LEÇONS D'ANALYSE NUMÉRIQUE

PREMIÈRE PARTIE

J.F. DEBONGNIE

© DEBONGNIE (JEAN-FRANÇOIS), Liège, 2009

ISBN-13 : 978-2-9600313-5-5
Dépôt légal : D/2009/0480/34

D E S E R R E U R S E T D E L E U R S C O N S E Q U E N C E S

1. Tous les nombres que l'on utilise en pratique sont entachés d'erreurs. Ces erreurs ont des provenances diverses. Rappelons pour mémoire que les problèmes physiques sont tous un tant soit peu idéalisés, ce qui fait naître dès le départ des différences entre le phénomène réel et son modèle. Bien que rarement ressenties par les calculateurs, ces différences peuvent mener dans certains cas à des divergences fondamentales de comportement où, par exemple, l'un est stable et l'autre, instable. Ainsi, par exemple, la théorie des plaques de KIRCHHOFF prévoit une déformée finie sous une charge concentrée, au contruire de théories plus fines.

La valeur des paramètres à faire entrer dans le modèle est souvent connue avec une précision très limitée et, d'ailleurs, qui peut garantir que son acier est parfaitement homogène, avec un module de YOUNG connu à $1^{o}/_{oo}$? De même, les résistances électriques, les pièces de machines et, en général, tous les objets fabriqués le sont avec certaines tolérances.

A ces erreurs, sur les quelles nous ne nous étendrons pas, mais qui doivent constamment rester à l'esprit du calculateur, il faut en ajouter d'autres, d'origine purement numérique. Soit à représenter le nombre 1/30 dans le système décimal. On sait que

$$1/30 = 0{,}033\ 333\ 3\ldots\ ,$$

la suite des chiffres 3 étant indéfinie. Néanmoins, une machine à calculer possède une capacité limitée de représenter les chiffres, par exemple, elle retiendra 5 chiffres significatifs, ce qui lui fera écrire

$$1/30 \approx 0{,}33333 \cdot 10^{-1}$$

dans le système de notation dit "scientifique". (Ceci suppose déjà que l'on dispose d'une machine à virgule flottante). Il est clair que ce nombre diffère de 1/30 par la chute des derniers chiffres. C'est ce que l'on appelle l'erreur d'arrondi.

Ces arrondis sont omniprésents. Tout d'abord, la machine calcule en binaire, mais communique le plus souvent avec l'utilisateur en décimal. Il en résulte des conversions qui, nécessairement, sont arrondies si le nombre n'est pas un multiple d'une puissance de

deux. De plus, les nombres ont leur développement tronqué à un nombre donné de chiffres binaires.

L'erreur d'arrondi dépend du soin apporté par le constructeur à son arithmétique. La bonne pratique des calculateurs manuels est de calculer avec un certain nombre de décimales supplémentaires, dites décimales de réserve, mettons deux, puis d'arrondir le résultat final à la valeur la plus proche. L'erreur d'arrondi est alors statistiquement aussi souvent par excès que par défaut, et toujours inférieure à $0{,}5.\ 10^{-p+1}$, où p est le nombre de chiffres conservés. Les ordinateurs se contentent souvent d'une simple troncature, ce qui double l'erreur maximale d'arrondi et la rend systématiquement par défaut, ce qui peut à la longue biaiser les calculs.

Un modèle élémentaire de l'erreur d'arrondi consiste à admettre que tout nombre x est représenté en machine par $x = x(1 + \varepsilon)$, où ε représente une erreur relative dépendant de la machine.

D'apparence anodine, les erreurs d'arrondi peuvent avoir des conséquences très fâcheuses, allant jusqu'à détruire totalement l'efficacité d'un algorithme pourtant valable sur le plan théorique. C'est pourquoi il est bon d'en connaître quelques notions avant même d'étudier les algorithmes eux-mêmes.

2. OPERATIONS ELEMENTAIRES

2.1 - Considérons d'abord l'addition de deux nombres positifs, arrondis $\tilde{x}_1$ et $\tilde{x}_2$. On a donc

$$\tilde{x}_1 = x_1(1 + \varepsilon_1) \quad , \quad \tilde{x}_2 = x_2(1 + \varepsilon_2) \ ,$$

ε_1 et ε_2 étant les erreurs relatives. La somme faite, le résultat est encore arrondi, ce qui donne, en l'appelant $\tilde{y}$,

$$\begin{aligned} \tilde{y} &= (x_1(1 + \varepsilon_1) + x_2(1 + \varepsilon_2))(1 + \varepsilon_3) \\ &= x_1 + x_2 + (\varepsilon_1 + \varepsilon_3)\, x_1 + (\varepsilon_2 + \varepsilon_3)\, x_2 + 0(\varepsilon^2). \end{aligned}$$

L'erreur maximale au premier ordre vaut donc

$$\Delta y = |\varepsilon_1|\, x_1 + |\varepsilon_2|\, x_2 + |\varepsilon_3|(x_1 + x_2) \ ,$$

ce qui correspond à une erreur relative

$$\frac{\Delta y}{y} = |\varepsilon_1| \frac{x_1}{x_1 + x_2} + |\varepsilon_2| \frac{x_2}{x_1 + x_2} + |\varepsilon_3| \ .$$

Les deux premiers termes constituent l'erreur propagée, et le troisième provient de l'arrondi final:

$$\Delta y = \Delta_1 y + \Delta_2 y \qquad \begin{cases} \Delta_1 y = |\varepsilon_1| x_1 + |\varepsilon_2| x_2 \\ \Delta_2 y = |\varepsilon_3| y . \end{cases}$$

On remarquera que l'erreur propagée vérifie

$$\frac{\Delta_1 y}{y} = |\varepsilon_1| \frac{x_1}{x_1 + x_2} + |\varepsilon_2| \frac{x_2}{x_1 + x_2} \leqslant \sup(|\varepsilon_1|, |\varepsilon_2|) .$$

Par conséquent l'erreur relative propagée ne peut que diminuer.

2.2 - Dans le cas de la soustraction de deux nombres positifs, les choses se présentent différemment. En effet, pour

$$y = x_1 - x_2 \quad , \quad x_1 > x_2 ,$$

on obtient

$$\tilde{y} = (x_1(1 + \varepsilon_1) - x_2(1 + \varepsilon_2))(1 + \varepsilon_3)$$

$$= x_1(1 + \varepsilon_1 + \varepsilon_3) - x_2(1 + \varepsilon_2 + \varepsilon_3) + 0(\varepsilon^2)$$

et

$$\Delta y = |\varepsilon_1| x_1 + |\varepsilon_2| x_2 + |\varepsilon_3|(x_1 - x_2) ,$$

ce qui donne

$$\frac{\Delta y}{y} = \frac{|\varepsilon_1| x_1 + |\varepsilon_2| x_2}{x_1 - x_2} + |\varepsilon_3| .$$

L'erreur propagée vérifie dans ce cas

$$\frac{\Delta_1 y}{y} = \frac{|\varepsilon_1|(x_1 - x_2) + (|\varepsilon_1| + |\varepsilon_2|) x_2}{x_1 - x_2}$$

$$= |\varepsilon_1| + \frac{(|\varepsilon_1| + |\varepsilon_2|) x_2}{x_1 - x_2} > |\varepsilon_1| .$$

Bien plus, dans le cas où x_1 et x_2 sont voisins, c'est-à-dire que

$$x_1 - x_2 = \eta \ll x_2 ,$$

il vient

$$\frac{\Delta_1 y}{y} = \frac{|\varepsilon_1|(x_2 + \eta) + |\varepsilon_2| x_2}{\eta} \geqslant \frac{x_2}{\eta}(|\varepsilon_1| + |\varepsilon_2|),$$

et ce nombre peut devenir très grand. La soustraction de nombres voisins est donc instable numériquement.

2.3 - Examinons à présent la multiplication de deux nombres positifs. Pour $y = x_1 x_2$, on calcule

$$\tilde{y} = x_1(1 + \varepsilon_1)\, x_2(1 + \varepsilon_2)\, (1 + \varepsilon_3) =$$

$$= x_1 x_2 (1 + \varepsilon_1 + \varepsilon_2 + \varepsilon_3) + 0(\varepsilon^2),$$

donc

$$\frac{\Delta y}{y} = |\varepsilon_1| + |\varepsilon_2| + |\varepsilon_3| .$$

La partie propagée de l'erreur relative est donc la somme des erreurs relatives de départ. Ce n'est qu'après un grand nombre d'opérations que l'erreur accumulée peut devenir importante.

2.4 - La division a un comportement semblable: pour $y = x_1/x_2$, on calcule

$$\tilde{y} = \frac{x_1(1 + \varepsilon_1)}{x_2(1 + \varepsilon_2)} (1 + \varepsilon_3) = \frac{x_1}{x_2} (1 + \varepsilon_1 - \varepsilon_2 + \varepsilon_3) + 0(\varepsilon^2) ,$$

d'où

$$\frac{\Delta y}{y} = |\varepsilon_1| + |\varepsilon_2| + |\varepsilon_3|$$

3. ERREURS DANS LE CALCUL DE FONCTIONS PLUS ELABOREES

Soit à calculer une fonction $f(x_1, \ldots, x_n)$. La valeur calculée $\tilde{f}(\tilde{x}_1, \ldots, \tilde{x}_n)$ est affectée des erreurs suivantes:

- $\Delta_1 f$ = <u>erreur propagée</u> $= |f(\tilde{x}_1, \ldots, \tilde{x}_n) - f(x_1, \ldots, x_n)|$
- $\Delta_2 f$ = <u>erreur de calcul de f</u> $= |f_{calc}(\tilde{x}_1, \ldots, \tilde{x}_n) - f(\tilde{x}_1, \ldots, \tilde{x}_n)|$

  C'est dans ce terme qu'apparaissent les approximations de calcul et les erreurs d'arrondi en cours de route.
- $\Delta_3 f$ = <u>erreur d'arrondi finale</u>.

L'erreur $\Delta_3 f$ est généralement petite. L'erreur $\Delta_2 f$ dépend du soin apporté à l'algorithme de calcul de f. Quant à l'erreur propagée, elle a un caractère inéluctable. Son maximum au premier ordre est

$$\Delta_1 f = \sum_{i=1}^{n} \left|\frac{\partial f}{\partial x_i}\right| \Delta x_i .$$

Lorsque les dérivées premières sont nulles, l'erreur propagée est nulle au premier ordre, ce qui traduit une stabilité maximale. Par contre, il existe des fonctions pathologiques pour lesquelles l'arrondi est catastrophique. Ainsi, la fonction

$$f(x) = \exp\left(\frac{1}{|x-1|}\right) ,$$

pour $x = 0,99$, vaut $f(0,99) = \exp(100) \approx 10^{43}$. On a

$$f'(x) = \frac{-\operatorname{sign}(x-1)}{|x-1|^2} \exp\left(\frac{1}{|x-1|}\right) = 10^4 \exp(100) ,$$

et, pour $\Delta x = 10^{-3}$,

$$\Delta_1 f = |f'(x)| \Delta x = 10 \exp(100) = 10\, |f(x)| .$$

4. ALTERATION DES RELATIONS ALGEBRIQUES

Dans les études théoriques, on utilise constamment des propriétés telles que l'associativité de l'addition. Or, numériquement, ces propriétés ne sont pas vérifiées. Considérons le cas très simple de l'addition de quatre nombres. Pour simplifier au maximum, nous supposerons que ces quatre nombres sont connus exactement.

Calculons d'abord

$$y_1 = ((x_1 + x_2) + x_3) + x_4 .$$

On a successivement en mémoire

$$(x_1 + x_2)(1 + \varepsilon_1) \quad , \quad \varepsilon_1 = \text{erreur d'arrondi du résultat;}$$

$$((x_1 + x_2)(1 + \varepsilon_1) + x_3)(1 + \varepsilon_2) ;$$

$$(((x_1 + x_2)(1 + \varepsilon_1) + x_3)(1 + \varepsilon_2) + x_4)(1 + \varepsilon_3) = y_1 ,$$

soit

$$y_1 = (x_1 + x_2 + x_3 + x_4) + \varepsilon_1(x_1 + x_2) + \varepsilon_2(x_1 + x_2 + x_3) + \varepsilon_3(x_1 + x_2 + x_3 + x_4) + 0(\varepsilon^2) .$$

On peut aussi calculer

$$y_2 = (x_1 + x_2) + (x_3 + x_4) .$$

Les étapes numériques sont

$$(x_1 + x_2)(1 + \varepsilon_4)$$

$$(x_3 + x_4)(1 + \varepsilon_5)$$

$$((x_1 + x_2)(1 + \varepsilon_4) + (x_3 + x_4)(1 + \varepsilon_5))(1 + \varepsilon_6) = y_2 ,$$

soit

$$y_2 = (x_1 + x_2 + x_3 + x_4) + \varepsilon_4(x_1 + x_2) + \varepsilon_5(x_3 + x_4) + \varepsilon_6(x_1 + x_2 + x_3 + x_4) + 0(\varepsilon^2) .$$

Les résultats sont donc différents. Si l'on suppose que toutes les erreurs relatives d'arrondi sont égales, il vient

$$\Delta y_1 = |\varepsilon|(3 x_1 + 3 x_2 + 2 x_3 + x_4)$$

$$\Delta y_2 = |\varepsilon|(2 x_1 + 2 x_2 + 2 x_3 + 2 x_4)$$

et, pour autant que

$$x_1 + x_2 > x_4 ,$$

le second procédé est meilleur.

5. ARTIFICES DE CALCUL

Dans bien des cas, on peut transformer un calcul instable en un calcul stable par quelques manipulations. Soit par exemple à résoudre l'équation du second degré

$$1,1.10^{-4} x^2 + 2 x - 1,30 = 0 .$$

Par la formule classique

$$x = \frac{-\frac{b}{2} \pm \sqrt{\left(\frac{b}{2}\right)^2 - ac}}{a} ,$$

on obtient, pour la racine positive,

$$x = \frac{-1 + \sqrt{1,000143}}{1,1.10^{-4}} = \frac{-1 + 1,0000714}{1,1.10^{-4}} = 0,64997272 .$$

Mais tant que l'on effectue les calculs avec moins de six chiffres significatifs, on obtient

$$\tilde{x} = 0 ,$$

soit une erreur de 100%. Une première manière de contourner cette difficulté consiste à négliger le premier terme du trinôme: on obtient alors

$$\tilde{x} = 0,650,$$

et la réintroduction de cette valeur dans l'équation montre qu'effectivement,

$$1,1.10^{-4} x^2 = 0,46.10^{-4} ,$$

c'est-à-dire que le terme négligé est bien petit. Mais d'une manière plus générale, on peut transformer la formule de calcul de x en multipliant le numérateur et le dénominateur par

$$\left(-\frac{b}{2} - \sqrt{\left(\frac{b}{2}\right)^2 - ac}\right) ,$$

ce qui donne

$$x = \frac{c}{\frac{b}{2} + \sqrt{\left(\frac{b}{2}\right)^2 - ac}} \quad ,$$

formule où la soustraction instable a disparu. Dans notre cas, ce procédé donne $x \approx 0{,}65$ si l'on calcule avec trois chiffres significatifs seulement.

6. PROBLEMES DE CONVERGENCE

Soit à calculer la somme d'une série convergente

$$S = \sum_{n=1}^{\infty} x_n \ .$$

Pratiquement, on calcule successivement les sommes partielles

$$S_p = \sum_{n=1}^{p} x_n$$

par la relation de récurrence évidente

$$S_{p+1} = S_p + x_{p+1} \ .$$

Mais tous ces calculs sont arrondis. A la place de S_p, on dispose donc d'une valeur approchée $\tilde{S}_p$; à la place de x_{p+1}, on calcule en fait $\tilde{x}_{p+1} = x_{p+1}(1 + \varepsilon_{1,p+1})$; enfin, à la place de S_{p+1}, on calcule

$$\begin{aligned} \tilde{S}_{p+1} &= (\tilde{S}_p + x_{p+1}(1 + \varepsilon_{1,p+1}))(1 + \varepsilon_{2,p+1}) \\ &= (S_p + S_p + x_{p+1} + \varepsilon_{1,p+1}\, x_{p+1})(1 + \varepsilon_{2,p+1}) \\ &\quad S_{p+1}(1 + \varepsilon_{2,p+1}) + \Delta S_p + \varepsilon_{1,p+1}\, x_{p+1} \ . \end{aligned}$$

On a donc, au premier ordre,

$$\Delta S_{p+1} = \Delta S_p + \varepsilon_{2,p+1}\, S_{p+1} + \varepsilon_{1,p+1}\, x_{p+1} \ .$$

Supposons que les erreurs relatives admettent les majorations suivantes:

- Erreurs d'évaluation des termes: $|\varepsilon_{1,p+1}| \leq \varepsilon_1$
- Arrondis de sommation : $|\varepsilon_{2,p+2}| \leq \varepsilon_2$

Il vient alors

$$|\Delta S_{p+1}| \leq |\Delta S_p| + |\varepsilon_2|\,|S_{p+1}| + |\varepsilon_1|\,|x_{p+1}| \ .$$

La série étant convergente, il existe une majoration du type

$$|S_{p+1}| \leq A \quad ,$$

et on obtient

$$|\Delta S_p| \leq p\,|\varepsilon_2|\,A + |\varepsilon_1| \sum_{n=1}^{p} |x_p| \quad .$$

On constate que

a) Le dernier terme de cette majoration ne converge que si la série est absolument convergente. Les séries simplement convergentes sont donc instables numériquement.

b) Le premier terme de la majoration diverge toujours, mais lentement. Il est donc nécessaire que la série converge assez vite pour que cette divergence par accumulation d'arrondis ne se fasse pas sentir.

7. RECHERCHE DES POINTS OU UNE FONCTION PREND UNE VALEUR DONNEE

Supposons que l'on désire connaître les points ξ où une certaine fonction f prend la valeur b. La fonction est calculée par un programme menant, au voisinage de ξ , à une erreur maximale Δf sur la fonction. On peut donc s'attendre à trouver les points $(x, \tilde{f}(x))$ dans une bande de hauteur $2\Delta f$ (fig. 1) . Il en découle que l'ensemble e(b) où f(x) peut valoir b est

$$e(b) = \left\{ x \mid |f(x) - b| \leq \Delta f \right\} \quad .$$

Or, si $f(\xi) = b$, et si f est différentiable, on a

$$f(x) = f(\xi) + (x - \xi)\, f'(x^*) \quad ,$$

avec x^* compris entre ξ et x. Il vient donc

$$|x - \xi| \leq \frac{|f(x) - b|}{\inf\limits_{\overline{V}(\xi)} |f'(x)|} \quad ,$$

$\overline{V}(\xi)$ étant un voisinage fermé de ξ contenant x. L'ensemble e(b) est donc contenu dans

$$e^*(b) = \left\{ x \mid |x - \xi| \leq \frac{\Delta f}{\inf\limits_{\overline{V}(\xi)} |f'(x)|} \right\} \quad .$$

Lorsque $f'(\xi)$ est nulle, il faut pousser le développement de TAYLOR à l'ordre 2; plus généralement, si les dérivées jusqu'à l'ordre (p-1) de f sont nulles en ξ , on a

$$f(x) = b + (x - \xi)^p\, f^{(p)}(x^*) \quad ,$$

d'où

$$|\tilde{\xi} - \xi| \leqslant \left(\frac{\Delta f}{\underset{v(\xi)}{\inf} |f^{(p)}(x)|} \right)^{1/p} ,$$

valeur en général beaucoup plus grande (fig. 2) .

Exercice 1 - Soit A une matrice symétrique. On veut calculer le produit

$$y = A\,x\ .$$

On appelle conditionnement de la matrice A le rapport de la plus grande en module à la plus petite en module des valeurs propres de A:

$$\text{conditionnement} = \mathcal{K} = \frac{|\lambda|_{max}}{|\lambda|_{min}}$$

Montrer que si x subit une perturbation δx, la perturbation correspondante δy de y vérifie

$$\frac{\|\delta y\|}{\|y\|} \leqslant \mathcal{K}\ \frac{\|\delta x\|}{\|x\|}$$

(il s'agit de normes euclidiennes)

Solution: voir chapitre relatif à la résolution des systèmes linéaires.

Exercice 2 - On donne une table des logarithmes des sinus à 5 décimales et une table des logarithmes des tangentes à 5 décimales. (On comptera, dans les deux cas, sur une erreur absolue maximale de 5.10^{-6}). Les procédés d'interpolation donnés dans la marge des tables permettent de préserver la même précision lors de l'interpolation. Des deux calculs $\log \operatorname{tg} x \longmapsto x$ et $\log \sin x \longmapsto x$, lequel permet de déterminer x avec le plus de précision, et que vaut cette dernière dans le cas le plus défavorable?

Solution: par application de la section 7 , calculons

$$\frac{d}{dx}(\log \operatorname{tg} x) = \frac{1}{\ln 10} \cdot \frac{1}{\operatorname{tg} x}(1 + \operatorname{tg}^2 x) = \frac{1}{\ln 10} \cdot \frac{2}{\sin 2x}$$

$$\frac{d}{dx}(\log \sin x) = \frac{1}{\ln 10} \cdot \frac{1}{\operatorname{tg} x}$$

Pour une erreur Δf , on a donc au premier ordre

$$\Delta x \approx \frac{\Delta f}{2} \sin 2x \ln 10 \quad \text{pour} \quad \log \operatorname{tg}$$

et

$$\Delta x \approx \Delta f \ \operatorname{tg} x \ \ln 10 \quad \text{pou} \quad \log \ \sin$$

Comme on a toujours $\frac{\sin 2x}{2} \leqslant \operatorname{tg} x$, c'est le log tg qui permet de trouver x avec le plus de précision. Pour le log sin, aucune précision ne peut être garantie si $x \approx \pi/2$ (Cela est dû au fait que le sinus est stationnaire en $\pi/2$). Pour le log tg, l'erreur sur x est toujours

inférieure à $\Delta f \cdot \frac{\ln 10}{2} = 5{,}756.10^{-6}$, ce qui est excellent.

Exercice 3 - Soit à résoudre l'équation $F(x) = x$. Le calcul de la fonction F est affecté d'une erreur relative maximale ε_o . Quelle précision peut-on attendre sur la solution ξ ?

Solution: La solution calculée vérifie

$$\tilde{\xi} = \tilde{F}(\tilde{\xi}) = F(\tilde{\xi}) + \delta F(\tilde{\xi}) \approx F(\xi) + F'(\xi)(\tilde{\xi} - \xi) + \delta F(\tilde{\xi}) ,$$

soit

$$\tilde{\xi}\,[\,1 - F'(\xi)\,] \approx \xi\,[\,1 - F'(\xi)\,] + \delta F(\tilde{\xi})$$

et

$$\tilde{\xi} - \xi \approx \frac{F(\tilde{\xi})}{1 - F'(\xi)} .$$

En première approximation, $|\delta F(\tilde{\xi})| \leq \varepsilon_o |F(\tilde{\xi})| \approx \varepsilon_o |\xi|$ et

$$|\delta\xi| \lesssim \frac{\varepsilon_o |\xi|}{|1 - F'(\xi)|}$$

soit

$$\left|\frac{\delta\xi}{\xi}\right| \lesssim \frac{\varepsilon_o}{|1 - F'(\xi)|}$$

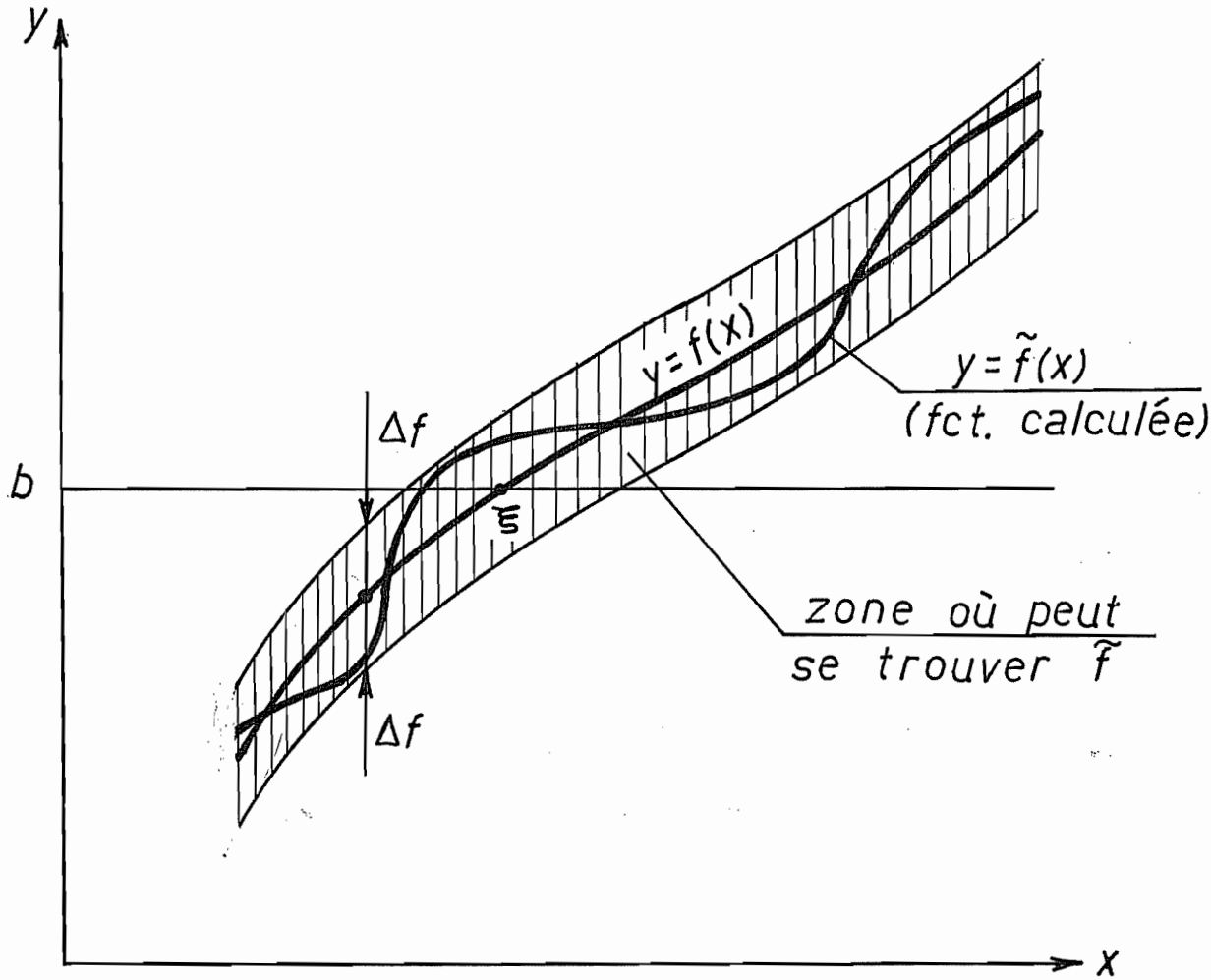


Fig. 1

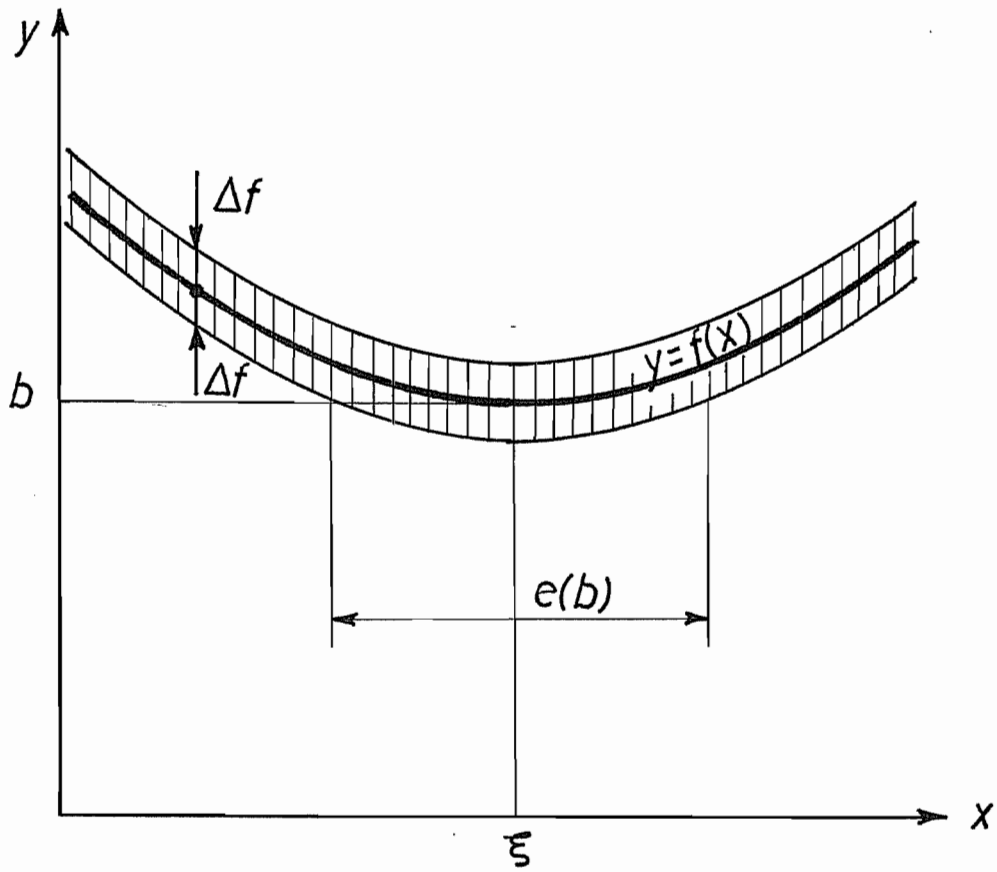


Fig. 2

I N T E R P O L A T I O N

1. Il n'est pas rare que l'on soit obligé de recourir à l'approximation d'une fonction. Un exemple familier est celui des fonctions spéciales. Il existe des tables de ces fonctions [5] , où l'on trouve leurs valeurs en un certain nombre de points. Souvent, l'établissement de ces tables est une tâche très ardue, et il ne peut être question de refaire leur calcul pour un besoin passager. C'est ainsi qu'il existe des tables de la fonction de BESSEL J_0 définie comme la solution de l'équation différentielle

$$\frac{d^2f}{dx^2} + \frac{1}{x}\frac{df}{dx} + f = 0 \ ,$$

avec $f(0) = 1$, $f'(0) = 0$. Supposons que l'on possède une table de $J_0(x)$ pour x allant de 0 à 15, par intervalles de 0,1 . Les tables sont établies avec 15 décimales. Comment obtenir, par exemple, $J_0(1{,}73)$? On peut évidemment dire

$$J_0(1{,}73) \quad J_0(1{,}7) = 0{,}39798\ldots \ ,$$

ce qui revient en fait à remplacer la fonctions par des escaliers (fig.1). Souvent, on a recours à l'interpolation linéaire:

$$J_0(1{,}73) \quad J_0(1{,}70) + \frac{0{,}03}{0{,}1}\left(J_0(1{,}80) - J_0(1{,}70)\right),$$

ce qui équivaut en fait à remplacer, entre 1,70 et 1,80, la fonction J_0 par un segment de droite (fig. 2). Mais on peut obtenir un meilleur résultat en remplaçant J_0 par la parabole passant par les points

$$(1{,}7 \ , \ J_0(1{,}7)) \quad , \ (1{,}8 \ , \ J_0(1{,}8)) \quad , \ (1{,}9 \ , \ J_0(1{,}9))$$

(fig.3). On peut d'ailleurs remplacer J_0 par des polynômes de degré plus élevé encore. Dans le cas considéré, l'erreur d'une interpolation linéaire aurait été de l'ordre de 6.10^{-4}. On retrouve la précision de la table en interpolant sur 11 points, à l'aide d'un polynôme de degré 10.

L'interpolation joue également un rôle important dans la résolution des équations et l'intégration numérique .

Bien que l'interpolation polynomiale soit la plus courante, il existe des situations où elle ne convient pas. Soit par exemple une fonction f dont on sait que (voir fig. 4)

$$\lim_{x \downarrow 0} \frac{f(x)}{\sqrt{x}} \quad \begin{cases} \neq 0 \\ \neq \infty \end{cases}$$

Il est naturel de chercher une expression du genre

$$\tilde{f}(x) = A\sqrt{x} + B\,x + C\,x\sqrt{x} \ ,$$

par exemple. Il s'agit en l'occurrence d'un polynôme en $\sqrt{x}$. On peut encore imaginer bien d'autres situations. Soit f une fonction telle que

$$\lim_{x \downarrow 0} f(x)/x \begin{cases} \neq 0 \\ \neq \infty \end{cases}$$

et

$$\lim_{x \to \infty} x^2 f(x) \begin{cases} \neq 0 \\ \neq \infty \end{cases}$$

(fig. 5). On peut chercher une expression de la forme

$$\tilde{f}(x) = \frac{x}{a + bx + cx^2 + dx^3}$$

qui a effectivement ces propriétés. On en déduit

$$\varphi(x) = \frac{x}{\tilde{f}(x)} = a + bx + cx^2 + dx^3 \quad ,$$

ce qui ramène le problème à une interpolation polynomiale après un changement de fonction.

Lorsque la fonction présente une asymptote horizontale, de bons résultats peuvent souvent être obtenus à l'aide de fractions rationnelles de la forme

$$\tilde{f}(x) = \frac{a_o + a_1 x + a_2 x^2 + \ldots + a_n x^n}{1 + b_1 x + b_2 x^2 + \ldots + b_n x^n} \quad ,$$

ce qui revient à écrire

$$\tilde{f}(x) = a_o + a_1 x + \ldots + a_n x^n - b_1 x\tilde{f}(x) - \ldots - b_n x^n \tilde{f}(x),$$

soit une combinaison linéaire des fonctions

$$1, x, \ldots, x^n, x\tilde{f}(x), \ldots, x^n\tilde{f}(x).$$

Les fonctions périodiques, de période $T = 2\pi/\omega$ peuvent être approchées par des expressions trigonométriques:

$$f(x) = a_o + a_1 \cos \omega x + b_1 \sin \omega x + a_2 \cos 2\omega x + b_2 \sin 2\omega x + \ldots + a_n \cos n\omega x + b_n \sin n\omega x.$$

Lorsque la fonction présente en x=b une asymptote verticale, avec

$$\lim_{x \to b} (x-b)^p f(x) \begin{cases} \neq 0 \\ \neq \infty \end{cases} \quad ,$$

on peut utiliser une expression de la forme

$$\tilde{f}(x) = \frac{a_o}{(x-b)^p} + \frac{a_1}{(x-b)^{p-1}} + \ldots + a_p + a_{p+1}(x-b) + \ldots + a_n(x-b)^{n-p} \quad ,$$

ce qui équivaut à une interpolation polynomiale de $(x-b)^p f(x)$.

Les situations sont donc multiples, et le choix des fonctions entrant dans l'approximation est avant tout une question de doigté et d'opportunité, dans laquelle une certaine expérience et quelques essais jouent souvent un rôle prépondérant.

2. INTERPOLATION PAR UNE COMBINAISON LINEAIRE DE FONCTIONS

Le plus souvent, on cherche une interpolée de la forme

$$\tilde{f}(x) = \alpha_o \varphi_o(x) + \ldots + \alpha_n \varphi_n(x)$$

où $\varphi_o, \ldots, \varphi_n$ sont des fonctions choisies d'avance selon les critères évoqués ci-dessus et $\alpha_o, \ldots, \alpha_n$, des coefficients inconnus.

Ces fonctions $\varphi_o, \ldots, \varphi_n$ forment la base d'interpolation. La raison de cette appellation est que l'ensemble de leurs combinaisons linéaires est un espace vectoriel, l'espace d'interpolation, dont elles forment une base. On détermine les (n+1) coefficients inconnus par la condition qu'en (n+1) points, l'interpolée ait la même valeur que la fonction f à interpoler:

$$\tilde{f}(x_o) = f(x_o)$$
$$\ldots\ldots$$
$$\tilde{f}(x_n) = f(x_n).$$

C'est la condition d'interpolation. L'ensemble des points d'interpolation $x_o, \ldots, x_n$ s'appelle support d'interpolation. Pour éviter toute difficulté, il convient de supposer f et les φ_i continues sur l'ensemble $[a,b]$ où l'on désire calculer f .

La condition d'interpolation s'écrit explicitement

$$\begin{cases} \alpha_o \varphi_o(x_o) + \alpha_1 \varphi_1(x_o) + \ldots + \alpha_n \varphi_n(x_o) = f(x_o) \\ \alpha_o \varphi_o(x_1) + \alpha_1 \varphi_1(x_1) + \ldots + \alpha_n \varphi_n(x_1) = f(x_1) \\ \qquad\qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \alpha_o \varphi_o(x_n) + \alpha_1 \varphi_1(x_n) + \ldots + \alpha_n \varphi_n(x_n) = f(x_n) \end{cases}$$

Appelant C la matrice de connexion

$$C = \begin{bmatrix} \varphi_o(x_o) & \varphi_1(x_o) & \ldots\ldots & \varphi_n(x_o) \\ \varphi_o(x_1) & \varphi_1(x_1) & \ldots\ldots & \varphi_n(x_1) \\ \vdots & \vdots & & \vdots \\ \varphi_o(x_n) & \varphi_1(x_n) & \ldots\ldots & \varphi_n(x_n) \end{bmatrix},$$

a et q, les vecteurs

$$a = \begin{bmatrix} \alpha_o \\ \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \qquad q = \begin{bmatrix} f(x_o) \\ f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} ,$$

il nous faut donc résoudre le système matriciel

$$C\ a = q$$

Ce système admettra une solution unique si et seulement si la matrice C est inversible. Supposons qu'elle ne le soit pas: il existe alors une relation linéaire entre ses colonnes, de la forme

$$\begin{cases} \alpha_o\ \varphi_o(x_o) + \alpha_1\ \varphi_1(x_o) + \ldots + \alpha_n\ \varphi_n(x_o) = 0 \\ \alpha_o\ \varphi_o(x_1) + \alpha_1\ \varphi_1(x_1) + \ldots + \alpha_n\ \varphi_n(x_1) = 0 \\ \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \alpha_o\ \varphi_o(x_n) + \alpha_1\ \varphi_1(x_n) + \ldots + \alpha_n\ \varphi_n(x_n) = 0 \end{cases} .$$

Ceci revient à dire qu'une certaine combinaison linéaire des fonctions de base s'annule en tous les points du support d'interpolation. La négation de cette propriété constitue la <u>condition de TCHEBICHEFF:</u>

<u>Il ne peut exister de combinaison linéaire des (n+1) fonctions de base qui s'annule en tous les points du support d'interpolation.</u>

Lorsque cette condition est remplie, les (n+1) paramètres $\alpha_o, \alpha_1, \ldots, \alpha_n$ s'obtiennent par

$$a = C^{-1} q .$$

Introduisant alors le vecteur

$$g(x) = \begin{bmatrix} \varphi_o(x) \\ \varphi_1(x) \\ \vdots \\ \varphi_n(x) \end{bmatrix} ,$$

on peut écrire

$$\tilde{f}(x) = \alpha_o\ \varphi_o(x) + \ldots + \alpha_n\ \varphi_n(x) = q^T\ C^{-T}\ g(x) ,$$

ce qui fait apparaître le nouveau vecteur

$$l(x) = C^{-T}\, g(x) = \begin{bmatrix} L_o(x) \\ L_1(x) \\ \cdot \\ \cdot \\ \cdot \\ L_n(x) \end{bmatrix} \quad ,$$

tel que

$$\tilde{f}(x) = q^T\, l(x) = f(x_o)\, L_o(x) + \ldots + f(x_n)\, L_n(x) \ .$$

Les fonctions $L_o, L_1, \ldots, L_n$ sont linéairement indépendantes, sans quoi il existerait un ensemble de valeurs $f(x_o), \ldots, f(x_n)$ non toutes nulles et telles que $f(x) \equiv 0$, ce qui contredirait la condition d'interpolation. Ces fonctions forment donc une nouvelle base de l'espace d'interpolation. Nous l'appellerons <u>base de LAGRANGE</u> (généralisée) Cette base possède la propriété fondamentale

$$L_i(x_j) = \delta_{ij} \qquad (1)$$

En effet, pour $f(x) = L_k(x)$, on doit avoir $f(x) = L_k(x)$, en vertu de l'unicité de l'interpolation. Il en découle

$$L_k(x) = f(x_o)\, L_o(x) + \ldots + f(x_k)\, L_k(x) + \ldots + f(x_n)\, L_n(x) \ ,$$

c'est-à-dire la nullité de la combinaison

$$f(x_o)\, L_o(x) + \ldots + f(x_{k-1})\, L_{k-1}(x) + (f(x_k) - 1)\, L_k(x) +$$

$$+ f(x_{k+1})\, L_{k+1}(x) + \ldots + f(x_n)\, L_n(x) = 0 \ ,$$

ce qui implique

$$f(x_i) = 0 \quad , \quad i \neq j$$

$$f(x_k) = 1 \quad ,$$

puisque les fonctions L_i sont linéairement indépendantes. On remarquera d'ailleurs que les conditions (1) définissent les L_i de manière univoque.

3. <u>INTERPOLATION POLYNOMIALE : FORMULE DE LAGRANGE</u>

On rencontre le plus souvent l'interpolation <u>polynomiale</u>, qui correspond à la base d'interpolation

$$\{ 1, x, x^2, \ldots, x^n \} \quad .$$

L'espace d'interpolation est donc l'ensemble $\mathcal{P}_n$ des polynômes de degré n (au plus) de la forme

$$\tilde{f}(x) = \alpha_o + \alpha_1 x + \alpha_2 x^2 + \ldots + \alpha_n x^n \quad .$$

La condition de Tchébicheff est automatiquement remplie si les points du support sont tous distincts, car un polynôme de degré n qui s'annule aux n points x_o , x_1, ..., x_{n-1} a nécessairement la forme

$$h(x) = A(x-x_o)\ldots(x-x_{n-1}) \quad , \quad A \in R \quad ;$$

pour qu'il s'annule encore en x_n, il faudra que

$$h(x_n) = A(x_n-x_o)\ldots(x_n-x_{n-1}) = 0 \ ,$$

ce qui implique $A = 0$, puisque tous les points du support sont différents.

En fait, on obtient très aisément la base de Lagrange en développant l'interpolée sous la forme

$$\tilde{f}(x) = \beta_o \prod_{\neq 0} (x-x_j) + \beta_1 \prod_{j\neq 1} (x-x_j) + \ldots + \beta_n \prod_{j\neq n} (x-x_j) \ .$$

On obtient alors

$$\tilde{f}(x_i) = \beta_i \prod_{j\neq i} (x_i-x_j) = f(x_i),$$

d'où

$$\tilde{f}(x) = \sum_{i=0}^{n} f(x_i) \frac{\prod_{j\neq i} (x-x_j)}{\prod_{j\neq i} (x_i-x_j)} = \sum_{i=0}^{n} L_i(x)\, f(x_i) \quad ,$$

avec

$$\boxed{L_i(x) = \frac{\prod_{j\neq i} (x-x_j)}{\prod_{j\neq i} (x_i-x_j)}}$$

C'est la base de Lagrange de l'interpolation polynomiale.

On peut encore donner une autre forme aux fonctions L_i. A cette fin, introduisons le polynôme de degré (n+1)

$$\prod(x) = \prod_{i=0}^{n} (x-x_i) \ .$$

Le numérateur de l'expression de L_i s'écrit

$$\prod_{j\neq i} (x-x_j) = \frac{\prod_{j=0}^{n} (x-x_j)}{x-x_i} = \frac{\prod(x)}{x - x_i} \quad ;$$

le dénominateur est sa valeur en $x = x_i$, soit

$$\prod_{j \neq i} (x_i - x_j) = \lim_{x \to x_i} \frac{\prod(x)}{x - x_i}$$

et, comme $\prod(x_i) = 0$, c'est encore

$$\lim_{x \to x_i} \frac{\prod(x) - \prod(x_i)}{x - x_i} = \prod'(x_i) .$$

Ainsi,

$$\boxed{L_i(x) = \frac{\prod(x)}{(x-x_i)\prod'(x_i)}}$$

4. INTERPOLATION POLYNOMIALE: FORMULE DE NEWTON

La formule de NEWTON est une autre expression de l'interpolée. Bien entendu, on obtient la même interpolée que par la formule de Lagrange, car l'interpolée est unique. Mais l'expression de Newton présente la particularité de se présenter comme une généralisation de la formule de Taylor. Cette dernière fait intervenir les dérivées qui, dans la formule de Newton, sont remplacées par des différences divisées.

4.1 - Différences divisées

$x_0, x_1, \ldots, x_n$ étant les points du support d'interpolation, on construit les différences divisées comme suit:

- ordre 0 : $f(x_0)$
- ordre 1 : $f(x_0, x_1) = \dfrac{f(x_0) - f(x_1)}{x_0 - x_1}$
- ordre 2 : $f(x_0, x_1, x_2) = \dfrac{f(x_0, x_1) - f(x_1, x_2)}{x_0 - x_2}$

. . . . . .

- ordre n : $f(x_0, x_1, \ldots, x_n) = \dfrac{f(x_0,x_1,\ldots,x_{n-1}) - f(x_1,\ldots,x_n)}{x_0 - x_n}$

4.2 - Formule de structure des différences divisées

Les différences divisées possèdent la propriété fondamentale suivante:

$$f(x_o, x_1, \ldots, x_n) = \sum_{k=0}^{n} \frac{f(x_k)}{\prod_{j\neq k} (x_k - x_j)}$$

On démontre cette propriété par récurrence sur l'ordre des différences divisées. Pour n = 1, c'est évident, car

$$f(x_o, x_1) = \frac{f(x_o) - f\, x_1)}{x_o - x_1} = \frac{f(x_o)}{x_o - x_1} + \frac{f(x_1)}{x_1 - x_o} \quad .$$

Supposons donc la propriété vraie jusqu'à lordre (n-1) et étendons-la à l'ordre n. On a par définition

$$f(x_o,x_1,\ldots,x_n) = \frac{f(x_o,x_1,\ldots,x_{n-1}) - f(x_1,\ldots,x_n)}{x_o - x_n} =$$

$$\frac{1}{x_o-x_n} \left\{ \sum_{k=0}^{n-1} \frac{f(x_k)}{\prod_{\substack{j=0 \\ \neq k}}^{n-1} (x_k-x_j)} - \sum_{k=1}^{n} \frac{f(x_k)}{\prod_{\substack{j=1 \\ \neq k}}^{n} (x_k-x_j)} \right\} =$$

$$\frac{1}{x_o-x_n} \left\{ \sum_{k=0}^{n-1} \frac{f(x_k)(x_k-x_n)}{\prod_{\substack{j=0 \\ \neq k}}^{n} (x_k-x_j)} - \sum_{k=1}^{n} \frac{f(x_k)(x_k-x_o)}{\prod_{\substack{j=0 \\ \neq k}}^{n} (x_k-x_j)} \right\} =$$

$$\frac{1}{x_o-x_n} \left\{ \frac{f(x_o)(x_o-x_n)}{\prod_{\substack{j=0 \\ \neq 0}}^{n} (x_o-x_j)} + \sum_{k=1}^{n-1} \frac{f(x_k)(x_o-x_n)}{\prod_{\substack{j=0 \\ \neq k}}^{n} (x_k-x_j)} - \frac{f(x_n)(x_n-x_o)}{\prod_{\substack{j=0 \\ \neq n}}^{n} (x_n-x_j)} \right\} =$$

$$= \sum_{k=0}^{n} \frac{f(x_k)}{\prod_{\substack{j=0 \\ \neq k}}^{n} (x_k-x_j)} \quad ,$$

ce qui démontre la proposition.

On en déduit les deux propriétés suivantes:

a) <u>L'ordre des arguments d'une différence divisée est indifférent</u>. En effet, les termes restent identiques et seul change leur ordre dans la somme.

b) <u>Les différences divisées sont linéaires:</u>

$$(\lambda f + \mu g)(x_o,x_1,\ldots,x_n) = \sum_{k=0}^{n} \frac{\lambda f(x_k) + \mu g(x_k)}{\prod \ldots\ldots} =$$

$$= \sum_{k=0}^{n} \frac{f(x_k)}{\prod \cdots} + \sum_{k=0}^{n} \frac{g(x_k)}{\prod \cdots} = f(x_0,\ldots,x_n) + g(x_0,\ldots,x_n).$$

4.3 - Tableau des différences divisées

Le calcul des différences divisées se systématise à l'aide du tableau suivant:

x	f_0	f_1	f_2	f_3	f_4
x_0	$f(x_0)$	$f(x_0,x_1)$	$f(x_0,x_1,x_2)$	$f(x_0,x_1,x_2,x_3)$	$f(x_0,x_1,x_2,x_3,x_4)$
x_1	$f(x_1)$	$f(x_1,x_2)$	$f(x_1,x_2,x_3)$	$f(x_1,x_2,x_3,x_4)$	
x_2	$f(x_2)$	$f(x_2,x_3)$	$f(x_2,x_3,x_4)$		
x_3	$f(x_3)$	$f(x_3,x_4)$			
x_4	$f(x_4)$				

On calcule d'abord les différences d'ordre 0, puis celles d'ordre 1, puis celles d'ordre 2, etc..., par les formules de définition.

4.4 - Développement d'une fonction en différences divisées

On a successivement

$$f(x) = f(x_0) + \frac{f(x)-f(x_0)}{x-x_0}(x-x_0) = f(x_0) + (x-x_0)\, f(x_0,x)$$

$$f(x_0,x) = f(x_0,x_1) + \frac{f(x_0,x)-f(x_0,x_1)}{x-x_1}(x-x_1) = f(x_0,x_1)+(x-x_1)f(x_0,x_1,x)$$

. . . . . .

$$f(x_0,\ldots,x_{n-1},x) = f(x_0,\ldots,x_{n-1},x_n) + \frac{f(x_0,\ldots,x_{n-1},x) - f(x_0,\ldots,x_{n-1},x_n)}{x - x_n}(x - x_n)$$

$$= f(x_0,\ldots,x_{n-1},x_n) + (x-x_n)f(x_0,\ldots,x_n,x) .$$

Rassemblant ces résultats, on obtient

$$f(x) = f(x_o) + (x-x_o)f(x_o,x_1) + (x-x_o)(x-x_1)f(x_o,x_1,x_2) + \ldots$$
$$\ldots + (x-x_o)\ldots(x-x_{n-1})f(x_o,\ldots,x_n) + (x-x_o)\ldots(x-x_n)f(x_o,\ldots,x_n,x)$$

4.5 - Polynôme d'interpolation de Newton

Examinons le dernier terme du développement: on a

$$f(x_o,\ldots,x_n,x) = \sum_{i=0}^{n} \frac{f(x_i)}{(x_i-x)\prod_{j\neq i}(x_i-x_j)} + \frac{f(x)}{\prod_j (x-x_j)} ,$$

ce qui entraîne

$$\prod_j (x-x_j) \cdot f(x_o,\ldots,x_n,x) = \sum_{i=0}^{n} f(x_i) \frac{x-x_i}{x_i-x} \prod_{j\neq i} \frac{x-x_j}{x_i-x_j} + f(x)$$
$$= -\tilde{f}(x) + f(x) ,$$

où $\tilde{f}$ est l'interpolée de Lagrange de f. Par identification, on obtient donc

$$\tilde{f}(x) = f(x_o) + (x-x_o)f(x_o,x_1) + (x_o-x_1)f(x_o,x_1,x_2) + \ldots$$
$$\ldots + (x-x_o)\ldots(x-x_{n-1})f(x_o,\ldots,x_n)$$

avec un reste

$$R(x) = (x-x_o)\ldots(x-x_n)f(x_o,\ldots,x_n,x)$$

On donne à l'interpolée ainsi construite le nom de polynôme d'interpolation de Newton. Cette expression se prête particulièrement bien au calcul numérique, par l'algorithme ci-dessous:

- On calcule une fois pour toutes
  $f_o = f(x_o)$, $f_1 = f(x_o,x_1)$, ... , $f_n = f(x_o,x_1,\ldots,x_n)$.
- Pour x donné, on calcule
  $X_o = (x-x_o)$, $X_1 = (x-x_1)$, ... , $X_n = (x-x_n)$
- On effectue les calculs comme suit:

$$\begin{cases} y \longleftarrow f_n \\ \text{Pour } k = (n-1), \ldots, 0 , \quad y \longleftarrow y \cdot X_k + f_k . \end{cases}$$

5. ERREUR D'INTERPOLATION

5.1 - Nous avons déjà que l'erreur d'interpolation s'écrit

$$R(x) = (x-x_0)\ldots(x-x_n)f(x_0,\ldots,x_n,x) \ .$$

Mais il est possible de l'exprimer en termes d'une dérivée de f, pour autant que $f \in C^{n+1}$ sur l'intervalle $[a,b]$ contenant $x_0,\ldots,x_n$ et x. A cette fin, considérons la fonction auxiliaire

$$g(y) = f(y) - f(y) - (y-x_0)\ldots(y-x_n)f(x_0,x_1,\ldots,x_n,x) \ .$$

On a visiblement $g(y)=0$ en $x_0,\ldots,x_n$ et x, soit en (n+2) points. Par des applications successives du théorème de Rolle, on obtient que

$g'(y)$ s'annule en (n+1) points intermédiaires
$g''(y)$ s'annule en n points intermédiaires
. . . . . .
$g^{(n+1)}(y)$ s'annule en un point intermédiaire au moins, soit ξ .

En ce point, comme f est un polynôme de degré n, on a

$$0 = g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - 0 - (n+1)!\ f(x_0,\ldots,x_n,x),$$

soit

$$\boxed{f(x_0,\ldots,x_n,x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \quad , \xi \in \ ]\inf(x_0,\ldots,x_n,x),\ \sup(x_0,\ldots,x_n,x)[}$$

Nous venons en fait de démontrer le résultat général que voici:
Une différence divisée d'ordre n est, pour autant que la fonction considérée soit de classe C^n, égale à 1/n! fois la valeur de sa dérivée d'ordre n en un point situé entre les deux points extrêmes de calcul de la différence divisée.

En posant

$$\Pi(x) = (x-x_0)\ldots(x-x_n) \ ,$$

on obtient

$$\boxed{R(x) = \prod(x)\ \frac{f^{(n+1)}(\xi)}{(n+1)!}} \quad ,$$

résultat dont on déduit la majoration

$$\boxed{|R(x)| \leq |\Pi(x)|\ \sup_{z \in [a,b]}\ \frac{f^{(n+1)}(z)}{(n+1)!}}$$

si $x_0, \ldots, x_n, x \in [a,b]$.

Remarque - Si f est un polynôme de degré n, $f^{(n+1)} = 0$, ce qui entraîne la nullité des différences divisées d'ordre $\geq (n+1)$.

5.2 - Convergence en h

Il est aisé de voir que, si h = b-a,

$$|\prod(x)| = |x-x_0|\ldots|x-x_n| \leq h^{n+1} .$$

Soit donc $f \in C^{n+1}(\,]a,b[\,)$ et supposons que l'on découpe cet intervalle en N sous-intervalles $e_1, \ldots, e_N$, de longueur $h = (b-a)/N$. Dans chacun de ceux-ci, on utilise une formule d'interpolation à (n+1) points. Alors, l'erreur sur le sous-intervalle e_k vérifie

$$|R_{e_k}(x)| \leq h^{n+1} \sup_{e_k} \frac{|f^{(n+1)}(z)|}{(n+1)!} \leq h^{n+1} \sup_{[a,b]} \frac{|f^{(n+1)}(z)|}{(n+1)!} .$$

Il est clair que pour $N \longrightarrow \infty$, soit $h \longrightarrow 0$, l'erreur d'interpolation tend vers zéro. C'est la convergence en h.

C'est ainsi que l'on interpole linéairement les logarithmes décimaux entre deux unités, à partir d'une table des logarithmes des nombres entiers de 1000 à 10 000, avec une erreur très faible.

5.3 - Convergence en n

Une croyance assez répandue consiste à admettre qu'en augmentant indéfiniment le degré des polynômes d'interpolation, on converge vers la fonction f. Rien n'est moins sûr, même si $f \in C^{\infty}$.

Définissons la longueur

$$l = \limsup_{n \longrightarrow \infty} \left(\sup_{[a,b]} \sqrt[n+1]{|\prod(x)|} \right) \qquad (\prod(x) = (x-x_0)\ldots(x-x_n))$$

(D'après ce qui précède, $l \leq h$). Alors, une condition suffisante de convergence en n est

$$\frac{l^n \sup_{[a,b]} |f^{(n)}(x)|}{n!} \longrightarrow 0 \qquad \text{pour } n \longrightarrow \infty . \qquad (1)$$

Cette condition est très forte: le théorème qui suit montre qu'elle implique l'analyticité de la fonction.

Théorème - Si la condition (1) est remplie, la fonction f admet, en chaque point $\xi \in [a,b]$, un développement en série de Taylor de la forme

$$f(x) = \sum_{n=0} a_n (x-\xi)^n$$

dont le rayon de convergence vaut au moins l.

En effet, le reste du développement de Taylor limité à l'ordre

(k-1) s'écrit

$$\mathcal{R}_k(x) = (x - \xi)^k \frac{f^{(k)}(x^*)}{k!} \quad , \quad x^* \text{ compris entre } \xi \text{ et } x,$$

et pour $|x-\xi| < 1$, avec la condition supplémentaire $x \in [a,b]$, on aura

$$|\mathcal{R}_k(x)| \leqslant \frac{1^k}{k!} \sup_{[a,b]} |f^{(k)}(x)| \longrightarrow 0 \quad ,$$

donc, en posant

$$a_n = \frac{f^{(n)}(\xi)}{n!} \quad ,$$

le développement

$$\sum_{n=0}^{\infty} a_n (x - \xi)^n$$

converge pour tout $x \in [a,b]$ tel que $|x - \xi| < 1$. Le terme général de la série doit dès lors tendre vers zéro et, donc, être borné:

$$|a_n| \, 1^n \leqslant C \, .$$

Par conséquent, chaque fois que $|x - \xi| < 1$, que le point x soit ou non situé dans l'intervalle $]a,b[$, la série

$$\sum_{n=0}^{\infty} a_n (x - \xi)^n$$

converge, car son terme général est majoré par

$$C \left(\frac{|x - \xi|}{1} \right)^n = C \cdot \theta^n \quad , \quad \theta < 1 \quad ,$$

terme général de la série géométrique.

Ce théorème admet une réciproque, qui constitue une <u>condition suffisante de convergence en n</u> des interpolations:

<u>Si la fonction f admet en chaque point de $]a,b[$ un développement en série de Taylor de rayon de convergence $\rho > 1$, on a</u>

$$\lim_{p \to \infty} 1^p \sup_{[a,b]} \left| \frac{f^{(p)}(x)}{p!} \right| = 0$$

Soit en effet L un nombre tel que $1 < L < \rho$ et soit m le plus petit entier vérifiant simultanément

$$m > \frac{h}{2(L - 1)} \qquad \text{et} \qquad m > \frac{h}{2l} \quad .$$

Découpons l'intervalle $[a,b]$ en m sous-intervalles

$$\left[a, \; a + \frac{h}{m}\right] \; , \; \left[a + \frac{h}{m} \; , \; a + 2\frac{h}{m}\right] \; , \; \ldots \; , \; \left[b - \frac{h}{m} \; , \; b\right] \; .$$

Considérons l'un quelconque de ces sous-intervalles, et notons c son

centre. Il est clair que si x est dans ce sous-intervalle,

$$| x - c | \leq \frac{h}{2m} < 1 .$$

La série

$$f(x) = \sum_{n=0}^{\infty} a_n (x - c)^n$$

converge pour $|x - c| < \rho$, et en particulier, pour $x = c + L$. La convergence de la série

$$\sum_{n=0}^{\infty} a_n L^n$$

entraîne la convergence vers zéro de son terme général, qui est donc borné:

$$|a_n| L^n \leq A ,$$

ce qui entraîne

$$|a_n| \leq A/L^n .$$

Il en résulte que la série $\sum_n |a_n| |x - c|^n$ converge uniformément dans le sous-intervalle, car son terme général vérifie

$$|a_n| |x - c|^n \leq A (1/L)^n ,$$

le majorant étant le terme général d'une série géométrique convergente. De même, la série dérivée p fois terme à terme,

$$\sum_{n=p}^{\infty} n(n-1)\ldots(n-p+1)(x-c)^{n-p}$$

converge absolument et uniformément sur le sous-intervalle, car son terme général est majoré par

$$n^p |a_n| |x - c|^{n-p} \leq \frac{A}{L^p} \left(\frac{h}{2mL} \right)^{n-p} n^p ,$$

et la série

$$\sum_{n \geq p} n^p \theta^{n-p} , \quad \theta < 1 ,$$

converge en vertu du critère du quotient, puisque

$$\left| \left(\frac{n+1}{n} \right)^p \frac{\theta^{n-p+1}}{\theta^{n-p}} \right| \longrightarrow \theta < 1 .$$

Cette série représente donc bien la dérivée de f :

$$f^{(p)}(x) = \sum_{n=p}^{\infty} a_n \frac{n!}{(n-p)!} (x-c)^{n-p}$$

et vérifie :

$$|f^{(p)}(x)| \leqslant A \sum_{n=p}^{\infty} \frac{n!}{(n-p)!} \frac{|x-c|^{n-p}}{L^n} \leqslant \frac{A}{L^p} \sum_{n=p}^{\infty} \frac{n!}{(n-p)!} \theta^{n-p} ,$$

avec $\theta = h/(2mL)$. Or, le majorant vaut encore

$$\frac{A}{L^p} \sum_{n=0} D_\theta^p \theta^n = \frac{A}{L^p} D_\theta^p \sum_{n=0} \theta^n = \frac{A}{L^p} D_\theta^p (1-\theta)^{-1} = \frac{A}{L^p} (-1)^p D_{1-\theta}^p (1-\theta)^{-1}$$

$$= \frac{A}{L^p} \frac{p!}{(1 - \frac{h}{2mL})^{p+1}}$$

(L'interversion de la dérivée et du signe Σ se justifie comme ci-dessus). On a donc, sur le sous-intervalle considéré,

$$\left| \frac{l^p f^{(p)}(x)}{p!} \right| \leqslant \frac{A}{1 - \frac{h}{2mL}} \left(\frac{1/L}{1 - \frac{h}{2mL}} \right)^p$$

et, pour la valeur de m choisie,

$$\frac{h}{2mL} < 1 - \frac{1}{L} ,$$

soit

$$\frac{1}{L} < 1 - \frac{h}{2mL} ,$$

si bien que, en notant

$$\varepsilon = \frac{1/L}{1 - \frac{h}{2mL}} \qquad (\varepsilon < 1) ,$$

on obtient

$$\left| \frac{l^p f^{(p)}(x)}{p!} \right| < \frac{A}{1 - \frac{h}{2mL}} \cdot \varepsilon^p \longrightarrow 0$$

Soient alors $A_1, A_2, \ldots, A_m$ les constantes A correspondant aux m sous-intervalles. On a évidemment

$$\sup_{[a,b]} \left| \frac{l^p f^{(p)}(x)}{p!} \right| < \frac{\sup(A_1, \ldots, A_m)}{1 - \frac{h}{2mL}} \varepsilon^p \longrightarrow 0 ,$$

comme annoncé.

5.4 - Conclusions

Il découle de ce qui précède que l'on ne peut garantir la convergence en n pour une famille quelconque de formules d'interpolation que dans des conditions extrêmement restrictives (analyticité). Il n'y a donc guère avantage à fort élever la valeur de n, et il est beaucoup plus intéressant de découper l'intervalle. Pour obtenir le taux de convergence maximal dans chaque sous-intervalle, on veillera à placer les discontinuités de la fonction à la frontière commune de deux sous-intervalles.

6. RECHERCHE DU SUPPORT OPTIMAL

6.1 - La formule de l'erreur fait apparaître deux grandeurs: la dérivée de la fonction à interpoler, et la fonction $\prod(x)$. La première est imposée par le problème. Mais pour n donné, on peut disposer les points de manière judicieuse, de façon à rendre $|\prod(x)|$ aussi petit que possible dans l'intervalle $[a,b]$.

Bien entendu, la grandeur de $|\prod(x)|$ dépend de la longueur de l'intervalle, et nous savons déjà que

$$|\prod(x)| \leq h^{n+1} \qquad \text{et} \qquad 1 \leq h.$$

Une première amélioration consiste à ne prendre que des formules à points symétriques par rapport au centre de l'intervalle. En effet, si

$$\alpha = c - \xi \quad , \qquad \beta = c + \xi \quad ,$$

on a

$$(x-\alpha)(x-\beta) = (x - c + \xi)(x - c - \xi) = (x-c)^2 - \xi^2 \quad .$$

Cette fonction peut être maximale en $x = c$, où sa valeur absolue est $\xi^2 < h^2/4$ ou à l'une des extrémités de l'intervalle, où sa valeur absolue est

$$\frac{h^2}{4} - \xi^2 < \frac{h^2}{4} \quad .$$

Elle est donc toujours inférieure à $h^2/4$. Dès lors, pour les formules symétriques à nombre pair de points,

$$|\prod(x)| \leq \left(\frac{h}{2}\right)^{n+1} \quad .$$

Si le nombre de points est impair, l'un de ceux-ci est le centre de l'intervalle, et $|x - c| \leq h/2$, ce qui donne encore

$$|\prod(x)| \leq (h/2)^{n+1} \quad .$$

Au total, les formules symétriques réalisent toujours

$$|\prod(x)| \leq (h/2)^{n+1} \quad , \qquad 1 \leq h/2 \quad .$$

Le problème d'optimalisation envisagé se ramène à celui de la minimisation du rapport

$$K_{n+1} = \frac{\sup_{[a,b]} |\prod(x)|}{h^{n+1}} \quad ,$$

ce qui élimine le rôle de la taille de l'intervalle et ne fait plus intervenir que la répartition des points. On peut donc se ramener à un intervalle de référence. Le changement de variable

$$x = \frac{a+b}{2} + \frac{b-a}{2} X$$

ramène à l'intervalle $[-1,+1]$, de longueur h=2. Sur cet intervalle, on aura donc

$$K_{n+1} = \frac{\sup_{[-1,+1]} |\Pi(x)|}{2^{n+1}} \quad .$$

Le problème de la minimisation de K_{n+1} a été étudié par Tchébicheff et fait intervenir des polynômes particuliers.

6.2 - Polynômes de Tchébicheff

On peut définir les polynômes de Tchébicheff à partir d'une formule liant les cosinus. On a d'une part

$$\cos(n+1)\theta = \cos n\theta \cos \theta - \sin n\theta \sin \theta$$

et, d'autre part,

$$\cos(n-1)\theta = \cos n\theta \cos \theta + \sin n\theta \sin \theta ,$$

ce qui entraîne

$$\cos(n+1)\theta + \cos(n-1)\theta = 2 \cos \theta \cos n\theta,$$

soit

$$\boxed{\cos (n+1)\theta = 2 \cos \theta \cos n\theta - \cos (n-1)\theta}$$

Pour n = 1, cette formule se réduit à

$$\cos 2\theta = 2 \cos \theta \cos \theta - 1 .$$

Posons donc $x = \cos \theta \in [-1,+1]$, et définissons les fonctions T_n par

$$T_n(x) = T_n(\cos \theta) = \cos n\theta .$$

D'après la formule que nous venos d'établir, ces fonctions vérifient les relations de récurrence

$$\begin{cases} T_o(x) = 1 \\ T_1(x) = x \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ T_{n+1}(x) = 2x\, T_n(x) - T_{n-1}(x) , \end{cases}$$

qui montrent clairement qu'il s'agit de polynômes. On les appelle polynomes de Tchébicheff. Ils jouissent de quelques propriétés spéciales.

a) Relations d'orthogonalité

Dans la formule connue

$$\int_0^{\pi} \cos n\theta \cos m\theta = \frac{\pi}{2} (1 + \delta_{m,o}) \delta_{mn} ,$$

effectuons le changement de variable $x = \cos\theta$. On a

$$dx = - \sin\theta \, d\theta = - (1 - x^2)^{\frac{1}{2}} \, d\theta ,$$

c'est-à-dire

$$d\theta = - \frac{dx}{\sqrt{1 - x^2}} \quad ;$$

x varie entre +1 et -1 lorsque θ va de 0 à π, ce qui donne

$$\int_0^{\pi} \cos n\theta \cos m\theta \, d\theta = \int_{-1}^{+1} \frac{T_n(x)\, T_m(x)}{\sqrt{1 - x^2}} \, dx = \frac{\pi}{2} \delta_{mn} (1 + \delta_{m0})$$

b) Emplacement des zéros

Les zéros de T_n sont les images des zéros de cos $n\theta$, soit

$$n\theta = (2k + 1) \frac{\pi}{2} \quad , \quad k = 0, 1, \ldots ,$$

c'est-à-dire

$$\theta_k = \frac{2k+1}{n} \frac{\pi}{2} \quad .$$

θ_k se situera entre 0 et π pour

$$\frac{2k+1}{n} < 2 \ ,$$

soit

$$2k < 2n - 1$$

ou encore,

$$k < n - \tfrac{1}{2} \ .$$

T_n admet donc dans $]-1, +1[$ exactement n zéros

$$x_k = \cos \frac{(2k+1)\pi}{2n} \quad , \quad k = 0, 1, \ldots, (n-1).$$

En d'autres termes, tous les zéros de T_n sont simples et contenus dans l'ouvert $]-1,+1[$.

c) Extrema

Les extrema de ces polynômes correspondent à cos $n\theta = \pm 1$, soit

$$n\theta = k\pi \quad , \quad k=0,1,\ldots ,$$

c'est-à-dire

$$\theta_k^* = k \frac{\pi}{n} \quad .$$

θ_k^* se trouvera dans le fermé $[\, 0, \pi \,]$ pour

$$k \frac{\pi}{n} \leq \pi \ ,$$

soit

$$k \leq n \ .$$

Il existe donc (n+1) extrema de T_n dans $[\, -1, +1 \,]$, tous égaux à ± 1 , et situés aux points

$$x_k^* = \cos \frac{k\pi}{n} \quad , \quad k = 0, 1, \ldots, n \ .$$

Remarquons que $x_0^* = +1$, $x_n^* = -1$.

d) Coefficient de tête

Le coefficient de tête (celui qui affecte le terme du degré le plus élevé) de T_n vaut 2^{n-1} pour $n \geq 1$, 1 pour n=0. Pour n = 0 et 1, c'est évident; si c'est vrai jusqu'à une certaine valeur de n, on a

$$T_{n+1}(x) = 2x\, T_n(x) - T_{n-1}(x)$$

$$= 2x\,(2^{n-1}x^n + \ldots) - T_{n-1}(x) = 2^n x^{n+1} + \ldots ,$$

donc c'est encore vrai pour T_{n+1}.

6.3 - Théorème de Tchébicheff

L'introduction des polynômes de Tchébicheff se justifie par l'important théorème suivant:

De tous les polynômes de degré n ayant leur coefficient de tête égal à 1, c'est $\hat{T}_n = T_n/(2^{n-1})$ qui a la plus petite borne supérieure en module sur $[-1,+1]$.

Supposons en effet le contraire: on peut donc trouver un polynôme

$$P_n(x) = x^n + \alpha_{n-1} x^{n-1} + \alpha_{n-2} x^{n-2} + \ldots + \alpha_0$$

dont la borne supérieure en module dans $[-1,+1]$ est inférieure à celle de $\hat{T}_n$, c'est-à-dire $\frac{1}{2^{n-1}}$. Alors (fig. 6), la différence

$$R_{n-1}(x) = \hat{T}_n(x) - P_n(x)$$

est un polynôme de degré (n-1). En chaque extrémum de $\hat{T}_n$, ce polynôme doit avoir le même signe que $\hat{T}_n$, sans quoi

$$| P_n(x_i^*) | = | \hat{T}_n(x_i^*) - R_{n-1}(x_i^*) | \geq | \hat{T}_n(x_i^*) | ,$$

en contradiction avec la définition de P_n. On a donc

$$R_{n-1}(1) > 0$$

$$R_{n-1}(x_1^*) < 0$$

$$R_{n-1}(x_2^*) > 0$$

.....

Comme les extréma de T_n sont au nombre de (n+1), il en résulte que R_{n-1} prend (n+1) signes différents. Il doit donc s'annuler en n points au moins, ce qui est impossible s'il n'est pas identiquement nul, puisqu'il s'agit d'un polynôme de degré (n-1).

6.4 - Support de Tchébicheff

Il résulte du théorème précédent que le support optimal à (n+1) points est constitué des (n+1) zéros de T_{n+1}, soit

$$x_k = \cos(2k+1)\frac{\pi}{2n} \quad , \qquad k = 0, 1, 2, \ldots, n.$$

Pour ce support particulier, dit support de Tchébicheff, on a

$$\sup_{[\,-1,+1\,]} |\Pi(x)| = \sup_{[\,-1,+1\,]} \left| \frac{T_{n+1}}{2^n} \right| = \frac{1}{2^n} \quad ,$$

ce qui donne à la constante K_{n+1} définie plus haut la valeur

$$K_{n+1} = \frac{\sup_{[\,-1,+1\,]} |\Pi(x)|}{2^{n+1}} = \frac{1}{2^{2n+1}} \quad .$$

En conséquence, pour une distribution identique des points d'interpolation sur un intervalle $[\,a,b\,]$ de longueur h, on a

$$\sup_{[\,a,b\,]} |\Pi(x)| = \frac{h^{n+1}}{2^{2n+1}} \quad ,$$

et

$$1 = \limsup_{n \to \infty} \sqrt[n+1]{(h^{n+1}/2^{2n+1})} = \frac{h}{4} \quad .$$

Nous verrons plus loin (§ 10) que, bien plus, l'interpolation de Tchébicheff converge en n dans des conditions fort générales.

7. INTERPOLATION AVEC CONDITIONS DE CONTACT

7.1 - Différences divisées avec arguments confondus

Nous avons démontré en section 5.1 la relation

$$f(x_0, x_1, \ldots, x_{k-1}, x) = \frac{f^{(k)}(\xi)}{k!} \quad ,$$

ξ étant un point situé entre le plus petit et le plus grand argument de la différence divisée. En particulier, pour $f \in C^k$,

$$f(\underbrace{x_0, x_0, \ldots, x_0, x_0}_{(k+1)}) = \lim_{\varepsilon \to 0} f(x_0, x_0+\varepsilon, x_0+2\varepsilon, \ldots, x_0+k\varepsilon)$$

$$= \lim_{\varepsilon \to 0} \frac{f^{(k)}(x_0 + \theta k \varepsilon)}{k!} = \frac{f^{(k)}(x_0)}{k!} \quad .$$

On peut d'ailleur en déduire le théorème de Taylor pour $f \in C^{n+1}$:

$$f(x) = f(x_0) + (x-x_0)f(x_0,x_0) + \ldots + (x-x_0)^n f(\underbrace{x_0, \ldots, x_0}_{(n+1)}) +$$

$$+ (x-x_0)^{n+1} f(\underbrace{x_0, \ldots, x_0}_{(n+1)}, x)$$

$$= f(x_o) + (x-x_o)f'(x_o) + \ldots + (x-x_o)^n \frac{f^{(n)}(x_o)}{n!} + (x-x_o)^{n+1} \frac{f^{(n+1)}(\xi)}{(n+1)!} \quad ,$$

étant compris entre x_o et x.

7.2 - Soit à interpoler f aux points $x_o,\ldots,x_n$, avec, de plus, les conditions de contact $\tilde{f}'(x_i) = f'(x_i)$. Il suffit de construire le tableau de différences qui suit (pour trois points):

$$\begin{array}{llllll}
x_o & & & & & \\
 & f(x_o,x_o) & & & & \\
x_o & & f(x_o,x_o,x_1) & & & \\
 & f(x_o,x_1) & & f(x_o,x_o,x_1,x_1) & & \\
x_1 & & f(x_o,x_1,x_1) & & f(x_o,x_o,x_1,x_1,x_2) & \\
 & f(x_1,x_1) & & f(x_o,x_1,x_1,x_2) & & f(x_o,x_o,x_1,x_1,x_2,x_2) \\
x_1 & & f(x_1,x_1,x_2) & & f(x_o,x_1,x_1,x_1,x_2) & \\
 & f(x_1,x_2) & & f(x_1,x_1,x_2,x_2) & & \\
x_2 & & f(x_1,x_2,x_2) & & & \\
 & f(x_2,x_2) & & & & \\
x_2 & & & & &
\end{array}$$

et d'écrire

$$\begin{aligned}
\tilde{f}(x) = {} & f(x_o) + (x-x_o)f(x_o,x_o) + (x-x_o)^2 f(x_o,x_o,x_1) + \\
& + (x-x_o)^2(x-x_1)f(x_o,x_o,x_1,x_1) + (x-x_o)^2(x-x_1)^2 f(x_o,x_o,x_1,x_1,x_2) + \\
& + (x-x_o)^2(x-x_1)^2(x-x_2)f(x_o,x_o,x_1,x_1,x_2,x_2) \quad ,
\end{aligned}$$

avec le reste

$$\begin{aligned}
R(x) &= f(x_o,x_o,x_1,x_1,x_2,x_2,x) \cdot (x-x_o)^2(x-x_1)^2(x-x_2)^2 \\
&= (x-x_o)^2(x-x_1)^2(x-x_2)^2 \frac{f^{(6)}(\xi)}{6!} \quad .
\end{aligned}$$

Cette procédure se généralise aisément au cas où des dérivées d'ordre supérieur doivent être respectées.

7.3 Formule d'Hermite

Nous avons jusqu'ici traité les conditions de contact en termes des différences divisées. Mais on peut aussi trouver des expressions à la Lagrange. Soit à interpoler une fonction f de façon à respecter, en (n+1) points, la valeur de f et de sa dérivée. Cela fait en tout (2n+2) conditions et, pour les respecter, il faudra un polynôme de degré (2n+1). Ecrivons-le sous la forme

$$\tilde{f}(x) = \sum_{i=0}^{n+1} \left[M_i(x)\, f(x_i) + N_i(x)\, f'(x_i) \right] \quad .$$

Les fonctions d'interpolation M_i et N_i doivent évidemment vérifier

$$\begin{cases} M_i(x_j) = \delta_{ij} \quad ; \quad M_i'(x_j) = 0 \\ N_i(x_j) = 0 \quad ; \quad N_i'(x_j) = \delta_{ij} \end{cases} .$$

On obtient automatiquement les conditions en x_j , $j \neq i$, en posant

$$M_i(x) = [\, A + B(x-x_i) \,]\, L_i^2(x)$$

$$N_i(x) = [\, C + D(x-x_i) \,]\, L_i^2(x) \quad ,$$

L_i étant le polynôme d'interpolation de Lagrange

$$L_i(x) = \prod_{\substack{j=0 \\ \neq i}}^{n} \frac{(x-x_j)}{(x_i-x_j)} \quad .$$

On détermine les quatre coefficients A, B, C et D par les conditions

$$1 = M_i(x_i) = A\, L_i^2(x_i) \quad , \quad \text{d'où} \quad A = 1 \quad ;$$

$$0 = M_i'(x_i) = B\, L_i^2(x_i) + 2A\, L_i(x_i) L_i'(x_i) \quad , \quad \text{d'où} \quad B = -2\, L_i'(x_i) \quad ;$$

$$0 = N_i(x_i) = C\, L_i^2(x_i) \quad , \quad \text{d'où} \quad C = 0 \quad ;$$

$$1 = N_i'(x_i) = D\, L_i^2(x_i) + 0 \quad , \quad \text{d'où} \quad D = 1 .$$

Il vient donc

$$\begin{cases} M_i(x) = [\, 1 - 2(x-x_i)\, L_i'(x_i) \,]\, L_i^2(x) \\ N_i(x) = (x-x_i)\, L_i^2(x) \end{cases}$$

et

$$\tilde{f}(x) = \sum_{i=0}^{n} \left\{ [\, 1 - 2(x-x_i)\, L_i'(x_i) \,]\, L_i^2(x)\, f(x_i) + (x-x_i)\, L_i^2(x)\, f'(x_i) \right\} .$$

C'est la <u>formule d'interpolation d'Hermite</u>.

On notera qu'il est possible d'exprimer $L_i'(x_i)$ en termes de la fonction $\prod(x) = (x-x_0)\ldots(x-x_n)$. En effet,

$$L_i'(x) = \frac{d}{dx} \left[\frac{\prod(x)}{(x-x_i)\prod'(x_i)} \right] = \frac{(x-x_i)\prod'(x) - \prod(x)}{(x-x_i)^2 \prod'(x_i)}$$

Tenant compte du développement de Taylor

$$\prod(x_i) = \prod(x) + (x_i-x)\prod'(x) + \frac{(x_i-x)^2}{2!}\prod''(x^*) \quad ,$$

avec x^* compris entre x et x_i , on a donc

$$(x-x_i)\,\Pi'(x) - \Pi(x) = \frac{(x-x_i)^2}{2!}\,\Pi''(x^*) - \Pi(x_i) = \frac{(x-x_i)^2}{2}\,\Pi''(x^*)\ .$$

et

$$L_i'(x) = \frac{\Pi''(x^*)}{2\Pi'(x_i)}\ .$$

Faisant tendre x vers x_i, on obtient

$$L_i'(x_i) = \frac{\Pi''(x_i)}{2\Pi'(x_i)}\ .$$

Quant à l'erreur d'interpolation, elle vaut évidemment

$$R(x) = (x-x_o)^2 \ldots (x-x_n)^2\, f(x_o,x_o,\ldots,x_n,x_n,x)$$

$$= (x-x_o)^2 \ldots (x-x_n)\ \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\ ,$$

avec ξ compris entre la plus petite et la plus grande des valeurs $x_o, \ldots, x_n$ et x.

8. RECURRENCE SUR LES INTERPOLATIONS : FORMULE D'AITKEN

Soient $\sigma = \{x_o,\ldots,x_n\}$ un support d'interpolation, a et b deux points complémentaires. Supposons que l'on connaisse l'interpolée $\tilde{f}_{\sigma,a}$ de f sur σ et a, et l'interpolée $\tilde{f}_{\sigma,b}$ de f sur σ et b. On se propose de calculer l'interpolée $\tilde{f}_{\sigma,a,b}$ de f sur σ, a, b.

On a

$$\tilde{f}_{\sigma,a} = f(x_o) + (x-x_o)f(x_o,x_1) + \ldots + (x-x_o)\ldots(x-x_n)f(x_o,\ldots,x_n,a)$$

$$= \tilde{f}_\sigma + (x-x_o)\ldots(x-x_n)f(x_o,\ldots,x_n,a)$$

et, de même,

$$\tilde{f}_{\sigma,b} = \tilde{f}_\sigma + (x-x_o)\ldots(x-x_n)f(x_o,\ldots,x_n,b),$$

d'où

$$(x-b)\tilde{f}_{\sigma,a} - (x-a)\tilde{f}_{\sigma,b} = (a-b)\tilde{f}_\sigma + (x-x_o)\ldots(x-x_n)\,[\,(x-b)f(x_o,\ldots,x_n,a) - (x-a)f(x_o,\ldots,x_n,b)\,]\ .$$

Le facteur entre crochets s'écrit encore

$$(x-b)f(x_o,\ldots,x_n,a) - (x-a)\,[\,f(x_o,\ldots,x_n,a) + (b-a)\,f(x_o,\ldots,x_n,a,b)\,]$$

$$= (a-b)f(x_o,\ldots,x_n,a) + (a-b)(x-a)f(x_o,\ldots,x_n,a,b).$$

Il en résulte

$$(x-b)\tilde{f}_{\sigma,a} - (x-a)\tilde{f}_{\sigma,b} = (a-b)\,[\,\tilde{f}_\sigma + (x-x_o)\ldots(x-x_n)f(x_o,\ldots,x_n,a) + (x-x_o)\ldots(x-x_n)(x-a)f(x_o,\ldots,x_n,a,b)\,]$$

$$= (a-b)\,\tilde{f}_{\sigma,a,b} \ ,$$

soit, au total,

$$\boxed{\tilde{f}_{\sigma,a,b} = \frac{(x-b)\tilde{f}_{\sigma,a} - (x-a)\tilde{f}_{\sigma,b}}{a-b}}$$

C'est la formule d'Aitken.

9. INTERPOLATION TRIGONOMETRIQUE

9.1 - Une fonction continue et périodique f, de période 2π, admet un développement de Fourier de la forme

$$f(\theta) = a_o + \sum_{k=1}^{\infty} (a_k \cos k\theta + b_k \sin k\theta) \ .$$

Les conditions d'orthogonalité dans $L^2(\,]0,2\pi[\,)$,

$$\int_0^{2\pi} \cos p\theta \cos q\theta \, d\theta = (1 + \delta_{p,0})\,\delta_{pq}$$

$$\int_0^{2\pi} \sin p\theta \sin q\theta \, d\theta = \delta_{pq}$$

$$\int_0^{2\pi} \sin p\theta \cos q\theta \, d\theta = 0$$

entraînent alors

$$\begin{cases} a_p = \dfrac{1}{\pi(1 + \delta_{p,0})} \displaystyle\int_0^{2\pi} f(\theta) \cos p\theta \, d\theta \\ b_p = \dfrac{1}{\pi} \displaystyle\int_0^{2\pi} f(\theta) \sin p\theta \, d\theta \end{cases}$$

Ce sont là des faits bien connus, sur lesquels nous ne reviendrons pas.

L'interpolation trigonométrique, encore appelée transformation de Fourier discrète, consiste à rechercher une interpolée de la forme

$$\tilde{f}(\theta) = \tilde{a}_o + \sum_{k=1}^{n} (\tilde{a}_k \cos k\theta + \tilde{b}_k \sin k\theta) \ .$$

On détermine les coefficients de Fourier discrets $\tilde{a}_o$, $\tilde{a}_k$, $\tilde{b}_k$ par les conditions d'interpolation

$$\tilde{f}(\theta_j) = f(\theta_j)$$

en $(2n+1)$ points $\theta_o, \ldots, \theta_{2n}$. Il est clair qu'au sens de l'interpolation, deux fonctions égales en ces $(2n+1)$ points sont équivalentes: elles ont la même interpolée. La norme naturelle du problème est donc

$$\| f \|^2_{int} = \sum_{j=0}^{2n} f^2(\theta_j) \ .$$

Elle est associée au produit scalaire

$$(f,g)_{int} = \sum_{j=0}^{2n} f(\theta_j)\, g(\theta_j) \ .$$

Alors,

$$(f,\ \cos p\theta)_{int} = \sum_{j=0}^{2n} f(\theta_j) \cos p\theta_j$$

$$(f,\ \sin p\theta)_{int} = \sum_{j=0}^{2n} f(\theta_j) \sin p\theta_j$$

et on se rend aisément compte que si l'on peut trouver un support d'interpolation tel que soient vérifiées les relations d'orthogonalité discrètes

$$(\cos p\theta,\ \cos q\theta)_{int} = \sum_{j=0}^{2n} \cos p\theta_j \cos q\theta_j = A_p \delta_{pq} \ ,$$

$$(\sin p\theta,\ \sin q\theta)_{int} = \sum_{j=0}^{2n} \sin p\theta_j \sin q\theta_j = B_p \delta_{pq}$$

et

$$(\sin p\theta,\ \cos q\theta)_{int} = \sum_{j=0}^{2n} \sin p\theta_j \cos q\theta_j = 0 \ ,$$

on aura tout simplement

$$\tilde{a}_p = \frac{1}{A_p} (f,\ \cos p\theta)_{int} = \frac{1}{A_p} \sum_{j=0}^{2n} f(\theta_j) \cos p\theta_j$$

et

$$\tilde{b}_p = \frac{1}{B_p} (f,\ \sin p\theta)_{int} = \frac{1}{B_p} \sum_{j=0}^{2n} f(\theta_j) \sin p\theta_j \ .$$

9.2 - Relations d'orthogonalité discrètes

Un tel support est donné par les points

$$\theta_j = \frac{2j+1}{2n+1}\pi \qquad , \ j = 0,\ 1,\ \ldots,\ 2n.$$

En effet, on a

$$\sum_{j=0}^{2n} \cos p\theta_j \cos q\theta_j = \tfrac{1}{2} \sum_{j=0}^{2n} (\cos (p-q)\theta_j + \cos (p+q)\theta_j)$$

$$= \tfrac{1}{2}\ \mathcal{R} \sum_{j=0}^{2n} (e^{i(p-q)\theta_j} + e^{i(p+q)\theta_j}) \ ,$$

$$\sum_{j=0}^{2n} \sin p\theta_j \sin q\theta_j = \tfrac{1}{2} \sum_{j=0}^{2n} (\cos (p-q)\theta_j - \cos (p+q)\theta_j)$$

$$= \tfrac{1}{2}\, \mathcal{R} \sum_{j=0}^{2n} \left(e^{i(p-q)\theta_j} - e^{i(p+q)\theta_j} \right)$$

et

$$\sum_{j=0}^{2n} \sin p\theta_j \cos q\theta_j = \tfrac{1}{2} \sum_{j=0}^{2n} (\sin (p+q)\theta_j + \sin(p-q)\theta_j)$$

$$= \tfrac{1}{2}\, \mathcal{J} \sum_{j=0}^{2n} (e^{i(p+q)\theta_j} + e^{i(p-q)\theta_j})) \ .$$

Or, pour $r \neq 0$,

$$\sum_{j=0}^{2n} e^{ir\theta_j} = \frac{e^{ir\frac{2n+3}{2n+1}\pi} - e^{ir\frac{1}{2n+1}\pi}}{e^{ir\frac{2}{2n+1}\pi} - 1} = 0$$

et, pour $r = 0$,

$$\sum_{j=0}^{2n} e^{ir\theta_j} = 2n+1 \ .$$

Donc,

$$\sum_{j=0}^{2n} e^{ir\theta_j} = (2n+1)\, \delta_{r,0} \quad ,$$

ce qui entraîne

$$\sum_{j=0}^{2n} \cos p\theta_j \cos q\theta_j = \frac{2n+1}{2}\, \delta_{pq}\, (1 + \delta_{p,0}) \quad ,$$

$$\sum_{j=0}^{2n} \sin p\theta_j \sin q\theta_j = \frac{2n+1}{2}\, \delta_{pq}$$

et

$$\sum_{j=0}^{2n} \sin p\theta_j \cos q\theta_j = 0 \ .$$

On en déduit que, sur ce support,

$$\tilde{a}_p = \frac{2}{2n+1}\, \frac{1}{1+\delta_{p,0}} \sum_{j=0}^{2n} f(\theta_j) \cos p\theta_j$$

$$\tilde{b}_p = \frac{2}{2n+1} \sum_{j=0}^{2n} f(\theta_j) \sin p\theta_j$$

9.3 - Cadre naturel de l'approximation

Pour les fonctions périodiques, il y a équivalence entre les points θ et $\theta + 2k\pi$, $k = 1, 2, \ldots$. Une telle fonction ne peut donc être qualifiée de continue que si, outre la condition évidente $f \in C^0([0,2\pi])$, on a en outre $f(0) = f(2\pi)$. Nous écrirons dans ce cas $f \in \hat{C}^0([0,2\pi])$.

On sait par ailleurs que le cadre naturel des séries de Fourier est l'espace $L^2( ]0,2\pi[ )$ des fonctions périodiques de carré intégrable, et que

$$\int_\alpha^{\alpha+2\pi} f^2(\theta)\, d\theta = \left[2a_0^2 + \sum_{k=1}^{\infty} (a_k^2 + b_k^2) \right] ,$$

quel que soit l'intervalle de période $]\alpha , \alpha + 2\pi[$ considéré. Mais dans L^2, la série ne converge que presque partout. Nous exigerons donc un peu plus, à savoir que la dérivée de f ait son carré intégrable, et que la série de Fourier soit dérivable terme à terme. Les coefficients de Fourier de f' étant donnés par

$$a'_p + i\, b'_p = \frac{1}{\pi} \int_0^{2\pi} f'(\theta)\, e^{ip\theta}\, d\theta$$

$$= \frac{1}{\pi} \left[f(2\pi)\, e^{2ip\pi} - f(0) \right] - \frac{ip}{\pi} \int_0^{2\pi} f(\theta)\, e^{ip\theta}\, d\theta ,$$

on a

$$a'_p + i\, b'_p = \frac{1}{\pi} (f(2\pi) - f(o)) + p(b_p - i\, a_p) .$$

Dès lors, si les relations

$$b'_p = - p\, a_p$$

sont toujours vraies, les relations

$$a'_p = p\, b_p$$

ne sont, quant à elles, vérifiées que si $f(0) = f(2\pi)$. Nous noterons en conséquence $\hat{H}^1( ]0,2\pi[ )$ l'espace des fonctions de carré intégrable telles que $f(0) = f(2\pi)$. Cette condition supplémentaire est naturelle dans le cadre des fonctions périodiques. En effet, il faut assurer la condition

$$\int_\alpha^{\alpha+2\pi} f'^2\, d\theta < \infty$$

sur tout intervalle de période. Or, cela implique la continuité, car

$$| f(\theta_2) - f(\theta_1) | = \left| \int_{\theta_1}^{\theta_2} f'(\theta)\, d\theta \right| \leq |\theta_2 - \theta_1|^{\frac{1}{2}} \left(\int_{\theta_1}^{\theta_2} f'^2(\theta)\, d\theta \right)^{\frac{1}{2}}$$

$$\leq |\theta_2 - \theta_1|^{\frac{1}{2}} \left(\int_\alpha^{\alpha+2\pi} f'^2(\theta)\, d\theta \right)^{\frac{1}{2}} .$$

En particulier, $f(0_+) = f(2\pi_+)$ doit être égal à $f(2\pi_-)$.

Soit donc $f \in \hat{H}^1(\,]0,2\pi[\,)$. En notant f_N la série de Fourier de f limitée à l'ordre N, on a encore

$$\sup_{\theta \in [\,0,2\pi\,]} |f(\theta) - f_N(\theta)| \leq |f(\theta_o) - f_N(\theta_o)| + (2\pi)^{\frac{1}{2}} \left(\int_0^{2\pi} (f-f_N)'^2 d\theta \right)^{\frac{1}{2}} .$$

Comme $f \in L^2(\,]0,2\pi[\,)$, sa série de Fourier converge presque partout. On peut donc choisir pour θ_o un point où elle converge. L'inégalité ci-dessus garantit alors la convergence uniforme de la série de Fourier vers f.

9.4 - Lemme - Soit $f \in \hat{H}^1(\,]0,2\pi[\,)$ et soit g_N une famille de fonctions du même espace, telles que g_N interpole f aux (N+1) points

$$\theta_o = \Delta\theta/2 \ , \ \theta_1 = \theta_o + \Delta\theta \ , \ \ldots, \ \theta_N = \theta_o + N\,\Delta\theta = 2\pi - \frac{\Delta\theta}{2} \ ,$$

ce qui implique

$$\Delta\theta = \frac{2\pi}{N+1} \ .$$

S'il est possible de trouver une majoration

$$\|f' - g_N'\| \leq C$$

en norme de $L^2(\,]0,2\pi[\,)$, indépendante de N, alors, les g_N convergent uniformément vers f.

En effet, on a, pour $\theta_k < \theta < \theta_{k+1}$,

$$|f(\theta) - g_N(\theta)| = \left| \int_{\theta_k}^{\theta} (f'(\tau) - g_N'(\tau))\, d\tau \right| \leq \left(\int_{\theta_k}^{\theta} (f'-g_N')^2 d\tau \right)^{\frac{1}{2}} (\theta-\theta_k)^{\frac{1}{2}}$$

$$\|f'-g_N'\| \sqrt{\Delta\theta} \leq C \left(\frac{2}{N+1}\right)^{\frac{1}{2}} \longrightarrow 0 \ .$$

A gauche de θ_o ou à droite de θ_N, on considère le prolongement périodique de f sur $]\theta_N - 2\pi \, , \theta_o[$, ce qui conduit à la même conclusion.

9.5 - Un problème auxiliaire

Etant donné les points

$$\theta_o = \frac{\Delta\theta}{2} \ , \ \theta_1 = \theta_o + \Delta\theta, \ \ldots, \ \theta_N = \theta_o + N\Delta\theta \ ,$$

avec

$$\Delta\theta = \frac{2\pi}{N+1} \ ,$$

considérons l'interpolation linéaire par morceaux

$$\hat{f}(\theta) = f(\theta_k)\,\frac{\theta_{k+1} - \theta}{\theta} + f(\theta_{k+1})\,\frac{\theta - \theta_k}{\theta} \qquad \text{entre } \theta_k \text{ et } \theta_{k+1} \text{ ,}$$

pour les intervalles $]\theta_o,\theta_1[$, ... , $]\theta_{N-1},\theta_N[$, ainsi que $]\theta_N,\theta_{N+1}[$, en posant $\theta_{N+1} = \theta_o + 2\pi$ et en jouant sur l'équivalence périodique (fig.7).

Alors, si f est élément de $\hat{H}^1(\,]0,2\pi[\,)$, les coefficients de Fourier de $\hat{f}$ vérifient

$$(1+\delta_{p,0})\ (\hat{a}_p + i\hat{b}_p) = \sum_{k=0}^{N} \int_{\theta_k}^{\theta_{k+1}} \hat{f}\, e^{ip\theta}\, d\theta$$

$$= -\sum_{k=0}^{N} \int_{\theta_k}^{\theta_{k+1}} \hat{f}'\, \frac{e^{ip\theta}}{ip}\, d\theta$$

$$= \sum_{k=0}^{N} \frac{f(\theta_{k+1}) - f(\theta_k)}{\Delta\theta} \cdot \frac{1}{p^2} \left(e^{ip\theta_{k+1}} - e^{ip\theta_k}\right)$$

$$= \frac{1}{p^2\,\Delta\theta} \left[\sum_{k+1=1}^{N+1} f(\theta_{k+1})\left(e^{ip\theta_{k+1}} - e^{ip\theta_k}\right) - \sum_{k=0}^{N} f(\theta_k)\left(e^{ip\theta_{k+1}} - e^{ip\theta_k}\right) \right]$$

Compte tenu de la périodicité des fonctions, cette expression se réduit à

$$\frac{1}{p^2\,\Delta\theta} \sum_{k=0}^{N} f(\theta_k)\, e^{ip\theta_k} \left(2 - e^{ip\Delta\theta} - e^{-ip\Delta\theta}\right)$$

$$= \sum_{k=0}^{N} f(\theta_k)\, e^{ip\theta_k}\, \frac{2 - 2\cos(p\Delta\theta)}{p^2\,\Delta\theta} = \sum_{k=0}^{N} f(\theta_k)\, e^{ip\theta_k}\, \frac{\sin^2\left(\frac{p\,\Delta\theta}{2}\right)}{\left(\frac{p\,\Delta\theta}{2}\right)^2}\ \theta$$

Notant que $\theta = \frac{2}{N+1}$, on obtient

$$\hat{a}_p + i\hat{b}_p = \left[\frac{2}{N+1}\, \frac{1}{1+\delta_{p,0}} \sum_{k=0}^{N} f(\theta_k)\, e^{ip\theta_k} \right] \cdot \frac{\sin^2\left(\frac{p\,\theta}{2}\right)}{\left(\frac{p\,\Delta\theta}{2}\right)^2}$$

L'interpolation linéaire par morceaux jouit d'une seconde propriété utile. On a

$$\|f' - \hat{f}'\|^2 = \sum_{k=0}^{N} \int_{\theta_k}^{\theta_{k+1}} (f' - \hat{f}')^2\, d\theta$$

et comme, entre θ_k et θ_{k+1}, $\hat{f}'$ est une constante,

$$\|f' - \hat{f}'\|^2 = \sum_{k=0}^{N} \left(\int_{\theta_k}^{\theta_{k+1}} f'^2\, d\theta - 2\,\hat{f}' \int_{\theta_k}^{\theta_{k+1}} f'\, d\theta + \hat{f}'^2 \Delta\theta \right)$$

$$= \|f'\|^2 - \|\hat{f}'\|^2 \ .$$

Il en résulte les deux inégalités suivantes:

$$\begin{cases} \|f' - \hat{f}'\| \leq \|f'\| \\ \|\hat{f}'\| \leq \|f'\| \end{cases}$$

9.6 - Majoration de la différence entre $\hat{f}'$ et $\tilde{f}'$

Il est également possible de majorer $\|\hat{f}' - \tilde{f}'\|$, à partir du développement de Fourier: en posant N = 2n,

$$\|\hat{f}' - \tilde{f}'\|^2 = \sum_{p=1}^{n} p^2 \left[(\tilde{a}_p - \hat{a}_p)^2 + (\tilde{b}_p - \hat{b}_p)^2 \right] + \sum_{p=n+1}^{\infty} (\hat{a}_p^2 + \hat{b}_p^2) .$$

Or, nous venons de voir que

$$\tilde{a}_p = \hat{a}_p \frac{(p\Delta\theta/2)^2}{\sin^2(p\Delta\theta/2)} \quad , \qquad b_p = \hat{b}_p \frac{(p\Delta\theta/2)^2}{\sin^2(p\Delta\theta/2)}$$

Comme on a toujours, pour $p \leq n$,

$$\frac{p\Delta\theta}{2} \leq \frac{n}{2} \frac{2\pi}{2n+1} \leq \frac{\pi}{2} ,$$

on vérifie aisément que

$$\frac{p\Delta\theta/2}{\sin(p\Delta\theta/2)} \leq \frac{\pi}{2} .$$

Dès lors,

$$\|\hat{f}' - \tilde{f}'\|^2 \leq \sum_{p=1}^{\infty} (\frac{\pi^2}{4} - 1)^2 p^2(\hat{a}_p^2 + \hat{b}_p^2) \leq (\frac{\pi^2}{4} - 1)^2 \|\hat{f}'\|^2$$

et, comme $\|\hat{f}'\| \leq \|f'\|$,

$$\|\hat{f}' - \tilde{f}'\| \leq (\frac{\pi^2}{4} - 1)\|f'\|$$

9.7 - Convergence en n de l'interpolation trigonométrique

On déduit des paragraphes précédents que

$$\|f' - \tilde{f}'\| \leq \|f' - \hat{f}'\| + \|\hat{f}' - \tilde{f}'\| \leq \|f'\| + (\frac{\pi^2}{4} - 1)\|f'\|$$
$$\leq \frac{\pi^2}{4} \|f'\| .$$

Dès lors, en vertu du lemme 9.4, nous avons démontré le résultat suivant:

Soit f une fonction périodique, appartenant à $\hat{H}^1(]0,2\pi[)$.
Les interpolées trigonométriques d'ordre n de f convergent uniformément vers f lorsque n tend vers l'infini.

10. INTERPOLATION EN COSINUS DES FONCTIONS NON PERIODIQUES

10.1 - Soit f une fonction continue sur l'intervalle $[\,0,\pi\,]$. On peut la prolonger en une fonction $F \in \hat{C}^0([\,-\pi\,,\pi\,])$ en posant

$$\begin{cases} F(\theta) = f(\theta) \text{ pour } \theta \in [\,0,\pi\,] \\ F(\theta) = f(-\theta) \text{ pour } \theta \in [\,-\pi\,,0[\end{cases}$$

Comme f est une fonction paire, elle admet un développement de Fourier en cosinus:

$$F(\theta) = \sum_{k=0}^{\infty} a_k \cos k\theta$$

et, bien entendu, comme sur $[\,0,\pi\,]$, $f(\theta) = F(\theta)$, on a encore

$$f(\theta) = \sum_{k=0}^{\infty} a_k \cos k\theta \ .$$

Ce résultat suggère d'utiliser, pour les fonctions continues sur $[\,0,\pi\,]$, une interpolation en cosinus, de la forme

$$\tilde{f}(\theta) = \sum_{k=0}^{n} \tilde{a}_k \cos k\theta \ .$$

En choisissant comme support d'interpolation les (n+1) points

$$\theta_j = \frac{(2j+1)\pi}{2n+1} \ ,$$

on obtient les relations d'orthogonalité discrètes

$$\sum_{j=0}^{n} \cos p\theta_j \cos q\theta_j = \frac{n+1}{2}(1+\delta_{p,0})\delta_{pq} \ ,$$

ce qui permet de calculer les coefficients $\tilde{a}_p$ par la formule simple

$$\tilde{a}_p = \frac{1}{1+\delta_{p,0}} \quad \frac{2}{n+1} \quad \sum_{j=0}^{n} f(\theta_j) \cos p\theta_j \ .$$

10.2 - Convergence en n des interpolations en cosinus

Les raisonnements relatifs aux séries de Fourier discrètes s'appliquent intégralement dans ce cas. Il suffit en effet de considérer le prolongement périodique de F à l'intervalle $[\,0,\,2\pi\,]$, et de considérer la fonction F qui l'interpole aux points

$$\theta_0 = \Delta\theta/2 \ , \quad \theta_1 = \theta_0 + \Delta\theta, \ \ldots, \ \theta_{2n+1} = \theta_0 + (2n+1)\,\Delta\theta \ ,$$

avec $\Delta\theta = \frac{2}{2n+2}$, soit N = 2n+1. Lors de l'interpolation en cosinus, chaque valeur de la fonction f sera ainsi utilisée deux fois, mais la valeur de a_p se calcule en divisant par 2n+2 au lieu de (n+1), ce qui ramène au même résultat. De même pour l'interpolation

linéaire par morceaux. Nous laissons au lecteur le soin de refaire la démonstration de la section 9.6 pour ce cas: il suffit de poser $N = 2n+1$ au lieu de $N = 2n$ et d'omettre tous les termes b_p, $\tilde{b}_p$ et $\hat{b}_p$.

Par conséquent, <u>si f est une fonction de classe $H^1(\,]0, 2\pi[\,)$, les interpolées en cosinus à (n+1) points de f convergent uniformément vers f lorsque n tend vers l'infini.</u>

10.3 - <u>Relation avec l'interpolation de Tchébicheff</u>

Si l'on pose $x = \cos\theta$, les interpolations ci-dessus prennent la forme

$$f(x) = \sum_{k=0}^{n} a_k T_k(x) ,$$

et le support sur $[\,-1,+1\,]$ est celui de Tchébicheff. Lorsque $f(\cos\theta) \in H^1(\,]0,\pi[\,)$, on a

$$\int_0^{\pi} f^2 \, d\theta = \int_{-1}^{+1} f^2(x) \frac{1}{\sqrt{1-x^2}} \, dx \; < \infty$$

et

$$\int_0^{\pi} \left(\frac{df}{d\theta}\right)^2 d\theta = \int_0^{\pi} \left(\frac{df}{d(\cos\theta)}\right)^2 . \sin^2\theta \, d\theta$$

$$= \int_{-1}^{+1} \left(\frac{df}{dx}\right)^2 \sqrt{1-x^2} \, dx \; < \infty .$$

Or, on peut montrer que ces deux conditions sont vérifiées si $f(x) \in H^1(\,]-1,+1[\,)$. Pour la condition portant sur les dérivées, c'est évident; pour l'autre,

$$\int_{-1}^{+1} \frac{f^2 dx}{\sqrt{1-x^2}} = \left[\, f^2 \arccos x \,\right]_{-1}^{+1} - 2 \int_{-1}^{+1} f \frac{df}{dx} \arccos x \, dx$$

$$\leq \pi f^2(-1) + 2 \int_{-1}^{+1} |f| \, |f'| \, \pi \, dx$$

$$\leq \pi \, f^2(-1) + 2\pi \, \|f\| \, \|f'\| .$$

Il suffit de majorer $f^2(-1)$. On a

$$f^2(-1) = f^2(x) + 2 \int_x^{-1} f \, f' \, dx \leq f^2(x) + 2 \, \|f\| \, \|f'\|$$

et, en intégrant entre -1 et +1,

$$2 \, f^2(-1) \leq \int_{-1}^{+1} f^2(x) dx + 4 \, \|f\| \, \|f'\| .$$

On peut en déduire les deux résultats suivants:

a) Théorème de développement en série de polynômes de Tchébicheff.

Toute fonction $f \in H^1(\,]-1, +1[\,)$ admet un développement en série de la forme

$$f = \sum_{k=0}^{\infty} a_k T_k(x) \quad ,$$

uniformément convergent, dont les coefficients sont donnés par

$$a_p = \frac{1}{1+\delta_{p,0}} \frac{2}{\pi} \int_{-1}^{+1} \frac{f(x)\, T_p(x)}{\sqrt{1-x^2}}\, dx.$$

b) Théorème de convergence en n de l'interpolation de Tchébicheff

Soit f une fonction de $H^1(\,]-1,+1[\,)$. Ses interpolées de Tchébicheff à (n+1) points convergent uniformément vers f lorsque n tend vers l'infini.

En outre, les coefficients du développement discret en polynômes de Tchébicheff sont donnés par

$$\tilde{a}_p = \frac{1}{1+\delta_{p,0}} \frac{2}{n+1} \sum_{k=0}^{n} f(x_k)\, T_p(x_k) \, .$$

11. ALGORITHME DE DESCENTE DE DEGRE

Il est assez fréquent que l'on cherche une approximation polynomiale aussi simple que possible d'une fonction. Soit, sur $[\,-1,+1\,]$, P_m un polynôme proche de f en ce sens que

$$\| f - P_m \|_{C^0} \leq \varepsilon$$

Quelle est la meilleure approximation uniforme de degré (m-1) de P_m? Supposons que l'on connaisse ce polynôme P_{m-1}. Alors, si α_m est le coefficient de tête de P_m, le polynôme

$$\frac{P_m - P_{m-1}}{\alpha_m}$$

est le polynôme de degré m de coefficient de tête égal à 1 qui a la plus petite borne supérieure en module. En vertu du théorème de Tchébicheff, on a donc

$$P_m - P_{m-1} = \alpha_m T_m^* = \frac{\alpha_m}{2^{m-1}} T_m$$

et

$$\sup_{[\,-1,+1\,]} |P_m - P_{m-1}| = |\alpha_m| / 2^{m-1} \, ,$$

soit

$$\| f - P_{m-1} \|_{C^0} \leq \varepsilon + |\alpha_m| / 2^{m-1} +$$

Il suffit donc, pour obtenir P_{m-1}, d'interpoler P_m aux points de Tchébicheff. On peut utiliser plusieurs fois ce procédé et descendre ainsi de degré, tout en contrôlant l'erreur.

Cette méthode peut être présentée autrement: si l'on développe P_m en polynômes de Tchébicheff,

$$P_m = \sum_{k=0}^{m} a_k T_k ,$$

les polynômes P_r, $r < m$, sont donnés par

$$P_r = \sum_{k=0}^{r} a_k T_k$$

et, d'une façon générale, P_{r-1} est la meilleure approximation uniforme de degré $(r-1)$ de P_r. L'erreur vérifie

$$| P_m(x) - P_r(x) | = | \sum_{k=r+1}^{m} a_k T_k(x) | \leq \sum_{k=r+1}^{m} | a_k | ,$$

si bien que

$$\| f - P_r \|_{C^0} \leq \varepsilon + \sum_{k=r+1}^{m} |a_k| .$$

Comme polynôme de départ, on peut, par exemple, choisir un développement de Taylor suffisamment poussé.

Exercice 1 - Chercher une approximation f de $f(x) = tgx$ dans $[o, \frac{\pi}{2}]$, vérifiant les conditions

$$\begin{cases} \lim\limits_{x \to \pi/2} \dfrac{\tilde{f}(x)}{f(x)} = 1 \\ \lim\limits_{x \to 0} \dfrac{\tilde{f}(x)}{f(x)} = 1 \\ \tilde{f}(\pi/4) = f(\pi/4) \end{cases}$$

Comparer les résultats obtenus par cette fonction aux vraies valeurs de tg x pour les angles suivants:

1°, 10°, 20°, 30°, 40°, 50°, 60°, 70°, 80°, 89° .

(Donner l'erreur en %).

Suggestion: Donner à $\tilde{f}$ la forme

$$\tilde{f}(x) = \frac{x}{\frac{\pi}{2} - x} g(x) ,$$

g(x) étant un polynôme.

Solution: $\tilde{f}(x) = -\frac{4x}{\pi}(x - \frac{\pi}{4}) + \frac{16}{\pi^2} x^2 + \frac{16}{\pi^3} x^2 \frac{x - \frac{\pi}{4}}{\frac{\pi}{2} - x}$.

θ (°)	$\tilde{f}(\theta)$	f(θ)	erreur %
1	0,01748	0,01746	0,2
10	0,1783	0,1763	1,1
20	0,3690	0,3640	1,4
30	0,5836	0,5774	1,1
40	0,8425	0,8391	0,4
50	1,187	1,192	-0,4
60	1,712	1,732	-1,2
70	2,704	2,747	-1,6
80	5,596	5,671	-1,4
89	57,18	57,29	-0,2

Exercice 2 [7] - Trouver la somme des carrés

$S(n) = 1^2 + 2^2 + \ldots + n^2$.

Suggestion: Considérer une interpolée S(x) de S(n) et calculer ses différences divisées, qui possèdent une propriété spéciale.

Solution: On notera que

$$S(n) = S(n)$$
$$S(n,n-1) = \frac{S(n) - S(n-1)}{1} = n^2$$

$$S(n,n-1,n-2) = \frac{S(n,n-1) - S(n-1,n-2)}{2} = \frac{n^2 - (n-1)^2}{2} = \frac{2n-1}{2}$$

$$S(n,n-1,n-2,n-3) = \frac{\frac{2n-1}{2} - \frac{2(n-1)-1}{2}}{3} = \frac{1}{3}$$

$$S(n,n-1,n-2,n-3,n-4) = 0$$

Par conséquent, S(x) est un polynôme de degré 3. On le développera sous la forme

$$S(x) = S(o) + x\, S(0,1) + x(x-1)\, S(0,1,2) + x(x-1)(x-2)\, S(0,1,2,3),$$

avec

$$S(0) = 0$$
$$S(0,1) = 1$$
$$S(0,1,2) = 3/2$$
$$S(0,1,2,3) = 1/3 .$$

Tous calculs faits, on obtient

$$S(x) = \frac{x\,(2x+1)\,(x+1)}{6}$$

et, en particulier,

$$S(n) = \frac{n\,(n+1)\,(2n+1)}{6} .$$

Exercice 3 [7] - Avec quelle précision peut-on calculer $\sqrt{115}$ par interpolation quadratique de la fonction $f(x) = \sqrt{x}$ entre les points $x_0 = 100$, $x_1 = 121$ et $x_2 = 144$? Calculer cette valeur.

Solution: La formule de l'erreur est ici

$$|R(x)| \leq |(x-100)(x-121)(x-144)| \sup_{[100,144]} \frac{|f^{(3)}(x)|}{3!} .$$

Or,

$$f(x) = x^{\frac{1}{2}} \quad , \quad f'(x) = \tfrac{1}{2} x^{-\frac{1}{2}} \quad , \quad f''(x) = -\frac{1}{4} x^{-3/2} \quad , \quad f^{(3)}(x) = \frac{3}{8} x^{-5/2} ,$$

d'où

$$\sup_{[100,\ 144]} |f^{(3)}(x)| = \frac{3}{8} \cdot 100^{-5/2} = \frac{3}{8} \cdot 10^{-5} .$$

Il vient donc, en x = 115,

$$|R(x)| \leq 15.6.29. \frac{(3/8).10^{-5}}{6} = \underline{1,631 \cdot 10^{-3}}$$

La solution est

$$\tilde{f}(x) = \frac{(x-121)(x-144)}{21.44}\ 10 + \frac{(x-100)(144-x)}{21.23}\ 11 + \frac{(x-100)(x-121)}{44.23}\ 12$$

soit, en x= 115,

$$f(115) = 10,722756.$$

Compte tenu de l'évaluation de l'erreur, on obtient donc

$$\sqrt{115} = 10{,}7227 \pm 0{,}0016 = \begin{cases} 10{,}7243 \text{ (max)} \\ 10{,}7211 \text{ (min)} \end{cases}$$

La réponse exacte est 10,722805. On remarquera la précision du procédé et l'exactitude de l'évaluation de l'erreur.

Exercice 4 - Soit une table de logarithmes décimaux à cinq décimales, donnant les logarithmes des entiers de 1000 à 10000 avec une borne d'erreur absolue de $\frac{1}{2}.10^{-5}$. Peut-on, avec une bonne précision, interpoler linéairement entre les unités?

Solution: L'erreur d'une interpolation entre x_0 et x_1 vaut

$$|R(x)| \leq |x-x_0|\,|x-x_1| \sup_{[x_0,x_1]} \frac{|f''(x)|}{2} .$$

Ici,

$$f(x) = \log_{10} x = \frac{\ln x}{\ln 10} \quad , \quad f'(x) = \frac{1}{x \ln 10} \quad , \quad f''(x) = -\frac{1}{x^2 \ln 10} ,$$

et

$$\sup_{[\,1000,\ 10000\,]} |f''(x)|/2 = 0{,}2172.\ 10^{-6}$$

Pour $x_1 = x_0+1$, on a, si x est une valeur intermédiaire,

$$|x-x_0|.|x-x_1| \leq 1/4 ,$$

ce qui donne

$$|R(x)| \leq 5{,}429.10^{-8} ,$$

soit une erreur bien inférieure à celle des tables. En conclusion, l'interpolation linéaire est plus que suffisante.

Exercice 5 - On utilise souvent avec succès, dans les problèmes de concentration de contraintes en élasticité, des interpolations de la forme

$$\tilde{f}(x) = \sqrt{Ax + B} + C \qquad \text{(Expressions de NEUBER)} ,$$

avec les conditions

$$f(x_1) = \alpha_1 \ , \ f(x_2) = \alpha_2 \ , \ f(x_3) = \alpha_3 .$$

Déterminer A, B, C dans le cas suivant:

point	1	2	3
x	0	10	500
α	1	2	17

Solution: Les conditions d'interpolation s'écrivent

$$\sqrt{Ax_1 + B} + C = \alpha_1$$
$$\sqrt{Ax_2 + B} + C = \alpha_2$$
$$\sqrt{Ax_3 + B} + C = \alpha_3 \quad ,$$

soit encore

$$\begin{cases} Ax_1 + B = (\alpha_1 - C)^2 = \alpha_1^2 - 2C\alpha_1 + C^2 \\ Ax_2 + B = (\alpha_2 - C)^2 = \alpha_2^2 - 2C\alpha_2 + C^2 \\ Ax_3 + B = (\alpha_3 - C)^2 = \alpha_3^2 - 2C\alpha_3 + C^2 \end{cases}$$

Soustrayant la première des deux suivantes, on obtient

$$\begin{cases} A(x_2 - x_1) = \alpha_2^2 - \alpha_1^2 - 2C(\alpha_2 - \alpha_1) \\ A(x_3 - x_1) = \alpha_3^2 - \alpha_1^2 - 2C(\alpha_3 - \alpha_1) \end{cases} ,$$

cé qui mène au système

$$\begin{cases} A(x_2 - x_1) + 2C(\alpha_2 - \alpha_1) = \alpha_2^2 - \alpha_1^2 \\ A(x_3 - x_1) + 2C(\alpha_3 - \alpha_1) = \alpha_3^2 - \alpha_1^2 \end{cases}$$

permettant de déterminer A et C. On tire B de la condition

$$B = (\alpha_1 - C)^2 - Ax_1 \ .$$

Application numérique: Le système s'écrit

$10\ A + 2.1.C = 3$ ..... $500\ A + 100\ C = 150$

$500\ A + 2.16.C = 288$ ..... $500\ A + 32\ C = 288$

$68\ C = -138$ $\quad C = -2,029$

$$A = \frac{3 - 2C}{10} = 0,7059$$

$$B = (1+2,029)^2 - 0,7059.0 = 9,175$$

On a donc

$$\tilde{f}(x) = \sqrt{0,7059\ x + 9,175} - 2,029 \ .$$

Exercice 6 - On donne les valeurs suivantes de l'intégrale de probabilité

$$y = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

x	0,45	0,46	0,47	0,48	0,49	0,50
y	0,4754818	0,4846555	0,4937452	0,5027498	0,5116683	0,5204999

En quel x a-t-on $y = \frac{1}{2}$?

Suggestion: Effectuer une interpolation inverse, où y est la variable, et x la fonction.

Solution: on obtient le tableau des différences suivant (x = g(y)) .

y	g_o	g_1	g_2	g_3	g_4	g_5
0,475 481 8	0,45					
		1,090 072				
0,484 655 5	0,46		0,551 594 9			
		1,100 146		0,843 695 9		
0,493 735 2	0,47		0,574 600 8		0,793 055 4	
		1,110 543		0,872 393 8		-16,072 22
0,502 749 8	0,48		0,598 166 6		0,069 514 34	
		1,121 264		0,874 885 5		
0,511 668 3	0,49		0,621 573 9			
		1,322 297				
0,520 499 9	0,50					

Les accroissements de y sont

$Y_0 = 24{,}518\ 20.10^{-3}$

$Y_1 = 15{,}344\ 50.10^{-3}$

$Y_2 = 6{,}254\ 800.10^{-3}$

$Y_3 = -2{,}749\ 800.10^{-3}$

$Y_4 = -11{,}668\ 30.10^{-3}$

$Y_5 = 20{,}499\ 9.10^{-3}$

On obtient $g(\frac{1}{2}) = 0{,}4769361$ (les six premières décimales sont exactes).

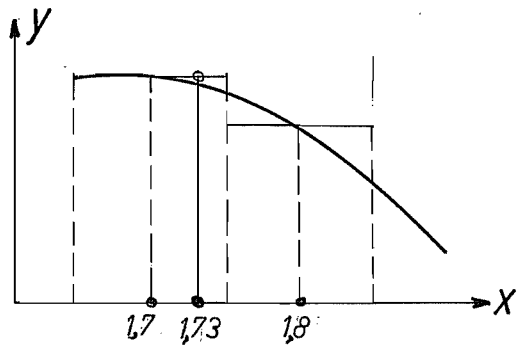


fig. 1

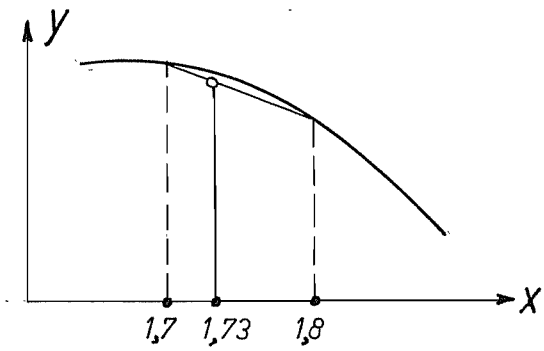


fig. 2

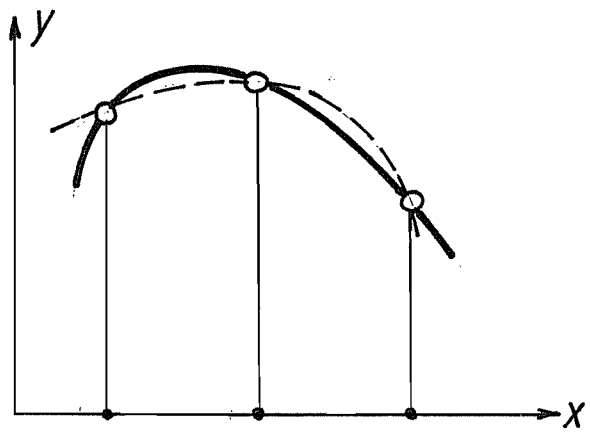


fig. 3

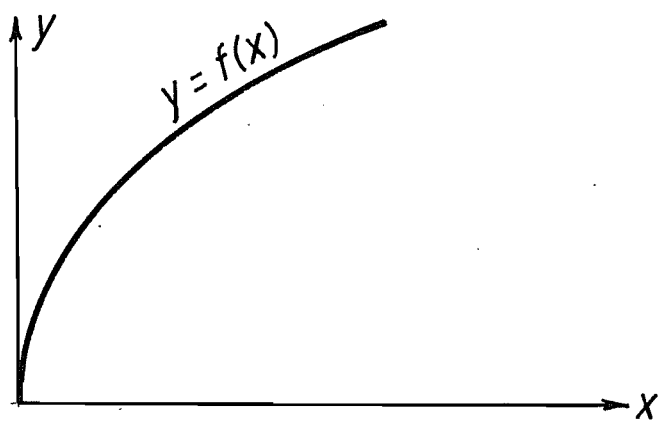


fig. 4

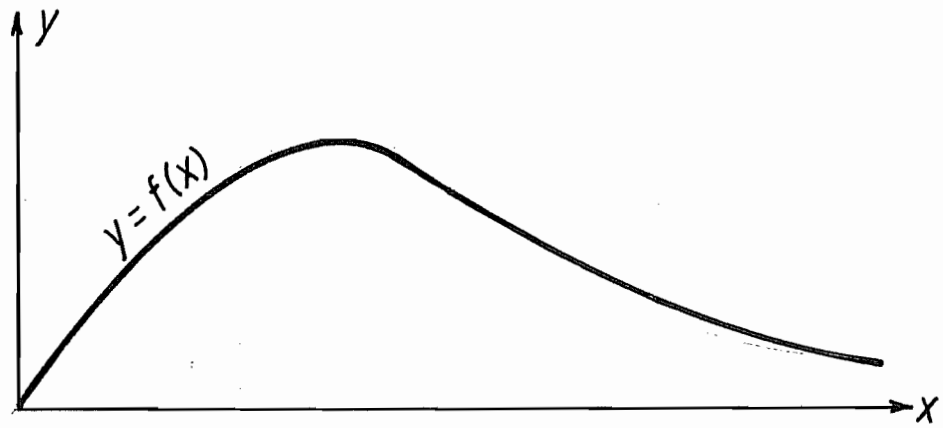


fig. 5

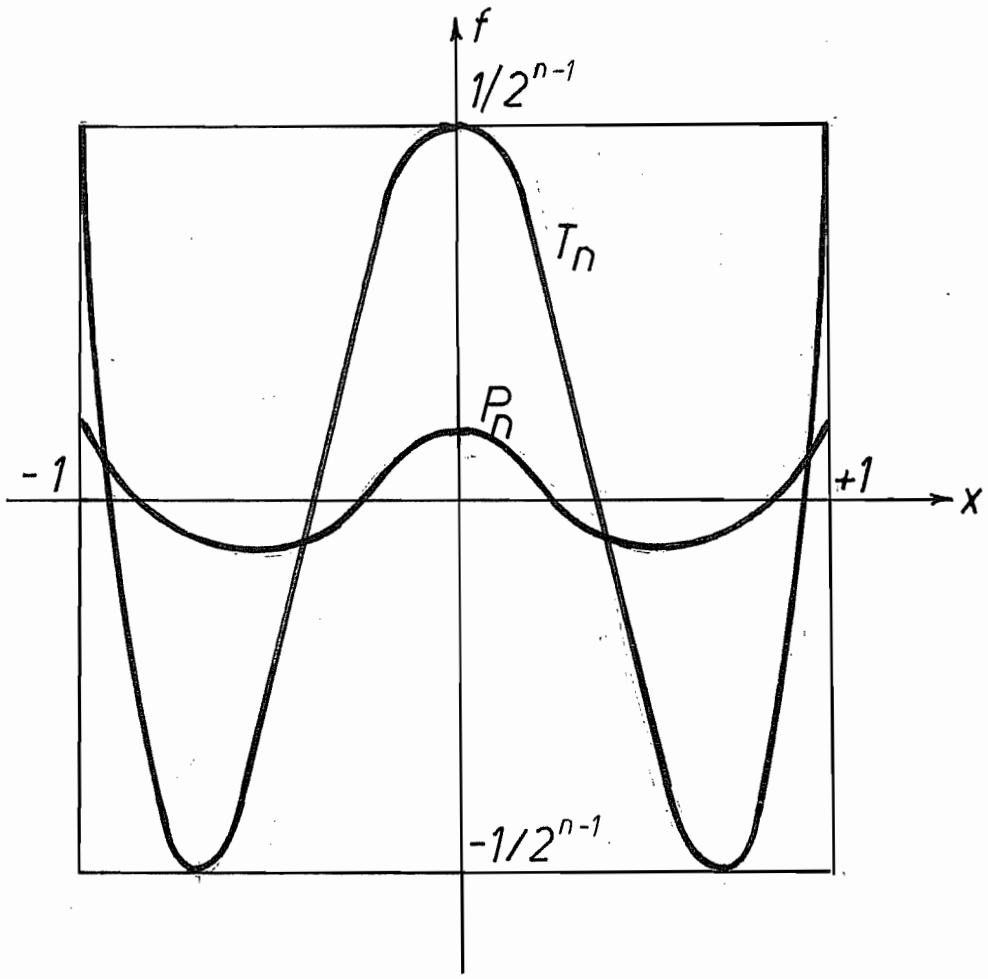


fig. 6

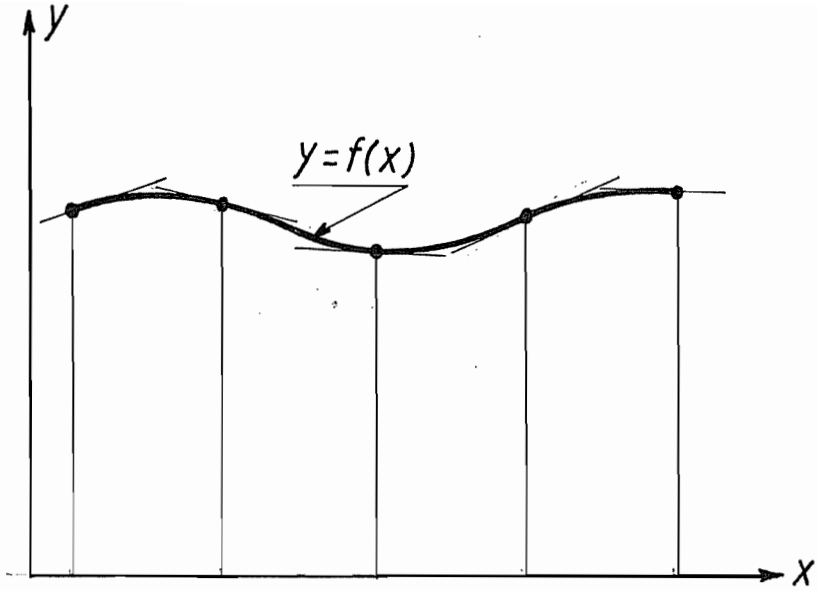


fig. 7

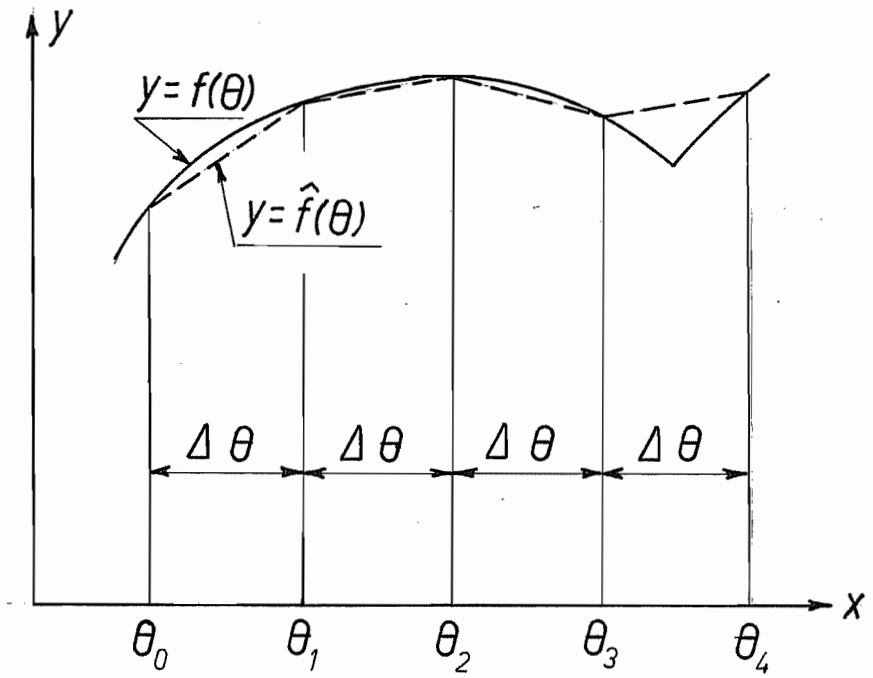


fig. 8

A P P R O X I M A T I O N A U S E N S D E S

M O I N D R E S C A R R E S

---

1. INTRODUCTION

L'interpolation permet, à partir de n points, de définir une fonction dépendant de n paramètres. Mais bien souvent, le problème pratique est tout autre. Ayant fait un certain nombre n d'expériences, on désire ajuster une loi théorique à ces expériences. La loi théorique dépend de $k < n$ paramètres et la question consiste à choisir ceux-ci "le mieux possible". Bien entendu, cette expression est vague et l'interprétation que nous en ferons ci-dessous est quelque peu arbitraire. C'est à l'utilisateur de la méthode de juger si elle correspond effectivement à son cahier de charges.

2. PRINCIPE DE LA METHODE

Etant donné des points expérimentaux $(x_i, \hat{y}_i)$, on admet que x_i est parfaitement connu et que $\hat{y}_i$ est éventuellement entaché d'une erreur de mesure. On veut ajuster une loi de la forme

$$x \longmapsto f(x; \alpha_1, \ldots, \alpha_k)$$

et, comme on ne pourra vérifier simultanément toutes les conditions

$$\hat{y}_1 = f(x_1; \alpha)$$
$$\hat{y}_2 = f(x_2; \alpha)$$
$$\ldots\ldots\ldots$$
$$\hat{y}_n = f(x_n; \alpha) ,$$

on s'efforcera de minimiser la somme des carrés des écarts, ce qui revient à écrire

$$G(\alpha) = \sum_{i=1}^{n} (\hat{y}_i - f(x; \alpha))^2 \qquad \min_{\alpha_1, \ldots, \alpha_k} \qquad (1)$$

Il est utile, dans la présentation de la méthode, de définir une notion de produit scalaire. Etant donné le support $(x_1, \ldots, x_n)$ et deux fonctions f et g, nous écrirons

$$(f,g) = \sum_{i=1}^{n} f(x_i)\, g(x_i) . \qquad (2)$$

Nous utiliserons aussi la semi-norme $\|f\|$ définie par

$$\|f\|^2 = (f,f) = \sum_{i=1}^{n} |f(x_i)|^2 . \qquad (3)$$

Il ne s'agit pas à proprement parler d'une norme, car la condition $\|f\| = 0$ n'entraîne pas $f = 0$. Mais si l'on convient de dire que deux

fonctions f et g sont <u>équivalentes</u>, ce que nous noterons $f \equiv g$, lorsque $f(x_i) = g(x_i)$, $i = 1, \ldots, n$, (ce qui signifie que g est une interpolée de f), la condition $\|f\| = 0$ implique $f \equiv 0$ (*)

La distance des deux fonctions $\hat{y}$ (valeurs expérimentales) et f est donc

$$d^2(\hat{y}, f) = \|\hat{y} - f\|^2 = \sum_{i=1}^{n} (\hat{y}_i - f(x_i; \alpha))^2 , \qquad (4)$$

ce qui permet de se faire une représentation géométrique du problème considéré: <u>il s'agit de minimiser la distance entre f et $\hat{y}$.</u>

On a évidemment

$$d^2(\hat{y}, f) = \|\hat{y} - f\|^2 = \|\hat{y}\|^2 - 2\,(\hat{y}, f) + \|f\|^2$$

et

$$\frac{\partial}{\partial \alpha_j}\, d^2(\hat{y}, f) = 2\,(f, \frac{\partial f}{\partial \alpha_j}) - 2\,(\hat{y}, \frac{\partial f}{\partial \alpha_j}) ,$$

ce qui mène aux <u>équations normales</u>

$$\boxed{(\hat{y} - f, \frac{\partial f}{\partial \alpha_j}) = 0 \quad , \quad j = 1, \ldots, k} \qquad (5)$$

exprimant l'orthogonalité entre l'erreur et les dérivées de f. On peut encore écrire, en utilisant la notion de <u>variation première de f</u>,

$$\delta f = \sum_{j=1}^{k} \frac{\partial f}{\partial \alpha_j}\, \delta \alpha_j ,$$

$$\boxed{(\hat{y} - f, \delta f) = 0} \qquad (6)$$

ce qui exprime que <u>l'erreur</u> $\hat{y} - f$ <u>est orthogonale à toute variation première de f</u>.

3. APPROXIMATION PAR UNE COMBINAISON LINEAIRE DE FONCTIONS DONNEES

Etant donné un ensemble de fonctions $\phi_1, \ldots, \phi_k$ indépendantes, cherchons une approximation de la forme

$$f = \sum_{j=1}^{k} \alpha_j \phi_j .$$

(*) Il s'agit là d'un procédé bien connu en analyse sous le nom de <u>séparation d'un espace</u> [20] .

Il vient alors

$$d^2(f,\hat{y}) = \|f - \hat{y}\|^2 = (f,f) - 2\,(f,\hat{y}) + \|\hat{y}\|^2$$

et

$$\frac{\partial}{\partial \alpha_j} d^2(f,\hat{y}) = 2\,(f,\phi_j) - 2\,(\hat{y},\phi_j) = 0 \ ,$$

soit

$$\sum_{l=1}^{k} \alpha_l\,(\phi_l,\phi_j) = (\hat{y},\phi_j) \ .$$

Il s'agit d'un système linéaire. La matrice

$$G_{lj} = (\phi_l,\phi_j)$$

est appelée matrice de GRAM. Elle est symétrique et définie positive, car si l'un au moins des α_j est non nul,

$$\sum_{lj} \alpha_l\,\alpha_j\,(\phi_l,\phi_j) = \|\sum_l \alpha_l\,\phi_l\|^2 > 0 \ .$$

L'inversion est donc toujours possible.

Exemple : Approximation cubique

$$f = \alpha_o + \alpha_1\,x + \alpha_2\,x^2 + \alpha_3\,x^3 \ .$$

$$d^2(f,\hat{y}) = \sum_i (\hat{y}_i - \alpha_o - \alpha_1\,x_i - \alpha_2\,x_i^2 - \alpha_3\,x_i^3)^2$$

$$\frac{\partial d^2}{\partial \alpha_o} = \alpha_o\,n + \alpha_1 \sum_i x_i + \alpha_2 \sum_i x_i^2 + \alpha_3 \sum_i x_i^3 - \sum_i \hat{y}_i = 0$$

$$\frac{\partial d^2}{\partial \alpha_1} = \alpha_o \sum_i x_i + \alpha_1 \sum_i x_i^2 + \alpha_2 \sum_i x_i^3 + \alpha_3 \sum_i x_i^4 - \sum_i x_i\,\hat{y}_i = 0$$

$$\frac{\partial d^2}{\partial \alpha_2} = \alpha_o \sum_i x_i^2 + \alpha_1 \sum_i x_i^3 + \alpha_2 \sum_i x_i^4 + \alpha_3 \sum_i x_i^5 - \sum_i x_i^2\,\hat{y}_i = 0$$

$$\frac{\partial d^2}{\partial \alpha_3} = \alpha_o \sum_i x_i^3 + \alpha_1 \sum_i x_i^4 + \alpha_2 \sum_i x_i^5 + \alpha_3 \sum_i x_i^6 - \sum_i x_i^3\,\hat{y}_i = 0$$

La matrice de GRAM est donc, en omettant l'indice de sommation,

$$\underset{\sim}{G} = \begin{bmatrix} n & \Sigma x & \Sigma x^2 & \Sigma x^3 \\ \Sigma x & \Sigma x^2 & \Sigma x^3 & \Sigma x^4 \\ \Sigma x^2 & \Sigma x^3 & \Sigma x^4 & \Sigma x^5 \\ \Sigma x^3 & \Sigma x^4 & \Sigma x^5 & \Sigma x^6 \end{bmatrix}$$

et le second membre,

$$\begin{bmatrix} \Sigma \hat{y} \\ \Sigma x\hat{y} \\ \Sigma x^2\hat{y} \\ \Sigma x^3\hat{y} \end{bmatrix} .$$

4. EVALUATION A POSTERIORI DE L'ERREUR

Etant donné la fonction expérimentale $\hat{y}$ et la fonction ajustée f, leur distance relative sera

$$s(\hat{y},f) = \frac{\|\hat{y} - f\|}{\|\hat{y}\|}$$

et peut être calculée a posteriori.

Lorsqu'il s'agit d'une approximation par combinaison linéaire de fonctions données, on a en outre la possibilité de poser, dans la relation

$$(\hat{y} - f, \delta f) = 0 ,$$

avec

$$\delta f = \sum_j \delta\alpha_j \phi_j ,$$

tous les coefficients $\delta\alpha_j$ égaux à α_j, ce qui donne $\delta f = f$ et

$$(\hat{y} - f, f) = 0 ,$$

ce qui signifie que l'erreur $\hat{y} - f$ est orthogonale à la fonction ajustée. On en déduit

$$\|\hat{y} - f\|^2 = \|\hat{y}\|^2 - 2(f,\hat{y}) + \|f\|^2$$

$$= \|\hat{y}\|^2 - 2(f,\hat{y}-f) - \|f\|^2 = \|\hat{y}\|^2 - \|f\|^2 ,$$

et

$$s(\hat{y},f) = 1 - \frac{\|f\|^2}{\|\hat{y}\|^2}$$

Par le théorème de Pythagore (fig. 2), $s(\hat{y},f)$ s'interprète comme le sinus de l'angle θ entre le vecteur $\hat{y}$ et le vecteur f. On utilise aussi le cosinus

$$c(\hat{y},f) = (1 - s^2(\hat{y},f))^{\frac{1}{2}} = \frac{\|f\|}{\|\hat{y}\|} ,$$

qui vaut 1 lorsque l'erreur est nulle et 0 lorsque f = 0 (cas tragique où $\hat{y}$ est orthogonal aux fonctions de base). Ce coefficient coïncide, dans le cas d'une approximation affine, avec le module du coefficient de corrélation des statisticiens.

5. MOINDRES CARRES CONTINUS

Il est parfois utile de remplacer une fonction connue y par une fonction simple f, avec la condition de minimiser la grandeur

$$I((y - f)^2),$$

où I est une mesure définie positive, de la forme

$$I(f) = \int_a^b w(x)\, f(x)\, dx \ ,$$

w étant une densité positive. Les méthodes précédentes s'appliquent encore, à condition de poser

$$(f,g) = I(fg) \quad , \quad \|f\|^2 = I(|f|^2) \ .$$

Exemple - Soit à approcher une fonction y entre 0 et 1, par une fonction affine, de manière à minimiser leur distance pour la norme

$$\|g\|^2 = \int_0^1 x\, |g(x)|^2\, dx.$$

On a

$$(f,g) = \int_0^1 x\, f(x)\, g(x)\, dx \ ,$$

et les conditions sont

$$\int_0^1 x \frac{\partial f}{\partial \alpha_j}\, f\, dx = \int_0^1 x\, y \frac{\partial f}{\partial \alpha_j}\, dx \ .$$

Exercice - On cherche une relation de la forme

$$y = \gamma(1 - x)$$

entre les deux variables x et y données dans le tableau suivant:

x	$\hat{y}$
0,5022	0,5156
0,6688	0,3410
0,8313	0,2073

Calculer par la méthode des moindres carrés, et donner la valeur du coefficient de mesure d'erreur s.

Solution : $= 1{,}048$, $s = 5{,}081.10^{-2}$

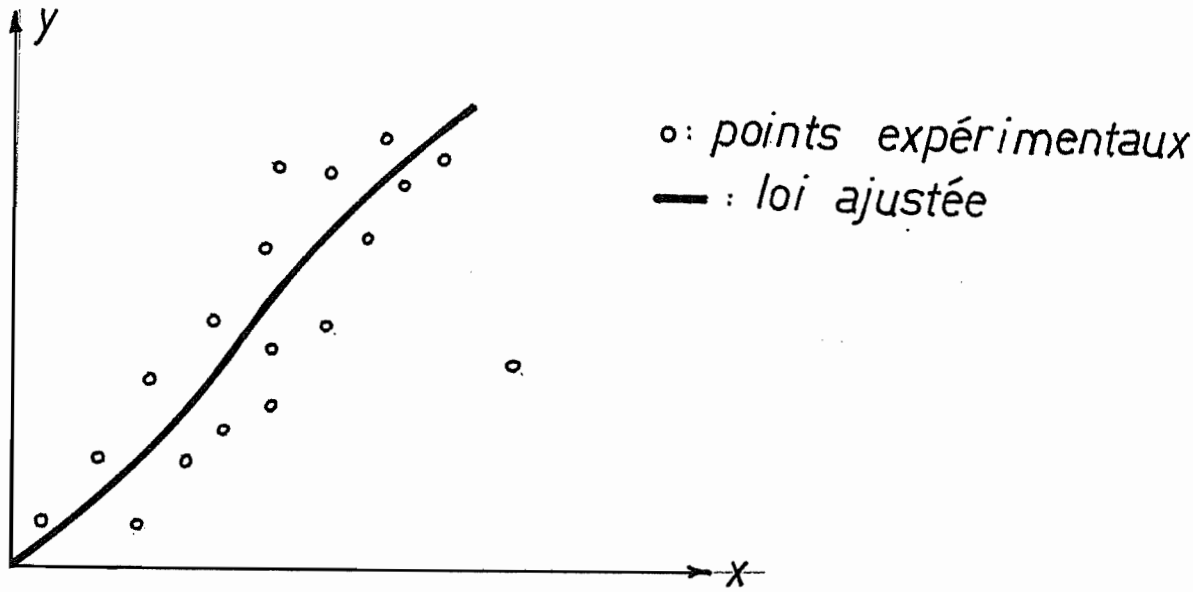


Fig. 1

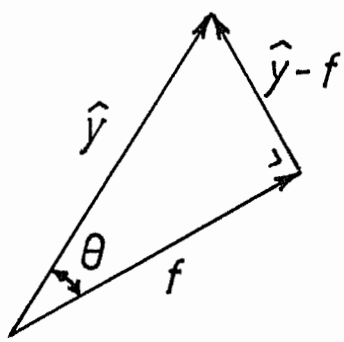


Fig. 2

I N T E G R A T I O N N U M E R I Q U E

1. PRINCIPE GENERAL

On désire calculer numériquement des intégrales de la forme

$$I(f) = \int_a^b w(x)\, f(x)\, dx \quad ,$$

avec a et b finis ou non, w étant une fonction positive donnée appelée densité. Le cas le plus courant correspond à a et b finis, $w(x) = 1$; mais certaines applications utiles nécessitent plus de généralité. Ainsi, les statisticiens s'intéressent beaucoup aux probabilités des formes suivantes:

$$P(f) = \int_0^\infty e^{-x} f(x)\, dx \qquad \text{(Loi de Poisson)}$$

et

$$P(f) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-x^2/2} f(x)\, dx \qquad \text{(Loi normale).}$$

Les méthodes d'intégration numérique les plus répandues sont fondées sur l'interpolation polynomiale et consistent à remplacer la fonction f par un polynôme $\tilde{f}$ qui l'interpole aux points $x_0, \ldots, x_n$:

$$\tilde{f}(x) = \sum_{i=0}^{n} L_i(x)\, f(x_i)$$

et à remplacer l'intégrale de f par celle de $\tilde{f}$, qui est plus simple. L'intégrale approchée $\tilde{I}(f)$ est donc donnée par

$$\tilde{I}(f) = I(\tilde{f}) = \sum_{i=0}^{n} f(x_i)\, I(L_i) = \sum_{i=0}^{n} H_i\, f(x_i) \quad ,$$

c'est-à-dire comme une somme des valeurs de f aux points d'interpolation, pondérée par les poids d'intégration

$$H_i = I(L_i) \ .$$

Pour pouvoir établir la théorie de l'intégration numérique, polynomiale, il faut faire les restrictions suivantes:

a) Quel que soit le polynôme P_m , on a $I(P_m) < \infty$
(régularité de l'intégrale considérée)

b) Chaque fois que $f \geqslant 0$, on a $I(f) \geqslant 0$, l'égalité à zéro n'ayant lieu que si $f = 0$ presque partout (positive définition)

En outre, nous supposerons toujours que les fonctions considérées sont définies partout (sauf éventuellement aux extrémités de l'intervalle).

Soit par exemple

$$I(f) = \int_{-1}^{+1} \frac{f(x)}{\sqrt{1 - x^2}} dx \ .$$

La positive définition est visiblement assurée, si bien que la seule question à examiner est la régularité. On a

$$\lim_{x \to \pm 1} \frac{|x \pm 1|^{\frac{1}{2}} |P_m|}{\sqrt{(1 + x)(1 - x)}} = \frac{|P_m(\pm 1)|}{\sqrt{2}}$$

donc $P_m(x) / \sqrt{(1-x^2)}$ est intégrable sur $(-1, +1)$.

2. CALCUL DES POIDS

Pour calculer les poids, on peut soit partir des relations

$$H_i = I(L_i) \ ,$$

soit noter que, de toute façon, si f est un polynôme P_k de degré k inférieur ou égal au degré n de l'interpolation, son interpolée lui est égale, si bien que

$$\tilde{I}(P_k) = I(\tilde{P}_k) = I(P_k)$$

Ecrivons donc que les fonctions $1, x, x^2, \ldots, x^n$ sont intégrées exactement:

$$\left\{ \begin{array}{llll} H_o & + H_1 & + \ldots\ldots & + H_n = I(1) \\ H_o x_o & + H_1 x_1 & + \ldots\ldots & + H_n x_n = I(x) \\ \vdots & \vdots & & \vdots \\ H_o x_o^n & + H_1 x_1^n & + \ldots\ldots & + H_n x_n^n = I(x^n) \end{array} \right.$$

Il suffit donc de résoudre ce système linéaire pour obtenir les $(n + 1)$ poids.

Cependant, ce système est très mal conditionné lorsque les nombres $x_o, x_1, \ldots, x_n$ sont très différents de l'unité. Lorsque a et b sont finis, on porte aisément remède à cette situation en effectuant une transformation de la forme

$$x = \alpha \hat{x} + \beta \quad ,$$

conduisant à $\hat{x} \in [0 , 1]$ ou $\hat{x} \in [-1 , +1]$. L'intervalle choisi est appelé <u>intervalle de référence</u>. Alors, si les $\hat{x}_i$ sont les images des x_i par cette transformation, on a, en notant en général

$$\hat{g}(\hat{x}) = g(\alpha \hat{x} + \beta) \quad ,$$

la relation

$$H_i = I(L_i) = \int_a^b w(x) L_i(x) dx = h \int_{\hat{a}}^{\hat{b}} \hat{w}(\hat{x}) \hat{L}_i(\hat{x}) dx = h \, \hat{I}(\hat{L}_i) = h \, \hat{H}_i \quad ,$$

ce qui ramène le problème à la résolution du système

$$\begin{cases} \hat{H}_o + \hat{H}_1 + \ldots + \hat{H}_n = \hat{I}(1) \\ \hat{H}_o \hat{x}_o + \hat{H}_1 \hat{x}_1 + \ldots + \hat{H}_n \hat{x}_n = \hat{I}(\hat{x}) \\ \vdots \qquad \vdots \qquad \qquad \vdots \qquad \vdots \\ \hat{H}_o \hat{x}_o^n + \hat{H}_1 \hat{x}_1^n + \ldots + \hat{H}_n \hat{x}_n^n = \hat{I}(\hat{x}^n) \end{cases} \quad ,$$

bien mieux conditionné.

Lorsque l'une au moins des deux limites d'intégration est infinie, on effectue une transformation linéaire ramenant l'intervalle $[x_o , x_n]$ à $[0 , 1]$ ou $[-1 , +1]$, avec le même bénéfice.

3. <u>EXEMPLES ELEMENTAIRES</u>

a) <u>Formule du rectangle</u> - Le plus simple des polynômes est la fonction constante. En interpolant f au centre de l'intervalle, on obtient

$$\int_a^b f(x) \, dx \approx H_o \, f(x_o) \, .$$

On calcule H_o de façon que les constantes soient intégrées exactement:

$$H_o = I(1) = b - a \quad ,$$

d'où

$$\tilde{I}(f) = (b - a) \, f(x_o) \, .$$

Comme le montre la figure 1 , cette formule revient à remplacer l'aire aAX_oBb sous la courbe représentative de la fonction par l'aire du rectangle $aA'B'b$, ce qui justifie le nom de <u>formule du rectangle</u>.

b) Formule du trapèze - En interpolant f linéairement entre a et b, on obtient

$$\tilde{I}(f) = H_o\, f(a) + H_1\, f(b) \quad .$$

Les conditions sont ici

$$\begin{cases} H_o + H_1 = I(1) = b - a \\ H_o a + H_1 b = I(x) = \dfrac{b^2 - a^2}{2} \end{cases} .$$

Il est plus simple de déterminer $\hat{H}_o$ et $\hat{H}_1$ sur $[0\ ,\ 1]$, ce qui donne:

$$\begin{cases} \hat{H}_o + \hat{H}_1 = \hat{I}(1) = 1 \\ \hat{H}_o \cdot 0 + \hat{H}_1 \cdot 1 = \hat{I}(\hat{x}) = \frac{1}{2} \end{cases} ,$$

soit

$$\hat{H}_1 = \tfrac{1}{2} \quad , \quad \hat{H}_o = 1 - \tfrac{1}{2} = \tfrac{1}{2} \quad .$$

Il vient donc

$$H_o = H_1 = \frac{b - a}{2}$$

et

$$\tilde{I}(f) = \frac{b - a}{2}\ (f(a) + f(b)) \quad .$$

Cette formule revient à remplacer, comme l'illustre la figure 2, l'aire aABb par le trapèze aABb.

c) Formule de Simpson - En interpolant f au second degré en

$$x_o = a \quad , \quad x_1 = \frac{a + b}{2} \quad , \quad x_2 = b \quad ,$$

on obtient

$$\tilde{I}(f) = H_o\, f(a) + H_1\, f(\frac{a + b}{2}) + H_2\, f(b) \quad .$$

Pour déterminer les poids, ramenons-nous à l'intervalle $[-1\ ,\ +1]$ par la transformation

$$x = \frac{b + a}{2} + \frac{b - a}{2}\ \hat{x} \quad ,$$

ce qui mène au système

$$\left\{\begin{array}{lllll} \hat{H}_o & + \hat{H}_1 & + \hat{H}_2 & = \hat{I}(1) = 2 \\ -\hat{H}_o & + 0 & + \hat{H}_2 & = \hat{I}(\hat{x}) = 0 \\ \hat{H}_o & + 0 & + \hat{H}_2 & = \hat{I}(\hat{x}^2) = \frac{2}{3} \end{array}\right. ,$$

d'où

$$\hat{H}_o = \hat{H}_2 = \frac{1}{3} \quad , \quad \hat{H}_1 = 2 - \frac{2}{3} = \frac{4}{3} \quad .$$

On en déduit

$$H_o = H_2 = \frac{b - a}{6} \quad , \quad H_1 = 4\,\frac{b - a}{6}$$

et

$$\tilde{I}(f) = \frac{b - a}{6}\,(f(a) + 4\,f(\frac{a + b}{2}) + f(b)) \ .$$

C'est la <u>formule de Simpson</u> (fig. 3).

4. <u>DEGRE D'UNE FORMULE D'INTEGRATION POLYNOMIALE</u>

Par construction, toute formule d'intégration polynomiale à $(n + 1)$ points vérifie, pour $k \leqslant n$,

$$\tilde{I}(P_k) = I(\tilde{P}_k) = I(P_k) \ ,$$

c'est-à-dire qu'elle est exacte pour tous les polynômes de degré n au plus. Mais il peut se faire qu'elle intègre exactement certains polynômes de degré plus élevé. Montrons par exemple que la formule de Simpson, construite avec des polynômes de degré 2, est en fait de degré 3. Il suffit, pour cela, de considérer, sur l'intervalle $(-1 \ , +1)$, la fonction $f(x)=x^3$, dont l'intégrale est nulle. Or,

$$\tilde{I}(f) = \frac{1}{6}((-1)^3 + 4.0 + 1^3) = 0 = I(f).$$

Pour exprimer ce fait, on dit que la formule de Simpson <u>est de degré 3</u>. En général, on dit qu'<u>une formule d'intégration polynomiale à $(n + 1)$ points est de degré p si, pour tout polynôme de degré $r \leqslant p$, on a</u>

$$\tilde{I}(P_r) = I(P_r) \ .$$

Evidemment, $p \geqslant n$. Cette notion joue un rôle important dans le calcul de l'erreur d'intégration.

Le cas de la formule de Simpson n'est pas isolé, comme le montre le résultat suivant: Soit (a,b) un intervalle fini, de centre $c = (a + b)/2$. La densité est dite <u>symétrique</u> si pour tout $y \in (0 \ , \frac{b - a}{2})$, on a

$$w(c + y) = w(c - y).$$

Une formule d'intégration polynomiale relative à une densité symétrique est dite <u>symétrique</u> si ses points d'interpolation x_i sont disposés

symétriquement par rapport au centre de l'intervalle. Dans ce cas, on peut numéroter les x_i de la manière suivante:

$$\left.\begin{array}{l} x_0^* = c \text{ (éventuellement)} \\ x_1^* = c + y_1 \quad , \quad x_1^{**} = c - y_1 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ x_m^* = c + y_m \quad , \quad x_m^{**} = c - y_m \end{array}\right\} - 0 < y_1 < y_2 < \ldots < y_m$$

et

$$n + 1 = \begin{cases} 2m \text{ si le nombre de points est pair} \\ 2m + 1 \text{ si le nombre de points est impair} \end{cases}$$

Alors, en posant $x = c + y$ dans la formule des polynômes de Lagrange, on a, pour $i \neq 0$,

$$L_i^*(c + y) = \prod_{\substack{j=1 \\ \neq i}}^{m} \frac{(y^2 - y_j^2)}{(y_i^2 - y_j^2)} \cdot \frac{y + y_i}{2y_i} \left[\cdot \frac{y}{y_i} \right]$$

$$L_i^{**}(c - y) = \prod_{\substack{j=1 \\ \neq i}}^{m} \frac{(y^2 - y_j^2)}{(y_i^2 - y_j^2)} \cdot \frac{-(y + y_i)}{-2y_i} \left[\cdot \frac{-y}{-y_i} \right]$$

les facteurs entre crochets ne devant être pris en considération que si $(n + 1)$ est impair. Visiblement,

$$L_i^*(c + y) = L_i^{**}(c - y)$$

et

$$H_i^* = \int_{-\frac{b-a}{2}}^{\frac{b-a}{2}} w(c+y)\, L_i^*(c+y)\,dy = \int_{\frac{b-a}{2}}^{-\frac{b-a}{2}} w(c-y)\, L_i^{**}(c-y)\,d(-y) = H_i^{**},$$

c'est-à-dire que <u>les poids d'une formule symétrique sont symétriques.</u>

Cela étant, <u>si $\tilde{I}$ représente une formule symétrique à $(n + 1)$ points, avec $(n + 1)$ IMPAIR, elle est au moins de degré $(n + 1)$</u>. En effet, en notant

$$\prod(x) = (x - x_0) \cdot \ldots \cdot (x - x_n),$$

tout polynôme de degré $(n+1)$ est de la forme

$$P_{n+1}(x) = \alpha \prod(x) + Q_n(x),$$

et le reste $Q_n(x)$, de degré n, est intégré exactement. Quant à la fonction $\prod(x)$, elle vérifie

$$\prod(c+y) = y \prod_{j=1}^{m} (y^2 - y_j^2) \quad , \quad \prod(c-y) = -y \prod_{j=1}^{m} (y^2 - y_j^2) \quad ,$$

d'où

$$\int_a^b w(x)\Pi(x)\,dx = \int_0^{\frac{b-a}{2}} w(c+y)\,(\Pi(c+y) + \Pi(c-y))\,dy = 0 \quad ,$$

ce que donne bien la formule numérique :

$$\tilde{I}(\Pi) = \sum_i H_i\,\Pi(x_i) = \sum_i H_i \cdot 0 = 0$$

5. ERREUR D'INTEGRATION

5.1 - Résultats fondamentaux

Soit f une fonction dont on calcule une intégrale à l'aide d'une formule de degré p, à (n + 1) points. (On a toujours $p \geq n$). Posons

$$\Pi^*_{p+1}(x) = (x - x_o) \ldots (x - x_n)\, Q_{p-n}(x) \quad ,$$

où

$$Q_{p-n}(x) = \begin{cases} 1 & \text{si } p = n \\ (x - x_{n+1}) \ldots (x - x_p) & \text{si } p \neq n \ , \end{cases}$$

les points x_{n+1} , ... , x_p étant choisis <u>arbitrairement</u> dans $]a\ ,\ b[$. Soit $P^*_p(x)$ l'interpolée de Newton de f aux points x_o, ... ,x_p . On a donc

$$f(x) = P^*_p(x) + \Pi^*_{p+1}(x)\, f(x_o, \ldots , x_p, x) \ ,$$

ce qui entraîne

$$\tilde{I}(f) = \tilde{I}(P^*_p) + \tilde{I}(\Pi^*_{p+1}\, f(x_o, \ldots ,x_p, x)).$$

Le dernier terme de cette expression est nul, car Π^*_{p+1} s'annule en x_o, ... , x_n. Comme, par ailleurs, on a par hypothèse

$$\tilde{I}(P^*_p) = I(P^*_p) \quad ,$$

il vient

$$\boxed{\Delta I(f) = I(f) - \tilde{I}(f) = I(\Pi^*_{p+1}\, f(x_o, \ldots, x_p, x))} \qquad (1)$$

Ce résultat est général. Moyennant l'hypothèse supplémentaire que $f \in C^{p+1}(\,]a, b[\,)$, on peut écrire

$$f(x_o, \ldots , x_p, x) = \frac{f^{(p+1)}(\xi)}{(p+1)!} \quad ,$$

ξ étant un certain point compris entre les extrêmes des arguments de la différence divisée. Il en découle

$$\boxed{\Delta I(f) = I\left(\Pi^*_{p+1}\ \frac{f^{(p+1)}(\xi)}{(p+1)!}\right)} \qquad (2)$$

Dans les cas où Π^*_{p+1} est de signe constant sur l'intervalle, il existe, par le théorème de la moyenne, un point $\eta \in]a, b[$ tel que

$$\boxed{\Delta I(f) = \frac{f^{(p+1)}(\eta)}{(p+1)!} \; I(\Pi^*_{p+1})} \qquad (3)$$

Cette relation reste d'ailleurs vraie si l'on a pu montrer par un moyen quelconque que

$$I(f) = K f^{(p+1)}(\eta) \quad ,$$

avec K indépendante de f : il suffit alors de considérer la fonction

$$f(x) = \frac{x^{p+1}}{(p+1)!}$$

pour obtenir, par la majoration (2), la valeur

$$K = \frac{I(\Pi^*_{p+1})}{(p+1)!} \quad .$$

Sur les intervalles finis, on a en outre

$$|\Delta I(f)| \leq I(|\Pi^*_{p+1}|) \sup_{(a,b)} \frac{|f^{p+1}(x)|}{(p+1)!} \quad .$$

Notons encore dans ce cas la majoration grossière

$$|\Delta I(f)| \leq h^{p+2} \sup_{(a,b)} \frac{|f^{p+1}(x)|}{(p+1)!} \quad , \quad h = b - a \quad .$$

5.2 - Calcul de l'erreur sur un intervalle de référence

Dans le cas où l'on connaît, sur un intervalle de référence de longueur $\hat{h}$, la valeur de $\hat{I}((\hat{x} - \hat{x}_o) \ldots (\hat{x} - \hat{x}_p))$, on a immédiatement

$$I((x - x_o) \ldots (x - x_p)) = \frac{h}{\hat{h}} \hat{I}(\frac{h}{\hat{h}}(\hat{x} - \hat{x}_o) \ldots \frac{h}{\hat{h}}(\hat{x} - \hat{x}_p))$$

$$= (\frac{h}{\hat{h}})^{p+2} \hat{I}((\hat{x} - \hat{x}_o) \ldots (\hat{x} - \hat{x}_p)).$$

De la même façon,

$$I(|x - x_o| \ldots |x - x_p|) = (\frac{h}{\hat{h}})^{p+2} \hat{I}(|\hat{x} - \hat{x}_o| \ldots |\hat{x} - \hat{x}_p|) \; .$$

5.3 - Exemple

Calculons l'erreur de la formule de Simpson. Nous savons que cette formule est de degré 3. Sur l'intervalle $]-1, +1[$ de longueur 2, choisissons les points

$$x_o = -1 \quad , \; x_1 = 0 \; , \quad x_2 = 1 \; , \quad x_3 = 0 \; ,$$

ce qui donne

$$\Pi^*_4(x) = x^2(x^2 - 1) \; ,$$

de signe négatif sur tout l'intervalle. On a donc à calculer

$$\int_{-1}^{+1} (x^4 - x^2)\, dx = \frac{2}{5} - \frac{2}{3} = -\frac{4}{15} .$$

Sur un intervalle de longueur h, l'erreur vaut donc

$$\Delta I(f) = -\frac{4}{15} \left(\frac{h}{2}\right)^5 \frac{f^{IV}(\xi)}{4!} = -\frac{h^5}{2880} f^{IV}(\xi) ,$$

ξ étant un point intérieur à l'intervalle.

6. CONVERGENCE PAR DECOUPAGE DE L'INTERVALLE

6.1 - Dans le cas d'une densité w = 1 sur l'intervalle (a, b) de longueur finie, on peut aisément découper l'intervalle en N sous-intervalles $e_1, e_2, \ldots, e_N$ de longueur

$$h = (b - a)/N$$

et appliquer dans chacun d'eux une formule unique de degré p. Si $\tilde{I}_k$ est l'intégrale approchée sur e_k, on a évidemment

$$\tilde{I} = \sum_{k=1}^{N} \tilde{I}_k .$$

Comme

$$|\Delta I_k| \leq h^{p+2} \sup_{e_k} \frac{|f^{p+1}(x)|}{(p+1)!} \leq h^{p+2} \sup_{(a,b)} \frac{|f^{p+1}(x)|}{(p+1)!} ,$$

on a évidemment

$$|\Delta I| \leq \sum_{k=1}^{N} |\Delta I_k| \leq N h^{p+2} \sup_{(a,b)} \frac{|f^{p+1}(x)|}{(p+1)!}$$

$$\leq (b - a) h^{p+1} \sup_{(a,b)} \frac{|f^{p+1}(x)|}{(p+1)!} ,$$

et, pour $h \longrightarrow 0$, $\Delta I(f) \longrightarrow 0$, c'est-à-dire que l'intégrale approchée converge vers l'intégrale exacte (convergence en h^{p+1}) .

6.2 - On peut même généraliser cette conclusion au cas bien plus général d'une fonction f intégrable au sens de Riemann, à condition que les poids H_i de la formule soient positifs.

Rappelons qu'une fonction f est intégrable au sens de Riemann si, pour tout $\varepsilon > 0$, on peut trouver des fonctions étagées φ et ψ telles que

$$\varphi(x) \leq f(x) \leq \psi(x) \qquad \text{sur } (a, b)$$

et

$$I(\psi - \varphi) \leq \varepsilon .$$

a) Commençons par envisager le cas d'une fonction caractéristique de semi-intervalle

$$\delta_{[c,\,d[}(x) = \begin{cases} 1 \text{ pour } & x \in [c,\,d[\\ 0 \text{ pour } & x \in \complement [c,\,d[\end{cases}$$

Dans ce cas (voir figure 4), on obtient l'intégrale exacte dans tous les sous-intervalles, excepté ceux qui contiennent les points c et d. Pour ceux-ci, on a évidemment

$$0 \leqslant \delta_{[c,\,d[}(x) \leqslant 1 \ ,$$

si bien que, en appelant e_c et e_d ces deux sous-intervalles, et en notant δ pour $\delta_{[c,\,d[}$, on a, pour $h = (b - a)/n$,

$$0 \leqslant I_{e_c}(\delta) \leqslant h \qquad \text{et} \qquad 0 \leqslant I_{e_d}(\delta) \leqslant h \ .$$

La même inégalité reste vraie pour les intégrales numériques, <u>pourvu que les poids soient positifs</u>, ce qui mène au respect des relations d'ordre:

$$0 \leqslant \tilde{I}_{e_c}(\delta) \leqslant h \qquad \text{et} \qquad 0 \leqslant \tilde{I}_{e_d}(\delta) \leqslant h \ .$$

En combinant ces relations d'inégalité, on obtient

$$\begin{cases} I_{e_c}(\delta) - \tilde{I}_{e_c}(\delta) \leqslant h - 0 \leqslant h \\ \tilde{I}_{e_c}(\delta) - I_{e_c}(\delta) \leqslant h - 0 \leqslant h \end{cases}$$

et de même pour I_{e_d} , ce qui entraîne

$$|I(\delta) - \tilde{I}(\delta)| \leqslant |I_{e_c}(\delta) - \tilde{I}_{e_c}(\delta)| + |I_{e_d}(\delta) - \tilde{I}_{e_d}(\delta)| \leqslant 2h$$

et il suffit de prendre h suffisamment petit (c.-à-d. N suffisamment grand) pour rendre la différence aussi petite que l'on veut.

b) La généralisation au cas des fonctions étagées est immédiate, puis qu'il s'agit de combinaisons linéaires finies de fonctions caractéristiques de semi-intervalles. On a, dans ce cas,

$$|\Delta I(f)| = |I(\sum_{i=1}^{K} \alpha_i \delta_i(x)) - \tilde{I}(\sum_{i=1}^{K} \alpha_i \delta_i(x))|$$

$$\sum_{i=1}^{K} |\alpha_i| |I(\delta_i) - \tilde{I}(\delta_i)| \ .$$

Fixons $\varepsilon > 0$. Pour un maillage composé de N sous-intervalles, avec

$$N \geqslant \frac{\sum_{i=1}^{K} |\alpha_i| \ . \ 2 \ . \ (b - a)}{\varepsilon} \ ,$$

on obtient

$$|\Delta I(f)| \leq \sum_{i=1}^{K} |\alpha_i| \cdot 2h = \sum_{i=1}^{K} |\alpha_i| \cdot 2 \cdot \frac{b-a}{N} \leq \varepsilon$$

c) Passons enfin au cas d'une fonction intégrable au sens de Riemann. Pour ε fixé, on peut trouver deux fonctions étagées φ et ψ telles que

$$\varphi \leq f \leq \psi \qquad \text{et} \qquad I(\psi) - I(\varphi) \leq \frac{\varepsilon}{2} \quad .$$

Comme l'intégrale respecte l'ordre, on a encore

$$I(\varphi) \leq I(f) \leq I(\psi),$$

ce qui entraîne

$$|I(f) - I(\varphi)| \leq \frac{\varepsilon}{2} \qquad \text{et} \qquad |I(\psi) - I(f)| \leq \frac{\varepsilon}{2} \quad .$$

Par ailleurs, en choisissant une maille h suffisamment petite, on peut obtenir simultanément

$$|I(\varphi) - \tilde{I}(\varphi)| \leq \frac{\varepsilon}{2} \qquad \text{et} \qquad |I(\psi) - \tilde{I}(\psi)| \leq \frac{\varepsilon}{2} \quad .$$

Enfin, si la formule $\tilde{I}$ est à poids positifs, elle respecte les relations d'ordre, ce qui implique

$$\tilde{I}(\varphi) \leq I(f) \leq \tilde{I}(\psi).$$

Mais alors,

$$\tilde{I}(f) \leq \tilde{I}(\psi) \leq I(\psi) + \frac{\varepsilon}{2} \leq I(f) + \varepsilon$$

et

$$\tilde{I}(f) \geq \tilde{I}(\varphi) \geq I(\varphi) - \frac{\varepsilon}{2} \geq I(f) - \varepsilon \quad ,$$

donc

$$I(f) - \varepsilon \leq \tilde{I}(f) \leq I(f) + \varepsilon \quad .$$

On a donc démontré le théorème suivant:

<u>La convergence par découpage de l'intervalle a lieu pour toute fonction intégrable au sens de Riemann, pour autant que l'on emploie une formule à poids positifs</u>.

6.3 - Peut-on élargir ce résultat aux fonctions intégrables au sens de Lebesgue? La réponse est négative, comme le montre le contre-exemple suivant.

Donnons-nous une formule pour un sous-intervalle. Son support, c'est-à-dire l'ensemble des points utilisés dans la formule, forme un ensemble de mesure nulle. Il en est de même de l'ensemble E_N des points utilisés lors d'une subdivision en N sous-intervalles. L'ensemble

$$E = \bigcup_{N=1}^{\infty} E_N \quad ,$$

union dénombrable d'ensembles de mesure nulle, est également de mesure nulle.

Considérons alors la fonction

$$f(x) = \begin{cases} 1 & \text{si } x \in E \\ 0 & \text{si } x \in \complement E \end{cases} .$$

Son intégrale est nulle (f est nulle presque partout), mais sur n'importe quel maillage, on trouve $\tilde{I}(f) = I(1) > 0$.

7. CONVERGENCE EN DEGRE

Si, en conservant un intervalle (a, b) fixe, on considère des formules de degrés de plus en plus élevés (avec de plus en plus de points), les intégrales approchées convergent-elles vers la valeur exacte de l'intégrale? La réponse à cette question dans le cas d'intervalles bornés repose sur le théorème de Weierstrass sur l'approximation uniforme des fonctions continues par des polynômes. Nous aurons aussi besoin ultérieurement de ce théorème sur un compact K de R^n, ce qui justifie sa démonstration dans ce cadre général.

7.1 - Soit K un compact de R^n. Notons $C^o(K)$ l'ensemble des fonctions continues sur K, muni de la norme

$$\|f\| = \sup_K |f(x)|$$

qui mène à la notion de convergence uniforme. Cette norme convient bien, car la convergence uniforme préserve la continuité et est passible du critère de CAUCHY. $C^o(K)$ est donc un espace de Banach.

Appelons $\mathcal{P}(K)$ l'ensemble des restrictions à K de polynômes de R^n. On a évidemment $\mathcal{P}(K) \subset C^o(K)$. Soit alors $\overline{\mathcal{P}}(K)$ son adhérence dans $C^o(K)$, c'est-à-dire l'ensemble des limites uniformes sur K de polynômes. $\overline{\mathcal{P}}(K)$ est évidemment fermé. Nous allons commencer par montrer quelques propriétés de $\overline{\mathcal{P}}(K)$

Lemme 1 - Soit $f \in \overline{\mathcal{P}}(K)$, avec $0 \leq f \leq 1$. Alors, $\sqrt{f} \in \overline{\mathcal{P}}(K)$

Considérons en effet la suite

$$\begin{cases} f_o = 0 \\ f_{m+1} = f_m + \frac{1}{2}(f - f_m^2) \end{cases}$$

Tous les éléments de cette suite vérifient $f_m \leq \sqrt{f}$, car s'il en est ainsi de f_m, on a

$$f_{m+1} = f_m + \tfrac{1}{2}(\sqrt{f} + f_m)(\sqrt{f} - f_m) \leqslant f_m + \tfrac{1}{2} \cdot 2 \cdot (\sqrt{f} - f_m) = \sqrt{f}$$

et c'est évidemment le cas pour f_o.

Par ailleurs, cette suite est croissante, car

$$f_{m+1} = f_m + \tfrac{1}{2}(\sqrt{f} + f_m)(\sqrt{f} - f_m) \geqslant f_m .$$

Il en découle qu'elle converge (croissante et bornée). Bien plus, par le théorème de Dini, une suite de fonctions croissante et convergeant simplement converge aussi uniformément. Donc, la limite g est élément de $\overline{\mathcal{P}}(K)$. Or, elle vérifie

$$g = g + \tfrac{1}{2}(f - g^2),$$

soit $g = \sqrt{f}$.

Ce lemme admet des conséquences importantes:

a) Si $f \in \overline{\mathcal{P}}(K)$, on a aussi $|f| \in \overline{\mathcal{P}}(K)$

En effet, la fonction

$$g = \frac{f^2}{\|f\|^2}$$

est élément de $\overline{\mathcal{P}}(K)$ et vérifie $0 \leqslant g \leqslant 1$, ce qui permet d'affirmer que $\sqrt{g} \in \overline{\mathcal{P}}(K)$. Or,

$$\sqrt{g} = \frac{|f|}{\|f\|} .$$

b) Soient f_1 et f_2 deux fonctions de $\overline{\mathcal{P}}(K)$. Alors, les fonctions $\sup(f_1, f_2)$ et $\inf(f_1, f_2)$ sont éléments de $\overline{\mathcal{P}}(K)$

En effet,

$$\sup(f_1, f_2) = \tfrac{1}{2}(f_1 + f_2) + \tfrac{1}{2}|f_1 - f_2|$$

$$\inf(f_1, f_2) = \tfrac{1}{2}(f_1 + f_2) - \tfrac{1}{2}|f_1 - f_2| .$$

Lemme 2 - Etant donné deux points différents a et b de K et deux nombres α et β , il existe un élément f de $\overline{\mathcal{P}}(K)$ tel que

$$f(a) = \alpha \qquad \text{et} \qquad f(b) = \beta$$

Il existe en effet une coordonnée au moins qui diffère d'un point à l'autre, soit $a_i \neq b_i$. Le polynôme

$$f(x) = \frac{b_i - x_i}{b_i - a_i}\,\alpha + \frac{x_i - a_i}{b_i - a_i}\,\beta$$

convient.

7.2 - Théorème de Weierstrass

Toute fonction f continue dans un compact K est limite uniforme dans K de polynômes.

Cet énoncé revient à dire que $C^o(K) = \overline{\mathcal{P}}(K)$. Comme ce dernier ensemble est fermé, il suffit de montrer que tout élément de $C^o(K)$ est limite uniforme de fonctions de $\overline{\mathcal{P}}(K)$.

a) Quel que soit ε, il existe un élément g de $\overline{\mathcal{P}}(K)$ tel que $g(y) = f(y)$, avec y fixé d'avance, et $g(x) \geq f(x) - \varepsilon$ sur K.

Associons à chaque x de K un polynôme P_x tel que $P_x(x)=f(x)$ et $P_x(y)=f(y)$. Par continuité, ce polynôme vérifie

$$P_x \geq f-\varepsilon$$

dans une boule ouverte B_x centrée en x. L'ensemble des boules B_x recouvrant K, il résulte du théorème de recouvrement que l'on peut recouvrir K par un nombre fini de ces ouverts, soit $B_{x_1}, B_{x_2}, \ldots, B_{x_r}$. Alors, la fonction

$$g(x) = \sup(P_{x_1}, \ldots, P_{x_r})$$

est élément de $\overline{\mathcal{P}}(K)$ et satisfait à la question.

b) Quel que soit ε, il existe un élément h de $\overline{\mathcal{P}}(K)$ tel que $f(x) - \varepsilon \leq h(x) \leq f(x) + \varepsilon$.

Associons à chaque y de K la fonction g_y définie ci-dessus. Par continuité, cette fonction vérifie

$$g_y \leq f + \varepsilon$$

dans une boule ouverte B_y. De toute façon, elle vérifie

$$g_y \geq f - \varepsilon$$

dans K tout entier. L'ensemble des boules B_y recouvre K: on peut donc en extraire un recouvrement fini $B_{y_1}, \ldots, B_{y_s}$. Alors la fonction

$$h(x) = \inf(g_{y_1}, \ldots, g_{y_s})$$

est élément de $\overline{\mathcal{P}}(K)$ et satisfait à la question, ce qui démontre le théorème.

7.3 - Remarque - On peut renforcer le théorème de Weierstrass de la manière suivante: Etant donné un compact K de R^n, un point $x_o \in K$ et une fonction $f \in C^o(K)$, pour tout $\varepsilon > 0$ fixé, on peut trouver un polynôme P_m tel que $P_m(x_o) = f(x_o)$ et que

$$\sup_K |f(x) - P_m(x)| \leq \varepsilon$$

Soit en effet P_m tel que $\|f - P_m\| \leq \varepsilon/2$. Le polynôme

$$P_m^*(x) = P_m(x) + f(x_o) - P_m(x_o)$$

vérifie pour tout $x \in K$

$$|f(x) - P_m^*(x)| \leq |f(x) - P_m(x)| + |f(x_o) - P_m(x_o)| \leq \varepsilon .$$

Nous utiliserons cette remarque au § 15.

7.4 - Théorème de convergence en degré

Soit une famille de formules d'intégration approchée sur le compact $[a, b]$, de degrés croissant à l'infini, et vérifiant la condition

$$\sum_{i=0}^{n} |H_i| \leq C$$

indépendemment de n. Alors, les intégrales approchées de toute fonction f continue sur $[a, b]$ convergent vers l'intégrale de f.

Soit en effet à obtenir $|\Delta I(f)| \leq \varepsilon$. On choisit un polynôme P_m tel que

$$\sup_{[a, b]} |f(x) - P_m(x)| \leq \frac{\varepsilon}{I(1) + C} \quad .$$

Si m est le degré de P_m , pour toute formule de degré $p \geq m$, on a

$$\begin{aligned} |\Delta I(f)| &= |I(f - P_m) - \tilde{I}(f - P_m) + I(P_m) - \tilde{I}(P_m)| \\ &\leq |I(f - P_m)| + |\tilde{I}(f - P_m)| \\ &\leq \frac{\varepsilon}{I(1) + C} \left(I(1) + \sum_i |H_i| \right) \leq \varepsilon \quad . \end{aligned}$$

On notera en particulier que les formules à poids positifs vérifient toujours

$$I(1) = \sum_i H_i = \sum_i |H_i| ,$$

si bien que les formules à poids positifs convergent toujours en degré sur un compact lorsque la fonction f à intégrer est continue.

7.5 - On peut étendre ce résultat à toute fonction intégrable au sens de Riemann pour la densité considérée. Ces fonctions se définissent comme en 6.2. Nous suivrons d'ailleurs la même démarche que pour la convergence par découpage de l'intervalle.

a) Nous commencerons par établir la convergence en degré pour une fonction caractéristique de semi-intervalle

$$\delta_{[c, d[}(x) = \begin{cases} 1 & \text{si} \quad x \in [c, d[\\ 0 & \text{sinon} \end{cases}$$

que nous noterons simplement δ dans la suite. Considérons les fonctions continues φ_n définies par

$$\varphi_n(x) = \begin{cases} 1 & \text{dans }]c, d[\\ 0 & \text{dans }]a, c - \frac{1}{n}[\cup]d + \frac{1}{n}, b[\\ \dfrac{x - c - \frac{1}{n}}{1/n} & \text{dans }]c - \frac{1}{n}, c[\\ \dfrac{d + \frac{1}{n} - x}{1/n} & \text{dans }]d, d + \frac{1}{n}[\end{cases}$$

(voir figure 5). Les fonctions φ_n sont majorées par la fonction intégrable φ_1 et convergent presque partout vers δ, donc, par le théorème de Lebesgue,

$$I(\varphi_n) \longrightarrow I(\delta).$$

Pour $\varepsilon > 0$ fixé, on peut donc trouver une certaine valeur n_o pour laquelle

$$|I(\varphi_{n_o}) - I(\delta)| \leq \frac{\varepsilon}{2}$$

La fonction φ_{n_o} étant continue, il existe un degré p_o à partir duquel les formules numériques vérifient

$$|I(\varphi_{n_o}) - \tilde{I}(\varphi_{n_o})| \leq \frac{\varepsilon}{2}$$

Alors,

$$\begin{aligned} \tilde{I}(\varphi_{n_o}) &= \tilde{I}(\varphi_{n_o}) - I(\varphi_{n_o}) + I(\varphi_{n_o}) - I(\delta) + I(\delta) \\ &\leq |\tilde{I}(\varphi_{n_o}) - I(\varphi_{n_o})| + |I(\varphi_{n_o}) - I(\delta)| + I(\delta) \\ &\leq I(\delta) + \varepsilon . \end{aligned}$$

Les formules considérées étant à poids positifs, la relation $\delta \leq \varphi_{n_o}$ entraîne

$$\tilde{I}(\delta) \leq \tilde{I}(\varphi_{n_o}) .$$

On a donc

$$\tilde{I}(\delta) \leq \tilde{I}(\varphi_{h_o}) \leq I(\delta) + \varepsilon .$$

De la même façon, les fonctions continues définies par

$$\psi_n(x) = \begin{cases} 1 \quad \text{dans } \left[c + \frac{1}{n}, d - \frac{1}{n} \right[\\ 0 \quad \text{dans } [a, c[\cup]d, b] \\ \frac{x - c}{1/n} \quad \text{dans } [c, c + \frac{1}{n}[\\ \frac{d - x}{1/n} \quad \text{dans }]d - \frac{1}{n}, d[\end{cases}$$

(voir figure 6) sont inférieures à δ et convergent presque partout vers δ, donc, par le théorème de Lebesgue, $I(\psi_n) \longrightarrow I(\delta)$, ce qui permet d'affirmer qu'il existe une valeur n_1 pour laquelle

$$|I(\delta) - I(\psi_{n_1})| \leqslant \frac{\varepsilon}{2} \quad .$$

La fonction ψ_{n_1} étant continue, il existe un degré p_1 à partir duquel les formules numériques vérifient

$$|I(\psi_{n_1}) - \tilde{I}(\psi_{n_1})| \leqslant \frac{\varepsilon}{2} \quad .$$

Alors,

$$\begin{aligned} \tilde{I}(\psi_{n_1}) &= \tilde{I}(\psi_{n_1}) - I(\psi_{n_1}) + I(\psi_{n_1}) - I(\delta) + I(\delta) \\ &\geqslant -|\tilde{I}(\psi_{n_1}) - I(\psi_{n_1})| - |I(\psi_{n_1}) - I(\delta)| + I(\delta) \\ &\geqslant I(\delta) - \varepsilon \quad . \end{aligned}$$

Le fait que $\psi_{n_1} \leqslant \delta$ entraîne encore $\tilde{I}(\psi_{n_1}) \leqslant \tilde{I}(\delta)$, si bien que

$$\tilde{I}(\delta) \geqslant \tilde{I}(\psi_{n_1}) \geqslant I(\delta) - \varepsilon \quad .$$

Au total, pour des formules de degré $p \geqslant \sup(p_o, p_1)$, on a

$$I(\delta) - \varepsilon \leqslant \tilde{I}(\psi_{n_1}) \leqslant \tilde{I}(\delta) \leqslant \tilde{I}(\varphi_{n_o}) \leqslant I(\delta) + \varepsilon \quad ,$$

soit

$$|\tilde{I}(\delta) - I(\delta)| \leqslant \varepsilon \quad .$$

b) La généralisation aux fonctions étagées est immédiate. En effet, si

$$f = \sum_{i=1}^{K} \alpha_i \, \delta_i(x) \quad ,$$

les δ_i étant des fonctions caractéristiques de semi-intervalles, il suffit de trouver pour chacun de ceux-ci le degré p_i à partir duquel

$$|\tilde{I}(\delta_i) - I(\delta_i)| \leqslant \frac{\varepsilon}{\sum_{i=1}^{K} |\alpha_i|} \quad .$$

Alors, toute formule de degré $p \geqslant \sup(p_1,\ldots,p_K)$ donne

$$|\Delta I(f)| \leqslant \sum_{i=1}^{K} \alpha_i |\Delta I(\delta_i)| \leqslant \varepsilon \quad .$$

c) Soit à présent f intégrable au sens de Riemann. Pour $\varepsilon > 0$ donné, il existe deux fonctions étagées φ et ψ telles que

$$\varphi \leqslant f \leqslant \psi \quad , \qquad I(\varphi - \psi) \leqslant \frac{\varepsilon}{2}$$

et, vu la positivité de l'intégrale,

$$I(\varphi) \leqslant I(f) \leqslant I(\psi) \quad ,$$

ce qui entraîne

$$|I(f) - I(\varphi)| \leqslant \frac{\varepsilon}{2} \quad , \qquad |I(f) - I(\psi)| \leqslant \frac{\varepsilon}{2} \quad .$$

Les fonctions φ et ψ étant étagées, il résulte du point b) ci-dessus que l'on peut trouver un degré p à partir duquel on a simultanément

$$|I(\varphi) - \tilde{I}(\varphi)| \leqslant \frac{\varepsilon}{2} \quad , \quad |I(\psi) - \tilde{I}(\psi)| \leqslant \frac{\varepsilon}{2} \quad .$$

Dès lors, comme la positivité des poids entraîne

$$\tilde{I}(\varphi) \leqslant \tilde{I}(f) \leqslant \tilde{I}(\psi) \quad ,$$

on obtient

$$\tilde{I}(f) \leqslant \tilde{I}(\psi) \leqslant I(\psi) + \frac{\varepsilon}{2} \leqslant I(f) + \varepsilon$$

et

$$\tilde{I}(f) \geqslant \tilde{I}(\varphi) \geqslant I(\varphi) - \frac{\varepsilon}{2} \geqslant I(f) - \varepsilon \quad ,$$

soit

$$I(f) - \varepsilon \leqslant \tilde{I}(f) \leqslant I(f) + \varepsilon \quad .$$

On a donc démontré le théorème suivant: Soit une famille de formules d'intégration approchée sur le compact $[a, b]$, de degrés croissant indéfiniment, et à poids positifs. Alors, les intégrales approchées de toute fonction f intégrable au sens de Riemann (pour la densité considérée) convergent vers l'intégrale de f.

7.6 - Peut-on étendre ce résultat aux fonctions intégrables au sens de Lebesgue? A nouveau, la réponse est négative. Pour le voir, donnons-nous une famille de formules à poids positifs de degrés croissant indéfiniment. Soit E_{n+1} l'ensemble des points d'intégration de la formule à (n + 1) points. Cet ensemble est de mesure nulle,

de même que

$$E = \bigcup_{n=0}^{\infty} E_{n+1}$$

union dénombrable d'ensembles de mesure nulle. Soit alors la fonction f définie sur (a, b) par

$$\begin{cases} f(x) = 1 & \text{si } x \in E \\ f(x) = 0 & \text{sinon} \end{cases}$$

Cette fonction est intégrable au sens de Lebesgue, car nulle presque partout, et I(f) = 0. Mais par toute formule approchée, quel que soit son degré, on trouvera

$$I(f) = \sum_i H_i \cdot 1 = I(1) \neq 0 ,$$

ce qui montre que $\tilde{I}(f)$ ne converge pas vers I(f).

8. FORMULES DE NEWTON-COTES

8.1 - Soit w(x) = 1 sur un intervalle (a, b) borné. Les points d'intégration sont choisis équidistants. On distingue les formules fermées, pour lesquelles

$$x_o = a,\ x_1 = a + e,\ x_2 = a + 2e,\ \ldots,\ x_n = a + ne = b ,$$

$$e = \frac{b - a}{n} = \frac{h}{n}$$

et les formules ouvertes, où

$$x_o = a + e,\ x_1 = a + 2e,\ \ldots,\ x_n = a + (n - 1)e,\ b = a + (n + 2)e,$$

$$e = \frac{b - a}{n + 2} = \frac{h}{n + 2} .$$

(figure 7)

Il s'agit de formules symétriques. Dès lors, les formules de Newton-Cotes à (n+1) points sont au moins de degré n si (n+1) est pair, et de degré (n+1) si ce nombre est impair. On peut montrer l'existence d'une constante K comme introduite en section 5.

8.2 - Calcul de l'erreur pour (n+1) impair

Prenant x_{n+1} comme centre de l'intervalle, on a

$$\Delta I(f) = \int_a^b \Pi_{n+1}(x).(x - c).\ f(x_o, \ldots, x_n, c, x)\, dx$$

Malheureusement, sauf cas particuliers, la fonction $(x - c)\Pi_{n+1}(x)$ n'est pas de signe constant. Cependant, comme

$$f(x_o, \ldots, x_n, c, x) = \frac{f(x_o, \ldots, x_n, x) - f(x_o, \ldots, x_n, c)}{x - c} ,$$

on a encore

$$\Delta I(f) = \int_a^b \Pi_{n+1}(x) f(x_o,\ldots,x_n,x)\,dx - f(x_o,\ldots,x_n,c)\int_a^b \Pi_{n+1}(x)\,dx\ ,$$

et le dernier terme est nul, car la fonctionn Π_{n+1} est antisymétrique par rapport au centre de l'intervalle. Considérant la fonction

$$\varphi(x) = \int_a^b \Pi_{n+1}(t)\,dt,$$

on a évidemment

$$\varphi(x) = 0 \quad , \quad \varphi(b) = 0.$$

Bien plus, cette fonction est <u>de signe constant</u>. Pour s'en assurer, on n notera d'abord que

$$\varphi(a+y) = \int_a^b \Pi_{n+1}(t)\,dt = \int_b^{b-y} \Pi_{n+1}(t)\,dt = \varphi(b-y)\ ,$$

si bien qu'il suffit de montrer que la fonction φ n'a pas de zéro entre a et c. Pour y arriver, notons

$$\mathcal{J}_k = \int_{x_k}^{x_{k+1}} \Pi_{n+1}(x)\,dx,$$

et, dans le cas des formules ouvertes,

$$\mathcal{J}_{-1} = \int_a^{x_o} \Pi_{n+1}(x)\,dx. \quad \text{(fig. 8)}$$

Leurs signes sont alternés. Montrons que l'on a toujours

$$|\mathcal{J}_{k+1}| < |\mathcal{J}_k|.$$

On a en effet, en posant $x = y + e$,

$$\mathcal{J}_{k+1} = \int_{x_{k+1}}^{x_{k+2}} (x-x_o)\ldots(x-x_n)\,dx = \int_{x_k}^{x_{k+1}} (y-x_o+e)(y-x_1+e)\ldots(y-x_n+e)\,dy$$

$$= \int_{x_k}^{x_{k+1}} \frac{y-x_o+e}{y-x_n}\,(y-x_o)\ldots(y-x_n)\,dy = \int_{x_k}^{x_{k+1}} \frac{y-x_o+e}{y-x_n}\,\Pi_{n+1}(y)\,dy\ .$$

Comme, entre x_k et x_{k+1}, $\Pi_{n+1}(y)$ garde un signe constant, le théorème de la moyenne permet d'affirmer que

$$\mathcal{J}_{k+1} = \frac{x_k+\theta h-x_o+e}{x_k+\theta h-x_k}\int_{x_k}^{x_{k+1}} \Pi_{n+1}(y)\,dy, \qquad 0 \leq \theta \leq 1\ .$$

Or, le facteur

$$\frac{x_k+\theta h-x_o}{x_k+\theta h-x_k} = \frac{x_k+\theta h-x_o+e}{x_n-x_k-\theta h}$$

est inférieur à 1 pour

$$x_k + \theta h - x_o + e < x_n - x_k - \theta h,$$

soit

$$2(x_k + \theta h) < x_n + x_o - e,$$

ce qui sera sûrement vérifié si

$$x_{k+1} \leqslant \frac{x_n + x_o}{2} - e = c - e,$$

soit si $x_{k+2} \leqslant c$. Ainsi, aux x_k, la fonction φ passe par des extrema qui valent

$$(\mathcal{I}_{-1}),\quad \mathcal{I}_{-1} + \mathcal{I}_o,\quad \mathcal{I}_{-1} + \mathcal{I}_o + \mathcal{I}_1,\ \cdots,\quad \mathcal{I}_{-1} + \mathcal{I}_o + \cdots + \mathcal{I}_{\frac{n}{2}-1}$$

et sont tous de même signe, donc φ ne peut s'annuler.

Cela étant, on a directement

$$I(f) = \int_a^b \Pi_{n+1}(x) f(x_o, \ldots, x_n, x)\,dt = \left[\varphi(x) f(x_o, \ldots, x_n, x)\right]_a^b$$

$$- \int_a^b \varphi(x) \frac{d}{dx} f(x_o, \ldots, x_n, x)\,dx$$

et le terme intégré est nul, car $\varphi(a) = \varphi(b) = 0$. Comme la fonction φ est de signe constant, il existe un point $\eta \in (a, b)$ où

$$\Delta I(f) = -\left[\frac{d}{dx} f(x_o, \ldots, x_n, x)\right]_\eta \int_a^b \varphi(x)\,dx$$

soit

$$\Delta I(f) = -f(x_o, \ldots, x_n, \eta, \eta) \int_a^b \varphi(x)\,dx$$

ou encore, un point $\xi \in]a, b[$ où

$$I(f) = \frac{f^{(n+2)}(\xi)}{(n+2)!} \int_a^b \varphi(x)\,dx.$$

D'après la théorie générale de l'erreur d'intégration, on a encore

$$\boxed{\Delta I(f) = \frac{f^{(n+2)}(\xi)}{(n+2)!} \int_a^b \Pi(x)(x - x_{n+1})\,dx},$$

x_{n+1} étant choisi arbitrairement dans $[a, b]$. En fait, <u>on peut même choisir</u> x_{n+1} <u>hors de</u> (a, b), <u>car</u>

$$\int_a^b x_{n+1} \Pi(x)\,dx = 0.$$

8.3 - Calcul de l'erreur pour (n+1) pair

Le raisonnement ci-dessus ne tient plus. Mais on peut écrire

$$\Delta I(f) = \Delta I_1(f) + \Delta I_2(f) ,$$

avec

$$\left.\begin{aligned} \Delta I_1(f) &= \int_a^z \Pi(x) f(x_o,\ldots,x_n,x)dx \\ \Delta I_2(f) &= \int_z^b \Pi(x) f(x_o,\ldots,x_n,x)dx \end{aligned}\right\} \quad z = \begin{cases} x_{n-1} & \text{(formule fermée)} \\ x_n & \text{(formule ouverte)} \end{cases}$$

On peut transformer la première intégrale en notant que

$$f(x_o,\ldots,x_{n-1},x_n,x) = \frac{f(x_o,\ldots,x_{n-1},x) - f(x_o,\ldots,x_{n-1},x_n)}{x - x_n} ,$$

d'où

$$\Delta I_1(f) = \int_a^z (x-x_o)\ldots(x-x_{n-1})\, f(x_o,\ldots,x_{n-1},x)dx$$

$$- f(x_o,\ldots,x_{n-1},x_n) \int_a^z (x-x_o)\ldots(x-x_{n-1})dx.$$

Comme le polynôme $(x-x_o)\ldots(x-x_{n-1})$ est antisymétrique sur $[a, z]$, le dernier terme s'annule. Quant au premier, il s'identifie avec l'erreur d'intégration sur $[a, z]$, pour le support$(x_o,\ldots,x_{n-1})$. Cette erreur vaut (en prenant le point supplémentaire x_n)

$$\Delta I_1(f) = \frac{f^{(n+1)}(\xi_1)}{(n+1)!} \int_a^z (x-x_o)\ldots(x-x_n)dx.$$

Cette intégrale a le même signe que le polynôme $(-(x-x_o)\ldots(x-x_{n-1}))$ au voisinage de a.

D'autre part, dans $[z, b]$, le signe de $\Pi(x)$ est constant. C'est le signe de $(x-x_o)\ldots(x-x_n)$ au voisinage de x=b. C'est aussi celui du même polynôme au voisinage de x=a (fig. 9). (Il est symétrique et, comme, en $x = a + \varepsilon$,

$$(x-x_o)\ldots(x-x_{n-1}) = \frac{(x-x_o)\ldots(x-x_n)}{x - x_n} ,$$

le dénominateur étant négatif, c'est encore le signe de

$$(-(x-x_o)\ldots(x-x_{n-1})).$$

On a donc

$$\Delta I_2(f) = \frac{f^{(n+1)}(\xi_2)}{(n+1)!} \int_z^b (x-x_o)\ldots(x-x_n)dx.$$

Comme les intégrales apparaissant dans ΔI_1 et ΔI_2 ont le même signe, il existe encore un point ξ dans $]a, b[$ tel que

$$\Delta I(f) = \Delta I_1(f) + \Delta I_2(f) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \int_a^b (x-x_o)\ldots(x-x_n)dx$$

8.4 - Les formules de Newton-Cotes sont données dans les tableaux 1 et 2. On notera que les poids ne sont pas nécessairement positifs: la convergence en degré n'est pas assurée.

9. POLYNOMES ORTHOGONAUX

9.1 - Comment construire des formules de degré élevé?

Il ressort de l'analyse de l'erreur que l'on a tout intérêt à disposer de formules polynomiales d'un degré aussi élevé que possible. Comment en obtenir? Supposons que l'on connaisse une formule de support $(x_o,\ldots,x_n)$ de degré $p>n$. Elle intègre donc exactement les polynômes

$$x^k \Pi(x) \qquad 0 \leq k \leq p-n-1$$

avec, comme d'habitude,

$$\Pi(x)=(x-x_o)\ldots(x-x_n).$$

Or, cette condition s'écrit

$$I(x^k \Pi(x)) = I(x^k \Pi(x)) = \sum_{i=0}^{n} H_i \, x_i^k \Pi(x_i) = 0 \ .$$

Par conséquent, on devra nécessairement avoir

$$I(x^k \Pi(x)) = 0 \quad , \quad 0 \leq k \leq p-n-1.$$

En définissant le produit scalaire

$$(f,g) = I(fg),$$

dont découle la norme

$$\|f\|^2 = I(f^2) \ ,$$

ces conditions reviennent à dire que le support doit être tel que la fonction Π soit orthogonale aux fonctions $1, \ldots, x^{p-n-1}$ (et par combinaison linéaire, elle doit être orthogonale à tous les polynômes de degré inférieur ou égal à (p-n-1)).

Cette condition justifie l'étude des polynômes orthogonaux.

Mais avant d'aller plus loin, posons-nous la question suivante: jusqu'où peut-on élever le degré d'une formule d'intégration? On s'aperçoit aisément qu'il existe une limite infranchissable: en aucun cas, une formule à (n+1) points ne peut excéder le degré (2n+1). Supposons en effet que l'on dispose d'une formule à (n+1) points et de degré $p > 2n+1$. Alors, la fonction Π^2, de degré (2n+2), devrait être

intégrée exactement, ce qui impliquerait la relation absurde

$$0 < I(\Pi^2) = \tilde{I}(\Pi^2) = \sum_{i=0}^{n} H_i \, \Pi^2(x_i) = 0.$$

9.2 - Nous appellerons <u>polynôme orthogonal de degré n</u> (pour l'intégrale I ou pour la densité w) <u>un polynôme φ_n de degré n dont le coefficient de tête vaut 1 , et tel que</u>

$$(\varphi_n, P_k) = 0$$

<u>pour tout polynôme P_k de degré $k < n$.</u>

Pour démontrer l'existence des φ_n, écrivons

$$\varphi_n = x^n + \sum_{k=0}^{n-1} \alpha_{nk} x^k .$$

Il suffit évidemment que φ_n soit orthogonal aux fonctions 1, x, ..., x^{n-1}, ce qui s'écrit

$$(\varphi_n, 1) = (x^n, 1) + \sum_{k=0}^{n-1} \alpha_{nk}(x^k, 1) = 0$$

$$(\varphi_n, x) = (x^n, x) + \sum_{k=0}^{n-1} \alpha_{nk}(x^k, x) = 0$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$(\varphi_n, x^{n-1}) = (x^n, x^{n-1}) + \sum_{k=0}^{n-1} \alpha_{nk}(x^k, x^{n-1}) = 0$$

C'est un système linéaire, de matrice

$$A = \begin{bmatrix} (1,1) & (1,x) & \cdots\cdots & (1,x^{n-1}) \\ \vdots & & & \\ (x^{n-1},1) & \cdots\cdots & \cdots\cdots & (x^{n-1},x^{n-1}) \end{bmatrix}$$

Cette matrice symétrique <u>ne peut être singulière</u>, car, à supposer que $Ab = 0$, avec

$$b = \begin{bmatrix} 0 \\ \vdots \\ n-1 \end{bmatrix} ,$$

on obtiendrait

$$0 = b^T A b = \sum_{i,j=0}^{n-1} \beta_i \beta_j (x^i, x^j) = \| \beta_0 + \beta_1 x + \cdots \beta_{n-1} x^{n-1} \| = 0 ,$$

soit

$$\beta_0 + \beta_1 x + \cdots + \beta_{n-1} x^{n-1} = 0 \qquad \text{p.p.} ,$$

cé qui suppose $\beta_o = \beta_1 = \ldots = \beta_{n-1} = 0$. Dès lors, les coefficients α_{nk} existent et sont uniques, et il en est de même des φ_n. (En fait, la matrice A est définie positive).

9.3 - La matrice A est une matrice de GRAM, très mal conditionnée. Aussi, vaut-il mieux, pour le calcul des polynômes orthogonaux, recourir aux formules de récurrence, qui sont fondées sur le théorème suivant:

Il existe une relation de la forme

$$\varphi_n = (x - \alpha_n)\,\varphi_{n-1} - \beta_n \varphi_{n-2},$$

α_n et β_n étant des nombre réels dépendant de n.

Divisons en effet φ_n par φ_{n-1}: on trouve

$$\varphi_n = (x - \alpha_n)\,\varphi_{n-1} + P_{n-2}$$

et, pour tput $k < n-2$, on a

$$0 = (\varphi_n, x^k) = (\varphi_{n-1}, x^{k+1}) - \alpha_n(\varphi_{n-1}, x^k) + (p_{n-2}, x^k)$$

et, comme $(k+1) < (n-1)$, cela donne

$$(P_{n-2}, x^k) = 0$$

si bien que

$$P_{n-2} = -\beta_n \varphi_{n-2},$$

comme annoncé.

Pour trouver les coefficients α_n et β_n, on note que

$$0 = (\varphi_{n-1}, \varphi_n) = (x\,\varphi_{n-1}, \varphi_{n-1}) - \alpha_n \|\varphi_{n-1}\|^2 - \beta_n(\varphi_{n-1}, \varphi_{n-2}),$$

d'où

$$\boxed{\alpha_n = \frac{(x\varphi_{n-1}, \varphi_{n-1})}{\|\varphi_{n-1}\|^2}},$$

puis que

$$0 = (\varphi_{n-2}, \varphi_n) = (x\,\varphi_{n-1}, \varphi_{n-2}) - \alpha_n(\varphi_{n-1}, \varphi_{n-2}) - \beta_n \|\varphi_{n-2}\|^2,$$

d'où

$$\boxed{\beta_n = \frac{(x\varphi_{n-1}, \varphi_{n-2})}{\|\varphi_{n-2}\|^2}}.$$

Il est encore possible d'exprimer ces coefficients en termes des coefficients des polynômes . En effet, si l'on écrit

$$\varphi_n(x) = x^n + b_n x^{n-1} + \dots ,$$

on obtient, en identifiant les coefficients de degré (n-1) de la relation de récurrence:

$$b_n = b_{n-1} - \alpha_n ,$$

soit

$$\alpha_n = b_{n-1} - b_n .$$

D'autre part, on a, en vertu de l'orthogonalité,

$$(\varphi_{n-1}, x\varphi_{n-2}) = (\varphi_{n-1}, x^{n-1} + b_{n-2} x^{n-2} + \dots) = (\varphi_n, x^{n-1})$$
$$= (\varphi_{n-1}, x^{n-1} + b_{n-1} x^{n-2} + \dots) = \|\varphi_{n-1}\|^2 ,$$

d'où

$$\beta_n = \frac{\|\varphi_{n-1}\|^2}{\|\varphi_{n-2}\|^2} ,$$

soit, en définitive,

$$\boxed{\varphi_n = (x - b_{n-1} + b_n)\,\varphi_{n-1} - \frac{\|\varphi_{n-1}\|^2}{\|\varphi_{n-2}\|^2}\,\varphi_{n-2}}$$

9.4 - Zéros des polynômes orthogonaux

Voici une propriété de grande importance pour l'intégration numérique: Le polynôme orthogonal φ_n possède exactement n zéros simples dans l'ouvert $]a, b[$

Soient en effet $x_1, \dots, x_k$ les zéros de φ_n dans $]a, b[$. Ils peuvent a priori avoir une multiplicité $m_1, \dots, m_k$. Formons le polynôme

$$P_r(x) = \prod_{\substack{\text{zéros de} \\ \text{multiplicité} \\ \text{impaire}}} (x - x_i) ,$$

dont le degré r est évidemment égal au nombre de zéros de multiplicité impaire de φ_n. Ce polynôme a les mêmes changements de signe que φ_n. En le multipliant par $\alpha = \pm 1$, on peut donc obtenir

$$\varphi_n \cdot \alpha P_r > 0$$

presque partout dans $]a, b[$, ce qui entraîne

$$(\varphi_n, \alpha P_r) = I(\varphi_n \alpha P_r) > 0 .$$

Mais si $r < n$, on a

$$(\varphi_n, \alpha P_r) = 0 ,$$

ce qui est contradictoire. Donc, φ_n a exactement n zéros simples dans $]a, b[$.

10. PROPRIETES COMPLEMENTAIRES DE CERTAINS POLYNOMES ORTHOGONAUX [16]

10.1 - Densité de Rodrigues

Soit β la fonction positive définie comme suit:

$$\beta = \beta_0 + \beta_1 x + \beta_2 x^2 = \begin{cases} (x - a)(b - x) & \text{si l'intervalle est borné} \\ (x - a) & \text{sur l'intervalle } [a, \infty[\\ (b - x) & \text{sur l'intervalle }]-\infty, b] \\ 1 & \text{sur l'intervalle }]-\infty, +\infty[\end{cases}$$

Nous dirons que la densité w est de Rodrigues si les fonctions

$$F_n = \frac{1}{w} D^n(w \beta^n)$$

sont des polynômes de degré n.

Il est aisé de dégager une condition nécessaire pour qu'il en soit ainsi. En effet, on doit avoir

$$F_1 = \frac{1}{w} (w \beta' + w' \beta) ,$$

soit

$$\frac{w'}{w} \beta + \beta' = F_1$$

ou encore,

$$\frac{w'}{w} = \frac{F_1 - \beta'}{\beta} = \frac{\alpha_0 + \alpha_1 x}{\beta} = \frac{\alpha(x)}{\beta(x)} ,$$

$\alpha(x)$ étant une fonction affine.

Montrons que cette condition est également suffisante. A cette fin, observons d'abord que

$$D^0(w \beta^n) = w \beta^n P_0 ,$$

et montrons que si

$$D^k(w \beta^n) = w \beta^{n-k} P_k ,$$

on a aussi

$$D^{k+1}(x \beta^n) = w \beta^{n-k-1} P_{k+1}$$

(Dans tout ceci, P_1 représente un polynôme de degré 1). On a en effet

$$D(D^k(w \beta^n)) = D(w \beta^{n-k} P_k)$$

$$= w' \beta^{n-k} P_k + w(n - k) \beta^{n-k-1} \beta' P_k + w \beta^{n-k} P_k'$$

$$= w\beta^{n-k-1}(\alpha P_k + (n-k)\beta' P_k + \beta P_k') = w\beta^{n-k-1} P_{k+1}.$$

Par écurrence, on obtient

$$D^n(w\beta^n) = w P_n ,$$

comme annoncé. Par conséquent, <u>la condition nécessaire et suffisante pour qu'une densité w soit de Rodrigues est qu'elle vérifie l'équation différentielle</u>

$$\frac{w'}{w} = \frac{\alpha(x)}{\beta(x)} ,$$

<u>α étant une fonction affine.</u>

10.2 - <u>Densités de Rodrigues pour a et b finis</u>

Pour a et b finis, on peut toujours se ramener à l'intervalle $[-1,+1]$. L'équation différentielle de la densité s'écrit alors

$$\frac{w'}{w} = \frac{\alpha_0 + \alpha_1 x}{(x+1)(1-x)} = \frac{-\lambda}{1-x} + \frac{\mu}{1+x} ,$$

avec

$$\alpha_0 + \alpha_1 x = -\lambda(x+1) + \mu(1-x),$$

soit

$$\alpha_0 = \mu - \lambda \quad , \quad \alpha_1 = -(\lambda+\mu)$$

et

$$\alpha(x) = (\mu-\lambda) - (\lambda+\mu)x.$$

Intégrant, on obtient

$$\ln w = \mu \ln(x+1) + \lambda \ln(1-x) + \ln C ,$$

soit

$$\boxed{w = C\,(1-x)^{\lambda}\,(x+1)^{\mu}}$$

Cette fonction n'est intégrable que si

$$\lambda > -1 , \quad \mu > -1 .$$

Les polynômes orthogonaux correspondants sont les <u>polynômes de Jacobi</u> $p^{(\lambda,\mu)}(x)$. Ils ont deux cas particuliers importants:

- Pour $\lambda = \mu = 0$, on a $w = 1$, $\alpha = 0$. On obtient les <u>polynômes de Legendre</u>.

- Pour $\lambda = \mu = -\frac{1}{2}$, on a

$$w = \frac{1}{\sqrt{1-x^2}} , \quad \alpha = -x .$$

On obtient les <u>polynômes de Tchébicheff</u>

10.3 - Densités de Rodrigues lorsque l'une des deux extrémités de l'intervalle est infinie

Dans ce cas, il est toujours possible, par un changement de variables affine, de se ramener à l'intervalle $[\,0, +\infty[$. L'équation différentielle de la densité est alors

$$\frac{w'}{w} = \frac{\alpha_o + \alpha_1 x}{x} = \frac{\alpha_o}{x} + \alpha_1 ,$$

ce qui s'intègre en

$$\ln w = \alpha_o \ln x + \alpha_1 x + \ln C ,$$

soit

$$w = C\, x^{\alpha_o}\, e^{\alpha_1 x} .$$

Cette fonction n'est intégrable que si α_1 est strictement inférieur à zéro. Moyennant un changement de variables, elle se ramène toujours à la forme

$$\boxed{w = C\, x^{\lambda}\, e^{-x}} .$$

Cette densité n'est intégrable que si $\lambda > -1$. On a

$$\alpha = \lambda - x \quad , \quad \beta = x .$$

Les polynômes orthogonaux correspondants sont les polynômes de Laguerre généralisés $l_n^{(\lambda)}(x)$. Dans le cas particulier important $\lambda = 0$, il s'agit des polynômes de Laguerre (classiques) $L_n(x)$.

10.4 - Densités de Rodrigues lorsque les deux extrémités de l'intervalle sont infinies.

La densité vérifie l'équation

$$\frac{w'}{w} = \alpha_o + \alpha_1 x ,$$

soit

$$\ln w = \alpha_o x + \alpha_1 \frac{x^2}{2} + C ,$$

ou encore

$$w = e^{\,C + \alpha_o x + \frac{\alpha_1 x^2}{2}}$$

Cette fonction n'est intégrable que si $\alpha_1 < 0$. En prenant pour constante C le nombre

$$C = -\frac{\alpha_o^2}{2\,|\alpha_1|} ,$$

on obtient

$$w = \exp\left(-|\alpha_1| \frac{(x - x_o)^2}{2}\right) ,$$

avec

$$x_o = \alpha_o / |\alpha_1| .$$

Un changement de variables affine permet toujours de se ramener à la forme

$$\boxed{w = e^{-x^2}}$$

Il vient alors

$$\alpha = -2x \quad , \quad \beta = 1$$

et les polynômes orthogonaux correspondants sont les polynômes d'Hermite.

10.5 - On le voit, les densités de Rodrigues sont très particulières. Elles possèdent toutre la propriété suivante: soient w une densité de Rodrigues sur l'intervalle (a,b), et P_r un polynôme de degré r. On a

$$[w(x)\,\beta(x)\,P_r(x)]_a^b = 0 \ .$$

En effet, si a et b sont finis, on a

$$w(x)\,\beta(x) = (b-x)^{\lambda+1}\,(x-a)^{\mu+1} \ ,$$

avec $(\lambda+1)>0$ et $(\mu+1)>0$;
pour $a = 0$ et $b = +\infty$, on a

$$w(x)\,\beta(x) = x^{\lambda+1}\,e^{-x}$$

et enfin, pour $a = -\infty$ et $b = +\infty$, on a

$$w(x)\,\beta(x) = e^{-x^2} \ .$$

10.6 - Dans les mêmes conditions, pour $k<n$, on a

$$[D^k(w\,\beta^n)\,P_r]_a^b = 0$$

Ceci résulte de la propriété précédente et du fait que

$$D^k(w\,\beta^n)P_r = w\,\beta^{n-k}\,P_{r+k} = w\,\beta\,P_{r+n-1} \ .$$

10.7 - L'intérêt que portent les physiciens aux polynômes orthogonaux résulte du fait qu'ils sont solutions de certaines équations différentielles linéaires du second ordre. Pour le démontrer, observons d'abord que

$$(\beta\,w\,\varphi_n')' = \beta' w\,\varphi_n' + \beta\,w'\,\varphi_n' + \beta\,w\,\varphi_n'' = w((\alpha+\beta')\,\varphi_n' + \varphi_n'') \qquad (a)$$

et calculons l'intégrale

$$J_k = \int_a^b (\beta\,w\,\varphi_n')'\,x^k\,dx \quad , \quad 0 \leq k < n \ .$$

Une intégration par parties donne

$$J_k = [\,w\,\varphi_n'\,x^k]_a^b - k\int_a^b \beta\,w\,\varphi_n'\,x^{k-1}\,dx \ ,$$

et le terme intégré disparaît, car $\varphi_n'\,x^k$ est un polynôme. Une seconde

intégration par parties conduit à l'expression

$$J_k = - k \left[\beta w \varphi_n x^{k-1}\right]_a^b + k \int_a^b \varphi_n (\beta w x^{k-1})' dx ,$$

dont le terme intégré est également nul. Par ailleurs, on calcule comme ci-dessus

$$(\beta w x^{k-1})' = w((\alpha + \beta') x^{k-1} + \beta(k-1) x^{k-2}) = w P_k ,$$

où P_k est un polynôme de degré k. On a donc

$$J_k = k \int_a^b w \varphi_n P_k dx = 0 \quad , \quad 0 \leq k < n.$$

Mais ceci revient à dire que le polynôme de degré n

$$(\alpha + \beta') \varphi_n' + \beta \varphi_n''$$

est orthogonal à x^k , $0 \leq k < n$: il doit donc s'identifier à un multiple de φ_n. On a donc

$$\boxed{\beta \varphi_n'' + (\alpha + \beta') \varphi_n' = - \gamma_n \varphi_n}$$

C'est l'équation différentielle des polynômes orthogonaux par rapport à une densité de Rodrigues. On peut encore, en la multipliant par w, la mettre sous la forme auto-adjointe que voici:

$$\boxed{(\beta w \varphi_n')' = - \gamma_n w \varphi_n}$$

A noter que les coefficients de φ_n'' et de φ_n' sont indépendants de n. Quant à γ_n , on peut le calculer en identifiant les coefficients de tête des deux membres de l'équation, ce qui donne

$$\beta_2 n(n-1) + (\alpha_1 + 2\beta_2) n = - \gamma_n \quad ,$$

soit

$$\boxed{\gamma_n = - n (\alpha_1 + (n+1) \beta_2)}$$

Nous montrerons plus loin que γ_n est toujours positif.

10.8 - La véritable motivation de l'introduction des densités de Rodrigues est la formule de Rodrigues: si w est une densité de Rodrigues, le n^e polynôme orthogonal est donné par

$$\boxed{\varphi_n = A_n \cdot \frac{1}{w} D^n (w \beta^n)} \quad ,$$

A_n étant un coefficient destiné à obtenir un coefficient de tête égal à 1.

(Cette formule a été démontrée par Rodrigues en 1814 pour les polynômes de Legendre).

Tout d'abord, l'expression proposée est un polynôme par définition des densités de Rodrigues. Il reste donc à démontrer les relations d'orthogonalité. Pour $k < n$, on a

$$(\varphi_n, x^k) = A_n \int_a^b w \frac{1}{w} D^n(w\beta^n)\, x^k\, dx$$

$$= A_n [x^k D^{n-1}(w\beta^n)]_a^b - A_n\, k \int_a^b x^{k-1} D^{n-1}(w\beta^n)\, dx,$$

et le terme intégré s'annule en vertu de la propriété 10.6. Poursuivant les intégrations par parties, on obtient de proche en proche

$$(\varphi_n, x^k) = A_n\, k! \int_a^b D^{n-k}(w\beta^n)\, dx = A_n\, k! \,[D^{n-k}(w\beta^n)]_a^b = 0 ,$$

pour autant que k soit strictement inférieur à n.

10.9 - Calcul de A_n et du deuxième coefficient b_n du polynôme φ_n.

Dérivant la relation

$$D^{k-1}(w\beta^n) = w\,\beta^{n-k+1}\,P_{k-1} ,$$

on obtient

$$D^k(w\beta^n) = w'\,\beta^{n-k+1}P_{k-1} + w\,(n-k+1)\,\beta^{n-k}\,\beta'\,P_{k-1} + w\,\beta^{n-k+1}\,P'_{k-1}$$

$$= w\,\beta^{n-k}((\alpha + (n-k+1)\,\beta')P_{k-1} + \beta P'_{k-1}) = w\,\beta^{n-k}\,P_k ,$$

d'où

$$P_k = (\alpha + (n-k+1)\,\beta')\,P_{k-1} + \beta P'_{k-1} . \qquad (a)$$

Utilisant la notation

$$P_k = \mathcal{A}_k\, x^k + \mathcal{B}_k\, x^{k-1} + \dots ,$$

on peut ainsi obtenir P_n par récurrence. Alors,

$$\varphi_n = A_n\, P_n = A_n(\mathcal{A}_n\, x^n + \mathcal{B}_n\, x^{n-1} + \dots) ,$$

d'où

$$A_n = \frac{1}{\mathcal{A}_n} \quad , \quad b_n = \frac{\mathcal{B}_n}{\mathcal{A}_n} .$$

Identifions donc les coefficients de x^k et x^{k-1} dans la relation de récurrence (a). On arrive à

$$\begin{cases} \mathcal{A}_k = (\alpha_1 + (2n-k+1)\,\beta_2)\,\mathcal{A}_{k-1} , & \mathcal{A}_0 = 1 \\ \mathcal{B}_k = (\alpha_1 + (2n-k)\,\beta_2)\,\mathcal{B}_{k-1} + (\alpha_0 + n\,\beta_1)\,\mathcal{A}_{k-1} , & \mathcal{B}_0 = 0 \end{cases}$$

Pour les $\mathcal{A}_k$, il est immédiat que

$$\mathcal{A}_n = (\alpha_1+(n+1)\,\beta_2)(\alpha_1+(n+2)\,\beta_2)\dots(\alpha_1+2n) = \prod_{j=1}^{2n}(\alpha_1 + j\,\beta_2) ,$$

c'est-à-dire

$$\boxed{A_n = \frac{1}{\prod_{j=n+1}^{2n} (\alpha_1 + j\beta_2)}}$$

En ce qui concerne les $\mathcal{B}_k$, montrons que l'on a

$$\mathcal{B}_k = k \frac{\alpha_o + n\beta_1}{\alpha_1 + 2n\beta_2} \mathcal{A}_k \ .$$

Pour k = 0, c'est immédiat, et nous allons montrer que si cette relation est vraie à l'ordre (k-1), elle l'est encore à l'ordre k. On a en effet

$$\mathcal{B}_k = ((\alpha_1+(2n-k)\beta_2)(k-1) \frac{\alpha_o+n\beta_1}{\alpha_1+n\beta_2} + (\alpha_o+n\beta_1)) \mathcal{A}_{k-1}$$

$$= \frac{\alpha_o+n\beta_1}{1+2n\ 2}(\alpha_1(k-1) + (2n-k-2n-k^2+k)\beta_2 + \alpha_1 + 2n\beta_2) \mathcal{A}_{k-1}$$

$$= \frac{\alpha_o+n\beta_1}{\alpha_1+2n\beta_2}(k\alpha_1 + k(2n-k+1)\beta_2) \mathcal{A}_{k-1} = k \frac{\alpha_o + n\beta_1}{\alpha_1 + 2n\beta_2} \mathcal{A}_k \ .$$

En conséquence,

$$\mathcal{B}_n = n \frac{\alpha_o + n\beta_1}{\alpha_1 + 2n\beta_2} \mathcal{A}_n$$

et

$$\boxed{b_n = n \frac{\alpha_o + n\beta_1}{\alpha_1 + 2n\beta_2}}$$

10.10 - <u>La norme des polynômes orthogonaux par rapport à une densité de Rodrigues est donnée par</u>

$$\boxed{\| \varphi_n \|^2 = (-1)^n\, n!\, A_n \int_a^b w \beta^n \, dx}$$

En effet,

$$\| \varphi_n \|^2 = \int_a^b w \varphi_n^2 \, dx = A_n \int_a^b \varphi_n D^n(w\beta^n) \, dx$$

$$= A_n [\varphi_n D^{n-1}(w\beta^n)]_a^b - A_n \int_a^b \varphi_n' D^{n-1}(w\beta^n) \, dx \ ,$$

et nous savons que le terme intégré s'annule. Continuant à intégrer par parties, on obtient

$$\| \varphi_n \|^2 = (-1)^n \int_a^b \varphi_n^{(n)} w \beta^n \, dx = (-1)^n\, n!\, A_n \int_a^b w \beta^n \, dx,$$

comme annoncé.

10.11 - Les dérivées des polynômes orthogonaux par rapport à une densité de Rodrigues sont également des polynômes orthogonaux: <u>les polynômes</u> $\psi_{n-1} = \frac{1}{n}\varphi_n'$ <u>sont les polynômes orthogonaux par rapport à la densité</u> $w\beta$, <u>qui est elle-même de Rodrigues</u>.

En effet, on a

$$(\beta\varphi_n', x^k) = \int_a^b w\beta\varphi_n' x^k \, dx$$

$$= [w\beta\varphi_n x^k]_a^b - \int_a^b \varphi_n (w'\beta x^k + w\beta' x^k + w\beta k x^{k-1}) \, dx$$

$$= -\int_a^b w\varphi_n ((\alpha+\beta') x^k + \beta k x^{k-1}) \, dx = -\int_a^b w\varphi_n P_{k+1} \, dx,$$

et cette intégrale est nulle chaque fois que $k < n-1$.

Le coefficient $1/n$ est évidemment destiné à maintenir le coefficient de tête de ψ_{n-1} égal à 1.

Enfin, la densité w vérifie

$$\frac{(w\beta)'}{w\beta} = \frac{w'\beta + w\beta'}{w\beta} = \frac{w'}{w} + \frac{\beta'}{\beta} = \frac{\alpha+\beta'}{\beta}$$

et correspond donc à $\beta^* = \beta$, $\alpha^* = \alpha + \beta'$.

10.12 - <u>Norme du polynôme</u> ψ_{n-1}

La norme des ψ_{n-1} dans la métrique donnée par la densité $w\beta$ est

$$\|\psi_{n-1}\|_*^2 = \int_a^b w\beta\psi_{n-1}^2 \, dx = \frac{1}{n}\int_a^b w\beta\varphi_n'\psi_{n-1} \, dx = \frac{1}{n}\int_a^b w\beta\varphi_n' x^{n-1} \, dx$$

$$= \frac{1}{n}\left[w\beta\varphi_n' \frac{x^n}{n}\right]_a^b - \frac{1}{n^2}\int_a^b (w\beta\varphi_n') x^n \, dx.$$

Tenant compte de l'équation différentielle des φ_n (sous forme autoadjointe), on obtient

$$\boxed{\|\psi_{n-1}\|_*^2 = \frac{\gamma_n}{n^2}\int_a^b w\varphi_n x^n \, dx = \frac{\gamma_n}{n^2}\|\varphi_n\|^2}$$

Ce résultat admet deux corollaires:

a) Le coefficient γ_n est nécessairement <u>positif</u>.

b) Par comparaison avec ler résultat obtenu en 10.10, on obtient, si A_{n-1}^* est le coefficient de la formule de Rodrigues pour ψ_{n-1} ,

$$(-1)^n n! A_n \cdot \frac{\gamma_n}{n^2} = (-1)^{n-1}(n-1)! A_{n-1}^* ,$$

soit

$$A_{n-1}^* = -\frac{\gamma_n}{n} A_n .$$

10.13 - Récurrence relative aux dérivées (pour une densité de Rodrigues)

Divisons $\beta\varphi_n'$ par φ_n : on obtient

$$\beta\varphi_n' = (\mathcal{A}_n x + \mathcal{B}_n)\,\varphi_n + R_{n-1} \ ,$$

R_{n-1} étant le reste. Pour $k < n-1$, on a

$$0 = (\beta\varphi_n', x^k) = \mathcal{A}_n(\varphi_n, x^{k+1}) + \mathcal{B}_n(\varphi_n, x^k) + (R_{n-1}, x^k) = (R_{n-1}, x^k) \ ,$$

ce qui implique que R_{n-1} est un multiple de φ_{n-1}. On a donc

$$\beta\varphi_n' = (\mathcal{A}_n x + \mathcal{B}_n)\,\varphi_n + \mathcal{C}_n \varphi_{n-1} \ .$$

Il nous reste à calculer les coefficients $\mathcal{A}_n$, $\mathcal{B}_n$ et $\mathcal{C}_n$. En identifiant les coefficients de x^{n+1} dans l'équation ci-dessus, on obtient immédiatement

$$\mathcal{A}_n = n\,\beta_2 \ .$$

A la puissance n, on trouve

$$\beta_2(n-1)b_n + \beta_1 n = \mathcal{A}_n b_n + \mathcal{B}_n = n\,\beta_2 b_n + \mathcal{B}_n ,$$

d'où

$$\mathcal{B}_n = n\,\beta_1 - \beta_2 b_n \ .$$

Enfin, pour obtenir $\mathcal{C}_n$, on note que

$$(\beta\varphi_n' , x^{n-1}) = \mathcal{A}_n(x^n , \varphi_n) + \mathcal{C}_n(x^{n-1}, \varphi_{n-1}) = \mathcal{A}_n \| \varphi_n \|^2 + \mathcal{B}_n \| \varphi_{n-1} \|^2$$

On sait par ailleurs que

$$(\beta\varphi_n', x^{n-1}) = n(\beta\psi_{n-1}, x^{n-1}) = n\,(\beta\psi_{n-1}, \psi_{n-1}) = \frac{\gamma_n}{n} \| \varphi_n \|^2 \ .$$

Par conséquent,

$$\mathcal{C}_n = -\frac{\| \varphi_n \|^2}{\| \varphi_{n-1} \|^2}\,(\mathcal{A}_n - \frac{\gamma_n}{n}) = -\frac{\| \varphi_n \|^2}{\| \varphi_{n-1} \|^2}(\alpha_1 + (2n+1)\,\beta_2).$$

Finalement, on obtient la formule

$$\boxed{\beta\varphi_n' = (n\,\beta_2 x + n\,\beta_1 - \beta_2 b_n)\,\varphi_n - (\alpha_1 + (2n+1)\,\beta_2)\,\frac{\| \varphi_n \|^2}{\| \varphi_{n-1} \|^2}\,\varphi_{n-1}}$$

11. LES POLYNOMES ORTHOGONAUX LES PLUS COURANTS

Nous allons rapidement énumérer les propriétés essentielles de tous les polynômes orthogonaux pour une densité de Rodrigues. On trouvera de nombreux renseignements complémentaires dans les tables de fonctions d'Abramowitz et Stegun [5] . Nous adoptons systématiquement un coefficient de tête égal à l'unité. Il convient de noter que ce choix n'est pas nécessairement conforme à la tradition. C'est pourquoi nous surmonterons le symbole du polynôme d'un accent circonflexe.

11.1 - Polynômes de Jacobi $\hat{p}_n^{(\lambda,\mu)}$ (Jacobi, 1859)

$a = + 1$; $b = 1$; $w = (1 - x)^{\lambda} (1 + x)^{\mu}$, $\lambda > -1$, $\mu > -1$

$\beta_0 = 1$; $\beta_1 = 0$; $\beta_2 = - 1$

$\alpha_0 = \mu - \lambda$; $\alpha_1 = - (\lambda + \mu)$

Equation de Rodrigues: $\varphi_n = \frac{1}{A_n} D^n((1 - x)^{n+\lambda} (1 + x)^{n+\mu})$

$$A_n = \frac{1}{\prod_{j=n+1}^{2n} (-(\lambda+\mu) - j)} = (-1)^n \frac{\Gamma(n + \lambda + \mu + 1)}{\Gamma(2n + \lambda + \mu + 1)}$$

$$b_n = n \frac{\lambda - \mu}{2n + \lambda + \mu}$$

Equation différentielle

$$\gamma_n = n(n + \lambda + \mu + 1)$$

$$\alpha + \beta' = (\mu - \lambda) - (\lambda + \mu + 2)$$

$$(1 - x^2) \varphi_n'' + ((\mu - \lambda) - (\lambda + \mu + 2)x) \varphi_n' + n(n + \lambda + \mu + 1) \varphi_n = 0$$

Norme

$$\| \varphi_n \|^2 = n! \frac{\Gamma(n + \lambda + \mu + 1)}{\Gamma(2n + \lambda + \mu + 1)} \int_{-1}^{+1} (1-x)^{n+\lambda} (1+x)^{n+\mu} dx.$$

Posant

$x = 2y - 1$, $0 \leq y \leq 1$,

on transforme l'intégrale en

$$2^{2n + \lambda + \mu + 1} \int_0^1 y^{n+\mu} (1-y)^{n+\lambda} dy = 2^{2n + \lambda + \mu + 1} B(n + \lambda + 1, n + \mu + 1)$$

$$= 2^{2n + \lambda + \mu + 1} \frac{\Gamma(n + \lambda + 1) \Gamma(n + \mu + 1)}{\Gamma(2n + \lambda + \mu + 2)} .$$

Il en découle

$$\| \varphi_n \|^2 = \frac{n!\ 2^{2n+\lambda+\mu+1}}{2n+\lambda+\mu+1} \cdot \frac{\Gamma(n+\lambda+\mu+1)\ \Gamma(n+\lambda+1)\ \Gamma(n+\mu+1)}{\Gamma^2(2n+\lambda+\mu+1)}$$

Relations de récurrence

$$\varphi_n = (x + B_n)\,\varphi_{n-1} + C_n \varphi_{n-2} \quad ,$$

avec

$$C_n = - \frac{2^2 (n-1)(n+\lambda+\mu-1)(n+\lambda-1)(n+\mu-1)}{(2n+\lambda+\mu-1)(2n+\lambda+\mu-2)^2(2n+\lambda+\mu-3)}$$

$$B_n = \frac{\lambda^2 - \mu^2}{(2n+\lambda+\mu)(2n+\lambda+\mu-2)}$$

Récurrence sur les dérivées

$$(1 - x^2)\,\varphi_n' = (\mathcal{A}_n x + \mathcal{B}_n)\,\varphi_n + \mathcal{C}_n \varphi_{n-1}$$

avec

$$\mathcal{A}_n = - n$$

$$\mathcal{B}_n = - n \frac{\lambda - \mu}{2n+\lambda+\mu}$$

$$\mathcal{C}_n = \frac{2^2\, n(n+\lambda+\mu)(n+\lambda)(n+\mu)}{(2n+\lambda+\mu+1)(2n+\lambda+\mu)^2} \quad .$$

11.2 - Polynômes de Legendre $\hat{P}_n$ (Legendre, 1785)

Les polynômes de Jacobi particuliers qui correspondent à $\lambda = \mu = 0$ sont appelés polynômes de Legendre. Classiquement, on utilise le polynôme

$$P_n = \frac{(2n)!}{2^n (n!)^2}\ \hat{P}_n \quad .$$

$w = 1$; $a = -1$; $b = 1$

Formule de Rodrigues: $\hat{P}_n = \frac{1}{A_n} D^n((1-x)^n(1+x)^n)$

$$A_n = (-1)^n\ \frac{n!}{(2n)!}$$

$$b_n = 0$$

Equation différentielle

$$\gamma_n = n(n+1)$$

$$(1-x^2)\hat{P}_n'' - 2x\,\hat{P}_n' + n(n+1)\,\hat{P}_n = 0$$

Norme

$$\|\hat{P}_n\|^2 = \frac{2^{2n+1}\,(n!)^4}{((2n)!)^2(2n+1)}$$

Relation de récurrence

$$\hat{P}_n = x\,\hat{P}'_{n-1} - \frac{(n-1)^2}{(2n-1)(2n-3)}\,\hat{P}_{n-2}$$

Récurrence sur les dérivées

$$(1-x^2)\hat{P}'_n = -\,nx\hat{P}_n + \frac{n^2}{2n-1}\,\hat{P}_{n-1}$$

11.3 - Polynômes de Tchébicheff $\hat{T}_n$ (Tchébicheff, 1859)

Les polynômes de Jacobi correspondant aux valeurs $\lambda = -\frac{1}{2}$ $\mu = -\frac{1}{2}$ sont appelés polynômes de Tchébicheff. Nous les avons déjà rencontrés dans le chapitre relatif à l'interpolation. La définition usuelle est

$$T_n = 2^{n-1}\,\hat{T}_n \quad ,$$

avec

$$T_n(\cos\theta) = \cos n\theta \,.$$

$$w = \frac{1}{\sqrt{1-x^2}} \qquad ; \quad a = -1 \quad ; \qquad b = 1$$

Formule de Rodrigues: $\hat{T}_n = \frac{1}{A_n} D^n((1-x^2)^{n-\frac{1}{2}}(1+x^2)^{n-\frac{1}{2}})$

$$A_n = (-1)^n\,\frac{(n-1)!}{(2n-1)!}$$

$$b_n = 0$$

Equation différentielle

$$\gamma_n = n^2$$

$$(1-x^2)\,\hat{T}''_n - x\,\hat{T}'_n + n^2\,\hat{T}_n = 0$$

Norme

$$\|\hat{T}_n\|^2 = \frac{n!\,2^n}{2n} \cdot \frac{(n-1)!\;\Gamma^2(n+\frac{1}{2})}{((2n-1)!)^2}$$

Notant que

$$\Gamma(n+\tfrac{1}{2}) = (n-\tfrac{1}{2})(n-\tfrac{3}{2})\ldots\,\frac{1}{2}\,\Gamma(\tfrac{1}{2}) = \frac{(2n-1)(2n-3)\ldots 2}{2^n}\,\sqrt{\pi} = \frac{(2n)!}{2^{2n}\,n!}\,\sqrt{\pi} \quad ,$$

on obtient

$$\|\hat{T}_n\|^2 = \frac{\pi}{2^{2n-1}}$$

Relation de récurrence

$$\hat{T}_n = x\,\hat{T}_{n-1} - \frac{1}{4}\,\hat{T}_{n-2}$$

Récurrence sur les dérivées

$$(1-x^2)\,\hat{T}'_n = -\,n\,x\,\hat{T}_n + \frac{n}{2}\,\hat{T}_{n-1}$$

11.4 - Polynômes de Laguerre généralisés $\hat{L}_n^{(\lambda)}$

(Tchébicheff, 1859; Laguerre, 1879)

Pour $\lambda = 0$, il s'agit des polynômes de Laguerre classiques, notés $\hat{L}_n$.
On utilise souvent les polynômes

$$L_n = \frac{(-1)^n}{n!} \hat{L}_n$$

$$a = 0 \quad ; \quad b = \infty \; ; \qquad w = x^{\lambda} e^{-x} \quad , \quad \lambda > -1$$

$$\beta_0 = 0 \quad ; \quad \beta_1 = 1 \quad ; \quad \beta_2 = 0$$

$$\alpha_0 = \lambda \quad ; \quad \alpha_1 = -1$$

Formule de Rodrigues : $\varphi_n = \frac{1}{A_n} D^n (x^{n+\lambda} e^{-x})$

$$A_n = \frac{1}{\prod_{j=n+1}^{2n} (-1)} = (-1)^n$$

$$b_n = -n(n+\lambda)$$

Equation différentielle

$$\gamma_n = n$$

$$x \varphi_n'' + (\lambda + 1 - x) \varphi_n' + n \varphi_n = 0$$

Norme

$$\| \varphi_n \|^2 = n! \int_0^{\infty} x^{n+\lambda} e^{-x} \, dx = n! \, \Gamma(n + \lambda + 1)$$

Relation de récurrence

$$\varphi_n = (x - 2n + 1 - \lambda) \varphi_{n-1} - (n - 1)(n + \lambda - 1) \varphi_{n-2}$$

Récurrence sur les dérivées

$$x \varphi_n' = n \varphi_n + n(n + \lambda) \varphi_{n-1}$$

11.5 - Polynômes d'Hermite $\hat{H}_n$

(Tchébicheff, 1859; Hermite, 1864)

La définition classique est $H_n = 2^n \hat{H}_n$.

$$a = -\infty \quad ; \quad b = \infty \quad ; \quad w = e^{-x^2}$$

$$\beta_0 = 1 \quad ; \quad \beta_1 = 0 \quad ; \quad \beta_2 = 0$$

$$\alpha_0 = 0 \quad ; \quad \alpha_1 = -2$$

Formule de Rodrigues: $\hat{H}_n = A_n \, e^{x^2} \, D^n(e^{-x^2})$

$$A_n = \frac{1}{\prod_{j=n+1}^{2n} (-2)} = (-1)^n \, 2^{-n}$$

$$b_n = 0$$

Equation différentielle

$$\gamma_n = - n \, (-2) = 2n$$

$$\hat{H}_n'' - 2x \, \hat{H}_n' + 2n \, \hat{H}_n = 0$$

Norme

$$\| \hat{H}_n \|^2 = n! \, 2^{-n} \int_{-\infty}^{+\infty} e^{-x^2} \, dx = n! \, 2^{-n} \sqrt{\pi}$$ (Intégrale de Poisson)

Relations de récurrence

$$\hat{H}_n = x \, \hat{H}_{n-1} - \frac{n-1}{2} \, \hat{H}_{n-2}$$

Récurrence sur les dérivées

$$\hat{H}_n' = n \, \hat{H}_{n-1}$$

12. POLYNOMES ORTHOGONAUX ET INTEGRATION NUMERIQUE

Le lien entre les polynômes orthogonaux et l'intégration numérique est exprimé par le théorème suivant [3] :

Soit ψ_{n+1} un polynôme de degré (n+1), dont tous les zéros sont distincts et contenus dans $[a, b]$. Si ψ_{n+1} est orthogonal à tous les polynômes de degré $\leq m$, la formule polynomiale ayant pour support les (n+1) zéros de ψ_{n+1} est de degré (n + m + 1).

(Gain de (m+1) unités)

Soit en effet P_{n+m+1} un polynôme de degré (n+m+1). On peut le diviser par ψ_{n+1}, ce qui donne

$$P_{n+m+1} = \psi_{n+1} \, Q_m + R_n \, ,$$

où Q_m et R_n, quotient et reste de la division, sont respectivement de degrés m et n. Intégrant, on obtient

$$I(P_{n+m+1}) = I(\psi_{n+1} \, Q_m) + I(R_n) = I(R_n) \, ,$$

puisque

$$(\psi_{n+1}, Q_m) = 0 \, .$$

Quant au calcul numérique, il donne

$$\tilde{I}(P_{n+m+1}) = \tilde{I}(\psi_{n+1}\, Q_m) + \tilde{I}(R_n) = \tilde{I}(R_n) ,$$

car

$$\tilde{I}(\psi_{n+1}\, Q_m) = \sum_{\substack{\text{zéros} \\ \text{de } \psi_{n+1}}} H_i\, \psi_{n+1}(x_i)\, Q_m(x_i) = 0 .$$

Or, une formule polynomiale à (n+1) points est toujours de degré n au moins, donc $\tilde{I}(R_n) = I(R_n)$, ce qui achève la démonstration.

13. FORMULES DE GAUSS

Les formules de Gauss s'obtiennent en prenant pour points d'intégration les zéros du polynôme orthogonal φ_{n+1}. Le degré de ces formules est donc (2n + 1). Les points d'intégration de ces formules sont souvent appelés points de Gauss et les poids H_i correspondants, poids de Gauss.

13.1 - Montrons que les poids de Gauss sont positifs. Considérons en effet les fonctions $L_i^2(x)$, où L_i est le i^e polynôme de Lagrange sur les points de Gauss:

$$L_i(x) = \prod_{\substack{j=0 \\ \neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)}$$

Evidemment, $L_i^2(x_j) = \delta_{ij}$. Comme c'est un polynôme de degré 2n, il est intégré exactement. C'est, de plus une fonction positive, donc

$$0 < I(L_i^2) = \tilde{I}(L_i^2) = \sum_{j=0}^{n} H_j\, L_i^2(x_j) = H_i .$$

Cette propriété nous permet d'affirmer que, sur un intervalle borné, les formules de Gauss convergent aussi bien en degré que par raffinement de l'intervalle pour w = 1.

13.2 - Optimalité des formules de Gauss

Les formules de Gauss sont optimales au sens suivant:

a) Leur degré est le plus élevé possible: nous avons vu en section 9 qu'il n'est pas possible d'excéder le degré (2n + 1).

b) Ce sont les seules formules de degré (2n+1)

En effet, nous avons vu en section 9.1 qu'une formule de degré (2n + 1) doit avoir son support tel que $\prod(x)$ soit orthogonal à tout polynôme de degré n. Mais il n'existe qu'un tel polynôme, c'est φ_{n+1}.

13.3 - Erreur d'intégration

Le degré étant (2n + 1), on peut, dans la formule générale de l'erreur, poser

$$x_{n+1} = x_0 \quad , \quad \cdots \quad , \quad x_{2n+1} = x_n \quad ,$$

ce qui donne

$$\boxed{\Delta I(f) = I\left(\frac{f^{(2n+2)}(\xi)}{(2n+2)!}\,\varphi_{n+1}^2\right) = \frac{f^{(2n+2)}(\eta)}{(2n+2)!}\,\|\varphi_{n+1}\|^2}$$

η étant un point de l'intervalle.

13.4 - Calcul direct des poids de Gauss

Le polynôme $\varphi_n L_i$ est de degré 2n, donc

$$I(\varphi_n L_i) = H_i\,\varphi_n(x_i);$$

d'autre part,

$$I(\varphi_n L_i) = \int_a^b w(x)\,\varphi_n(x)\,\frac{\varphi_{n+1}(x)}{(x-x_i)\,\varphi'_{n+1}(x_i)}\,dx$$

$$= \int_a^b w\,\varphi_n\,(\varphi_n + \mathcal{A}\varphi_{n-1} + \cdots) = \frac{1}{\varphi'_n(x_i)}\,\|\varphi_n\|^2 \,.$$

Il en découle la formule générale

$$\boxed{H_i = \frac{\|\varphi_n\|^2}{\varphi_n(x_i)\,\varphi'_{n+1}(x_i)}}$$

Dans le cas où la densité w est de Rodrigues, cette formule peut être améliorée de manière à ne plus faire apparaître que φ_{n+1} et à être indépendante du coefficient de tête de ce polynôme. Appliquant en effet la formule de récurrence aux dérivées à φ_{n+1} , on obtient

$$\varphi'_{n+1} = ((n+1)\,\beta_2 x + (n+1)\,\beta_1 - \beta_2 b_{n+1})\,\varphi_{n+1} - (\alpha_1 + (2n+3)\,\beta_2)\,\frac{\|\varphi_{n+1}\|^2}{\|\varphi_n\|^2}\,\varphi_n \,.$$

Aux zéros x_i de φ_{n+1} , il vient donc

$$\frac{1}{\varphi_n(x_i)} = -\frac{(\alpha_1 + (2n+3)\,\beta_2)}{\beta(x_i)\,\varphi'_{n+1}(x_i)} \,.\, \frac{\|\varphi_{n+1}\|^2}{\|\varphi_n\|^2} \,,$$

de qui donne

$$\boxed{H_i = -\frac{(\alpha_1 + (2n+3)\,\beta_2)\,\|\varphi_{n+1}\|^2}{\beta(x_i)\,(\varphi'_{n+1}(x_i))^2}}$$

14. QUELQUES FORMULES D'INTEGRATION DE GAUSS

14.1 - Intégration de Gauss-Legendre (Gauss classique)

$a = -1$, $b = 1$, $w = 1$

Points de Gauss: zéros des polynômes de Legendre

Poids: $$H_i = \frac{2^{2n+2}((n+1)!)^4}{((2n+2)!)^2} \cdot \frac{2}{(1-x_i^2)(\hat{P}'_{n+1}(x_i))^2}$$

Erreur: $$I(f) = \varepsilon_n \, 2^{2n+3} \, f^{(2n+2)}(\xi) \quad ,$$

avec

$$\varepsilon_n = \frac{((n+1)!)^4}{((2n+2)!)^3 \, (2n+3)} \quad .$$

Pour la formule transformée sur l'intervalle $[a, b]$ de longueur h, on a

$$I(f) = \varepsilon_n \, h^{2n+3} \, f^{(n+2)}(\xi) \quad .$$

Voir tableau 3

14.2 - Intégration de Gauss-Tchébicheff

$a = -1$, $b = 1$, $w = \frac{1}{\sqrt{(1-x^2)}}$

Points de Gauss: zéros des polynômes de Tchébicheff

$$x_k = \cos \frac{(2k+1)}{2n+2} \qquad k = 0, \ldots, n \qquad (\theta_k = \arccos x_k)$$

Poids: $$H_i = \frac{(2n+2) \; \|\hat{T}_{n+1}\|^2}{(1-x_i^2) \; \hat{T}'^2_{n+1}(x_i)} \quad .$$

Il est plus aisé de travailler avec les $T_{n+1}(x)$, qui vérifient

$$T_{n+1}(\cos\theta) = \cos (n+1) \, \theta \quad .$$

Il vient en effet

$$\|T_{n+1}\|^2 = \int_0^{\pi} \cos^2(n+1) \, \theta \, d\theta = \frac{\pi}{2}$$

$$1 - x_i^2 = \sin^2 \theta_i$$

$$\frac{d \, T_{n+1}}{dx} = \frac{d(\cos(n+1)\theta)}{d\theta} \cdot \frac{d\theta}{d(\cos \theta)} = - \frac{(n+1) \sin(n+1)\theta}{\sin\theta}$$

et

$$\sin(n+1)\theta_i = \pm 1 \quad ,$$

d'où

$$H_i = \frac{2n+2}{\sin^2\theta_i} \cdot \frac{\pi}{2} \cdot \frac{\sin^2\theta_i}{(n+1)^2 \cdot 1} = \frac{\pi}{n+1} \qquad \text{(tous égaux)}$$

Erreur $$I(f) = \frac{\pi}{2^{2n+1}} \cdot \frac{f^{(2n+2)}(\xi)}{(2n+2)!} = \varepsilon_n \cdot 2^{2n+3} \cdot f^{(2n+2)}(\xi) ,$$

avec $$\varepsilon_n = \frac{\pi}{2^{4n+4}(2n+2)!}$$

Pour un intervalle de longueur h, la formule transformée donne une erreur

$$I(f) = \varepsilon_n \; h^{2n+3} \; f^{(2n+2)}(\xi) .$$

Voir tableau 4

14.3 - Intégration de Gauss-Laguerre

$a = 0 \quad , \quad b = \infty \; , \; w = e^{-x}$

Points de Gauss: zéros des polynômes de Laguerre.

Poids: $$H_i = \frac{((n+1)!)^2}{x_i \; (\hat{L}'_{n+1}(x_i))^2}$$

Erreur: $$I(f) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \cdot ((n+1)!)^2 = \varepsilon_n \; f^{(2n+2)}(\xi)$$

avec $$\varepsilon_n = \frac{((n+1)!)^2}{(2n+2)!}$$

Voir tableau 5

14.4 - Intégration de Gauss-Hermite

$a = -\infty \; , \quad b = \infty \; , \; w = e^{-x^2}$

Points de Gauss: zéros des polynômes d'Hermite.

Poids: $$H_i = \frac{2^{-n}(n+1)! \; \sqrt{\pi}}{\hat{H}'_{n+1}(x_i)}$$

Erreur : $$I(f) = \varepsilon_n \; f^{(n+2)}(\xi) ,$$

avec $$\varepsilon_n = \frac{(n+1)! \; \sqrt{\pi}}{2^{n+1}(2n+2)!}$$

Voir tableau 6

15. - CONVERGENCE EN DEGRE DES FORMULES DE LAGUERRE ET HERMITE

Les formules de Laguerre et Hermite ayant trait à des intervalles non bornés, on ne peut leur appliquer la démonstration fondée sur le théorème de Weierstrass. Cependant, on peut démontrer la convergence en degré de ces formules moyennant des hypothèses raisonnables, en se fondant sur cerains théorèmes d'approximation sur $[0, \infty[$ ou $]-\infty, +\infty[$ que nous allons établir. Ces développements, quelque peu techniques, peuvent être passés en première lecture.

15.1 - Lemme 1 [17] - Soit N un entier positif, et soit P(x) un polynôme donné. Il existe des polynômes p_m tels que

$$p_m(x)\, e^{-\alpha x^2} \underset{]-\infty,+\infty[}{\Longrightarrow} P(x)\, e^{-N\alpha x^2} \quad , \quad \alpha > 0 \quad \text{quelconque.}$$

La démonstration de ce lemme se fait en trois étapes:

a) Soit d'abord $P_o(x) = 1$ et soit $N = 2$: il faut montrer que

$$p_m(x)\, e^{-\alpha x^2} \Longrightarrow e^{-2\alpha x^2} \quad .$$

On remarque simplement que, pour $X = \alpha x^2$,

$$= \sup_{X \in [0,\infty[} \left| e^{-2X} - e^{-X} \sum_{k=0}^{K-1} \frac{(-X)^k}{k!} \right| = \sup_{X \in [0,\infty[} e^{-X} \left| \sum_{k=K}^{\infty} \frac{(-X)^k}{k!} \right|$$

et, en utilisant la majoration du reste des séries alternées,

$$\sup_{X \in [0,\infty[} e^{-X} \frac{1}{K!}\, X^k = \frac{K^k\, e^{-K}}{K!} \approx \frac{1}{\sqrt{2\pi K}} \longrightarrow 0 \quad ,$$

par la formule de Stirling.

b) Toujours pour $P_o = 1$, on peut faire une récurrence sur $\mathbb{N}$: vu a), il existe des polynômes p_m tels que

$$p_m(x)\, e^{-\alpha \frac{N+1}{2} x^2} \Longrightarrow e^{-\alpha(N+1)\, x^2}$$

Or,

$$e^{-\alpha \frac{N+1}{2} x^2} = e^{-\frac{\alpha}{2}(N+1)\, x^2} = e^{-\frac{\alpha}{2} x^2}\, e^{-\frac{\alpha}{2} N x^2} \quad .$$

Supposant la proposition vraie jusqu'à la valeur N, il existe des polynômes q_n tels que

$$q_n(x)\, e^{-\frac{\alpha}{2} x^2} \underset{]-\infty,+\infty[}{\Longrightarrow} e^{-\frac{\alpha}{2} N x^2}$$

et, dès lors,

$$p_m(x)\, q_n(x)\, e^{-\alpha x^2} \underset{]-\infty,\infty[}{\Longrightarrow} e^{-\alpha(N+1)x^2} \quad .$$

c) Dans le cas d'un polynôme quelconque, on effectue la décomposition

$$P(x)\, e^{-N\alpha x^2} = (P(x)\, e^{-\frac{\alpha}{2}x^2})\, e^{-\frac{2N-1}{2}\alpha x^2} .$$

Alors,

$$\sup_{]-\infty,\infty[} |P(x)\, e^{-\frac{\alpha}{2}x^2}| \leq M \quad \text{borné},$$

et, pour $\varepsilon > 0$ donné d'avance, il existe un polynôme p_m tel que

$$\sup_{]-\infty,\infty[} |p_m(x)\, e^{-\frac{\alpha}{2}x^2} - e^{-\frac{2N-1}{2}\alpha x^2}| \leq \frac{\varepsilon}{M} ,$$

ce qui entraîne

$$\sup_{]-\infty,\infty[} |P(x)\, p_m(x)\, e^{-\alpha x^2} - P(x)\, e^{-N\alpha x^2}| \leq \varepsilon .$$

15.2 - Lemme 2 - Soit N un entier positif, et soit P(x) un polynôme donné. Il existe des polynômes p_m tels que

$$p_m(x)\, e^{-\alpha x} \underset{[0,\infty[}{\Longrightarrow} P(x)\, e^{-N\alpha x} , \quad \alpha > 0 \text{ quelconque}$$

Il suffit en effet de paraphraser la démonstration ci-dessus en remplaçant x^2 par x.

15.3 - Premier théorème d'approximation [17] - Toute fonction $f \in C^0(]-\infty, \infty[)$ telle que

$$\lim_{|x| \longrightarrow \infty} f(x) = 0$$

est limite uniforme dans R d'expressions de la forme

$$e^{-\alpha x^2}\, p_m(x) ,$$

p_m désignant un polynôme, $\alpha > 0$ arbitraire.

En effet, l'ensemble de R^2 défini par

$$K = \{ (e^{-\alpha x^2}, x\, e^{-\alpha x^2}),\ x \in R \} \cup \{0\}$$

est un compact de R^2 (*) dont les points X sont en correspondance biunivoque avec $\overline{R}$. La fonction

---

(*) Soit $\{ (e^{-\alpha x_m^2}, x_m\, e^{-\alpha x_m^2}) \}$ une suite de points de K. Si la suite $\{x_m\}$ est bornée, on peut en extraire une sous-suite convergente $\{x_{m'}\}$ et $e^{-\alpha x_{m'}^2} \longrightarrow$, $x_{m'}\, e^{-\alpha x_{m'}^2} \longrightarrow$. Si la suite $\{x_m\}$ n'est pas bornée, on peut en extraire une sous-suite qui tend vers $+\infty$ ou $-\infty$, et $e^{-\alpha x_{m'}^2} \longrightarrow 0$, $x_{m'}\, e^{-\alpha x_{m'}^2} \longrightarrow 0$. Au total, toute suite de K a un point d'accumulation, donc K est compact (Théorème de Bolzano-Weierstrass).

$$\varphi(X) = \begin{cases} f(X_2/X_1) & \text{si } X \neq 0 \\ 0 & \text{si } X = 0 \end{cases}$$

est continue sur K. Dès lors, par le théorème de Weierstrass, il existe des polynômes $p_m(X)$ qui convergent uniformément vers $\varphi(X)$ sur K, et on peut même imposer que $p_m(0) = 0$. Ceci revient à dire que

$$p_m(e^{-\alpha x^2}, x\, e^{-\alpha x^2}) \underset{R}{\Longrightarrow} f(x) ,$$

et le théorème en découle, vu le lemme 1.

15.4 - Second théorème d'approximation [17]

Toute fonction $f \in C^0([\,0, \infty\,[\,)$ telle que

$$\lim_{x \to \infty} f(x) = 0$$

est limite uniforme sur $[\,0, \infty\,[$ d'expressions de la forme

$$e^{-\alpha x}\, p_m(x) ,$$

p_m étant un polynôme, $\alpha > 0$ arbitraire.

En effet, pour $y = e^{-\alpha x}$, posons

$$\varphi(y) = \begin{cases} 0 \text{ si } y = 0 \\ f\left(\dfrac{-\ln y}{\alpha}\right) \text{ si } y \in]0, 1\,] \end{cases} .$$

Cette fonction est continue dans $[\,0, 1\,]$, donc, par le théorème de Weierstrass, il existe des polynômes $p_m(y)$ qui convergent uniformément vers $\varphi(y)$ dans $[\,0, 1\,]$, et on peut même imposer que $p_m(y) = 0$. Ceci revient à dire que

$$p_m(e^{-\alpha x}) \underset{[\,0, \infty\,[}{\Longrightarrow} f(x) ,$$

d'où le théorème, par le lemme 2.

15.5 - Nous somme à présent en mesure de démontrer le théorème suivant: Soit f une fonction continue sur $\left\{ \begin{array}{l} [0, \infty\,[\\ ]-\infty, +\infty\,[\end{array} \right\}$ et telle que

$$\left\{ \begin{array}{l} \lim\limits_{x \to \infty} e^{-\alpha x} f(x) = 0 \\ \lim\limits_{|x| \to \infty} e^{-\alpha x^2} f(x) = 0 \end{array} \right\} \qquad 0 \leq \alpha < 1$$

Alors, les intégrales numériques de degré croissant par la formule de $\left\{ \begin{array}{l} \text{Laguerre} \\ \text{Hermite} \end{array} \right\}$ convergent vers $\left\{ \begin{array}{l} \int_0^{\infty} e^{-x} f(x)\, dx \\ \int_{-\infty}^{+\infty} e^{-x^2} f(x)\, dx \end{array} \right\}$

Il suffit en effet de choisir un polynôme P_m tel que

$$\left\{ \begin{array}{l} \sup_{[0,\infty[} |e^{-\alpha x} P_m(x) - e^{-\alpha x} f(x)| \leq \frac{1-\alpha}{2}\varepsilon \\ \sup_{]-\infty,\infty[} |e^{-\alpha x^2} P_m(x) - e^{-\alpha x^2} f(x)| \leq \sqrt{\frac{1-\alpha}{\pi}}\,\frac{\varepsilon}{2} \end{array} \right\}$$

car, alors, pour toute formule de degré m au moins,

$$|I(f) - I(f)| = |I(f - P_m) - I(f - P_m) + I(P_m) - I(P_m)|$$
$$\leq |I(f - P_m)| + |I(f - P_m)|.$$

Or,

$$|I(f-P_m)| \leq \left\{ \begin{array}{l} \int_0^{\infty} e^{-(1-\alpha)x} |e^{-\alpha x}(f-P_m)|\,dx \leq \frac{1-\alpha}{2}\int_0^{\infty} e^{-(1-\alpha)x}dx \leq \frac{\varepsilon}{2} \\ \int_{-\infty}^{+\infty} e^{-(1-\alpha)x^2} |e^{-\alpha x^2}(f-P_m)|\,dx \leq \sqrt{\frac{1-\alpha}{\pi}}\,\frac{\varepsilon}{2}\int_{-\infty}^{+\infty} e^{-(1-\alpha)x^2}dx \leq \frac{\varepsilon}{2} \end{array} \right\}$$

et

$$|\tilde{I}(f-P_m)| \leq \left\{ \begin{array}{l} \sum_i H_i e^{\alpha x_i} |e^{-\alpha x_i}(f-P_m)| \leq \frac{1-\alpha}{2}\varepsilon\sum_i H_i e^{\alpha x_i} \leq \frac{\varepsilon}{2} \\ \sum_i H_i e^{\alpha x_i^2} |e^{-\alpha x_i^2}(f-P_m)| \leq \sqrt{\frac{1-\alpha}{\pi}}\,\frac{\varepsilon}{2}\sum_i H_i e^{\alpha x_i^2} \leq \frac{\varepsilon}{2} \end{array} \right\}$$

car

$$\left\{ \begin{array}{l} \tilde{I}(e^{\alpha x}) = I(e^{\alpha x}) - \Delta I(e^{\alpha x}) \leq I(e^{\alpha x}) = \frac{1}{1-\alpha} \\ \tilde{I}(e^{\alpha x^2}) = I(e^{\alpha x^2}) - \Delta I(e^{\alpha x^2}) \leq I(e^{\alpha x^2}) = \sqrt{\frac{\pi}{1-\alpha}} \end{array} \right\}$$

puisque l'erreur d'intégration est de la forme

$$\Delta I(F) = \varepsilon_n F^{(2n+2)}(\xi) \quad , \quad \varepsilon_n > 0$$

et que les dérivées paires de $e^{\alpha x}$ et $e^{\alpha x^2}$ sont toutes positives.

15.5 - On peut étendre ce résultat de convergence aux <u>fonctions caractéristiques de semi-intervalles</u> $\delta_{[c,\ d[}$, avec c et d finis. Pour cela, on considère les fonctions continues φ_n et ψ_n définies en section 7.5, et on note que ces fonctions satisfont aux exigences du théorème ci-dessus. Cela étant, la démonstration est la même qu'en 7.5.

15.6 - Partant de là et paraphrasant les démonstration de la section 7, on peut généraliser le théorème de convergence aux <u>fonctions étagées</u>, puis aux <u>fonctions à support compact, intégrables au sens de Riemann</u>.

15.7 - Voici encore un résultat plus général: <u>Soit f une fonction telle que</u>

$$\left\{ \begin{array}{l} \lim_{x \to \infty} \sup |e^{-\alpha x} f| = 0 \\ \lim_{|x| \to \infty} \sup |e^{-\alpha x^2} f| = 0 \end{array} \right\} \qquad 0 \leq \alpha < 1$$

et telle que les fonctions tronquées

$$\begin{cases} f_N(x) = f \ \delta_{[0,\ N[} \\ f_N(x) = f \ \delta_{[-N,\ N[} \end{cases}$$

soient, pour tout N fini, intégrables au sens de Riemann. Alors, les intégrales numériques de $\left\{\begin{matrix}\text{Laguerre}\\ \text{Hermite}\end{matrix}\right.$ de f convergent en degré.

Fixons $\varepsilon > 0$. Par hypothèse, on peut trouver une valeur N_o de N telle que

$$\left\{ \begin{array}{l} \sup\limits_{x \geqslant N_o} |e^{-\alpha x} f| \leqslant (1-\alpha)\,\frac{\varepsilon}{4} \\ \sup\limits_{|x| \geqslant N_o} |e^{-\alpha x^2} f| \leqslant \sqrt{\frac{1-\alpha}{\pi}} \cdot \frac{\varepsilon}{8} \end{array} \right\} ,$$

ce qui implique, quel que soit le degré de la formule,

$$|I(f-f_{N_o})| \leqslant \left\{ \begin{array}{l} (1-\alpha)\,\frac{\varepsilon}{4} \int_{N_o}^{\infty} e^{-(1-\alpha)x}\,dx \leqslant \frac{\varepsilon}{4} \\ \sqrt{\frac{1-\alpha}{\pi}}\ \frac{\varepsilon}{8} \cdot 2 \int_{N_o}^{\infty} e^{-(1-\alpha)x^2}\,dx \leqslant \frac{\varepsilon}{4} \end{array} \right\}$$

et

$$|\tilde{I}(f-f_{N_o})| \leqslant \left\{ \begin{array}{l} (1-\alpha)\,\frac{\varepsilon}{4}\ \tilde{I}(e^{\alpha x}) \leqslant \frac{\varepsilon}{4} \\ \sqrt{\frac{1-\alpha}{\pi}} \cdot \frac{\varepsilon}{8} \cdot \tilde{I}(e^{\alpha x^2}) \leqslant \frac{\varepsilon}{4} \end{array} \right\} .$$

On en déduit

$$|\Delta I(f-f_{N_o})| \leqslant |I(f-f_{N_o})| + |\tilde{I}(f-f_{N_o})| \leqslant \frac{\varepsilon}{2} .$$

Par ailleurs, f_{N_o} est à support compact et intégrable au sens de Riemann. Il existe donc un certain degré à partir duquel les formules d'intégration numérique vérifient

$$|\Delta I(f_{N_o})| \leqslant \frac{\varepsilon}{2} .$$

Mais alors,

$$|\Delta I(f)| \leqslant |\Delta I(f_{N_o})| + |\Delta I(f-f_{N_o})| \leqslant \varepsilon .$$

16. FORMULES DE LOBATTO

16.1 - Sur l'intervalle de référence $[-1, +1]$, considérons l'intégrale particulière

$$L(f) = I(\beta f),$$

avec

$$\beta(x) = 1 - x^2 .$$

On peut construire une suite de polynômes orthogonaux $\psi_0, \ldots, \psi_n, \ldots$ pour cette intégrale de densité w . Soit alors P_{2n-1} un polynôme de degré (2n-1). divisons-le par ψ_{n-1} : il vient, en notant les degrés en indices,

$$P_{2n-1} = \psi_{n-1} Q_{n-2} + R_n ,$$

ce qui entraîne

$$I(P_{2n-1}) = I(\psi_{n-1} Q_{n-2}) + I(R_n) = I(R_n) ,$$

en vertu des propriétés d'orthogonalité de ψ_{n-1}. Pour intégrer numériquement R_n, il suffit de (n+1) points disposés de manière quelconque. Le résultat de l'intégration numérique,

$$\tilde{I}(P_{2n-1}) = \tilde{I}(\psi_{n-1} Q_{n-2}) + I(R_n) ,$$

sera donc exact si

$$\tilde{I}(\psi_{n-1} Q_{n-2}) = 0,$$

quel que soit Q_{n-2}. Or, ceci est réalisé si l'on choisit le support constitué:

- <u>des deux extrémités de l'intervalle</u>, où $\beta = 0$
- <u>des (n-1) zéros de</u> ψ_{n-1}

Les formules ainsi construites portent le nom de <u>formules de Lobatto</u>. Bien que ne garantissant que le degré (2n-1) pour (n+1) points, elles sont parfois plus performantes que les formules de Gauss dans les problèmes à plusieurs dimensions [3, 4] .

16.2 - Il est aisé de montrer que <u>les poids de Lobatto sont positifs</u>. Pour $i = 1, \ldots, (n-1)$, d'abord, on remarque que la fonction

$$F_i(x) = \frac{\beta(x)}{\beta(x_i)} \prod_{\substack{j=1 \\ \neq i}} \frac{(x-x_j)^2}{(x_i-x_j)^2}$$

est positive, et de degré $2 + 2(n-2) = 2n-2$. Elle s'annule en tous les points du support, à l'exception de x_i où elle vaut l'unité. Par conséquent,

$$0 < I(F_i) = \tilde{I}(F_i) = H_i .$$

Pour les points $x_o=-1$ et $x_n = +1$, on tient le même raisonnement avec les polynômes

$$F_o(x) = \frac{1-x}{2} \psi^2_{n-1}(x) \qquad \text{et} \qquad F_n(x) = \frac{1+x}{2} \psi^2_{n-1}(x).$$

16.3 - Erreur d'intégration

Conformément à la théorie générale, on peut choisir les points

$$x_{n+2} = x_1 \ , \ldots , \ x_{2n} = x_{n-1} \ ,$$

ce qui conduit à la fonction négative

$$\Pi^*_{2n}(x) = (x^2 - 1) \ \psi^2_{n-1} = - \beta \psi^2_{n-1} \ .$$

Il en découle l'expression suivante de l'erreur:

$$\boxed{\Delta I(f) = - \frac{f^{(2n)}(\xi)}{(2n)!} \ I(\beta \psi^2_{n-1})}$$

16.4 - Poids de Lobatto

a) Pour i = 1,..., (n-1), calculons l'intégrale de la fonction

$$F_i(x) = \beta(x) \frac{\psi_{n-1}(x)}{x-x_i} \ \psi_{n-2}(x) \ .$$

Comme

$$\frac{\psi_{n-1}(x)}{x-x_i} = x^{n-2} + \ldots . \ ,$$

on a

$$I(F_i) = I(\beta x^{n-2} \ \psi_{n-2}) = I(\beta \psi^2_{n-2}) \ .$$

La fonction F_i étant un polynôme de degré (2n-2), son intégrale numérique est exacte:

$$I(F_i) = I(F_i) = H_i \ \beta(x_i) \ \psi'_{n-1}(x_i) \ \psi_{n-2}(x_i).$$

On en déduit

$$\boxed{H_i = \frac{I(\beta \ \psi^2_{n-2})}{\beta(x_i) \ \psi'_{n-1}(x_i) \ \psi_{n-2}(x_i)}}$$

b) Pour i=0, on a

$$I((1-x)\psi^2_{n-1}) = I((1-x)\psi^2_{n-1}) = 2 \ H_o \psi^2_{n-1}(-1)$$

soit

$$\boxed{H_o = \frac{I((1-x)\psi^2_{n-1})}{2 \psi^2_{n-1}(-1)}}$$

c) De la même façon, pour i = n ,

$$I((1+x)\psi_{n-1}^2) = I((1+x)\psi_{n-1}^2) = 2\, H_n \psi_{n-1}^2(1)$$

et

$$\boxed{H_n = \frac{I((1+x)\psi_{n-1}^2)}{2\psi_{n-1}^2(1)}}$$

16.5 - Cas d'une densité de Rodrigues

Comme on sait, lorsque w est une densité de Rodigues, les polynômes ψ_{n-1} ne sont autres que

$$\psi_{n-1} = \frac{1}{n}\ \varphi_n' \quad ,$$

où les φ_n sont les polynômes orthogonaux pour w. La densité est nécessairement de la forme

$$w(x) = C\ (1-x)^{\lambda}\ (1+x)^{\mu} \qquad , \quad \lambda\ , \mu > 1$$

et

$$\varphi_n = \hat{p}_n^{(\lambda\ ,\mu)} \qquad , \quad \psi_{n-1} = \hat{p}_{n-1}^{(\lambda+1,\ \mu+1)} .$$

On peut alors transformer la formule donnant les poids de Lobatto pour i= 1,..., (n-1) en partant de la relation de récurrence aux dérivées

$$\beta\psi_{n-1}' = (\mathcal{A}_{n-1}\ x + \mathcal{B}_{n-1})\ \psi_{n-1} + \mathcal{C}_{n-1}\ \psi_{n-2}$$

qui, aux zéros x_i de ψ_{n-1}, se réduit à

$$\beta\psi_{n-1}'(x_i) = \mathcal{C}_{n-1}\ \psi_{n-2}(x_i),$$

ce qui donne

$$\boxed{H_i = \frac{I(\psi_{n-2}^2)}{\mathcal{C}_{n-1}\psi_{n-2}^2(x_i)}} \quad ,$$

avec

$$\mathcal{C}_{n-1} = \mathcal{C}_{n-1}(\lambda+1, \mu+1) = \frac{4(n-1)(n+\lambda+\mu+1)(n+\lambda)(n+\mu)}{(2n+\lambda+\mu+1)(2n+\lambda+\mu)^2} .$$

On obtient ainsi des expressions des poids indépendantes des coefficients de tête des ψ_n.

Il est également possible, et souvent préférable, de tout exprimer en fonction des polynômes φ_n. Tout d'abord, on sait que

$$I(\beta\psi_{n-1}^2) = \frac{\gamma_n}{n^2}\ \|\varphi_n\|^2 \quad ,$$

ce qui donne pour l'erreur d'intégration

$$\Delta I(f) = - \frac{f^{(2n)}(\xi)}{(2n)!} \quad \frac{\gamma_n}{n^2} \quad \| \varphi_n \|^2$$

Pour obtenir les poids H_i , $i = 1,\ldots,(n-1)$, considérons la fonction

$$F_i = \beta \frac{\varphi_n'}{x-x_i} \quad \varphi_n \ .$$

Comme

$$\beta \frac{\varphi_n'}{x-x_i} = - n\, x^n + \ldots \ ,$$

on a directement

$$I(F_i) = - n\, I(\varphi_n\, x^n) = - n \| \varphi_n \|^2 \ .$$

L'intégrale numérique de cette fonction vaut

$$\tilde{I}(F_i) = H_i\, \beta(x_i)\, \varphi_n''(x_i)\, \varphi_n(x_i) \ .$$

Or, l'équation différentielle régissant les φ_n s'écrit

$$\beta \varphi_n'' + (\alpha + \beta') \varphi_n' = - \gamma_n \varphi_n$$

et, aux zéros x_i de φ_n' , elle se réduit à

$$\beta \varphi_n''(x_i) = - \gamma_n \varphi_n(x_i) .$$

On a donc

$$\tilde{I}(F_i) = - \gamma_n H_i \varphi_n^2(x_i) .$$

Comme la fonction F_i est un polynôme de degré (2n), l'intégrale numérique est entachée d'une erreur que l'on peut calculer: puisque

$$F_i(x) = - n\, x^{2n} + \ldots \ ,$$

on a

$$F_i^{(2n)}(x) = -n\ (2n)!$$

et

$$\Delta I(F_i) = \frac{\gamma_n}{n} \quad \| \varphi_n \|^2 \ .$$

Dès lors, l'équation

$$I(F_i) = \tilde{I}(F_i) + \ I(F_i)$$

conduit, après remplacement par les valeurs explicites, à

$$H_i = \frac{n(1 + \frac{\gamma_n}{n^2}) \| \varphi_n \|^2}{\gamma_n \varphi_n^2(x_i)}$$

Pour trouver H_o , considérons la fonction

$$F_o(x) = (x-1)\, \varphi_n\, \varphi_n' \ ,$$

qui est de la forme

$$F_o = n \varphi_n (\varphi_n + P_{n-1})$$

et a donc pour intégrale

$$I(F_o) = n \, \| \varphi_n \|^2 \, .$$

Son intégrale numérique vaut par ailleurs

$$\tilde{I}(F_o) = - 2 H_o \varphi_n(-1) \, \varphi_n'(-1) \, .$$

On peut éliminer la dérivée à partir de l'équation différentielle

$$\beta \varphi_n'' + (\alpha + \beta') \varphi_n' = - \gamma_n \varphi_n$$

dont, en x = -1, le premier terme disparaît. Il vient

$$\varphi_n'(-1) = - \frac{\gamma_n \varphi_n(-1)}{\alpha_o - \alpha_1 + 2}$$

et

$$\tilde{I}(F_o) = \frac{2 H_o \gamma_n \varphi_n^2(-1)}{\alpha_o - \alpha_1 + 2} \, .$$

Comme

$$F_o(x) = n \, x^{2n} + \ldots . \, ,$$

on a

$$F_o^{(2n)} = n \, (2n)!$$

et

$$\Delta I(F_o) = - \frac{\gamma_n}{n^2} \, \| \varphi_n \|^2 \, .$$

Identifiant $I(F_o)$ et $\tilde{I}(F_o) + \Delta I(F_o)$, on obtient

$$H_o = \frac{n(1 + \frac{\gamma_n}{n^2}) \, (\alpha_o - \alpha_1 + 2) \| \varphi_n \|^2}{2 \gamma_n \varphi_n^2(-1)} = \frac{n(1 + \frac{\gamma_n}{n^2})(\mu+1) \| \varphi_n \|^2}{\gamma_n \varphi_n^2(-1)}$$

Enfin, H_n se calcule en intégrant la fonction

$$F_n(x) = (x+1) \varphi_n \varphi_n'$$

qui a pour intégrale

$$I(F_n) = n \, \| \varphi_n \|^2 \, ,$$

et, pour intégrale numérique,

$$\tilde{I}(F_n) = 2 H_n \varphi_n(1) \varphi_n'(1) \, .$$

On peut encore éliminer $\varphi_n'(1)$ par la relation

$$\varphi_n'(1) = - \frac{\gamma_n \varphi_n(1)}{\alpha_o + \alpha_1 - 2} \, ,$$

ce qui donne

$$\tilde{I}(F_n) = - \frac{2 H_n \gamma_n \varphi_n^2(1)}{\alpha_o + \alpha_1 - 2} \, .$$

Enfin,

$$\Delta I(F_n) = \frac{\gamma_n}{n} \, \| \varphi_n \|^2 \, ,$$

et on trouve finalement

$$H_n = \frac{n(1+\frac{\gamma_n}{n^2})(2-\alpha_1-\alpha_o)\,\|\varphi_n\|^2}{2\,\gamma_n\,\varphi_n^2(1)} = \frac{n(1+\frac{\gamma_n}{n^2})(\lambda+1)\,\|\varphi_n\|^2}{\gamma_n\,\varphi_n^2(1)}$$

On notera que, si la densité est symétrique, $\lambda = \mu$ et $\varphi_n(1) = \pm\,\varphi_n(-1)$, si bien que $H_o = H_n$.

16.6 - Formule de Lobatto classique

C'est celle qui correspond à une densité unitaire. On a donc

$$\lambda = \mu = 0$$
$$\gamma_n = n(n+1)$$
$$1 + \frac{\gamma_n}{n^2} = \frac{2n+1}{n} .$$

Les polynômes orthogonaux φ_n sont ceux de Legendre. On sait que

$$\|\hat{P}_n\|^2 = \frac{2^{2n+1}(n!)^4}{((2n)!)^2(2n+1)} ,$$

ce qui permet de calculer

$$H_i = \frac{2^{2n}\,(n!)^4}{(2n!)^2} \cdot \frac{2}{n(n+1)\hat{P}_n^2(x_i)} ,$$

pour i = 0, ... , n. Pour H_o et H_n, on peut calculer a priori $\hat{P}_n(\pm 1)$ de la façon suivante:

$$A_n\,\hat{P}_n = D^n((1-x^2)^n) = \sum_{k=0}^{n} C_n^k\, D^k(1-x^2)\, D^{n-k}(1-x^2)^{n-1}$$

$$= 0 - 2n\,x\,D^{n-1}((1-x^2)^{n-1}) - 2\,C_n^2\,D^{n-2}((1-x^2)^{n-1})$$

et, en $x = \pm 1$,

$$A_n\,\hat{P}_n(\pm 1) = \mp\,(-1)^n\,2n\,A_{n-1}\,\hat{P}_{n-1}(\pm 1) .$$

Comme

$$A_o\,\hat{P}_o = 1,$$

on a

$$\hat{P}_n(\pm 1) = \mp\,\frac{2^n\,n!\,(-1)^n}{A_n} = \mp\,\frac{2^n\,(n!)}{(2n)!}$$

et

$$H_o = H_n = \frac{2}{n(n+1)}$$

L'erreur d'intégration est donnée par

$$I(f) = -\,\varepsilon_n\,2^{2n+1}\,f^{(2n)}(\xi) ,$$

avec

$$\varepsilon_n = \frac{n+1}{n(2n+1)}\,\frac{(n!)^4}{(2n!)^3} .$$

La formule adaptée à un intervalle de longueur h conduit à une erreur

$$I(f) = - \varepsilon_n \, h^{2n+1} \, f^{(2n)}(\xi)$$

Voir tableau 7

17. PARTIE PRINCIPALE DE L'ERREUR POUR w = 1 SUR UN INTERVALLE BORNE

Lorsque l'intervalle d'intégration a une longueur h petite, les puissances successives de h vont en décroissant. Il est alors intéressant de développer l'erreur d'intégration en puissances de h. Supposons $f \in C^q(\,[a, b]\,)$ et développons cette fonction en série limitée de puissances autour du centre c de l'intervalle:

$$f(x) = f(c) + (x-c)f'(c) + \frac{(x-c)^2}{2} f''(c) + \ldots + \frac{(x-c)^q}{q!} f^{(q)}(x^*).$$

En intégrant, on obtient

$$I(f) = f(c)I(1) + f'(c)I(x-c) + f''(c)I(\frac{(x-c)^2}{2}) + \ldots + I(\frac{(x-c)^q}{q!} f^{(q)}(x^*)) .$$

De même, l'intégrale numérique se développe en

$$\tilde{I}(f) = f(c)\tilde{I}(1) + f'(c)\tilde{I}(x-c) + f''(c)\tilde{I}(\frac{(x-c)^2}{2}) + \ldots + \tilde{I}(\frac{(x-c)^q}{q!} f^{(q)}(x^*))$$

et, par différence, on obtient

$$\Delta I(f) = f(c)\,\Delta I(1) + f'(c)\,\Delta I(x-c) + f''(c)\,\Delta I(\frac{(x-c)^2}{2}) + \ldots$$
$$\ldots + \Delta I(\frac{(x-c)^q}{q!} f^{(q)}(x^*)) .$$

Les termes jusqu'à l'ordre (q-1) se ramènent avec profit à l'intervalle de référence $[\,-1, +1\,]$ par la transformation

$$x = c + \frac{h}{2}\xi ,$$

ce qui donne

$$\Delta I(\frac{(x-c)^k}{k!}) = (\frac{h}{2})^{k+1} \Delta \hat{I}(\frac{\xi^k}{k!}) = h^{k+1} \gamma_k ,$$

en introduisant la notation

$$\boxed{\gamma_k = \frac{1}{2^{k+1}} \hat{I}(\frac{\xi^k}{k!})}$$

Quant au dernier terme, il est majoré par

$$\gamma_q^o \, h^{q+1} \sup_{[a, b]} |f^{(q)}| ,$$

avec

$$\gamma_q^o = \hat{I}(\frac{|\xi|^q}{q!}) + \tilde{\hat{I}}(\frac{|\xi|^q}{q!}) .$$

On obtient ainsi

$$\Delta I(f) = \gamma_0 h\, f(c) + \gamma_1 h^2\, f'(c) + \ldots + \gamma_{q-1}\, h^q\, f^{(q-1)}(c) + \delta_q ,$$

avec

$$|\delta_q| \leq \gamma_q^0\, h^{q+1} \sup_{(a,\,b)} |f^{(q)}| .$$

Cela étant, la <u>partie principale de l'erreur</u> est le premier terme non nul du développement de l'erreur. Pour une formule de degré p, on a par définition $\gamma_0 = \gamma_1 = \ldots = \gamma_p = 0$, d'où

$$\Delta I(f) = \gamma_{p+1}\, h^{p+2}\, f^{(p+1)}(c) + \gamma_{p+2}\, h^{p+3}\, f^{(p+2)}(c) + \ldots$$

et la partie principale de l'erreur est

$$\Delta I_{princ}(f) = \gamma_{p+1}\, h^{p+2}\, f^{(p+1)}(c) .$$

Le cas courant des formules <u>symétriques</u> mérite une certaine attention. En effet, pour ces formules, on a, pour k <u>impair</u>,

$$\hat{I}\left(\frac{\xi^k}{k!}\right) = 0 \quad , \qquad \tilde{\tilde{I}}\left(\frac{\xi^k}{k!}\right) = 0 ,$$

ce qui entraîne la nullité de tous les γ_k d'ordre impair. Ceci implique en particulier qu'<u>une formule symétrique est toujours de degré impair</u>, en accord avec les résultats de la section 4. Son erreur est de la forme

$$\Delta I(f) = \gamma_{p+1}\, h^{p+2}\, f^{(p+1)}(c) + \gamma_{p+3}\, h^{p+4}\, f^{(p+3)}(c) + \ldots$$

En particulier, la partie principale de l'erreur est exacte à $O(h^{p+4})$ près, c'est-à-dire avec une erreur relative $O(h^2)$.

18. <u>EXTRAPOLATION DE RICHARDSON</u>

Soit à calculer l'intégrale de f sur (a, b). On subdivise l'intervalle en N sous-intervalles de longueur $h = (b-a)/N$ et de centres respectifs $c_1, \ldots, c_N$. En utilisant dans le sous-intervalle n° k une formule de degré p, on a

$$\Delta I_k(f) = \gamma_{p+1}\, h^{p+2}\, f^{(p+1)}(c_k) + \gamma_{p+3}\, h^{p+4}\, f^{(p+2)}(c_k) + \ldots$$

et, pour l'intervalle complet,

$$\Delta I(f) = \gamma_{p+1} h^{p+2} \sum_{k=1}^{N} f^{(p+1)}(c_k) + \gamma_{p+3} h^{p+4} \sum_{k=1}^{N} f^{(p+3)}(c_k) + \ldots$$

Le sommes effectuées sur les centres des sous-intervalles peuvent être considérées comme le calcul d'une intégrale par la formule des rectangles, qui est une formule symétrique de degré 1. Notant $\overline{\gamma}_k$ les nombres γ_k de cette dernière formule, on a, pour toute fonction φ,

$$I_k(\varphi) = h\,\varphi(c_k) + \bar{\gamma}_2\, h^3\,\varphi''(c_k) + \bar{\gamma}_4\, h^5\,\varphi^{IV}(c_k) + \dots ,$$

d'où

$$I(\varphi) = h \sum_k \varphi(c_k) + \bar{\gamma}_2 h^3 \sum_k \varphi''(c_k) + \bar{\gamma}_4 h^5 \sum_k \varphi^{IV}(c_k) + \dots$$

On peut en déduire le système linéaire

$$\left\{\begin{aligned}
\Delta I(f) &= \gamma_{p+1} h^{p+2} \sum_k f^{(p+1)}(c_k) + \gamma_{p+3} h^{p+4} \sum_k f^{(p+3)}(c_k) + \gamma_{p+5} h^{p+6} \sum_k f^{(p+5)}(c_k) + \dots \\
h^{p+1} I(f^{(p+1)}) &= h^{p+2} \sum_k f^{(p+1)}(c_k) + \bar{\gamma}_2 h^{p+4} \sum_k f^{(p+3)}(c_k) + \bar{\gamma}_4 h^{p+6} \sum_k f^{(p+5)}(c_k) + \dots \\
h^{p+3} I(f^{(p+3)}) &= \qquad\qquad h^{p+4} \sum_k f^{(p+3)}(c_k) + \bar{\gamma}_2 h^{p+6} \sum_k f^{(p+5)}(c_k) + \dots \\
&\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots
\end{aligned}\right.$$

qui, par élimination des sommes, donne, en supposant $f \in C^{q+2}$, q pair,

$$\Delta I(f) = \beta_{p+1} h^{p+1} (f^{(p+1)}) + \beta_{p+3} h^{p+3} I(f^{(p+3)}) + \dots + \beta_q h^q I(f^{(q)}) + O(h^{q+2}) .$$

Ceci peut aussi s'écrire

$$I(f) = I(f) + h^{p+1} (\beta_{p+1} I(f^{(p+1)}) + \beta_{p+3} h^2 I(f^{(p+3)}) + \dots) + O(h^{q+2}$$

ou encore,

$$I(f;h) = I(f) + a_{p+1} h^{p+1} + a_{p+3} h^{p+3} + \dots + a_q h^q + O(h^{q+2}) ,$$

les coefficients a_k dépendant de la formule utilisée et de la fonction, mais non de h. Il est clair qu'à la limite, $I(f;0) = I(f)$. Ceci suggère une <u>procédure d'extrapolation</u>. Partant d'abord de deux calculs avec des valeurs différentes de h, h_1 h_2, on écrira

$$I_1 = I(f;h_1) \approx I(f) + a_{p+1}\, h_1^{p+1}$$

$$I_2 = I(f;h_2) \approx I(f) + a_{p+1}\, h_2^{p+1} .$$

Par différence, on obtient

$$I_1 - I_2 \approx a_{p+1} (h_1^{p+1} - h_2^{p+1}) ,$$

ce qui permet d'écrire, à partir de la première relation,

$$I(f) \approx I_1 - \frac{(I_1 - I_2) h_1^{p+1}}{h_1^{p+1} - h_2^{p+1}} = I_1 - \frac{(I_1 - I_2)}{1 - \left(\frac{h_2}{h_1}\right)^{p+1}} ,$$

soit

$$\boxed{I(f) \approx \frac{1}{1 - \left(\frac{h_2}{h_1}\right)^{p+1}} \left(I_2 - \left(\frac{h_2}{h_1}\right)^{p+1} I_1 \right)}$$

Ceci permet d'ailleurs d'évaluer l'erreur d'intégration de I_1 ou I_2 (Règle de RUNGE) :

$$\Delta I_2 = I(f) - I_2 \approx \frac{(h_2/h_1)^{p+1}}{1 - (h_2/h_1)^{p+1}}(I_2 - I_1)$$

$$\Delta I_1 = I(f) - I_1 \approx \frac{I_2 - I_1}{1 - (h_2/h_1)^{p+1}}$$

Quelle est l'erreur résiduelle? Comme

$$I_1 \approx I(f) + a_{p+1}h_1^{p+1} + a_{p+3}h_1^{p+3}$$

$$I_2 \approx I(f) + a_{p+1}h_2^{p+1} + a_{p+3}h_2^{p+3} \quad ,$$

la valeur extrapolée I_1^1 vérifie

$$I_1^1 = \frac{1}{1 - (h_2/h_1)^{p+1}}(I_2 - (h_2/h_1)^{p+1} I_1)$$

$$\approx I(f) - a_{p+3}\, h_1^{p+3}\, (h_2/h_1)^{p+1}\, \frac{1 - (h_2/h_1)^2}{1 - (h_2/h_1)^{p+1}} \quad ,$$

c'est-à-dire que l'on a gagné deux degrés et deux ordres de h, puisque a_{p+1} (et, donc $f^{(p+1)}$) n'intervient plus.

A titre d'exemple, examinons ce que devient la formule des trapèzes dans ce processus, lorsque $h_2/h_1 = \frac{1}{2}$. Au départ, on calcule

$$I_1(f) = (h/2)\,(f_o + f_2) \,.$$

Avec des intervalles deux fois plus petits,

$$I_2(f) = (h/4)\,(f_o + 2\,f_1 + f_2) \,.$$

Par l'extrapolation, on obtient

$$I_1^1(f) = \frac{1}{1 - (1/4)}(I_2(f) - (1/4)\, I_1(f))$$

$$= (4/3)\,(\,(h/4)(f_o + 2\,f_1 + f_2) - (h/8)\,(f_o + f_2))$$

$$= (h/6)\,(f_o + 4\,f_1 + f_2) \quad ,$$

c'est-à-dire la formule de Simpson qui est, en effet, de degré 3.

19. METHODES A LA ROMBERG

Considérons une formule de degré 1, symétrique, à poids positifs. On a

$$I(f;h) = I(f) + a_2h^2 + a_4h^4 + \ldots$$

Il est tentant, à partir d'un certain nombre (m+1) de valeurs calculées

pour différentes valeurs de h, d'extrapoler sous la forme d'un polynôme d'interpolation en h^2. Ayant deux valeurs h_1 et h_2 conduisant à I_1 et I_2, on peut calculer une extrapolation

$$I_1^1 = \frac{I_2 - (h_2/h_1)\, I_1}{1 - (h_2/h_1)^2}$$

Ayant donc $I_1, \ldots, I_{m+1}$, on peut former de la sorte $I_1^1, \ldots, I_1^m$. De là, les extrapolations sur trois valeurs de h^2 s'obtiennent par la formule d'Aitken:

$$I_i^2 = I_{(h_i^2, h_{i+1}^2, h_{i+2}^2)} = \frac{(0-h_{i+2}^2) I_{(h_i^2, h_{i+1}^2)} - (0-h_i^2) I_{(h_{i+1}^2, h_{i+2}^2)}}{h_i^2 - h_{i+2}^2}$$

soit

$$I_i^2 = \frac{h_i^2\, I_{i+1}^1 - h_{i+2}^2\, I_i^1}{h_i^2 - h_{i+2}^2} = \frac{I_{i+1}^1 - (h_{i+2}/h_i)^2\, I_i^1}{1 - (h_{i+2}/h_i)^2}$$

Plus généralement, l'interpolation I_i^j sur $h_i^2, h_{i+1}^2, \ldots, h_{i+j}^2$ se calcule par

$$\boxed{I_i^j = \frac{I_{i+1}^{j-1} - (h_{i+j}/h_i)^2\, I_i^{j-1}}{1 - (h_{i+j}/h_i)^2}}$$

On peut donc former un tableau d'approximations successives

$$\begin{array}{lllll} I_1 & & & & \\ I_2 & I_1^1 & & & \\ I_3 & I_2^1 & I_1^2 & & \\ I_4 & I_3^1 & I_2^2 & I_1^3 & \\ \vdots & & & & \\ I_n & I_{n-1}^1 & I_{n-2}^2 & I_{n-3}^3 & \ldots\ I_1^{n-1} \end{array}$$

d'autant meilleures que l'on est plus bas et plus à droite dans le tableau. L'application de ce schéma à la formule des trapèzes, avec $h_i/h_{i+1} = 2$ porte le nom de <u>méthode de Romberg</u>

Examinons l'ordre de grandeur de l'erreur en supposant, pour la simplicité, $f \in C^\infty$. I_n^m est la valeur obtenue par extrapolation entre $h_n^2, h_{n+1}^2, \ldots, h_{n+m}^2$. L'erreur est donc, en posant $F(f) = I(f;h)$,

$$\begin{aligned} I(f) - I_n^m &= (0-h_n^2)\ldots(0-h_{n+m}^2)\, F(h_n^2, \ldots, h_{n+m}^2, 0) \\ &= (-1)^{m+1} h_n^2 \ldots h_{n+m}^2\, F(h_n^2, \ldots, h_{n+m}^2, 0) \end{aligned}$$

$$= (-1)^{m+1} h_n^2 \dots h_{n+m}^2 \frac{D^m F(\xi^2)}{m!} \quad , \quad 0 < \xi^2 < h_n^2$$

$$= (-1)^{m+1} h_n^2 \dots h_{n+m}^2 \, (a_{2m} + (m+1)\, a_{2m+2}\, \xi^2 + \dots) \quad .$$

Elle est donc au pire $O(h_m^{2m+2})$. Dans le cas où la fonction à intégrer est un polynôme de degré $r < 2m$, l'erreur s'annule, puisque

$$a_k = \beta_k \, I(f^{(k)})$$

s'annule pour $k > r$. Ainsi, les approximations de la diagonale finissent par être exactes pour tout polynôme donné.

Que dire du signe des poids? Nous allons montrer que si le calcul de I_{i+1} reprend tous les points de I_i, chacun à la même place dans les sous-intervalles, les poids des I_n^m sont tous positifs.

Cette circonstance se produit dans la méthode de Romberg, car chaque sous-intervalle servant au calcul de I_i contient exactement deux sous-intervalles de I_{i+1} (fig. 10).

Si l'on veut utiliser la formule des rectangles, ce qui présente l'avantage d'une formule ouverte, permettant de traiter certaines singularités ou indéterminations aux extrémités de l'intervalle, il convient de diviser les sous-intervalles en trois parties ($h_1/h_2 = 3$), comme l'illustre la fig. 11.

Démontrons donc notre assertion relative aux poids. Les poids des I_i, tout d'abord, sont positifs par hypothèse. Montrons donc que si les poids des I_n^{m-1} sont positifs, il en est de même de ceux des I_n^m (récurrence sur les colonnes du tableau). Tout repose sur la formule

$$I_n^m = \frac{I_{n+1}^{m-1} - (h_{n+m}/h_n)^2 \, I_n^n}{1 - (h_{n+m}/h_n)^2} \quad .$$

Pour les points du support de I_{n+1}^{m-1} non repris dans le support de I_n^{m-1}, c'est évident, car seule intervient I_{n+1}^{m-1}, à poids positifs. Pour les autres points, la contribution de I_{n+1}^{m-1} est de la forme $K\, h_{n+1}$, celle de I_n^{m-1} étant alors $K\, h_n$. Il en résulte, pour (h_{n+1}/h_n) constant, la valeur suivante du poids:

$$K \frac{h_{n+1} - (h_{n+m}/h_n)^2 \, h_n}{1 - (h_{n+m}/h_n)^2} = K \frac{h_{n+1} - (h_{n+1}/h_n)^{2m} \, h_n}{1 - (h_{n+1}/h_n)^{2m}}$$

$$> K \frac{h_{n+1} - (h_{n+1}/h_n)\, h_n}{1 - (h_{n+1}/h_n)^{2m}} = 0$$

Nous avons donc obtenu le résultat suivant: la méthode des extrapolations successives, obtenues à partir

de la formule des trapèzes, avec $h_n/h_{n+1} = 2$	de la formule des rectangles, avec $h_n/h_{n+1} = 3$

donne sur la diagonale du tableau des valeurs pouvant être considérées comme obtenues par des formules de degrés croissant à l'infini et à poids positifs: il y a donc convergence pour toute fonction continue et même pour toute fonction intégrable au sens de Riemann.

En pratique, on se limite souvent à construire quelques colonnes du tableau, soit les I_n^m , $m \leq k$. Alors, les I_n^k sont des intégrales approchées par des formules à poids positifs sur des sous-intervalles décroissants, ce qui assure la convergence à l'ordre $0(h^{2k+2})$ si la fonction est régulière et même, en général, la convergencd (mais sans garantie d'ordre) pour toute fonction intégrable au sens de Riemann.

20. INTERPOLATION DES FONCTIONS FORTEMENT OSCILLANTES

Il arrive parfois que l'on ait à intégrer des fonctions qui oscillent rapidement. L'exemple caractéristique est le calcul des coefficients de Fourier d'ordre élevé

$$a_n = \frac{1}{\pi} \int_0^{2\pi} f(\theta) \cos n\theta \, d\theta \quad , \quad b_n = \frac{1}{\pi} \int_0^{2\pi} f(\theta) \sin n\theta \, d\theta$$

Dans ce cas, les méthodes classiques ne peuvent être appliquées directement. Supposons par exemple que l'on utilise une formule de degré p, sur un découpage de l'intervalle $]0, 2\pi[$ en sous-intervalles de longueur h. Pour une fonction g, on a alors

$$I(g) = K \, h^{p+1} \, g^{(p+1)}(\xi),$$

le coefficient K dépendant de la formule utilisée. Mais la dérivée d'ordre (p+1) croît rapidement avec l'ordre de l'harmonique considéré:

$$D^{p+1}(f \, e^{in\theta}) = \sum_{k=0}^{p+1} C_{p+1}^k (in)^{p+1-k} \, D^k f = 0(n^{p+1}).$$

Dès lors, lorsque n croît, la précision d'intégration se détériore très vite.

Pour contourner cette difficulté, il faut interpoler f seulement. Imaginons que l'on utilise une interpolation linéaire par morceaux. Dans ce cas, on aura, dans un sous-intervalle $[\theta_k, \theta_{k+1}]$

$$\tilde{f}(\theta) = \frac{\theta_{k+1} - \theta}{\theta_{k+1} - \theta_k} f(\theta_k) + \frac{\theta - \theta_k}{\theta_{k+1} - \theta_k} f(\theta_{k+1})$$

et

$$\int_{\theta_k}^{\theta_{k+1}} \tilde{f}(\theta)\, d\theta = \frac{f(\theta_k) + f(\theta_{k+1})}{2} (\theta_{k+1} - \theta_k)$$

$$\int_{\theta_k}^{\theta_{k+1}} \tilde{f}(\theta) \cos n\theta\, d\theta = \frac{1}{n}(f(\theta_{k+1}) \sin n\theta_{k+1} - f(\theta_k)\sin n\theta_k) - \frac{1}{n} \int_{\theta_k}^{\theta_{k+1}} \tilde{f}'(\theta) \sin n\theta\, d\theta$$

$$= \frac{1}{n}(f(\theta_{k+1})\sin n\theta_{k+1} - f(\theta_k)\sin n\theta_k) + \frac{1}{n^2} \frac{f(\theta_{k+1})-f(\theta_k)}{\theta_{k+1}-\theta_k} (\cos n\theta_{k+1} - \cos n\theta_k)$$

$$\int_{\theta_k}^{\theta_{k+1}} \tilde{f}(\theta) \sin n\theta\, d\theta = -\frac{1}{n}(f(\theta_{k+1})\cos n\theta_{k+1} - f(\theta_k)\cos n\theta_k) + \frac{1}{n^2} \frac{f(\theta_{k+1})-f(\theta_k)}{\theta_{k+1}-\theta_k} (\sin n\theta_{k+1} - \sin n\theta_k)$$

Etant donné une subdivision de l'intervalle $[\,0, 2\pi\,]$ en N sous-intervalles $]\theta_0, \theta_1[,]\theta_1, \theta_2[, \ldots,]\theta_{N-1}, \theta_N[$, avec $\theta_0 = 0$ et $\theta_N = 2\pi$, on peut sommer sur ces sous-intervalles en distinguant éventuellement les valeurs à droite et à gauche des θ_k, ce qui permet de traiter les fonctions présentant des sauts. Dans ce cas, on a

$$a_n = O(1/n) \quad , \quad b_n = O(1/n) \quad .$$

Lorsque la fonction f est continue, avec $f(0)=f(2\pi)$, les termes en $(1/n)$ se compensent et

$$a_n = O(1/n^2) \quad , \quad b_n = O(1/n^2) \quad ,$$

si bien que la série calculée converge uniformément vers $\tilde{f}$. Tout dépend donc de l'erreur d'interpolation entre f et $\tilde{f}$.

Les critères de convergence sont:

- Convergence dans L^2 : $\quad (2a_o^2 + \sum_{n=1}^{M} (a_n^2 + b_n^2)) \longrightarrow \int_0^{2\pi} \tilde{f}^2\, d\theta$,

avec

$$\int_0^{2\pi} \tilde{f}^2\, d\theta = \sum_{k=0}^{N-1} \frac{\theta_{k+1} - \theta_k}{3} (f^2(\theta_k) + f(\theta_k)f(\theta_{k+1}) + f^2(\theta_{k+1}))$$

- Convergence éventuelle des dérivées:

$$\sum_{n=1}^{M} (a_n^2 + b_n^2) \longrightarrow \int_0^{2\pi} \tilde{f}'^2 d\theta = \sum_{k=0}^{N-1} \frac{(f(\theta_{k+1}) - f(\theta_k))^2}{\theta_{k+1} - \theta_k}$$

Il est donc aisé de mesurer l'erreur de troncature de la série.

Par ailleurs, on peut mesurer l'erreur en norme de l'interpolation $\|\Delta f\|_{L^2} = \|f - \tilde{f}\|_{L^2}$ comme suit. On suppose le découpage en sous-intervalles réalisé de façon que pour tout k,

$$\int_{\theta_k}^{\theta_{k+1}} f''^2 \, d\theta \quad ,$$

ce qui suppose que tout saut ou tout point anguleux soit situé à une frontière de sous-intervalle. Alors, dans chaque sous-intervalle, Δf est une fonction nulle aux deux extrémités et admet donc un développement en série de sinus: si $h_k = \theta_{k+1} - \theta_k$, on a

$$\Delta f = \sum_n A_n \sin \frac{n}{h_k}(\theta - \theta_k)$$

$$\Delta f'' = f'' = \sum_n \frac{n^2 \pi^2}{h_k^2} A_n \sin \frac{n}{h_k}(\theta - \theta_k)$$

Il en résulte

$$\int_{\theta_k}^{\theta_{k+1}} (\Delta f)^2 \, d\theta = \sum_n A_n^2 \quad ; \quad \int_{\theta_k}^{\theta_{k+1}} f''^2 \, d\theta = \sum_n \frac{n^4 \pi^4}{h_k^4} A_n^2$$

et

$$\int_{\theta_k}^{\theta_{k+1}} (\Delta f)^2 \, d\theta \leq \frac{h_k^4}{\pi^4} \int_{\theta_k}^{\theta_{k+1}} f''^2 \, d\theta$$

Sommant sur les intervalles, on obtient

$$\|\Delta f\|_{L^2} \leq \frac{1}{\pi^2} \left(\sum_k h_k^4 \int_{\theta_k}^{\theta_{k+1}} f''^2 \, d\theta \right)^{\frac{1}{2}} = O(h^2) \quad , \; h = \sup_k h_k$$

Alors, l'erreur sur un harmonique particulier vérifie

$$|\Delta a_n| = \left| \int_0^{2\pi} \Delta f \cos n\theta \, d\theta \right| \leq \|\Delta f\|_{L^2} \pi^{\frac{1}{2}}$$

et, de même

$$|\Delta b_n| \leq \|\Delta f\|_{L^2} \pi^{\frac{1}{2}}$$

$$|\Delta a_o| \leq \|\Delta f\|_{L^2} (2\pi)^{\frac{1}{2}}$$

Des formules plus élaborées que ci-dessus ont été développées par FILON [5] . Pour illustrer l'avantage des formules ci-dessus par rapport à une interpolation classique de f cos nθ ou f sin nθ, envisageons le cas d'un maillage uniforme $\theta_{k+1} = \theta_k + h$. Alors, la formule

classique des trapèzes donne, pour une fonction continue vérifiant $f(0)=f(2\pi)$,

$$(\hat{a}_n + i\,\hat{b}_n) = h \sum_{k=0}^{N-1} f(\theta_k)\, e^{in\theta_k}$$

Les formules proposées ci-dessus donnent

$$(a_n + i\,b_n) = \frac{1}{n^2} \sum_{k=0}^{N-1} (f(\theta_{k+1}) - f(\theta_k)) \frac{e^{in\theta_{k+1}} - e^{in\theta_k}}{h}$$

$$= \frac{1}{n^2 h} \sum_{k=0}^{N-1} f(\theta_k)\, e^{in\theta_k} (2 - e^{inh} - e^{-inh})$$

$$= \frac{2(1 - \cos nh)}{n^2 h} \sum_{k=0}^{N-1} f(\theta_k)\, e^{in\theta_k}$$

$$= \frac{\sin^2 \frac{nh}{2}}{(\frac{nh}{2})^2}\, h \sum_{k=0}^{N-1} f(\theta_k)\, e^{in\theta_k} \quad .$$

On constate que la différence réside dans un <u>facteur correctif</u> $\sin^2 \frac{nh}{2} / (\frac{nh}{2})^2$. Pour situer l'ordre de grandeur de ce facteur, supposons que $h = \pi/9 = 20°$, ce qui équivaut à 18 sous-intervalles. En fonction de n, on obtient:

n	facteur correctif
1	0,9899
2	0,9600
5	0,7706
10	0,3184
20	0,0096
50	0,0054
100	0,0032

Déjà pour n = 5, la différence est de 23%. A partir de n = 20, il n'y a plus de commune mesure entre les coefficients calculés par la formule des trapèzes et les coefficients améliorés. La correction apportée est donc bien nécessaire. Mais il y a plus: pour une fonction f, on a

$$c_n = a_n + i\,b_n = \frac{1}{\pi}(f,\, e^{in\theta}) \quad , \quad c_o = \frac{1}{2\pi}(f,\, 1) \ ,$$

où $(f, *)$ est la fonctionnelle définie à partir de f par

$$(f, \varphi) = \int_0^{2\pi} f\varphi \, d\theta \ .$$

La mesure de Dirac en θ_o est la fonctionnelle définie par $\delta_{\theta_o}(\varphi) = \varphi(\theta_o)$. Ses coefficients de Fourier sont donc

$$c_n(\delta_{\theta_o}) = \frac{1}{\pi}\ \delta_{\theta_o}(e^{in\theta}) = \frac{1}{\pi}\ e^{in\theta_o}\ ;\quad c_o = \frac{1}{2\pi}\ .$$

Dès lors, par application classique de la formule des trapèzes, on a

$$\hat{c}_n = c_n\Big(\sum_{k=0}^{N-1} f(\theta_k)\ \delta_{\theta_k}\Big)\ ,$$

c'est-à-dire que le développement, loin de converger vers f, converge au sens des distributions vers $\sum_k f(\theta_k)\ \delta_{\theta_k}$.

Exercice 1 - Soit P_{n+1} un polynôme de degré (n+1) sur $[a,b]$. Montrer que le polynôme Q_n de degré n qui l'interpole aux points de Gauss, zéros de φ_{n+1} sur $[a,b]$ réalise le minimum de $I((P_{n+1} - R_n)^2)$ sur l'ensemble des polynômes R_n de degré n.

Solution - Posant $R_n = a_o x^n + a_1 x^{n-1} + \ldots$, on obtient, en dérivant par rapport à a_o, a_1, ... , les conditions

$$I((P_{n+1} - R_n)x^k) = 0 \quad , \quad k = 1,\ldots,n \ ,$$

qui expriment que $P_{n+1} - R_n$ est un multiple de φ_{n+1} .

Exercice 2 - On considère l'intégrale $I = \int_0^1 \sqrt{1+2x}\, dx$.

a) On donne les points et poids de Gauss sur $[-1,+1]$:

$$x_o = -(3/5)^{\frac{1}{2}} \quad , \quad x_1 = 0 \quad , \quad x_2 = (3/5)^{\frac{1}{2}}$$

$$H_o = 5/9 \quad , \quad H_1 = 8/9 \quad , \quad H_2 = 5/9$$

Transformer ces points et poids pour qu'ils conviennent à l'intervalle $]0,1[$.

b) Etant donné la formule de l'erreur:

$$\Delta I(f) = 496{,}0 \cdot 10^{-9} h^7 f^{(6)}(\xi) \ ,$$

déterminer le nombre de chiffres nécessaires aux points et aux poids pour que l'erreur numérique soit inférieure au dixième de l'erreur d'algorithme. Déterminer simultanément la précision exigée du calcul de la fonction.

c) Faire le calcul dans ces circonstances.

d) Calculer la vraie valeur de I et vérifier l'évaluation de ΔI.

Solution

a) On effectue la transformation $y = \frac{1}{2}(1 + x)$, $\bar{H}_i = \frac{1}{2} H_i$, ce qui donne

$$y_o = \tfrac{1}{2}(1 - (3/5)^{\frac{1}{2}}) \quad , \quad y_1 = \tfrac{1}{2} \quad , \quad y_2 = \tfrac{1}{2}(1 + (3/5)^{\frac{1}{2}})$$

$$\bar{H}_o = 5/18 \quad , \quad \bar{H}_1 = 4/9 \quad , \quad \bar{H}_2 = 5/18 \ .$$

b) $f(x) = (1 + 2x)^{\frac{1}{2}}$

$$f^{(6)}(x) = \tfrac{1}{2}(-\tfrac{1}{2})(-\tfrac{3}{2})(-\tfrac{5}{2})(-\tfrac{7}{2})(-\tfrac{9}{2}) \cdot 2^6 \cdot (1+2x)^{-11/2}$$

$$= 945\ (1 + 2x)^{-11/2} .$$

Dans l'intervalle $[0,1]$, cette fonction évolue entre 2,453 et 945. Dès lors,

$$1{,}114 \cdot 10^{-6} \leq I(f) \leq 4{,}687 \cdot 10^{-4} .$$

L'erreur sur le calcul de

$$\tilde{I}(f) = \sum_i H_i f(x_i)$$

est

$$|\delta\tilde{I}| \leq \sum_i |H_i| \max_i |\delta f(x_i)| + \max_i |f(x_i)| . 3 \max_i |\delta H_i|$$

$$\max_i |\delta f(x_i)| + 3\sqrt{3} \max_i |\delta H_i| \quad .$$

En admettant une erreur identique sur f et sur les poids, on obtient la condition

$$\eta (1 + 3\sqrt{3}) \leq 1{,}114 \cdot 10^{-7} \ ,$$

soit

$$\eta \leq 1{,}797 \cdot 10^{-8} \quad .$$

Il faut donc faire les calculs avec huit chiffres après la virgule. Pour obtenir

$$|\delta f| \ 1{,}797 \cdot 10^{-8} \ ,$$

il faut que

$$|\delta f| = (1+2x)^{-\frac{1}{2}} \ |\delta x| \leq 1{,}797 \cdot 10^{-8} \ ,$$

ce qui aura lieu si

$$|\delta x| \leq 1{,}797 \cdot 10^{-8} ,$$

ce qui demande également 8 chiffres après la virgule. L'erreur d'évaluation de la racine doit être inférieure à cette valeur, ce que nous supposerons (Cela dépend de la machine utilisée).

c)

POINT	y	H	f	Hf
0	0,112 701 67	0,277 777 78	1,106 979 38	0,307 494 27
1	0,500 000 00	0,444 444 44	1,414 213 56	0,628 539 35
2	0,887 298 33	0,277 777 78	1,665 712 06	0,462 697 80
			I	1,398 731 42

d) Posant

$$\sqrt{1+2x} = 1 + y \ , \quad y \in [0, \sqrt{3} - 1] \ ,$$

on obtient

$$1+2x = (1+y)^2 \quad \text{et} \quad 2dx = 2(1+y)dy \ ,$$

d'où

$$\sqrt{1+2x}\, dx = (1+y)^2 \, dy$$

et

$$I = \int_0^{\sqrt{3}-1} (1+2y+y^2)dy = \sqrt{3} - \frac{1}{3} = 1{,}398\ 717\ 48 \ .$$

L'erreur vaut $1{,}394 \cdot 10^{-4}$. Elle est bien comprise entre les deux bornes prévues.

Exercice 3 - On considère l'intégrale

$$I = \int_0^1 \frac{dx}{1+x}$$

On donne les points et poids de Gauss:

$$x_1 = \tfrac{1}{2}(1 - (3/5)^{\frac{1}{2}}) \ , \ x_2 = \tfrac{1}{2} \ , \ x_3 = \tfrac{1}{2}(1 + (3/5)^{\frac{1}{2}})$$

$$H_1 = H_3 = 5/18 \ , \ H_2 = 8/18$$

On demande:

a) De calculer l'intégrale approchée.

b) De comparer le résultat à l'intégrale exacte.

c) De vérifier la formule

$$|\Delta I(f)| \leq 496.10^{-9}\, h^7 \sup_{[a,b]} |f^{(6)}(x)|$$

Solution

a) 0,69312

b) ln 2 = 0,69315

c) $f'(x) = (1+x)^{-1}$; $f''(x) = -(1+x)^{-2}$ .....

$$|f^{(6)}(x)| = |(-1)(-2)(-3)(-4)(-5)(1+x)^{-6}| \leq 120$$

$$|\Delta I(f)| \leq 496.10^{-9}.120 = 0,00006$$

Exercice 4 - Calculer la constante d'Euler

$$\gamma = - \int_0^\infty e^{-x} \ln x \, dx$$

Suggestion : $-\gamma = \int_0^1 e^{-x} \ln x \, dx + \int_1^\infty e^{-x} \ln x \, dx = I_1 + I_2$.

Pour calculer I_1 , on note que

$$e^{-x} = \sum_{k=0}^{\infty} (-1)^k \frac{x^k}{k!}$$

et que les fonctions

$$f_n = \sum_{k=0}^{n} (-1)^k \frac{x^k}{k!} \ln x$$

sont majorées en module par la fonction intégrable $e^{-x}.|\ln x|$, et $f_n \xrightarrow[pp]{} e^{-x} \ln x$. Dès lors, par le théorème de Lebesgue, on peut intégrer la série terme à terme. Or,

$$\int_0^1 \frac{x^k}{k!} \ln x \, dx = \left[\frac{x^{k+1}}{(k+1)!} \ln x \right]_0^1 - \int_0^1 \frac{x^{k+1}}{(k+1)!} \cdot \frac{1}{x} \, dx = - \frac{1}{(k+1)!(k+1)} \ ,$$

ce qui permet d'écrire

$$I_1 = \sum_{k=0} \frac{(-1)^{k+1}}{(k+1)(k+1)!} \ ,$$

série dont la convergence est très rapide. On obtient aisément

$$I_1 = -0,796599599\ldots .$$

Pour le calcul de I_2 , en faisant $x = 1 + y$, on obtient

$$\int_1^{\infty} e^{-x} \ln x \, dx = \int_0^{\infty} e^{-(1+y)} \ln(1+y) \, dy = (1/e) \int_0^{\infty} e^{-y} \ln(1+y) \, dy$$

La fonction $\ln(1+y)$ jouit des conditions voulues pour que les formules de Laguerre convergent en degré. Cependant, on ne peut majorer l'erreur, car

$$D^{2n+2} \ln(1+x) = (-1)^{2n+1}(2n+1)!(1+x)^{-2n-2},$$

ce qui donne la majoration

$$|\Delta I| \leq \frac{((n+1)!)^2}{(2n+2)!} (2n+1)! = \frac{((n+1)!)^2}{2n+2}$$

qui tend vers l'infini. Ceci résulte du fait que la fonction considérée n'a pas de développement en série entière sur $[0, \infty]$. Le seul recours est de comparer les intégrales numériques successives:

n	I	I/e	γ
1	0,61100 5058	0,22477 6199	0,57182 3400
4	0,59674 0086	0,21952 8409	0,57707 1190
7	0,59637 7846	0,21939 5149	0,57720 4450
14	0,59634 7721	0,21938 4066	0,57721 5533

La valeur exacte est [5]

$$= 0,57721\ 56649\ 0\ldots .$$

Exercice 5 - Calculer

$$I = \int_{-1}^{+1} \frac{\sqrt{1-x^2}}{2+x} \, dx$$

avec 6 décimales exactes.

Solution - On a

$$\frac{\sqrt{1-x^2}}{2+x} = \frac{1}{\sqrt{1-x^2}} \cdot \frac{1-x^2}{2+x} .$$

On peut donc utiliser une formule de Gauss-Tchébicheff. On a

$$f(x) = \frac{1-x^2}{2+x} = -x + \frac{3}{x+2} \quad ; \quad f'(x) = -1 - 3(x+2)^{-2} \quad ;$$

$$f''(x) = 3.2(x+2)^{-3} \quad ; \quad \ldots \quad ; \quad f^{(k)}(x) = (-1)^k (k-1)!(x+2)^{-k},$$

$$|f^{(k)}(x)| \leq (k-1)!$$

Les erreurs vérifient donc

$$|\Delta I| \leq \frac{\pi}{2^{2n+1}} \quad \frac{(2n+1)!}{(2n+2)!} = \frac{\pi}{(2n+2)2^{2n+1}}$$

n	0	2	4	6	7	8	9
$\lvert\Delta I\rvert \leq$	0,785	0,016	6.10^{-3}	27.10^{-6}	6.10^{-6}	$1,33.10^{-6}$	$299,6.10^{-9}$

La formule n=9 (à 0 points) convient donc. On fera les calculs avec 2 chiffres de réserve:

x	f(x)
-0,987 688 341	0,024 174 12
-0,891 006 524	0,185 850 84
-0,707 106 781	0,386 729 54
-0,453 990 500	0,513 510 83
-0,156 434 465	0,529 153 01
00,156 434 465	0,452 380 20
0,453 990 500	0,323 510 88
0,707 106 781	0,184 699 03
0,891 006 524	0,071 292 60
0,987 688 341	0,008 190 86

$\Sigma = 2,679\ 491\ 92 \qquad \tilde{I} = \frac{\pi}{10}\,\Sigma = 0,841\ 787\ 21 \approx 0,841\ 787$

Vérifions en calculant la valeur exacte de l'intégrale. On a, en posant $x = \cos\varphi$,

$$I = \int_0^{\pi} \frac{\sin^2\varphi}{2+\cos\varphi}\,d\varphi = \int_0^{\pi} \frac{1-\cos^2\varphi}{2+\cos\varphi}\,d\varphi = \int_0^{\pi} (2-\cos\varphi)\,d\varphi - 3\int_0^{\pi} \frac{d\varphi}{2+\cos\varphi}$$

$$= 2\pi - I_2$$

$$I_2 = 3\int_0^{\pi} \frac{1+\operatorname{tg}^2\frac{\varphi}{2}}{3+\operatorname{tg}^2\frac{\varphi}{2}}\,d\varphi = 6\int_{\varphi=0}^{\varphi=\pi} \frac{d(\operatorname{tg}\frac{\varphi}{2})}{3+\operatorname{tg}^2\frac{\varphi}{2}} = 2\int_0^{\infty} \frac{dy}{1+(y^2/3)}$$

$$= 2\sqrt{3}\int_0^{\infty} \frac{d(y/\sqrt{3})}{1+y^2/3} = 2\sqrt{3}\left[\operatorname{arctg}(y/\sqrt{3})\right]_0^{\infty} = \pi\sqrt{3},$$

d'où

$$I = \pi\,(2-\sqrt{3}) = 0,841\ 787\ 213\ .$$

On a donc très largement les 6 chiffres exacts après la virgule.

Exercice 6 - Calculer

$$I = \int_0^1 \frac{e^x\,dx}{\sqrt{1-x}}$$

avec 6 décimales exactes au moins.

Solution - En posant $\sqrt{1-x} = u$, on obtient

$$I = 2e\int_0^1 e^{-u^2}\,du\ ,$$

qui se calcule aisément par une formule de Gauss. Il est encore plus rapide de noter que

$$e^{-u^2} = \sum_{k=0} (-1)^k \frac{u^{2k}}{k!} ,$$

ce qui donne

$$I = 2e \sum_{k=0} (-1)^k \frac{1}{k!(2k+1)} ,$$

série très rapidement convergente: le reste, après N termes, est majoré par

$$\frac{1}{(N+1)!} \frac{1}{2N+2} .$$

Avec 13 termes, on obtient

$$I/e = 0,746\ 824\ 133,$$

tous chiffres stabilisés. Il en découle

$$I = 4,060\ 156\ 938.$$

TABLEAU 1 - FORMULES DE NEWTON-COTES FERMEES

$$\int_a^b f(x)dx = \frac{h}{A} \sum_{i=0}^{n} B_i \, f(x_i) - E_n \, h^{p+2} \, f^{(p+1)}(\xi)$$

$h = b - a$

n	A	B 0	B 1	B 2	B 3	B 4	B 5	B 6	B 7	B 8	B 9	B 10	p	E n	NOM
1	2	1	1										1	1/12	TRAPEZE
2	6	1	4	1									3	1/2880	Simpson
3	8	1	3	3	1								3	1/6480	Newton
4	90	7	32	12	32	7							5	516,70E-09	Bode-Villarceau
5	288	19	75	50	50	75	19						5	26,455E-09	-
6	840	41	216	27	272	27	216	41					7	637,90E-12	Hardy
7	17280	751	3577	1323	2989	2989	1323	3577	751				7	391,17E-12	-
8	28350	989	5888	-928	10496	-4540	10496	-928	5888	989			9	589,33E-15	-
9	89600	2857	15741	1080	19344	5778	5778	19344	1080	15741	2857		9	377,08E-15	-
10	598752	16067	106300	-48525	272400	-260550	427368	-260550	272400	-48525	106300	16067	11	411,83E-18	-

TABLEAU 2 -FORMULES DE NEWTON-COTES OUVERTES

$$\int_a^b f(x)dx = \frac{h}{A}\sum_{i=0}^{n} B_i f(x_i) + E_n h^{p+2} f^{(p+1)}(\xi)$$

$h = b - a$

n	A	B 0	B 1	B 2	B 3	B 4	B 5	B 6	p	E n	NOM
0	1	1							1	1/24	Poncelet (RECTANGLE)
1	2	1	1						1	1/36	-
2	3	2	-1	2					3	303,8E-06	-
3	24	11	1	1	11				3	211,1E-06	-
4	20	11	-14	26	-14	11			5	1,046E-06	-
5	1440	611	-453	562	562	-453	611		5	738,8E-09	-
6	945	460	-954	2196	-2459	2196	-954	460	7	2,079E-09	-

TABLEAU 3 - INTEGRATION DE GAUSS-LEGENDRE [5]

$$\int_{-1}^{+1} f(x)\,dx = \sum_{i=0}^{n} H_i\, f(x_i) + E_n\, h^{2n+3}\, f^{(2n+2)}(\xi) \qquad h = 2$$

$\pm x_i$	H_i
n=1	
0,57735 02691 89626	1,00000 00000 00000
n=2	
0,00000 00000 00000	0,88888 88888 88889
0,77459 66692 41483	0,55555 55555 55556
n=3	
0,33998 10435 84856	0,65214 51548 62546
0,86113 63115 94053	0,34785 48541 37454
n=4	
0,00000 00000 00000	0,56888 88888 88889
0,53846 93101 05683	0,47862 86704 99366
0,90617 98459 38664	0,23692 68850 56189
n=5	
0,23861 91860 83197	0,46791 39345 72691
0,66120 93864 66265	0,36076 15730 48139
0,93246 95142 03152	0,17132 44923 79170
n=6	
0,00000 00000 00000	0,41795 91836 73469
0,40584 51513 77397	0,38183 00505 05119
0,74153 11855 99394	0,27970 53914 89277
0,94910 79123 42759	0,12948 49661 68870
n=7	
0,18343 46424 95650	0,36268 37833 78362
0,52553 24099 16329	0,31370 66458 77887
0,79666 64774 13627	0,22238 10344 53374
0,96028 98564 97536	0,10122 85362 90376
n=8	
0,00000 00000 00000	0,33023 93550 01260
0,32425 34234 03809	0,31234 70770 40003
0,61337 14237 00590	0,26061 06964 02935
0,83603 11073 26636	0,18064 81606 94857
0,96816 02395 07626	0,08127 43883 61574
n=9	
0,14887 43389 81631	0,29552 42247 14753
0,43339 53941 29647	0,26926 67193 09996
0,67940 95682 99024	0,21908 63625 15982
0,86506 33666 88985	0,14945 13491 50581
0,97390 65285 17172	0,06667 13443 08688
n=11	
0,12523 34085 11469	0,24914 70458 13403
0,36783 14989 98180	0,23349 25365 38355
0,58731 79542 86617	0,20316 74267 23066
0,76990 26741 94305	0,16007 83285 43346
0,90411 72563 70475	0,10693 93259 95318
0,98156 06342 46719	0,04717 53363 86512

$\pm x_i$	H_i
n=15	
0,09501 25098 37637 440185	0,18945 06104 55068 496285
0,28160 35507 79258 913230	0,18260 34150 44923 588867
0,45801 67776 57227 386342	0,16915 65193 95002 538189
0,61787 62444 02643 748447	0,14959 59888 16576 732081
0,75540 44083 55003 033895	0,12462 89712 55533 872052
0,86563 12023 87831 743880	0,09515 85116 82492 784810
0,94457 50230 73232 576078	0,06225 35239 38647 892863
0,98940 09349 91649 932596	0,02715 24594 11754 094852
n=19	
0,07652 65211 33497 333755	0,15275 33871 30725 850698
0,22778 58511 41645 078080	0,14917 29864 72603 746788
0,37370 60887 15419 560673	0,14209 61093 18382 051329
0,51086 70019 50827 098004	0,13168 86384 49176 626898
0,63605 36807 26515 025453	0,11819 45319 61518 417312
0,74633 19064 60150 792614	0,10193 01198 17240 435037
0,83911 69718 22218 823395	0,08327 67415 76704 748725
0,91223 44282 51325 905868	0,06267 20483 34109 063570
0,96397 19272 77913 791268	0,04060 14298 00386 941331
0,99312 85991 85094 924786	0,01761 40071 39152 118312
n=23	
0,06405 68928 62605 626085	0,12793 81953 46752 156974
0,19111 88674 73616 309159	0,12583 74563 46828 296121
0,31504 26796 96163 374387	0,12167 04729 27803 391204
0,43379 35076 26045 138487	0,11550 56680 53725 601353
0,54542 14713 88839 535658	0,10744 42701 15965 634783
0,64809 36519 36975 569252	0,09761 86521 04113 888270
0,74012 41915 78554 364244	0,08619 01615 31953 275917
0,82000 19859 73902 921954	0,07334 64814 11080 305734
0,88641 55270 04401 034213	0,05929 85849 15436 780746
0,93827 45520 02732 758524	0,04427 74388 17419 806169
0,97472 85559 71309 498198	0,02853 13886 28933 663181
0,99518 72199 97021 360180	0,01234 12297 99987 199547

n	E_n	$E_n \cdot 2^{2n+3}$
0	41,67 E-03	333,3 E-03
1	231,5 E-06	7,407 E-03
2	496,0 E-09	63,49 E-06
3	562,4 E-12	287,9 E-09
4	394,5 E-15	807,9 E-12
5	188,1 E-18	1,541 E-12
6	64,92 E-21	2,127 E-15
7	16,97 E-24	2,225 E-18
8	3,478 E-27	1,823 E-21
9	573,4 E-33	1,203 E-24
11	8,816 E-39	295,8 E-33
15	318,7 E-57	2,738 E-45
19	1,573 E-72	3,459 E-60
23	1,581 E-90	890,0 E-78

TABLEAU 4 - INTEGRATION DE GAUSS-TCHEBICHEFF

$$\int_{-1}^{+1} \frac{f(x)}{\sqrt{1 - x^2}}\,dx = H_n \sum_{i=0}^{n} f(x_i) + E_n\, h^{2n+3} f^{(2n+2)}(\xi)$$

$h = 2$

n	$\pm x_i$	H_n	E_n	$E_n \cdot 2^{2n+3}$
0	0,000 000 000		98,17 E-03	0,7854
1	0,707 106 781	$\pi/2$	511,3 E-06	16,36 E-03
2	0,866 025 404	$\pi/3$	1,065 E-06	136,3 E-06
	0,000 000 000			
3	0,923 879 533	$\pi/4$	1,189 E-09	608,8 E-09
	0,382 683 433			
	0,951 056 516			
4	0,587 785 252	$\pi/5$	825,6 E-15	1,691 E-09
	0,000 000 000			
	0,965 925 826			
5	0,707 106 781	$\pi/6$	390,9 E-18	3,127 E-15
	0,258 819 045			
	0,974 927 912			
6	0,781 831 482	$\pi/7$	134,2 E-21	4,397 E-15
	0,433 883 739			
	0,000 000 000			
	0,980 785 280			
7	0,831 469 612	$\pi/8$	34,96 E-24	4,582 E-18
	0,555 570 233			
	0,195 090 322			
	0,984 807 753			
	0,866 025 404			
8	0,642 787 610	$\pi/9$	7,141 E-27	3,744 E-21
	0,342 020 143			
	0,000 000 000			
	0,987 688 341			
	0,891 006 524			
9	0,707 106 781	$\pi/10$	1,174 E-30	2,464 E-24
	0,453 990 500			
	0,156 434 465			

TABLEAU 5 - INTEGRATION DE GAUSS-LAGUERRE [5]

$$\int_0^{\infty} e^{-x} f(x)\,dx = \sum_{i=0}^{n} H_i f(x_i) + E_n f^{(2n+2)}(\xi)$$

x_i	H_i	E_n
n = 1		
0,58578 64376 27	8,53553 390593 E-01	0,1667
3,41421 35623 73	1,46446 609407 E-01	
n = 2		
0,41577 45567 83	7,11093 009929 E-01	
2,29428 03602 79	2,78517 733569 E-01	5,000 E-02
6,28994 50829 37	1,03892 565016 E-02	
n = 3		
0,32254 76896 19	6,03154 104342 E-01	
1,74576 11011 58	3,57418 692438 E-01	1,429 E-02
4,53662 02969 21	3,88879 085150 E-02	
9,39507 09123 01	5,39294 705561 E-04	
n = 4		
0,26356 03197 18	5,21755 610583 E-01	
1,41340 30591 07	3,98666 811083 E-01	
3,59642 57710 41	7,59424 496817 E-02	3,968 E-03
7,08581 00058 59	3,61175 867992 E-03	
12,64080 08442 76	2,33699 723858 E-05	
n = 5		
0,22284 66041 79	4,58964 673950 E-01	
1,18893 21016 73	4,17000 830772 E-01	
2,99273 63260 59	1,13373 382074 E-01	1,082 E-03
5,77514 35691 05	1,03991 974531 E-02	
9,83746 74183 83	2,61017 202815 E-04	
15,98287 39806 02	8,98547 906430 E-07	
n = 6		
0,19304 36765 60	4,09318 951701 E-01	
1,02666 48953 39	4,21831 277862 E-01	
2,56787 67499 51	1,47126 348658 E-01	
4,90035 30845 26	2,06335 144687 E-02	2,914 E-04
8,18215 34445 63	1,07401 014328 E-03	
12,73418 02917 98	1,58654 643486 E-05	
19,39572 78622 63	3,17031 547900 E-08	
n = 7		
0,17027 96323 05	3,69188 589342 E-01	
0,90370 17767 99	4,18786 780814 E-01	
2,25108 66298 66	1,75794 986637 E-01	
4,26670 01702 88	3,33434 922612 E-02	7,770 E-05
7,04590 54023 93	2,79453 623523 E-03	
10,75851 60101 81	9,07650 877336 E-05	
15,74067 86412 78	8,48574 671627 E-07	
22,86313 17368 89	1,04800 117487 E-09	

x_i	H_i	E_n
n = 8		
0,15232 22277 32	3,36126 421798 E-01	
0,80722 00227 42	4,11213 980424 E-01	
2,00513 51556 19	1,99287 525371 E-01	
3,78347 39733 31	4,74605 627657 E-02	
6,20495 67778 77	5,59962 661079 E-03	2,057 E-05
9,37298 52516 88	3,05249 767093 E-04	
13,46623 69110 92	6,59212 302608 E-06	
18,83359 77889 92	4,11076 933035 E-08	
26,37407 18909 27	3,29087 403035 E-11	
n = 9		
0,13779 34705 40	3,08441 115765 E-01	
0,72945 45495 03	4,01119 929155 E-01	
1,80834 29017 40	2,18068 287612 E-01	
3,40143 36978 55	6,20874 560987 E-02	
5,55249 61400 64	9,50151 697518 E-03	5,413 E-06
8,33015 27467 64	7,53008 388588 E-04	
11,84378 58379 00	2,82592 334960 E-05	
16,27925 78313 78	4,24931 398496 E-07	
21,99658 58119 81	1,83956 482398 E-09	
29,92069 70122 74	9,91182 721961 E-13	
n = 11		
0,11572 21173 58	2,64731 371055 E-01	
0,61175 74845 15	3,77759 275873 E-01	
1,51261 02697 76	2,44082 011320 E-01	
2,83375 13377 44	9,04492 222117 E-02	
4,59922 76394 18	2,01023 811546 E-02	
6,84452 54531 15	2,66397 354187 E-03	3,698 E-07
9,62131 68424 57	2,03231 592663 E-04	
13,00605 49933 06	8,36505 585682 E-06	
17,11685 51874 62	1,66849 387654 E-07	
22,15109 03793 97	1,34239 103052 E-09	
28,48796 72509 84	3,06160 163504 E-12	
37,09912 10444 67	8,14807 746743 E-16	
n = 14		
0,09330 78120 17	2,18234 885940 E-01	
0,49269 17403 02	3,42210 177923 E-01	
1,21559 54120 71	2,63027 577942 E-01	
2,26994 95262 04	1,26425 818106 E-01	
3,66762 27217 51	4,02068 649210 E-02	
5,42533 66274 14	8,56387 780361 E-03	
7,56591 62266 13	1,21243 614721 E-03	
10,12022 85680 19	1,11674 392344 E-04	6,447 E-09
13,13028 24821 76	6,45992 676202 E-06	
16,65440 77083 30	2,22631 690710 E-07	
20,77647 88994 49	4,22743 038498 E-09	
25,62389 42267 29	3,92189 726704 E-11	
31,40751 91697 54	1,45651 526407 E-13	
38,53068 33064 86	1,48302 705111 E-16	
48,02608 55726 86	1,60059 490621 E-20	

TABLEAU 6 - INTEGRATION DE GAUSS-HERMITE [5]

$$\int_{-\infty}^{+\infty} e^{-x^2} f(x)\,dx = \sum_{i=0}^{n} H_i f(x_i) + E_n f^{(2n+2)}(\xi)$$

$+x_i$ / $-x_i$	H_i	E_n
	n = 1	
0,70710 67811 86548	8,86226 92545 28 E-01	3,693 E-02
	n = 2	
0,00000 00000 00000	1,18163 59006 04 E 00	1,846 E-03
1,22474 48713 91589	2,95408 97515 09 E-01	
	n = 3	
0,52464 76232 75290	8,04914 09000 55 E-01	6,594 E-05
1,65068 01238 85785	8,13128 35447 25 E-02	
	n = 4	
0,00000 00000 00000	9,45308 72048 29 E-01	
0,95857 24646 13819	3,93619 32315 22 E-01	1,832 E-06
2,02018 28704 56086	1,99532 42059 05 E-02	
	n = 5	
0,43607 74119 27617	7,24629 59522 44 E-01	
1,33584 90740 13697	1,57067 32032 29 E-01	4,163 E-08
2,35060 49736 74492	4,53000 99055 09 E-03	
	n = 6	
0,00000 00000 00000	8,10264 61755 68 E-01	
0,81628 78828 58965	4,25607 25261 01 E-01	8,005 E-10
1,67355 16287 67471	5,45155 82819 13 E-02	
2,65196 13568 35233	9,71781 24509 95 E-04	
	n = 7	
0,38118 69902 07322	6,61147 01255 82 E-01	
1,15719 37124 46780	2,07802 32581 49 E-01	1,334 E-11
1,98165 67566 95843	1,70779 83007 41 E-02	
2,93063 74202 57244	1,99604 07221 14 E-04	
	n = 8	
0,00000 00000 00000	7,20235 21650 61 E-01	
0,72355 10187 52838	4,32651 55900 26 E-01	
1,46855 32892 16668	8,84745 27394 38 E-02	1,962 E-13
2,26658 05845 31843	4,94362 42755 37 E-03	
3,19099 32017 81528	3,96069 77263 26 E-05	

$+x_i$ / $-x_i$	H_i	E_n
	n = 9	
0,34290 13272 23705	6,10862 63373 53 E-01	
1,03661 08297 89514	2,40138 61108 23 E-01	
1,75668 36492 99882	3,38743 94455 48 E-02	2,582 E-15
2,53273 16742 32790	1,34364 57467 81 E-03	
3,43615 91188 37738	7,64043 28552 33 E-06	
	n = 11	
0,31424 03762 54359	5,70135 23626 25 E-01	
0,94778 83912 40164	2,60492 31026 42 E-01	
1,59768 26351 52605	5,16079 85615 88 E-02	3,341 E-19
2,27950 70805 01060	3,90539 05846 29 E-03	
3,02063 70251 20890	8,57368 70435 88 E-05	
3,88972 48978 69782	2,65855 16843 56 E-07	
	n = 15	
0,27348 10461 3815	5,07929 47901 66 E-01	
0,82295 14491 4466	2,80647 45852 85 E-01	
1,38025 85391 9888	8,38100 41398 99 E-02	
1,95178 79909 1625	1,28803 11535 51 E-02	2,151 E-27
2,54620 21578 4748	9,32284 00862 42 E-04	
3,17699 91619 7996	2,71186 00925 38 E-05	
3,86944 79048 6012	2,32098 08448 65 E-07	
4,68873 89393 0582	2,65480 74740 11 E-10	
	n = 19	
0,24534 07083 009	4,62243 66960 06 E-01	
0,73747 37285 454	2,86675 50536 28 E-01	
1,23407 62153 953	1,09017 20602 00 E-01	
1,73853 77121 166	2,48105 20887 46 E-02	
2,25497 40020 893	3,24377 33422 38 E-03	5,040 E-36
2,78880 60584 281	2,28338 63601 63 E-04	
3,34785 45673 832	7,80255 64785 32 E-06	
3,94476 40401 156	1,08606 93707 69 E-07	
4,60368 24495 507	4,39934 09922 73 E-10	
5,38748 08900 112	2,22939 36455 34 E-13	

TABLEAU 7 - INTEGRATION DE LOBATTO [5]

$$\int_{-1}^{+1} f(x)\,dx = \sum_{i=0}^{n} H_i\, f(x_i) - E_n\, h^{2n+1}\, f^{(2n)}(\xi) \qquad h = 2$$

n	± x - i	H i	E n
2	1 0	1/3 4/3	347,2 E-06
3	1 0,44721 360	1/6 5/6	661,4 E-09
4	1 0,65465 367 0	0,10000 000 0,54444 444 0,71111 111	703,0 E-12
5	1 0,76505 532 0,28523 152	0,06666 667 0,37847 496 0,55485 838	473,4 E-15
6	1 0,83022 390 0,46884 879 0	0,04761 904 0,27682 604 0,43174 538 0,48761 904	219,4 E-18
7	1 0,87174 015 0,59170 018 0,20929 922	0,03571 428 0,21070 422 0,34112 270 0,41275 880	74,20 E-21
8	1 0,89975 79954 0,67718 62795 0,36311 74638 0	0,02777 77778 0,16549 53616 0,27453 87126 0,34642 85110 0,37151 92744	19,10 E-24
9	1 0,91953 39082 0,73877 38651 0,47792 49498 0,16527 89577	0,02222 22222 0,13330 59908 0,22488 93420 0,29204 26836 0,32753 97612	3,864 E-27

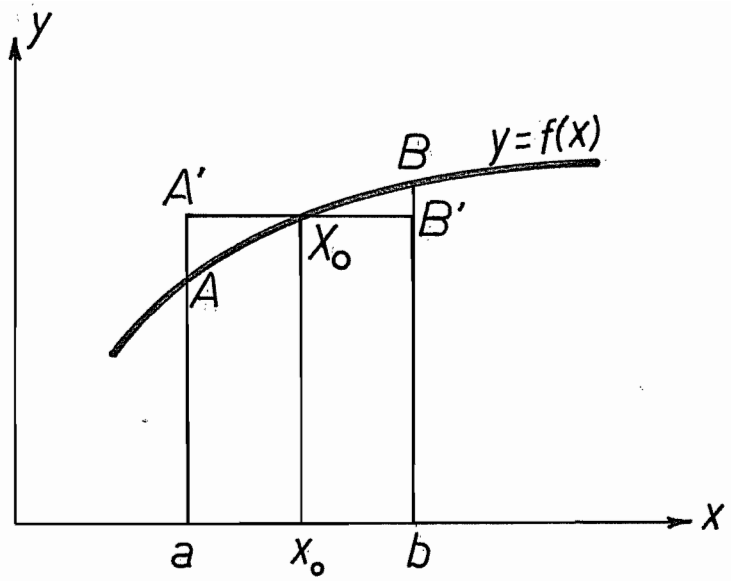


Fig.1

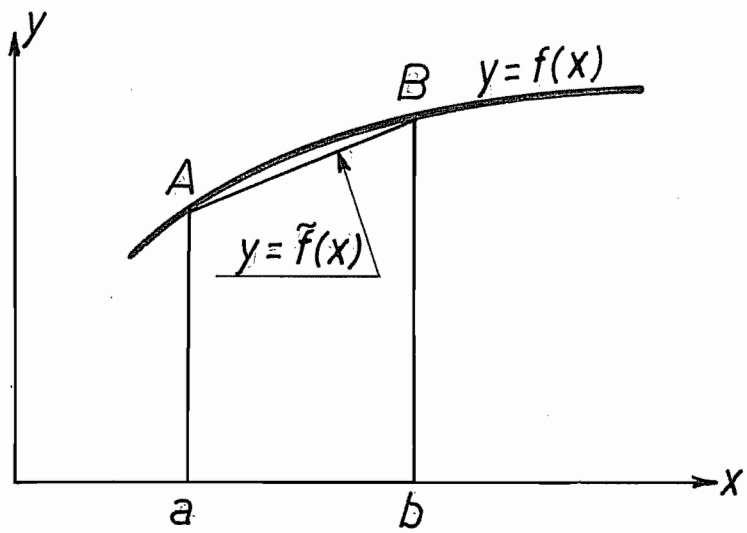


fig. 2

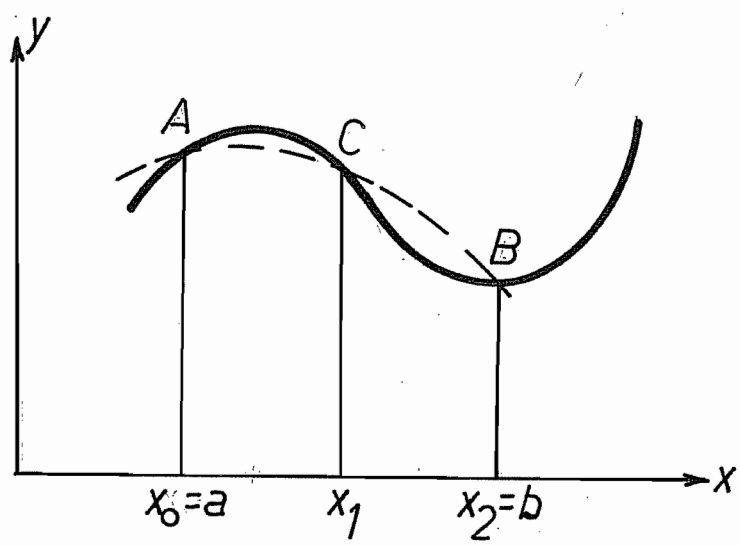


fig. 3

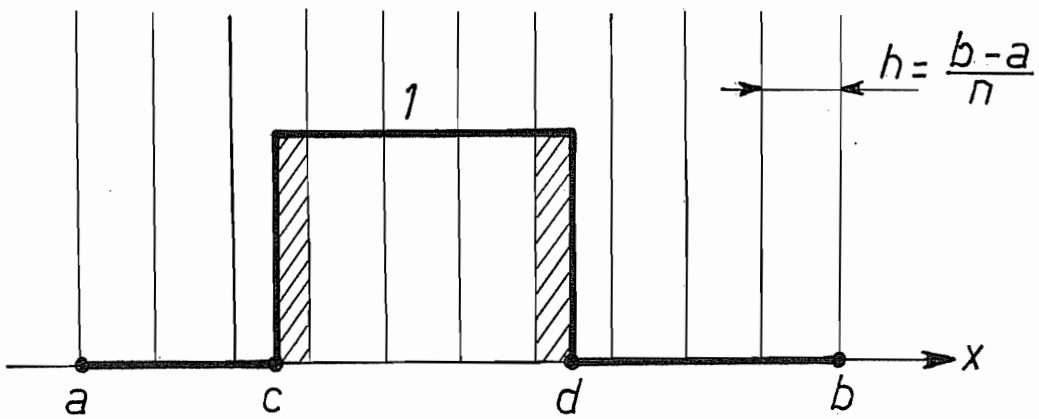


fig. 4

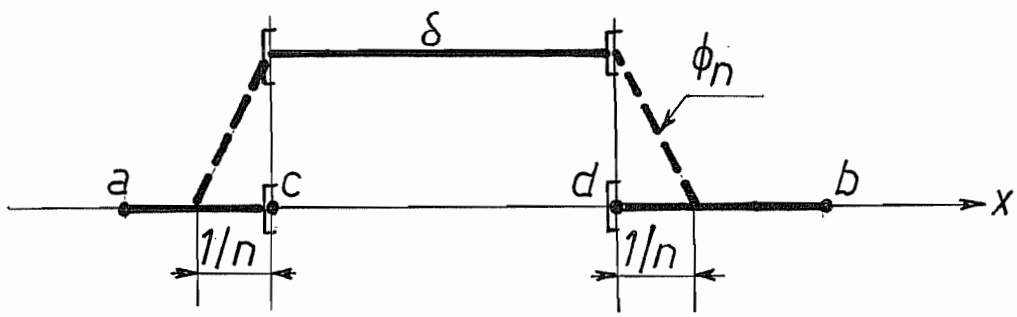


fig. 5

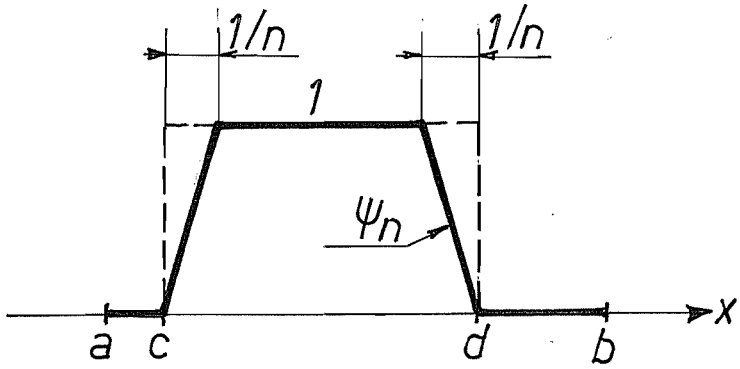
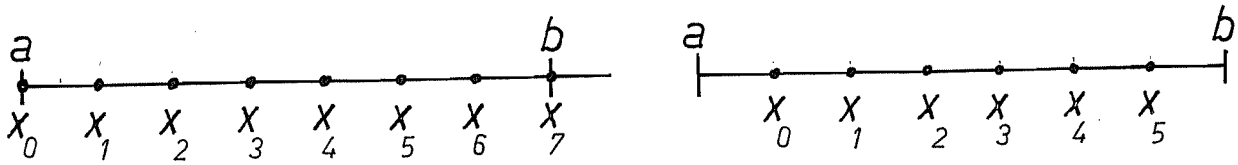


fig. 6



formule fermée

formule ouverte

fig. 7

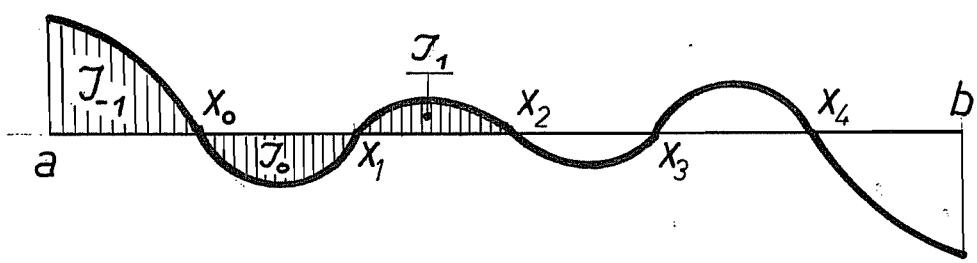
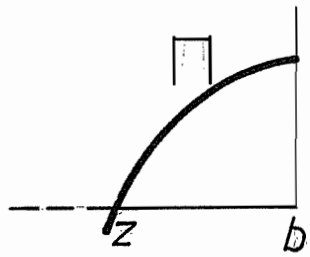
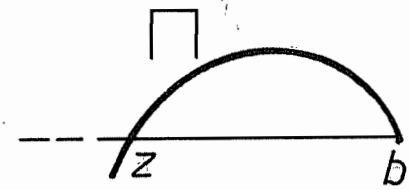


fig. 8



formule ouverte



formule fermée

fig.9

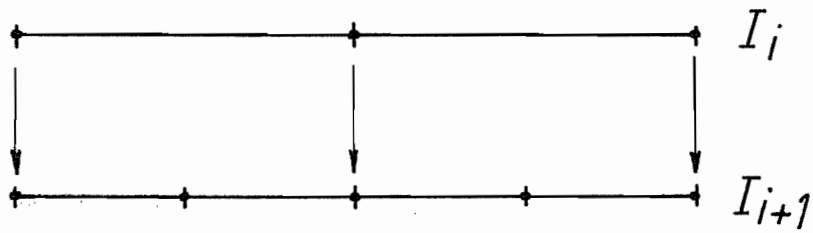


fig.10

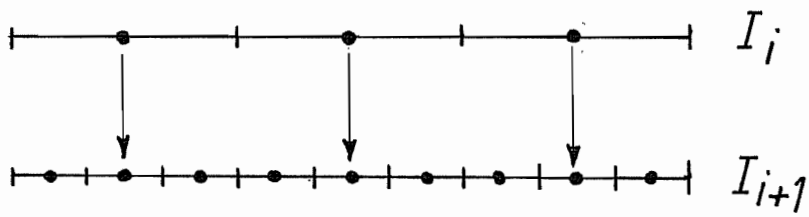


fig.11

R E S O L U T I O N D E S E Q U A T I O N S A L G E B R I Q U E S

Les équations algébriques peuvent être traitées par un certain nombre d'algorithmes simples spécialement appropriés.

1. REPRESENTATION DES POLYNOMES EN MACHINE. CALCULS SUR LES POLYNOMES

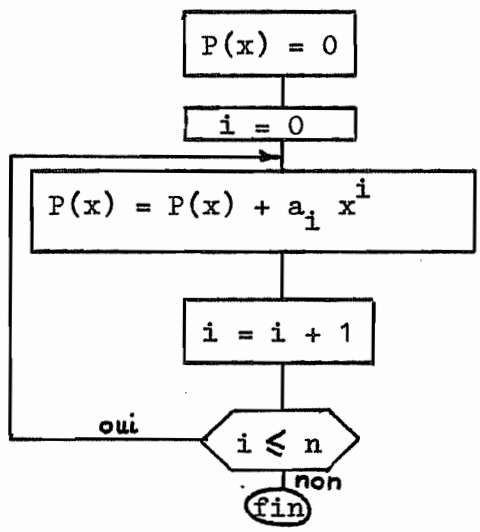
1.1 - Le polynôme

$$P_n(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_o$$

peut être représenté en machine par la suite des (n+1) nombres

$$(a_o, a_1, \ldots, a_n) .$$

Nous ne considérerons ici que le cas des polynômes à coefficients réels.
Pour calculer la valeur de $P_n(x)$, il ne faut pas faire comme suit:



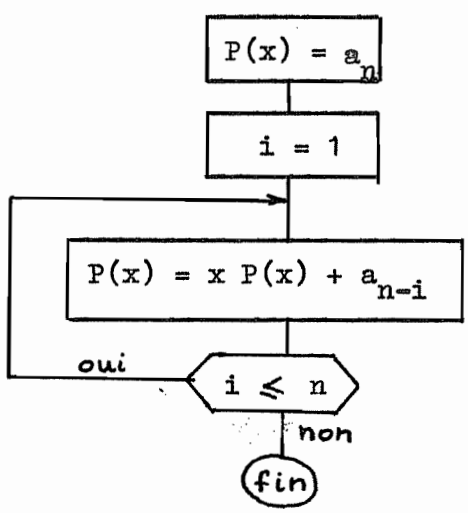
car cet algorithme nécessite:

- (n + 1) exponentiations (avec des problèmes éventuels si x = 0)
- (n + 1) produits
- (n + 1) additions.

La seule méthode efficace consiste à calculer

$$P(x) = ((a_n x + a_{n-1}) x + a_{n-2}) x + \ldots ,$$

soit



On notera que la combinaison linéaire de deux polynômes de même degré s'obtient simplement en combinant leurs tableaux de coeffidients.

1.2 - Division par $(x - p)$: algorithme de HORNER

Soit à diviser le polynôme

$$P_n(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_o$$

par $(x - p)$. On a

$$P_n(x) = (x - p)\, Q_{n-1}(x) + R ,$$

où Q_{n-1} est le polynôme quotient, et R, le reste, qui est un nombre. Nous chechons les coefficients de Q_{n-1} :

$$Q_{n-1}(x) = b_{n-1} x^{n-1} + b_{n-2} x^{n-2} + \ldots + b_o .$$

On a

$$P_n(x) = b_{n-1} x^n + (b_{n-2} - p\, b_{n-1}) x^{n-1} + \ldots + (b_o - p\, b_1) x + R - p\, b_o ,$$

ce qui donne

$$\begin{cases} b_{n-1} = a_n \\ b_{n-2} - p\, b_{n-1} = a_{n-1} & \text{soit} \quad b_{n-2} = a_{n-1} + p\, b_{n-1} \\ \ldots\ldots\ldots\ldots \\ b_o - p\, b_1 = a_1 & \text{soit} \quad b_o = a_1 + p\, b_1 \\ R - p\, b_o = a_o & \text{soit} \quad R = a_o + p\, b_o \end{cases}$$

C'est l'algorithme de Horner.

1.3 - Division par un trinôme du second degré $(x^2 - p x - q)$

On a ici

$$P_n(x) = (x^2 - px - q)\, Q_{n-2}(x) + \underbrace{Rx + S}_{\text{reste}}$$

soit explicitement

$$a_n x^n + a_{n-1} x^{n-1} + \ldots + a_o = (b_{n-2} x^{n-2} + \ldots + b_o)(x^2 - px - q) + Rx + S$$

ce qui donne

$$\begin{cases} a_n = b_{n-2} & \text{soit} \quad b_{n-2} = a_n \\ a_{n-1} = b_{n-3} - p\, b_{n-2} & \text{soit} \quad b_{n-3} = a_{n-1} + p b_{n-2} \\ a_{n-2} = b_{n-4} - p\, b_{n-3} - q b_{n-2} & \text{soit} \quad b_{n-4} = a_{n-2} + p b_{n-3} + q b_{n-2} \\ \ldots\ldots\ldots \\ a_2 = b_o - p b_1 - q b_2 & \text{soit} \quad b_o = a_2 + p b_1 + q b_2 \end{cases}$$

$$\left\{\begin{array}{ll} a_1 = R - pb_o - qb_1 & \text{soit} \quad R = a_1 + pb_o + qb_1 \\ a_o = S - qb_o & \text{soit} \quad S = a_o + qb_o \end{array}\right.$$

2. LOCALISATION DES ZEROS

Il existe un certain nombre de théorèmes permettant de localiser les zéros d'un polynôme. Nous en citerons trois qui sont particulièrement simples et efficaces.

2.1 - Borne supérieure du module

Tout zéro x_k (réel ou complexe) du polynôme

$$P_n(x) = a_n x^n + \ldots + a_o$$

vérifie la relation

$$|x_k| < 1 + \frac{\alpha}{|a_n|} \quad , \quad \alpha = \sup(|a_{n-1}|, \ldots, |a_o|)$$

Démonstration : Si $|x_k| \leqslant 1$, le théorème est satisfait. Examinons donc le comportement du polynôme pour $|x| > 1$. On a successivement

$$\begin{aligned} |P_n(x)| &= |a_n x^n + \ldots + a_o| \\ &\geqslant |a_n x^n| - |a_{n-1}|\,|x_{n-1}| - \ldots - |a_o| \\ &\geqslant |a_n|\,|x^n| - \sup(|a_{n-1}|, \ldots, |a_o|)(|x^{n-1}| + \ldots + 1) \\ &\geqslant |a_n|\,|x^n| - \frac{|x^n| - 1}{|x| - 1}\alpha \\ &> |a_n|\,|x^n| - \frac{|x|^n}{|x| - 1}\alpha \end{aligned}$$

soit

$$|P_n(x)| > \left\{|a_n| - \frac{\alpha}{|x| - 1}\right\} |x|^n$$

Pour

$$|a_n| \geqslant \frac{\alpha}{|x| - 1} \quad ,$$

on a donc $|P_n(x)| > 0$. Dès lors, on ne peut avoir $|P_n(x)| = 0$ que si

$$|a_n| < \frac{\alpha}{|x| - 1} \quad ,$$

soit si

$$|x| - 1 < \frac{\alpha}{|a_n|} \quad ,$$

comme annoncé.

2.2 - Borne inférieure du module

Tout zéro x_k du polynôme $P_n(x) = a_n x^n + \ldots + a_o$ vérifie la relation

$$|x_k| > \frac{|a_o|}{|a_o| + \beta} \quad , \quad \beta = \sup(|a_n|, \ldots, |a_1|)$$

Démonstration : Posant $x = \frac{1}{y}$, on obtient

$$y^n P_n(1/y) = a_o y^n + \ldots + a_n = Q_n(y) .$$

Les racines de ce polynôme vérifient donc, en vertu du théorème précédent,

$$|y_k| < 1 + \frac{\sup(|a_1|, \ldots, |a_n|)}{|a_o|} = \frac{|a_o| + \beta}{|a_o|} .$$

Or, ce sont les inverses des racines de P_n . Donc,

$$|x_k| > \frac{|a_o|}{|a_o| + \beta} .$$

2.3 - Localisation d'un zéro proche d'un point

Voici encore un résultat intéressant: Pour tout polynôme $P_n(z)$ de degré n, et pour tout nombre complexe z_o, une racine au moins de P_n se trouve dans le disque

$$|z - z_o| \leq \sqrt[n]{\left|\frac{P_n(z_o)}{a_n}\right|}$$

Remarquons d'abord que $P_n(z) = a_n(z - z_1)\ldots(z - z_n)$, ce qui implique $a_o = a_n z_1 \ldots z_n (-1)^n$. Dès lors, la racine de module minimal vérifie

$$\inf_k |z_k| \leq \sqrt[n]{\left|\frac{a_o}{a_n}\right|} = \sqrt[n]{\left|\frac{P_n(0)}{a_n}\right|} ,$$

ce qui démontre le théorème en $z_o = 0$. Mais on peut écrire le polynôme $P_n(z)$ sous la forme

$$P_n(z) = a_n(z - z_o)^n + a^*_{n-1}(z - z_o)^{n-1} + \ldots + a^*_o ,$$

ce qui donnera de même

$$\inf_k |z_k - z_o| \leq \sqrt[n]{\left|\frac{a^*_o}{a_n}\right|} = \sqrt[n]{\left|\frac{P_n(z_o)}{a_n}\right|} ,$$

comme annoncé.

3. METHODE DE GRAEFFE

Le principe de cette méthode est que le polynôme $P_n(x)$ a la forme

$$P_n(x) = a_n(x - x_1) \ldots (x - x_n) = a_n\Big(x^n - x^{n-1} \sum_{i=1}^{n} x_i +$$

$$+ x^{n-2} \sum_{i \neq j} x_i x_j + \ldots + (-1)^n \sum_{i=1}^{n} x_i\Big)$$

c'est-à-dire que

$$\left\{ \frac{a_{n-1}}{a_n} = - \sum_{i=1}^{n} x_i \right.$$

$$\left\{\begin{array}{l}
\dfrac{a_{n-2}}{a_n} = (-1)^2 \sum_{i\neq j} x_i x_j \\[2ex]
\dfrac{a_{n-3}}{a_n} = (-1)^3 \sum_{\substack{i,j,k \\ \text{différents}}} x_i x_j x_k \\[2ex]
\cdots\cdots\cdots \\[2ex]
\dfrac{a_{n-p}}{a_n} = (-1)^p \sum_{\substack{i_1,\ \ldots,i_p \\ \text{différents}}} x_{i_1} x_{i_2} x_{i_3} \cdots x_{i_p} \\[2ex]
\cdots\cdots\cdots\cdots \\[2ex]
\dfrac{a_o}{a_n} = (-1)^n x_1 \cdots x_n
\end{array}\right.$$

Supposons que les racines du polynôme soient de modules très différents en ordre de grandeur:

$$|x_1| \gg |x_2| \gg |x_3| \gg \cdots\cdots \gg |x_n|$$

lors, les relations précédentes admettent les approximations

$$\left\{\begin{array}{l}
\dfrac{a_{n-1}}{a_n} = -\, x_1(1 + 0(x_2/x_1)) \\[2ex]
\dfrac{a_{n-2}}{a_n} = x_1 x_2(1 + 0(x_3/x_2)) \\[2ex]
\dfrac{a_{n-3}}{a_n} = -\, x_1 x_2 x_3(1 + 0(x_4/x_3)) \\[2ex]
\cdots\cdots\cdots\cdots \\[2ex]
\dfrac{a_{n-p}}{a_n} = (-1)^p x_1 \cdots x_p(1 + 0(x_{p+1}/x_p)) \\[2ex]
\cdots\cdots\cdots\cdots \\[2ex]
\dfrac{a_o}{a_n} = (-1)^n x_1 \cdots x_n
\end{array}\right.$$

dont on déduit

$$x_1 \approx -\frac{a_{n-1}}{a_n}\ ,\quad x_2 \approx -\frac{a_{n-2}}{a_{n-1}}\ ,\quad x_3 \approx -\frac{a_{n-3}}{a_{n-2}}\ ,\quad \ldots\ ,\quad x_n \approx -\frac{a_{n-1}}{a_n}\ ,$$

avec des erreurs relatives dont les ordres de grandeur sont respectivement

$$|x_2/x_1|\ ,\ |x_2/x_1| + |x_3/x_2|,\ |x_3/x_2| + |x_4/x_3|,\ \ldots,\ |x_{n-1}|/|x_n|\ .$$

Malheureusement, même si toutes les racines d'un polynôme sont distinctes, il est rare qu'elles soient aussi bien séparées. Par contre, si l'on parvient à construire un polynôme ayant comme racines les carrés

des x_i, les rapports x_i^2/x_{i+1}^2 seront plus petits. Or, c'est chose simple, car le polynôme $P_n(-x)$ a pour expression

$$P_n(-x) = a_n(-x - x_1) \ldots (-x - x_n) \quad ,$$

si bien que le produit

$$P_n^{(1)}(x^2) = (-1)^n P_n(x) P_n(-x) = a_n^2(x^2 - x_1^2) \ldots (x^2 - x_n^2)$$

admet les x_i^2 comme racines. Il suffit donc de calculer ses coefficients. On a

$$P_n(x^2) = b_n x^{2n} + b_{n-1} x^{2n-2} + \ldots + b_o$$

$$= (a_n x^n + a_{n-1} x^{n-1} + \ldots + a_o)(a_n x^n - a_{n-1} x^{n-1} + a_{n-2} x^{n-2} + \ldots + (-1)^n a_o$$

ce qui permet de calculer b_p par identification:

$$b_p = (-1)^n \left\{ a_p^2 + 2 \sum_{\{k>o \mid p+k \leq n \text{ et } p-k \geq 0\}} (-1)^k a_{p+k}\, a_{p-k} \right\}$$

Une meilleure approximation des racines est alors

$$x_1^2 = - \frac{b_{n-1}}{b_n} \quad , \quad x_2^2 = - \frac{b_{n-2}}{b_{n-1}} \quad , \quad \ldots \quad , \quad x_n^2 = - \frac{b_o}{b_1} \quad .$$

Après un certain nombre de mises au carré successives, les racines des nouveaux polynômes s'écartant de plus en plus, le calcul converge pour autant que toutes les racines diffèrent en module.

On notera que dans ce procédé, les grands coefficients croissent de plus en plus, ce qui nécessite de <u>diviser à chaque itération tous les coefficients par le plus grand d'entre eux, afin d'éviter le dépassement de capacité</u>. Même ce faisant, on est limité par le fait qu'au bout d'un certain nombre d'itérations, les plus petits coefficients deviennent, au sens de la machine, assimilables à 0. ("underflow" sur certaines machines).

Lorsque le processus est suffisamment avancé, on a

$$\left| \frac{a_{p+1}}{a_p} \cdot \frac{a_{p-1}}{a_p} \right| \approx \left| \frac{x_{n-p+1}}{x_{n-p}} \right| \ll 1$$

$$\left| \frac{a_{p+2} \cdot a_{p-2}}{a_p^2} \right| \approx \left| \frac{x_{n-p+2}}{x_{n-p+1}} \cdot \frac{x_{n-p+1}}{x_{n-p}} \right| \ll 1$$

etc... , c'est-à-dire que les doubles produits deviennent négligeables. Le critère de convergence est donc la <u>petitesse des doubles produits</u>.

Il existe un certain nombre de cas où la méthode de Graeffe ne converge

pas:

a) une racine est multiple

b) deux racines ont le même module

c) deux racines sont complexes conjuguées.

Il est possible d'aménager cette méthode pour certains de ces cas, mais c'est au prix d'une plus grande complication ! 7 !.

On obtient finalement le module des racines. On obtient les signes par des essais.

4. EQUATIONS RECURRENTES - METHODE DE BERNOULLI

Une méthode très simple de recherche des zéros des polynômes, due à Bernoulli, est fondée sur la théorie des équations de récurrence à coefficients constants. On appelle équation récurrente à n pas une équation de la forme

$$a_n y(m) + a_{n-1} y(m-1) + \ldots + a_o y(m-n) = 0 \qquad (1)$$

avec $a_o \neq 0$, $a_n \neq 0$. Les solutions de ces équations sont des suites.

4.1 - Solutions linéairement indépendantes

Des solutions $y_1(m), \ldots, y_k(m)$ de l'équation (1) sont linéairement dépendantes s'il existe k nombres $\lambda_1, \ldots, \lambda_k$ non tous nuls tels que, pour tout m,

$$\lambda_1 y_1(m) + \lambda_2 y_2(m) + \ldots + \lambda_k y_k(m) = 0 . \qquad (2)$$

Dans le cas contraire, les k solutions sont linéairement indépendantes.

En fait, il suffit de vérifier la dépendance linéaire de k solutions d'une équation récurrente à n pas sur n points consécutifs. En effet, si les solutions sont linéairement dépendantes, la condition (2) est certainement vérifiée sur n points consécutifs. A l'inverse, considérons k solutions vérifiant en un m donné

$$\begin{cases} \lambda_1 y_1(m-1) + \ldots + \lambda_k y_k(m-n) = 0 \\ \quad \ldots\ldots \\ \lambda_1 y_1(m-n) + \ldots\ldots + \lambda_k y_k(m-n) = 0 \end{cases}$$

On a d'une part

$$\lambda_1 y_1(m) + \ldots + \lambda_k y_k(m) =$$

$$-\frac{1}{a_n}\left\{ a_{n-1}\Big(\sum_{i=1}^{k} \lambda_i y_i(m-1)\Big) + a_{n-2}\Big(\sum_{i=1}^{k} \lambda_i y_i(m-2)\Big) + \ldots + a_o\Big(\sum_{i=1}^{k} \lambda_i y_i(m-n)\Big)\right\} = 0$$

et, d'autre part,

$$\lambda_1 y_1(m-n-1) + \ldots + \lambda_k y_k(m-n-1) =$$
$$-\frac{1}{a_o}\left\{a_n\left(\sum_{i=1}^{k}\lambda_i y_i(m-1)\right) + a_{n-1}\left(\sum_{i=1}^{k}\lambda_i y_i(m-2)\right) + \ldots + a_1\left(\sum_{i=1}^{k}\lambda_i y_i(m-n)\right)\right\} = 0$$

ce qui permet de conclure par récurrence.

4.2 - Solution générale de l'équation homogène

Supposons que l'on connaisse n solutions indépendantes $y_1(m), \ldots, y_n(m)$ de l'équation récurrente (1). Alors, toute solution $y_o(m)$ de cette équation est de la forme

$$y_o(m) = \lambda_1 y_1(m) + \ldots + \lambda_n y_n(m) .$$

En effet, on a nécessairement

$$\begin{cases} a_n y_1(m) + a_{n-1} y_1(m-1) + \ldots + a_o y_1(m-n) = 0 \\ \ldots\ldots\ldots\ldots \\ a_n y_n(m) + a_{n-1} y_n(m-1) + \ldots + a_o y_n(m-n) = 0 \\ a_n y_o(m) + a_{n-1} y_o(m-1) + \ldots + a_o y_o(m-n) = 0 \end{cases}$$

On peut considérer qu'il s'agit d'un système homogène de (n+1) équations portant sur les (n+1) inconnues $a_n, \ldots, a_o$. Pour que ce système soit compatible, il faut que le déterminant de la matrice

$$\begin{bmatrix} y_1(m) & y_1(m-1) & \ldots\ldots & y_1(m-n) \\ y_2(m) & y_2(m-1) & \ldots\ldots & y_2(m-n) \\ \ldots\ldots\ldots\ldots & & & \\ y_n(m) & y_n(m-1) & \ldots\ldots & y_n(m-n) \\ y_o(m) & y_o(m-1) & \ldots\ldots & y_n(m-n) \end{bmatrix}$$

soit nul, ce qui implique l'existence d'une relation de la forme

$$\lambda_1 y_1(s) + \lambda_2 y_2(s) + \ldots + \lambda_n y_n(s) + \lambda_o y_o(s) = 0$$

aux (n+1) points s = m, m-1, ..., (m-n). Cette relation subsiste donc partout (voir section 4.1). On ne peut avoir $\lambda_o = 0$, car cela contredirait la condition d'indépendance des solutions $y_1, \ldots, y_n$. Dès lors, pour tout m, on a

$$y_o(m) = -\frac{\lambda_1}{\lambda_o}\, y_1(m) - \ldots\ldots - \frac{\lambda_n}{\lambda_o}\, y_n(m) \quad .$$

4.3 - Conditions de départ. Unicité de la solution

Montrons que si l'on se donne les valeurs $y(1) = \bar{y}(1), \ldots, y(n) = \bar{y}(n)$, l'équation récurrente (1) admet une solution unique. A cette fin, considérons deux solutions $y_1(m)$ et $y_2(m)$ vérifiant ces conditions de départ. On a alors

$$y_1(1) - y_2(1) = 0, \ \ldots\ , y_1(n) - y_2(n) = 0 \ ,$$

d'où

$$y_1(n+1) - y_2(n+1) = -\frac{1}{a_n}\left\{a_{n-1}(y_1(n)-y_2(n)) + \ldots + a_o(y_1(1)-y_2(1))\right\} = 0$$

et ainsi de suite, ce qui implique $y_1(m) = y_2(m)$.

4.4 - Solutions particulières des équations à coefficients constants

Soient $x_1, \ldots, x_n$ les racines de l'équation polynomiale

$$P_n(x) = a_n x^n + \ldots + a_o = 0 \ ,$$

et supposons d'abord qu'elles sont toutes distinctes. Alors, n solutions particulières de l'équation de récurrence

$$a_n y(m) + \ldots + a_o y(m-n) = 0$$

sont données par

$$y(m) = x_k^m \ ,$$

car

$$a_n x_k^m + \ldots + a_o x_k^{m-n} = x_k^{m-n} P_n(x_k) = 0 \ .$$

La solution générale de l'équation de récurrence est donc

$$y(m) = C_1 x_1^m + \ldots + C_n x_n^m \ .$$

Lorsque les racines ne sont pas toutes distinctes, il manque des solutions particulières de ce type. Examinons le cas d'une racine $x_k \neq 0$ de multiplicité p. On a donc

$$P_n(x_k) = 0 \ , \ P_n'(x_k) = 0 \ , \ \ldots \ , \ P_n^{(p-1)}(x_k) = 0.$$

Le polynôme

$$f_o(x) = x^{m-n} P_n(x) = a_n x^m + a_{n-*} x^{m-1} + \ldots + a_o x^{m-n}$$

possède le zéro x_k avec la même multiplicité p. Dès lors, sa dérivée

$$f_o'(x) = a_n\, m\, x^{m-1} + a_{n-1}\,(m-1)\, x^{m-1} + \ldots + a_o(m-n)\, x^{m-n}$$

possède en x_k un zéro de multiplicité (p-1) , de même que le polynôme

$$f_1(x) = x\, f_o'(x) = a_n \cdot m\, x^m + a_{n-1} \cdot (m-1)\, x^{m-1} + \ldots + a_o(m-n)\, x^{m-n},$$

dont la dérivée possède un zéro de multiplicité (p-2) en x_k. Poursuivant de la sorte, on obtient que tout polynôme

$$f_r(x) = x\, f_{r-1}'(x) = a_n\, m^r\, x^m + a_{n-1}(m-1)^r\, x^{m-1} + \ldots + a_o(m-n)^r\, x^{m-n}$$

avec $r < p$, possède en x_k un zéro de multiplicité (p-r). Mais cela signifie précisément qu'il existe p solutions indépendantes de l'équation récurrente, données par

$$\begin{cases} y_{ko}(m) = x_k^m \\ y_{k1}(m) = m\, x_k^m \\ \cdots\cdots\cdots\cdots \\ y_{k(p-1)}(m) = m^{p-1}\, x_k^m \end{cases} .$$

Plus généralement, s'il existe q zéros de multiplicités respectives $p_1, \ldots, p_q$, la solution générale de l'équation de récurrence est de la forme

$$y(m) = (C_{1,0} + \ldots + C_{1,p_1} m^{p_1}) x_1^m + (C_{2,0} + \ldots + C_{2,p_2} m^{p_2}) x_2^m + \ldots + (C_{q,0} + \ldots + C_{q,p_q} m^{p_q}) x_q^m .$$

4.5 - Méthode de Bernoulli simple

Supposons que les racines $x_1, \ldots, x_n$ du polynôme P_n soient séparées en module:

$$|x_1| > |x_2| > |x_3| > \ldots > |x_n| .$$

Dans ce cas, la solution de l'équation de récurrence (1) a la forme

$$y(m) = C_1 x_1^m + \ldots + C_n x_n^m$$

et, pour autant que $C_1 \neq 0$,

$$\frac{y(m)}{y(m-1)} = x_1 \frac{1 + \frac{C_2}{C_1}\left(\frac{x_2}{x_1}\right)^m + \ldots + \frac{C_n}{C_1}\left(\frac{x_n}{x_1}\right)^m}{1 + \frac{C_2}{C_1}\left(\frac{x_2}{x_1}\right)^{m-1} + \ldots + \frac{C_n}{C_1}\left(\frac{x_n}{x_1}\right)^{m-1}} = x_1 (1 + O((x_2/x_1)^{m-1}) ,$$

ce qui montre que le rapport y_m/y_{m-1} converge vers la racine de plus grand module.

Dans le cas où la première racine est multiple, la convergence est encore assurée, mais elle est beaucoup plus lente. En effet, si p est la multiplicité de x_1, on a

$$\frac{y(m)}{y(m-1)} = \frac{(C_{1,0} + \ldots + C_{1,p-1} m^{p-1}) x_1^m + (C_{2,0} + \ldots) x_2^m + \ldots}{(C_{1,0} + \ldots + C_{1,p-1}(m-1)^{p-1}) x_1^{m-1} + (C_{2,0} + \ldots) x_2^{m-1}}$$

$$= \left(\frac{m}{m-1}\right)^{p-1} x_1 \frac{(C_{1,p-1} + O(1/m)) + (C_{2,0} + \ldots)(x_2/x_1)^m + \ldots}{(C_{1,p-1} + O(1/(m-1)) + (C_{2,0} + \ldots)(x_2/x_1)^m + \ldots}$$

$$= x_1 (1 + O(1/m))$$

Dans les cas suivants: deux racines complexes conjuguées, deux racines différentes mais de même module, la méthode de Bernoulli simple ne converge pas.

Remarques

a) Un mauvais choix des conditions initiales pourrait conduire à un coefficient C_{1,p_1-1} nul, ce qui détruirait nos conclusions. HILDEBRAND [15] a proposé de choisir y(1), ..., y(m) de manière que tous les coefficients C_i soient égaux à 1 si les racines sont simples. Mais il n'est pas besoin d'utiliser de tels raffinements, car on peut montrer que le choix

$$y(1) = ... = y(n-1) = 0 \quad , \quad y(n) = 1$$

conduit nécessairement à $C_{1,p-1} \neq 0$.

En effet, supposons le contraire: alors, les (n-1) premières conditions ont la forme

$$\begin{cases} C_{1,0}\, x_1 + \ldots + C_{1,p_1-2} x_1 + C_{2,0}\, x_2 + \ldots = 0 \\ \ldots\ldots\ldots \\ C_{1,0} x_1^{n-1} + \ldots + C_{1,p_1-2}(n-1)^{p_1-2} x_1^{n-1} + C_{2,0}\, x_2^{n-1} + \ldots = 0 \end{cases}$$

Ce système homogène de (n-1) équations à (n-1) inconnues n'admet de solution non nulle que si son déterminant est nul, ce qui implique l'existence de (n-1) coefficients $b_1, \ldots, b_{n-1}$ non tous nuls et tels que

$$\begin{cases} b_1 x_1 + b_2 x_1^2 + \ldots + b_{n-1} x_1^{n-1} = 0 \\ \ldots\ldots\ldots\ldots \\ b_1 x_1 + b_2 2^{p_1-2} x_1^2 + \ldots + b_{n-1} (n-1)^{p_1-2} x_1^{n-1} = 0 \\ b_1 x_2 + b_2 x_2^2 + \ldots + b_{n-1} x_2^{n-1} = 0 \\ \ldots\ldots \end{cases}$$

Mais cela revient à dire que le polynôme de degré (n-2)

$$Q(x) = b_1 + b_2\, x + \ldots + b_{n-1}\, x^{n-2}$$

admet en x_1 un zéro de multiplicité $(p_1 - 1)$, en les autres x_k un zéro de multiplicité p_k . Or, la somme de ces multiplicités est (n-1), supérieure au degré du polynôme, ce qui est impossible. Par conséquent, tous les coefficients autres que C_{1,p_1-1} sont également nuls. Mais alors, tous les coefficients étant nuls, on ne pourra vérifier y(n) = 1.

b) Il est nécessaire de surveiller l'évolution de l'ordre de grandeur des y(m). Lorsqu'ils deviennent trop grands ou trop petits, on divise y(m), ..., y(m-n) par un même nombre, de manière à les ramener à une grandeur raisonnable. Cette division ne modifie en rien la convergence puisque seuls interviennent les rapports y(m)/y(m-1) .

c) On obtient la plus petite racine en appliquant la méthode de Bernoulli au polynôme

$$P_n^*(x) = x^n\, P_n(1/x) = a_o\, x^n + \ldots + a_n$$

dont les racines sont les inverses de celles de P_n .

d) Pour trouver les racines suivantes, on recommence les opérations sur le quotient $P_n(x)/(x - x_1)$.

4.6 - Méthode de Bernoulli double

Lorsque deux racines distinctes ont le même module, ou lorsqu'on est en présence de deux racines complexes conjuguées, la méthode de Bernoulli simple ne converge pas; dans le cas d'une racine multiple, elle converge lentement. Les racines multiples peuvent être éliminées par un artifice (voir section 5). Les autres cas courants sont:

- $|x_1| > |x_2| = |x_3| > |x_4|$: la méthode de Bernoulli simple permet d'obtenir x_1, mais non x_2 et x_3 .
- $|x_1| = |x_2| > |x_3|$: la méthode de Bernoulli simple ne converge pas pour x_1 .

Examinons d'abord le second cas. Après un nombre suffisant d'itérations,

$$y(m) \approx C_1 x_1^m + C_2 x_2^m ,$$

où x_1 et x_2 sont solutions de l'équation du second degré

$$x^2 - px - q = 0 \quad , \quad p = (x_1 + x_2) \quad , \quad q = - x_1 x_2 \ .$$

On a donc asymptotiquement le

$$y(m) - p\, y(m-1) - q\, y(m-2) = 0$$

ce qui, pris sur deux valeurs successives de m, donne le système linéaire en p et q

$$\begin{cases} p\, y(m-1) + q\, y(m-2) = y(m) \\ p\, y(m-2) + q\, y(m-3) = y(m-1) \end{cases}$$

Posant

$$\delta_{m-1} = y(m-1)\, y(m-3) - y^2(m-2) ,$$

on a donc à l'itération m

$$q(m) = \frac{1}{\delta_{m-1}} \left(y^2(m-1) - y(m)\, y(m-2) \right) = - \frac{\delta_m}{\delta_{m-1}}$$

et

$$p(m) = \frac{1}{\delta_{m-1}} \left(y(m)\, y(m-3) - y(m-1)\, y(m-2) \right) ,$$

ce qui ramène à la résolution d'une équation du second degré.

En général, deux cas sont possibles:

a) $|x_1| > |x_2| = |x_3|$, auquel cas $y(m)/y(m-1) \longrightarrow x_1$,

tandis que q(m) et p(m) ne convergent pas. Dans ce cas, $\delta_m \longrightarrow 0$.

b) Si $\delta_m \not\to 0$ et $|x_2| > |x_3|$, il y a convergence de la méthode double.

Une méthode fort générale résulte donc de la combinaison de la méthode simple et de la double, avec les tests appropriés.

Dans le cas d'une racine <u>double</u>, la méthode double converge beaucoup plus vite que la méthode simple.

<u>Remarques</u>

a) Le test de convergence des q(m) et p(m) est assez délicat, car il peut arriver que la même valeur se répète systématiquement deux fois (cas $x_2 = -x_1$, voir exercice). Il faut donc en vérifier la stabilisation sur un certain nombre d'itérations.

b) Il faut se garder de calculer δ_m / δ_{m-1} dans vérifier la non-nullité du dénominateur. Dans les tout premiers pas, celle-ci peut être accidentelle. Là encore, il faut éviter d'en déduire erronément que cela signifie $\delta_m \to 0$.

5. SUPPRESSION DES RACINES MULTIPLES

Les racines multiples peuvent être éliminées par l'artifice suivant [7].
Soit le polynôme

$$P_n(x) = a\,(x - x_1)^{m_1} \ldots\ldots (x - x_p)^{m_p} .$$

Montrons que <u>sa dérivée a la forme</u>

$$P_n'(x) = (x - x_1)^{m_1 - 1} \ldots (x - x_p)^{m_p - 1} \; H(x)$$

<u>où H(x) est un polynôme qui ne s'annule en aucun des zéros de $P_n(x)$.</u>

En effet, en supposant le contraire, c'est-à-dire $H(x_i) = 0$ pour un certain i, on a nécessairement

$$H(x) = (x - x_i)\,K(x) \quad , \; K(x) = \text{polynôme}$$

et

$$P_n'(x) = (x - x_1)^{m_1 - 1} \ldots (x - x_i)^{m_i} \ldots (x - x_p)^{m_p - 1} \; K(x)$$

$$= (x - x_i)^{m_i} \; J(x).$$

On peut alors calculer par la formule de LEIBNITZ

$$P_n^{(1+k)}(x) = \sum_{l=1}^{k} C_{1+k}^{l} \; m_i \ldots (m_i - l + 1)\,(x - x_i)^{m_i - l} \; J^{(k+1-l)}(x) .$$

Dans cette expression, on peut mettre le facteur $(x - x_i)$ en évidence tant que $k \leq m_i - 1$. Mais cela signifie qu'en x_i, P_n s'annule avec toutes ses dérivées jusqu'à l'ordre m_i ; en d'autres termes, que le zéro x_i est de

multiplicité $(m_i + 1)$ au moins. Or, ceci contrdit les hypothèses de départ.

Ce résultat signifie que <u>le plus grand commun diviseur G de P_n et P_n' est</u>

$$G(x) = (x - x_1)^{m_1 - 1} \dots (x - x_p)^{m_p - 1}$$

<u>si bien que le quotient P(x)/G(x) n'a plus que des zéros simples</u>.

Or, le calcul du plus grand commun diviseur peut être effectué par <u>l'algorithme d'EUCLIDE</u>, dont le principe est le suivant: si G(x) divise $P_n(x)$ et $P_n'(x)$, effectuons la division de ces deux polynômes:

$$P_n(x) = P_n'(x)\, Q(x) + R(x) ,$$

R(x) étant le reste. Comme G(x) divise $P_n(x)$ et $P_n'(x)Q(x)$, il doit diviser leur différence R(x). De la même façon, G(x) divisera le reste de la division de $P_n'(x)$ et R(x) et ainsi de suite, jusqu'à trouver une division exacte. Alors, le dernier diviseur est le p.g.c.d..

<u>Exemple</u> - Eliminons les racines multiples du polynôme

$$f(x) = x^3 - 4x^2 + 5x - 2$$

On a

$$f'(x) = 3x^2 - 8x + 5$$

<u>Divisons</u>:

$$\begin{array}{l|l}
x^3 - 4x^2 + 5x - 2 & 3x^2 - 8x + 5 \\
\cline{2-2}
-x^3 + \frac{8}{3}x^2 - \frac{5}{3}x & \frac{1}{3}x - \frac{4}{9} \\
\cline{1-1}
-\frac{4}{3}x^2 + \frac{10}{3}x - 2 & \\
\frac{4}{3}x^2 - \frac{32}{9}x + \frac{16}{9} & \\
\cline{1-1}
\text{RESTE:} \quad -\frac{2}{9}x + \frac{2}{9} & \\
= -(2/9)(x - 1) &
\end{array}$$

<u>Deuxième division</u>: on s'aperçoit aisément par l'algorithme de HORNER que

$$3x^2 - 8x + 5 = (x - 1)(3x - 5) + 0 \quad .$$

Le p.g.c.d. est donc (x-1) . La division de f(x) par (x-1), qui peut être effectuée par l'algorithme de HORNER, donne le polynôme

$$f^*(x) = x^2 - 3x + 2$$

qui donne de toute évidence 2 racines simples, à savoir, x = 1 et x = 2 .

L'application de ce procédé d'élimination des zéros multiples pose cependant la question d'un <u>algorithme général de division des polynômes</u>.

Soient

$(a_n, a_{n-1}, \ldots, a_o)$ les coefficients du dividende

$(b_p, b_{p-1}, \ldots, b_o)$ les coefficients du diviseur

$(c_q, c_{q-1}, \ldots, c_o)$ les coefficients du quotient.

On aura, bien entendu, $p + q = n$. On procède comme suit: partant de

$$c_q = a_n / b_p ,$$

on calcule

$$\begin{cases} a_n^{(1)} = 0 \\ a_{n-1}^{(1)} = a_{n-1} - c_q\, b_{p-1} \\ \cdots\cdots \\ a_{n-p}^{(1)} = a_{n-p} - c_q\, b_o \end{cases} .$$

Ensuite, partant de

$$c_{q-1} = a_{n-1}^{(1)} / b_p ,$$

on calcule

$$\begin{cases} a_{n-1}^{(2)} = a_{n-1}^{(1)} - c_{q-1}\, b_{p-1} \\ \cdots\cdots\cdots \\ a_{n-p-1}^{(2)} = a_{n-p-1}^{(2)} - c_{q-1}\, b_o \end{cases}$$

et ainsi de suite, jusques et y compris la q^e étape. Les $a_i^{(q)}$ sont alors les coefficients du reste, dont l'expression commence au premier coefficient non nul.

* 6. <u>METHODE DE BAIRSTOW</u>

Le principe de cette méthode est de chercher des diviseurs du polynôme de la forme $(x^2 - px - q)$. On sait que dans le cas général,

$$P_n(x) = (x^2 - px - q)(b_{n-2}\, x^{n-2} + \ldots + b_o) + R\, x + S .$$

Les coefficients R et S du reste seront nuls si $(x^2 - p\, x - q)$ est un diviseur de P_n . On peut les considérer comme des fonctions de p et q, dont on cherchera les zéros par la méthode de Newton-Raphson: étant donné une approximation de départ (p_o, q_o), on écrira

$$p = p_o + \Delta p \quad , \quad q = q_o + \Delta q$$

et

$$\begin{cases} R(p,q) \approx R(p_o, q_o) + \dfrac{\partial R}{\partial p_o} \Delta p + \dfrac{\partial R}{\partial q_o} \Delta q = 0 \\ S(p,q) \approx S(p_o, q_o) + \dfrac{\partial S}{\partial p_o} \Delta p + \dfrac{\partial S}{\partial q_o} \Delta q = 0 \end{cases}$$

Le déterminant de ce système de deux équations aux inconnues Δp et Δq vaut

$$\delta = \frac{\partial R}{\partial p_o}\frac{\partial S}{q_o} - \frac{\partial R}{\partial q_o}\frac{\partial S}{\partial p_o} \quad .$$

La solution est donc

$$\Delta p = \frac{1}{\delta}\left(\frac{\partial R}{\partial q_o} S(p_o,q_o) - \frac{\partial S}{\partial p_o} R(p_o,q_o)\right)$$

$$\Delta q = \frac{1}{\delta}\left(\frac{\partial S}{\partial p_o} R(p_o,q_o) - \frac{\partial S}{\partial q_o} S(p_o,q_o)\right)$$

Mais comment calculer les dérivées de R et S? L'algorithme général de division par un trinôme donne

$$b_{n-2} = a_n$$

$$b_{n-3} = a_{n-1} + p\, b_{n-2}$$

$$b_{n-4} = a_{n-2} + p\, b_{n-3} + q\, b_{n-2}$$

$$\cdots\cdots$$

$$b_o = a_2 + p\, b_1 + q\, b_2$$

$$R = a_1 + p\, b_o + q\, b_1$$

$$S = a_o + p\, b_o$$

Introduisant les notations

$$c_k = \frac{\partial b_k}{\partial p} \qquad \text{et} \qquad d_k = \frac{\partial b_k}{\partial q} \quad ,$$

on obtient

$$c_{n-2} = 0$$

$$c_{n-3} = b_{n-2} + p\, c_{n-2} = b_{n-2}$$

$$c_{n-4} = b_{n-3} + p\, c_{n-3} + q\, c_{n-2}$$

$$\cdots\cdots$$

$$c_o = b_1 + p\, c_1 + q\, c_2$$

$$\frac{\partial R}{\partial p} = b_o + p\, c_o + q\, c_1$$

$$\frac{\partial S}{\partial p} = q\, c_o$$

et

$$d_{n-2} = 0$$

$$d_{n-3} = p\, d_{n-2} = 0$$

$$d_{n-4} = p\, d_{n-3} + q\, d_{n-2} + b_{n-2}$$

$$\cdots\cdots$$

$$d_o = p\, d_1 + q\, d_2 + b_2$$

$$\frac{\partial R}{\partial q} = p\ d_o + q\ d_1 + b_1$$

$$\frac{\partial S}{\partial q} = q\ d_o + b_o \quad .$$

On recommence les calculs jusqu'à convergence de p et q. Deux racines sont ainsi obtenues (elles peuvent être confondues en module ou complexes conjuguées). On divise alors le polynôme initial par $(x^2 - p\ x - q)$ (on a déjà calculé les coefficients du quotient!), et on recommence le processus.

Exercice 1 - Vérifier l'indépendance des solutions $m^l \, x_k^m$ de l'équation de récurrence

$$a_n \, y(m) + \ldots + a_o \, y(m-n) = 0 \, .$$

Exercice 2 - On considère l'équation de récurrence non homogène

$$a_n \, y(m) + \ldots + a_o \, y(m-n) = b \quad , \quad b = \text{cte.}$$

a) Montrer que si l'on se donne les valeurs $\bar{y}(1), \ldots, \bar{y}(m)$ de la solution pour m = 1 à n, elle est univoquement définie.

b) Montrer que la solution générale de l'équation est la somme d'une solution particulière de l'équation complète et de la solution générale de l'équation homogène.

c) Dans quelles conditions existe-t-il une solution particulière constante ? Quelle est-elle?

Solution

a) Soient deux solutions y_1 et y_2 . Leur différence $z = y_1 - y_2$ vérifie l'équation homogène et s'annule aux n premiers points. Elle est donc nulle.

b) Soit y_o une solution particulière, et soit y une quelconque autre solution de l'équation. Leur différence z vérifie évidemment l'équation homogène.

c) soit $y_o(m) = z = \text{cte}$. On a donc

$$(a_n + \ldots + a_o) \, z = b \, ,$$

ce qui donne

$$z = \frac{b}{a_n + \ldots + a_o} \, ,$$

à condition que le quotient ne soit pas nul, ce qui a lieu si l'équation

$$a_n \, x^n + \ldots + a_o = 0$$

n'admet pas le nombre 1 comme racine.

Exercice 3 - Soit le polynôme

$$f(x) = 231 \, x^6 - 315 \, x^4 + 105 \, x^2 - 5$$

On demande :

a) De chercher un borne supérieure du module des racines.

b) De chercher une borne inférieure des racines.

c) De déterminer à quelle distance de l'origine on est certain de trouver un zéro de f(x).

d) De calculer par la méthode de Bernoulli la racine de plus grand module.

Solution

a) $\alpha = 315$, $|a_n| = 231$ $\qquad |x_k| < 1 + \frac{315}{231} = 2{,}364$

b) $\beta = 315$, $|a_o| = 5$ $\qquad |x_k| > \frac{5}{315 + 5} = 0{,}01563$

c) on a $\inf_k |x_k - 0| \leq \sqrt[6]{\left|\frac{5}{231}\right|} = 0,5279$

d) Le polynôme étant pair, on aura automatiquement

$f(x_k) = f(-x_k) = 0$,

donc il faut employer la méthode de Bernoulli double. Après 46 itérations en partant des données $y(1) = \ldots = y(5) = 0$, $y(6) = 1$, on obtient pour la troisième fois

$q(m) = -(\delta_m / \delta_{m-1}) = 0,8694995$, et on calcule alors $p = 0$.

On a donc

$x_{1,2} = \pm \frac{1}{2} \sqrt{4q} = \pm 0,9324695.$

Il faut noter que δ_m / δ_{m-1} garde systématiquement la même valeur pendant deux pas.

Exercice 4 - Ne peut on imaginer un artifice permettant de ne garder systématiquement qu'une des deux racines symétriques de l'équation de l'exercice précédent et de le traiter par la méthode de Bernoulli simple?

Solution - En posant $z = x^2$, on obtient l'équation

$231 z^3 - 315 z^2 + 105 z - 5 = 0$.

Partant de $y(1) = y(2) = 0$, $y(3) = 1$, on obtient

$z_1 = 0,8694995$

après 23 itérations. Il en découle

$x_{1,2} = \pm 0,9324695.$

R E S O L U T I O N D E S E Q U A T I O N S
A L G E B R I Q U E S E T T R A N S C E N D A N T E S

1. REPERAGE DES ZEROS

Le problème de la résolution d'une équation de la forme

$$f(x) = 0$$

se pose fréquemment. La première étape de sa résolution consiste à situer approximativement les zéros. Dans bien des cas, un diagramme peut rendre de précieux services. Soit par exemple à résoudre l'équation

$$f(x) = \cos x + \frac{1}{\operatorname{ch} x} = 0 \ ,$$

que l'on rencontre dans l'étude des vibrations d'une poutre encastrée-libre (fig. 1). Il est aisé de se faire une idée de l'emplacement des zéros en remarquant qu'il s'agit des points de rencontre des graphes de cos x et $(-1/\operatorname{ch} x)$. A l'aide d'un diagramme (fig. 2), on peut observer les faits suivants:

a) Il y a une infinité de solutions positives $x_1, \ldots, x_n, \ldots$

b) Pour x suffisamment grand, $1/\operatorname{ch} x \approx 0$, ce qui signifie que pour les grandes valeurs de n,

$$\cos x_n \approx 0 \ ,$$

soit

$$x_n \approx \frac{\pi}{2} + (n-1)\pi \ .$$

c) La première racine est contenue entre $\frac{\pi}{2}$ et π , la seconde, entre π et $\frac{3\pi}{2}$, la troisième entre $\frac{5\pi}{2}$ et 2π , etc .... En général, une racine d'ordre impair x_{2k+1} vérifie

$$\frac{\pi}{2} + 2k\pi < x_{2k+1} < (2k+1)\pi \ ,$$

et une racine d'ordre pair x_{2k} admet l'encadrement

$$(2k-1)\pi < x_{2k} < \frac{\pi}{2} + (2k-1)\pi \ .$$

Nous avons donc obtenu un encadrement séparé de chaque racine ou, comme on dit, nous les avons séparées.

2. MULTIPLICITE DES ZEROS

On dit qu'un zéro ξ est simple ou de multiplicité 1 si $f'(\xi)$ est un nombre fini ou nul. Dans ce cas, on a

$$f'(\xi) = \lim_{x \to \xi} \frac{f(x)}{x - \xi} ,$$

ce qui entraîne

$$\lim_{x \to \xi} \frac{(x-\xi)f'(x)}{f(x)} = 1.$$

Un zéro de multiplicité entière m se caractérise par les conditions

$$f(\xi) = 0 , f'(\xi) = 0, \ldots, f^{(m-1)}(\xi) = 0 , f^{(m)}(\xi) \neq 0 \text{ et fini},$$

ce qui permet d'écrire, si $f \in C^{m+1}$,

$$f(x) = f^{(m)}(\xi) (x-\xi)^m + f^{(m+1)}(x^*) (x-\xi)^{m+1} ,$$

et, par dérivation,

$$f'(x) = m f^{(m)}(\xi) (x-\xi)^{m-1} + 0(x-\xi)^m.$$

Il en découle la relation

$$\lim_{x \to \xi} \frac{(x-\xi) f'(x)}{f(x)} = m .$$

Nous dirons qu'en général, une fonction $f \in C^1(\xi)$ admet un zéro de multiplicité m, m réel >0, éventuellement infini, si

$$\frac{(x-\xi) f'(x)}{f(x)} = m + g(x) ,$$

avec $g(x)$ tendant vers zéro pour $x \to \xi$ de telle façon que

$$\frac{g(x)}{x-\xi}$$

soit intégrable au voisinage de ξ. Cette dernière condition est certainement vérifiée s'il existe un nombre $\theta > 0$ tel que

$$g(x) = 0((x-\xi)^\theta).$$

La définition ci-dessus implique un comportement particulier au voisinage de ξ, comme le montre le théorème suivant:

Théorème 1 - Si $f \in C^1(\xi)$ admet en $x = \xi$ un zéro de multiplicité m, $0 < m < \infty$, il existe une constante finie A telle que

$$\lim_{x \to \xi} \frac{|f(x)|}{|x-\xi|^m} = A .$$

En effet, on a

$$\frac{f'(x)}{f(x)} = \frac{m}{x - \xi} + \frac{g(x)}{x - \xi} \quad .$$

Intégrant entre un point $x_o \neq \xi$ et x, on obtient

$$\ln \left| \frac{f(x)}{f(x_o)} \right| = m \ln \left| \frac{x - \xi}{x_o - \xi} \right| + \int_{x_o}^{x} \frac{g(t)}{t - \xi} \, dt$$

$$= m \ln \left| \frac{x - \xi}{x_o - \xi} \right| + G(x) - G(x_o) \, ,$$

où G est la fonction continue définie par

$$G(x) = \int_{\xi}^{x} \frac{g(t)}{t - \xi} \, dt \, .$$

On en déduit

$$\frac{|f(x)|}{|x - \xi|^m} = \frac{|f(x_o)|}{|x_o - \xi|^m} \exp \left[G(x) - G(x_o) \right] ,$$

et

$$\lim_{x \to \xi} \frac{|f(x)|}{|x - \xi|^m} = \frac{|f(x_o)|}{|x_o - \xi|^m} \exp \left[- G(x_o) \right] = A.$$

Ce théorème admet une réciproque, à savoir:

Théorème 2 - Si la fonction f vérifie

$$\frac{|f(x)|}{|x - \xi|^m} = G(x) \, ,$$

où G est une fonction dont le logarithme est absolument continu au voisinage de ξ , f admet en ξ un zéro de multiplicité m.

En effet, on a dans ce cas

$$\ln \frac{|f(x)|}{|x - \xi|^m} = \ln G(x) = \int_{\xi}^{x} h(t) \, dt \, ,$$

avec h intégrable, soit

$$\ln |f(x)| = m \ln |x - \xi| + \int_{\xi}^{x} h(t) \, dt \, ,$$

ce qui entraîne

$$\frac{f'(x)}{f(x)} = \frac{m}{x - \xi} + h(x)$$

et

$$\frac{(x - \xi) f'(x)}{f(x)} = m + (x - \xi) h(x) \, ,$$

c'est-à-dire que f admet un zéro de multiplicité m.

Notons que l'on peut trouver des zéros de multiplicité infinie. C'est le cas de la fonction

$$f(x) = \exp(-\frac{1}{|x|})$$

en x = 0. On a en effet

$$f'(x) = \frac{1}{|x|^2} \exp(-\frac{1}{|x|}) ,$$

et

$$\frac{x\, f'(x)}{f(x)} = \frac{1}{|x|} \to \infty .$$

Il s'agit d'un contact infiniment intime avec l'axe des x. A l'inverse, on peut trouver des multiplicités nulles. La fonction

$$f(x) = \frac{\text{sign } x}{\ln |\frac{1}{x}|} = - \frac{\text{sign } x}{\ln |x|}$$

s'annule en x = 0 . On a

$$f'(x) = \frac{\text{sign } x}{\ln^2 |x|} \cdot \frac{1}{x}$$

et

$$\frac{x\, f'(x)}{f(x)} = - \frac{1}{\ln |x|} \to 0 .$$

Le graphe de cette fonction épouse l'axe des y de manière parfaitement intime.

4. SEPARATION DU ZERO PAR LES VALEURS DE LA FONCTION

Comment déterminer si une approximation x du zéro ξ est bonne ou mauvaise? On est tenté de dire qu'il suffit de vérifier que |f(x)| est très petit. Malheureusement, les choses ne sont pas si simples. Tout d'abord, il faudrait pouvoir dire par rapport à quoi |f(x)| est petit. Ainsi, la racine positive de l'équation

$$f(x) = x^2 - 1 = 0$$

est égale à 1. Pour x = 1,001, on a

$$f(1{,}001) = 0{,}002 ;$$

la même valeur de x intoduite dans la fonction

$$g(x) = 10^6 x - 10^6$$

donne

$$g(1,001) = 2000,$$

assurément petit devant 10^6, mais non dans l'absolu.

D'autre part, le comportement même de la fonction au voisinage du zéro joue un grand rôle: lorsque le zéro a une multiplicité supérieure à l'unité, il y a contact entre le graphe de la fonction et l'axe des x, ce qui signifie que $f(x) \approx 0$ pour x déjà assez éloigné du zéro ξ (fig. 3).

Etudions plus précisément cette relation. En supposant que le zéro soit de multiplicité m, on a

$$\lim_{x \to \xi} \frac{|f(x)|}{|x - \xi|^m} = a ,$$

et il existe un voisinage où

$$\frac{|f(x)|}{|x - \xi|^m} \geqslant \frac{a}{2} = \alpha \quad .$$

Dans ce voisinage, on a donc

$$|x - \xi| \leqslant \left(\frac{|f(x)|}{\alpha} \right)^{1/m} .$$

La relation

$$y = \left(\frac{|f(x)|}{\alpha} \right)^{1/m}$$

est représentée en figure 4 pour diverses valeurs de m. On constate aisément que pour les fortes valeurs de m, une petite valeur de $|f(x)|$ correspond à une grande valeur de $|x - \xi|$.

Cette circonstance s'aggrave encore du fait des erreurs d'arrondi. Supposons le calcul de f entaché d'une erreur maximale $\pm \eta$, et que l'on arrête le calcul dès que la valeur calculée $\tilde{f}(x)$ vérifie

$$|\tilde{f}(x)| \leqslant \varepsilon \quad .$$

On ne pourra rien affirmer de plus que

$$|f(x)| \leqslant \varepsilon + \eta ,$$

ce qui donne

$$|x - \xi| \leqslant \left(\frac{\varepsilon + \eta}{\alpha} \right)^{1/m} ,$$

valeur d'autant plus grande que la multiplicité m du zéro est plus élevée. Même si le processus de calcul mène à $\tilde{f}(x) = 0$, on ne pourra jamais garantir mieux que

$$|x - \xi| \leqslant \left(\frac{\eta}{\alpha} \right)^{1/m} .$$

Il est donc difficile de tester la valeur d'un zéro de multiplicité élevée. Pour illustrer ce fait, supposons que l'on calcule, avec une précision de 10^{-6}, le zéro positif de

$$f(x) = x^2 - 10.$$

La solution est

$$= \sqrt{10} = 3{,}162277660...$$

Les nombres entrant en jeu étant voisins de 10, on aura $\Delta f \approx 2.10^{-5}$. Pour $f(x) = 2.\ 10^{-5}$, on obtient

$$x = \sqrt{10 - 2.10^{-5}} = \sqrt{10}\sqrt{1 - 2.10^{-6}} \approx \sqrt{10}\,(1 - 10^{-6}),$$

soit

$$|x - \xi| \approx 3.10^{-6}\ .$$

Soit à présent la fonction

$$g(x) = \tfrac{1}{2}(x^2 - 2\sqrt{10}\,x + 10),$$

admettant le même zéro, mais avec une multiplicité 2. Pour une précision de 10^{-6}, on aura $\Delta g \approx 2.10^{-5}$, et, pour $g(x) = -2.\ 10^{-5}$, il vient

$$\frac{x^2}{2} - \sqrt{10}\,x + 10 - 2.10^{-5} = 0\ ,$$

soit

$$x = \sqrt{10} + \sqrt{10 - 10 + 2.10^{-5}} = \sqrt{10} + 4{,}5.10^{-3}\ ,$$

ce qui donne

$$|x - \xi| = 4{,}5.10^{-3}\ ,$$

soit une erreur quinze cents fois plus grande sur la racine, pour une même erreur sur la fonction.

La figure 5 permet de visualiser l'influence de l'erreur de calcul de f.

5. MÉTHODES DE SUBDIVISION DE L'INTERVALLE

Les méthodes de subdivision de l'intervalle ne s'appliquent que si la fonction passe par la valeur zéro en changeant de signe.

5.1 - Subdivision binaire (fig. 6)

Partant d'un encadrement $(g_n\ ,\ d_n)$ tel que

$$f(g_n)\ .\ f(d_n) < 0\ ,$$

on calcule

$$x_{n+1} = \frac{g_n + d_n}{2}\ .$$

Si ce point est à droite du zéro, il remplacera avantageusement d_n; dans le cas contraire, il remplacera g_n. Ces conditions s'écrivent explicitement

- Si $f(x_{n+1}) \cdot f(g_n) < 0$, alors

$$\rightarrow \begin{cases} d_{n+1} = x_{n+1} \\ g_{n+1} = g_n \end{cases}$$

- Si $f(x_{n+1}) \cdot f(g_n) > 0$, alors

$$\rightarrow \begin{cases} g_{n+1} = x_{n+1} \\ d_{n+1} = d_n \end{cases}$$

- Dans le cas fortuit où $f(x_{n+1}) = 0$, on a évidemment $\xi = x_{n+1}$.

La convergence de cet algorithme est évidente, puisque

$$| d_n - g_n | = 2^{-n} | d_o - g_o | \quad ,$$

ce qui entraîne la convergence des deux extrémités de l'intervalle:

$$| d_n - \xi | \leq | d_n - g_n | = 2^{-n} | d_o - g_o | \longrightarrow 0$$

$$| \xi - g_n | \leq | d_n - g_n | = 2^{-n} | d_o - g_o | \longrightarrow 0 \quad .$$

Cependant, tout ceci suppose qu'il n'y a pas d'erreur de décision: supposons que, dans le cas du premier diagramme de la figure 6, les erreurs d'arrondi entraînent $\tilde{f}(x_{n+1}) < 0$. On cherchera alors le zéro dans l'intervalle $[x_{n+1} , d_n]$. Or, ceci est possible dès que

$$| x_{n+1} - \xi | < | \eta / \alpha |^{1/m} \quad ,$$

η étant l'erreur de calcul de f et m, la multiplicité du zéro. On peut donc s'attendre à une erreur

$$| \Delta\xi | \approx | \eta / \alpha |^{1/m} \quad ,$$

ce qui signifie que ce procédé est pratiquement limité aux zéros simples ou de multiplicité inférieure à 1.

5.2 - Méthode de la sécante (fig. 7)

Le principe de la méthode de la sécante est d'essayer de construire un découpage plus efficace que le découpage binaire. A cette fin, on remplace la fonction f par son interpolée linéaire

$$f(x) = f(g_n) + \frac{x_n - g_n}{d_n - g_n} (f(d_n) - f(g_n))$$

et on calcule le zéro de cette dernière, qui est donné par

$$x_{n+1} = \frac{g_n f(d_n) - d_n f(g_n)}{f(d_n) - f(g_n)} \quad .$$

Pour le reste, on procède comme dans le cas de la subdivision binaire.

L'étude de la convergence de cette méthode nécessite l'hypothèse que le zéro soit simple, ce qui implique l'existence de deux constantes α et A telles que

$$\alpha\,|x-\xi| \leqslant |f(x)| \leqslant A\,|x-\xi|$$

dans un certain voisinage du zéro ξ . Supposons, pour fixer les idées, que $x_{n+1} > \xi$. Alors (fig. 8) $d_{n+1} = x_{n+1}$, $g_{n+1} = g_n$. Examinons quelle est la relation entre $|d_{n+1} - \xi|$ et $|d_n - \xi|$. On a

$$(d_{n+1} - \xi) = (d_n - \xi) - (d_n - d_{n+1}) \quad ,$$

soit

$$|d_{n+1} - \xi| = |d_n - \xi| - |d_n - d_{n+1}| \; . \qquad (1)$$

Cherchons donc une borne inférieure de $|d_n - d_{n+1}|$:

$$|d_n - d_{n+1}| = |d_n - x_{n+1}|$$

$$= \left| d_n - \frac{g_n f(d_n) - d_n f(g_n)}{f(d_n) - f(g_n)} \right| = \left| \frac{(d_n - g_n) f(d_n)}{f(d_n) - f(g_n)} \right|$$

$$= \frac{|d_n - g_n|\,|f(d_n)|}{|f(d_n) - f(g_n)|} \; .$$

Or, on a d'une part

$$|f(d_n)| \geqslant \alpha\,|d_n - \xi|$$

et, d'autre part,

$$|f(d_n) - f(g_n)| \leqslant |f(d_n)| + |f(g_n)| \leqslant A\,|d_n - \xi| + A\,|g_n - \xi| = A\,|d_n - g_n|$$

ce qui entraîne

$$|d_n - d_{n+1}| \geqslant \frac{\alpha}{A}\,|d_n - \xi|$$

et, par (1),

$$\boxed{|d_{n+1} - \xi| \leqslant \left(1 - \frac{\alpha}{A}\right) |d_n - \xi|}$$

Ainsi, chaque fois que la borne droite se déplace, sa distance au zéro est multipliée par $\left(1 - \frac{\alpha}{A}\right)$. De la même manière, on montrerait aisément que tout mouvement de la borne gauche la rapproche du zéro d'un facteur $\left(1 - \frac{\alpha}{A}\right)$.

Sur n itérations, on observera n_d mouvements de la borne droite et n_g mouvements de la borne gauche, avec

$$n_g + n_d = n \ ,$$

si bien que

$$|d_n - \xi| \leq (1 - \frac{\alpha}{A})^{n_d} |d_o - \xi| \leq (1 - \frac{\alpha}{A})^{n_d} |d_o - g_o|$$

$$|\xi - g_n| \leq (1 - \frac{\alpha}{A})^{n_g} |\xi - g_o| \leq (1 - \frac{\alpha}{A})^{n_g} |d_o - g_o|$$

On peut imaginer quatre cas:

a) On obtient ξ après un nombre fini d'itérations. Ce cas exceptionnel n'exige aucun commentaire.

b) Les itérations se passent de telle manière que

$$n_g \longrightarrow \infty \quad , \quad n_d \leq N_d \text{ fini.}$$

Alors, $g_n \longrightarrow \xi$. De plus, il existe nécessairement une valeur N de n au-delà de laquelle seule la borne inférieure de meut. Donc $x_n \longrightarrow \xi$.

c) Au contraire, n_d tend vers l'infini, tandis que n_g reste fini. C'est exactement le cas symétrique du précédent et on a toujours $x_n \longrightarrow \xi$.

d) Quel que soit N, il existe toujours une itération $n_1 \geq N$ où l'extrémité gauche se meut et une itération $n_2 \geq N$ où la droite se meut. alors, $g_n \longrightarrow \xi$, $d_n \longrightarrow \xi$, $x_n \longrightarrow \xi$.

Comment évaluer l'erreur après une itération $(n + 1)$? On peut se fonder sur la relation

$$|d_n - \xi| \leq \frac{A}{\alpha} |d_{n+1} - d_n|$$

ou son équivalent à gauche, selon que l'une ou l'autre des extrémités est en mouvement.

6. METHODE DES APPROXIMATIONS SUCCESSIVES

6.1 - Imaginons que l'on ait à résoudre l'équation

$$x^3 + 5x - 1 = 0 \ .$$

Il est tentant de mettre cette équation sous la forme

$$x = \frac{1}{5} (1 - x^3)$$

et d'essayer de la résoudre itérativement en posant

$$x_{n+1} = \frac{1}{5} (1 - x_n^3) \ .$$

Partant de $x = 0$, on obtient successivement:

$$x_o = 0$$

$$x_1 = 0{,}2000$$

$$x_2 = 0{,}1984$$

$$x_3 = 0{,}1984$$

et

$$x_3^3 + 5\,x_3 - 1 = 4{,}499 \cdot 10^{-6} .$$

On le voit, la convergence est rapide. Par contre, si l'on part de $x_o = 10$, on obtient

$$x_o = 10$$
$$x_1 = -199{,}8$$
$$x_2 = 1{,}595.10^6$$
$$x_3 = -\,811{,}9.10^5,$$

soit une suite nettement divergente.

On conçoit donc l'utilité d'une étude approfondie des processus comme ci-dessus, permettant de déterminer les conditions de convergence.

6.2 - Le principe de la méthode exposée ci-dessus est de construire une suite de la forme

$$x_{n+1} = F(x_n) ,$$

qui devra converger vers la racine ξ de l'équation. La fonction F est supposée continue.

Une première condition <u>nécessaire</u> est $F(\xi) = \xi$. En effet, si $x_n \longrightarrow \xi$, on a

$$|F(\xi) - \xi| \leq |F(\xi) - F(x_n)| + |x_{n+1} - \xi| \longrightarrow 0 .$$

Mais cette condition ne suffit pas.

6.3 - <u>Critère de l'application contractante</u>

Une condition suffisante est le critère de l'application contractante, encore connu sous le nom de théorème du point fixe. Ce résultat est très général et s'applique bien au-delà du seul problème des racines des équations. Citons par exemple

- les systèmes non linéaires dans R^n
- les solutions d'équations différentielles dans $C^o([a , b])$.

Aussi travaillerons-nous dans le cadre général des <u>espaces métriques complets</u>. Rappelons qu'un espace E est <u>métrique</u> si à tout couple (x,y) de points de E, on peut associer une distance d(x,y) telle que:

$$\begin{cases} d(x,y) > 0 \quad \text{si } x \neq y \\ d(x,x) = 0 \quad , \quad d(x,y) = d(y,x) \\ d(x,z) \leq d(x,y) + d(y,z) \quad \text{pour tous } x, y, z \in E . \end{cases}$$

Cet espace est <u>complet</u> si la condition de CAUCHY

$$\lim_{p,q \longrightarrow \infty} d(x_p , x_q) = 0$$

sur une suite $\{ x_n \}$ implique l'existence d'une limite ξ telle que

$$\lim_{m \longrightarrow \infty} d(x_m , \xi) = 0.$$

Dans un espace métrique complet E, une application F de E dans E est dite lipschitzienne s'il existe une constante L, positive et finie, telle que

$$d(F(x),F(y)) \leqslant L\, d(x,y) ,$$

quels que soient x et $y \in E$. Elle est contractante si $L < 1$.

On a les théorèmes suivants:

Théorème 1 : Soit ξ une solution de l'équation $F(x) = x$. Alors, si F est contractante dans une boule $\overline{B}_R(\xi) = \{ x \in E \mid d(x,\xi) \leqslant R \}$, toute suite de la forme $x_{n+1} = F(x_n)$, avec $x_o \in \overline{B}_R(\xi)$, converge vers ξ .

On a en effet

$$d(x_{n+1},\xi) = d(F(x_n), F(\xi)) \leqslant L\, d(x_n,\xi) \leqslant \ldots \leqslant L^{n+1}\, d(x_o,\xi)$$

$$\leqslant L^{n+1} R \longrightarrow 0 .$$

Théorème 2 (d'existence de la limite) : Soit F une application contractante dans une boule $\overline{B}_R(c) = \{ x \in E \mid d(x,c) \leqslant R \}$, et soit L la constante de Lipschitz correspondante. Si

$$d(F(c),c) \leqslant (1 - L)R ,$$

l'équation admet une et une seule racine $\xi \in \overline{B}_R(c)$, et c'est la limite de la suite $x_{n+1} = F(x_n)$, avec un point de départ quelconque dans la boule fermée.

Démonstration

a) La suite ne sort pas de la boule $\overline{B}_R(c)$, car si $x \in \overline{B}_R(c)$, on a

$$d(F(x),c) \leqslant d(F(x),F(c)) + d(F(c),c)$$

$$\leqslant L\, d(x,c) + (1 - L)\, R \leqslant R$$

b) La suite est de CAUCHY. En effet, pour $p < q$, on a

$$d(x_p, x_q) \leqslant \sum_{k=p}^{q-1} d(x_k, x_{k+1}) \leqslant \sum_{k=p}^{q-1} L^k\, d(x_1, x_o)$$

$$\leqslant d(x_1, x_o) \sum_{k=p}^{\infty} L^k = d(x_1, x_o)\, \frac{L^p}{1 - L} \longrightarrow 0$$

pour $\inf (p,q) \longrightarrow \infty$.

c) La limite ξ vérifie $F(\xi) = \xi$. En effet,

$$d(F(\xi),\xi) \leqslant d(F(\xi), F(x_n)) + d(x_{n+1},\xi)$$

$$\leqslant L\, d(x_n,\xi) + d(x_{n+1},\xi) \longrightarrow 0 .$$

6.4 - Un critère de divergence

Le critère suivant permet d'écarter des processus qui ne peuvent converger:

Théorème 3 : Si la fonction F vérifie

$$d(F(x_1), F(x_2)) \geqslant G\, d(x_1, x_2) \qquad , \qquad G \geqslant 1$$

dans un ouvert Ω , une suite de la forme

$$x_{n+1} = F(x_n)$$

ne peut converger en aucun point de Ω , à moins que $F(x_o) = x_o$ (cas exceptionnel de la suite $x_o, x_o, x_o, \ldots$)

Nous démontrerons ce théorème par l'absurde: supposons que la suite

$$x_{n+1} = F(x_n)$$

converge vers un point $\xi \in \Omega$, sans que $F(x_o) = x_o$. Posons (fig. 9)

$$\rho_1 = d(x, \complement\Omega) > 0 \qquad (\Omega \text{ est ouvert! })$$

$$\rho_2 = d(x_o, F(x_o)) > 0 \qquad (x_o \neq F(x_o) \ ! \ )$$

et

$$\rho = \frac{1}{3} \inf(\rho_1, \rho_2) \neq 0$$

Comme la suite $\{x_n\}$ converge vers ξ , il doit exister un nombre N tel que pour $n \geqslant N$, $x_n \in \overline{B}_\rho(\xi)$. Mais si x_n et x_{n+1} vérifient cette condition, on a aussi

$$2\rho \geqslant d(x_n, x_{n+1}) \geqslant G^n\, d(x_o, x_1) \geqslant d(x_o, x_1) = \rho_2 \geqslant 3\rho ,$$

soit

$$2 \geqslant 3,$$

ce qui est contradictoire.

6.5 - Estimation pratique de l'erreur dans un prcessus contractant

Après n itérations, on a

$$d(x_n, \xi) \leqslant \sum_{k=n} d(x_k, x_{k+1}) \leqslant d(x_n, x_{n+1}) \sum_{k=0} L^k ,$$

soit

$$\boxed{d(x_n, \xi) \leqslant \frac{d(x_n, x_{n+1})}{1 - L}}$$

Il suffit donc de vérifier que deux itérés successifs sont suffisamment voisins.

6.6 - Stabilité numérique de la contraction

Supposons qu'à un instant quelconque du processus apparaisse une erreur, de telle sorte que l'on traite un point $\tilde{x}_n \neq x_n$:

$$e_n = d(\tilde{x}_n , x_n) \neq 0 .$$

Mais on a

$$d(F(\tilde{x}_n) , F(x_n)) \leq L\, d(\tilde{x}_n , x_n) = L\, e_n \leq e_n ,$$

si bien que l'erreur <u>diminue</u> au cours de l'itération.

7. APPLICATION AUX FONCTIONS NUMERIQUES D'UNE VARIABLE

7.1 - Pour une fonction F(x) dérivable, on a, par le théorème des accroissements finis,

$$|F(x) - F(y)| = |x - y|\, |F'(z)| ,$$

avec z situé strictement entre x et y. Dès lors,

- <u>La fonction F est contractante sur $[a, b]$ si</u>

$$L = \sup_{]a, b[} |F'(z)| < 1 ;$$

- <u>Dans un ouvert où $|F'(z)| \geq 1$, il ne peut y avoir convergence.</u>

Un examen plus approfondi permet d'améliorer ces conclusions. En effet, si ξ est la racine, on a

$$F(x) - F(\xi) = F'(z).(x - \xi) ,$$

ce qui donne, au cours de l'itération,

$$x_{n+1} - \xi = F'(z).(x_n - \xi) .$$

<u>Dès lors, si dans $]x_o , \xi[$ (ou $]\xi, x_o[$) , $F'(z) \in [0, 1[$, la suite est monotone et bornée, donc convergente.</u> Au contraire, si $F'(z) \in]-1, 0[$, la suite oscille autour de ξ , et il convient de s'assurer que l'on ne sortira pas de l'intervalle de convergence. Cette condition de stabilité sera garantie si $|F'(z)| < 1$ dans $[a, b]$, avec ξ contenu dans $[a + \frac{h}{3} , b - \frac{h}{3}]$, où $h = b - a$, à condition de partir dans ce sous-intervalle. En effet (fig. 10), le cas le plus défavorable est alors celui où $\xi = b - h/3$ et $x_o = a + h/3$ (ou l'inverse). Alors, $|x_o - \xi| \leq h/3$ et x_1 sera dans $[a, b]$.

La figure 11 donne une interprétation graphique des considérations qui précèdent.

7.2 - Imaginons que l'on veuille résoudre l'équation

$$tg\ x = x ,$$

qui apparaît dans certains problèmes thermiques. L'examen de la figure 12 montre que les racines positives sont encadrées par

$$n\pi < \xi_n < n\pi + \frac{\pi}{2}$$

et, pour les grandes valeurs de n et, donc de x, ces racines finissent par se confondre avec celles de l'équation

$$tg\ x = \infty ,$$

ce qui donne

$\xi_n \approx \frac{\pi}{2} + n\pi$ (n grand) .

Partant de $n\pi$, on est tenté d'essayer l'algorithme

$$x_{k+1} = \text{tg } x_k$$

ce qui revient à écrire $F(x) = \text{tg } x$. On a alors

$$F'(x) = 1 + \text{tg}^2 x \geqslant 1 ,$$

si bien que cet algorithme <u>diverge toujours</u>. En partant par exemple de $x_o = 4$, on obtient

$x_1 = 1,158$	$x_5 = 1,127$	$x_9 = -14,06$
$x_2 = 2,282$	$x_6 = 2,103$	$x_{10} = -13,33$
$x_3 = -1,160$	$x_7 = -1,696$	.
$x_4 = -2,297$	$x_8 = 7,925$	.
		.

Mais on peut aussi écrire

$$\xi_n = n\pi + \text{arctg } \xi_n$$

(Le terme $n\pi$ est nécessaire, car la fonction arctg a ses valeurs entre $-\frac{\pi}{2}$ et $\frac{\pi}{2}$). L'algorithme de recherche de ξ_n est alors

$$x_{k+1} = n\pi + \text{arctg } x_k ,$$

ce qui correspond à

$$F_n(x) = n\pi + \text{arctg } x .$$

On a ici

$$F_n'(x) = \frac{1}{1 + x^2} < 1 ,$$

et l'algorithme converge dans $\left]0$. Voyons par exemple comment convergent les suites construites pour ξ_1, ξ_2, ξ_3 en partant de $n\pi$.

k	n = 1	n = 2	n = 3
0		2	3
1	4,4042	7,6962	10,890
2	4 4,4891	7,7248	10,904
3	4,4932	7,7252	10,904
4	4,4934	7,7253	
5	4,4934	7,7253	

La situation que nous venons de constater est générale:
<u>Si la fonction F possède une dérivée supérieure en valeur absolue à l'unité, la fonction inverse, si elle existe, possède une dérivée de valeur absolue inférieure à l'unité</u>.

En effet, de

$$y = F(x) \qquad \text{et} \quad x = F^{-1}(y)$$

on déduit

$$\frac{dy}{dx} = F'(x) \qquad \text{et} \quad \frac{dx}{dy} = (F^{-1})'(y) \ ,$$

d'où

$$F'(x) \ . \ (F^{-1})'(y) = 1 \quad .$$

8. COMPARAISON DES VITESSES DE CONVERGENCE. ORDRE ET TAUX DE CONVERGENCE

8.1-Un bon algorithme est avant tout un algorithme qui converge vite, ce qui signifie que l'on en obtient un résultat bien approché après un petit nombre d'itérations. Il est donc naturel de définir des critères permettant de comparer les algorithmes sur ce point.

Dès qu'un algorithme est contractant, on a au moins

$$d(x_{n+1}, \xi) \leq L \, d(x_n, \xi) \qquad , \qquad L < 1 \ ,$$

ce qui signifie que la distance à la solution est multipliée par L à chaque itération. Plus précisément, notons

$$A_n = \frac{d(x_{n+1}, \xi)}{d(x_n, \xi)} \quad .$$

Les nombres A_n devront être aussi petits que possible si l'on veut que la suite des x_n converge vite. On s'intéresse souvent, à ce propos, au comportement de la suite pour n très grand. Si la suite des nombres A_n converge, on définit le taux de convergence asymptotique (à l'ordre un) par

$$\tau = \lim_{n \to \infty} A_n$$

Cette notion est intéressante en ce que, quel que soit ε , il existe une valeur de n à partir de laquelle

$$A_n < \tau + \varepsilon$$

Mais en général, rien ne garantit la convergence de la suite des A_n. Cependant, on peut la remplacer par la suite des B_n définis par

$$B_n = \sup_{k \geq n} A_k \quad .$$

Comme cette suite est décroissante et bornée inférieurement par zéro, elle admet une limite qui constitue la définition la plus générale du taux de convergence asymptotique (à l'ordre 1) :

$$\tau = \lim_{n \to \infty} \left\{ \sup_{k \geq n} \frac{d(x_{k+1}, \xi)}{d(x_k, \xi)} \right\} \ ,$$

ce que l'on note souvent

$$\tau = \limsup_{n \to \infty} \frac{d(x_{k+1}, \xi)}{d(x_k, \xi)} \quad .$$

Cette définition remplit le même office que la précédente, car pour tout ε, il existe une valeur N de n à partir de laquelle

$$\sup_{k \geqslant N} A_k < \tau + \varepsilon \quad ,$$

si bien que l'on a encore pour tout n N

$$A_n < \tau + \varepsilon \quad .$$

(La convergence donnerait plus: $\tau - \varepsilon < A_n < \tau + \varepsilon$, mais précisément, nous n'avons pas besoin de la minoration).

8.2 - Mais on peut imaginer des algorithmes où A_n décroît rapidement avec n, donnant ainsi une convergence de plus en plus rapide à mesure que l'on avance. Il arrive par exemple que

$$A_n \leqslant B\, d^{p-1}(x_n, \xi) \quad , \quad p > 1 \quad ,$$

ce qui implique

$$d(x_{n+1}, \xi) \leqslant B\, d^p(x_n, \xi).$$

Dans ce cas, on dit que l'algorithme est d'ordre p. On définit alors le taux de convergence asymptotique à l'ordre p par

$$= \limsup_{n \to \infty} \frac{d(x_{n+1}, \xi)}{d^p(x_n, \xi)} \quad .$$

On examinera avec attention la figure 13, qui donne une comparaison graphique d'un algorithme du premier ordre avec un algorithme du second ordre.

8.3 - Dans R, on a les lemmes suivants:

Lemme 1 - Soit F une fonction telle que

$$\lim_{x \to \xi} \frac{|F(x) - \xi|}{|x - \xi|^p} = \tau \quad \begin{cases} \neq 0 \\ \neq \infty \end{cases} \quad ,$$

avec $p > 1$. Alors, l'algorithme

$$x_{n+1} = F(x_n)$$

converge vers ξ si le point de départ x_o est suffisamment proche de ξ, son ordre est p et son taux de convergence asymptotique à l'ordre p est τ.

Il existe en effet un nombre η_1 tel que pour $|x - \xi| \leqslant \eta_1$, on ait

$$\frac{|F(x) - \xi|}{|x - \xi|^p} \leqslant \tau + 1 \quad .$$

Soit alors $L < 1$. On a

$$(\tau + 1)\, |x - \xi|^p \leqslant L\, |x - \xi|$$

pour

$$|x - \xi| \leq L(\tau + 1)^{-\frac{1}{p-1}} = \eta_2 .$$

Pour $|x_o - \xi| \leq \eta = \inf(\eta_1, \eta_2)$, la suite $\{x_n\}$ vérifiera donc

$$|x_{n+1} - \xi| = |F(x_n) - \xi| \leq L\,|x_n - \xi| ,$$

ce qui implique la convergence. Dès lors,

$$\lim_{n \to \infty} \frac{|x_{n+1} - \xi|}{|x_n - \xi|^p} = \lim_{n \to \infty} \frac{|F(x_n) - \xi|}{|x_n - \xi|^p} = \tau .$$

Lemme 2 - Soit F une fonction telle que

$$\lim_{x \to \xi} \frac{|F(x) - \xi|}{|x - \xi|} = \tau < 1 .$$

Alors, l'algorithme

$$x_{n+1} = F(x_n)$$

converge vers ξ si le point de départ x_o est suffisamment proche de ξ, son ordre est 1 et son taux de convergence asymptotique est τ.

Soit en effet L un nombre vérifiant

$$\tau < L < 1 .$$

Il existe un nombre η tel que pour $|x - \xi| \leq \eta$, on ait

$$\frac{|F(x) - \xi|}{|x - \xi|} \leq L < 1 .$$

Pour $|x_o - \xi| \leq \eta$, la suite $\{x_n\}$ vérifiera alors

$$|x_{n+1} - \xi| \leq L\,|x_n - \xi|$$

et sera donc convergente. Il en découle

$$\lim_{n \to \infty} \frac{|x_{n+1} - \xi|}{|x_n - \xi|} = \lim_{n \to \infty} \frac{|F(x_n) - \xi|}{|x_n - \xi|} = \tau .$$

8.4 - Voici un critère relatif à l'ordre des algorithmes itératifs dans R : Soit F une fonction dérivable dans ξ, telle que $(F(x) - \xi)$ admette en ξ un zéro de multiplicité $p \geq 1$. Alors, l'algorithme

$$x_{n+1} = F(x_n)$$

est d'ordre p.

En effet, il existe alors une constante τ telle que

$$\lim_{x \to \xi} \frac{|F(x) - \xi|}{|x - \xi|^p} = \tau .$$

8.5 - Si F est p fois différentiable, $(F(x) - \xi)$ admet un zéro de multiplicité p si et seulement si

$$F(\xi) = \xi , \quad F'(\xi) = \ldots = F^{(p-1)}(\xi) = 0 , \quad F^{(p)}(\xi) \begin{cases} \neq 0 \\ \neq \infty \end{cases} .$$

Dans ce cas, on a

$$F(x) - \xi = \frac{(x - \xi)^p}{p!} F^{(p)}(\xi) + o(|x - \xi|^p) ,$$

ce qui entraîne visiblement

$$\tau = \left| \frac{F^{(p)}(\xi)}{p!} \right|$$

9. METHODE DE NEWTON

Pour résoudre l'équation $f(x) = 0$, utilisons le développement

$$f(\xi) = f(x) + (\xi - x)\, f'(x) + o(|\xi - x|) .$$

pour $|\xi - x|$ suffisamment petit, le dernier terme est négligeable et on aura donc $f(\xi) = 0$ pour

$$0 \approx f(x) + (\xi - x)\, f'(x) ,$$

c'est-à-dire

$$\xi \approx x - \frac{f(x)}{f'(x)} = F(x) .$$

La méthode de NEWTON consiste à utiliser cette approximation sous forme itérative:

$$\boxed{x_{n+1} = F(x_n) = x_n - \frac{f(x_n)}{f'(x_n)}}$$

L'interprétation graphique de la méthode de NEWTON est illustrée par la figure 14 : partant de x_0, on progresse suivant la tangente au graphe de f, jusqu'à rencontrer l'axe des x au point x_1; de là, on recommence le processus.

Convergence de la méthode de NEWTON

Il convient d'abord de montrer que $F(\xi) = \xi$, ce qui n'a rien d'évident, car on peut avoir $f'(\xi) = 0$. Ecrivons F(x) sous la forme

$$F(x) = \xi + (x - \xi)\left(1 - \frac{f(x)}{(x - \xi) f'(x)}\right) .$$

Par définition de la multiplicité du zéro, il existe un nombre η tel que pour $|x - \xi| \leq \eta$, on ait

$$\left| \frac{f(x)}{(x - \xi) f'(x)} \right| \leq \frac{1}{m} + 1$$

(pour autant, bien entendu, que $m \neq 0$). Donc, pour m non nul, on a, dans un voisinage de ξ , la relation

$$|F(x) - \xi| \leq |x - \xi| \left(\frac{1}{m} + 2 \right) \longrightarrow 0$$

Cela étant, on a

$$F(x) - F(\xi) = x - \xi - \frac{f(x)}{f'(x)} ,$$

soit

$$\frac{F(x) - F(\xi)}{x - \xi} = 1 - \frac{f(x)}{(x - \xi) f'(x)} ,$$

ce qui entraîne

$$F'(\xi) = \lim_{x \to \xi} \left| 1 - \frac{f(x)}{(x-\xi)f'(x)} \right| = 1 - \frac{1}{m} .$$

Pour un <u>zéro simple</u>, $F'(\xi) = 0$, donc la méthode est du second ordre. Pour $m \neq 1$, on ne peut avoir $|F'(x)| < 1$ dans un voisinage de ξ que si

$$-1 < 1 - \frac{1}{m} < 1$$

soit, d'une part,

$$\frac{1}{m} < 0 \quad , \quad \text{c.-à-d.} \quad m < \infty$$

et, d'autre part,

$$\frac{1}{m} < 2 \quad , \quad \text{c.-à-d.} \quad m > \tfrac{1}{2} .$$

<u>La convergence de la méthode de Newton a donc lieu dans un certain voisinage de la racine pour autant que sa multiplicité m vérifie</u>

$$\tfrac{1}{2} < m < \infty .$$

<u>Pour m = 1 , la convergence est quadratique</u>.

Le taux de convergence à l'ordre 1 pour $m \neq 1$ vaut donc $1 - \frac{1}{m}$.

Le cas où $m = 1$, on calcule successivement

$$F'(x) = 1 - \frac{f'^2(x) - f(x)f''(x)}{f'^2(x)} = \frac{f(x)\, f''(x)}{f'^2(x)} \quad , \quad F'(\xi) = 0$$

$$F''(x) = \frac{(f'(x)f''(x) + f(x)f'''(x))f'^2(x) - 2\, f(x)f'(x)f''^2(x)}{f'^4(x)} ,$$

$$F''(\xi) = \frac{f''(\xi)}{f'(\xi)}$$

ce qui donne finalement le taux de convergence au second ordre

$$\tau = \tfrac{1}{2} \frac{f''(\xi)}{f'(\xi)}$$

10. METHODE DE SHROEDER

10.1 - La méthode de Schröder permet d'obtenir des algorithmes d'un degré aussi élevé que l'on veut. L'idée est la suivante: le graphe $(x, f(x))$ peut tout aussi bien être exprimé sous la forme inverse

$$x = g(y) ,$$

au moins dans le voisinage du zéro, si celui-ci est <u>simple</u>. Alors, la racine cherchée n'est autre que

$$\xi = g(0) .$$

Développons la fonction g en série de Taylor à partir du point y : on aura

$$g(0) = g(y) - y\, g'(y) + \frac{y^2}{2} g''(y) + \ldots + \frac{(-1)^p\, y^p}{p!} g^{(p)}(y) + \ldots$$

et en revenant à f,

$$g(0) = g(y) - g'(y)\, f(x) + \tfrac{1}{2}\, g''(y)\, f^2(x) + \ldots + \frac{(-1)^p}{p!}\, g^{(p)}(y)\, f^p(x) + \ldots$$

Il est aisé d'exprimer les dérivées de g en termes de f, en notant que

$$g'(y) = \frac{1}{f'(x)}$$

$$g''(y) = \frac{d}{dx}(g') \cdot \frac{dx}{dy} = \left(-\frac{f''}{f'^2}\right) \cdot \frac{1}{f'} = -\frac{f''}{f'^3}$$

$$g'''(y) = \frac{1}{f'}\,\frac{d}{dx}(g'') = -\frac{1}{f'}\left(\frac{f''' f'^3 - 3 f'^2 f''^2}{f'^6}\right) = \frac{3 f''^2 - f' f'''}{f'^5}$$

etc ... ,

ce qui donne

$$g(0) = x - \frac{f(x)}{f'(x)} - \frac{f''(x)\, f^2(x)}{2 f'^3(x)} - \frac{3 f''^2(x) - f'(x)\, f'''(x)}{6\, f'^5(x)}\, f^3(x) - \ldots$$

L'algorithme de Schröder numéro m est donné par

$$x_{n+1} = F_m(x_n) \quad ,$$

où F_m est le développement de g(o) arrêté au terme de degré (m-1) en f(x). Pour m = 2, on retrouve la méthode de Newton.

11.2 - Montrons que l'algorithme de Schröder numéro m est d'ordre m. On a, par la formule générale du reste de la formule de Taylor,

$$|F_m(x) - \xi| = |F_m(x) - g(0)|$$

$$= \left| \frac{(-1)^m}{m!}\, g^{(m)}(y^*) \cdot f^m(x) \right| \quad ,$$

y^* étant un point intermédiaire entre y et 0. Pour un zéro de multiplicité égale à 1, on a

$$|f(x)| \leq A\, |x - \xi| \quad ,$$

d'où

$$|F_m(x) - \xi| \leq \left| \frac{g^{(m)}(y^*)}{m!} \right| A^m\, |x - \xi|^m$$

$$\left[\sup_{f(a),\, f(b)} \right] \left| \frac{g^{(m)}(y)}{m!} \right| A^m\, |x - \xi|^m \quad ,$$

a et b étant les bornes de l'intervalle de variation de x.
Pour autant que le zéro soit simple, l'expression de $g^{(m)}$ est finie si $f \in C^m$, ce qui achève la démonstration.

11. METHODE DE LA SECANTE MODIFIEE

Dans certains cas, le dérivée de la fonction f n'est pas calculable ou est trop longue à calculer. On ne peut alors utiliser la méthode de NEWTON. Mais on peut en construire un succédané en remplaçant $f'(x_n)$ par la différence finie (fig. 16)

$$\frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} \quad ,$$

ce qui donne

$$x_{n+1} = x_n - \frac{(x_n - x_{n-1})\, f(x_n)}{f(x_n) - f(x_{n-1})}$$

soit, tous calculs faits,

$$\boxed{x_{n+1} = \frac{x_{n-1}\, f(x_n) - x_n\, f(x_{n-1})}{f(x_n) - f(x_{n-1})}}$$

On retrouve la méthode de la sécante, à cette différence près que, dans le cas présent, on ne s'interdit pas l'extrapolation. Celle-ci constitue, bien entendu, un risque d'instabilité de l'algorithme, mais on obtient à ce prix une amélioration du taux de convergence.

Calcul de l'ordre de convergence asymptotique

Comme il s'agit d'une récurrence à plusieurs points, nous partirons de la définition

$$\tau = \lim_{n \to \infty} \frac{|x_{n+1} - \xi|}{|x_n - \xi|^p} \quad , \text{ fini et non nul.}$$

Il nous faut déterminer p et τ . A cette fin, posons $\varepsilon_n = x_n - \xi$. On a

$$x_{n+1} - \xi = \frac{x_{n-1}\, f(x_n) - x_n\, f(x_{n-1}) - \xi(f(x_n) - f(x_{n-1}))}{f(x_n) - f(x_{n-1})}$$

soit

$$\varepsilon_{n+1} = \frac{\varepsilon_{n-1}\, f(x_n) - \varepsilon_n\, f(x_{n-1})}{f(x_n) - f(x_{n-1})}$$

Supposant alors f deux fois différentiable, on peut écrire

$$f(x) = f'(\xi).\varepsilon + \frac{\varepsilon^2}{2} f''(\xi) + o(\varepsilon^2)$$

Dès lors,

$$\varepsilon_{n-1}\, f(x_n) - \varepsilon_n f(x_{n-1}) =$$

$$\varepsilon_{n-1}\, \varepsilon_n\, f'(\xi) - \varepsilon_n\, \varepsilon_{n-1} f'(\xi) + \tfrac{1}{2}\, \varepsilon_{n-1}\, \varepsilon_n^2\, f''(\xi) - \tfrac{1}{2}\, \varepsilon_n\, \varepsilon_{n-1}^2 f''(\xi)$$

$$+ \varepsilon_{n-1}\, \varepsilon_n^2\, o(\varepsilon_n) - \varepsilon_n\, \varepsilon_{n-1}^2\, o(\varepsilon_{n-1})$$

$$= -\tfrac{1}{2}\, \varepsilon_n\, \varepsilon_{n-1} (\varepsilon_n - \varepsilon_{n-1}) f''(\xi) + o(\varepsilon^3) \quad ;$$

d'autre part,

$$f(x_n) - f(x_{n-1}) = (\varepsilon_n - \varepsilon_{n-1}) f'(\xi) + o(\varepsilon^2) \quad .$$

On en déduit

$$\varepsilon_{n+1} = - \tfrac{1}{2} \varepsilon_n \varepsilon_{n-1} \frac{f''(\xi)}{f'(\xi)} + o(\varepsilon^2) \ ,$$

et nous négligerons le dernier terme. Passant aux logarithmes, on obtient, en notant $d_n = \ln | \varepsilon_n |$,

$$\boxed{d_{n+1} - d_n - d_{n-1} = K \ ,}$$

avec $K = \ln(|f''(\xi)/f'(\xi)|)$. Il s'agit d'une équation aux différences. Avec son second membre, elle admet visiblement la solution particulière

$$d_n = - K \ .$$

La solution générale de l'équation homogène est la somme de deux solutions de la forme

$$d_n = p^n \ .$$

Introduisant cette expression dans l'équation, on trouve

$$p^{n+1} - p^n - p^{n-1} = 0 \ ,$$

soit

$$p^2 - p - 1 = 0 \ ,$$

ce qui mène aux deux valeurs

$$p_1 = \frac{1 + \sqrt{5}}{2} = 1,618... \quad \text{(nombre d'or)}$$

$$p_2 = \frac{1 - \sqrt{5}}{2} = - 0,618...$$

Au total, la solution générale de l'équation complète est

$$d_n = - K + A \, p_1^n + B \, p_2^n$$

et pour n suffisamment grand,

$$d_n \approx A \, p_1^n - K \ .$$

On en déduit aisément

$$d_{n+1} - p_1 \, d_n \approx (p_1 - 1)K \ ,$$

soit, en prenant les exponentielles des deux membres,

$$\frac{| \varepsilon_{n+1} |}{| \varepsilon_n |^{p_1}} \doteq \exp ((p_1 - 1)K) = \tau \ , \text{ fini et non nul.}$$

<u>L'ordre de convergence asymptotique est donc</u> $p_1 = 1,618...$

Bien qu'il ne s'agisse pas de convergence quadratique, c'est de loin supérieur à une convergence d'ordre 1. Le <u>taux de convergence</u> à l'ordre p_1 se calcule aisément à partir de la définition de K:

$$\tau = \left| \frac{f''(\xi)}{f'(\xi)} \right| 0,618...$$

12. METHODE DE LA PARABOLE

Par les points x_n, x_{n-1}, x_{n-2}, on fait passer une parabole d'équation

$$\tilde{f}(x) = f(x_n) + a(x - x_n) + b(x - x_n)^2 \quad .$$

On calcule aisément les coefficients a et b par

$$b = \frac{\dfrac{f(x_{n-1}) - f(x_n)}{x_{n-1} - x_n} - \dfrac{f(x_{n-2}) - f(x_n)}{x_{n-2} - x_n}}{x_{n-1} - x_{n-2}}$$

$$a = \frac{f(x_{n-1}) - f(x_n)}{x_{n-1} - x_n} - b(x_{n-1} - x_n) \quad .$$

On cherche alors le point de rencontre x_{n+1} du graphe de $\tilde{f}$ avc l'axe des x : il s'agit de résoudre l'équation du second degré

$$b(x_{n+1} - x_n)^2 + a(x_{n+1} - x_n) + f(x_n) = 0 \quad .$$

Des deux solutions

$$x_{n+1} - x_n = \frac{-a \pm \sqrt{a^2 - 4\,b\,f(x_n)}}{2\,b}$$

il faut évidemment choisir celle dont le radical est affecté du signe positif, car à la limite, si $f(x_n) = 0$, on doit obtenir $x_{n+1} = x_n$. Du reste, on améliore le conditionnement numérique en multipliant haut et bas par

$$-a - \sqrt{a^2 - 4\,b\,f(x_n)} \quad ,$$

ce qui donne finalement

$$x_{n+1} = x_n - \frac{2f(x_n)}{a + \sqrt{a^2 - 4\,b\,f(x_n)}}$$

13. RESOLUTION DES SYSTEMES D'EQUATIONS NON LINEAIRES

Dans ce qui suit, nous ferons les conventions de notations suivantes: les vecteurs de R^n seront notés par ses minuscules latines: a, b, ... Ce sont toujours des vecteurs-colonnes. Pour désigner les mêmes vecteurs écrits sous forme de lignes, nous utiliserons la notation a^T, b^T, ... Les lettres latines majuscules désignent des matrices. Enfin, pour les scalaires, nous utiliserons des lettres grecques.

Soit donc à résoudre le système

$$f(x) = 0 .$$

On utilisera souvent une méthode itérative de la forme

$$x^{(k+1)} = g(x^{(k)}) .$$

Mais pour obtenir des conditions suffisantes de convergence, il ne suffit pas, comme dans R, d'invoquer le théorème de Taylor, car on aurait

$$g_i(y) - g_i(x) = \sum_k D_k g_i(x + \theta_i(y - x)) ,$$

avec θ_i différent pour chaque i. La bonne méthode est donnée par le théorème suivant [7]: Soit $J_{ij}(x) = D_j g_i(x)$. Alors, si g est continûment différentiable sur le segment (x, x + Δx) , on a

$$\| g(x + \Delta x) - g(x) \|_1 \leq \| J(x + \theta \Delta x) \|_1 \, \| \Delta x \|_1$$

pour un certain $\theta \in]0 , 1[$.

Considérons en effet la fonction

$$\phi(\tau) = \sum_{i=1}^{n} \varepsilon_i \left[g_i(x + \tau \Delta x) - g_i(x) \right] ,$$

les ε_i étant des nombres fixes $\in [0 , 1]$. On a $\phi \in C^1([0 , 1])$, d'où, par le théorème des accroissements finis, comme $\phi(0) = 0$, il existe $\theta \in]\, 0, 1 \,[$ tel que

$$\sum_{i=1}^{n} \varepsilon_i \left[g_i(x + \Delta x) - g_i(x) \right] = \phi(1) - \phi(0) = \phi'(\theta) .$$

Or,

$$\phi'(\theta) = \sum_{i=1}^{n} \varepsilon_i \left[\sum_{j=1}^{n} D_j g_i(x + \theta \Delta x) \, \Delta x_j \right] = \sum_{i=1}^{n} \varepsilon_i \sum_{j=1}^{n} J_{ij}(x + \theta \Delta x) \, \Delta x_j$$

et, comme les $|\varepsilon_i| \in [0 , 1]$,

$$|\phi'(\theta)| \leq \sum_{i=1}^{n} |\varepsilon_i| \, |(J(x + \theta \Delta x) \Delta x)_i|$$

$$\leq \sum_{i=1}^{n} |(J(x + \theta \Delta x) \Delta x)_i| = \| J(x + \theta \Delta x) \Delta x \|_1$$

$$\| J(x + \theta \Delta x) \|_1 \, \| \Delta x \|_1 .$$

Dans le cas particulier où l'on choisit

$$\varepsilon_i = \text{sign} \left[g_i(x + \Delta x) - g_i(x) \right] ,$$

on obtient

$$\sum_{i=1}^{n} |g_i(x + \Delta x) - g_i(x)| \leq ||J(x + \theta \Delta x)||_1 \; ||\Delta x||_1 ,$$

soit

$$||g(x + \Delta x) - g(x)||_1 \leq ||J(x + \theta \Delta x)||_1 \; ||\Delta x||_1$$

comme annoncé.

Il résulte de ce théorème que <u>la contraction est assurée dans un ensemble V si</u> $||J(x)||_1 \leq \alpha < 1$ <u>dans V.</u>

<u>14. METHODE DE NEWTON-RAPHSON</u>

Soit un système $f(x)=0$ admettant une solution $s \in R^n$. Développant en série de Taylor limitée, on a

$$0 = f_i(s) = f_i(x) + \sum_j D_j f_i(x) \, (s_j - x_j) + o(||s - x||)$$

Définissant la matrice des dérivées

$$F_{ij}(x) = D_j f_i(x) ,$$

on aura donc

$$F(x) \, (s - x) \approx - f(x) ,$$

soit, si F est inversible,

$$s - x \approx - F^{-1}(x) \, f(x) .$$

Cette relation suggère l'algorithme

$$\boxed{x^{(k+1)} = x^{(k)} - F^{-1}(x^{(k)}) \, f(x^{(k)})}$$

dit de <u>Newton-Raphson</u>. C'est la généralisation à n dimensions de l'algorithme de Newton.

Nous démontrerons la convergence de cet algorithme dans le cas où l'on sait qu'il existe une solution s et que, dans une certaine boule $B_\rho(s)$, on a les relations suivantes:

$$\begin{cases} a) \quad ||F^{-1}(x)||_1 \leq \frac{1}{\alpha} \quad , \quad \alpha > 0 \\ b) \quad ||F(x) - F(y)||_1 \leq \lambda \; ||x - y||_1 \quad , \quad 0 < \lambda < \infty \end{cases}$$

La première de ces conditions signifie que la matrice des dérivées F est inversible au voisinage de la solution (zéro simple). La seconde est une condition de Lipschitz sur ces dérivées.

La démonstration se fait en deux étapes:

A. On montre d'abord que

$$\|f(x+\Delta x) - f(x) - F(x)\Delta x\|_1 \leq \|F(x+\theta\Delta x) - F(x)\|_1 \, \|\Delta x\|_1 \ ,$$

avec $\theta \in]0,1[$. Il suffit pour cela de s'inspirer de la démonstration du 13 ci-dessus, et de considérer la fonction auxiliaire

$$\phi(\tau) = \sum_i \varepsilon_i \left[f_i(x+\tau\Delta x) - f_i(x) - \sum_j F_{ij}(x)\,\tau\,\Delta x_j \right]$$

avec $-1 \leq \varepsilon_i \leq 1$. Cette fonction est continûment dérivable et s'annule en $\tau = 0$. Par le théorème des accroissements finis, il existe donc un θ strictement compris entre 0 et 1, tel que $\phi(1) = \phi'(\theta)$. Or,

$$\phi'(\theta) = \sum_i \varepsilon_i \left[\sum_j D_j f_i(x+\theta\Delta x)\,\Delta x_j - \sum_j F_{ij}(x)\,\Delta x_j \right]$$

$$= \sum_i \varepsilon_i \left[(F(x+\theta\Delta x) - F(x)) \right]_i$$

et, comme tous les ε_i sont inférieurs en valeur absolue à 1,

$$|\phi'(\theta)| \leq \sum_i |((F(x+\theta\Delta x) - F(x))\Delta x)_i| = \|(F(x)+\theta\Delta x) - F(x))\,\Delta x\|_1$$

$$\leq \|F(x+\theta\Delta x) - F(x)\|_1 \, \|\Delta x\|_1 \ .$$

Choisissant alors

$$\varepsilon_i = \operatorname{sign}(f_i(x+\Delta x) - f_i(x) - \sum_j F_{ij}(x)\Delta x_j) \ ,$$

on obtient

$$\phi(1) = \|f(x+\Delta x) - f(x) - F(x)\Delta x\|_1 \ ,$$

d'où l'inégalité annoncée.

B. On a alors, en posant

$$g(x) = x - F^{-1}(x)\, f(x) \ ,$$

l'égalité

$$g(x) - s = x - s - F^{-1}(x)\, f(x)$$

soit, comme $f(s) = 0$,

$$g(x) - s = F^{-1}(x) \left[f(s) - f(x) - F(x)\,(s-x) \right] .$$

On en déduit, en utilisant l'inégalité démontrée en A,

$$\|g(x) - s\|_1 \leq \|F^{-1}(x)\|_1 \, \|f(s) - f(x) - F(x)\,(s-x)\|_1$$

$$\leq \|F^{-1}(x)\|_1 \, \|F(x+\theta(s-x)) - F(x)\|_1 \, \|s-x\|_1 \ .$$

Tenant enfin compte des hypothèses a et b , on obtient

$$\|g(x) - s\|_1 \leq \frac{\lambda}{\alpha}\,\theta\,\|s - x\|_1^2 \leq \frac{\lambda}{\alpha}\,\|s - x\|_1^2 \quad ,$$

ce qui montre que l'algorithme est <u>du second ordre</u>. La convergence sera assurée si au départ,

$$\|s - x^{(0)}\|_1 \leq \frac{\alpha}{\lambda}\beta \quad , \qquad 0 < \beta < 1 \quad .$$

15. METHODE DE NEWTON-RAPHSON MODIFIEE

Le calcul de la matrice F à chaque itération étant généralement très long, on préfère bien souvent ne le faire que de temps en temps, et utiliser la récurrence simplifiée

$$x^{(k+1)} = x^{(k)} - F^{-1}(x^{(0)})\, f(x^{(k)}) \quad .$$

Eventuellement, on réactualise de temps en temps la matrice F pour accélérer (ou simplement garantir) la convergence. Le choix de la stratégie de calcul (mettre à jour fréquemment ou rarement) demande une certaine expérience du genre de problèmes à traiter.

* La convergence de cet algorithme (lorsque l'on ne remet pas à jour) peut être démontrée dans le cadre des hypothèses suivantes:

$$\begin{cases} \text{a)}\ \|F^{-1}(x^{(0)})\|_1 = \alpha < \infty \quad , \\ \text{b) Dans une boule } B_\rho(s), \text{ on a } \|F(x) - F(y)\| \leq \lambda\,\|x - y\| \quad , \\ \qquad\qquad \text{avec} \quad 0 < \lambda < \infty \, . \end{cases}$$

La démonstration est très voisine de celle de la convergence de la méthode de Newton-Raphson.

A. <u>On a la relation</u>

$$\|f(x) - F(x^{(0)})\,(x - s)\|_1 \leq \|F(s + \theta(x - s)) - F(x^{(0)})\|_1\,\|x - s\|_1 \quad ,$$

<u>avec $0 < \theta < 1$.</u> Il suffit, pour le montrer, de considérer la fonction

$$\phi(\tau) = \sum_i \varepsilon_i \left[f_i(s + \tau(x - s)) - \sum_j F_{ij}(x^{(0)})\,\tau\,(x_j - s_j) \right] \quad ,$$

avec $-1 \leq \varepsilon_i \leq 1$. Cette fonction est nulle en $\tau = 0$, ce qui permet d'affirmer l'existence d'un θ strictement compris entre 0 et 1 tel que $\phi(1) = \phi'(\theta)$. Or,

$$\phi'(\theta) = \sum_i \varepsilon_i \sum_j \left[F_{ij}(s + \theta(x - s)) - F_{ij}(x^{(0)}) \right] (x_j - s_j)$$

et, comme tous les ε_i sont inférieurs en module à l'unité,

$$|\phi'(\theta)| \leq \sum_i \left| \left[(F(s + \theta(x - s)) - F(x^{(0)}))\,(x - s) \right]_i \right|$$

$$= \|(F(s + \theta(x - s)) - F(x^{(0)}))\,(x - s)\|_1$$

$$\leq \|F(s+\theta(x-s)) - F(x^{(0)})\|_1 \, \|x - s\|_1 \ .$$

Choisissant d'autre part

$$\varepsilon_i = \operatorname{sign}(f_i(x) - \sum_j F_{ij}(x^{(0)}) \, (x_j - s_j)) \ ,$$

on obtient

$$\phi(1) = \|f(x) - F(x^{(0)}) \, (x - s)\|_1 \ ,$$

d'où l'inégalité annoncée.

B. On a alors

$$g(x) - s = x - s - F(x^{(0)}) \, f(x) = F^{-1}(x^{(0)}) \, [\, F(x^{(0)}) \, (x - s) - f(x) \,] \ ,$$

ce qui entraîne

$$\begin{aligned} \|g(x) - s\|_1 &\leq \alpha \, \|F(s+\theta(x-s)) - F(x^{(0)})\|_1 \, \|x - s\|_1 \\ &\leq \alpha \lambda \, \|s + \theta(x - s) - x^{(0)}\|_1 \, \|x - s\|_1 \\ &\leq \alpha \lambda \, [\, \|s - x^{(0)}\|_1 + \theta \, \|s - x\|_1 \,] \, \|x - s\|_1 \\ &\qquad [\, \|s - x^{(0)}\|_1 + \|s - x\|_1 \,] \, \|x - s\|_1 \ . \end{aligned}$$

En particulier, pour $x = x^{(0)}$, on a $g(x) = x^{(1)}$ et

$$\|x^{(1)} - s\|_1 \leq \alpha . \, 2\lambda \, \|x^{(0)} - s\|_1 \, \|x^{(0)} - s\|_1 \leq \beta \, \|x^{(0)} - s\|_1$$

avec $\beta < 1$ dès que

$$\|x^{(0)} - s\|_1 \leq \frac{\beta}{2 \lambda \alpha} \ .$$

Si cette condition est vérifiée, on a automatiquement

$$\|x^{(k)} - s\|_1 \leq \beta \|x^{(0)} - s\|_1$$

pour tout k. En effet, c'est vrai pour k = 1. Si c'est vrai pour la valeur (k-1) de l'indice, on a

$$\begin{aligned} \|x^{(k)} - s\|_1 &\leq \alpha\lambda \, [\, \|x^{(0)} - s\|_1 + \|x^{(k-1)} - s\|_1 \,] \, \|x^{(k-1)} - s\|_1 \\ &\leq \alpha\lambda \, (1 + \beta) \, \|x^{(0)} - s\|_1 . \beta \, \|x^{(0)} - s\|_1 \\ &\leq 2 \lambda \, \|x^{(0)} - s\|_1 \, \alpha \, \|x^{(0)} - s\|_1 \leq \beta \, \|x^{(0)} - s\|_1 \ . \end{aligned}$$

Dès lors, on a tout au cours de l'itération

$$\|x^{(k+1)} - s\|_1 \leq 2\alpha\lambda \, \|x^{(0)} - s\|_1 \, \|x^{(k)} - s\|_1 \leq \beta \, \|x^{(k)} - s\|_1 \ ,$$

ce qui garantit la convergence à l'ordre 1 dès que le point de départ vérifie

$$\|x^{(0)} - s\|_1 \leq \frac{\beta}{2\alpha\lambda} \qquad , \quad 0 < \beta < 1 \ .$$

Exercice 1 - Soit f une fonction admettant en un zéro d'ordre $p \neq 1$, mais connu. Montrer que l'algorithme

$$x_{n+1} = x_n - p \frac{f(x_n)}{f'(x_n)}$$

est d'ordre deux.

Solution - On a donc

$$F(x) = x - p \frac{f(x)}{f'(x)}$$

et, comme $F(\xi) = \xi$,

$$F(x) - F(\xi) = (x - \xi) - p \frac{f(x)}{f'(x)} \quad ,$$

soit

$$\frac{F(x) - F(\xi)}{x - \xi} = 1 - p \frac{f(x)}{(x-\xi)f'(x)}$$

et

$$F'(x) = 1 - \frac{p}{p} = 0 \ .$$

Exercice 2 - On considère l'équation

$$x^2 = \ln (x + 1) \ .$$

On demande

a) De situer les racines, s'il y en a.

b) D'en chercher une approximation à 4 chiffres significatifs par une méthode d'itération spécialement conçue à cet effet, dont on aura au préalable démontré la convergence

Exercice 3 - On considère l'itération suivante:

$$x_{n+1} = x_n + \tfrac{1}{2} (y - x_n)^2 \quad ,$$

avec $y \in]\, 0 \ , \ 1 \,]$, $x_0 = 0$. Démontrer la convergence et trouver la limite.

Solution - Si le processus converge, la limite doit vérifier

$$\xi = \xi + \tfrac{1}{2} (y - \xi^2) \quad ,$$

soit $\xi = \pm \sqrt{y}$. Posant $x_{n+1} = F(x_n)$, on a ici

$$F'(x) = 1 - x \quad ,$$

donc $|F'(x)| > 1$ pour $x < 0$. On ne peut donc converger que vers $+\sqrt{y}$. Pour cette dernière racine, on remarque que $0 < F'(x) < 1$ dans $]0 \ , \ 1[$, donc la convergence est assurée.

Exercice 4 - Résoudre l'équation $x \ln x = 1$

Solution - Situons d'abord les racines. $x \ln x$ n'a de sens réel que pour $x \geqslant 0$. Pour $x = 0$,

$$\lim_{x \to 0} \; x \ln x = \lim \frac{\ln x}{1/x} = \lim \frac{1/x}{\frac{-1}{x^2}} = \lim (-x) = 0 \; .$$

On a d'ailleurs $x \ln x = 0$ en $x = 1$. Comme

$$D_x(x \ln x) = \ln x + 1 > 0 \quad \text{pour} \quad x > \frac{1}{e} \; , \quad < 0 \text{ pour } x < \frac{1}{e} \; ,$$

$$= -\infty \text{ pour } x = 0 \; ,$$

comme par ailleurs , $x \ln x \longrightarrow \infty$ pour $x \longrightarrow \infty$, son graphe a la forme représentée en figure 17. On en déduit que la racine cherchée se situe en un point $x > 1$. Elle est unique. Pour la trouver, on peut utiliser la méthode de Newton: comme

$f(x) = x \ln x - 1$, $f'(x) = \ln x + 1$, on a

on a

$$x_{n+1} = x_n - \frac{x_n \ln x_n - 1}{\ln x_n + 1} = \frac{x_n + 1}{\ln x_n + 1}$$

Partant de $x_o = 1$, on obtient successivement

$x_1 = 2$, $x_2 = 1{,}772$, $x_3 = 1{,}763$, $X_4 = 1{,}763$.

Exercice 5 - Résoudre l'équation $x - \sin x = 1/4$

Solution - Un diagramme (fig. 18) montre qu'il n'existe qu'une racine ξ , qui est positive. Elle est enoutre inférieure à $\pi/2$, car $\pi/2 - 1 = 0{,}57\ldots > 0{,}25$.

On utilise la méthode suivante:

$$x_{n+1} = \sin x_n + 1/4 = F(x_n) \; ,$$

pour laquelle

$$F'(x) = \cos x \in \;]0, \; 1[\quad \text{pour} \quad x \in \;]0, \; \pi/2 \; [$$

Partant de $x_o = \pi/2$, on obtient

$x_1 = 1{,}250$	$x_5 = 1{,}173$
$x_2 = 1{,}199$	$x_6 = 1{,}173$
$x_3 = 1{,}182$	$x_7 = 1{,}171$
$x_4 = 1{,}175$	$x_8 = 1{,}171$

Exercice 6 - Trouver les trois premières racines strictement positives de l'équation $\operatorname{tg} x = \operatorname{th} x$

Solution - $x_1 = 3{,}9266023$

$x_2 = 7{,}0685827$

$x_3 = 10{,}210176$

Exercice 7 - Montrer que si, entre la racine ξ et le point de départ x_o, il n'y a pas de second zéro de f et $f(x).f''(x) > 0$, la convergence de la méthode de Newton est monotone.

Solution: Développons f en série de Taylor limitée au second ordre autour du point x_o:

$$0 = f(\xi) = f(x_o) + (\xi - x_o)\, f'(x_o) + \frac{(\xi - x_o)^2}{2}\, f''(z)$$

avec z compris entre x_o et ξ . Divisons par $f(x_o)$: il vient

$$0 = 1 + (\xi - x_o)\frac{f'(x_o)}{f(x_o)} + \frac{(\xi - x_o)^2}{2}\,\frac{f''(z)}{f(x_o)} \quad ,$$

et le dernier terme est positif, ce qui entraîne

$$(\xi - x_o)\frac{f'(x_o)}{f(x_o)} < -1 \quad .$$

Supposons d'abord $f(x_o).f'(x_o) > 0$. Il vient

$$\xi - x_o < -\frac{f(x_o)}{f'(x_o)}$$

et

$$\xi < x_o - \frac{f(x_o)}{f'(x_o)} < x_o \quad .$$

A l'inverse, si $f(x_o).f'(x_o) < 0$, on a

$$\xi - x_o > -\frac{f(x_o)}{f'(x_o)}$$

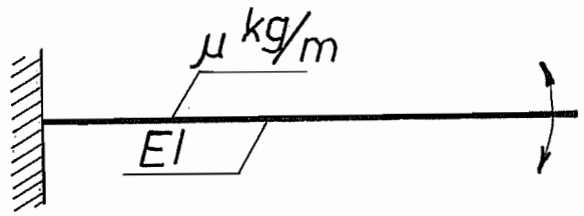
et

$$\xi > x_o - \frac{f(x_o)}{f'(x_o)} > x_o \quad .$$

Le même raisonnement reste valable pour les itérations suivantes. La suite des x_k est donc monotone et bornée, donc elle converge. Puisqu'à la limite, on doit avoir

$$x = x - \frac{f(x)}{f'(x)} = 0 \quad ,$$

on a donc certainement $f(x) = 0$, pour autant que la dérivée ne soit pas nulle (zéro simple) .



$$\beta^4 = \frac{\mu}{EI}\,\omega^2$$

$$\cos\beta l = -\frac{1}{ch\,\beta l}$$

Fig. 1

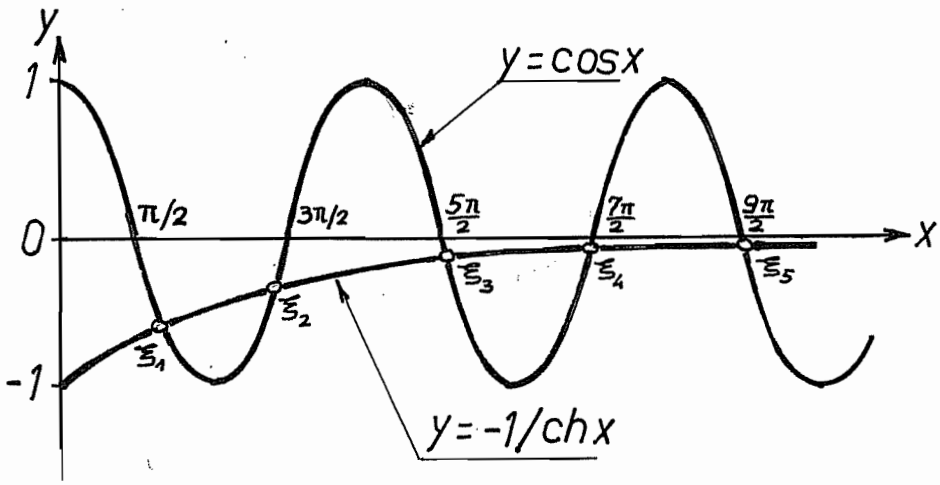


Fig. 2

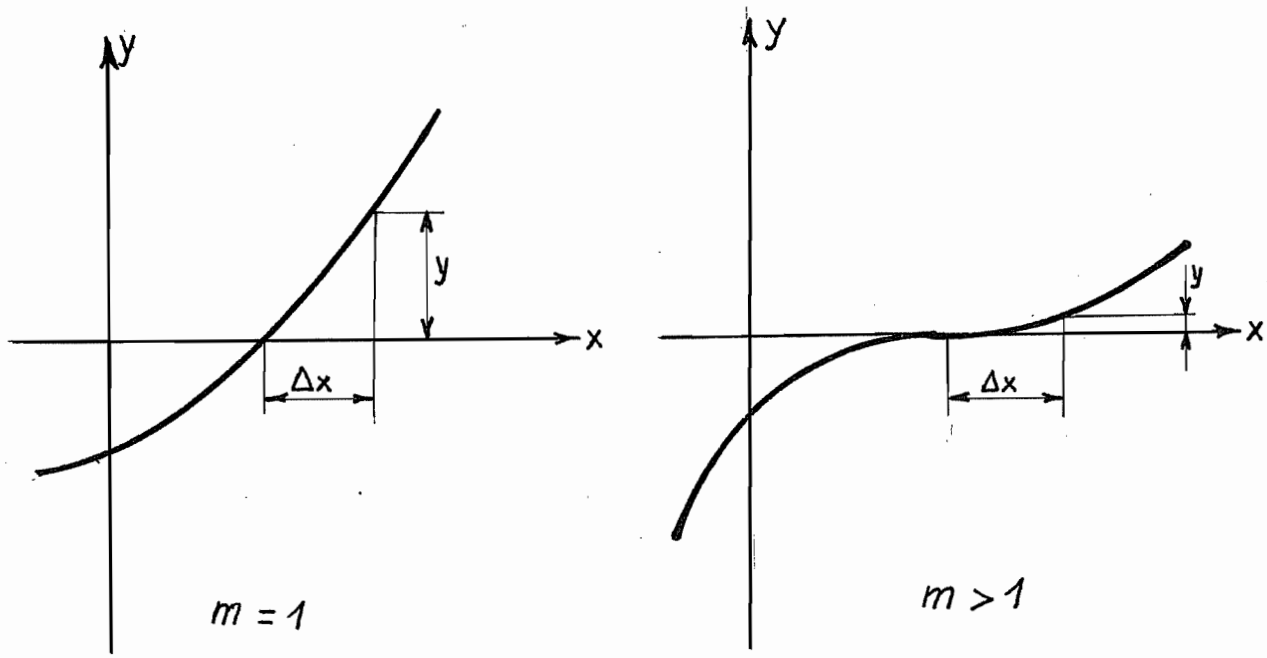


Fig. 3

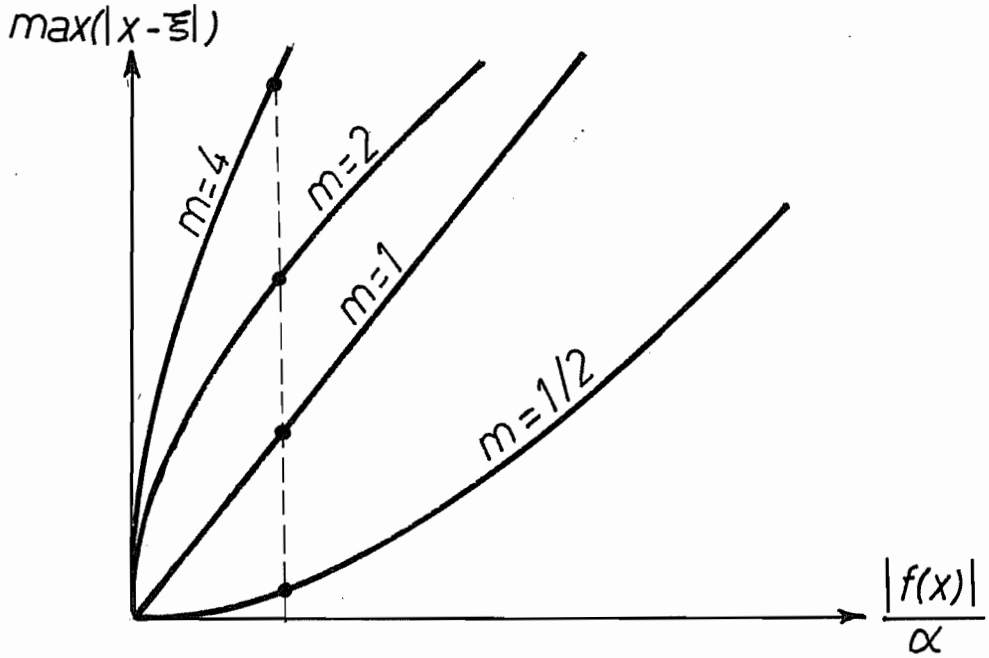
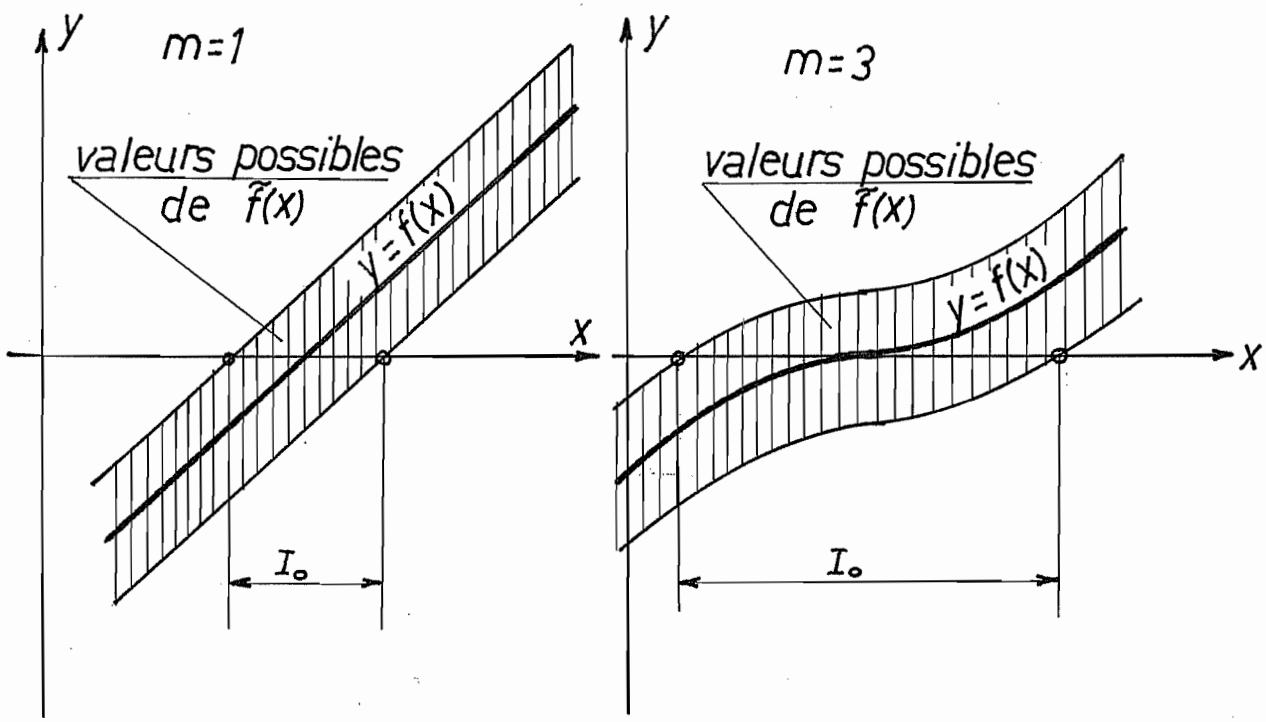


Fig. 4



I_o = intervalle où $\tilde{f}(x)$ peut = 0

Fig. 5

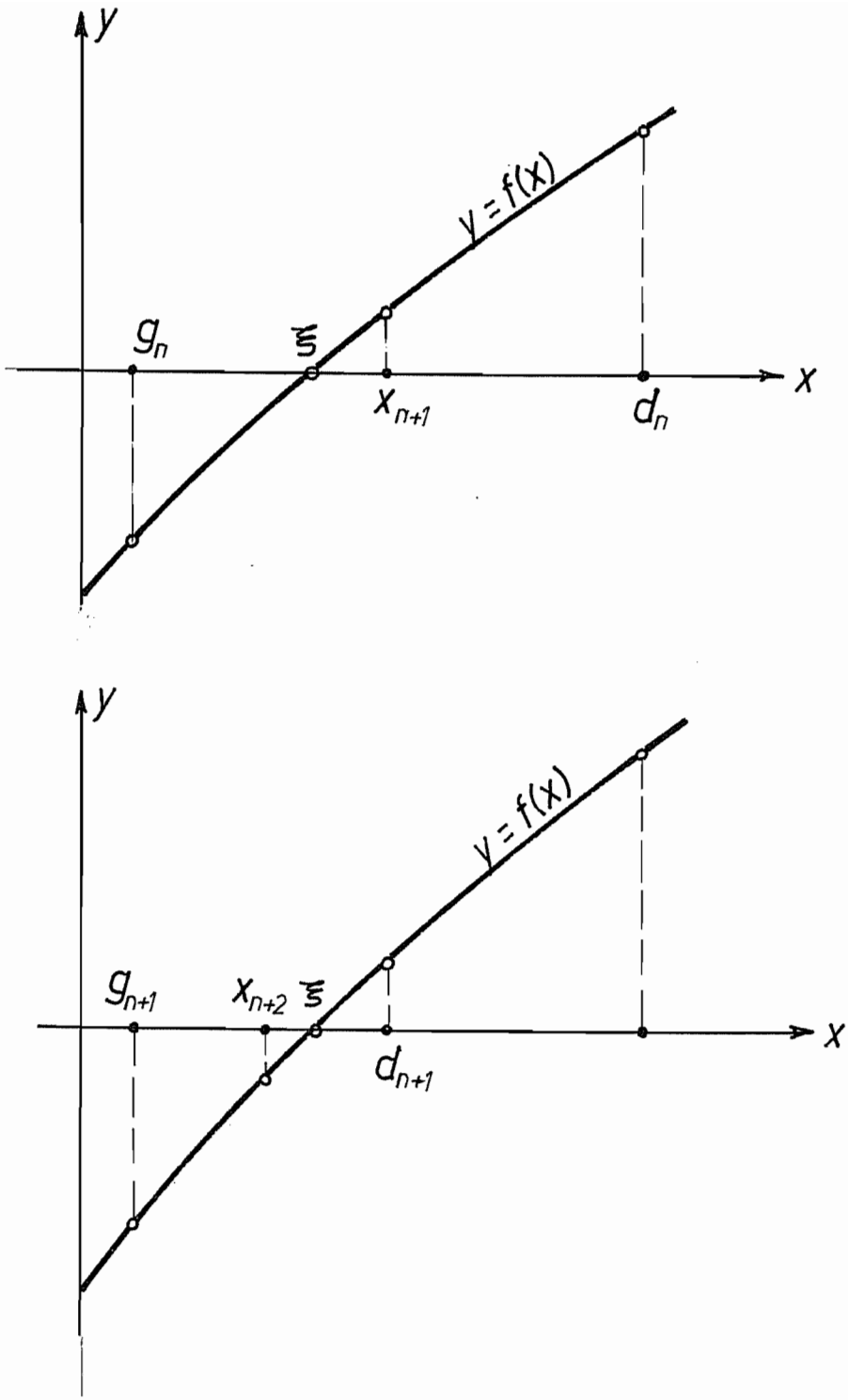


Fig. 6

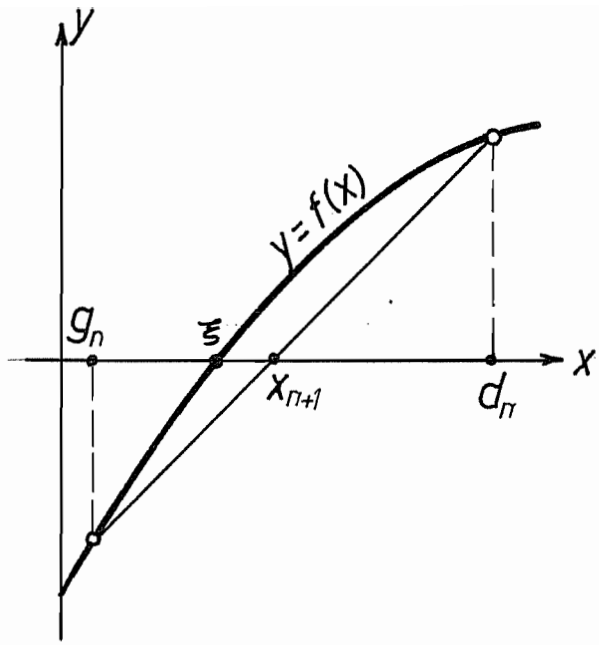


Fig. 7

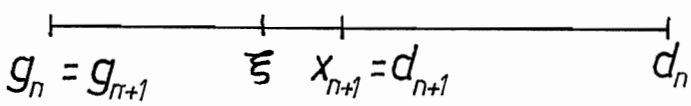


Fig. 8

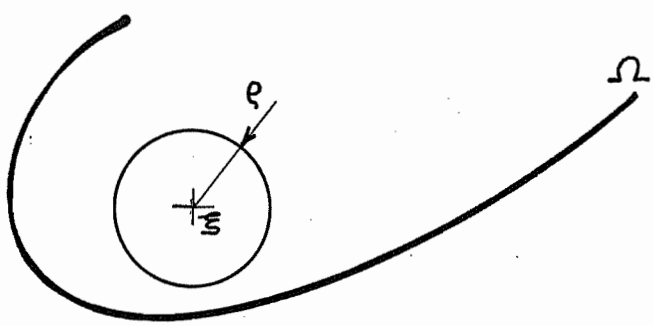


Fig. 9

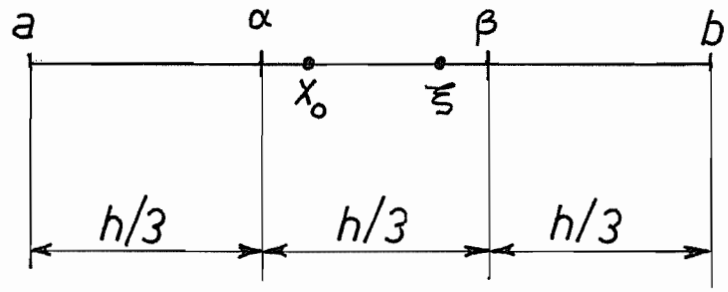
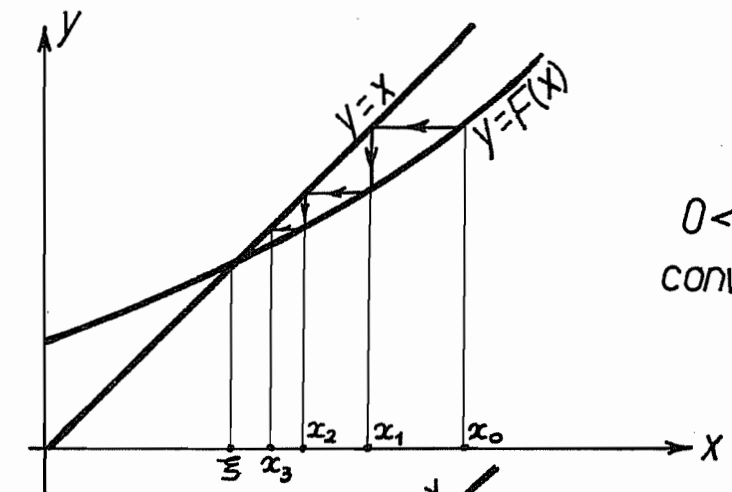
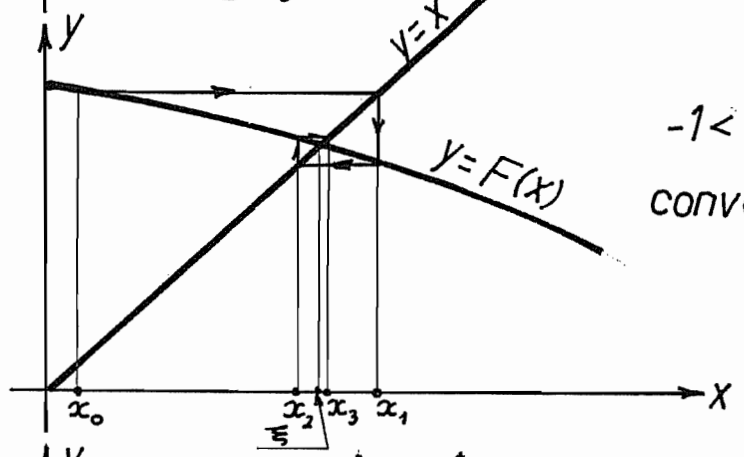


Fig. 10



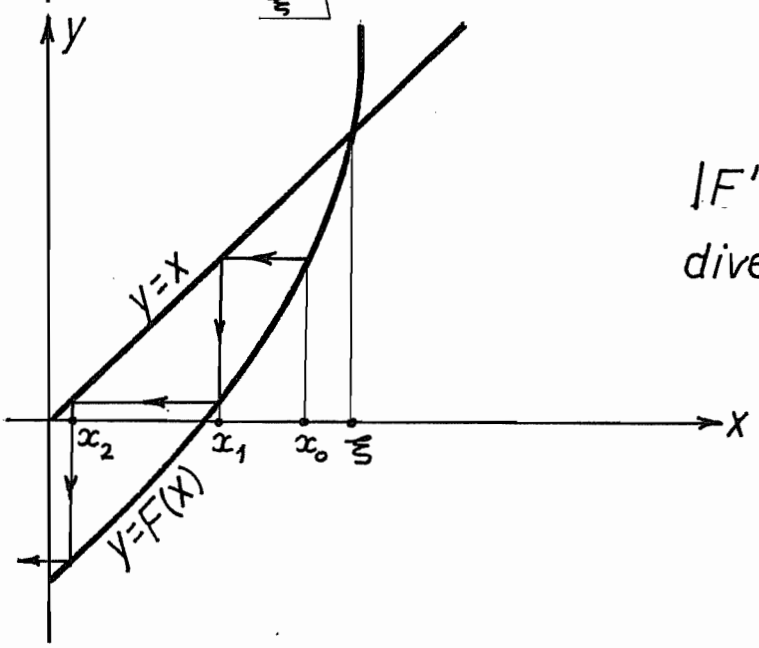
$0<F'(x)<1$

convergence monotone



$-1<F'(x)<0$

convergence oscillante



$|F'(x)| > 1$

divergence

Fig. 11

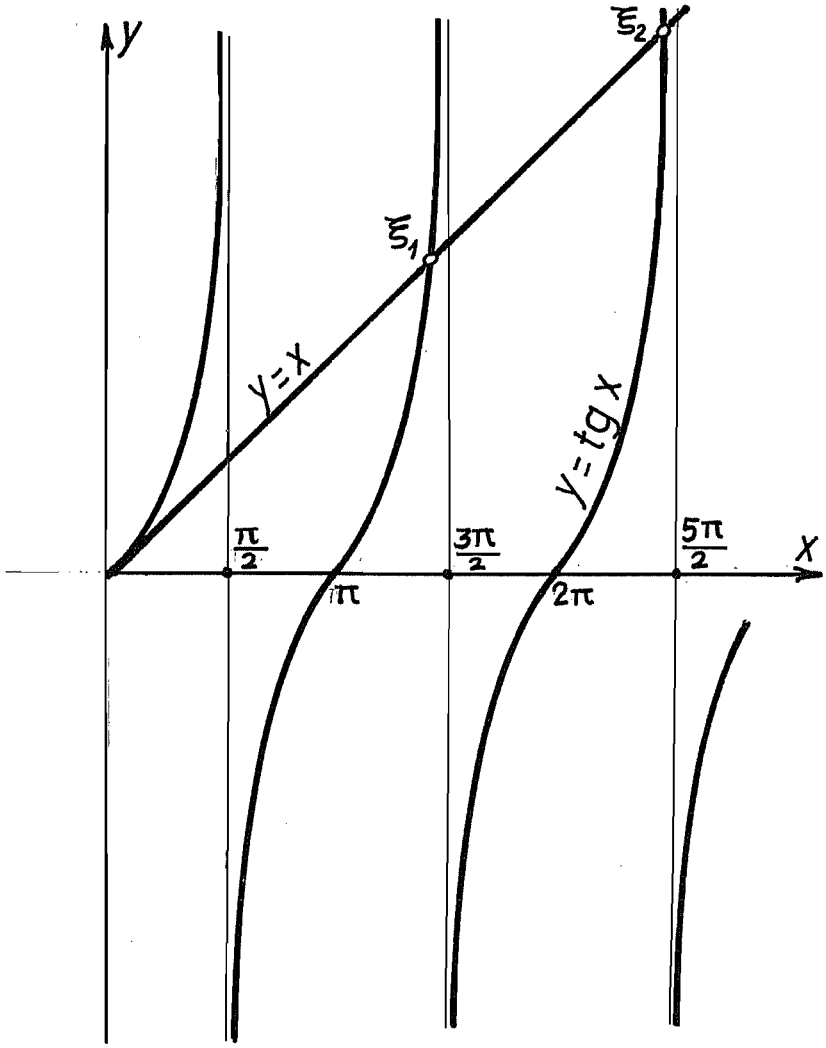
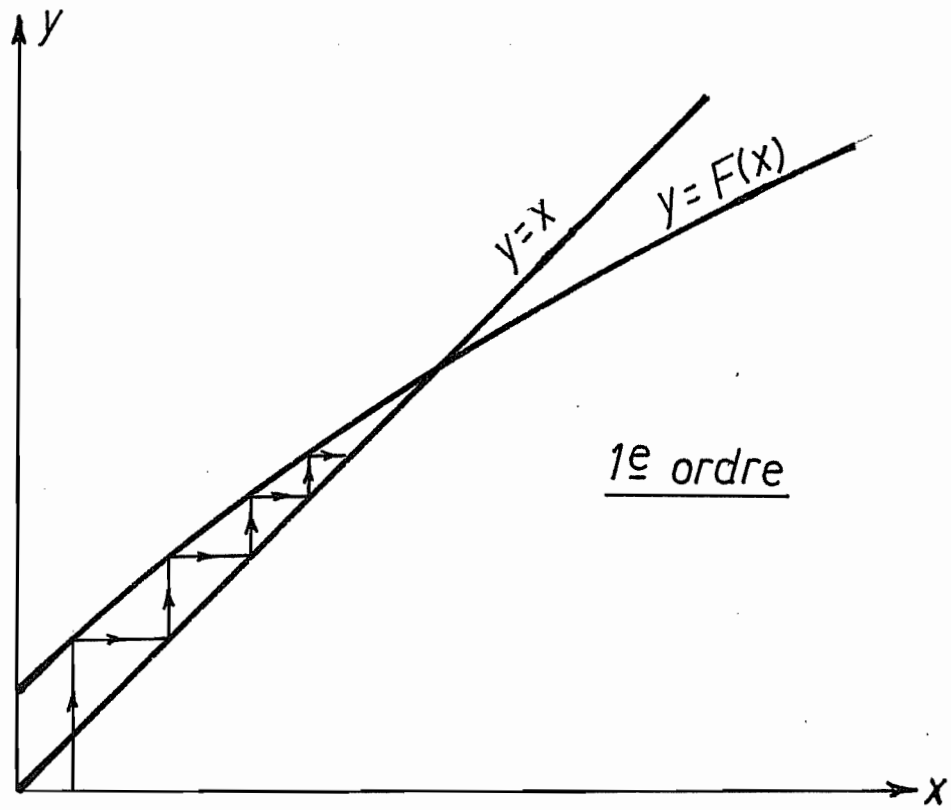


Fig. 12



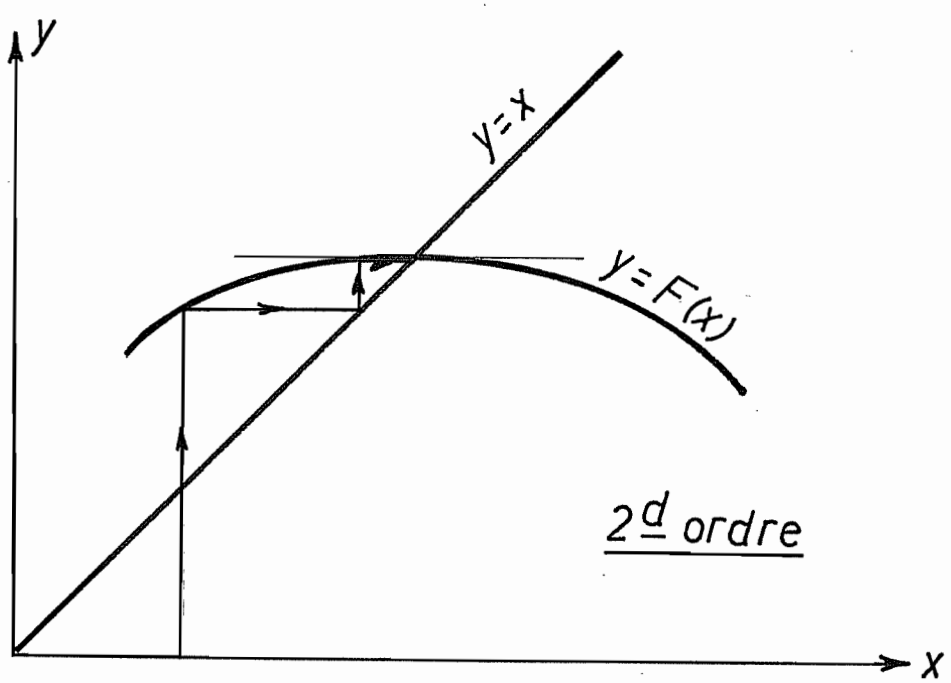


Fig. 13

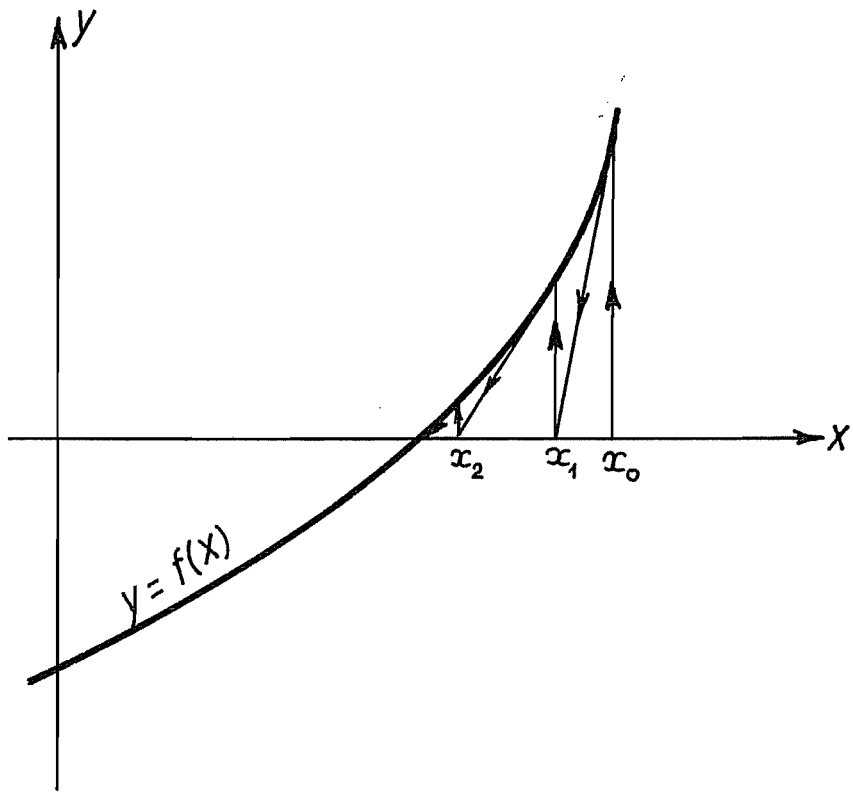


Fig. 14

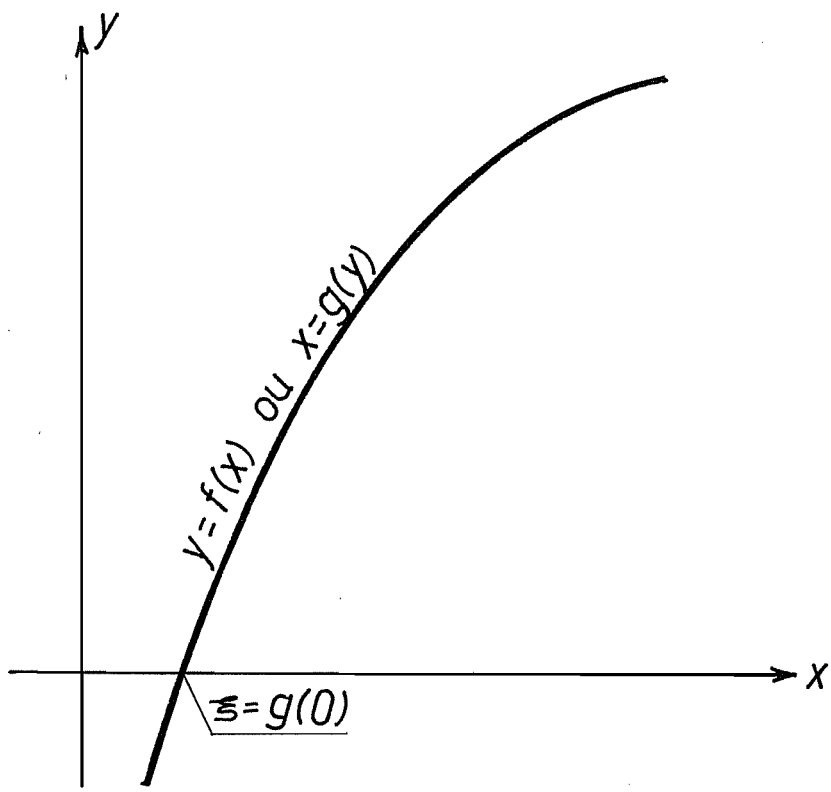


Fig. 15

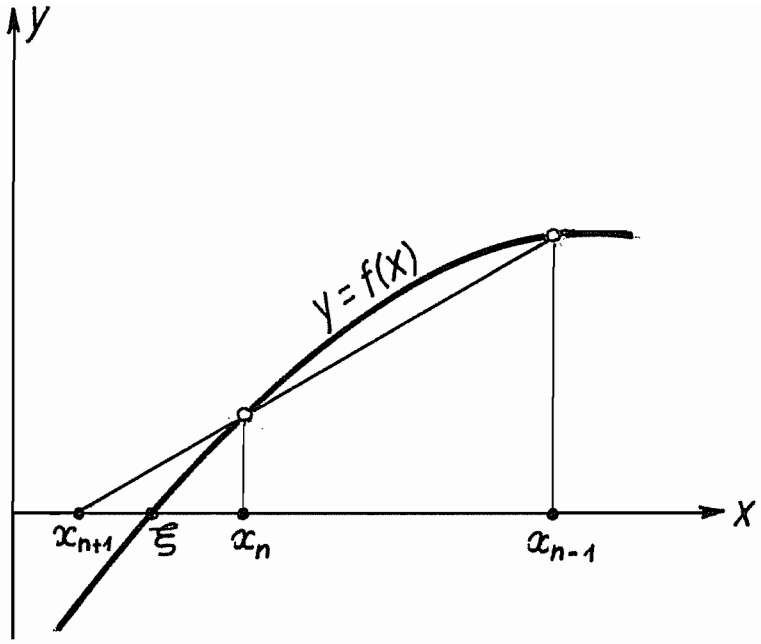


Fig. 16

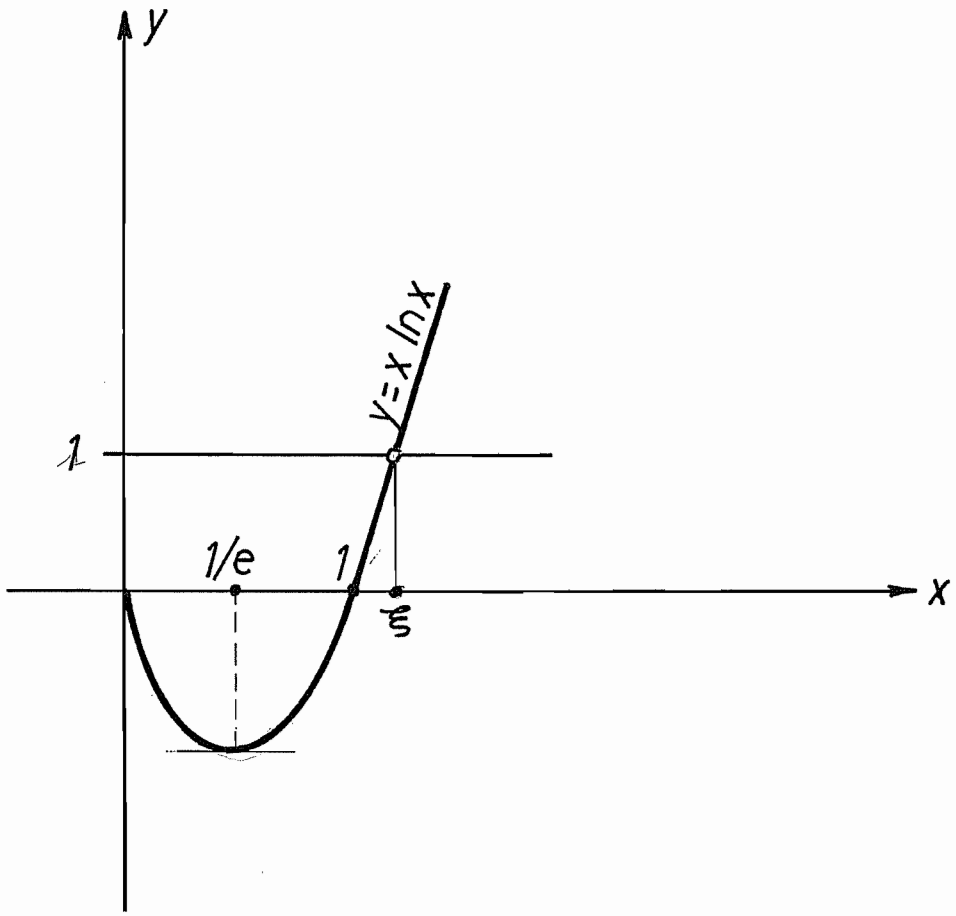


Fig. 17

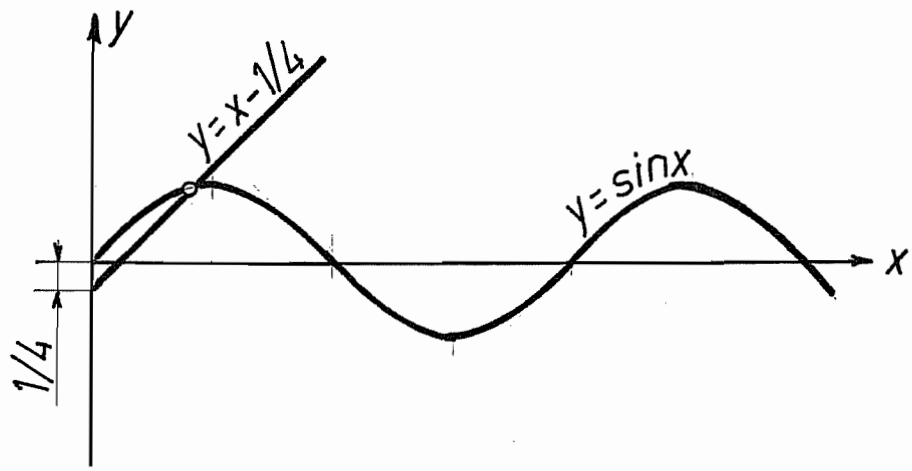


Fig. 18

ISBN-13 : 978-9600313-5-5