_j - \bar{C})(Y_j - \bar{Y})}{\sum_j (C_j - \bar{C})^2}$$

(*) Sampling methods applied to differential and différence equation. Seminar on scientific computation. I B M corp. (1949).
Monte-Carlo Methods for the iteration of linear operators. J. of Math. and Phys. Vol 32 n° 4 pp 209 - 232 (1954).

Remarquons qu'en posant $Y_i = a + b\,C_i + E_i$ et en cherchant les a et b (méthode des moindres carrés) qui minimisent $\sum_i E_i^2$ on trouve pour b l'expression précédente. En pratique, la variable C est difficile à trouver et le gain assez faible.

La méthode de la variable de contrôle est peu utilisable pour les méthodes de Monte-Carlo.

δ) Méthode des variables compensées : (Antithetic variates, Hammerley and Morton [17], Morton [39])

Soient Y^1 et Y^2 deux variables aléatoires de même moyenne I, de variances v_1 et v_2 et de coefficient de corrélation p. On forme le nouvel estimateur de I :

$$Y_A = b_1 Y^1 + b_2 Y^2 \quad \text{avec } b_1 + b_2 = 1 \tag{7.2.4}$$

L'estimateur Y_A sera avantageux si p est suffisamment proche de - 1.

Plus précisément, si Y_A coûte k fois Y^1 il faut avoir :

$$b_1^2 v_1 + b_2^2 v_2 + 2 b_1 b_2 p \sqrt{v_1 v_2} \leqslant \frac{v_1}{k}$$

Le membre de gauche est minimum pour :

$$\begin{cases} b_1 = \dfrac{1 - p\sqrt{\lambda}}{1 + \lambda - 2p\sqrt{\lambda}} \\ b_2 = \dfrac{\lambda - p\sqrt{\lambda}}{1 + \lambda - 2p\sqrt{\lambda}} \\ \lambda = \dfrac{v_1}{v_2} \end{cases}$$

Et la condition de rentabilité devient :

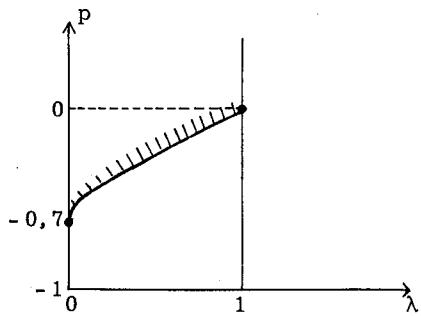
$$\frac{(1 - p^2)}{1 + \lambda - 2p\sqrt{\lambda}} < \frac{1}{k}$$

On peut supposer $\lambda \leqslant 1$: Si $\lambda > 1$, on comparerait plutôt Y_A à Y^2 dont la variance serait inférieure à celle de Y^1, d'où permutation des roles de Y^1 et Y^2, ce qui ramène à un λ inférieur à 1. On trouve finalement la condition :

$$-1 \leqslant p \leqslant \frac{1}{k}\left[\sqrt{\lambda} - \sqrt{(k - 1)(k - \lambda)}\right] \tag{7.2.5}$$

Dans le cas raisonnable où k = 2, λ = 1 il suffit que p soit négatif pour que Y_A soit avantageux.

Pour k = 2 le point (λ, p) doit se trouver sous la courbe suivante :



Si f est une fonction croissante, $Y^1 = f(\xi)$ et $Y^2 = f(1 - \xi)$ fournissent un exemple simple de telles variables compensées. Bien que cette méthode semble plus efficace que la méthode de la variable de contrôle, elle n'est pas d'un emploi facile : même avec une étude préalable on ne trouve pas souvent deux variables compensées efficaces.

Un procédé, introduit par Hammersley et Morton [17] se rattache à la même idée de compensation :

$$Y_{AT} = \alpha f(\alpha\,\xi) + (1 - \alpha)\, f(1 - (1 - \alpha)\,\xi) = T_\alpha\, f(\xi) \tag{7.2.6}$$

Pour certaines fonctions (monotones, par exemple) Y_{AT} peut être avantageux avec un coefficient α convenable. La détermination du α qui rend var (Y_{AT}) minimum est trop compliquée : on peut se contenter de rendre $T_\alpha\, f(x)$ aussi constant que possible ; par exemple :

$$T_\alpha\, f(0) = T_\alpha\, f(1) \quad \text{soit } f(\alpha) = (1 - \alpha)\, f(1) + \alpha\, f(0) \tag{7.2.7}$$

Les auteurs envisagent également la variable aléatoire :

$$Y_{SY} = \alpha\, f(\alpha\,\xi) + (1 - \alpha)\, f(\alpha + (1 - \alpha)\,\xi) = S_\alpha\, f(\xi) \tag{7.2.8}$$

avec α tel que $S_\alpha\, f(1) = S_\alpha\, f(0)$ soit :

$$(1 - 2\alpha)\, f(\alpha) = (1 - \alpha)\, f(1) - \alpha\, f(0) \tag{7.2.9}$$

La détermination des α qui satisfont (7.2.7) ou (7.2.9) n'étant pas toujours facile et même possible, on conçoit que l'emploi successif des opérateurs T_α et S_α, par exemple $T_\beta\, T_\alpha$ ou $T_\beta\, S_\alpha$, soit encore plus délicat. Néanmoins, on constate la propriété suivante : Si α est tel que (7.2.7) est vérifiée alors :

$$S_{1/2}\, T_\alpha\, f(1) = S_{1/2}\, T_\alpha\, f(0)$$

et par conséquent :

$$S^m_{1/2}\, T_\alpha\, f(1) = S^m_{1/2}\, T_\alpha\, f(0) \tag{7.2.10}$$

Si l'on pose :

$$U_n\, f(x) = \frac{1}{n} \sum_{j=0}^{n-1} f\left(\frac{x + j}{n}\right) \tag{7.2.11}$$

$S^m_{1/2}$ représente en fait U_n avec $n = 2^m$. Cela suggère l'idée suivante, qui sera la base du paragraphe 7.3 : appliquer l'opérateur U_n sur une fonction f^* transformée de f d'une manière convenable. $T_\alpha f$ est un exemple simple de f^* ; la seule exigence est de conserver $E(f^*(\xi)) = E(f(\xi)) = I$.

D'autres techniques ont été proposées. Citons le procédé de Roulette russe et éclatement (Splitting) [21], surtout employé dans les problèmes de transport de particules, l'utilisation de valeurs moyennes etc. Toutes ces méthodes demandent une étude préalable de la fonction ; nous développons dans le paragraphe suivant un procédé qui n'a pas cette exigence mais qui suppose en revanche l'existence des premières dérivées.

§ 7.3 - ECHANTILLONNAGE SYSTEMATIQUE SUR UNE FONCTION MODIFIEE

L'intégrale sur [0,1] de la fonction $U_n\, f$ définie en (7.2.11) vaut I. La variable aléatoire $U_n\, f(\xi)$ est donc un estimateur de I.

Reprenant les notations du début du § 7.2, on a divisé la population de N éléments en n strates égales. Appelons X_{ij} le $j^{ème}$ élément de la $i^{ème}$ strate, $X_{ij} = f\left(\frac{i}{n} + \frac{j}{N}\right)$, $i = 0, 1, 2, \ldots, n - 1$; $j = 0, 1, \ldots, \frac{N}{n} - 1$. L'estimateur $U_n\, f(\xi)$ consiste à choisir j au hasard de 0 à $\frac{N}{n} - 1$ et à prendre $Y_{SY} = \frac{1}{n} \sum_{i=0}^{n-1} X_{ij}$. Ce procédé est utilisé en échantillonnage classique sous le nom *d'échantillonnage*

systématique, Cochran [5]. On peut donc transposer tous les théorèmes qui s'y rapportent. Nous n'en citerons qu'un :

L'échantillonnage systèmatique conduit à une variance plus faible que l'échantillonnage simple (à nombre égal de calculs de f) si :

$$\frac{1}{n-1}\int_0^1 \sum_{j=0}^{n-1} \left(f\left(\frac{x+j}{n}\right) - U_n\, f(x)\right)^2 dx > \int_0^1 (f(x) - I)^2\, dx \qquad (7.3.1)$$

(On peut également comparer à l'échantillonnage stratifié).

Nous nous proposons d'exprimer la variance de $U_m\, f(\xi)$ sous forme d'un développement limité suivant les puissances croissantes de $\frac{1}{m}$ en utilisant la formule de sommation d'Euler-Mac-Laurin (Fort [13], Kuntzmann [26]) :

$$U_m\, f(x) = \frac{1}{m}\sum_{j=0}^{m-1} f\left(\frac{x+j}{m}\right)$$

$$= \int_0^1 f(x)\, dx + \sum_{\nu=1}^{n} \frac{B_\nu(x)\, \Delta f^{\nu-1}}{\nu!\, m^\nu} - R_n(x) \qquad (7.3.2)$$

$$R_n(x) = \int_0^1 \frac{\bar{B}_n(x-t)}{n!\, m^{n+1}} \left[\sum_{j=0}^{m-1} f^{(n)}\left(\frac{t+j}{m}\right)\right] dt$$

avec :

$$\Delta f^\nu = f^{(\nu)}(1) - f^{(\nu)}(0)$$

$B_\nu(x)$: polynôme de Bernoulli de degré ν

$\bar{B}_\nu(x)$: fonction de période 1 coïncidant avec $B_\nu(x)$ sur l'intervalle semi ouvert $[0,1[$, (Favard [9], [10]).

Cette formule suppose l'existence des n premières dérivées de f sur l'intervalle $[0,1]$.

Comme $f^{(n)}$ est continue sur $[0,1]$ on a :

$$\frac{1}{m}\sum_{j=0}^{m-1} f^{(n)}\left(\frac{t+j}{m}\right) = \int_0^1 f^{(n)}(t)\, dt + \varepsilon_m(t)$$

ε_m convergeant uniformément vers 0 quand m tend vers l'infini. On a donc :

$$R_n(x) = \int_0^1 \frac{\bar{B}_n(x-t)}{n!\, m^n} \left(\int_0^1 f^{(n)}(t)\, dt + \varepsilon_m(t)\right) dt$$

D'après les propriétés des polynômes de Bernoulli, le premier terme de la somme est nul et on trouve :

$$R_n(x) = \frac{1}{m^n}\int_0^1 \frac{\bar{B}_n(x-t) \times \varepsilon_m(t)}{n!}\, dt = \frac{\varepsilon_m^*(x)}{m^n} \qquad (7.3.3)$$

ε_m^* convergeant uniformément vers 0 quand m tend vers l'infini.

D'après (7.3.2) et (7.3.3) on a :

$$\mathrm{Var}\ [U_m\, f(\xi)] = \int_0^1 \left[\sum_{\nu=1}^{n} \frac{B_\nu(x)\, \Delta f^{\nu-1}}{\nu!\, m^\nu} - \frac{\varepsilon_m^*(x)}{m^n}\right]^2 dx$$

on sait que :

$$\int_0^1 B_\mu(x)\, B_\nu(x)\, dx = (-1)^{\nu-1} \frac{\nu!\, \mu!\, B_{\mu+\nu}}{(\mu+\nu)!}\ ,\ \mu \geq 1\ ,\ \nu \geq 1 \qquad (7.3.4)$$

B_i désignant le ième nombre de Bernoulli, $B_i = B_i(0)$.

(La dissymétrie entre μ et ν dans le membre de droite n'est qu'apparente ; on peut remplacer $(-1)^{\nu-1}$ par $(-1)^{\mu-1}$, car si μ et ν sont de parité différente, $\mu+\nu$ est impair (avec $u+\nu>2$) donc $B_{\mu+\nu}=0$).

On trouve finalement (en supposant n pair) :

$$\text{Var}\,[U_m\, f(\xi)] = \sum_{\nu,\mu=1}^{\mu+\nu\leq n} \frac{(-1)^{\nu-1}\,\Delta f^{\nu-1}\,\Delta f^{\mu-1}\,B_{\mu+\nu}}{m^{\nu+\mu}\,(\mu+\nu)!} + \frac{1}{m^{n+1}}\,\varepsilon_m^+ \tag{7.3.5}$$

où ε_m^+ est un nombre qui tend vers 0 quand m tend vers l'infini. Explicitons pour n = 8 :

$$\begin{aligned}\text{Var}\,[U_m\, f(\xi)] = {} & \frac{(\Delta f^{(0)})^2}{12\,m^2} + \frac{(\Delta f^{(1)})^2 - \Delta f^{(0)}\,\Delta f^{(2)}}{720\,m^4} + \frac{(\Delta f^{(2)})^2 - 2\,\Delta f^{(1)}\,\Delta f^{(3)} + 2\,\Delta f^{(0)}\,\Delta f^{(4)}}{30\,240\,m^6} \\ & + \frac{(\Delta f^{(3)})^2 - 2\,\Delta f^{(2)}\,\Delta f^{(4)} + 2\,\Delta f^{(1)}\,\Delta f^{(5)} - 2\,\Delta f^{(0)}\,\Delta f^{(6)}}{1.209.600\,m^8} + \frac{1}{m^9}\,\varepsilon_m^+\end{aligned} \tag{7.3.6}$$

Il est alors visible que pour diminuer la variance de $U_m f(\xi)$, (ou au moins pour obtenir qu'elle tende vers 0 comme l'inverse d'une puissance plus élevée de m) il faut effectuer une transformation préalable de la fonction f qui n'altère pas la valeur de l'intégrale et telle que la nouvelle fonction g vérifie $\Delta g^{(j)} = 0$ jusqu'à un rang en j aussi élevé que possible.

On rappelle que $g = T_\alpha f$ défini par (7.2.6) avec α choisi tel que (7.2.7) soit satisfait vérifie $\Delta g^{(0)} = 0$.

Nous essayerons plutôt de trouver des transformations qui soient indépendantes de la fonction f.

Nous allons utiliser la propriété suivante (Halton and Handscomb [16]) :

$$\begin{aligned}\Delta\,(U_s\ f)^{(r)} &= \frac{1}{s^{r+1}}\,\Delta f^{(r)} \\ \Delta\,(U_s\ k)^{(r)} &= (-1)^{r+1}\frac{1}{s^{r+1}}\,\Delta f^{(r)} \quad \text{si } k(x) \equiv f(1-x)\end{aligned} \tag{7.3.7}$$

Or, $U_s f$ comme $U_s k$ sont des fonctions dont l'intégrale sur [0,1] vaut I quel que soit s ; il en sera de même pour la combinaison linéaire $g = a\,U_s f + b\,U_t f$ si $a + b = 1$. On peut choisir a et b pour que :

$$\Delta g^{(r)} = 0$$

soit :

$$\frac{a}{s^{r+1}}\,\Delta f^{(r)} + \frac{b}{t^{r+1}}\,\Delta f^{(r)} = 0$$

On obtient :

$$g = (s^{r+1}\,U_s\ f - t^{r+1}\,U_t\ f) \,/\, (s^{r+1} - t^{r+1})$$

Il faut remarquer que si l'on avait $\Delta f^{(r')} = 0$ on a toujours après transformation $\Delta g^{(r')} = 0$. Il est donc possible d'appliquer le procédé plusieurs fois de suite. Considérons par exemple :

$$S^r f = \frac{2^{r+1}\,U_2 f - f}{2^{r+1} - 1} \tag{7.3.8}$$

$$\Delta(S^r f)^{(r)} = 0 \quad \text{et} \quad \Delta(S^r f)^{(r')} = \frac{2^{r-r'} - 1}{2^{r+1} - 1}\,\Delta f^{(r')}$$

L'application successive d'opérateurs S^r permet d'obtenir une fonction $S_n f$ ayant la même intégrale que f et telle que $\Delta (S_n f)^{(j)} = 0$ pour $j < n - 1$:

$$S_n f = S^{n-2}\ S^{n-3}\ \ldots\ S^1\ S^0\ f \tag{7.3.9}$$

D'une manière analogue on a, d'après (7.3.7) :

$$D^s f = \frac{1}{2} (U_s f + U_s k)$$

$$\Delta (D^s f)^{(t)} = \begin{cases} 0 \text{ si } t \text{ est pair} \\ \frac{1}{s^{t+1}} \Delta f^{(t)} \quad \text{si } t \text{ est impair} \end{cases} \tag{7.3.10}$$

(On n'utilisera en pratique que $D^1 f = D f$).

En combinant les opérateurs S^r et D on obtient :

$$F_n f = S^{2n-3} \ldots\ldots S^3 S^1 D f$$
$$\Delta (F_n f)^{(j)} = 0 \quad \text{pour } j < 2n - 1 \tag{7.3.11}$$

On vérifie aisément que la fonction $S_n f$ est en fait de la forme :

$$S_n f = \sum_{h=1}^{n} \lambda_h^n U_{2^{h-1}} f \tag{7.3.12}$$

où les λ_h^n sont des constantes convenables.

De même $F_n f$ s'écrit :

$$F_n f = \sum_{h=1}^{n} \mu_h^n \times \frac{(U_{2^{h-1}} f + U_{2^{h-1}} k)}{2} \tag{7.3.13}$$

Il n'est pas du tout obligatoire de prendre des combinaisons de $U_m f$ où m est une puissance de 2 : Posons :

$$V_n f = \sum_{h=1}^{n} \theta_h^n U_{m_h} f \tag{7.3.14}$$

où les m_h sont des entiers distincts rangés par ordre croissant.

Pour que l'intégrale de $V_n f$ soit égale à I il faut :

a) $$\sum_{k=1}^{n} \theta_k^n = 1$$

En utilisant (7.3.7) on voit que les conditions :

b) $$\Delta (V_n f)^{(j)} = 0 \quad (j = 0, 1, \ldots, n - 2)$$

aboutissent à un système linéaire de n - 1 équations en θ_k^n indépendantes de la fonction f. Le système a) + b) admet toujours une solution unique.

De même on peut trouver les φ_h^n pour que :

$$W_n f = \sum_{h=1}^{n} \varphi_h^n \frac{(U_{m_h} f + U_{m_h} k)}{2} \quad \text{vérifie}$$
$$\Delta (W_n f)^{(j)} = 0 \quad (j = 0, 1, \ldots, 2n - 2) \tag{7.3.15}$$

Dans le cas particulier où $m_h = h$ on obtient les premières fonctions suivantes ([29], [30]) :

$$V_2^* f(x) = -\left[f(x) \right] + \left[f\left(\frac{x}{2}\right) + f\left(\frac{x+1}{2}\right) \right]$$
$$V_3^* f(x) = \frac{1}{2}\left[f(x) \right] - 2 \left[f\left(\frac{x}{2}\right) + f\left(\frac{x+1}{2}\right) \right] + \frac{3}{2} \left[f\left(\frac{x}{3}\right) + f\left(\frac{x+1}{3}\right) + f\left(\frac{x+2}{3}\right) \right] \tag{7.3.16}$$
$$V_4^* f(x) = -\frac{1}{6}\left[f(x) \right] + 2 \left[f\left(\frac{x}{2}\right) + f\left(\frac{x+1}{2}\right) \right] - \frac{9}{2} \left[f\left(\frac{x}{3}\right) + f\left(\frac{x+1}{3}\right) + f\left(\frac{x+2}{3}\right) \right]$$
$$+ \frac{8}{3} \left[f\left(\frac{x}{4}\right) + f\left(\frac{x+1}{4}\right) + f\left(\frac{x+2}{4}\right) + f\left(\frac{x+3}{4}\right) \right]$$

$$W_1^* f(x) = \frac{1}{2}\left[f(x) + f(1 - x) \right]$$

$$W_2^* f(x) = -\frac{1}{6}\left[f(x) + f(1 - x) \right] + \frac{1}{3}\left[f\left(\frac{x}{2}\right) + f\left(\frac{x+1}{2}\right) + f\left(\frac{1-x}{2}\right) + f\left(\frac{2-x}{2}\right) \right]$$

$$W_3^* f(x) = \frac{1}{48}\left[f(x) + f(1 - x) \right] - \frac{4}{15}\left[f\left(\frac{x}{2}\right) + f\left(\frac{x+1}{2}\right) + f\left(\frac{1-x}{2}\right) + f\left(\frac{2-x}{2}\right) \right]$$

$$+ \frac{27}{80}\left[f\left(\frac{x}{3}\right) + f\left(\frac{x+1}{3}\right) + f\left(\frac{x+2}{3}\right) + f\left(\frac{1-x}{3}\right) + f\left(\frac{2-x}{3}\right) + f\left(\frac{3-x}{3}\right) \right] \qquad (7.3.17)$$

$$W_4^* f(x) = -\frac{1}{720}\left[f(x) + f(1 - x) \right] + \frac{4}{45}\left[f\left(\frac{x}{2}\right) + \ldots \right]$$

$$- \frac{243}{560}\left[f\left(\frac{x}{3}\right) + \ldots \right] + \frac{128}{315}\left[f\left(\frac{x}{4}\right) + \ldots \right]$$

§ 7.4 - EXTRAPOLATION APPLIQUEE A UNE FORMULE DE QUADRATURE D'ABCISSES ALEATOIRES

La quantité $U_m f(x)$ est en fait le résultat de la formule de quadrature approchée dite "des rectangles", d'abcisse x et appliquée avec un pas $h = \frac{1}{m}$ pour calculer $I = \int_0^1 f(x)\, dx$.

L'estimateur $U_m f(\xi)$ est donc la formule des rectangles appliquée avec une abcisse aléatoire. C'est ce point de vue qui sera étudié dans ce paragraphe.

Posons :

$$U_{m,x} f = \frac{1}{m} \sum_{j=0}^{m-1} f\left(\frac{x+j}{m}\right) \qquad (7.4.1)$$

$$U_{1,x} f = \int_0^1 f(t)\, dt + \sum_{\nu=1}^{n} \frac{B_\nu(x)\, \Delta f^{(\nu-1)}}{\nu!} - \int_0^1 \frac{\bar{B}_n(x - t)}{n!} f^{(n)}(t)\, dt$$

$$= \int_0^1 f(t)\, dt + \sum_{\nu=1}^{n-1} \frac{B_\nu(x)\, \Delta f^{(\nu-1)}}{\nu!} - \int_0^1 \frac{(\bar{B}_n(x - t) - B_n(x))}{n!} f^{(n)}(t)\, dt$$

$$U_{m,x} f = \int_0^1 f(t)\, dt + \sum_{\nu=1}^{n-1} \frac{h^\nu B_\nu(x)\, \Delta f^{(\nu-1)}}{\nu!} - \frac{h^n}{n!} \int_0^1 \left(\bar{B}_n\left(x - \frac{t}{h}\right) - B_n(x)\right) f^{(n)}(t)\, dt \qquad (7.4.2)$$

On peut donc appliquer l'extrapolation de Richardson, comme au chapitre 3, à la fonction $v(h) = U_{m,x} f$ pour un x fixé. Les puissances impaires de h n'étant pas nulles, c'est la procédé L_n qui convient. On suppose que L_n est relatif aux abcisses $h_1, h_2, \ldots, h_n$ avec $h_k = \frac{h}{m_k}$, les m_k étant des entiers rangés par ordre croissant.

$$L_n(v) = L_n\,(U_{m,x}\, f) = P_{n,x}\, f =$$

$$= \int_0^1 f(t)\, dt - \frac{h^n}{n!} \int_0^1 \left[\sum_{k=1}^{n} A_k^n \left(\frac{1}{m}\right)^n \left(\bar{B}_n\left(x - \frac{t}{h_k}\right) - B_n(x)\right) \right] f^{(n)}(t)\, dt \qquad (7.4.3)$$

Le reste de la formule $P_{n,x}$ s'exprime sous forme d'une intégrale portant sur $f^{(n)}$.

On montre aisément que l'opérateur $U_{m,x}\left(m = \frac{1}{h}\right)$ appliqué à la fonction $V_n f$ définie en (7.3.14) donne en fait $P_{n,x} f$.

$$P_{n,x}\,(f) = U_{m,x}\,(V_n\, f) \qquad (7.4.4)$$

Il revient donc au même d'appliquer l'échantillonnage systématique U_m sur une fonction préalablement modifiée V_n f comme au § 7.3 ou de considérer l'extrapolation de Richardson appliquée à la formule des rectangles d'abscisse ξ tirée au hasard.

Essayons de borner la variance de $P_{n,\xi}$ f :

$$\begin{aligned}
\text{Var}\,(P_{n,\xi}\ f) &= \int_0^1 (P_{n,x}\ f - \int_0^1 f(t)\ dt)^2\ dx \\
&= \frac{h^{2n}}{(n!)^2} \int_0^1 \left\{ \int_0^1 \left[\sum_{k=1}^{n} A_k^n \left(\frac{1}{m_k}\right)^n \left(\bar{B}_n\left(x - \frac{t}{h_k}\right) - B_n(x) \right) \right] f^{(n)}(t)\ dt \right\} dx \\
&\leqslant \frac{h^{2n}}{(n!)^2} \int_0^1 \int_0^1 \left\{ \sum_{k=1}^{n} A_k^n \left(\frac{1}{m_k}\right)^n \left(\bar{B}_n\left(x - \frac{t}{h_k}\right) - B_n(x) \right) \right\}^2 \times (f^{(n)}(t))^2\ dt\ dx \\
&= \frac{h^{2n}}{(n!)^2} \int_0^1 (f^{(n)}(t))^2 \times \left(\int_0^1 \left\{ \sum_{k=1}^{n} A_k^n \left(\frac{1}{m_k}\right)^n \left(\bar{B}_n\left(x - \frac{t}{h_k}\right) - B_n(x) \right) \right\}^2 dx \right) dt
\end{aligned} \tag{7.4.5}$$

Evaluons séparément l'intégrale intérieure :

$$\int_0^1 \left\{ \sum_{k=1}^{n} A_k^n \left(\frac{1}{m_k}\right)^n \left(\bar{B}_n\left(x - \frac{t}{h_k}\right) - B_n(x) \right) \right\}^2 dx =$$

$$\sum_{j,k=1}^{n} \frac{A_j^n\ A_k^n}{(m_j)^n\ (m_k)^n} \int_0^1 \left(\bar{B}_n\left(x - \frac{t}{h_j}\right) - B_n(x) \right) \left(\bar{B}_n\left(x - \frac{t}{h_k}\right) - B_n(x) \right) dx$$

En utilisant (7.3.4) et la périodicité de $\bar{B}_n$ on montre que :

$$\int_0^1 \bar{B}_n\left(x - \frac{t}{h_j}\right) \bar{B}_n\left(x - \frac{t}{h_k}\right) dx = (-1)^{n-1} \frac{(n!)^2}{(2n)!} \bar{B}_{2n}\left(\frac{t}{h_j} - \frac{t}{h_k}\right)$$

$$\int_0^1 \bar{B}_n\left(x - \frac{t}{h_j}\right) \bar{B}_n(x)\ dx = \frac{(-1)^{n-1}\ (n!)^2}{(2n)!} \bar{B}_{2n}\left(\frac{t}{h_j}\right) \quad \text{etc.}$$

On a donc finalement :

$$\begin{aligned}
\text{Var}\,(P_{n,\xi}\ f) &\leqslant \frac{h^{2n}}{(n!)^2} \left(\text{Max}_t\ |f^{(n)}(t)|\right)^2 \int_0^1 \left\{ \sum_{\substack{k=1 \\ j=1}}^{n} \frac{A_j^n\ A_k^n}{(m_j)^n\ (m_k)^n} \right. \\
&\left. \left[\frac{(-1)^{n-1}\ (n!)^2}{(2n)!} \right] \times \left[\bar{B}_{2n}\left(\frac{t}{h_j} - \frac{t}{h_k}\right) + B_{2n} - \bar{B}_{2n}\left(\frac{t}{h_j}\right) - \bar{B}_{2n}\left(\frac{t}{h_k}\right) \right] \right\} dt \\
\text{Var}\,(P_{n,\xi}\ f) &\leqslant \frac{h^{2n}}{(2n)!} \left(\text{Max}_t\ |f^{(n)}(t)|\right)^2 \times \left(\sum_{\substack{k=1 \\ j=1}}^{n} \frac{|A_j^n| \times |A_k^n|}{(m_j)^n\ (m_k)^n} \right) \times \\
&\int_0^1 \left| \bar{B}_{2n}\left(\frac{t}{h_j} - \frac{t}{h_k}\right) + B_{2n} - \bar{B}_{2n}\left(\frac{t}{h_j}\right) - \bar{B}_{2n}\left(\frac{t}{h_k}\right) \right| dt \\
\text{Var}\,(P_{n,\xi}\ f) &\leqslant \frac{h^{2n}}{(2n)!} \left(\sum_{k=1}^{n} \frac{|A_k^n|}{(m_k)^n} \right)^2 (4\ |B_{2n}|)\ \left(\text{Max}_t\ |f^{(n)}(t)|\right)^2
\end{aligned} \tag{7.4.6}$$

<u>Remarque</u> : en utilisant le fait que pour tout x :

$$|P_{n,x}\ f - I| \leqslant \frac{h^n}{n!} \left(\sum_{k=1}^{n} \frac{|A_k^n|}{(m_k)^n} \right) |2B_n|\ \left(\text{Max}_t\ |f^{(n)}(t)|\right)$$

On obtient une borne pour Var $(P_{n,\xi}\ f)$ mais elle est moins bonne que la précédente.

<u>Valeur asymptotique de la variance</u> :

$$P_{n,x}\ f = I + \frac{h^n\ B_n(x)}{n!} \Delta f^{(n-1)} \left(\sum_{k=1}^{n} \frac{A_k^n}{(m_k)^n} \right) - \frac{h^{n+1}}{(n+1)!} \int_0^1 \sum_{k=1}^{n} \frac{A_k^n}{(m_k)^{n+1}} \left(\bar{B}_{n+1}\left(x - \frac{t}{h_k}\right) - B_{n+1}(x) \right) f^{(n+1)}(t)\ dt$$

Posons :

$$S(h, x) = \frac{-1}{(n+1)!} \int_0^1 \sum_{k=1}^{n} \frac{A_k^n}{(m_k)^{n+1}} \left[\bar{B}_{n+1}\left(x - \frac{t}{h_k}\right) - B_{n+1}(x) \right] f^{(n+1)}(t)\, dt$$

On a alors :

$$|S(h, x)| < \frac{2\, B_{n+1}}{(n+1)!} \left(\sum_{k=1}^{n} \frac{|A_k^n|}{(m_k)^{n+1}} \right) \left(\operatorname{Max}_t |f^{(n+1)}(t)| \right) = \Phi_n$$

$$\operatorname{Var}(P_{n,\xi} f) = \int_0^1 \left[\frac{h^n B_n(x)}{n!} \Delta f^{(n-1)} \left(\sum_{k=1}^{n} \frac{A_k^n}{(m_k)^n} \right) + h^{n+1} S(h, x) \right]^2 dx$$

$$= \frac{h^{2n}}{(n!)^2} (\Delta f^{(n-1)})^2 \left(\sum_{k=1}^{n} \frac{A_k^n}{(m_k)^n} \right)^2 \int_0^1 (B_n(x))^2\, dx$$

$$+ h^{2n+1} \left[\frac{\Delta f^{(n-1)}}{n!} \left(\sum_{k=1}^{n} \frac{A_k^n}{(m_k)^n} \right) \int_0^1 B_n(x)\, S(h, x)\, dx + h \int_0^1 (S(h, x))^2\, dx \right]$$

$$\operatorname{Var}(P_{n,\xi} f) = \frac{h^{2n}}{(2n)!} (\Delta f^{n-1})^2 \left(\frac{1}{\prod_{k=1}^{n} m_k} \right)^2 |B_{2n}| + h^{2n+1} R(h) \qquad (7.4.7)$$

avec :

$$|R(h)| < \frac{|\Delta f^{n-1}|}{n!} \times \left(\frac{1}{\prod_{k=1}^{n} m_k} \right) |B_n|\, \Phi_n + (\Phi_n)^2$$

Cas particuliers :

α) Prenons $m_k = 2^{k-1}$;

On a alors :

$$\sum_{k=1}^{n} \frac{A_k^n}{(2^{k-1})^n} = (-1)^{n+1} \frac{1}{2^{\frac{(n-1)n}{2}}}$$

Dans ce cas $P_{n,\xi} f = U_{m,\xi}(S_n f)$, (voir (7.3.9)) avec $m = \frac{1}{h}$:

$$\operatorname{Var}[U_{m,\xi}(S_n f)] = \frac{h^{2n} (\Delta f^{(n-1)})^2 |B_{2n}|}{(2n)!\, 2^{(n-1)n}} + h^{2n+1} R(h) \qquad (7.4.8)$$

où R(h) est une fonction bornée de h.

β) Prenons maintenant $m_k = k$

On obtient alors :

$$\sum_{k=1}^{n} \frac{A_k^n}{(k)^n} = (-1)^{n+1} \frac{1}{n!}$$

L'estimateur $P_{n,\xi} f$ représente alors $U_{m,\xi}(V_n^* f)$ (voir (7.3.14) et (7.3.16) :

$$\operatorname{Var}[U_{m,\xi}(V_n^* f)] = \frac{h^{2n} (\Delta f^{n-1})^2 |B_{2n}|}{(2n)!\,(n!)^2} + h^{2n+1} R(h) \qquad (7.4.9)$$

R(h) étant encore une fonction bornée de h.

L'opérateur $U_{m,x}$ appliqué à la fonction Df (voir 7.3.10) correspond à l'application de la formule d'intégration à 2 abcisses : $\frac{1}{2}(f(x) + f(1-x))$ appliquée avec un pas $h = \frac{1}{m}$ pour calculer $\int_0^1 f(t)\, dt$. Cette formule est du même ordre que la formule des trapèzes :

$$U_{m,x}(Df) = \int_0^1 f(t)\,dt + \sum_{\nu=1}^{n-1} \frac{h^{2\nu}\, B_{2\nu}(x)\, \Delta f^{(2\nu-1)}}{(2\nu)!}$$

$$- \frac{h^{2n}}{(2n)!} \int_0^1 \left[\frac{\bar{B}_{2n}\left(x - \frac{t}{h}\right) + \bar{B}_{2n}\left(x + \frac{t}{h}\right)}{2} - B_{2n}(x) \right] f^{(2n)}(t)\,dt \qquad (7.4.10)$$

On peut donc appliquer l'extrapolation M_n (chapitre 3) à la fonction $w(h) = U_{m,x}(Df)$ dont le développement en h est pair.

M_n est relatif aux abcisses $h_k = \frac{h}{m_k}$.

Le reste de la formule d'intégration obtenue s'écrit :

$$M_n(w) = M_n(U_{m,x}(Df)) = Q_{n,x} f = \int_0^1 f(t)\,dt - \frac{h^{2n}}{(2n)!} \int_0^1 \sum_{k=1}^{n} B_k^n \left(\frac{1}{m_k}\right)^{2n} \times$$

$$\left[\frac{\bar{B}_{2n}\left(x - \frac{t}{h_k}\right) + \bar{B}_{2n}\left(x + \frac{t}{h_k}\right)}{2} - B_{2n}(x) \right] f^{2n}(t)\,dt \qquad (7.4.12)$$

On remarque que $U_{m,x}$ appliqué à $W_n f$ défini en (7.3.15) représente en fait $Q_{n,x} f$:

$$U_{m,x}(W_n f) = Q_{n,x} f \qquad (7.4.12)$$

L'échantillonnage systématique U_m appliqué à la fonction $W_n f$ correspond donc à la formule de quadrature $Q_{n,\xi}$ d'abcisses aléatoires.

On peut borner la variance de $Q_{n,\xi}$ comme en (7.4.5) ; on obtient :

$$\mathrm{Var}(Q_{n,\xi} f) \leq \frac{h^{4n}}{(4n)!} \left(\sum_{k=1}^{n} \frac{|B_k^n|^2}{(m_k)^{2n}} \right) (4\,|B_{4n}|)\, (\mathrm{Max}_t\, |f^{(2n)}(t)|)^2 \qquad (7.4.13)$$

On a de même qu'en (7.4.7) l'expression asymptotique :

$$\mathrm{Var}(Q_{n,\xi} f) = \frac{h^{4n}}{(4n)!} (\Delta f^{2n-1})^2 \left(\frac{1}{\prod_{k=1}^{n} m_k} \right)^4 (-B_{4n}) + h^{4n+2} R(h) \qquad (7.4.14)$$

où R(h) est une fonction bornée de h.

Cas particuliers :

α) $$m_k = 2^{k-1} \quad ; \quad \sum_{k=1}^{n} \frac{B_k^n}{(2^{k-1})^{2n}} = \frac{1}{2^{(n-1)n}}$$

Dans ce cas $Q_{n,\xi} f$ représente $U_{m,\xi}(F_n f)$ (voir 7.3.11) :

$$\mathrm{Var}[U_{m,\xi}(F_n f)] = \frac{h^{4n} (\Delta f^{2n-1})^2 |B_{4n}|}{(4n)!\, 2^{2(n-1)n}} + h^{4n+2} R(h)$$

β) $$m_k = k \quad ; \quad \sum_{k=1}^{n} \frac{B_k^n}{(k)^{2n}} = \frac{1}{(n!)^2}$$

$Q_{n,\xi} f$ représente alors $U_{m,\xi}(W_n^* f)$ (voir 7.3.15 et 7.3.17) :

$$\mathrm{Var}[U_{m,\xi}(W_n^* f)] = \frac{h^{4n} (\Delta f^{2n-1})^2 |B_{4n}|}{(4n)!\,(n!)^4} + h^{4n+2} R(h)$$

CHAPITRE 8

APPLICATION A LA RÉSOLUTION D'ÉQUATIONS INTÉGRALES PAR UNE MÉTHODE DE MONTE-CARLO

§ 8.1 - ETABLISSEMENT D'UNE SOLUTION APPROCHEE AU MOYEN D'UNE SOMME D'INTEGRALES

Considérons l'équation de Fredholm, 2ème espèce :

$$\Phi(s) = g(s) + \int_0^1 K(s, t)\ \Phi(t)\ dt \qquad (8.1.1)$$

dans laquelle g et K sont des fonctions continues et même, à premières dérivées continues ; K est supposé symétrique.

Pour simplifier l'écriture on notera :

$$\begin{aligned} J\ \Phi(s) &= \int_0^1 K(s, t)\ \Phi(t)\ dt \\ A\ \Phi(s) &= (I - J)\ \Phi = \Phi(s) - \int_0^1 K(s, t)\ \Phi(t)\ dt \end{aligned} \qquad (8.1.2)$$

L'équation devient $A\ \Phi = g$.

La méthode de Richardson ([28], [51], [55]) pour les systèmes linéaires peut-être adaptée aux équations intégrales.

Nous cherchons une solution approchée de la forme :

$$\Phi^{(I)} = a_0\ g + a_1\ J\ g + a_2\ J^{(2)}\ g + \ldots + a_{I-1}\ J^{(I-1)}\ g \qquad (8.1.3)$$

avec :

$$J^{(i)}\ g = \int_0^1 K^{(i)}\ (s, t)\ g(t)\ dt$$

où $K^{(i)}$ est le $i^{ème}$ itéré du noyau K.

Appelons r_I le résidu correspondant :

$$r_I = A(\Phi - \Phi^{(I)}) = g - A\ \Phi^{(I)}$$

Le résidu peut encore s'écrire :

$$r_I = Q_I\ (J)\ g = R_I\ (A)\ g \qquad (8.1.4)$$

les polynômes R_I et Q_I du $I^{ème}$ degré étant liés par la relation :

$$R_I\ (1 - x) \equiv Q_I\ (x) \quad \text{avec } R_I\ (0) = 1$$

Soient λ_i les valeurs propres et Φ_i les fonctions propres correspondantes, solutions de :

$$A\ \Phi_i = \lambda_i\ \Phi_i$$

En utilisant le théorème de Hilbert-Schmidt on a :

$$J^{(n)}\ g = \sum_{i=1}^{\infty} (1 - \lambda_i)^n\ g_i\ \Phi_i \quad (n \geqslant 1)$$

avec $g_i = \int_0^1 g(s)\ \Phi_i(s)\ ds$. La série converge uniformément. Après quelques transformations on obtient la formule suivante :

$$r_I = R_I(1)\ g + \sum_i [R_I(\lambda_i) - R_I(1)]\ g_i\ \Phi_i \qquad (8.1.5)$$

(la série converge uniformément).

Si la série $\sum_i g_i\ \Phi_i$ converge uniformément vers g (sauf cas particulier, cela nécessite que les Φ_i forment un système complet de fonctions) alors la formule précédente peut s'écrire plus simplement :

$$r_I = \sum_i R_I(\lambda_i)\ g_i\ \Phi_i$$

Il suffit alors que R_I tende vers 0 uniformément par rapport aux abcisses $\{\lambda_i\}$ quand I tend vers l'infini pour que r_I tende uniformément vers 0.

Les conditions précédentes ne sont pas toujours réalisées. En particulier, elles le sont rarement si les valeurs propres sont en nombre fini. Nous avons utilisé dans les exemples traités le théorème suivant :

Si $\sum_i |g_i\ \Phi_i|$ converge ponctuellement vers une fonction g^* bornée sur [0,1], il suffit alors que R_I tende vers 0 uniformément par rapport aux abcisses $\{\lambda_i, 1\}$ quand I tend vers l'infini pour que r_I converge uniformément vers 0.

Choix de R_I

Supposons que l'on puisse déterminer deux nombres α et β strictement positifs tels que $\alpha \leqslant \lambda \leqslant \beta$ pour tout i. Dans les exemples traités (n° 1 et 3) ci-dessous on a : $\alpha = 2$, $\beta = 5$; (la série de Neumann ne converge donc pas). On considère un nouvel intervalle V = [a, b] qui contienne $[\alpha, \beta]$ et l'abcisse 1 ; (ici a = 1, b = 5). On prendra pour R_I le polynôme de degré I admettant entre a et b l'extremum le plus petit possible en valeur absolue tout en vérifiant $R_I(0) = 1$.

La solution est fournie par les polynômes de Tchebycheff ([28], [35], [55]) :

$$R_I(\lambda) = \frac{\cos(I\varphi)}{\mathrm{ch}(I\varphi_0)} \quad \text{avec} \quad \cos(\varphi) = \frac{b + a - 2\lambda}{b - a} \qquad \lambda \in [a,b]$$

$$\mathrm{ch}(\varphi_0) = \frac{b + a}{b - a} \qquad (8.1.6)$$

Connaissant R_I on remontera facilement aux coefficients a_i de la formule (8.1.3) qui sera en fait utilisée pour le calcul.

Méthode de Monte-Carlo élémentaire :

On désire évaluer :

$$\Phi^{(I)} = a_0\ g + a_1\ J\ g + a_2\ J^{(2)}\ g + \ldots + a_{I-1}\ J^{(I-1)}\ g$$

à l'abcisse $s = s_0$.

Les ξ_i étant des variables aléatoires indépendantes uniformément distribuées sur [0,1], posons :

$$\begin{aligned}
Z_0 &= a_0\ g(s_0) \\
Z_1 &= a_1\ K(s_0, \xi_1) \times g(\xi_1) \\
Z_2 &= a_2\ K(s_0, \xi_1) \times K(\xi_1, \xi_2) \times g(\xi_2) \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
Z_{I-1} &= a_{I-1}\ K(s_0, \xi_1) \times K(\xi_1, \xi_2) \times \ldots \times K(\xi_{I-2}, \xi_{I-1}) \times g(\xi_{I-1})
\end{aligned} \qquad (8.1.7)$$

Les Z_i sont des variables aléatoires telles que :

$$E(Z_i) = a_i \; J^i \quad g(s_o)$$

$Z = \sum_{i=0}^{I-1} Z_i$ vérifie donc $E(Z) = \Phi^{(I)}(s_o)$.

Utilisation des opérateurs du chapitre 7 :

Considérons l'application de l'échantillonnage systèmatique de pas $\frac{1}{4}$ sur la fonction modifiée $W_1^* f$. Cela revient à employer la variable aléatoire :

$$U_4 \; W_1^* f(\xi) = \frac{1}{8} \sum_{i=0}^{3} \left[f\left(\frac{\xi + i}{4}\right) + f\left(1 - \frac{\xi + i}{4}\right) \right]$$

pour évaluer $\int_o^1 f(x)\, dx$.

$$U_4 \; W_1^* f(\xi) = \sum_i \alpha_i \; f(\beta_i)$$

les β_i étant des fonctions simples de ξ.

Pour calculer $\int_o^1 \int_o^1 h(x, y)\, dx\, dy$, on employera de même l'estimateur :

$$(U_4 \; W_1^*)^y \; (U_4 \; W_1^*)^x \; h(\xi_1, \xi_2) = \sum_i \sum_j \alpha_i \; \alpha_j \; h(\beta_i^*, \beta_j^{**})$$

où les β_i^* sont des fonctions simples de ξ_1 et les β_j^{**} les mêmes fonctions mais portant sur ξ_2. On procède de même pour calculer

$$J^{(k)} \; g(s_o) = \int_o^1 \ldots \int_o^1 K(s_o, t_1) \times K(t_1, t_2) \times \ldots \times K(t_{k-1}, t_k) \; g(t_k) \; dt_1 \ldots dt_k$$

§ 8.2 - RESULTATS NUMERIQUES -

DESCRIPTION DES EQUATIONS INTEGRALES TRAITEES

Equation n° 1 :

$$K(s, t) = - e^{s+t} \left[\frac{3}{\mu_1} + \frac{7}{2} (a\, s + b)\, (a\, t + b) \right]$$

avec $\mu_1 = e^2 - 1$; $\mu_2 = e^2 + 1$;

$$a = 2 \sqrt{\frac{2\, \mu_1}{2\, \mu_1^2 - \mu_2^2}}$$

$$b = \frac{\mu_2}{\mu_1} \times \frac{a}{2}$$

$f(s) = s$.

Les valeurs propres sont situées dans l'intervalle [2,5] ; on choisit le polynôme $R_4(\lambda)$ relatif à l'intervalle [1,5] :

$$\Phi^{(4)} = a_o \; g + a_1 \; J \; g + a_2 \; J^{(2)} \; g + a_3 \; J^{(3)} \; g$$

$a_o = 0,957\,446\,808$

$a_1 = 0,617\,021\,276$

$a_2 = 0,191\,489\,362$

$a_4 = 0,021\,276\,595\,7$

$R_4(\lambda)$ reste inférieur à 0,045 dans l'intervalle [1,5]. On s'intéresse à la solution pour s = 0,5. On a $\Phi^{(4)}(0,5) = 0,239$ alors que $\Phi(0,5) = 0,246$.

Equation n° 2 :

$$K(s, t) = \frac{\pi}{4} \cos(\pi(s + t))$$

$$f(s) = \pi s$$

Les valeurs propres sont situées dans l'intervalle [0,5 ; 2] ; on choisit le polynôme $R_4(\lambda)$ relatif à l'intervalle [0,5 ; 2] :

$$\Phi^{(4)} = a_0\, g + a_1\, J\, g + a_2\, J^{(2)}\, g + a_3\, J^{(3)}\, g$$

$$a_0 = 0,994\ 818\ 652$$

$$a_1 = 1,131\ 362\ 389$$

$$a_2 = 1,247\ 180\ 737$$

$$a_3 = 0,624\ 199\ 932$$

On s'intéresse à la solution pour s = 0,5. On trouve $\Phi^{(4)}(0,5) = 0,983$ pour $\Phi(0,5) = 1,007$.

Equation n° 3 :

$$K(s, t) = -\frac{1}{2} [9(2s - 1)(2t - 1) + 25(2s - 1)^2 (2t - 1)^2]$$

$$f(s) = \exp(2s - 1)$$

Les valeurs propres sont situées dans l'intervalle [2,5] ; on prendra R_4 et les a_i comme pour l'exemple 1.

On cherche la solution pour s = 0,7. On a $\Phi^{(4)}(0,7) = 0,940$ alors que $\Phi(0,7) = 0,976$.

Dans tous les exemples, on pourrait obtenir une précision meilleure en retenant plus de termes : exemple :

$\Phi^{(5)}(0,7) = 0,963$. Il faut remarquer que la méthode de Monte-Carlo ne donnera que 2 ou 3 chiffres significatifs.

Les 7 procédés suivants ont été comparés (voir § 7.3).

1/ Monte-Carlo élémentaire

2/ $U_4\ W_1^*$

3/ W_2^*

4/ $U_2\ W_3^*$

5/ $U_2\ V_3^*$

6/ W_3^*

7/ $U_4\ V_2^*$

Pour évaluer les intégrales multiples on applique ces opérateurs successivement sur chaque dimensions avec des ξ_i différents et indépendants. On obtient ainsi 7 estimateurs différents pour $\Phi^{(4)}(s_o)$.

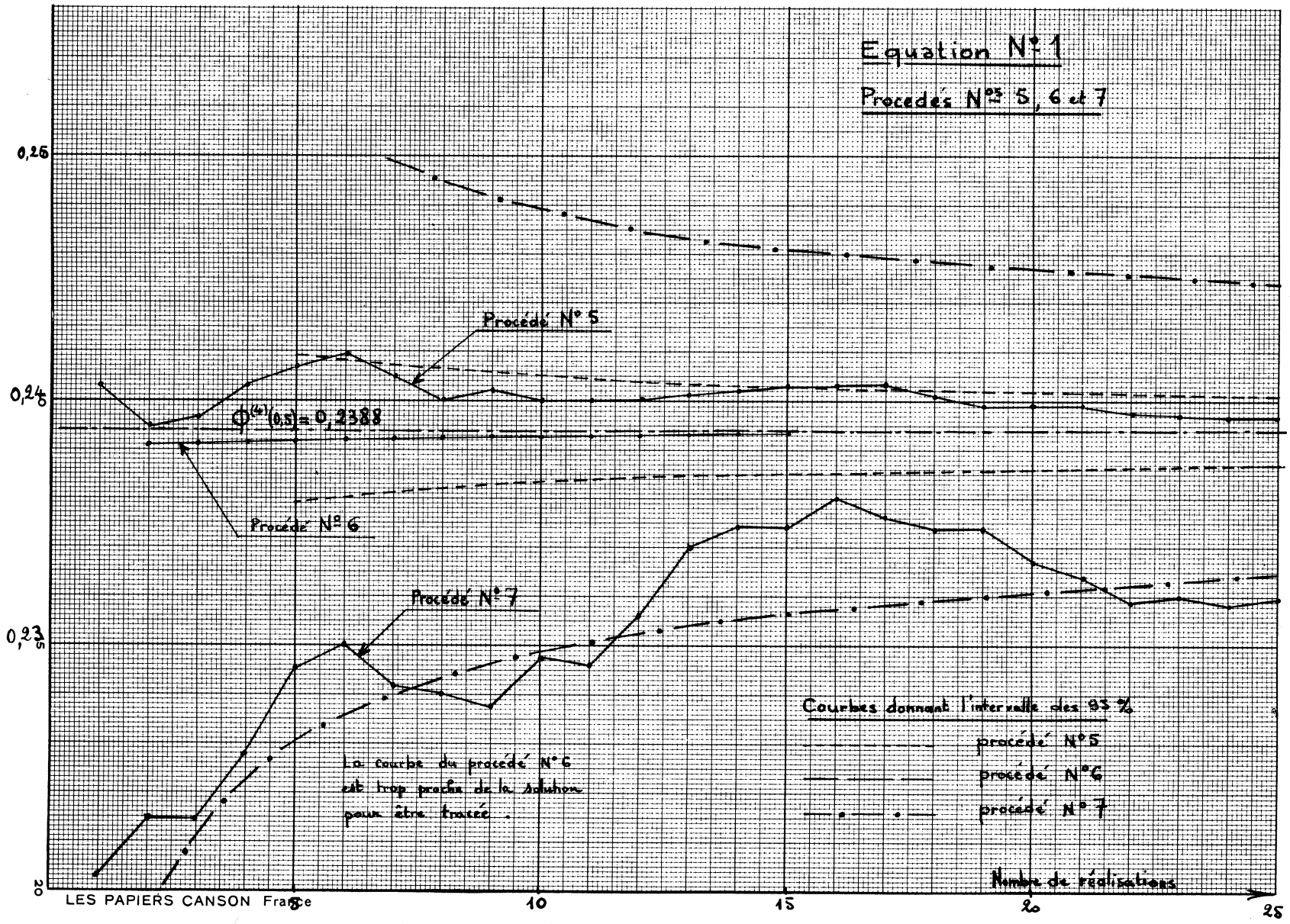
Pour chacun de ces estimateurs, on effectue un certain nombre de réalisations et on en fait la moyenne. On peut estimer que cette moyenne est une variable aléatoire sensiblement gaussienne. L'écart-type est estimé expérimentalement. On a alors une probabilité peu différente de 0,95 de se trouver dans l'intervalle ± 2 × écart-type autour de la moyenne. Les 7 estimateurs précédents ne font pas intervenir le même nombre de calculs de K et g. En jouant sur le nombre de réalisations nous avons comparé les 7 procédés à égalité de coût (même nombre de calculs de K et g). Cela correspondait d'ailleurs à des temps-machine sensiblement égaux.

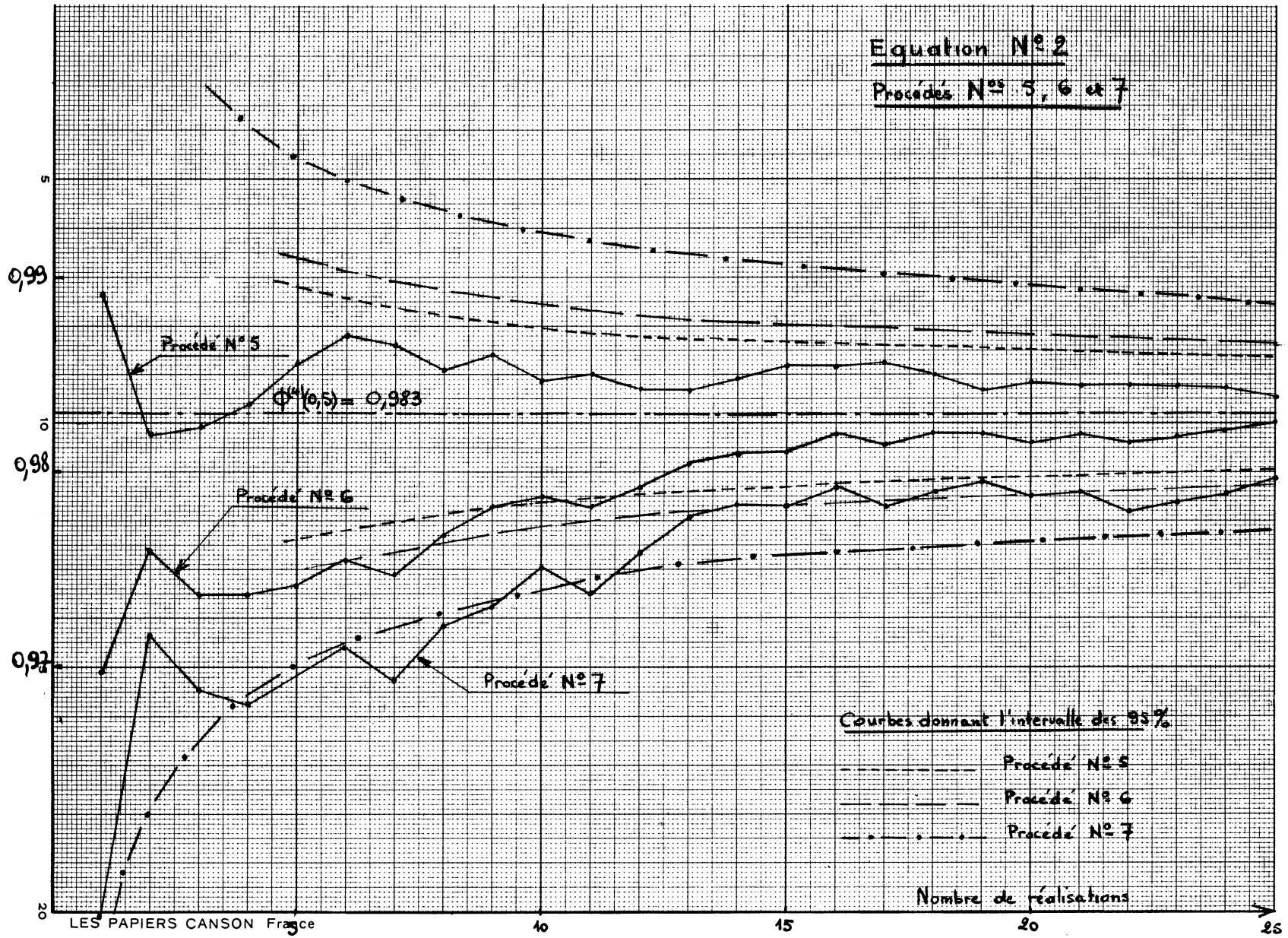
Le tableau suivant donne 2 x écart-type x 10^4 pour chaque méthode et chaque équation :

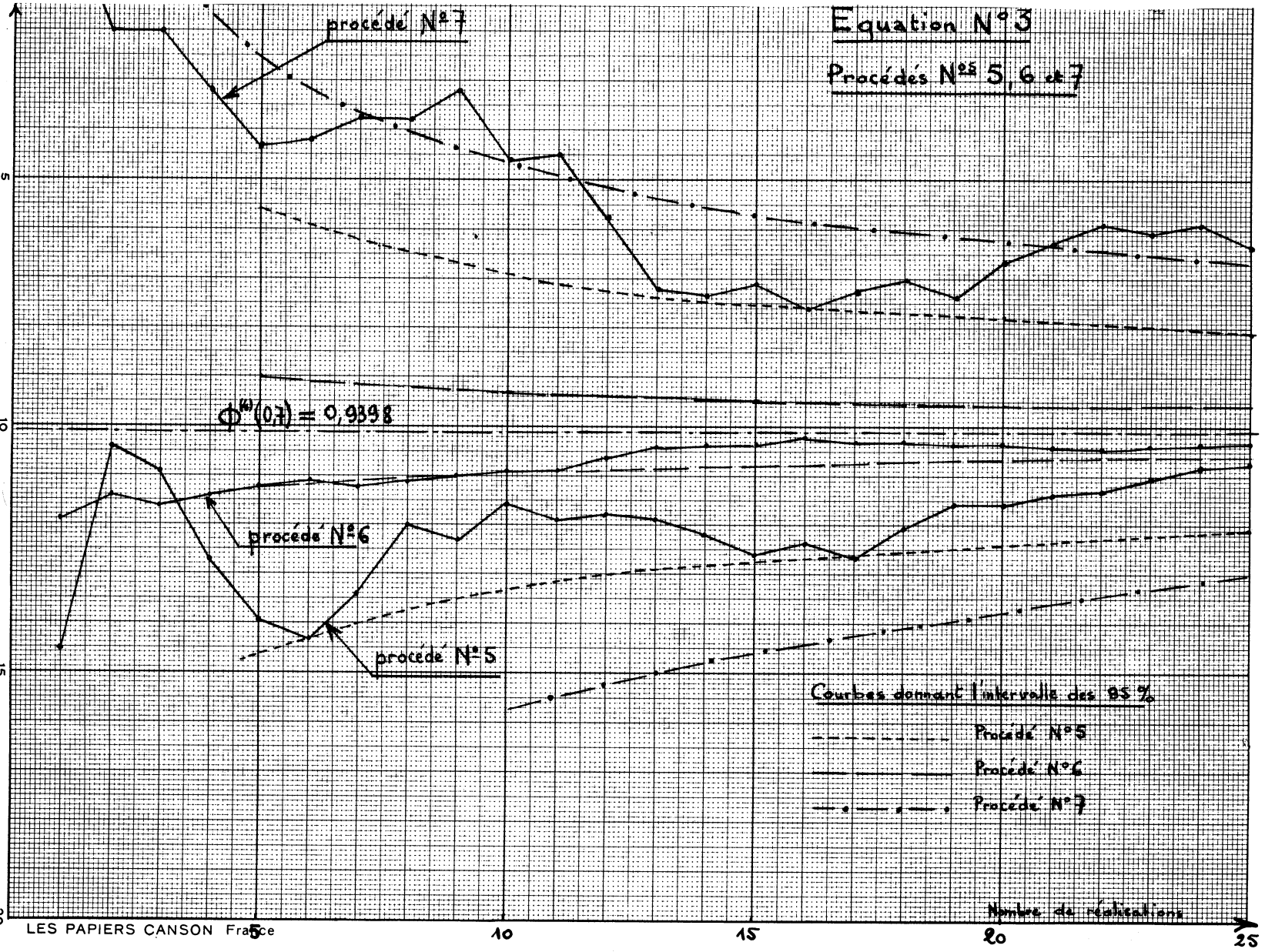
méthode → / Equation ↓	1	2	3	4	5	6	7
1	610	98	19	3	16	2	66
2	445	78	116	7	32	42	65
3	1 340	114	144	14	45	11	74

Les 6 méthodes introduites (2 à 7) donnent dans les 3 cas un résultat plus précis que la méthode de Monte-Carlo élémentaire (1).

Les graphiques suivants montrent l'évolution de la solution en fonction du temps. Les deux courbes symétriques par rapport à la moyenne donnent l'évolution de l'intervalle ± 2 x écart-type dans lequel on a 95 % de chance de se trouver.




Equation N° 2
Procédés N°s 5, 6 et 7
0,99
0,98
0,97
Procédé N° 5
Φ(u)(0,5) = 0,983
Procédé N° 6
Procédé N° 7
Courbes donnant l'intervalle des 95 %
Procédé N° 5
Procédé N° 6
Procédé N° 7
Nombre de réalisations
5
10
15
20
25
LES PAPIERS CANSON France



BIBLIOGRAPHIE

[1] N. I. ACHIESER - Theory of Approximation (traduit du Russe). New-York (1956).

[2] F. L. BAUER - La méthode d'intégration numérique de Romberg. Colloque sur l'Analyse numérique, Mons, (1961), pp 119 - 128.

[3] N. BOGOLOUBOFF et N. KRYLOFF - On the Rayleigh's Principle in the Theory of the Differential Equations of the Mathematical Physics and upon the Euler's Method in the Calculus of Variations. Acad. des Sci. de l'Ukraine, Phys. Math., tome 3 fasc. 3 (1926).

[4] L. BOLLIET, N. GASTINEL et P. J. LAURENT - Manuel ALGOL. Hermann (1964).

[5] W. G. COCHRAN - Sampling Techniques. John Wiley, N. Y., (1959).

[6] L. COLLATZ - Numerische Behandlung von Differentialgleichungen. Springer Verlag (1955).

[7] H. CRAMER - Mathematical Methods of Statistics, Princeton University Press (1946).

[8] W. E. DEMING - Some Theory of Sampling. John Wiley, (1950).

[9] J. FAVARD - Sur les quadratures mécaniques. Enseignement mathématique, tome III, fasc. 4 (1957), pp 263 - 275.

[10] J. FAVARD - Cours d'analyse de l'Ecole Polytechnique. Gauthier-Villars (1960).

[11] E. C. FIELLER and H. O. HARTLEY - Sampling with Control Variables. Biometrika 41 (1954), pp 494 - 501.

[12] G. E. FORSYTHE and W. R. WASOW - Finite Difference Methods for Partial Differential Equations. J. Wiley (1959).

[13] T. FORT - Finite Differences. Clarendon Press (1948).

[14] L. FOX - Numerical Solution of Ordinary and Partial Differential Equations. Pergamon Press (1962).

[15] M. GOLOMB - Lectures on Theory of Approximation. Argonne National Laboratory (1962).

[16] J. H. HALTON and D. C. HANDSCOMB - A Method for Increasing the Efficiency of Monte-Carlo Integration. J. Assoc. Comput. Mach. 4, (1957), pp 329 - 340.

[17] J. M. HAMMERSLEY and K. W. MORTON - A new Monte-Carlo Technique : Antithetic Variates. Proc. Camb. Phil. Soc. 52, (1956), pp 449 - 475.

[18] P. HENRICI - Discrete Variable Methods in Ordinary Differential Equations. J. Wiley (1962).

[19] F. B. HILDEBRANDT - Introduction to Numerical Analysis. Mc Graw-Hill (1956).

[20] A. HOUSEHOLDER - Principles of Numerical Analysis. Mc Graw-Hill (1953).

[21] H. KAHN - Use of Different Monte-Carlo Sampling Techniques. Symposium on Monte-Carlo Methods. A. Meyer (March 1954), pp 146 - 190.

[22] L. V. KANTOROVITCH et AKYLOV - Analyse fonctionnelle dans les espaces normés (en russe). Phyzmatguiz Moscou (1959), pp 229 - 234.

[23] V. I. KRYLOV - Approximate Calculation of Integrals. A. C. M. Monograph Series (1962).

[24] J. KUNTZMANN - Etude de représentations approchées dans le cas de deux variables. Chiffres 1 n° 1 (1958) pp 35 - 40.

[25] J. KUNTZMANN et F. CESCHINO - Problèmes differentiels de conditions initiales. Dunod (1963).

[26] J. KUNTZMANN - Methodes numériques. Interpolation. Dérivées. Dunod (1959).

[27] R. O. KUZMIN - On the Theory of Mechanical Quadrature. Izv. Leningrad. Polytechn. In-ta. Otd. Estest. Mat., vol. 32, (1931) (en russe), pp 5 - 14.

[28] C. LANCZOS - Chebyshev Polynomials in the Solution of Large-Scale Linear Systems. Proc. Assoc. Comput. Mach. ; Toronto Meeting (1956).

[29] P. J. LAURENT - Remarque sur l'évaluation d'intégrales par la méthode de Monte-Carlo. Comptes Rendus Acad. Sci. Paris, t. 253, pp 610 - 612 (1961).

[30] P. J. LAURENT - Evaluations d'integrales par une méthode de Monte-Carlo. Application à la résolution d'équations intégrales. 2ème Congrès de l'AFCALTI. Paris (oct. 1961).

[31] P. J. LAURENT - Un théorème de convergence pour le procédé d'extrapolation de Richardson. Comptes Rendus Acad. Sci., tome 256, pp 1435 - 1437, (1963).

[32] P. J. LAURENT - Convergence du procédé d'extrapolation de Richardson. 3ème Congrès de l'AFCALTI. Toulouse (1963).

[33] P. J. LAURENT - Application du procédé d'extrapolation de Richardson à l'intégration approchée. Colloque sur les méthodes probabilistes en Analyse numérique. Clermont-Ferrand (25-26 oct. 1963).

[34] P. J. LAURENT - Formules de quadrature approchée sur domaines rectangulaires convergentes pour toute fonction intégrable Riemann. Comptes rendus Acad. Sci. Paris t. 258, pp 798 - 801, (janv. 1964).

[35] W. MARKOW - Über Polynome, die in einem gegebenen. Intervalle möglichst wenig von Null abweichen. Math. Ann., vol. 77, pp 213 - 258 (1916) (traduit de l'original russe de 1892).

[36] A. MICHEL - Etude de l'erreur dans la représentation approchée des intégrales doubles. Chiffres vol. 3 n° 2 (1961), pp 79 - 84.

[37] A. MICHEL - Représentation approchée des intégrales doubles. Thèse. Grenoble (dec. 1959).

[38] J. C. MIELLOU - Contribution à l'étude de l'erreur dans l'intégration multiple. Thèse. Grenoble (Juin 1961).

[39] K. W. MORTON - A Generalisation of the Antithetic Variate Technique for Evaluating Integrals. J. Math. Phys. 36 (1957) pp 289 - 293.

[40] I. P. NATANSON - Konstruktive Funktionentheorie (traduit du russe). Berlin (1955).

[41] M. R. OSBORNE - h^2 - Extrapolation in Eigenvalue Problems. Quart. J. Mech. 13, (1960), pp 156 - 168.

[42] G. POLYA - Über die Konvergenz von Quadraturverfahren. Math. Z., vol. 37, (1933) pp. 264 - 86.

[43] L. F. RICHARDSON - The Deferred Approach to the Limit. Phil. Trans. Roy. soc. London vol. 226, (1927), pp 299 - 361.

[44] L. F. RICHARDSON - The Approximate Arithmetical Solution by Finite Differences of Physical Problems etc. Phil. Trans. Roy. Soc. London, vol. 210, (1911), pp 307 - 357.

[45] W. ROMBERG - Vereinfachte numerische Integration. Det Kong. Norske Videnskabers Selskab Forhandlinger, Band 28 n° 7, (1955), pp 30 - 36.

[46] H. RUTISHAUSER - Ausdehnung des Rombergschen Prinzips. Numerische Mathematik 5 (1963), pp 48 - 54.

[47] M. G. SALVADORI - Extrapolation Formulas in Linear Difference Operators. Proceedings of the First U. S. National Congress of Applied Mechanics, (1951), pp 15 - 18.

[48] A. SARD - Integral Representations of Remainders. Duke Math. J., vol. 15, n° 2, (1948), pp 333 - 345.

[49] A. SARD - Remainders as Integrals of Partial Derivatives. Proc. Amer. Math. Soc., vol 3, (1952), pp 732 - 741.

[50] V. A. STEKLOV - On the Approximate Calculation of Definite Integrals with the Aid of For-

mulas of Mechanical Quadrature. Izv. Akad. Nauk SSSR 6, vol. 10 (1916) pp 169 - 186, (en russe).

[51] E. STIEFEL - On Solving Fredholm Integral Equations. J. Soc. Ind. Appl. Math. 4 (1956) pp 63 - 85.

[52] E. STIEFEL et H. RUTISHAUSER - Remarques concernant l'intégration numérique. Comptes Rendus Acad. Sci. (1961), pp 1899 - 1900.

[53] E. STIEFEL - Altes und Neues über numerische Quadratur. Z. A. M. M. 41, Heft 10/11, (1961), pp 408 - 413.

[54] J. TODD - Survey of Numerical Analysis. Mc Graw-Hill.

[55] R. S. VARGA - A Comparaison of the Successive Overrelaxation Method and Semi-iterative Methods Using Chebyshev Polynomials. J. Soc. Ind. Appl. Math. 5, (1957), pp 39 - 46.

[56] B. L. VAN DER WÄRDEN - Mathematische Statistik, Springer Verlag (1957).

[57] W. WASOW - Discrete Approximations to Elliptic Differential Equations Z. A. M. P. 6, (1955), pp 81 - 97.

[58] A. C. ZAANEN - Linear Analysis. North-Holland Publishing Co. (1960).

PUBLICATIONS RECENTES

[1]* F. L. BAUER, H. RUTISHAUSER et E. STIEFEL - New Aspects in numerical Quadrature - Proc. of Symposia in Applied Mathematics 15, Am. Math. Soc. (1963).

[2]* R. BULIRSCH - Bemerkungen zur Romberg Integration. Num. Math. 6, p. 6 - 16 (1964).

[3]* R. BULIRSCH and J. STOER - Über Fehlerabschätzung und Extrapolation mit rationalen Funktionen bei Verfahren vom Richardson - Typus. (à paraître dans Num. Math.).

[5]* W. GRAGG - Repeated Extrapolation to the Limit in the Numerical Solution of Ordinary Differential Equations. Thesis UCLA (1963) (à paraître).

[5]* H. J. STETTER - Asymptotic Expansions for the Error of Discretisation Algorithms for Non-linear Functional Equations. (à paraître dans Arch. for Rat. Mech. and Anal.).

TABLE DES MATIÈRES

	Pages
CHAPITRE I - LE PROCEDE D'EXTRAPOLATION DE RICHARDSON	5
1.1 - Introduction	5
1.2 - Extrapolation d'une fonction. Erreurs	7
1.3 - Abcisses en progression géométrique de raison 1/2	11
1.4 - Extrapolation basée sur les polynômes pairs	12
1.5 - Procédures ALGOL pour le procédé d'extrapolation de Richardson	13
CHAPITRE II - CONVERGENCE ET STABILITE	15
2.1 - Un théorème de convergence	15
2.2 - Rappel d'un théorème de base	16
2.3 - Démonstration du théorème de convergence	17
2.4 - Stabilité du procédé d'extrapolation	19
CHAPITRE III - APPLICATION DU PROCEDE D'EXTRAPOLATION DE RICHARDSON A L'INTEGRATION NUMERIQUE	21
3.1 - Extrapolation sur la formule des trapèzes	21
3.2 - Choix des pas en progression géométrique de raison 1/2 ; méthode de Romberg	22
3.3 - Etude du noyau d'erreur	25
3.4 - Signe des coefficients	27
3.5 - Procédure ALGOL pour l'intégration simple et exemples numériques	29
3.6 - Autres procédés analogues	30
3.7 - Formule des trapèzes à deux dimensions	32
3.8 - Extrapolation sur la formule des trapèzes à deux dimensions	35
3.9 - Procédure ALGOL pour l'intégration double et exemples numériques	38
CHAPITRE IV - ETUDE DE LA DERIVATION APPROCHEE	40
4.1 - Formules utilisant des points tous du même côté	40
4.2 - Formules utilisant des points symétriques	43
4.3 - Programme ALGOL et exemple numérique	45
4.4 - Quelques extensions possibles	46
CHAPITRE V - APPLICATION A L'INTEGRATION DES EQUATIONS DIFFERENTIELLES	48
5.1 - Etude du comportement de la solution approchée en fonction du pas h pour un problème de conditions initiales	48

Pages

5.2 - Extrapolation sur la formule d'Euler, procédure ALGOL et exemples numériques..... 53

5.3 - Etude du comportement de la solution approchée en fonction du pas h pour un problème de conditions aux limites..... 56

5.4 - Exemple d'extrapolation pour un problème de conditions aux limites... 60

5.5 - Exemple d'extrapolation pour un problème de valeurs propres..... 61

CHAPITRE VI - APPLICATION AUX EQUATIONS INTEGRALES ET AUX EQUATIONS AUX DERIVEES PARTIELLES..... 64

6.1 - Etude du comportement de la solution approchée pour une équation intégrale de Fredholm de 2ème espèce..... 64

6.2 - Exemple d'extrapolation pour une équation intégrale..... 66

6.3 - Etude du comportement de la solution approchée pour l'équation aux dérivées partielles de Poisson..... 67

6.4 - Exemple d'extrapolation pour l'équation de Laplace..... 68

CHAPITRE VII - EVALUATION D'INTEGRALES PAR UNE METHODE DE MONTE-CARLO APPLICATION DU PROCEDE DE RICHARDSON..... 71

7.1 - Introduction..... 71

7.2 - Quelques procédés de réduction de variance..... 71

7.3 - Echantillonnage symétrique sur une fonction modifiée..... 75

7.4 - Extrapolation appliquée à une formule de quadrature d'abcisses aléatoires. 79

CHAPITRE VIII - APPLICATION A LA RESOLUTION D'EQUATIONS INTEGRALES PAR UNE METHODE DE MONTE-CARLO..... 83

8.1 - Etablissement d'une solution approchée au moyen d'une somme d'intégrales..... 83

8.2 - Résultats numériques..... 85

BIBLIOGRAPHIE..... 91

DEUXIEME THESE

Propositions données par la Faculté

LE THEOREME DU GRAPHE FERME DANS LES ESPACES DE BANACH

Vu,
Grenoble, le 19 mai 1964
Le Président de la thèse,
J. FAVARD

Vu,
Grenoble, le 20 mai 1964
Le Doyen de la Faculté des Sciences,
L. WEIL

Vu
et Permis d'imprimer
Le Recteur de l'Académie de Grenoble,
R. TREHIN

IMP. LOUIS-JEAN – GAP
Dépôt légal n° 181 – 1964