ANALYSE NUMÉRIQUE

ANALYSE NUMÉRIQUE

JEAN-PAUL CALVI

0.6.1

UPS
Université de Toulouse

JEAN-PAUL CALVI

Première mise en ligne, à la version 0.5.0, le 2 décembre 2008, sur la page
http ://www.math.univ-toulouse.fr/˜calvi

0.5.1	18 mai 2009	(corrections mineures)
0.5.2	28 mai 2009	(corrections mineures)
0.5.3	30 mai 2009	(corrections mineures)
0.6.0	12 juin 2009	(révision des trois premiers chapitres et insertion des corrections de certains exercices)
0.6.1	29 juillet 2009	(Corrections et compléments au premier chapitre)

Développements

0.7	Compléments au chapitre 3
0.8	Révision générale, mise à jour de la bibliographie et de l'index
1	insertion des codes (scilab) / prévu courant 2009-2010

At my hut
All that I have to offer you,
Is that the mosquitoes are small.

Bashô

PRÉFACE

Ce cours est une introduction aux méthodes fondamentales de l'analyse numérique. Il devrait être accessible à tout étudiant ayant suivi une première année d'études supérieures scientifiques. Il s'adressait à l'origine à des étudiants de deuxième année de licence d'informatique de l'université Paul Sabatier et son contenu a été plusieurs fois remanié pour qu'il s'adapte raisonnablement aux connaissances de ces derniers.

J'ai inclus d'assez nombreux exercices, souvent élémentaires, y compris dans le cours de texte, dans l'espoir d'aider à la compréhension des non-spécialistes. Ces exercices reprennent en particulier les sujets d'examen que j'ai proposés à mes étudiants. Si le survol du cours et compréhension générale des méthodes veut être accessible à un large public, la lecture de l'ensemble des démonstrations et le traitement de quelques développements proposés en exercice s'adressera plutôt à des étudiants avec une plus solide formation mathématique.

Les bases sur lesquelles je m'appuie sont modestes. Une connaissance raisonnable de l'analyse des fonctions d'une variable réelle, disons, du théorème des valeurs intermédiaires jusqu'à la formule de Taylor (qui sera rappelée) et je suppose aussi une certaine familiarité avec les bases de l'algèbre linéaire (systèmes linéaires, applications linéaires, matrices et déterminants).

Ces connaissances permettent un traitement assez substantiel de l'interpolation polynomiale, du calcul approché des intégrales et de l'approximation des racines des équations, trois thèmes qui forment souvent l'essentiel d'une introduction à l'analyse numérique. Par contre, dans le quatrième chapitre, consacré aux méthodes de résolution directe des systèmes linéaires, j'ai dû me limiter à une esquisse, essentiellement à la méthode de Gauss, faute de quoi je serais resté hors de portée de mes étudiants. Ici,

ce sont les exercices qui donneront aux lecteurs intéressés une approche plus réaliste du sujet.

La question de la complexité et de la stabilité des procédés numériques (disons, leur sensibilité aux erreurs d'arrondis) est introduite de manière concrète et informelle et abordée chaque fois que c'est possible. Je crois qu'il n'y a pas de plus mauvaise manière de commencer un cours d'analyse numérique que par un chapitre sur l'étude des erreurs.

J'espère que ce texte ne constituera que la première partie d'un cours plus ambitieux. Les premiers développements projetés, qui devraient conduire à la version 1.0 de ce texte sont indiqués sur la page des mentions légales.

Toulouse, Novembre 2008
Jean-Paul Calvi *

Renvois. Lorsque le texte renvoie à un objet (théorème, section, exercice, etc) du même chapitre, seul le numéro de l'objet est indiqué. Par contre si le texte renvoie à un objet d'un autre chapitre, le numéro du chapitre apparaît aussi. Ainsi, si au cours chapitre 2, on renvoie au théorème 20 du chapitre 1, on écrira théorème I.20. Pour utiliser les liens, il suffit de sélectionner le second, ici 20.

∗. Université de Toulouse, UPS, Institut de Mathématiques de Toulouse (CNRS UMR 5219), F-31062 Toulouse, France. Courriel : jpcmath@netscape.net

TABLE DES MATIÈRES

INTERPOLATION DE LAGRANGE

§ 1 INTRODUCTION À L'INTERPOLATION POLYNOMIALE

1.1 Espaces de polynômes

Nous rappelons quelques résultats sur les polynômes (ou fonctions polynomiales). Un **monôme** de degré k est une fonction de la forme $x \in \mathbb{R} \to cx^k$ où $c \in \mathbb{R}^\star$ et $k \in \mathbb{N}$. Un **polynôme** est une somme (finie) de monômes. La fonction nulle est aussi considérée comme un polynôme. L'ensemble $\mathscr{P}$ des polynômes forme alors un espace vectoriel quand on utilise l'addition habituelle des fonctions $(p+q)$ ainsi que la multiplication par une constante (λp). Le produit de deux polynômes (pq) est encore un polynôme. Les fonctions polynômes sont indéfiniment dérivables. Tout polynôme p *non nul* s'écrit d'une manière et d'une seule sous la forme

$$p(x) = c_0 + c_1 x + \cdots + c_m x^m = \sum_{i=0}^{m} c_i x^i \tag{1.1}$$

avec $c_m \neq 0$. L'unicité provient de ce que $c_k = p^{(k)}(0)/k!$. Les nombres c_i s'appellent les **coefficients** de p. L'entier non nul m dans (1.1) est le **degré** de p et le coefficient c_m est le **coefficient dominant** de p. On convient que $\deg 0 = -\infty$. Avec cette convention,

quels que soient les polynômes p et q, nous avons

$$\deg(pq) = \deg p + \deg q \tag{1.2}$$
$$\deg(p+q) \leq \max(\deg p, \deg q). \tag{1.3}$$

E 1 Ecrire une formule donnant les coefficients d'un produit de polynômes pq en fonction des coefficients des facteurs p et q.

Lorsque $\lambda \in \mathbb{R}^\star$,

$$\deg \lambda p = \deg p, \tag{1.4}$$

c'est un cas particulier de (1.2). En réalité le degré de $p+q$ coïncide toujours avec $\max(\deg p, \deg q)$ sauf lorsque les deux polynômes ont même degré et leurs coefficients dominants sont opposés l'un de l'autre. Nous noterons $\mathscr{P}_m$ l'ensemble des polynômes de degré inférieur ou égal à m. Les propriétés (1.3) et (1.4) montrent que $\mathscr{P}_m$ est un sous-espace vectoriel de $\mathscr{P}$ dont la base canonique est $\mathscr{B} = (x \to 1 = x^0, x \to x^1, \ldots, x \to x^m)$. En particulier, sa dimension est $m+1$.

Si r est une racine de p (c'est-à-dire $p(r) = 0$) alors p est divisible par $(\cdot - r)$. Cela signifie qu'il existe un polynôme q tel que $p(x) = (x-r)q(x)$ pour tout $x \in \mathbb{R}$. Nous disons que r est une racine de **multiplicité** m lorsque $(\cdot - r)^m$ divise p mais $(\cdot - r)^{m+1}$ ne divise pas p. On montre en algèbre que cela est équivalent à

$$0 = p(r) = p'(r) = \cdots = p^{(m-1)}(r) \quad \text{et} \quad p^{(m)}(r) \neq 0.$$

Un polynôme $p \in \mathscr{P}_m$ *non nul* admet au plus m racines en tenant compte de la multiplicité. Cela signifie que si r_i est racine de multiplicité m_i de $p \neq 0$ pour $i = 1, \ldots, l$ alors $m_1 + \cdots + m_l \leq m$. On dit alors que le nombre de racine de p est en tenant compte de la multiplicité plus petite ou égale au degré du polynôme p [*]. Nous utiliserons plusieurs fois que si p est un polynôme de degré au plus m qui admet au moins $m+1$ racines en tenant compte de la multiplicité alors p est nécessairement le polynôme nul ; autrement dit,

$$\left.\begin{array}{l} z_i \text{ racine de } p \text{ de multiplicité} \geq m_i,\ i = 1, \ldots, l, \\ \sum_{i=1}^{l} m_i > m, \\ p \in \mathscr{P}_m \end{array}\right\} \Longrightarrow p = 0. \tag{1.5}$$

E 2 Peut-on retrouver un polynôme de degré m quand on sait que $w_1, \ldots, x_m$ sont ses racines ?

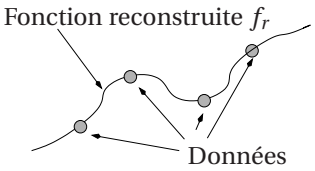
∗. Dans le cas complexe, c'est-à-dire, lorsque'on accepte de considérer les racines complexes (et mêmes les polynômes à coefficients complexes), le théorème fondamental de l'algèbre dit que le nombre de racines d'un polynôme non nul est, en tenant compte de la multiplicité, exactement égal au degré du polynôme.

E 3 En combien de points une droite peut-elle couper le graphe d'un polynôme ?

E 4 Combien d'axe de symétrie le graphe d'un polynôme peut-il admettre ? ($y = a$ est un axe de symétrie du graphe de p si $p(a-x) = p(a+x)$ pour tout $x \in \mathbb{R}$.)

1.2 Le problème général de l'interpolation polynomiale

En analyse numérique, une fonction f n'est souvent connue que par ses valeurs f_i en un nombre fini de points a_i, $f_i = f(a_i)$, (en réalité, en pratique f_i est seulement une approximation de $f(a_i)$). Cependant, dans la plupart des cas, il est nécessaire d'effectuer des opérations sur des fonctions *globales* (dérivation, intégration, ...) et nous sommes donc conduit à *reconstruire* une fonction globale f_r à partir d'un nombre fini de données (a_i, f_i).



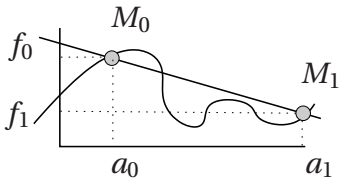
Sauf cas très simple, la fonction f_r ne coïncidera pas avec la fonction "idéale" f mais il faut faire en sorte qu'elle n'en soit pas trop éloignée.

Le problème de l'interpolation polynomiale consiste à choisir comme fonction reconstruite une fonction polynomiale. C'est la méthode la plus ancienne, la plus élémentaire et encore la plus utile. Mais il y en a d'autres. Nous verrons plus loin, au § 4, une seconde méthode employant les polylignes. Dans la figure ci-dessus la fonction reconstruite f_r est obtenue à partir de quatre données par un procédé voisin (spline d'interpolation) mais différent.

D'une manière précise, étant donnés $d+1$ points d'abscisses distinctes $M_j = (a_j, f_j)$, $j = 0, \dots, d$, dans le plan (pour des raisons de commodité d'écriture les points seront toujours indicés à partir de 0), le problème consiste à trouver un polynôme $p \in \mathscr{P}_m$ dont le graphe passe par les $d+1$ points M_j. En formule, nous devons avoir

$$p \in \mathscr{P}_m \quad \text{et} \quad p(a_j) = f_j \qquad j = 0, \dots, d. \tag{1.6}$$

Ce problème est bien facile à résoudre lorsque nous disposons de deux points M_0 et M_1 et cherchons un polynôme de degré 1 car il suffit alors de choisir l'unique polynôme dont le graphe est la droite (M_0M_1) comme indiqué sur la figure.



En effet, posant $p(x) = \alpha x + \beta$, nous déterminons α et β grâce aux équations $p(a_0) = f_0$ et $p(a_1) = f_1$. Il vient

$$p(x) = \frac{f_1 - f_0}{a_1 - a_0}(x - a_0) + f_0 \tag{1.7}$$

que nous pouvons aussi écrire

$$p(x) = f_0 \frac{x - a_1}{a_0 - a_1} + f_1 \frac{x - a_0}{a_1 - a_0}. \tag{1.8}$$

Il est à peine plus compliqué lorsque nous dispsons de trois points $M_i(a_i, f_i)$, $i = 0,1,2$ avec $a_0 < a_1 < a_2$ et cherchons un polynôme du second degré. Le graphe cherché est en général une parabole (correspondant à un polynôme de degré 2). Cependant dans le cas particulier où les trois points sont alignés le graphe est à nouveau une droite (correspondant à un polynôme de degré 1).

Ceci dit, s'il n'est pas davantage précisé, le problème (1.6) peut n'avoir aucune solution ou bien en avoir une infinité.

E 5 (a) Montrer qu'il existe une infinité de polynômes $p \in \mathscr{P}_2$ dont le graphe passe par les points $M_0(0,0)$ et $M_1(1,1)$. (b) Trouver quatre points M_i ($i = 1,2,3,4$) d'abscisses respectives $-1,0,1,2$ qui ne se trouvent sur le graphe d'aucun polynôme de $\mathscr{P}_2$.

1.3 Détermination du polynôme d'interpolation

Nous devinons aisément que pour qu'un seul polynôme satisfasse aux conditions (1.6), une relation doit exister entre le nombre de points $d+1$ et le degré m du polynôme cherché. Cette relation est facile à mettre en évidence. Pour déterminer $p \in \mathscr{P}_m$, nous devons connaître l'ensemble de ses coefficients et ceux-ci sont au nombre de $m+1$. Or, pour les obtenir, nous disposons des $d+1$ informations $p(a_i) = f_i$, $i = 0,\dots,d$. De manière précise, posant $p(x) = \sum_{i=0}^{m} c_i x^i$, nous devons calculer les $m+1$ coefficients c_i à l'aide des $d+1$ équations

$$\sum_{i=0}^{m} c_i \, {a_j}^i = f_j, \quad 0 \le j \le d. \tag{1.9}$$

Le cours d'algèbre linéaire nous dit alors que pour espérer une solution unique, il nous faut supposer que $m = d$ — ce que nous ferons à partir de maintenant — et, dans ce cas, le système admettra une solution et une seule si et seulement si son déterminant sera différent de 0. Nous pourrons alors obtenir une expression plus ou moins explicite pour chaque c_i en utilisant les formules de Cramer (voir IV.1.3). S'il n'est pas trop difficile, le calcul du déterminant de ce système est cependant assez long (il sera proposé en exercice) et nous suivrons ici une autre démarche, assez courante en mathématiques. Elle consiste à décomposer le problème en un grand nombre de micro-problèmes puis de superposer les solutions de ces micro problèmes pour obtenir une solution du problème de départ. L'idée est la suivante. Nous supposons dans un premier temps que nous connaissons pour chaque $i \in \{0,\dots,d\}$ un polynôme $\ell_i \in \mathscr{P}_d$ qui satisfasse $\ell_i(a_i) = 1$ et $\ell_i(a_j) = 0$ pour $j \neq i$. Il est commode de présenter cette propriété en utilisant le symbole de Kronecker δ_{ij} qui vaut 1 lorsque $i = j$ et 0 lorsque $i \neq j$. Ainsi, nos polynômes ℓ_i vérifient $\ell_i(a_j) = \delta_{ij}$. Nous formons ensuite le polynôme $p := \sum_{i=0}^{d} f_i \ell_i$. Puisque chaque $\ell_i \in \mathscr{P}_d$ et que $\mathscr{P}_d$ est un espace vectoriel nous avons

$p \in \mathscr{P}_d$. De plus $p(a_j) = \sum_{i=1}^{d} f_i \ell_i(a_j) = \sum_{i=1}^{d} f_i \delta_{ij} = f_j$ de sorte que le polynôme p satisfait les conditions demandées. Le problème sera donc résolu si nous établissons l'existence des polynômes ℓ_i. Cherchons donc à déterminer ℓ_i en exploitant les conditions que nous lui avons imposées. Puisque $\ell_i(a_j) = 0$ pour $j \neq i$, ℓ_i est factorisable par $(x - a_j)$ pour $j \neq i$ et comme les a_j sont supposés deux à deux distincts, il vient

$$\ell_i(x) = (x - a_0) \cdots (x - a_{i-1})(x - a_{i+1}) \cdots (x - a_d) R(x), \tag{1.10}$$

où R est un polynôme qu'il nous reste à déterminer. Puisqu'il y a dans (1.10) $d+1-1 = d$ facteurs $(x - x_j)$ qui donnent un polynôme de degré d et que ℓ_i lui-même appartient à $\mathscr{P}_d$, le polynôme R est nécessairement constant de sorte que pour un certain $K \in \mathbb{R}$,

$$\ell_i(x) = K(x - a_0)(x - a_1) \cdots (x - a_{i-1})(x - a_{i+1}) \cdots (x - a_d). \tag{1.11}$$

Mais il est aussi demandé que $\ell_i(a_i)$ soit égal à 1 et cette condition permet immédiatement d'obtenir la constante K,

$$K = \{(a_i - a_0) \cdots (a_i - a_{i-1})(a_i - a_{i+1}) \cdots (a_i - a_d)\}^{-1}. \tag{1.12}$$

Nous avons donc établi l'existence des polynômes ℓ_i et presque entièrement démontré le théorème suivant.

Théorème 1. *Soit $A = \{a_0, \ldots, a_d\}$ un ensemble de $d+1$ nombres réels (deux à deux) distincts. Quelles que soient les valeurs $f_0, f_1, \ldots, f_d$, il existe un et un seul polynôme $p \in \mathscr{P}_d$ tel que $p(a_i) = f_i$, $i = 0, 1, \ldots, d$. Ce polynôme, est donné par la formule*

$$p = \sum_{i=0}^{d} f_i \ell_i, \tag{1.13}$$

avec

$$\ell_i(x) = \frac{(x - a_0) \cdots (x - a_{i-1})(x - a_{i+1}) \cdots (x - a_d)}{(a_i - a_0) \cdots (a_i - a_{i-1})(a_i - a_{i+1}) \cdots (a_i - a_d)}. \tag{1.14}$$

Démonstration. La seule affirmation que nous n'avons a pas encore établie est l'unicité. Nous avons trouvé un polynôme p satisfaisant les conditions demandées mais nous n'avons pas montré qu'il n'y a pas d'autre solution que celle que nous avons trouvée. Supposons que q_1 et q_2 soient deux solutions et posons $q = p_1 - p_2$. En utilisant à nouveau le fait que $\mathscr{P}_d$ est un espace vectoriel, nous avons $q \in \mathscr{P}_d$. De plus, pour $i = 0, \ldots, d$, $q(a_i) = f_i - f_i = 0$. Nous avons donc un polynôme q de degré au plus d qui admet au moins $d+1$ racines. En vertu de la relation (1.5) sur les racines d'un polynôme, la seule possibilité est $q = 0$ qui entraîne $p_1 = p_2$ et l'unicité s'ensuit. ∎

1.4 Terminologie et notations

Les nombres a_i s'appellent les **points d'interpolation** ou encore **noeuds d'interpolation**. Lorsque $f_i = f(a_i)$, la fonction f est la **fonction interpolée**. Nous disons aussi que les valeurs $f(a_i)$ sont les **valeurs d'interpolation** ou **valeurs interpolées** L'unique polynôme $p \in \mathscr{P}_d$ vérifiant $p(a_i) = f(a_i)$ $(i = 0, 1, \ldots, d)$ s'appelle alors le polynôme d'**interpolation de Lagrange** de f aux points a_i. Il est noté $\mathbf{L}[a_0, \ldots, a_d; f]$ ou bien $\mathbf{L}[A; f]$.

Cette dernière notation est cohérente car le polynôme d'interpolation de Lagrange dépend uniquement de l'ensemble des points $A = \{a_0, \ldots, a_d\}$ et non du $d+1$-uplet $(a_0, \ldots, a_{d+1})$. Autrement dit, le polynôme d'interpolation de Lagrange ne dépend pas de la manière dont les points sont ordonnés. Une autre manière un peu sophistiquée de traduire cette propriété est la suivante : si σ est une permutation * quelconque des indices $0, 1, \ldots, d$ alors

$$\mathbf{L}[a_0, \ldots, a_d; f] = \mathbf{L}[a_{\sigma(0)}, \ldots, a_{\sigma(d)}; f].$$

Les polynômes ℓ_i s'appellent les **polynômes fondamentaux de Lagrange** . En utilisant le symbole $\prod$ qui est l'équivalent pour le produit de ce que $\sum$ est pour la somme, nous obtenons la formule suivante qui est une variante compacte de (1.14).

$$\ell_i(x) = \prod_{j=0, j\neq i}^{d} \frac{x - a_j}{a_i - a_j}. \tag{1.15}$$

Avec ces nouvelles notations, l'expression (1.13) devient

$$\mathbf{L}[a_0, \ldots, a_d; f](x) = \sum_{i=0}^{d} f(a_i) \prod_{j=0, j\neq i}^{d} \frac{x - a_j}{a_i - a_j}. \tag{1.16}$$

Cette expression de $\mathbf{L}[A; f]$ est connue sous le nom de **formule d'interpolation de Lagrange**.

1.5 Propriétés algébriques et linéarité

Il est essentiel de retenir l'équivalence suivante

$$\left.\begin{array}{l} p \in \mathscr{P}_d \\ p(a_i) = f(a_i) \quad i = 0, \ldots, d \end{array}\right\} \Leftrightarrow p = \mathbf{L}[a_0, \ldots, a_d; f]. \tag{1.17}$$

En particulier,

$$\text{si } p \in \mathscr{P}_d \text{ alors } \mathbf{L}[a_0, \ldots, a_d; p] = p. \tag{1.18}$$

*. Une permutation des indices $0, 1, \ldots, d$ est une bijection de l'ensemble$\{0, 1, \ldots, d\}$ dans lui-même.

Il faut prendre garde que cette propriété n'est valable que lorsque le degré de p est inférieur où égal à d *. La relation (1.17) signifie que pour établir qu'un polynôme donné p est égal au polynôme d'interpolation de Lagrange d'une fonction f aux points $a_0, \dots, a_d$, il suffit de vérifier que le degré de q est inférieur ou égal à d et que $q(a_i) = f(a_i)$ pour $i = 0, \dots, d$.

La relation (1.18) entraîne des propriétés algébriques intéressantes sur les polynômes ℓ_i. Par exemple, en utilisant que, quel que soit le nombre de points, le polynôme constant égal à 1 est son propre polynôme d'interpolation, nous obtenons

$$\sum_{i=0}^{d} \ell_i = 1. \tag{1.19}$$

E 6 Vérifiez la propriété ci-dessus par le calcul dans le cas où $d = 1$ (deux points d'interpolation) et $d = 2$ (trois points d'interpolation).

Théorème 2. *L'application qui à f définie (au moins) sur $A = \{a_0, \dots, a_d\}$ fait correspondre $\mathbf{L}[A; f] \in \mathscr{P}$,*

$$\mathbf{L}[A; \cdot] : f \in \mathscr{F}(A) \to \mathbf{L}[A; f] \in \mathscr{P}_d,$$

est une application linéaire de l'espace vectoriel $\mathscr{F}(A)$ des fonctions réelles définies sur A à valeurs dans l'espace vectoriel des polynômes de degré au plus d. Cela signifie qu'elle satisfait les deux propriétés suivantes

$$\begin{cases} \mathbf{L}[A; f+g] &= \mathbf{L}[A; f] + \mathbf{L}[A; g], \quad f, g \in \mathscr{F}(A) \\ \mathbf{L}[A; \lambda f] &= \lambda \mathbf{L}[A; f], \quad f \in \mathscr{F}(A),\ \lambda \in \mathbb{R} \end{cases}. \tag{1.20}$$

E 7 Montrer les propriétés (1.20).

E 8 Soit pour tout $n \in \mathbb{N}$, $M_n(x) = x^n$. Déterminer $\mathbf{L}[-1, 0, 1; M_n]$ et en déduire, pour tout polynôme p, une formule pour $\mathbf{L}[-1, 0, 1; p]$ en fonction des coefficients de p.

§ 2 ALGORITHME DE CALCUL ET EXEMPLES GRAPHIQUES

2.1 Algorithme basé sur la formule d'interpolation de Lagrange

L'algorithme suivant est une traduction directe de la formule d'interpolation de Lagrange (1.16). S'il est le plus simple, il n'est pas, loin s'en faut, le meilleur et il nous servira surtout à mettre en évidence les problèmes numériques liées à l'utilisation d'un algorithme. Un meilleur algorithme (de Neville-Aitken) est donné plus loin et une troisième méthode est esquissée dans l'exercice 30.

*. Pour la calcul de $\mathbf{L}[a_0, \dots, a_d; p]$ lorsque le degré de p est strictement supérieur à d, voir l'exercice 25.

Algorithme 3. *Les données de l'algorithme sont (a) le vecteur $a = (a_0, \dots, a_d)$ formé des points d'interpolation, (b) le vecteur $f = (f_0, \dots, f_d)$ formé des valeurs d'interpolations (c) le point t en lequel nous voulons calculer* $\mathbf{L}[a; f]$. *Le résultat est dans P.*

(a) $P := 0$

(b) Pour $i \in [0 : d]$ *faire*

(a) $L := 1$

(b) Pour $j \in [0 : i-1; i+1 : d]$, $L := L \times (t - a_j)/(a_i - a_j)$

(c) $P := P + L \times f_i$.

2.2 Exemples

Sur les graphiques de la table 1, nous pouvons comparer la fonction $f(x) = x\sin(\pi x)$ (tracée en bleu) et ses polynômes d'interpolation (tracés en rouge) de degré d par rapport aux $d+1$ **points équidistants** $a_i = -1 + 2i/d$, $i = 0, 1, \dots, d$ lorsque $d = 3$, 4, 5 et 6. Par exemple lorsque $d = 4$, les 5 noeuds d'interpolation sont -1, -0.6, -0.2, 0.2, 0.6, 1. Remarquons que les polynômes approchent si bien la fonction que les graphes sont confondus sur $[-1, 1]$ dès que $d = 6$. Par contre, le résultat est mauvais en dehors de l'intervalle $[-1, 1]$. En réalité, avec la fonction choisie, qui est très régulière * en augmentant d, nous obtiendrions aussi une excellente approximation en dehors de l'intervalle. Nous verrons plus loin des exemples de fonctions pour lesquelles les polynômes d'interpolation construits aux points équidistants ne fournissent pas une bonne approximation.

E 9 Remarquons que, dans le cas $d = 3$, le graphe du polynôme d'interpolation est une parabole, c'est à dire le graphe d'un polynôme du second degré (et non pas d'un polynôme de degré 3). Expliquer cela.

2.3 Le coût de l'algorithme

Le **coût** ou encore la **complexité** d'un algorithme est le nombre d'opérations élémentaires $(+, -, \times, \div)$ employé par cet l'algorithme pour produire son résultat. Parfois, comme dans le théorème ci-dessous, on se limite à compter le nombre de multiplications et de divisions en supposant que le travail demandé par les additions et les soustractions est négligeable devant celui demandé par une multiplication ou une division.

∗. Le mot *régulier* est un mot passe-partout en mathématiques qui n'a de sens que dans le contexte. Ici, le sens serait celui de *fonction analytique* mais cette notion est trop délicate pour être introduite dans ce cours. Nous disons plus simplement que la fonction $f(x) = x\sin(\pi x)$ est très régulière parce qu'elle est indéfiniment dérivable et les valeurs absolues de ses dérivées $|f^{(s)}(x)|$ ne croissent pas trop vite lorsque $x \in [-1, 1]$ et s devient grand. Cette idée est développée dans l'exercice 36.

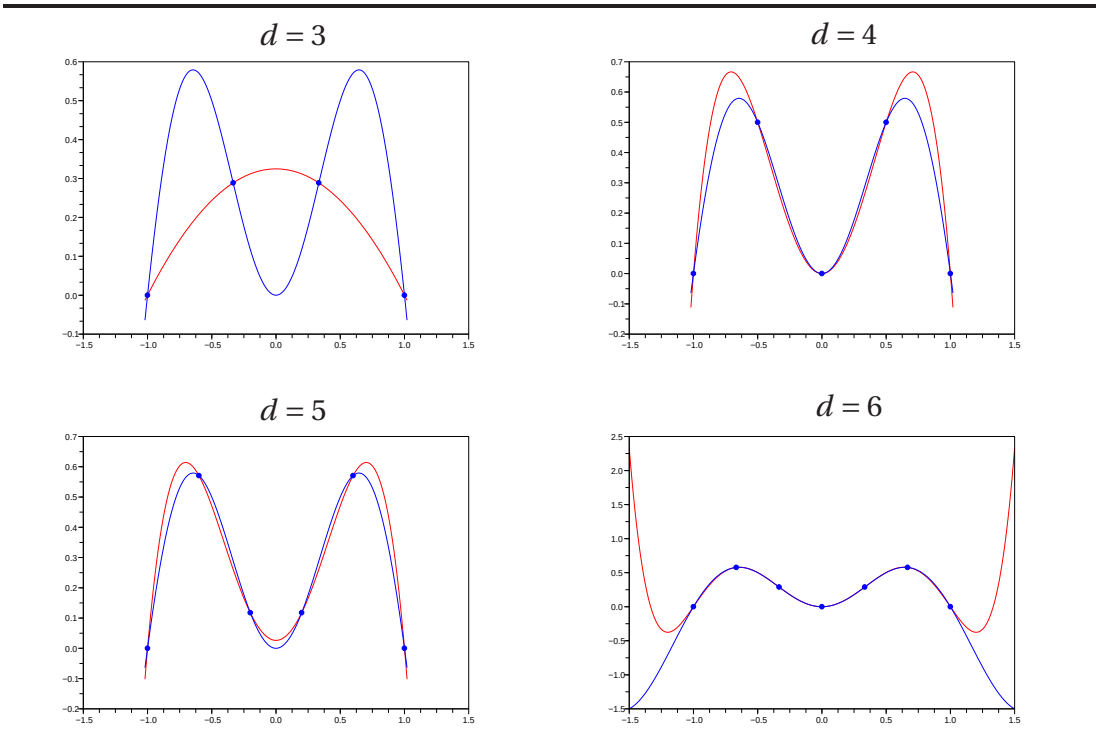


TABLE 1 – Quelques polynômes d'interpolation de la fonction $f(x) = x \sin \pi x$.

Ces nombres d'opérations ne donnent qu'une information partielle sur la rapidité (et l'utilité pratique, la faisabilité) de l'algorithme. D'autres éléments entrent en ligne de compte. Le nombre de mémoires (de registres, de variables) occupées par l'algorithme est un autre élément important. Dans le calcul de la complexité, il n'est pas tenu compte des actions de changement d'affectation de nombres dans des variables, non plus que des tris sur ces des listes de nombres comme par exemple la recherche du nombre le plus grand. Ces actions consomment une énergie et un temps importants qui peuvent suffire à dissuader de l'emploi d'un algorithme. Cette notion de coût/complexité joue un rôle fondamental en analyse numérique matricielle dans laquelle, sont presque uniquement utilisées les opérations élémentaires. Dans les autres parties de l'analyse numérique, il est souvent nécessaire d'utiliser les opérations élémentaires sur des valeurs de fonctions usuelles (elles-mêmes conservées en mémoire à disposition ou évaluées avec un nombre fini d'opérations élémentaires) et il est alors souvent plus naturel de compter les évaluations de ces fonctions usuelles parmi les opérations élémentaires.

Le théorème suivant donne une première idée d'un calcul de complexité.

Théorème 4. *Le calcul de* $\mathbf{L}[a_0,\dots,a_d;f](x)$ *par l'algorithme 3 nécessite* $(2d+1)(d+1)\approx 2d^2$ *multiplications-divisions.*

Ici, comme il est naturel, la complexité est une fonction croissante de d, c'est-à-dire essentiellement du nombre de points d'interpolation, ou encore du nombre de données à utiliser.

Le symbole $\approx$ est mis pour indiquer l'**équivalence** de deux suites. Nous disons que deux suites u_d et v_d sont équivalentes (lorsque $d\to\infty$) lorsqu'il existe une troisième suite ϵ_d qui tend vers 0 et telle que $u_d = v_d(1+\epsilon_d)$. Lorsque l'une des deux suites ne s'annule plus à partir d'un certain rang, disons v_d, nous pouvons diviser par v_d lorsque d est assez grand, et la définition se traduit alors par $\lim_{d\to\infty} u_d/v_d = 1$. Dans l'énoncé du théorème, prenant $u_d = (2d+1)(d+1) = 2d^2+3d+1$ et $v_d = 2d^2$, nous avons bien $u_d/v_d = 1+3/(2d)+2/(2d^2)\to 0$ lorsque $d\to\infty$.

Démonstration. Le nombre N_d de multiplications-divisions est donné par

$$\sum_{i=0}^{d}\left(\Big\{\sum_{j=0,j\neq i}^{d} 2\Big\}+1\right) = \sum_{i=0}^{d}(2d+1) = (2d+1)(d+1). \tag{2.1}$$

Le premier $\sum_{i=0}^{d}$ provient de la boucle *pour* $i\in[0,d]$ dans l'algorithme 3 tandis que la seconde somme provient de la boucle *pour* $j\in[0:i-1;i+1:d]$. L'imbrication des signes $\sum$ traduit le fait que la seconde boucle s'effectue à l'intérieur de la première. ■

2.4 La stabilité de l'algorithme

Prenons 7 points équidistants dans $[-1,1]$, $a_i = -1+2/6\cdot i$, $i=0\dots,6$ et calculons l'interpolant de Lagrange de la fonction $M_3 : x\to x^3$. La relation (1.18) montre que

$$\mathbf{L}[a_0,\dots,a_6;M_3](x) = x^3.$$

L'algorithme 3, correctement modifié pour produire un polynôme, fournit le résultat donné dans la table 2. Les résultats qui ne sont pas nuls mais qui devraient l'être sont tellement petit que nous pouvons sans hésiter les éliminer * de sorte que les résultats sont acceptables. Les difficultés apparaissent pour un nombre de points supérieur. Pour $d=30$ et la fonction polynôme $p(x) = 6x^2+2x^3+x^4+x^5$, nous obtenons les coefficients donnés dans la table 3. Il n'est pas difficile d'expliquer l'inexactitude du résultat. Un calculateur ne travaille qu'avec une famille finie (très étendue) de nombres F et le résultat

*. Les logiciels de calculs sont munis d'opérateur de "nettoyage" qui remplacent par 0 les données extrêmement petites.

n	coef. de x^n	n	coef. de x^n	n	coef. de x^n
1	$-2.776D-17$	3	1	5	$6.661D-16$
2	$-2.776D-16$	4	$5.551D-16$	6	$-1.110D-15$

TABLE 2 – Coefficients de $\mathbf{L}[_0,\ldots,a_6;M_3](x)=x^3$ calculés par l'algorithme 3.

n	coef. de x^n	n	coef. de x^n	n	coef. de x^n
1	$-7.154D-16$	11	0.0000102	21	0.0312387
2	6	12	−0.0000084	22	−0.0128725
3	2	13	0.0000207	23	−0.0083289
4	1	14	−0.0003492	24	0.0210372
5	1	15	0.0021007	25	−0.0186058
6	$2.804D-09$	16	−0.0073588	26	0.0090390
7	$3.535D-09$	17	0.0175530	27	−0.0024166
8	0.0000001	18	−0.0304909	28	0.0003477
9	0.0000011	19	0.0399997	29	0.0000052
10	−0.0000046	20	−0.0407562	30	−0.0000115

TABLE 3 – Coefficients de l'interpolant de Lagrange $\mathbf{L}[a_0,\ldots,a_29;p](x)=p$, $p(x)=6x^2+2x^3+x^4+x^5$, calculés par l'algorithme 3 avec 30 points équidistants dans $[-1,1)$.

de toutes les opérations qu'il effectue doit être sélectionné parmi ces nombres. Si a et b sont deux nombres réels et $*$ désigne une opération quelconques alors le résultat de l'opération $a*b$ sera $\mathtt{F}(a*b)$ avec

$$F(a*b)=\boxed{\boxed{a}*\boxed{b}},$$

où $\boxed{x}$ désigne l'élément de F le plus proche, en un certain sens, de x. Traditionnellement, même si cela ne correspond plus aux fonctionnement des calculateurs modernes, la différence entre le résultat exact $a*b$ et le nombre retenu par le calculateur $\mathtt{F}(a*b)$ est appelée **erreur d'arrondi**. De manière informelle,nous disons qu'un algorithme est stable lorsque les erreurs au niveaux des données et les erreurs d'arrondis n'induisent des erreurs au niveau du résultat comparables à celles sur les données. Pour être correctement analysée, cette idée doit être formalisée : nous devons estimer la différence entre le résultat théorique et le résultat fourni par le calculateur en tenant compte des erreurs sur les données et des propriétés techniques des calculateurs. Cette analyse, en général est délicate et, dans ce cours, nous n'aurons l'occasion d'en consi-

dérer que quelques exemples très simples qui ne concerneront que la propagation des erreurs sur les données.

Dans le cas du calcul de l'interpolant de Lagrange qui nous intéresse ici, l'algorithme ne calcule pas $\mathbf{L}[_0,\ldots,a_{29};p]$ mais une approximation $\tilde{\mathbf{L}}$. Si nous examinons le coefficient de x^{17} dans la table 3, qui théoriquement devrait être nul, nous trouvons une erreur de l'ordre d'un centième qui est une erreur extrêmement grande - si nous savons que tous les calcul sont effectués avec une précision de l'ordre de 10^{-12}. Nous dirons que l'algorithme est instable. En général, la stabilité dépend : (a) des points d'interpolation $a_0\ldots,a_d$ – de ce point de vue les points équidistants constituent un mauvais choix ; (b) de la fonction interpolée ; en particulier les risques d'erreur augmentent si la fonction admet des variations importantes, c'est-à-dire lorsque $f(x+\epsilon)$ peut être très différent de $f(x)$ pour ϵ petit – c'est le cas des fonctions avec $f'(x)$ grand ; (c) de l'algorithme utilisé, dont les qualités dépendent de la méthode mathématique dont il découle, du programme ou langage à l'intérieur duquel l'algorithme est programmé, de l'habilité du traducteur. Les problèmes de la complexité et de la stabilité ne sont pas indépendants puisqu'en général plus le nombre d'opérations sera grand plus le risque de propagation des erreurs d'arrondis sera élevé.

2.5 La formule de récurrence de Neville-Aitken

Théorème 5 (Neville-Aitken)**.** *Soit $A=\{a_0,a_1,\ldots,a_d\}$ un ensemble de $d+1$ nombres réels distincts et f une fonction définie (au moins) sur A. Alors*

$$\begin{aligned}(a_0-a_d)&\mathbf{L}[a_0,a_1,\ldots,a_d;f](x)\\ &=(x-a_d)\mathbf{L}[a_0,a_1,\ldots,a_{d-1};f](x)-(x-a_0)\mathbf{L}[a_1,a_2,\ldots,a_d;f](x).\end{aligned}\tag{2.2}$$

Démonstration. Considérons $Q\in\mathscr{P}_d$ défini par

$$Q(x)=\frac{1}{a_0-a_d}\{\underbrace{\mathbf{L}[a_0,\ldots,a_{d-1};f](x)}_{\in\mathscr{P}_{d-1}}\underbrace{(x-a_d)}_{\in\mathscr{P}_1}-\underbrace{\mathbf{L}[a_1,\ldots,a_d;f](x)}_{\in\mathscr{P}_{d-1}}\underbrace{(x-a_0)}_{\in\mathscr{P}_1}\},\tag{2.3}$$

et calculons ses valeurs en les points a_i. Nous avons, en indiquant par un point d'interrogation une valeur inconnue mais sans influence,

$$\begin{aligned}Q(a_0)&=\tfrac{1}{a_0-a_d}\{f(a_0)(a_0-a_d)-[?]\times 0\}=f(a_0),\\ Q(a_i)&=\tfrac{1}{a_0-a_d}\{f(a_i)(a_i-a_d)-f(a_i)(a_i-a_0)\}=f(a_i)\ (1\le i\le d-1),\\ Q(a_d)&=\tfrac{1}{a_0-a_d}\{[?]\times 0-f(a_d)(a_d-a_0)\}=f(a_d).\end{aligned}$$

Maintenant, de $Q\in\mathscr{P}_d$ et $Q(a_i)=f(a_i)$ pour $i=0,\ldots,d$, nous déduisons que Q n'est autre que $\mathbf{L}[a_0,\ldots,a_d;f](x)$ et c'est ce qu'il fallait établir. ∎

Corollaire. *Sous les mêmes hypothèses, pour tout couple d'indice* (i, j) *dans* $\{0,\dots,d\}$ *avec* $i \neq j$,

$$\begin{aligned}(a_i - a_j)\mathbf{L}[a_0, a_1, \dots, a_d; f](x) \\ = (x - a_j)\mathbf{L}[a_0, \dots, a_{j-1}, a_{j+1}, \dots a_d; f](x) \\ - (x - a_i)\mathbf{L}[a_0, \dots, a_{i-1}, a_{i+1}, \dots a_d; f](x). \quad (2.4)\end{aligned}$$

2.6 L'algorithme de Neville-Aitken

Posons $A = \{x_1, x_2, \dots, x_{d+1}\}$ un ensemble de $d+1$ réels distincts. Il faut prendre garde que les points sont ici indicés à partir de 1 et non pas, comme jusqu'à présent, à partir de 0. Nous définissons une famille de polynômes $p_{i,m}$ par récurrence sur $m \in \{0, 1, \dots, d\}$ comme suit

$$p_{i,0}(x) = f(x_i), \quad 1 \leq i \leq d+1; \tag{2.5}$$

puis,

$$p_{i,m+1}(x) = \frac{(x_i - x)p_{m+1,m}(x) - (x_{m+1} - x)p_{i,m}(x)}{x_i - x_{m+1}}, \quad m+2 \leq i \leq d+1. \tag{2.6}$$

Les polynômes $p_{i,m}$ sont définis seulement pour les couples d'indices (i, m) vérifiant la condition $d+1 \geq i > m \geq 0$, nous disons que nous avons construit une famille triangulaire de polynômes.

Théorème 6. *Pour* $0 \leq m \leq d$, *nous avons*

$$p_{i,m} = \mathbf{L}[x_1, x_2, \dots, x_m, x_i; f], \quad m+1 \leq i \leq d+1. \tag{2.7}$$

Lorsque $m = 0$ *l'écriture* $\mathbf{L}[x_1, x_2, \dots, x_m, x_i; f]$ *doit être comprise comme* $\mathbf{L}[x_i; f]$.

Remarquons que le cas $m = d$ dans la relation (2.7) donne $p_{d+1,d} = \mathbf{L}[x_1, x_2, \dots, x_{d+1}; f]$.

Démonstration. Nous démontrons (2.7) par récurrence sur m. Le résultat est vrai pour $m = 0$ à cause de la définition (2.5). Supposant que le résultat est vrai pour m, nous le montrons pour $m+1$. Appelons $Q(x)$ le terme de droite dans l'équation (2.6). L'hypothèse de récurrence nous permet d'écrire

$$p_{m+1,m} = \mathbf{L}[x_1, \dots, x_m, x_{m+1}; f] \quad \text{et} \quad p_{i,m} = \mathbf{L}[x_1, \dots, x_m, x_i; f],$$

de sorte que

$$Q(x) = \frac{(x_i - x)\mathbf{L}[x_1, \dots, x_m, x_{m+1}; f](x) - (x_{m+1} - x)\mathbf{L}[x_1, \dots, x_m, x_i; f](x)}{x_i - x_{m+1}}$$

et nous déduisons de la relation de Neville-Aitken (par la formule (2.4) du corollaire) que

$$Q(x) = \mathbf{L}[x_1, \dots, x_m, x_{m+1}, x_i; f](x)$$

qui est la formule annoncée dans le cas $m+1$. ■

Algorithme 7. *Les données de l'algorithme sont (a) le vecteur $x = (x_1, \dots, x_{d+1})$ formé des points d'interpolation, (b) le vecteur $f = (f_1, \dots, f_{d+1})$ formé des valeurs d'interpolation (c) le point t en lequel nous voulons calculer $\mathbf{L}[x; f]$. On utilise une matrice P de dimension $(d+1) \times (d+1)$ que l'on initialise à 0. Le résultat est dans $P(d+1, d+1)$.*

(a) Pour $j \in [1 : d+1]$, $P(j,1) = y(j)$.

(b) Pour $m \in [2 : d+1]$

Pour $i \in [m : d+1]$

$$p(i,m) = \frac{\big(x(i) - x\big) \times p(m-1, m-1) - \big(x(m-1) - x\big) \times p(i, m-1)}{(x(i) - x(m-1))}. \tag{2.8}$$

E 10 Déterminer le nombre de multiplications-divisions employé par l'algorithme de Neville-Aitken.

La table 4 reprend l'exemple de la table 3 précédent et compare les six plus mauvais coefficients obtenus par l'algorithme de Lagrange (Lag) aux coefficients correspondants produits par l'algorithme de Neville-Aitken (N-A) ci-dessus. Celui-ci améliore le résultat en moyenne par un facteur 10. L'algorithme reste instable (et le restera toujours s'agissant de points équidistants) mais cet exemple illustre bien le fait que l'algorithme lui-même et non seulement les données influe sur la stabilité.

	N-A	Lag		N-A	Lag		N-A	Lag
n	coef. x^n	coef. x^n	n	coef. x^n	coef. x^n	n	coef. x^n	coef. x^n
17	0.0024	0.0175	21	−0.002	0.0312	24	−0.0019	0.021
18	0.0023	−0.0305	22	0.0026	−0.0128	26	0.00226	0.00903

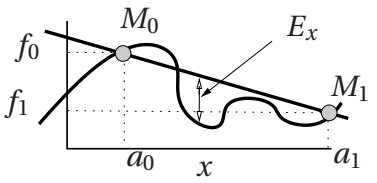
TABLE 4 –

§ 3 ETUDE DE L'ERREUR

3.1 L'énoncé du théorème

Comme le polynôme d'interpolation $\mathbf{L}[a_0, \dots, a_d; f]$ est égal à la fonction f en tous les points a_i, $i = 0, \dots, d$, il est naturel d'espérer que la différence entre f et ce poly-

nôme aux autres points sera petite c'est-à-dire, que $\mathbf{L}[a_0,\dots,a_d\,;f]$ fournira une bonne approximation de $f(x)$, au moins en les points x pas trop éloignés des a_i.



Pour mesurer la qualité de cette approximation, nous devons estimer l'erreur E_x entre $f(x)$ et $\mathbf{L}[a_0,\dots,a_d;f](x)$, c'est-à-dire trouver une majoration de la valeur absolue de E_x. La figure ci-dessous fait apparaître cette erreur dans le cas $d=1$. Cette erreur est une distance,

$$E_x = |f(x) - \mathbf{L}[a_0,\dots,a_d;f](x)|.$$

Nous devinons facilement qu'elle dépendra à la fois de la fonction f et de la position des points a_i. Le théorème suivant, et surtout son corollaire, fournissent une estimation simple de l'erreur.

Rappelons d'abord une notation. On dit qu'une fonction f définie sur un intervalle $[a,b]$ est de classe C^{d+1} sur cet intervalle et on écrit $f\in\mathrm{C}^{d+1}[a,b]$ lorsque f est $d+1$ fois dérivable et que $f^{(d+1)}$, la dérivée $(d+1)$-ième est continue. Au point a (resp. b) il s'agit de dérivées à droite (resp. à gauche).

Théorème 8. *Soient $f\in\mathrm{C}^{d+1}[a,b]$ et $A=\{a_0,a_1,\dots,a_d\}\subset[a,b]$. Nous supposons que $a_0<a_1<a_2<\dots<a_{d-1}<a_d$. Pour tout $x\in[a,b]$, il existe $\xi=\xi_x\in]a,b[$ tel que*

$$f(x)-\mathbf{L}[A;f](x)=\frac{f^{(d+1)}(\xi)}{(d+1)!}(x-a_0)(x-a_1)\dots(x-a_d). \tag{3.1}$$

Lorsque $d=0$ et $A=\{a_0\}$, nous avons $\mathbf{L}[a_0\,;f](x)=f(a_0)$ de sorte que le théorème 8 affirme que, pour un x fixé dans [a,b], il existe un point ξ – dépendant de x – tel que

$$f(x)-f(a_0)=f'(\xi)(x-a_0).$$

Il s'agit du théorème des accroissements finis dont le théorème 8 est par conséquent une généralisation.

E 11 Soit $a\le a_0<a_1\le b$. Montrer que si f est une fonction strictement convexe deux fois dérivable sur $[a,b]$ alors $f(x)-\mathbf{L}[a_0,a_1\,;f](x)<0$ pour tout $x\in]a_0,a_1[$. Que dire en dehors de l'intervalle $[a_0,a_1]$? Même question dans le cas des fonctions deux fois dérivables strictement concaves. Étudier le problème sans supposer que les fonctions soient deux fois dérivables.

Dans la pratique, le corollaire suivant est très souvent suffisant.

Corollaire.

$$|f(x)-\mathbf{L}[A;f](x)|\le\frac{1}{(d+1)!}|x-a_0||x-a_1|\dots|x-a_d|\max_{|t\in[a,b]}|f^{(d+1)}(t)|. \tag{3.2}$$

En particulier,

$$\max_{x\in[a,b]} |f - \mathbf{L}[A;f]| \leq \frac{1}{(d+1)!} \max_{x\in[a,b]} |f^{(d+1)}| \cdot \max_{x\in[a,b]} |w_A(x)|, \tag{3.3}$$

où w_A est le polynôme de degré $d+1$ défini par

$$w_A(x) = (x-a_0)(x-a_1)\cdots(x-a_d). \tag{3.4}$$

Une conséquence de ce résultat sur le choix des points d'interpolation est esquissée à la partie 3.4.

E 12 * Considérons les réels $a_0 = 100$, $a_1 = 121$ et $a_2 = 144$ et la fonction f définie de $\mathbb{R}^+$ dans lui-même par $f(x) = \sqrt{x}$. Calculer $\mathbf{L}[a_0, a_1, a_2; f](115)$ et montrer que

$$\left|\sqrt{115} - \mathbf{L}[a_0, a_1, a_2; f](115)\right| < 1,8 \cdot 10^{-3}.$$

(Sol. 1 p. 96.)

La démonstration du théorème 8, assez délicate, sera donnée un peu plus loin après que nous nous serons munis de l'outil nécessaire qui est une généralisation du théorème de Rolle habituel.

3.2 Le théorème de Rolle généralisé

Rappelons que théorème de Rolle ordinaire affirme que si f est une fonction continue sur $[a,b]$ et dérivable sur $]a,b[$ telle que $f(a) = f(b) = 0$ alors il existe c tel que $f'(c) = 0$. Ici, nous aurons besoin du

Théorème 9 (de Rolle généralisé)**.** *Si u est un fonction continue sur $[a,b]$ et k fois dérivable sur $]a,b[$ qui s'annule en $k+1$ points x_i, $i = 0,\ldots,k$, alors il existe $c \in]a,b[$ tel que $u^{(k)}(c) = 0$.*

L'énoncé habituel du théorème de Rolle correspond au cas $k = 1$.

Démonstration. Elle consiste à appliquer un grand nombre de fois le théorème de Rolle ordinaire. Nous supposerons, sans perte de généralité que $a \leq x_0 < x_1 < \cdots < x_k \leq b$.

Étape 1. Puisque $u(x_0) = 0 = u(x_1)$, le théorème de Rolle nous dit qu'il existe $c_0 \in]x_0, x_1[$ tel que $u'(c_0) = 0$. De $u(x_1) = 0 = u(x_2)$ nous tirons l'existence de $c_1 \in]x_1, x_2[$ tel que $u'(c_2) = 0$ et, en continuant ainsi, nous construisons k réels $c_i \in]x_i, x_{i+1}[$, $i = 0,\ldots,k-1$, tels que $u'(c_i) = 0$.

*. [Démidovitch & Maron 1979]

Étape 2. Nous reprenons le même raisonnement mais en l'appliquant à la fonction u'. De $u'(c_0) = u'(c_1) = 0$, nous tirons l'existence de $c_0^2 \in]c_0, c_1[$ tel que $u''(c_0^2) = 0$ et, en exploitant de la même manière tous les points c_i, nous obtenons $k-1$ réels $c_i^2 \in]c_i, c_{i+1}[$, $i = 0, \dots, k-1$.

Aux étapes suivantes, nous appliquerons le théorème de Rolle à u'' puis u''' jusqu'à l'appliquer à l'étape $k+1$ à $u^{(k)}$ et arriver à l'existence d'un réel c_1^{k+1} dans $]a, b[$ tel que $u^{(k+1)}(c_1^{k+1}) = 0$ et le théorème est établi. ■

E 13 Rédiger une démonstration par récurrence du théorème 9.

3.3 Démonstration du théorème 8

Démonstration. Fixons $x \in [a, b]$. Si $x \in A$, n'importe quel ξ convient. (Dans ce cas, la formule donne seulement $0 = 0$). Nous supposerons que $x \notin A$. Notons w le polynôme défini par $w(t) = (t - a_0) \dots (t - a_d)$ et prenons $K = K(x) \in \mathbb{R}$ tel que

$$f(x) - \mathbf{L}[A; f](x) = K(x)\, w(x). \tag{3.5}$$

Un tel nombre existe ; il suffit de prendre

$$K(x) = \frac{f(x) - \mathbf{L}[A; f](x)}{w(x)}, \tag{3.6}$$

qui est bien défini car, comme $x \neq a_i$ $(i = 0, \dots, d)$, le dénominateur ne s'annule pas.

Considérons maintenant le fonction u définie sur l'intervalle $[a, b]$ par la relation

$$u(t) = f(t) - \mathbf{L}[A; f](t) - K(x)\, w(t), t \in [a, b]. \tag{3.7}$$

Il faut prendre garde ici que x est un paramètre fixé et t est la variable. Puisque $f \in C^{d+1}[a, b]$, $u \in C^{d+1}[a, b]$. De plus,

$$u(a_i) = f(a_i) - \mathbf{L}[A; f](a_i) - K(x) \times 0 = f(a_i) - f(a_i) = 0, \quad 0 \leq i \leq d;$$

et, par définition de $K(x)$,

$$u(x) = f(x) - \mathbf{L}[A; f](x) - K(x)\, w(x) = 0.$$

Il suit que u s'annule en $d+2$ points, à savoir les $d+1$ points de A auxquels s'ajoute le point x. Le théorème 9 de Rolle généralisé nous permet d'affirmer l'existence de $\xi \in]a, b[$ tel que $u^{(d+1)}(\xi) = 0$. Nous pouvons facilement calculer la dérivée $d+1$-ième de

u. D'abord, puisque $\mathbf{L}[A;f](t)$ est un polynôme de degré d, sa dérivée $d+1$-ième est nulle. Quant à $w(t)$, puisque

$$w(t) = (t-a_0)(t-a_1)\dots(t-a_d) = t^{d+1} + (\text{polynôme de degré} \leq d).$$

Sa dérivée $d+1$-ième, est la constante $(d+1)!$. Finalement,

$$u^{d+1}(t) = f^{(d+1)}(t) - (d+1)!K(x). \tag{3.8}$$

En prenant $t = \xi$, il vient

$$0 = f^{(d+1)}(\xi) - (d+1)!K(x). \tag{3.9}$$

Revenant à la définition de $K(x)$, nous obtenons

$$f(x) - \mathbf{L}[A;f](x) = K(x)w(x) = \frac{f^{(d+1)}(\xi)}{(d+1)!}(x-a_0)\dots(x-a_d). \tag{3.10}$$

Nous avons bien trouvé un nombre ξ dans $]a,b[$ vérifiant la formule annoncée. ■

3.4 Conséquence de la formule d'erreur sur le choix des points d'interpolation

Le second corollaire du theorème 8 montre que si vous nous voulons rendre l'erreur entre la fonction f et son polynôme interpolation la plus petite possible et que nous sommes libres de choisir les points d'interpolation a_i, $i = 0,\dots,a_d$ comme nous le voulons dans $[a,b]$, alors nous avons intérêt à choisir ces points de telle sorte que la quantité $\max_{x\in[a,b]}|w_A|$ soit la plus petite possible, voir (3.4). Il existe un unique ensemble de points qui minimise cette quantité. On les appelle les **points de Chebyshev** en hommage au mathématicien russe qui les a déterminés pour la première fois en 1874. Lorsque $[a,b] = [-1,1]$ ces points sont donnés par la formule

$$a_i = \cos\left(\frac{2i+1}{2(d+1)}\pi\right), \quad i = 0,\dots,d. \tag{3.11}$$

L'exercice 33 montre comment ces points sont obtenus. La figure 1 compare la répartition des points de Chebyshev et des points équidistants, donnés lorsque $[a,b] = [-1,1]$ par la formule $a_i = 1 + 2i/d$, lorsque $d = 50$. Remarquons que les premiers tendent à se densifier lorsqu'on approche des extrémités de l'intervalle.

3.5 Précision de l'interpolant et nombre de points d'interpolation

Il est naturel de penser que plus nous augmenterons le nombre de points d'interpolation, meilleure sera la précision de l'approximation fournie par le polynôme

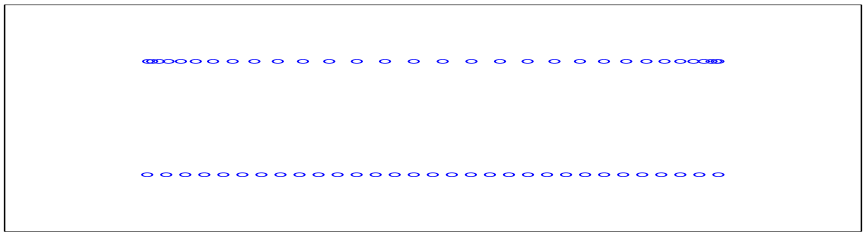


FIGURE 1 – Répartition des points de chebyshev et des points equidistants ($d = 30$)

d'interpolation de Lagrange. Cette intuition est renforcée par les exemples de la table 1 (p. 9). Pourtant, si cette idée reste correcte pour une classe importante de fonctions * et pour des points d'interpolation correctement choisis, elle est fausse dans le cas général. L'exemple classique a été donné par le mathématicien Runge qui a montré en 1901 que les polynômes d'interpolation aux points équidistants de la fonction f définit par $f(x) = 1/(1+x^2)$ donnaient des résultats très mauvais. La table 5 donne le graphe de la fonction d'erreur entre le polynôme interpolation aux points équidistants et la fonction de Runge modifiée $f(x) = 1/(1+100x^2)$ pour quelques valeurs de d. Ici nous avons modifié la fonction de Runge classique pour accélérer le phénomène de divergence. Le problème n'est évidemment pas limité à la fonction de Runge. On peut démontrer que, quels que soient les points d'interpolation choisis, il existe une fonction continue qui ne se laisse pas approcher par ses polynômes interpolation.

Par contre, il est possible de montrer que les polynômes d'interpolation aux points de Chebyshev convergent vers la fonction interpolée, lorsque le nombre de points croît indéfiniment, sous la seule condition que la fonction soit dérivable, de dérivée bornée. Il s'agit ici de **convergence uniforme** des fonctions. Cela signifie que la suite de nombres réels positifs $\max_{x\in[a,b]} |f(x) - \mathbf{L}[a_0, \cdots, a_d\,; f](x)|$, $d \in \mathbb{N}$, converge vers 0 lorsque $d \to \infty$. La convergence cependant peut être lente. La table 6 reprend l'exemple de la fonction de Runge et donne la fonction d'erreur entre cette fonction et le polynôme d'interpolation aux points de Chebyshev.

*. Le lecteur trouvera à l'exercice 36 une classe assez simple de fonctions pour lesquelles les polynômes d'interpolation de Lagrange fournissent toujours d'excellentes approximations.

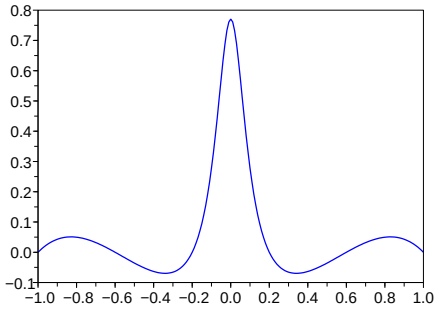
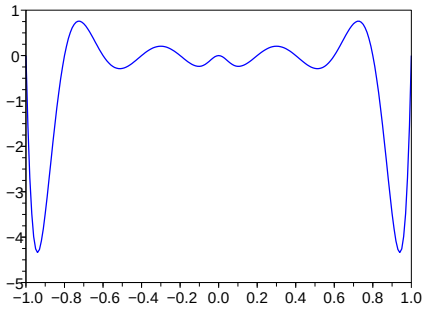
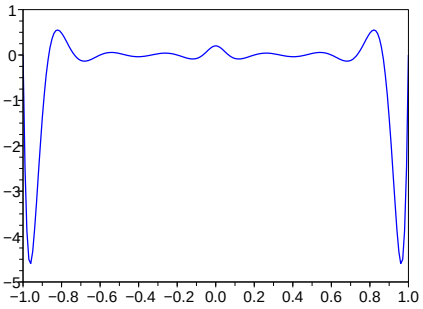
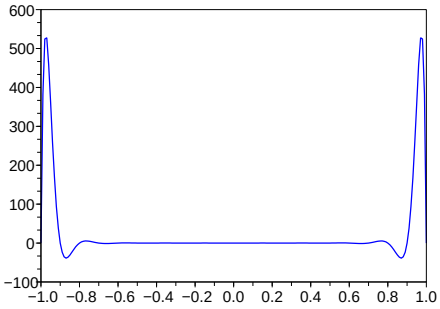
$d=5$	$d=10$
	
$d=15$	$d=20$
	

TABLE 5 – Graphe de la fonction d'erreur entre la fonction $f(x)=\frac{1}{1+100\,x^2}$ et ses polynôme d'interpolation de Lagrange aux points équidistants lorsque $d=5,10,15$ et 20

§ 4 POLYLIGNES

4.1 Subdivisions

Nous appelons **subdivision de longueur** d de $I=[a,b]$ une suite (strictement) croissante de $d+1$ éléments de I, $\sigma=(a_0,\cdots,a_d)$ telle que $a_0=a$ et $a_d=b$. Autrement dit,

$$a=a_0<a_1<a_2<\cdots<a_{d-1}<a_d=b. \tag{4.1}$$

A chaque subdivision σ de longueur d de $[a,b]$ est associée une **partition** de l'intervalle $[a,b]$,

$$[a,b]=[a_0\,,\,a_1[\cup[a_1\,,\,a_2[\cup\ldots[a_{d-2}\,,\,a_{d_1}[\cup[a_{d-1}\,,\,a_d]. \tag{4.2}$$

[TH 9]

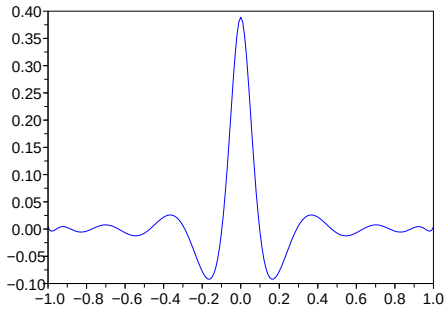
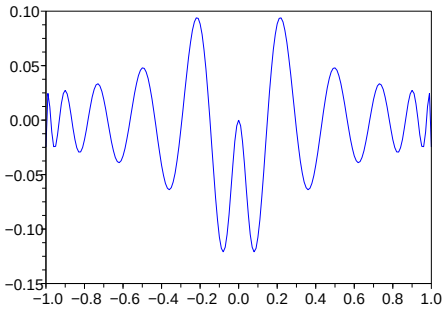
$d = 15$	$d = 20$
	

TABLE 6 – Graphe de la fonction d'erreur entre la fonction $f(x) = \frac{1}{1+100\,x^2}$ et ses polynômes d'interpolation de Lagrange aux points de Chebyshev lorsque $d = 15$ et $d = 20$

La distance entre deux points successifs a_i et a_{i+1} est noté h_i et l'**écart** h de la subdivision σ est la plus grande des distances entre deux points successifs,

$$h = \max_{i=0,\dots,d} h_i = \max_{i=0,\dots,d-1} (a_{i+1} - a_i). \tag{4.3}$$

Ces définitions sont mises en évidence sur la figure 2.

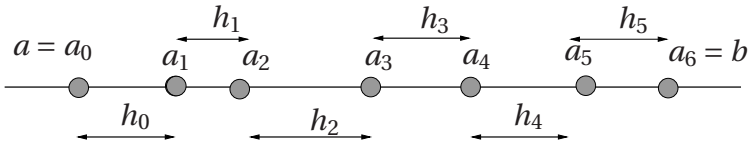


FIGURE 2 – Subdivision et écart d'une subdivision.

Lorsque les distances h_i sont constantes, $h_i = (b-a)/d$, la subdivision est formée des points équidistants

$$\sigma = \left(a + i\frac{b-a}{d} \,:\, i = 0, \dots, d \right).$$

Nous disons qu'une fonction g est **affine par morceaux** sur l'intervalle I s'il existe une subdivision $\sigma = (a_0, \dots, a_d)$ de l'intervalle I telle que la restriction de g à chacun des sous-intervalles défini par **sigma** soit une fonction affine, c'est-à-dire pour $i = 0, \dots, d-1$, il existe des coefficients α_i et β_i tels que

$$x \in [a_i\,,\, a_{i+1}[\implies g(x) = \alpha_i\, x + \beta_i.$$

Remarquons que cette définition n'impose aucune condition sur la valeur de g à l'extrémité $b = a_d$ de l'intervalle.

E 14 A quelles conditions (sur les nombres α_i et β_i) la fonction g est-elle continue ? (continue et convexe) ? Que dire de la dérivabilité des fonctions affines par morceaux ?

4.2 Fonctions polylignes

Soit σ une subdivision de $[a,b]$ et $f = (f_0,\dots,f_d)$ une suite de $d+1$ valeurs quelconques. Nous pouvons construire les polynômes de Lagrange

$$\mathbf{L}[a_i\,,\,a_{i+1}\,;\,f_i\,,\,f_{i+1}],\quad i=0,\dots,d-1,$$

c'est-à-dire les uniques polynômes de degrés inférieur ou égal à 1 qui prennent les valeurs f_i au point a_i et f_{i+1} au point a_{i+1}. La fonction **polyligne** associée à la subdivision σ et aux valeurs f, notée $\mathbf{PL}[\sigma,f]$, est définie sur chacun des sous-intervalles de la subdivision par la relation

$$\begin{cases} \mathbf{PL}[\sigma\,,\,f](x) = \mathbf{L}[a_i\,,\,a_{i+1}\,;\,f_i\,,\,f_{i+1}](x), \quad x\in[a_i\,,\,a_{i+1}[,\\ \mathbf{PL}[\sigma\,,\,f](b) = f_d. \end{cases} \tag{4.4}$$

Lorsque les valeurs f_i sont les valeurs d'une fonction f aux points a_i,

$$y_i = f(a_i),\quad i=0\,,\dots,d,$$

nous disons que $\mathbf{PL}[\sigma;f]$ est la (fonction) polyligne interpolant la fonction f aux points de la subdivision σ.

Les deux schémas dans le tableau 7 font apparaître en rouge les graphes des fonctions $\mathbf{PL}[\sigma;f]$ lorsque $f(x)=x^3$ (tracé en bleu) et (1) $\sigma=(-1,-0.5,1)$ puis (2) $\sigma=(-1,-0.5,0,0.5,1)$.

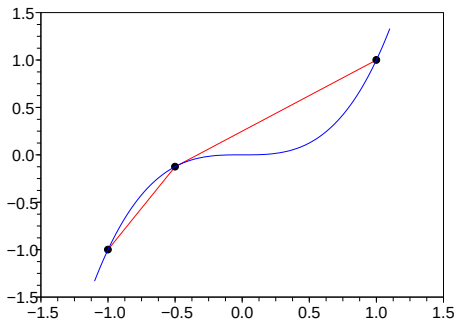
Théorème 10. $\mathbf{PL}[\sigma,f]$ *est une fonction affine par morceaux continue satisfaisant*

$$\mathbf{PL}[\sigma,f](a_i) = f_i,\quad i=0,\dots,d. \tag{4.5}$$

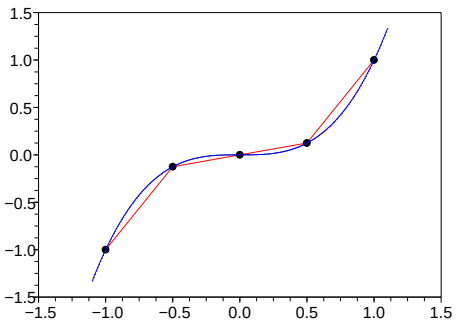
Démonstration. D'après la définition même, $\mathbf{PL}[\sigma,f]$ est une fonction affine par morceaux. D'autre part,

$$i\in\{0,\dots,d-1\} \implies \mathbf{PL}[\sigma,f](a_i) = \mathbf{L}[a_i\,,\,a_{i+1}\,;\,f_i\,,\,f_{i+1}](a_i) = f_i,$$

et nous avons aussi, toujours par définition, $\mathbf{PL}[\sigma,f](a_d) = f_d$. La seule propriété que nous devons démontrer est la continuité. Les éventuels problèmes de continuité d'une fonction affine par morceaux se trouvent aux points de jonction des sous-intervalles de



(1)



(2)

TABLE 7 – Deux exemples de polylignes

la subdivision, ici, aux points a_i, $i = 1, \dots, d$. Commençons par prendre un point a_i avec $1 \le i \le d-1$ de sorte que nous excluons le cas $a_i = a_d$. Pour montrer la continuité de $\mathbf{PL}[\sigma, f]$ en ce point, il suffit de s'assurer que les limites à gauche et à droite coïncident (et sont égales à la valeur de la fonction f au point). Or, d'une part,

$$\begin{aligned}\lim_{x \to a_i^+} \mathbf{PL}[\sigma, f](x) &= \lim_{x \to a_i^+} \mathbf{L}[a_i\,, a_{i+1}\,; f_i\,, f_{i+1}](x) \\ &= \mathbf{L}[a_i\,, a_{i+1}\,; f_i\,, f_{i+1}](a_i) = f(a_i) = \mathbf{PL}[\sigma; f](a_i),\end{aligned}$$

et, d'autre part,

$$\begin{aligned}\lim_{x \to a_i^-} \mathbf{PL}[\sigma, f](x) &= \lim_{x \to a_i^-} \mathbf{L}[a_{i-1}\,, a_i\,; f_{i-1}\,, f_i](x) \\ &= \mathbf{L}[a_{i-1}\,, a_i\,; f_{i-1}\,, f_i](a_i) = f(a_i) = \mathbf{PL}[\sigma, f](a_i).\end{aligned}$$

Les deux limites coïncident et sont égales à $f(a_i)$ donc la fonction est bien continue au point a_i. Il reste à étudier le cas $i = d$, c'est-à-dire $a_d = b$, qui se traite de la même manière, mis à part le fait que nous étudions uniquement la limite à gauche. Notons qu'il n'y a pas de problème de continuité au point a_0. ■

Théorème 11. *Si f est un polynôme de degré au plus* 1*, c'est-à-dire une fonction affine, alors* $\mathbf{PL}[\sigma; f] = f$.

Démonstration. Cela provient du fait que lorsque f est un polynôme de degré au plus 1, nous avons $\mathbf{L}[a_i, a_{i+1}; f] = f$ si bien que f elle-même vérifie les conditions (4.4) de la définition. ■

E 15 Montrer que l'application qui à une fonction f définie sur I — $f \in \mathscr{F}(I)$ — fait correspondre $\mathbf{PL}[A; f] \in \mathscr{F}(I)$ est une application linéaire.

E 16 Montrer que si f est une fonction polynomiale telle que $\mathbf{PL}[\sigma; f] = f$ alors f est nécessairement de degré au plus 1.

E 17 Expliquer pourquoi la fonction $\mathbf{PL}[\sigma, f]$ n'est pas la seule fonction affine par morceaux A vérifiant $A(a_i) = f(a_i)$, $i = 0, \ldots, d$. Quelle propriété supplémentaire, non formulée dans le théorème, caractérise-t-elle $\mathbf{PL}[\sigma, f]$?

4.3 Approximation des fonctions continûment dérivables par les fonctions polylignes

A la différence des polynômes d'interpolation de Lagrange, les fonctions polylignes fournissent une bonne approximation de toutes les fonctions continues, pour peu que l'écart de la subdivision soit suffisamment petit. Cela n'est pas surprenant. Les polylignes sont des objets beaucoup plus souples que les polynômes. Nous pouvons changer la valeur d'une polyligne au point a sans rien changer à la valeur au point b ; par contre une petite modification de la valeur d'un polynôme en a peut provoquer un grand écart de valeur en b. Nous pouvons dire que les valeurs d'un polynôme sont solidaires les unes des autres tandis que celles d'une polyligne – en des points suffisamment éloignés – sont complètement indépendantes. Le prix de la souplesse des polylignes est cependant lourd à payer : ce sont des fonctions très peu régulières, elles sont continues mais non dérivables sur $[a, b]$. Plus précisément, sauf cas exceptionnel, une fonction polyligne n'est dérivable en aucun des points de jonction. Un autre inconvénient peut-être plus sérieux est que la précision des interpolants polylignes est limitée. Quelle soit la fonction non affine considérée, l'erreur globale entre la fonction interpolée et la fonction polyligne ne pourra jamais décroître vers 0 que comme la suite $1/d^2$ où d dénote la longueur de la subdivision utilisée [*]. Au contraire, les polynômes d'interpolation bénéficient des propriétés de la fonction interpolée et, si les fonctions sont suffisamment régulières, l'erreur pourra décroître aussi vite qu'une suite géométrique r^d avec $0 < r < 1$ où d est le degré du polynôme d'interpolation [†].

Nous nous limiterons ici démontrer un théorème sur l'approximation des fonctions dérivables, de dérivées continues. Le cas des fonctions seulement continues sera traité plus bas (4.5) en complément. Un autre estimation, concernant les fonctions deux fois continûment dérivables est proposée à l'exercice 39.

Théorème 12. *Soit f une fonction continûment dérivable sur $[a, b]$ et σ une subdivi-*

*. Pour une explication de ce phénomène, le lecteur pourra consulter le commentaire de l'exercice 39
†. Voir l'exercice 36 et le commentaire qui le suit.

sion de $[a,b]$ *d'écart* h. *Pour tout* $x \in [a,b]$,

$$\left|f(x) - \mathbf{PL}[\sigma; f](x)\right| \leq h \cdot \max_{t\in[a,b]} |f'(t)|. \tag{4.6}$$

Démonstration. L'inégalité a démontrer est évidente lorsque x est l'un des points de la subdivision $\sigma = (a_0, \ldots, a_d)$ car alors $f(x) = \mathbf{PL}[\sigma; f](x)$. Nous supposerons que $x \neq a_i$, $i = 0, \ldots, d+1$. Dans ce cas, x appartient à un et un seul des sous-intervalles (ouverts) définis par la subdivision, disons, $x \in]a_j, a_{j+1}[$. Il suit que

$$f(x) - \mathbf{PL}[\sigma; f](x) = f(x) - \mathbf{L}[a_j, a_{j+1}; f](x) = f(x) - f(a_j)\ell_0(x) - f(a_{j+1})\ell_1(x), \tag{4.7}$$

où ℓ_0 et ℓ_1 sont les polynômes fondamentaux de Lagrange ,

$$\ell_0(x) = \frac{x - a_{j+1}}{a_j - a_{j+1}} \quad \text{et} \quad \ell_1(x) = \frac{x - a_j}{a_{j+1} - a_j}.$$

Mais nous avons vu – c'est l'équation (1.19) – que la somme des polynômes fondamentaux de Lagrange est toujours égale à 1, ici, $\ell_0 + \ell_1 = 1$. En utilisant cette relation, nous obtenons

$$f(x) - \mathbf{PL}[\sigma; f](x) = [f(x) - f(a_j)]\,\ell_0(x) + [f(x) - f(a_{j+1})]\,\ell_1(x) \tag{4.8}$$

$$\Longrightarrow |f(x) - \mathbf{PL}[\sigma; f](x)| \leq |f(x) - f(a_j)||\ell_0(x)| + |f(x) - f(a_{j+1})||\ell_1(x)|. \tag{4.9}$$

Maintenant, il résulte de $x \in]a_j, a_{j+1}[$ que $|x - a_j| \leq |a_{j+1} - a_j|$ et ceci entraîne $|\ell_1(x)| \leq 1$. Un argument similaire assure que $|\ell_0(x)| \leq 1$. En reportant ces deux nouvelles informations dans l'inégalité ci-dessus, il vient

$$|f(x) - \mathbf{PL}[\sigma; f](x)| \leq |f(x) - f(a_j)| + |f(x) - f(a_{j+1})|. \tag{4.10}$$

L'inégalité des accroissements finis donne finalement

$$|f(x) - \mathbf{PL}[\sigma; f](x)| \leq |x - a_j| \max_{t\in[a,b]} |f'(t)| + |x - a_{j+1}| \max_{t\in[a,b]} |f'(t)| \tag{4.11}$$

$$\leq (x - a_j) \max_{t\in[a,b]} |f'(t)| + (a_{j+1} - x) \max_{t\in[a,b]} |f'(t)| \tag{4.12}$$

$$\leq (a_{j+1} - a_j) \max_{t\in[a,b]} |f'(t)| \leq h \cdot \max_{t\in[a,b]} |f'(t)|. \tag{4.13}$$

La dernière égalité provenant de la définition de l'écart d'une subdivision. L'inégalité annoncée a été établie. ∎

Corollaire. *Si* σ^d, $d \in \mathbb{N}$, *est une suite de subdivisions de longueur* d *de* $[a,b]$ *dont l'écart tend vers* 0 *lorsque* d *tend vers* ∞ *alors*

$$\lim_{d\to\infty} \mathbf{PL}[\sigma^d; f] = f(x), \quad x \in [a,b]. \tag{4.14}$$

S'agissant d'une suite de subdivisions, à chaque changement de d, les points de la subdivision changent, excepté le premier qui doit toujours être égal à a et le dernier qui doit toujours être égal à b,

$$\sigma^d = (a, a_1^d, a_2^d, \dots, a_{d-1}^d, b).$$

Naturellement, l'écart de la subdivision σ^d dépend de d.

Corollaire. *Lorsque σ^d est la subdivision formée des points équidistants*

$$a_i = a + i \cdot \frac{b-a}{d}, \quad i = 0, \dots, d+1, \quad d \in \mathbb{N}^\star,$$

alors

$$\left| f(x) - \mathbf{PL}[\sigma^d; f](x) \right| \leq \frac{b-a}{d} \cdot \max_{t \in [a,b]} |f'(t)| \overset{d\to\infty}{\to} 0, \quad x \in [a,b]. \tag{4.15}$$

E 18 Les convergences des deux résultats précédents sont-elles aussi des convergences uniformes . Autrement dit, a-t-on

$$\lim_{d\to\infty} \max_{x\in[a,b]} |f(x) - \mathbf{PL}[\sigma^d; f](x)| = 0 \ ?$$

4.4 Représentation

Nous allons déterminer des fonctions $b_i = b_i^\sigma$ adaptées à la subdivision σ qui permettent une représentation simple du polyligne $\mathbf{PL}[\sigma; f]$. Pour que tous les points a_i, $i = 0, \dots, d$ jouent un rôle semblable, Nous sommes amené à compléter la subdivision σ par deux points a_{-1} et a_{d+1} comme indiqué sur la figure 3. Ces points peuvent être choisis librement sous les seules conditions que $a_1 < a = a_0$ et $a_{d+1} > a_d = b$.

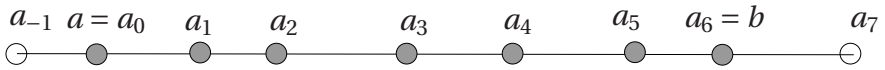


FIGURE 3 – Sudvision complétée des points a_{-1} et a_{d+1}.

Une fois la subdivision complétée, nous définissons pour $i = 0, \dots, d$ la fonction b_i sur $\mathbb{R}$ par le graphe donné dans la figure 4.4.

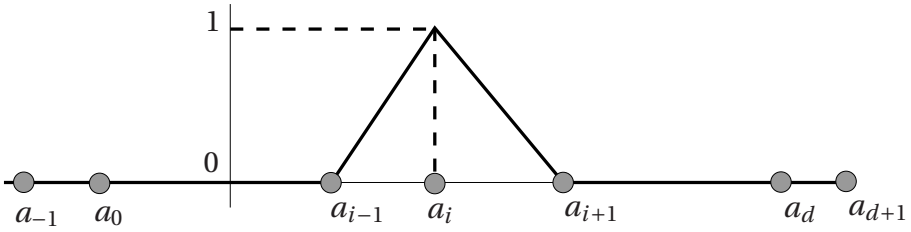


FIGURE 4 – graphe de la fonction b_i.

En formule, la fonction b_i est définie par

$$b_i(x) = \begin{cases} 0 & \text{if } x \le a_{i-1}, \\ \frac{x-a_{i-1}}{a_i-a_{i-1}} & \text{if } a_{i-1} \le x \le a_i, \\ \frac{x-a_{i+1}}{a_i-a_{i+1}} & \text{if } a_i \le x \le a_{i+1}, \\ 0 & \text{if } x \ge a_{i+1}. \end{cases} \tag{4.16}$$

Les fonctions b_i, $i = 0, \dots, d$, sont affines par morceaux, continues, positives ou nulles et bornées par 1,

$$0 \le b_i(x) \le 1, \quad x \in [a, b], \tag{4.17}$$

et s'annulent en tout les points a_j sauf lorsque $j = i$ auquel cas nous avons $b_i(a_i) = 1$. Remarquons aussi qu'elles nulles en dehors de l'intervalle $[a_{i-1}, a_{i+1}]$. Cet intervalle est appelé le **support** de la fonction b_i.

Théorème 13. $\mathbf{PL}[\sigma; f] = \sum_{i=0}^{d} f_i b_i$.

Démonstration. Appelons g la fonction définie par la partie droite de l'égalité à démontrer. Il suffit d'établir que pour tous $j = 1, \dots, d-1$ et $x \in [a_j, a_{j+1}[$ nous avons $g(x) = \mathbf{L}[a_j, a_{j+1}; f](x)$ ainsi que $g(a_d) = f(a_d)$ car la définition même de $\mathbf{PL}[\sigma; f]$ nous donnera alors $g = \mathbf{PL}[\sigma; f]$. Prenons donc $j \in \{1, \dots, d-1\}$ et $x \in]a_j, a_{j+1}[$. Remarquons que nous considérons ici l'intervalle ouvert $]a_j, a_{j+1}[$ alors que l'égalité doit être établie sur l'intervalle semi fermé $[a_j, a_{j+1}[$. Nous traiterons à part le cas $x = a_j$. Puisque $x \in]a_j, a_{j+1}[$, nous avons $b_i(x) = 0$ sauf si $i = j$ ou $i = j+1$. Nous en déduisons en tenant compte de (4.16) que

$$g(x) = f_j b_j(x) + f_{j+1} b_{j+1}(x) = f_j \frac{x - a_{j+1}}{a_j - a_{j+1}} + f_{j+1} \frac{x - a_j}{a_{j+1} - a_j} = \mathbf{L}[a_j, a_{j+1}; f](x),$$

qui est bien l'égalité obtenir. Pour la cas particulier où $x = a_j$, il suffit de remarquer que $b_i(a_j) = \delta_{ij}$, par conséquent $g(a_j) = f_j = \mathbf{L}[a_j, a_{j+1}; f](a_j)$. Le même raisonnement est valable pour a_d et la démonstration est terminée. ∎

Corollaire. $\sum_{i=0}^{d} b_i = 1$.

Démonstration. Il suffit de prendre la fonction constant égale à 1 (qui est un polynôme de degré ≤ 1) dans le théorème ci-dessus et d'utiliser ensuite le théorème 11. ■

COMPLÉMENT

4.5 Approximation des fonctions continues par des fonctions polylignes

Théorème 14. *Soient f une fonction continue sur $[a,b]$ et σ^d, $d \in \mathbb{N}$, une suite de subdivisions de $[a,b]$. Nous supoosons que la longueur est de σ_d est d et que l'écart de σ_d tend vers 0 lorsque d tend vers ∞. La suite des polylignes $\mathbf{PL}[\sigma^d; f]$ converge uniformément vers f sur $[a,b]$. Autrement dit,*

$$\lim_{d\to\infty} \max_{x\in[a,b]} |f(x) - \mathbf{PL}[\sigma^d; f](x)| = 0. \tag{4.18}$$

La démonstration utilise la notion d'**uniforme continuité**. Rappelons qu'une fonction f définie sur un intervalle I est uniformément continue sur I si la distance entre les valeurs $f(x)$ et $f(y)$ est petite chaque fois que la distance entre x et y est assez petite. De manière précise, pour tout réel positif ϵ, il doit exister un réel positif η, dépendant de ϵ, de telle sorte que $|f(x) - f(y)| \leq \epsilon$ pour tout couple de valeurs (x,y) dans I satisfaisant $|x-y| \leq \eta$. Il y a une manière plus commode de présenter la propriété d'uniforme continuité. Notons

$$\omega_f(\eta) = \sup\{|f(x) - f(y)| \,:\, (x,y) \in I \text{ et } |x-y| \leq \eta\},$$

de sorte que $\omega_f(\eta)$ est la plus grande des distances possibles entre $f(x)$ et $f(y)$ lorsque les deux éléments x et y de I sont distants d'au plus η. Dans ces conditions, nous avons

$$f \text{ uniformément continue sur } I \iff \lim_{\eta\to 0} \omega_f(\eta) = 0.$$

La fonction ω_f s'appelle le **module de continuité** de f. Elle joue un rôle important en analyse.

Dans la définition de la continuité ordinaire, nous parlons de continuité de f en un point x_0 qui est fixe et cherchons à vérifier l'existence d'un η, dépendant de ϵ et de x_0, tel que $|f(y) - f(x_0)| \leq \epsilon$ lorsque $|y - x_0| \leq \eta$. Dans le cas de la continuité uniforme, deux valeurs varient, x et y, contre une seule, y, dans le cas de la continuité ordinaire. Malgré cette différence importante, un théorème fondamental de l'analyse, connu sous le nom de **théorème de Heine**, établit que *si I est un intervalle fermé borné, c'est-à-dire de la forme $I = [a,b]$, alors toute fonction continue est aussi uniformément continue.* La condition sur la forme de l'intervalle I est importante et la propriété est fausse en général dans le cas des autres types d'intervalle.

Démonstration. Nous devrons montrer que pour tout $\epsilon > 0$ fixé à l'avance, il existe $d_0 \in \mathbb{N}$, d_0 dépendant de ϵ, tel que

$$d \geq d_0 \implies \max_{x\in[a,b]} |f(x) - \mathbf{PL}[\sigma^d; f](x)| \leq \epsilon. \tag{4.19}$$

Nous noterons $\sigma^d = (a_0^d, a_1^d, \ldots, a_{d-1}^d, a_d^d)$ avec $a_0^d = a$ et $a_d^d = b$. L'écart de σ^d sera noté h_d. Soit x un élément quelconque de $[a,b]$, x se trouve dans un unique intervalle $[a_j^d, a_{j+1}^d[$. Reprenons la relation (4.10),

$$|f(x) - \mathbf{PL}[\sigma^{\mathbf{d}}; f](x)| \leq |f(x) - f(a_j^d)| + |f(x) - f(a_{j+1}^d)|.$$

Puisque $|x - a_j^d| \leq h_d$ et $|x - a_{j+1}^d| \leq h_d$, l'inégalité ci-dessus et la définition du module de continuité ω_f donne

$$|f(x) - \mathbf{PL}[\sigma^{\mathbf{d}}; f](x)| \leq 2\omega_f(h_d).$$

Puisque cette estimation est valable pour tout x dans $[a,b]$, nous avons aussi

$$\max_{x\in[a,b]} |f(x) - \mathbf{PL}[\sigma^{\mathbf{d}}; f](x)| \leq 2\omega_f(h_d). \tag{4.20}$$

Puisque f est continue, elle est aussi, d'après le théorème de Heine, uniformément continue, de sorte que $\lim_{\eta\to 0}\omega_f(\eta) = 0$ et comme $\lim_{d\to\infty} h_d = 0$ par hypothèse, par composition des limites, $\lim_{d\to\infty} 2\omega_f(h_d) = 0$ ce qui entraîne à son tour, en vue de (4.20)

$$\lim_{d\to\infty} \max_{x\in[a,b]} |f(x) - \mathbf{PL}[\sigma^{\mathbf{d}}; f](x)| = 0,$$

ce qu'il fallait établir. ■

E 19 Dans la démonstration précédente, il faudrait en toute rigueur écrire $a_{j_d}^d$ et $a_{j_d+1}^d$ plutôt que a_j^d et a_{j+1}^d. Pourquoi ?

E 20 Soit, pour tout $d \geq 2$, $\sigma^d = (a = a_0^d, a_1^d, \ldots, a_{d-1}^d, a_d^d = b)$ une subdivision de longueur d de $[a,b]$ et d'écart h_d. On suppose que h_d tend vers 0 lorsque $d \to \infty$. Soient $y^d = (y_0^d, \ldots, y_d^d)$ une suite de $d+1$ valeurs et f une fonction continue sur $[a,b]$. Montrer que les deux conditions suivantes sont équivalents.

(a) $\lim_{d\to\infty} \max_{x\in[-1,1]} |f(x) - \mathbf{PL}[\sigma^d; y^d]| = 0$.

(b) $\lim_{d\to\infty} \max_{i=0,\ldots,d} |y_i^d - f(a_i^d)| = 0$.

On prendra garde que $\mathbf{PL}[\sigma^d; y^d]$ n'est pas $\mathbf{PL}[\sigma^d; f]$.

4.6 Extension

Plutôt que de se limiter à des fonctions affines par morceaux, c'est-à-dire polynomiales de degré 1 par morceaux, il est naturel de considérer des fonctions polynomiales de degré d par morceaux. La construction donnée ci-dessus s'étend immédiatement à ce cas. Il suffit simplement de remplacer les polynômes de Lagrange $\mathbf{L}[a_i, a_{i+1}; f]$ par des polynômes $\mathbf{L}[a_i, a_{i,1}, \ldots, a_{i,d-1}, a_{i+1}; f]$ où les $a_{i,j}$, $j = 1, \ldots, d-1$ sont des points

intérieurs à l'intervalle $[a_i, a_{i+1}]$. Cette extension de la théorie n'a pas beaucoup d'intérêt. Il est préférable de se concentrer sur l'insuffisance majeure des polylignes à savoir d'être non différentiable. En augmentant le degré des polynômes sur chaque sous-intervalle $[a_i, a_{i+1}]$, il est possible de les raccorder pour obtenir des fonctions plusieurs fois dérivables sur $[a, b]$. C'est la théorie des fonctions splines qui joue un rôle important en analyse numérique mais que nous ne pourrons pas étudier dans ce cours.

§ 5 EXERCICES ET PROBLÈMES

21 Un problème d'interpolation général. Trouver une condition sur la paire $(a, b) \in \mathbb{R}^2$ pour que la proposition suivante soit vraie : Quel que soit le triplet $(\alpha, \beta, \gamma) \in \mathbb{R}^3$ il existe un et une seul $p \in \mathscr{P}_2$ tel que $p(a) = \alpha$, $p(b) = \beta$, $p(a) + p'(b) = \gamma$.

22 Un problème d'interpolation des dérivées. Soient a, b et c trois nombres réels. Montrer que quels que soient les réels α, β, γ il existe un et un seul polynôme $p \in \mathscr{P}_2$ tel que $p(a) = \alpha$, $p'(b) = \beta$ et $p''(c) = \gamma$. (Sol. 4 p. 97)

23 Un exemple. Déterminer le polynôme d'interpolation de Lagrange de $f(x) = 1/(1+x)$ par rapport aux points $0, 3/4, 1$. Représenter sur un même graphique le polynôme et la fonction interpolée. Comparer, a l'aide d'une calculatrice, $f(1/2)$ et $\mathbf{L}[0, 3/4, 1; f](1/2)$.

24 Propriétés générales de l'interpolation. Soit $I = [a, b]$, $f : \mathbb{R} \to \mathbb{R}$ et $A = \{a_0, \ldots, a_d\} \in I$. Les assertions suivantes sont-elles vraies ou fausses ?

(a) i) si $\mathbf{L}[A; f]$ est un polynôme constant alors d est nécessairement égal à 0. ii) si $d > 1$ et $\mathbf{L}[A; f]$ est un polynôme constant alors f est nécessairement constante.

(b) Si $d = 1$ et f est une fonction croissante (resp. décroissante) sur I alors $\mathbf{L}[A; f]$ est croissante (resp. décroissante) sur I.

(c) Même question lorsque $d = 2$.

25 Interpolation et division euclidienne. Rappellons que la division euclidienne d'un polynôme V par un polynôme W non nul consiste à écrire (de manière unique) V sous la forme $V = qW + r$ où q et r sont deux polynômes, le second vérifiant $\deg(r) < \deg(W)$. On appelle q le quotient de la division euclidienne de V et W et r le reste de cette division.

(a) Montrer que si $W(x) = (x - a_0)(x - a_1) \cdots (x - a_d)$ alors $r = \mathbf{L}[A; V]$.

(b) Utiliser une division euclidienne pour calculer le polynôme d'interpolation de Lagrange de $V(x) = x^5 - 3x^4 + x - 3$ aux points $-1, 0, 1, 2$. Vérifier le résultat obtenu.

26 Formule de Simpson. Soient $a < b$ deux réels distincts. On pose $m = \frac{a+b}{2}$.

(a) Donner la formule de Lagrange pour le polynôme d'interpolation $\mathbf{L}[a, m, b; f]$.

(b) Démontrer que

$$\int_a^b \mathbf{L}[a, m, b; f](x)\,dx = \frac{b-a}{6}[f(a) + 4f(m) + f(b)].$$

NOTE. — Le résultat obtenu s'appelle la **formule de Simpson**. Nous l'étudierons dans le chapitre suivant (II.2.3).

27 Invariance des polynômes d'interpolation par les bijections affines. Soit h une bijection affine, $h(x) = \alpha x + \beta$, $(\alpha, \beta) \in \mathbb{R}^\star \times \mathbb{R}$. Soient $A = \{a_0, \dots, a_d\}$ et f une fonction définie sur $\mathbb{R}$.

A) Montrer que

$$\mathbf{L}[a_0, \dots, a_d; f \circ h](x) = \mathbf{L}[h(a_0), \dots, h(a_d); f](h(x)).$$

On commencera par expliciter et vérifier cette relation dans le cas où $d = 1$.

B) A quelle(s) condition(s) sur l'ensemble A les assertions suivantes sont-elles vraies ?

(a) Si f est une fonction paire alors $\mathbf{L}[A; f]$ est un polynôme pair.

(b) Si f est une fonction impaire alors $\mathbf{L}[A; f]$ est un polynôme impair.

28 Coefficients des polynômes d'interpolation. Montrer que si $L[a_0, \dots, a_d; f](x) = \sum_{j=0}^{d} c_j x^j$ alors les coefficients c_j sont solution du système

$$\begin{cases} c_0 + c_1 a_0 + c_2 a_0^2 + \cdots + c_d a_0^d = f(a_0) \\ c_0 + c_1 a_1 + c_2 a_1^2 + \cdots + c_d a_1^d = f(a_1) \\ \qquad\qquad \cdots \\ c_0 + c_1 a_d + c_2 a_d^2 + \cdots + c_d a_d^d = f(a_d) \end{cases}$$

Ecrire puis calculer le déterminant de ce système. On pourra se limiter au cas où $d = 2$. Retrouver la formule d'interpolation de Lagrange à l'aide des formules de Kramer.

29 Groupement des points d'interpolation. Soit $X = \{x_1, x_2, \dots, x_n\}$ et $Y = \{y_1, y_2, \dots, y_m\}$ deux ensembles respectivement de n et m nombres réels (deux à deux distincts). On suppose que X et Y n'ont aucun point en commun, autrement dit $X \cap Y = \emptyset$. On pose

$$p(x) = (x - x_1)(x - x_2)\dots(x - x_n) \quad \text{et} \quad q(x) = (x - y_1)(x - y_2)\cdots(x - y_m).$$

Pour toute fonction f définie (au moins) sur $X \cup Y$ on considère le polynôme

$$R_f(x) = q(x)\mathbf{L}\left[X; \frac{f}{q}\right](x) + p(x)\mathbf{L}\left[Y; \frac{f}{p}\right](x).$$

Comme d'habitude la notation $\mathbf{L}[X; f/q](x)$ (resp. $\mathbf{L}[Y; f/p](x)$) désigne le polynôme d'interpolation de Lagrange de la fonction f/q (resp. f/p) par rapport aux points de X (resp. de Y).

(a) Que peut-on dire du degré de $R_f(x)$ en fonction de n et m ?

(b) Calculer $R_f(x_i)$, $i = 1, \dots, n$ et $R_f(y_j)$, $j = 1, \dots, m$.

(c) En déduire que R_f est un polynôme d'interpolation de Lagrange que l'on précisera.

(Sol. 5 p. 98.)

30 Formule de Lagrange barycentrique. Soient $A = \{a_0, \dots, a_d\} \in I = [a, b]$ et f une fonction définie sur I. On note

$$w_A(x) = (x - a_0)(x - a_1) \cdots (x - a_d).$$

et pour $i = 0, \dots, d$,

$$w_{A,i} = w_A(x)/(x - a_i).$$

On remarquera que $w_{A,i}$ est un polynôme de degré (exactement) d.

(a) Montrer que $w'_A(a_i) = w_{A,i}(a_i)$. (Dériver la relation $w_A(x) = (x - a_i) \cdot w_{A,i}(x)$).

(b) En déduire en partant de la formule d'interpolation de Lagrange que

$$\mathbf{L}[A; f](x) = \sum_{i=0}^{d} f(a_i) \frac{w_A(x)}{w'_A(a_i)(x - a_i)}.$$

On note

$$\Delta_i = \frac{1}{w'_A(a_i)}.$$

(c) Montrer, en utilisant la relation $\mathbf{L}[A; 1](x) = 1$ que

$$w_A(x) \cdot \sum_{i=1}^{d} \frac{\Delta_i}{x - a_i} = 1.$$

(d) Montrer que

$$\mathbf{L}[A; f](x) = \frac{\sum_{i=0}^{d} f(a_i) \frac{\Delta_i}{x - a_i}}{\sum_{i=0}^{d} \frac{\Delta_i}{x - a_i}}. \tag{5.1}$$

(e) Ecrire un algorithme basé sur la formule ci-dessus pour calculer les valeurs du polynôme d'interpolation de Lagrange. Calculer le nombre d'opérations employé par cet algorithme.

NOTE. — La formule (5.1) s'appelle la **formule de Lagrange barycentrique**.

31 Un exemple. On souhaite obtenir une approximation de $\cos(\pi/5)$ connaissant $\cos(\pi/4) = \sqrt{2}/2$, $\cos(\pi/6) = \sqrt{3}/2$ et $\cos 0$. Pour cela on considère $f(x) = \cos(\pi x)$ et son polynôme d'interpolation de Lagrange $\mathbf{L}[0, 1/6, 1/4; f]$.

(a) Calculer $\alpha = \mathbf{L}[0, 1/6, 1/4; f](1/5)$.

(b) Donner une estimation de l'erreur $|\cos(\pi/5) - \alpha|$.

(Sol. 2 p. 96.)

32 Polynômes d'interpolation d'une fraction rationnelle. On considère $d + 1$ nombres réels $a_0, a_1, \dots, a_d$ deux à deux distincts et λ un paramètre réel différent de chacun des a_i c'est-à-dire $\lambda \neq a_i$ pour $i = 0, 1, \dots, a_d$. On pose $w(x) = (x - a_0)(x - a_1) \cdots (x - a_d)$ et on considère la fonction $f_\lambda(x)$ définie par

$$f_\lambda(x) = \frac{w(\lambda) - w(x)}{w(\lambda)(\lambda - x)}.$$

(a) Montrer que le polynôme $r(x) =_{def} (\lambda - x)$ divise le polynôme $q(x) =_{def} w(\lambda) - w(x)$. En déduire que f_λ est un polynôme. Préciser le degré de f_λ.

(b) Calculer $f_\lambda(a_i)$ pour $i = 0, 1, \ldots, d$ et en déduire que f_λ est le polynôme d'interpolation de Lagrange par rapport aux points $a_0, a_1, \ldots, a_d$ d'une fraction rationnelle g_λ que l'on précisera i.e. $f_\lambda = \mathbf{L}[a_0, a_1, \ldots, a_d; g_\lambda]$. (Sol. 3 p. 97.)

33 Polynômes de Chebyshev. On définit une suite de polynômes $T_d(x)$ par la relation de récurrence

$$\begin{cases} T_0(x) = 1, \quad T_1(x) = x \\ T_{d+1}(x) = 2xT_d(x) - T_{d-1}(x), \quad d \geq 1 \end{cases}.$$

Les polynômes T_d forment la suite des polynômes de Chebyshev.

(a) Déterminer T_2, T_3 et T_4 ?

(b) Montrer que pour tout $d \in \mathbb{N}$, T_d est un polynôme de degré d et son coefficient de plus haut degré est 2^{d-1} pour $d \geq 1$.

(c) Montrer que si d est pair alors T_d est un polynôme pair et si d est impair, T_d est un polynôme impair.

(d) Montrer que pour tout $d \in \mathbb{N}$ on a

$$T_d(\cos\theta) = \cos(d\theta), \quad \theta \in \mathbb{R}.$$

On pourra utiliser les relations trigonométriques suivantes :

$$\cos(a+b) = \cos a \cos b - \sin a \sin b$$
$$\cos(a-b) = \cos a \cos b + \sin a \sin b.$$

(e) Montrer que le polynôme T_{d+1} possède exactement $d+1$ racines r_i, toutes dans dans $[-1, 1]$, données par

$$r_i = \cos\frac{(2i+1)\pi}{2(d+1)}, \quad , i = 0, \ldots, d.$$

Ces nombres r_i sont les **points de Chebyshev**.

(f) Montrer que pour tout $x \in [-1, 1]$ on a

$$|T_d(x)| \leq 1.$$

On pourra utiliser que pour tout $x \in |-1, 1]$, il existe $\theta \in \mathbb{R}$ tel que $x = \cos(\theta)$.

34 Interpolation aux points de Chebyshev ← 33 . On pose $A = \{r_0, \ldots, r_d\}$.

(a) Quel est le lien entre $T_{d+1}(x)$ et $(x - r_0)(x - r_1) \cdots (x - r_d)$?

(b) Calculer les nombres Δ_i introduits dans la partie précédente. On pourra procéder comme suit : i) On montrera d'abord que

$$1/\Delta_i = \frac{1}{2^d} T'_{d+1}(r_i),$$

ii) puis on montrera

$$T_{d+1}(x) = \cos((d+1)\arccos x)$$

et on utilisera cette relation pour calculer $T'_{d+1}(r_i)$. On rappelle que

$$\arccos'(x) = \frac{-1}{\sqrt{1-x^2}}.$$

(c) En déduire la formule de Lagrange barycentrique pour les points de Chebyshev.

(d) Montrer en appliquant un théorème du cours et les résultats précédents que si f est une fonction $d+1$ continûment dérivable sur $[-1,1]$ alors pour tout $x \in [-1,1]$, on a

$$|f(x) - \mathbf{L}[r_0,\dots,r_d;f](x)| \leq \frac{1}{2^d \cdot (d+1)!} \max_{t\in[-1,1]} |f^{(d+1)}(t)|.$$

35 Une majoration de l'erreur entre la fonction interpolée et le polynôme d'interpolation.

Dans cette partie on considère un ensemble $A = \{a_0, a_1, a_2, a_3, a_4\} \subset [a,b]$ avec $a_0 < a_1 < a_2 < a_3 < a_4$. On définit $h_0 = a_0 - a$ puis pour $i = 1,2,3,4$, $h_i = a_i - a_{i-1}$ et enfin $h_5 = b - a_4$. On va majorer la valeur absolue du polynôme w_A défini par la relation

$$w_A(x) = (x-a_0)(x-a_1)(x-a_2)(x-a_3)(x-a_4).$$

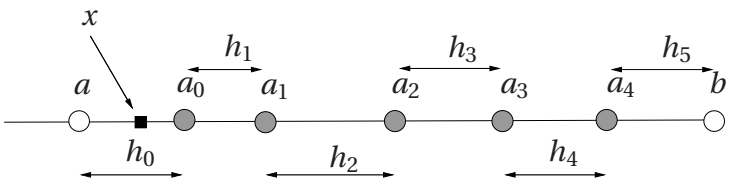
CAS OÙ x EST COMPRIS ENTRE a ET a_0.

A) On suppose que x est compris entre a et a_0 comme dans la figure ci-dessous. Montrer que

$$|w_A(x)| \leq h_0 \times (h_0+h_1) \times (h_0+h_1+h_2) \times (h_0+h_1+h_2+h_3) \times (h_0+h_1+h_2+h_3+h_4).$$

En déduire que

$$|w_A(x)| \leq 5! h^5 \quad \text{avec} \quad h = \max_{0\leq i\leq 5} h_i.$$



CAS OÙ x EST DANS UN INTERVALLE $[a_i, a_{i+1}[$.

B) On suppose maintenant que $x \in]a_0, a_1[$. Montrer, après avoir dessiné la figure correspondante et en utilisant la même idée que dans la question précédente que

$$|w_A(x)| \leq h_1 \times h_1 \times (h_1+h_2) \times (h_1+h_2+h_3) \times (h_1+h_2+h_3+h_4)$$

et en déduire que

$$|w_A(x)| \leq 4! h^5.$$

C) Démontrer en suivant les mêmes idées les majorations suivantes

(a) si $x \in]a_1, a_2[$ alors $|w_A(x)| \leq 2! \cdot 3! \cdot h^5$,

(b) si $x \in]a_2, a_3[$ alors $|w_A(x)| \leq 3! \cdot 2! \cdot h^5$,

(c) si $x \in]a_3, a_4[$ alors $|w_A(x)| \leq 1! \cdot 4! \cdot h^5$,

(d) si $x \in]a_4, b[$ alors $|w_A(x)| \leq 5! \cdot h^5$.

D) Déduire des résultats précédents que

$$\max_{x \in [a,b]} |w_A(x)| \leq 5! h^5$$

puis que pour toute fonction f de classe C^5 sur $[a, b]$ et tout $x \in [a, b]$, on a

$$|f(x) - \mathbf{L}[a_0, a_1, a_2, a_3, a_4; f](x)| \leq \max_{x \in [a,b]} |f^{(5)}| \cdot h^5.$$

CAS OÙ OÙ $a_0 = a$ ET $a_4 = b$.

Dans cette partie, on améliore l'inégalité précédente dans le cas où $a_0 = a$ et $a_4 = b$.

E) Soit $i \in \{0, \ldots, 3\}$. Montrer en étudiant la fonction $x \to (x - a_i)(x - a_{i+1})$ que

$$\max_{x \in [a_i, a_{i+1}]} |(x - a_i)(x - a_{i+1})| = \frac{h_{i+1}^2}{4}.$$

F) Soit $x \in [a_0, a_1]$. Montrer que

$$|w_A(x)| \leq \frac{h_1^2}{4} \times (h_1 + h_2) \times (h_1 + h_2 + h_3) \times (h_1 + h_2 + h_3 + h_4) \leq 4! \frac{h^5}{4}.$$

G) Montrer plus généralement, en considérant les intervalles $[a_1, a_2]$, $[a_2, a_3]$ et $[a_3, a_4]$ que

$$\max_{x \in [a,b]} |w_A(x)| \leq 4! \frac{h^5}{4}$$

et en déduire une nouvelle majoration pour $|f(x) - \mathbf{L}[a_0, a_1, a_2, a_3, a_4; f](x)|$.

H) Expliquer brièvement comment les résultats obtenus se généralisent au cas où

$$A = \{a_0, \ldots, a_n\} \subset [a, b] \quad a_i < a_{i+1}, \quad i = 0, \ldots, n-1.$$

(Sol. 6 p. 98)

36 Interpolation des fonctions de dérivées à croissance lente. Soient $I = [-1, 1]$ et $A = \{a_i : i \in \mathbb{N}\}$ une suite de points deux à deux distincts dans I. On note $A^d = \{a_0, \ldots, a_d\}$. Nous étudions une condition sur la fonction f pour que, pour tout $x \in I$, on ait

$$\lim_{d \to \infty} \mathbf{L}[A^d; f](x) = f(x).$$

A) Montrer que si f est $d+1$ fois continûment dérivable sur I et si $x \in I$ alors

$$|f(x) - \mathbf{L}[A^d; f](x)| \leq \frac{2^{d+1}}{(d+1)!} \max_{t\in[-1,1]} |f^{(d+1)}(t)|.$$

On note $\mathscr{G}$ l'ensemble des fonctions f indéfiniment dérivables sur I pour lesquelles il existe des nombres $M = M_f$ et $r = r_f$ tels que

$$|f^{(d)}(x)| \leq M \cdot r^d, \quad d \in \mathbb{N}, \quad x \in I.$$

B) Parmi les fonctions suivantes, lesquelles sont dans l'ensemble $\mathscr{G}$: $f_1(x) = \sin(\alpha x)$, $f_2(x) = \cos(\alpha x)$, $f_3(x) = \exp(\alpha x)$, $f_4(x) = \exp(\exp \alpha x)$, $f_5(x) = 1/(x+a)$ où $\alpha \in \mathbb{R}$ et $a > 1$?

C) Montrer les propriétés suivantes : (i) Si $f \in \mathscr{G}$ alors $f' \in \mathscr{G}$ (on déterminera $M_{f'}$ et $r_{f'}$ en fonction de M_f et r_f). (ii) Montrer que si f et g sont deux éléments de $\mathscr{G}$ et $\lambda \in \mathbb{R}$ alors $f + \lambda g \in \mathscr{G}$ (ceci signifie que $\mathscr{G}$ est un espace vectoriel). On déterminera $M_{f+\lambda g}$ et $r_{f+\lambda g}$ en fonction de $M_f, M_g, r_f, r_g, \lambda$.

D) Démontrer, par récurrence sur d la **formule de Leibniz** sur les dérivées d'un produit de deux fonctions (indéfiniment dérivables) f et g :

$$(f \cdot g)^{(d)} = \sum_{j=0}^{d} \binom{d}{j} f^{(j)} \cdot g^{(d-j)}$$

où $\binom{d}{j}$ désigne le coefficient binomial – aussi noté C_j^d – défini par

$$\binom{d}{j} = \frac{d!}{j!(d-j)!}.$$

E) Démontrer que si f et g sont deux éléments de $\mathscr{G}$ alors $f \cdot g$ est aussi un élément de $\mathscr{G}$. On déterminera $M_{f.g}$ et $r_{f.g}$ en fonction de M_f, M_g, r_f, r_g.

F) Démontrer que si $f \in \mathscr{G}$ alors pour tout $x \in I$, on a

$$\lim_{d\to\infty} \mathbf{L}[A^d; f](x) = f(x).$$

COMMENTAIRE. — * Pour une fonction $f \in \mathscr{G}$, nous avons

$$\max_{x\in[-1,1]} |f(x) - \mathbf{L}[A^d; f](x)| \leq M_f \frac{(2r)^{d+1}}{(d+1)!}.$$

Or pour tout $v \in \mathbb{R}$, la suite $v^{d+1}/(d+1)!$ converge vers 0 lorsque $d \to \infty$. Cela implique que si $\Delta > 1$ alors

$$\lim_{d\to\infty} \Delta^{d+1} \max_{x\in[-1,1]} |f(x) - \mathbf{L}[A^d; f](x)| = 0 \tag{5.2}$$

puisque

$$\Delta^{d+1} \max_{x\in[-1,1]} |f(x) - \mathbf{L}[A^d; f](x)| \leq M_f \frac{(2\Delta r)^{d+1}}{(d+1)!} \overset{d\to\infty}{\longrightarrow} 0.$$

La relation (5.2) signifie que l'erreur entre f et son polynôme d'interpolation converge uniformément vers 0 plus vite que n'importe quelle suite géométrique de raison moindre que 1.*

37 Effet d'une composition par une bijection affine sur les polylignes. Etudier les propriétés démontrées à l'exercice 27 dans le cas des interpolants polylignes $\mathbf{PL}[\mathbf{s}; f]$).

38 Propriétés générales des polylignes. Les implications suivantes sont-elles vraies ? Les fonctions sont considérées sur un intervalle $[a,b]$ et $\mathbf{s}$ désigne une subdivision quelconque de cet intervalle.

(a) Si f croissante (resp. décroissante) sur $[a,b]$ alors $\mathbf{PL}[\mathbf{s}; f]$ est croissante (resp. décroissante) sur $[a,b]$.

(b) Si $\mathbf{PL}[\mathbf{s}; f]$ croissante (resp. décroissante) sur $[a,b]$ alors f est croissante (resp. décroissante) sur $[a,b]$.

(c) Si f est convexe (resp. concave) sur $[a,b]$ alors $\mathbf{PL}[\mathbf{s}; f]$ est convexe (resp. concave) sur $[a,b]$.

(d) Si $\mathbf{PL}[\mathbf{s}; f]$ convexe (resp. concave) sur $[a,b]$ alors f est convexe (resp. concave) sur $[a,b]$.

39 Erreur entre polyligne et fonction interpolée de classe C^2. Montrer que si f est une fonction de classe C^2 sur l'intervalle $[a,b]$ et σ est une subdivison de $[a,b]$ d'écart h alors pour tout $x \in [a,b]$ on a

$$|f(x) - \mathbf{PL}[\sigma; f](x)| \leq \frac{h^2}{8} \cdot \max_{t\in[a,b]} |f^{(2)}(t)|.$$

Quelle inégalité obtient-on dans le cas où $f(x) = \sin x$?

COMMENTAIRE. — ∗ Supposons que $\sigma^d = (a, a_1^d, \ldots, a_{d-1}^d, b)$ soit la subdivision de $[a,b]$ formée des points équidistants. Dans ce cas, si

$$\mu = (a_j^d + a_{j+1}^d)/2$$

alors la définition des fonctions polylignes et le théorème 8 sur l'erreur la fonction et son polynôme d'interpolation dans le cas $d = 1$ donnent

$$|f(x)_{\mathbf{PL}}[\sigma^d; f](\mu)| = |f(x) - \mathbf{L}[a_j^d, a_{j+1}^d; f](\mu)| = \frac{|f^{(2)}(\xi)|}{2}|\mu - a_j^d||\mu - a_{j+1}^d| = \frac{|f^{(2)}(\xi)|}{d^2},$$

où ξ est un certain réel compris entre a_j^d et a_{j+1}^d. Il suit que

$$d^2 \max_{x\in[a,b]} |f(x) - \mathbf{PL}[\sigma^d; f](x)| \geq m_2,$$

où $m_2 = \inf_{x\in[a,b]} |f^{(2)(x)}|$. En particulier, l'inégalité montre que lorsque f est une fonction pour m_2 ne s'annulle pas alors l'erreur entre la fonction interpolée et la fonction polyligne ne saurait décroître plus vite que la suite $1/d^2$. ∗

40 Un exemple. On veut approcher la fonction $f(x) = 1/(1+x^2)$ sur $[-1,1]$ par une polyligne. Comment faut-il choisir la subdivision si l'erreur doit être moindre que 10^{-2} ?

41 Ajout d'un point à une subdivision Soit $\sigma = (a = a_0, a_1, \ldots, a_{d-1}, a_d = b)$ une subdivision de longueur d de l'intervalle $[a,b]$. On complète cette subdivision par deux points $a_{-1} < a$ et $a_{d+1} > b$ et on définit les fonctions b_i^σ, $i = 0, \ldots, d$ comme dans le cours par le graphe de la figure 4.4.

(a) Trouvez la relation entre deux fonctions b_i et b_j, $i, j \in \{0, \dots, d\}$ lorsque la subdivision σ est déterminée par les points équidistants.

(b) On rajoute un point a^+ à la subdivision σ pour obtenir une subdivision σ_+ de longueur $d+1$. Comment calculer les fonctions $b_i^{\sigma_+}$ à l'aide des fonctions b_i^{σ}.

42 Interpolation et calcul approché des dérivées. On étudie une méthode de calcul approché des dérivées des fonctions à partir des valeurs d'une fonction.

Soit f une fonction dérivable sur un intervalle fermé borné $I = [a, b]$ et $X = \{x_0, \dots, x_d\}$ un ensemble de $d+1$ point deux à deux distincts dans I. Etant donnée $y \in I$, on cherche une formule $Q_y(f)$ de la forme

$$Q_y(f) = A_0 f(x_0) + A_1 f(x_1) + \cdots + A_d f(x_d) \tag{5.3}$$

où les A_i sont des nombres réels indépendant de f, telle que

$$f'(y) \approx Q_y(f). \tag{5.4}$$

(a) Montrer que l'application Q_y définie ci-dessus est une application linéaire de E dans $\mathbb{R}$ où E désigne l'espace vectoriel des fonctions dérivables sur I.

(b) Dans cette partie, on cherche à déterminer les nombres A_i de sorte (5.4) se réduise à une égalité lorsque f est un polynôme de degré inférieur ou égal à d, autrement dit,

$$Q_y(p) = p'(y) \quad \text{pour tout polynôme } p \text{ de degré} \leq d. \tag{5.5}$$

Pour tout $i = 0, \dots d$, on note ℓ_i le polynôme fondamental de Lagrange pour $X = \{x_0, \dots, x_d\}$ correspondant au point x_i,

$$\ell_i(x) = \prod_{j=0, j\neq i}^{d} \frac{x - x_j}{x_i - x_j}. \tag{5.6}$$

(c) Montrer que la condition (5.5) est satisfaite *si et seulement si* pour tout $i = 0, \dots, d$, A_i est la dérivée de ℓ_i en y, autrement dit

$$A_i = \ell_i'(y). \tag{5.7}$$

(d) En déduire que la condition (5.5) est satisfaite *si et seulement si*

$$Q_y(f) = (\mathbf{L}[X, f])'(y) \tag{5.8}$$

où $\mathbf{L}[X, f]$ désigne le polynôme d'interpolation de Lagrange de f par rapport aux points de X.
Nota Bene : Dans la suite du problème on suppose que l'égalité (5.8) est satisfaite.

(e) Donner l'expression de $Q_0(f)$ lorsque $X = \{-1, 0, 1\}$.

(f) On suppose maintenant que f est $d+1$ dérivable sur I et que $y = x_i \in X$. On cherche une estimation de l'erreur $|f'(y) - Q_y(f)|$. On note $\epsilon(x) = f(x) - \mathbf{L}[X, f](x)$ et $w(x) = (x - x_0)(x - x_1) \dots (x - x_d)$

(g) Montrer que $w'(x_i) = \prod_{j=1, j\neq i}^{d} (x_i - x_j)$.

(h) On admet que pour tout $x \in I$ on a $\epsilon(x) = w(x)g(x)$ où g est une fonction dérivable telle que $g(x) = (1/(d+1)!)f^{(d+1)}(\xi_x)$. Montrer que

$$f'(x_i) - Q_{x_i}(f) = \frac{1}{(d+1)!} \prod_{j=1, j\neq i}^{d} (x_i - x_j) f^{d+1}(\xi_{x_i}). \tag{5.9}$$

En déduire une majoration de l'erreur $|f'(x_i) - Q_{x_i}(f)|$.

II

CALCUL APPROCHÉ DES INTÉGRALES

§ 1 FORMULES DE QUADRATURES ÉLÉMENTAIRES

1.1 Problème

Soit f une fonction continue sur un intervalle $[a,b]$. Nous voulons calculer l'intégrale $\int_a^b f(x)dx$. C'est un des calculs parmi les plus communs dans les applications des mathématiques. Le théorème fondamental du calcul intégral nous dit que $\int_a^b f(x)dx = F(b)-F(a)$ où F est une primitive de f. Pour appliquer ce résultat, nous disposons de divers outils théoriques dont les plus fondamentaux sont le théorème de changement de variable et le théorème d'intégration par partie. Cependant, il n'est possible de déterminer explicitement une primitive F que pour une classe relativement restreinte de fonctions f et, lorsque cette détermination est à notre disposition, l'expression de F est souvent si compliquée que l'évaluation de $F(b)-F(a)$ nécessite l'emploi d'un processus d'approximation. Dans ce cas, il est encore plus naturel et généralement moins coûteux de chercher directement une approximation de l'intégrale.

1.2 Présentation générale

L'idée consiste à utiliser une approximation $\int_a^b f(x)dx \approx \int_a^b g(x)dx$ où g est une fonction qui, d'une part, est proche de f et, d'autre part, possède des primitives aisément calculables. Le choix le plus naturel est celui du polynôme d'interpolation de

Lagrange,

$$g = \mathbf{L}[x_0, \dots, x_d; f],$$

où $A = \{x_0, \dots, x_d\} \subset [a,b]$ car les polynômes d'interpolation sont proches de la fonction qu'ils interpolent et, étant des polynômes, il est raisonnable d'espérer que leurs primitives seront facilement calculables. Nous appelons **formule de quadrature** (élémentaire) d'**ordre** d, toute expression

$$\mathtt{Q}(f) = \int_a^b \mathbf{L}[x_0, \dots, x_d; f](x)\,dx = \sum_{i=0}^{d} f(x_i) \int_a^b \ell_i(x)\,dx \tag{1.1}$$

où ℓ_i est le polynôme fondamental de Lagrange correspondant au point a_i, voir. L'application Q ainsi définie est une forme linéaire sur $\mathsf{C}[a,b]$, autrement dit, elle vérifie

$$\mathtt{Q}(\lambda_1 f_1 + \lambda_2 f_2) = \lambda_1 \mathtt{Q}(f_1) + \lambda_2 \mathtt{Q}(f_2) \quad \lambda_1, \lambda_2 \in \mathbb{R}, \quad f_2, f_2 \in \mathbb{C}[a,b]).$$

Pour savoir si $\mathtt{Q}(f)$ est effectivement proche de $\int_a^b f(x)\,dx$, nous devrons étudier l'erreur

$$\mathsf{E}^{\mathtt{Q}}(f) := \left| \int_a^b f(x)\,dx - \mathtt{Q}(f) \right|. \tag{1.2}$$

Remarquons que si $\mathtt{Q}$ est une formule de quadrature d'ordre d alors pour tout $p \in \mathscr{P}_d$ on a $\int_a^b p(x)\,dx = \mathtt{Q}(p)$. En effet,

$$\begin{aligned} p \in \mathscr{P}_d \quad &\Rightarrow \quad p = \mathbf{L}[x_0, \dots, x_d; p] \\ &\Rightarrow \quad \mathtt{Q}(p) \overset{def}{=} \int_a^b \mathbf{L}[x_0, \dots, x_d; p](x)\,dx = \int_a^b p(x)\,dx. \end{aligned}$$

Nous verrons que dans certains cas l'égalité ci-dessus peut continuer à être vérifiée pour des polynômes de degré plus grand que d.

Une réciproque est vraie.

Théorème 1. *Si $R(f)$ est une expression de la forme $R(f) = \sum_{i=0}^{d} \lambda_i f(a_i)$ telle que $R(p) = \int_a^b p(x)\,dx$ pour tout $p \in \mathscr{P}_d$ alors $\lambda_i = \int_a^b \ell_i(x)\,dx$ où ℓ_i est le polynôme fondamental de Lagrange correspondant à $x_i \in \{x_0, \dots, x_d\}$.*

Démonstration. Il suffit d'utiliser la relation $\sum_{i=0}^{d} \lambda_i p(a_i) = \int_a^b p(x)\,dx$ avec $p = \ell_j$. En effet, puisque $\ell_j(a_i) = 0$ sauf lorsque $i = j$ pour lequel nous avons $\ell_j(a_j) = 1$, on a $\sum_{i=0}^{d} \lambda_i \ell_j(a_i) = \lambda_j$. ∎

Remarquons que puisque R est une forme linéaire, pour s'assurer que

$$R(p) = \int_a^b p(x)\,dx, \quad \text{pour tout } p \in \mathscr{P}_d,$$

il suffit de vérifier l'identité lorsque p parcourt une base de $\mathscr{P}^d$. En particulier, si $M_i(x) = x^i$ il suffit de vérifier que $R(M_i) = \int_a^b x^i\,dx$ pour $i = 0, 1, \ldots, d$.

E 43 * On cherche une approximation de $\int_{-1}^{1} f(x)\,dx$ par une formule du type

$$\int_{-1}^{1} f(x)\,dx \approx f(t_1) + f(t_2)$$

de telle sorte que la formule soit *exacte* pour tous les polynômes de degré inférieur ou égal à 2. Montrer qu'il existe une et une seule paire $\{t_1, t_2\}$ satisfaisant la propriété demandée et la déterminer.

Dans la pratique, grâce au procédé de composition, des résultats très précis sont souvent obtenus en employant seulement des méthodes d'ordre $d \leq 2$. Nous étudierons en détail trois de ces méthodes : la **méthode du point milieu** ($d = 0$), la **méthode des trapèzes** ($d = 1$) et la **méthode de Simpson** ($d = 2$). D'autres exemples sont proposés en exercice.

§ 2 EXEMPLES FONDAMENTAUX

2.1 La formule du point milieu

Nous utilisons un polynôme d'interpolation de degré $d = 0$ avec le point $x_0 = \frac{a+b}{2}$. Dans ce cas, $\mathbf{L}[x_0; f](x) = f(\frac{a+b}{2})$ et l'approximation

$$\int_a^b f(x)\,dx \approx \int_a^b \mathbf{L}[x_0; f](x)\,dx \quad \text{devient} \quad \int_a^b f(x)\,dx \approx (b-a) f(\frac{a+b}{2}). \tag{2.1}$$

L'expression $\mathtt{Q}(f) = (b-a) f(\frac{a+b}{2})$ s'appelle la **formule du point milieu**. Lorsque $f(c) > 0$, $\mathtt{Q}(f)$ est l'aire du rectangle de sommets les points de coordonnées $(a, 0)$, $(b, 0)$, $(a, f(c))$ et $(b, f(c))$, voir la figure 1 A).

2.2 La formule du trapèze

Soit $f \in \mathrm{C}[a, b]$. Nous prenons $d = 1$ et $A = \{a, b\}$. L'approximation

$$\int_a^b f(x)\,dx \approx \int_a^b \mathbf{L}[a, b; f](x)\,dx \quad \text{devient} \quad \int_a^b f(x)\,dx \approx \frac{(b-a)}{2}(f(a) + f(b)). \tag{2.2}$$

*. [Démidovitch & Maron 1979]

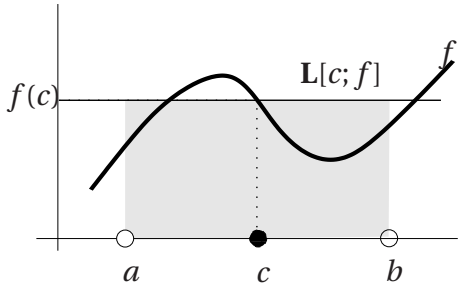
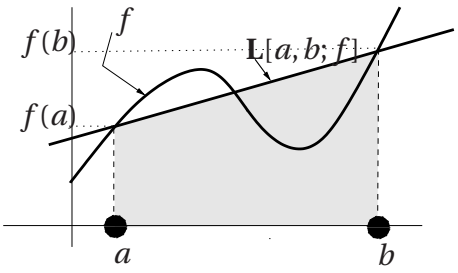
A) POINT MILIEU	B) TRAPÈZE
	
Nous avons posé $c = (a+b)/2$. L'aire de la partie grisée est égale à $\mathtt{Q}(f)$.	L'aire de la partie grisée est égale à $\mathtt{Q}(f)$.

TABLE 1 – Méthode du point milieu et du trapèze.

En effet, $\mathbf{L}[a,b;f](x) = f(a) + \frac{f(b)-f(a)}{b-a}(x-a)$, voir (I.1.7), d'où

$$\begin{aligned}
\int_a^b \mathbf{L}[a,b;f](x)\,dx &= \int_a^b f(a) + \left\{\frac{f(b)-f(a)}{b-a}(x-a)\right\} dx \\
&= f(a)(b-a) + \frac{f(b)-f(a)}{b-a}\int_a^b (x-a)\,dx \\
&= f(a)(b-a) + \frac{f(b)-f(a)}{b-a}\left[\frac{(x-a)^2}{2}\right]_a^b \\
&= f(a)(b-a) + \frac{f(b)-f(a)}{b-a}\cdot\frac{(b-a)^2}{2} \\
&= \frac{(b-a)}{2}\cdot[f(a)+f(b)].
\end{aligned}$$

L'expression $\mathtt{Q}(f) = \frac{(b-a)}{2}(f(a)+f(b))$ s'appelle la **formule du trapèze**. Lorsque $f(a)$ et $f(b)$ sont positifs, elle n'est autre que l'aire du trapèze de sommets les points de coordonnées $(a,0)$, $(b,0)$, $(a,f(a))$ et $(b,f(b))$ comme illustré par la figure 1 B).

2.3 La formule de Simpson

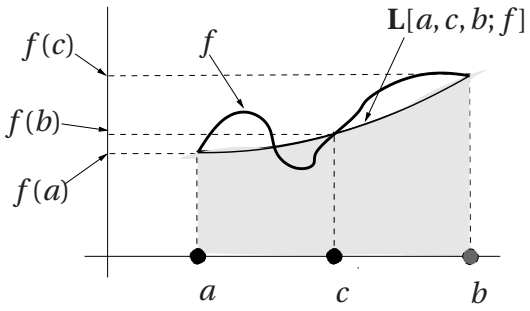
Nous prenons cette fois $d=2$ et $A=\{a,c,b\}$ où $c=\frac{a+b}{2}$. L'approximation

$$\int_a^b f(x)\,dx \approx \int_a^b \mathbf{L}[a,c,b;f](x)\,dx$$

devient

$$\int_a^b f(x)\,dx \approx \frac{(b-a)}{6}\big(f(a)+4f(c)+f(b)\big). \tag{2.3}$$

Le calcul est proposé à l'exercice I.26. L'expression $\mathtt{Q}(f) = \frac{(b-a)}{6}\big(f(a)+4f(c)+f(b)\big)$ s'appelle la **formule de Simpson**.



L'aire de la partie grisée est égale à $\mathtt{Q}(f)$.

FIGURE 1 – Méthode de Simpson.

§ 3 ETUDE DE L'ERREUR

3.1 Estimation de l'erreur dans la formule du point milieu

Théorème 2. *Soit $f \in C^2([a,b])$. Il existe $\xi \in [a,b]$ tel que*

$$\int_a^b f(t)\,dt - (b-a)f\left(\frac{a+b}{2}\right) = \frac{(b-a)^3}{24}.f^{(2)}(\xi).$$

En particulier,

$$\left|\int_a^b f(t)\,dt - (b-a)f\left(\frac{a+b}{2}\right)\right| \leq \frac{(b-a)^3}{24}.\max_{x\in[a,b]}|f^{(2)}(x)|.$$

Comme de nombreux résultats d'analyse numérique,la démonstration de ce théorème est basée sur la **formule de Taylor**. Nous la rappelons dans un cadre suffisamment général pour pouvoir servir dans la suite du cours.

Théorème 3 (Formule de Taylor)**.** *Soit f une fonction continue sur $[\alpha,\beta]$ et $d+1$ fois dérivable sur $]\alpha,\beta[$. Si u_0 et v sont dans $[\alpha,\beta]$ alors il existe $\xi \in]\alpha,\beta[$ tel que*

$$f(v) = f(u_0) + f'(u_0)(v-u_0) + \cdots + \frac{f^{(d)}(u_0)}{d!}(v-u_0)^d + \frac{f^{(d+1)}(\xi)}{d!}(v-u_0)^{d+1}. \tag{3.1}$$

[TH 3]

Cette égalité s'appelle la formule de Taylor de f en u_0 à l'ordre d.

Le polynôme $T^d(f,\cdot) = f(u_0) + f'(u_0)(\cdot - u_0) + \cdots + \frac{f^{(d)}(u_0)}{d!}(\cdot - u_0)^d$ s'appelle le **polynôme de Taylor** de f en u_0 à l'ordre d. Le vocabulaire ici est assez malheureux, il faudrait parler de *degré* plutôt que d'*ordre*.

Dans ce cours nous appliquerons toujours ce théorème avec une fonction f de classe C^{d+1} sur un intervalle contenant α et β de sorte que les conditions du théorème seront largement satisfaites.

Démonstration du Théorème 2. Posons $c = \frac{a+b}{2}$. Pour tout $x \in [a,b]$, une application de la formule de Taylor ci-dessus avec $u_0 = c$ donne l'existence de ξ_x tel que

$$f(x) = f(c) + f'(c)(x-c) + \frac{f''(\xi)}{2!}(x-c)^2.$$

D'où nous tirons l'inégalité

$$f(c) + f'(c)(x-c) + \frac{m_2}{2}(x-c)^2 \le f(x) \le f(c) + f'(c)(x-c) + \frac{M_2}{2}(x-c)^2$$

où $m_2 = \inf[a,b] f''$ et $M_2 = m_2 = \max[a,b] f''$. En intégrant la première inégalité, il vient

$$\begin{aligned} \int_a^b f(x)dx &\ge \int_a^b (f(c) + f'(c)(x-c) + m_2(x-c)^2)dx \\ &\ge \mathtt{Q}(f) + \tfrac{1}{2}[(x-c)^2]_a^b + \tfrac{m_2}{2\cdot 3}[(x-c)^3]_a^b \\ &\ge \mathtt{Q}(f) + 0 + \tfrac{m_2}{2\cdot 3}\left(\tfrac{(b-a)^3}{2^3}\right) \end{aligned}$$

et il suit

$$m_2 \frac{(b-a)^3}{24} \le \int_a^b f(x)dx - \mathtt{Q}(f).$$

De la même manière, en intégrant la seconde inégalité, nous obtenons

$$\int_a^b f(x)dx - \mathtt{Q}(f) \ge M_2 \frac{(b-a)^3}{24}.$$

Regroupant les deux estimations, nous tirons

$$m_2 \le \frac{24}{(b-a)^3}\left\{\int_a^b f(x)dx - \mathtt{Q}(f)\right\} \le M_2.$$

Maintenant puisque f'' est une fonction continue, d'après le théorème des valeurs intermédiaires, tout nombre compris entre sa plus grande valeur M_2 et sa plus petite valeur m_2 est encore une valeur de f''. Autrement dit, il existe $\theta \in [a,b]$ tel que

$$\int_a^b f(t)dt - (b-a) f\left(\frac{a+b}{2}\right) = \frac{(b-a)^3}{24} f''(\theta). \quad \blacksquare$$

3.2 Estimation de l'erreur dans la formule du trapèze

Théorème 4. *Soit $f \in C^2([a,b])$. Il existe $\theta \in [a,b]$ tel que*

$$\int_a^b f(t)\,dt - \frac{(b-a)}{2}\left[f(a)+f(b)\right] = -\frac{(b-a)^3}{12} f^{(2)}(\theta).$$

En particulier,

$$\left|\int_a^b f(t)\,dt - \frac{(b-a)}{2}\left[f(a)+f(b)\right]\right| \leq \frac{(b-a)^3}{12}.\max_{[a,b]}|f^{(2)}|.$$

Démonstration. Nous devons estimer $\int_a^b \{f(x)-\mathbf{L}[a,b;f](x)\}dx$. Nous commençons par obtenir une estimation du terme sous l'intégrale. D'après le Théorème I.8, pour tout $x \in [a,b]$, il existe $\xi_x \in [a,b]$ tel que

$$f(x)-\mathbf{L}[a,b;f](x) = \frac{f^{(2)}(\xi_x)}{2}(x-a)(x-b).$$

Puisque la fonction $x \to (x-a)(x-b)$ est négative ou nulle sur $[a,b]$, nous avons

$$\frac{M_2}{2}(x-a)(x-b) \leq f(x)-\mathbf{L}[a,b;f](x) \leq \frac{m_2}{2}(x-a)(x-b),$$

où $m_2 = \min_{[a,b]} f^{(2)}$ et $M_2 = \max_{[a,b]} f^{(2)}$. Nous intégrons ces deux inégalités en utilisant le résultat suivant

$$\int_a^b (x-a)(x-b)\,dx = -(b-a)^3/6, \tag{3.2}$$

qui se vérifie immédiatement pour obtenir

$$-M_2(b-a)^3/12 \leq \int_a^b \{f(x)-\mathbf{L}[a,b;f](x)\}dx \leq -m_2(b-a)^3/12.$$

En raisonnant comme dans la demonstration du Théorème 2, nous déduisons qu'il existe θ tel que

$$\int_a^b \{f(x)-\mathbf{L}[a,b;f](x)\}dx = -\frac{f(\theta)}{12}(b-a)^3.$$

■

E 44 Démontrer la relation (3.2).

3.3 Estimation de l'erreur dans la formule de Simpson

Nous ne connaissons pas de démonstration simple du résultat suivant. Une démonstration élémentaire mais peu naturelle sera proposée à l'exercice 56.

Théorème 5 (‡)**.** *Soit $f \in C^4([a,b])$. On a*

$$\left| \int_a^b f(t)\,dt - \frac{(b-a)}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right] \right| \leq \frac{(b-a)^5}{2880} \cdot \max_{[a,b]} |f^{(4)}|.$$

Les point important de cette estimation est qu'elle fait intervenir la dérivée 4-ième de la fonction (et un facteur $(b-a)^5$). La constante 2880 naturellement est purement anecdotique.

E 45 D'après sa construction la formule de Simpson est d'ordre 2 mais l'estimation ci-dessus montre qu'elle est en réalité d'ordre 3. Donner une démonstration directe de cette propriété.

§ 4 COMPOSITION

4.1 Idée générale

Nous savons que le polynôme $\mathbf{L}[x_0,\dots,x_d;f]$ a d'autant plus de chance d'être proche de la fonction interpolée f que l'intervalle $[a,b]$ est petit et les formules d'erreur données dans la partie précédente confirment l'intuition que plus l'intervalle $[a,b]$ sera petit plus l'approximation sera précise. Dans ces conditions, il est naturel de découper l'intervalle de départ en une famille de sous-intervalles beaucoup plus petits et d'appliquer les formules de quadrature à ces petits intervalles avant de regrouper les approximations obtenues. De manière précise, choisissons une subdivision $\sigma = (a = a_0, a_1, \dots, a_n = b)$ de $[a,b]$ et, dans chaque intervalle $[a_i, a_{i+1}]$, $d+1$ points distincts $X^i = \{x_0^i, x_1^i, \dots, x_d^i\}$ pour construire la formule d'approximation

$$\int_{a_i}^{a_{i+1}} f(x)\,dx \approx \mathtt{Q}_{[a_i,a_{i+1}]}(f) \tag{4.1}$$

avec

$$\mathtt{Q}_{[a_i,a_{i+1}]}(f) = \sum_{i=0}^{d} f(x_i^k) \int_{a_k}^{a_{k+1}} \ell_i^k(x)\,dx \quad \text{et} \quad \ell_i^k(x) = \prod_{j=0, j\neq i}^{d} \frac{x - x_j^k}{x_i^k - x_j^k}. \tag{4.2}$$

La localisation des points x_j^i dans chaque sous-intervalle est illustrée dans la figure 2 La relation de Chasle pour les intégrales nous donne

$$\int_a^b f(x)\,dx = \sum_{k=0}^{n-1} \int_{a_k}^{a_{k+1}} f(x)\,dx,$$

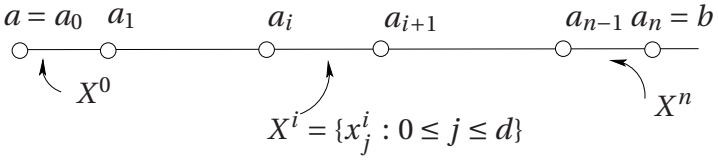


FIGURE 2 – Localisation des points dans une formule de quadrature composée

de sorte que pour approximer l'intégrale globale il suffit d'approximer les n termes de la somme

$$\int_a^b f(x)dx \approx \sum_{k=0}^{n-1}\sum_{i=0}^{d} \mathtt{Q}_{[a_k,a_{k+1}]}(f) \tag{4.3}$$

Toute expression $\mathtt{Qc}$ de la forme

$$\mathtt{Qc}(f) = \sum_{k=0}^{n-1}\sum_{i=0}^{d} \mathtt{Q}_{[a_k,a_{k+1}]}(f)$$

s'appelle une **formule de quadrature composée** d'**ordre** d. L'application $\mathtt{Qc}$ définit une forme linéaire sur $\mathsf{C}[a,b]$. L'erreur $|\int_a^b f(x)dx - \mathtt{Qc}(f)|$ est notée $\mathsf{E}^{\mathtt{Qc}}(f)$.

Théorème 6 (Principe d'addition des erreurs).

$$\mathsf{E}^{\mathtt{Qc}}(f) \le \sum_{i=0}^{n-1} \mathsf{E}^{\mathtt{Q}}_{[a_i,a_{i+1}]}(f). \tag{4.4}$$

Démonstration. Avec les notations précédentes, nous avons

$$\begin{aligned}
\mathsf{E}^{\mathtt{Qc}}(f) &= \Big|\sum_{i=0}^{n-1}\int_{a_i}^{a_{i+1}} f(x)dx - \sum_{i=0}^{n-1} \mathtt{Q}_{[a_i,a_{i+1}]}(f)\Big| \\
&= \left|\sum_{i=0}^{n-1}\left\{\int_{a_i}^{a_{i+1}} f(x)dx - \mathsf{E}^{\mathtt{Q}}_{[a_i,a_{i+1}]}(f)\right\}\right| \\
&\le \sum_{i=0}^{n-1}\left\{\left|\int_{a_i}^{a_{i+1}} f(x)dx - \mathsf{E}^{\mathtt{Q}}_{[a_i,a_{i+1}]}(f)\right|\right\} = \sum_{i=0}^{n-1} \mathsf{E}^{\mathtt{Q}}_{[a_i,a_{i+1}]}(f). \quad \blacksquare
\end{aligned}$$

§ 5 EXEMPLES FONDAMENTAUX DE FORMULES COMPOSÉES

Soit $n \in \mathbb{N}^\star$, $I = [a,b]$ et $f \in \mathsf{C}[a,b]$. Ecrivons $h(n) = (b-a)/n$ et $a(i,n) = a + ih(n)$. L'application du principe d'addition ci-dessus aux exemples fondamentaux des méthodes du point milieu, du trapèze et de Simpson donne les résultats regroupés dans

la table 2. Lorsque $n \to \infty$, dans les trois cas, l'erreur commise tend toujours vers 0, autrement dit, quelle que soit la précision choisie ϵ, en prenant n suffisamment grand, chacune des méthodes fournira une valeur approchée de l'intégrale avec une erreur moindre que ϵ. Pour connaître une valeur de n assurant la précision ϵ il faut cependant au moins disposer d'un majorant de $\max_{[a,b]} |f^{(2)}|$ pour la méthode des trapèzes ou du point milieu et de $\max_{[a,b]} |f^{(4)}|$ pour la méthode de Simpson.

E 46 Montrer que si $\mathtt{Qc}(f)$ désigne la formule des trapèzes composées avec $n+1$ points équidistants $a(i,n) = a + i(b-a)/n$ alors

$$\mathtt{Qc}(f) = \int_a^b \mathbf{PL}[f,\sigma](x)\,dx,$$

où σ est la subdivision $a = a(0,n) < a(1,n) < \cdots < a(n,n) = b$.

La table 3 montre l'erreur obtenue en utilisant les méthodes pour approcher $\int_0^1 4/(1+x^2)$ qui n'est autre que le nombre π. L'exécution est très rapide. Pour la méthode de Simpson avec $n = 700$, l'algorithme ne demande que 0.125 seconde d'attente.

§ 6 EXERCICES ET PROBLÈMES

47 Une caractérisation de la formule du point milieu. Déterminer tous les points $p \in [a,b]$ tels que l'approximation

$$\int_a^b f(x)\,dx = (b-a)f(p)$$

soit exacte (i.e., soit une égalité) pour tous les polynômes de degré ≤ 1.

48 Une expression de l'erreur dans la formule des trapèzes. Soit f une fonction de classe C^2 dans $[a,b]$. Montrer que

$$\int_a^b f(x)\,dx = \frac{b-a}{2}(f(a)+f(b)) - \int_a^b (x-a)(b-x)\frac{f''(x)}{2}\,dx.$$

On pourra calculer la partie intégrale dans le terme de droite en effectuant une ou plusieurs intégrations par parties.

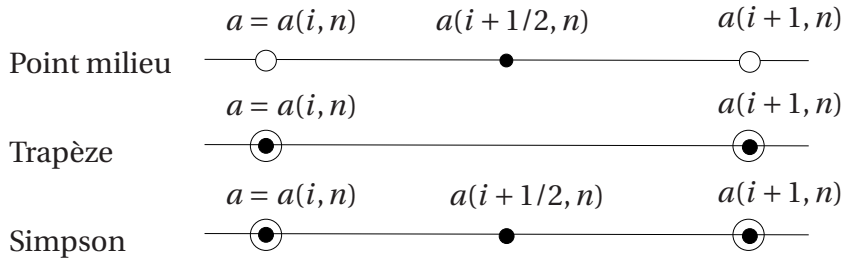
49 Une formule de quadrature avec points intérieurs. Soient $a < b$ et pour $i = 0,\ldots,3$, $x_i = a + i\frac{b-a}{3}$ de sorte que $x_0 = a$ et $x_3 = b$, f désigne une fonction continue sur $[a,b]$. Démontrer que

$$\int_a^b \mathbf{L}[x_1,x_2;f](x)\,dx = \frac{b-a}{2}[f(x_1)+f(x_2)] \tag{6.1}$$

où $\mathbf{L}[x_1,x_2;f]$ désigne le polynôme d'interpolation de f par rapport aux points x_1 et x_2.

PRINCIPALES FORMULES DE QUADRATURES COMPOSÉES

$I = [a,b]$, $h(n) = (b-a)/n$, $a(i,n) = a + ih(n)$, $f \in \mathsf{C}(I)$



	Formule : $\mathtt{Qc}(f)$	Erreur : $\mathsf{E}^{\mathtt{Qc}}(f)$	Type de fonctions
Point milieu	$h(n)\cdot\sum_{i=0}^{n-1} f(a(i+1/2,n))$	$\frac{(b-a)^3}{12n^2}\cdot\max_{[a,b]}\lvert f^{(2)}\rvert$	$f\in\mathsf{C}^2(I)$
Trapèze	$\frac{h(n)}{2}\cdot\left[f(a))+f(b)+2\sum_{i=1}^{n-1} f(a(i,n))\right]$	$\frac{(b-a)^3}{24n^2}\cdot\max_{[a,b]}\lvert f^{(2)}\rvert$	$f\in\mathsf{C}^2(I)$
Simpson	$\frac{h(n)}{6}\left\{f(a)+f(b)+2\sum_{i=1}^{n-1} f(a(i,n))+4\sum_{i=0}^{n-1} f(a(i+1/2,n))\right\}$	$\frac{(b-a)^5}{2880n^4}\cdot\max_{[a,b]}\lvert f^{(4)}\rvert$	$f\in\mathsf{C}^4(I)$

TABLE 2 – Exemples fondamentaux de formules de quadrature composées

n	Point milieu	Trapèze	Simpson
2.	- 0.0207603	0.0415927	0.0000240
4.	- 0.0052079	0.0104162	0.0000002
6.	- 0.0023148	0.0046296	1.328D-08
8.	- 0.0013021	0.0026042	2.365D-09
10.	- 0.0008333	0.0016667	6.200D-10
70.	- 0.0000170	0.0000340	5.329D-15
930.	- 9.635D-08	0.0000002	- 4.441D-16
2300.	- 1.575D-08	3.151D-08	4.441D-16

TABLE 3 – comparasion des diverses méthodes pour $\pi = \int_0^1 4/(1+x^2)\,dx$

50 La seconde formule de Simpson. Soient x_i, $i = 0,\dots,x_3$ les points équidistants de l'intervalle $[a,b]$, $x_i = a + ih$ avec $h = (b-a)/3$, $i = 0,1,2,3$.

(a) Montrer que

$$\int_a^b \mathbf{L}[x_0,x_1,x_2,x_3;f](x)\,dx = \frac{3h}{8}[f(x_0)+3f(x_1)+3f(x_2)+f(x_3)].$$

(b) L'expression

$$\mathtt{Q}(f) = \frac{3h}{8}[f(x_0)+3f(x_1)+3f(x_2)+f(x_3)]$$

s'appelle la **seconde formule de Simpson** ou la **formule de Newton** [Newton 1711].

(c) Illustrer graphiquement l'approximation

$$\int_a^b f(x)\,dx \approx \frac{3h}{8}[f(x_0)+3f(x_1)+3f(x_2)+f(x_3)].$$

(d) Donner la formule de quadrature composée correspondante (avec n sous-intervalles).

51 Un exemple. Soit $I = \int_0^1 x^{1/2}\,dx$.

(a) Donner la valeur exacte de I.

(b) Donner une approximation de I en utilisant la méthode de Simpson avec deux sous-intervalles.

(c) Le théorème du cours permettait-il de prédire l'erreur commise ?

(UPS, L2, 2005)

52 Formule des trapèzes et fonctions convexes. On souhaite calculer une valeur approchée de ln(2) à partir de la relation

$$\ln(2) = \int_1^2 \frac{dx}{x}.$$

Nous considérerons la fonction f définie sur $]0,\infty[$ par $f(x) = \frac{1}{x}$.

(a) Montrer que pour tout $(a,b) \in]0,\infty[\times]0,\infty[$ et pour tout $t \in [0,1]$ on a

$$f(ta+(1-t)b) \leq tf(a)+(1-t)f(b). \tag{6.2}$$

(b) On suppose $0 < a < b < \infty$. Soit $x \in [a,b]$. Montrer que $\frac{b-x}{b-a} \in [0,1]$. Montrer en prenant $t = \frac{b-x}{b-a}$ dans (6.2) que

$$f(x) \leq \mathbf{L}[a,b;f](x).$$

(c) Trouver une approximation de $\int_1^2 \frac{dx}{x}$ en appliquant la méthode des trapèzes combinée avec 2 sous-intervalles. Faire un schéma illustrant le calcul.

(d) Expliquer pourquoi quel que soit le nombre de sous-intervalles, le nombre trouvé par la méthode des trapèzes combinée fournira toujours une approximation par excès (c'est-à-dire supérieure à la valeur exacte ln(2).)

(e) On approche maintenant $\int_1^2 \frac{dx}{x}$ en utilisant la méthode Simpson combinée. Combien de sous-intervalles faut-il utiliser pour commettre une erreur inférieure ou égale à 10^{-10} ?

(UPS, L2, 2003, sol 7 p. 100.)

NOTE. — Voir l'exercice I.11.

53 Un exemple. Estimer, à l'aide des théorèmes du cours, le nombre de sous-intervalles n nécessaire pour obtenir une approximation de

$$I = \int_0^1 \frac{4}{1+x^2} dx$$

avec une erreur moindre que 10^{-6}, en utilisant (a) la méthode du point milieu combinée, (b) la méthode des trapèzes combinée, (c) la méthode de Simpson combinée ? Comparer les estimations trouvées avec les résultats donnés dans le tableau 4 (voir cours).

n	Point milieu	Trapèze	Simpson
2.	- 0.0207603	0.0415927	0.0000240
4.	- 0.0052079	0.0104162	0.0000002
6.	- 0.0023148	0.0046296	1.328D-08
8.	- 0.0013021	0.0026042	2.365D-09
10.	- 0.0008333	0.0016667	6.200D-10
70.	- 0.0000170	0.0000340	5.329D-15
930.	- 9.635D-08	0.0000002	- 4.441D-16
2300.	- 1.575D-08	3.151D-08	4.441D-16

TABLE 4 – Erreur dans l'approximation de $\int_0^1 \frac{4}{1+x^2} dx$ avec les méthodes du point milieu, du trapèze et de Simpson.

54 Sensibilité de la formule des trapèzes composée aux erreurs sur les valeurs de la fonction. On considère la fonction f définie sur $\mathbb{R}$ par $f(x) = \exp(x^2)$. On souhaite calculer une valeur approchée de

$$I = \int_0^1 f(x)dx$$

par la méthode de Simpson combinée.

On note $A(n, f)$ la valeur approchée fournie par la méthode de Simpson combinée avec n sous-intervalles.

(a) Déterminer une valeur de n aussi petite que possible assurant $\left|I - A(n, f)\right| \leq 10^{-3}$. On notera ν la valeur de n trouvée.

Le calcul de $A(n, f)$ nécessite l'utilisation d'un certain nombre de valeurs de la fonction f. Or on ne peut disposer que d'une approximation de cette fonction, une approximation donnée, disons, par la fonction $\tilde{f}$. Il est donc impossible de calculer exactement $A(n, f)$: on ne peut disposer que de $A(n, \tilde{f})$.

(b) On suppose que pour tout $x \in [0, 1]$ on a $|f(x) - \tilde{f}(x)| \leq \varepsilon$ où ε est un réel strictement positif. Montrer que

$$\left|A(n, f) - A(n, \tilde{f})\right| \leq \varepsilon$$

(c) En déduire une majoration pour $\left|I - A(\nu, \tilde{f})\right|$.

(d) Que peut-on dire de la perte de précision entraînée par le calcul de la formule donnant $A(\nu, f)$ sur une calculatrice travaillant avec une précision de 10^{-12} ?

(UPS, L2, 2004, sol. 8 p. 101.)

55 Méthode des parabole à chevauchement. Soit $n \geq 1$. On considère $n+1$ points x_i dans l'intervalle $[a, b]$ de sorte que

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b.$$

(a) Pour $i = 1, \ldots, n-2$, on utilise l'approximation

$$\int_{x_i}^{x_{i+1}} f(x)dx \approx Q_i(f)$$

où

$$Q_i(f) = \frac{1}{2}\left\{\int_{x_i}^{x_{i+1}} \mathbf{L}[x_{i-1}, x_i, x_{i+1}; f](x)dx + \int_{x_i}^{x_{i+1}} \mathbf{L}[x_i, x_{i+1}, x_{i+2}; f](x)dx\right\}.$$

(b) On définit les nombres a_i, b_i et c_i pour $i = 1, \ldots, n-1$ par la relation

$$\mathbf{L}[x_{i-1}, x_i, x_{i+1}; f](x) = a_i x^2 + b_i x + c_i.$$

Démontrer que

$$Q_i(f) = \frac{a_i + a_{i+1}}{2}\left(\frac{x_{i+1}^3 - x_i^3}{3}\right) + \frac{b_i + b_{i+1}}{2}\left(\frac{x_{i+1}^2 - x_i^2}{2}\right) + \frac{c_i + c_{i+1}}{2}(x_{i+1} - x_i).$$

(c) Montrer, en utilisant le théorème mesurant l'erreur entre une fonction et son polynôme d'interpolation, que si $f \in C^3[a,b]$ alors il existe une constante C_i que l'on précisera telle que

$$\left| \int_{x_i}^{x_{i+1}} f(x)dx - Q_i(f) \right| \leq C_i \cdot \sup_{[x_{i-1},x_{i+2}]} |f^{(3)}|.$$

On considère maintenant l'approximation

$$\int_a^b f(x)dx \approx Q(f)$$

avec

$$Q(f) = \int_{x_0}^{x_1} \mathbf{L}[x_0,x_1,x_2;f](x)dx + \sum_{i=1}^{n-2} Q_i(f) + \int_{x_{n-1}}^{x_n} \mathbf{L}[x_{n-2},x_{n-1},x_n;f](x)dx.$$

(d) Montrer que si f est un polynôme de degré ≤ 2 alors $Q(f) = \int_a^b f(x)dx$.

(e) Donner une majoration de l'erreur $\left| \int_a^b f(x)dx - Q_(f) \right|$ lorsque $f \in C^3[a,b]$.

(UPS, L2, 2005, sol. 9 p. 102.)

56 Une démonstration * des formules d'erreur pour les méthodes du point milieu , des trapèzes et de Simpson . Rappelons les théorèmes

Théorème. *(a) Si $f \in C^2([a,b])$ alors*

$$\left| \int_a^b f(t)dt - (b-a)f\left(\frac{a+b}{2}\right) \right| \leq \frac{(b-a)^3}{24} . \max_{[a,b]} |f^{(2)}| \tag{6.3}$$

$$\left| \int_a^b f(t)dt - \frac{(b-a)}{2}\left[f(a)+f(b)\right] \right| \leq \frac{(b-a)^3}{12} . \max_{[a,b]} |f^{(2)}|. \tag{6.4}$$

(b) Si $f \in C^4([a,b])$,

$$\left| \int_a^b f(t)dt - \frac{(b-a)}{6}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right] \right| \leq \frac{(b-a)^5}{2880} . \max_{[a,b]} |f^{(4)}|. \tag{6.5}$$

Nous construisons une démonstration de chaque inégalité en suivant un principe commun. Posons $c := \frac{a+b}{2}$, $a = c-h$ et $b = c+h$ (de sorte que $h = \frac{b-a}{2}$) et considérons

$$\Psi_{\text{Simp}}(t) = \Psi(t) = \int_{c-t}^{c+t} f(x)dx - \frac{t}{3}\left\{f(c+t)+4f(c)+f(c-t)\right\}$$

puis

$$\Phi(t) = \Psi(t) - \left(\frac{t}{h}\right)^5 \Psi(h)$$

∗. [Hardy 1952]

(a) Calculer les trois premières dérivées de Φ et montrer qu'il existe $\xi(t) \in [a,b]$ tel que

$$\Phi^{(3)}(t) = \frac{-2t^2}{3}\left\{f^{(4)}(\xi(t)) + \frac{90}{h^5}\Psi(h)\right\}.$$

(b) Montrer en appliquant plusieurs fois le théorème de Rolle qu'il existe $\overline{t} \in [a,b]$ tel que $\Phi^{(3)}(\overline{t}) = 0$.

(c) En déduire l'estimation (6.5).

(d) Démontrer (6.4) 4 en s'inspirant de la démonstration précédente. On considérera

$$\Psi_{\text{trap}} = \Psi(t) = \int_{c-t}^{c+t} f(x)dx - t\left\{f(c+t) + f(c-t)\right\}$$

puis

$$\Phi(t) = \Psi(t) - \left(\frac{t}{h}\right)^3 \Psi(h)$$

(e) Démontrer (6.3) toujours en suivant la même technique : quelle est la fonction Ψ_{milieu} appropriée ?

III

SOLUTIONS APPROCHÉES DES ÉQUATIONS

§ 1 INTRODUCTION

Le problème de construire une suite convergente vers la solution d'une équation est certainement à l'origine des plus anciens algorithmes mathématiques dont l'exemple le plus fameux est l'algorithme attribué à Héron d'Alexandrie et possiblement connu des mathématiciens babyloniens qui donne une approximation très rapide de $\sqrt{a}$, $a > 0$, l'unique solution positive de l'équation $x^2 - a = 0$. Les équations polynomiales sont les plus communément considérées mais des équations plus générales interviennent couramment comme par exemple dans le recherche des extremums d'une fonction lorsqu'il faut d'abord déterminer les points critiques, c'est-à-dire les valeurs en lesquelles la dérivée s'annule.

De manière précise, étant donnée une fonction continue $f : [a, b] \to \mathbb{R}$, nous cherchons les solutions de l'équation $f(x) = 0$. Les questions auxquelles il faut répondre sont les suivantes.

(i) L'équation a-t-elle des solutions ?

(ii) Si oui, combien en a-t-elle ?

(iii) Déterminer des valeurs aussi proches que nécessaire de ces solutions, étant entendu que les cas pour lesquels une solution exacte exploitable peut être obtenue sont très rares.

Dans ce cours, les solutions des points (i) et (ii) ne seront pas abordés, sauf dans la

dernière partie. Nous supposerons en général que l'équation admet une et une seule solution dans $[a,b]$. Rappelons ici simplement que, dans les cas assez simples, une étude élémentaire de fonction (avec tableau de variation) permet de s'assurer si cette hypothèse est satisfaite. Parmi le grand nombre de méthodes disponibles pour répondre au troisième point, nous étudierons quatre techniques très classiques.

(1) La méthode de dichotomie.

(2) Les méthodes de la sécante et de Newton qui consistent à remplacer l'équation $f(x)=0$ par $p(x)=0$ où p est un polynôme du premier degré – c'est-à-dire une fonction affine – proche de f.

(3) La méthode dite du point fixe ou des approximations successives .

§ 2 MÉTHODE DE DICHOTOMIE (OU DE BISSECTION)

2.1 Définition

La méthode repose uniquement sur le théorème des valeurs intermédiaires. Soit f une fonction continue sur $[a,b]$ satisfaisant les deux conditions suivantes ;

i) f admet une et une seule racine r dans $[a,b]$,

ii) $f(a)f(b)<0$.

Posons $c=(a+b)/2$. Trois cas de figure seulement sont possibles. Ou bien $f(c)=0$ auquel cas la solution de l'équation est trouvée puisque $r=c$, ou bien $f(c)\neq 0$ auquel cas $f(b)f(c)$ est soit négatif soit positif. Si $f(b)f(c)<0$, f change de signe en passant de c à b et, d'après le théorème des valeurs intermédiaires, f s'annule entre c et b. Comme f s'annule une seule fois, cela signifie que $r\in]c,b[$. Maintenant si $f(b)f(c)>0$, puisque $f(a)f(b)<0$, nous avons nécessairement $f(a)f(c)<0$ et le même théorème des valeurs intermédiaires nous donne $r\in]a,c[$.

Ce raisonnement simple nous a permis de nettement préciser la localisation de la solution puisque nous savions au départ que $r\in]a,b[$ et nous connaissons maintenant un intervalle de longueur deux fois moindre contenant r. En itérant le test, nous obtenons une suite qui converge vers la solution de l'équation. Cette itération est décrite dans l'algorithme suivant.

Algorithme 1. *Sous les hypothèses i et ii ci-dessus, il construit trois suites (a_n), (b_n) et (c_n) de la manière suivante.*

(a) $a_1=a$; $b_1=b$.

(b) Pour $n\geq 1$,

  (a) $c_n=\frac{a_n+b_n}{2}$

  (b) i. Si $f(c_n)=0$ *alors* c_n *est la racine de* f *et le processus est arrêté*

ii. Sinon
- *Si* $f(c_n)f(b_n) < 0$ *alors* $a_{n+1} = c_n$ *et* $b_{n+1} = b_n$.
- *Si* $f(c_n)f(b_n) > 0$ *alors* $a_{n+1} = a_n$ *et* $b_{n+1} = c_n$.

L'algorithme ci-dessus s'appelle l'**algorithme de dichotomie** ou **algorithme de bissection**.

2.2 Etude de la convergence

Théorème 2. *Soit* f *continue sur* $[a,b]$. *Nous supposons que* $f(a)f(b) < 0$ *et que l'équation* $f(x) = 0$ *admet une et une seule solution* r *dans* $[a,b]$. *Si l'algorithme de dichotomie arrive jusqu'à l'étape* $n+1$ *(de sorte que* $c_i \neq r$, $0 \leq i \leq n$*) alors*

$$|r - c_{n+1}| \leq \frac{b-a}{2^{n+1}}.$$

Démonstration. Remarquons que

$$b_{n+1} - a_{n+1} = \left\{ \begin{array}{ll} b_n - c_n = & b_n - \frac{a_n+b_n}{2} \\ \text{ou} & \\ c_n - a_n = & \frac{a_n+b_n}{2} - a_n \end{array} \right\} = \frac{b_n - a_n}{2}.$$

En continuant,

$$b_{n+1} - a_{n+1} = \frac{b_{n-1} - a_{n-1}}{4} = \cdots = \frac{b-a}{2^n}.$$

Ensuite, c_{n+1} étant le milieu de $[a_{n+1}, b_{n+1}]$, pour tout $x \in [a_{n+1}, b_{n+1}]$ nous avons

$$|x - c_{n+1}| \leq (b_{n+1} - a_{n+1})/2 = (b-a)/2^{n+1}.$$

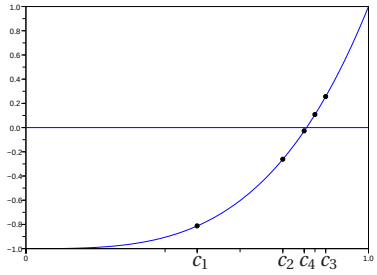
Mais, par définition, la racine r se trouve dans $[a_{n+1}, b_{n+1}]$ car $f(a_{n+1})f(b_{n+1}) < 0$, nous pouvons donc prendre $x = r$ dans l'inégalité précédente pour obtenir

$$|r - c_{n+1}| \leq (b_{n+1} - a_{n+1})/2 = (b-a)/2^{n+1}. \quad \blacksquare$$

Exemple 1.

(a) L'équation $x^4 + x^3 - 1 = 0$ admet une solution (unique) dans $]0,1[$ comme le montre une étude bien simple de fonction. Une approximation $\tilde{r}$ de la racine r avec une erreur moindre que 10^{-6} est obtenu en moins de 0.2 seconde, $\tilde{r} = 0.8191729$, par la méthode de dichotomie. La figure dans la première colonne table 1 représente les quatre premier termes de la suite et la seconde colonne en reporte les valeurs.

[TH 2]



Quatre première valeurs données par l'algorithme de dichotomie pour résoudre l'équation $x^4 + x^3 - 1 = 0$ dans $[0,1]$.

n	c_n	n	c_n
1.	0.5	1.	0.7853982
2.	0.75	2.	1.1780972
3.	0.875	3.	0.9817477
4.	0.8125	4.	1.0799225
5.	0.84375	5.	1.1290099
6.	0.828125	6.	1.1535536
16.	0.8191681	16.	1.1712183
17.	0.8191757	17.	1.1712303
18.	0.8191719	18.	1.1712243
19.	0.8191738	19.	1.1712273
20.	0.8191729	20.	1.1712288
	$x^4 + x^3 - 1 = 0$		$x - \sin x - 1/4 = 0$
	$x \in [0,1]$		$x \in [0, \pi/2]$

TABLE 1 – Exemple d'applications de la méthode de dichotomie

(b) De même, l'équation $x-\sin x-1/4=0$ admet une solution unique dans $]0,\pi/2[$. Une approximation $\tilde{r}$ de la racine r avec une erreur moindre que 10^{-6} est obtenu en moins de 0.2 seconde : $\tilde{r}=1.1712288$ par l'algorithme de dichotomie dont ses valeurs sont indiquées dans la troisième colonne de la table 1..

E 57 Montrer que l'algorithme 1 fonctionne encore si on retire la première hypothèse sur f à savoir que f admet une unique racine dans $[a,b]$ et vérifier qu'il converge toujours vers une racine de f. Dans les deux schémas de la table, dite vers laquelle des racines convergera l'algorithme. Il sera peut-être nécessaire d'utiliser une règle graduée.

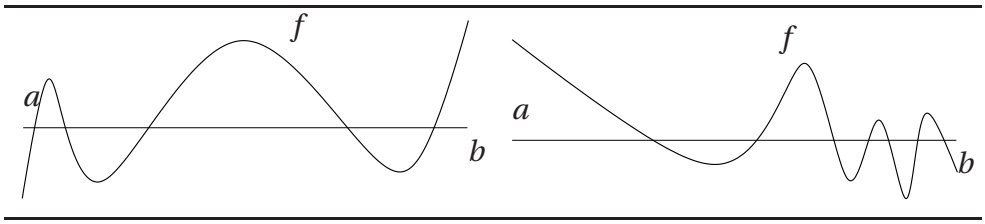


FIGURE 1 – Vers laquelle des racines convergera l'algorithme de dichotomie ?

§ 3 MÉTHODE DE NEWTON

3.1 Construction

Supposons que $f\in C^1[a,b]$ et que l'équation $f(x)=0$ admette une et une seule racine, notée r, dans $[a,b]$. L'idée de la **méthode de Newton** consiste à remplacer l'équation $f(x)=0$ par l'équation $T_1(x)=0$ où T_1 est un polynôme de Taylor de f de degré 1 en un point x_1.

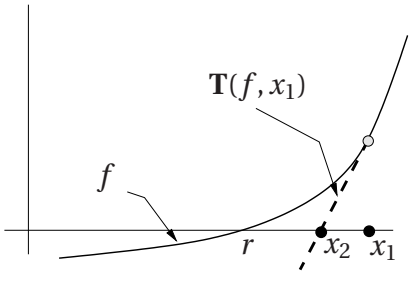


FIGURE 2 – Schéma de Newton.

Puisque $T_1(x)$ est un polynôme du premier degré, le calcul de sa racine est immédiat et il est naturel d'espérer que cette racine sera proche de celle de f. Nous avons

$$T_1(x)=f(x_1)+f'(x_1)(x-x_1),$$

et l'équation $T_1(x)=0$ a pour racine le nombre x_2 donné par

$$x_2=x_1-\frac{f(x_1)}{f'(x_1)}.$$

[TH 2]

Comme toujours, le principe est assez grossier puisqu'un polynôme du premier degré, quel qu'il soit, n'approchera que très imparfaitement la fonction f. C'est en itérant le procédé que nous obtiendrons une bonne approximation de la racine. Une condition nécessaire, qui n'est pas obligatoirement satisfaite, pour effectuer cette itération est le point x_2 appartienne bien à l'intervalle $[a,b]$ faute de quoi, cela n'aurait pas de sens de parler du polynôme de Taylor de f en x_2. Lorsque cette condition est satisfaite, en remplaçant $f(x)=0$ par $T_2(x)=0$ où $T_2(x)=f(x_2)+f'(x_2)(x-x_2)$ et crésolvant cette dernière équation, nous obtenons

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)} \quad \text{avec} \quad r \approx x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}.$$

Nous construisons ainsi par récurrence, *sous réserve que* $x_n \in [a,b]$, la suite

$$\begin{cases} x_1 &= b \\ x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)} \quad (n \geq 0) \end{cases}.$$

Cette relation de récurrence le schéma de Newton. Le mot *schéma* est un synonyme ancien du mot *algorithme*.

E 58 Donner (graphiquement) un exemple de fonction pour laquelle la suite (x_n) n'est pas définie. Il s'agit de construire une fonction pour laquelle il existe une valeur n avec $x_n \notin [a,b]$ de sorte qu'il ne soit pas possible de calculer $f(x_n)$ et donc x_{n+1}.

3.2 Etude de la convergence

Nous devons répondre aux trois questions suivantes.

i) La suite (x_n) est-elle bien définie ?

ii) Si oui, converge-t-elle vers la racine r ?

iii) Si oui, quelle est la rapidité de convergence ?

Les réponses dépendent naturellement des propriétés de la fonction f considérée. De nombreux théorèmes apportent des réponses. Le suivant est l'un des plus simples. Ses hypothèses correspondent à la figure 2.

Théorème 3. *Soit f une fonction de classe C^2 sur un intervalle ouvert I contenant $[a,b]$ telle que f' et f'' soient strictement positives sur I (f est strictement croissante convexe). Nous supposons que $f(b)>0$, $f(a)<0$ et nous appelons r l'unique solution de l'équation $f(x)=0$ dans $[a,b]$.*

(a) La suite de Newton

$$\begin{cases} x_1 &= b \\ x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)} \quad (n \geq 0) \end{cases}$$

est bien définie.

(b) Elle converge vers r en décroissant.

(c) L'estimation suivante est vraie

$$|x_n - r| \leq \frac{M_2}{2m_1}(x_n - r)^2$$

où $M_2 = \sup_{[a,b]} f''$ *et* $m_1 = \inf_{[a,b]} f'$.

Exemple 2. L'équation

$$3x^5 - x^4 - 1 = 0 \tag{3.1}$$

admet une et une seule racine r dans $[0,1]$. En effet la fonction $f(x) = 3x^5 - x^4 - 1$ a pour dérivée $f'(x) = x^3(15x-4)$ et, sur $[0,1]$ elle décroît de $f(0) = -1$ jusqu'à $f(4/15) \approx -1,001$ puis crôit jusqu'à $f(1) = 1$. En particulier $r \in]4/15, 1[$. Par ailleurs, puisque $f''(x) = 12x^5(5x-1)$, f est strictement convexe sur $[1/5, 1]$ en particulier sur $[4/15, 1]$ puique $4/15 > 1/5$. Nous pouvons appliquer le théorème 3 sur l'intervalle $[4/15, 1]$ en prenant Comme point de départ $x_0 = 1$. Les dix premiers termes de la suite de Newton sont donnés dans le tableau 2. Remarquons que nous obtenons les six premières décimales de r dès le quatrième terme de la suite.

$3x^5 - x^4 - 1 = 0$			$f(x) = x - \sin(x) - 1/4$		
n	x_n	$x_n - x_{n-1}$	n	x_n	$x_n - x_{n-1}$
1.	1.		1.	1.5707963 ($\pi/2$)	
2.	0.9090909	- 0.0909091	2.	1.25	- 0.3207963
3.	0.8842633	- 0.0248276	3.	1.1754899	- 0.0745101
4.	0.8826212	- 0.0016421	4.	1.1712433	- 0.0042467
5.	0.8826144	- 0.0000068	5.	1.1712297	- 0.0000136
6.	0.8826144	- 1.161D-10	6.	1.1712297	- 1.397D-10
7.	0.8826144	0.	7.	1.1712297	2.220D-16
8.	0.8826144	0.	8.	1.1712297	- 2.220D-16
9.	0.8826144	0.	9.	1.1712297	2.220D-16
10.	0.8826144	0.	10.	1.1712297	- 2.220D-16

TABLE 2 – Premiers termes de deux suites de Newton

E 59 Justifier l'emploi de la suite de Newton pour l'équation $x - \sin(x) - 1/4 = 0$.

Démonstration du théorème 3. Nous décomposons la démonstration en plusieurs étapes.
Etape 1. Montrons que $r < x_1 < x_0 = b$. Nous avons

$$x_0 = b \Rightarrow \left.\begin{array}{c} f(x_0) > 0 \\ f' > 0 \text{ (hyp.)} \end{array}\right\} \Rightarrow \frac{f'(x_0)}{f(x_0)} > 0 \Rightarrow x_0 - \frac{f'(x_0)}{f(x_0)} < x_0 \Rightarrow x_1 < x_0.$$

[TH 3]

Ensuite, en utilisant la formule de Taylor (th. II.3), il vient

$$f(r) = f(x_0) + (r - x_0)f'(x_0) + \frac{(r-x_0)^2}{2}f''(c)$$

où $c \in]r, x_0[$. Puisque $f(r) = 0$, cette relation devient

$$\begin{aligned} -f(x_0) &= (r-x_0)f'(x_0) + \tfrac{(r-x_0)^2}{2}f''(c) \\ \Rightarrow \quad \tfrac{-f(x_0)}{f'(x_0)} &= (r-x_0) + \tfrac{(r-x_0)^2}{2}\tfrac{f''(c)}{f'(x_0)} \\ \Rightarrow \quad x_0 - \tfrac{f(x_0)}{f'(x_0)} &= r + \tfrac{(r-x_0)^2}{2}\tfrac{f''(c)}{f'(x_0)} \\ \Rightarrow \quad x_1 &= r + \tfrac{(r-x_0)^2}{2}\tfrac{f''(c)}{f'(x_0)} \\ \Rightarrow \quad x_1 &> r \quad \text{car } f'' > 0, f' > 0 \end{aligned}$$

Etape 2. Supposons que nous ayons démontré que

$$(P_n) \qquad r < x_{n+1} < x_n \leq b \qquad \text{(Hypothèse de récurrence).}$$

D'abord, puisque x_{n+1} se trouve dans l'intervalle $[a, b]$, nous pouvons calculer x_{n+2} par la définition, $x_{n+2} = x_{n+1} - \frac{f(x_{n+1})}{f'(x_{n+1})}$. Nous allons établir la promrioété P_{n+1}, que

$$(P_{n+1} \qquad r < x_{n+2} < x_{n+1} \leq b.$$

Remarquons d'abord que la dernière inégalité est déjà contenue dans l'hypothèse de récurrence de sorte que nous devons simplement obtenir $x_{n+2} < x_{n+1}$ et $r < x_{n+2}$. Puisque f est strictement croissante et que, en vertue hypothèse de récurrence, $x_{n+1} > r$, nous avons aussi $f(x_{n+1}) > f(r) = 0$. Ensuite,

$$\left.\begin{matrix} f(x_{n+1}) > 0 \\ f' > 0 \end{matrix}\right\} \Rightarrow -\frac{f'(x_{n+1})}{f(x_{n+1})} < 0 \Rightarrow x_{n+1} - \frac{f'(x_{n+1})}{f(x_{n+1})} < x_{n+1} \Rightarrow x_{n+2} < x_{n+1}.$$

En utilisant à nouveau la formule de Taylor (th. II.3), nous pouvons écrire

$$f(r) = f(x_{n+1}) + (r - x_{n+1})f'(x_{n+1}) + \frac{(r-x_{n+1})^2}{2}f''(c)$$

où $c \in]r, x_{n+1}[$. Utilisant $f(r) = 0$, nous obtenons avec les mêmes calculs que précédemment

$$\begin{aligned} x_{n+1} - \frac{f(x_{n+1})}{f'(x_{n+1})} &= r + \frac{(r-x_{n+1})^2}{2}\frac{f''(c)}{f'(x_{n+1})} \\ \text{d'où} \quad x_{n+2} &= r + \frac{(r-x_{n+1})^2}{2}\frac{f''(c)}{f'(x_{n+1})} \\ \text{d'où} \quad x_{n+2} &> r \qquad \text{car } f'' > 0 \text{ et } f' > 0. \end{aligned}$$

Les étapes 1 et 2 montrent par récurrence que la suite (x_n) est bien définie et vérifie

$$r < x_{n+1} < x_n \leq b \quad n \geq 1.$$

En particulier, étant décroissante et minorée par r la suite (x_n) est convergente. Appelons l sa limite. Nous devons nous assurer que $l = r$. Faisons $n \to \infty$ dans la relation

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{3.2}$$

Nous avons à la fois $x_{n+1} \to l$ et $x_n \to l$. La continuité de f entraîne que $f(x_n) \to f(l)$ et celle de f' que $f'(x_n) \to f'(l)$. Observons que $f'(l)$ est non nul car, par hypothèse, f' ne s'annule jamais. Le passage à la limite $(n \to \infty)$ dans (3.2) donne donc

$$l = l - \frac{f(l)}{f'(l)} \Rightarrow f(l) = 0 \Rightarrow l = r, \tag{3.3}$$

la dernière implication étant justifiée par le fait que f admet une et une seule racine. Enfin, revenant à la relation

$$x_{n+2} = r + \frac{(r - x_{n+1})^2}{2} \frac{f''(c)}{f'(x_{n+1})},$$

étavlie au dessus, nous obtenons

$$|x_{n+2} - r| \leq \left| \frac{(r - x_{n+1})^2}{2} \right| \left| \frac{f''(c)}{f'(x_{n+1})} \right|,$$

$$\text{d'où} \quad |x_{n+2} - r| \leq \left| \frac{(r - x_{n+1})^2}{2} \right| \left| \frac{M_2}{m_1} \right|.$$

■

A cause de la relation $|x_{n+2} - r| \leq C(r - x_{n+1})^2$, nous disons que la méthode de Newton est d'**ordre** 2. Une telle propriété implique une convergence très rapide. Par exemple, si, au rang n l'erreur comparable à 10^{-3}, au rang $n+1$, elle sera au pire comparable à 10^{-6}, au rang $n+2$, 10^{-12} et ainsi de suite.

E 60 Donner une estimation de $|x_{n+2} - r|$ en fonction de C et $|x_1 - r|$. Montrer que si $|x_1 - r| < 1$ alors il existe $\delta < 1$ tel que

$$|x_{n+1} - r| \leq \delta^{2^n}.$$

3.3 Autres versions

Il est facile d'adapter le théorème précédent pour traiter toutes les équations de la forme $f(x) = 0$ lorsque la fonction f et sa dérivée sont toutes deux strictement monotones. Il y a quatre cas à considérer, donnés dans la figure 3.

[TH 3]

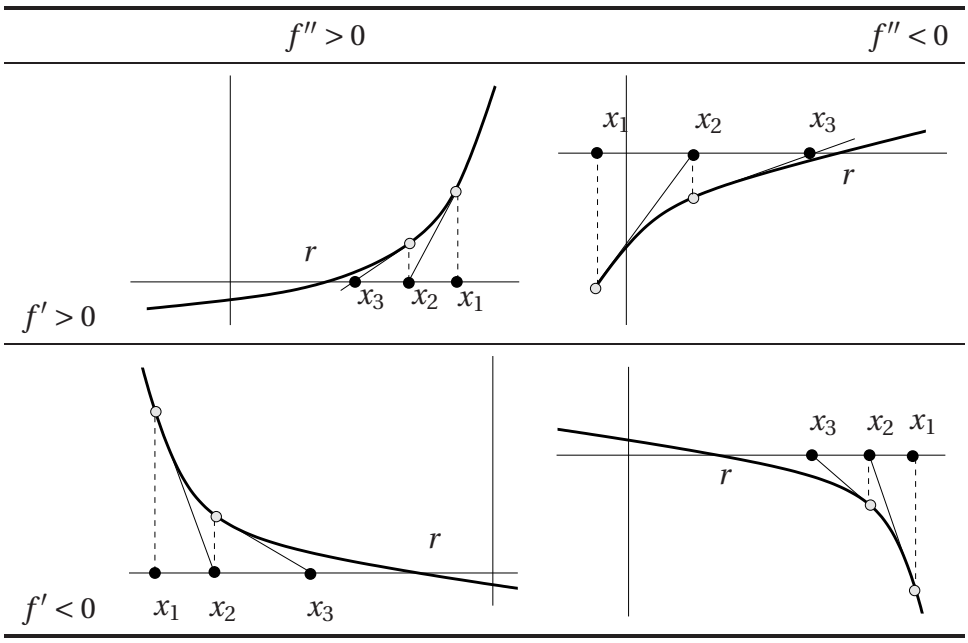


TABLE 3 – Quatre schémas de Newton convergents.

§ 4 MÉTHODE DE LA SÉCANTE

4.1 Construction

Supposons à nouveau que $f \in \mathsf{C}[a,b]$, que l'équation $f(x) = 0$ admette une et une seule solution r dans $[a,b]$ et enfin que $f(a) < 0$, $f(b) > 0$. Comme dans la méthode de Newton, la **méthode de la sécante** consiste à remplacer l'équation $f(x) = 0$ par une équation polynomiale du premier degré en choisissant un polynôme aussi voisin que possible de f. Ici, le choix se porte sur le polynôme d'interpolation de Lagrange $\mathbf{L}[a,b;f](x) = 0$ qui prend donc le rôle du développement de Taylor dans la méthode de Newton. Du point de vue de calcul, cette méthode a l'avantage de ne pas requérir le calcul d'une dérivée. Nous verrons cependant que sa rapidité de convergence est sensiblement plus faible. Puisque

$$\mathbf{L}[a,b;f](x) = f(a) + \frac{f(b)-f(a)}{b-a}(x-a),$$

l'unique solution de l'équation $\mathbf{L}[a,b;f](x)=0$ qui vient se substituer à $f(x)=0$ est donnée par

$$\begin{aligned} x_1 &= -f(a)\frac{b-a}{f(b)-f(a)}+a \\ &= \frac{-f(a)b+af(a)+af(b)-af(a)}{f(b)-f(a)} \\ &= \frac{af(b)-bf(a)}{f(b)-f(a)}. \end{aligned}$$

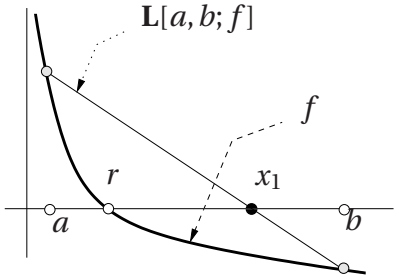


FIGURE 3 –

Si $x_1 \in [a,b]$ — condition, soulignons-le encore une fois, qui n'est pas obligatoirement satisfaite — le procédé peut être itéré en remplaçant $f(x)=0$ par $\mathbf{L}[x_1,b;f](x)=0$, autrement dit, nous faisons jouer à x_1 le rôle que tenait précédemment a. Nous pourrions évidemment envisager l'autre stratégie : garder a et remplacer b par x_1. Le choix, comme nous le verrons, est dicté par les propriétés de la fonction f et dans une variante de la méthode, on garde parfois l'extrémité droite, parfois l'extrémité gauche. Bornons-nous à considérer $\mathbf{L}[x_1,b;f](x)=0$ donc nous noterons la racine x_2. nous obtenons

$$x_2 = \frac{x_1 f(b)-bf(x_1)}{f(b)-f(x_1)}.$$

En poursuivant, nous construisons par récurrence, *sous réserve* que $x_n \in [a,b]$, la suite

$$\begin{cases} x_0 &= a \\ x_{n+1} &= \frac{x_n f(b)-bf(x_n)}{f(b)-f(x_n)} \quad (n\geq 0) \end{cases}.$$

La suite récurrente ainsi construire s'appelle le schéma de la sécante.

E 61 Donner sur un exemple graphique une équation $f(x)=0$ admettant une unique solution mais pour laquelle la suite de la sécante ne peut pas être construite. Voir aussi l'exercice 58.

4.2 Etude de la convergence

Nous devons répondre aux mêmes questions que pour la méthode de Newton. Les hypothèses du théorème suivant correspondent à la figure 4.1.

Théorème 4. *Soit f une fonction de classe C^2 sur un intervalle ouvert I contenant $[a,b]$ telle que f' et f'' soient strictement positives sur I (f est strictement croissante*

convexe). Nous supposons que $f(b) > 0$, $f(a) < 0$ et nous notons r l'unique solution de l'équation $f(x) = 0$ dans l'intervalle $[a, b]$.

(a) La suite

$$\begin{cases} x_0 &= a \\ x_{n+1} &= \frac{x_n f(b) - b f(x_n)}{f(b) - f(x_n)} \end{cases} \quad (n \geq 0)$$

est bien définie.

(b) Elle converge en croissant vers r et

(c) nous avons l'estimation

$$|x_n - r| \leq \frac{M_2}{2m_1}(x_n - x_{n-1})(b - x_n)$$

où $M_2 = \max_{[a,b]} f''$ et $m_1 = \min_{[a,b]} f'$.

Un énoncé plus simple (moins précis) est proposé à l'exercice 71 avec une démonstration différente de celle qui est esquissée ci-dessous.

Démonstration. Elle est similaire en principe à celle du théorème 3. Les détails sont laissés au lecteur. Nous établirions par récurrence sur n que

$$a < x_n \leq x_{n+1} < r < b, \quad n \in \mathbb{N}^\star. \tag{4.1}$$

La relation $x_n \leq x_{n=1}$ s'obtient facilement à partir de l'expression de x_{n+1} en fonction de x_n. Pour le reste, il suffit d'observer que par définition de x_n, $\mathbf{L}[x_{n-1}, b; f](x_n) = 0$ et d'autre part, en utilisant le théorème des accroissements finis

$$f(x_n) = f(x_n) - f(r) = (x_n - r) f'(\theta_n)$$

pour un certain θ_n compris strictement entre x_n et r. Il suit que

$$x_n - r = \frac{f(x_n) - L[x_{n-1}, b; f](x_n)}{f'(\theta_n)}. \tag{4.2}$$

Il reste à utiliser le théorème sur l'erreur dans l'interpolation de Lagrange (th. I.8) qui nous donne

$$f(x_n) - L[x_{n-1}, b; f](x_n) = \frac{1}{2}(x_n - x_{n-1})(x_n - b) f''(\xi) \tag{4.3}$$

pour un ξ compris entre x_{n-1} et b. ■

Exemple 3. Nous reprenons dans la table 4 les exemples étudiés ci-dessus avec la méthode de Newton.

$3x^5 - x^4 - 1 = 0$			$f(x) = x - \sin(x) - 1/4$		
n	x_n	$x_n - x_{n-1}$	n	x_n	$x_n - x_{n-1}$
1.	0.15	- 1.	1.	0.15	
2.	0.5750592	0.4250592	2.	1.4921931	1.3421931
3.	0.7787569	0.2036977	3.	1.1336931	- 0.3585000
4.	0.8533380	0.0745812	4.	1.1767653	0.0430721
5.	0.8749467	0.0216086	5.	1.170437	- 0.0063282
8.	0.8824870	0.0003733	6.	1.1713436	0.0009066
9.	0.8825820	0.0000950	9.	1.1712293	- 0.0000027
10.	0.8826062	0.0000242	10.	1.1712297	0.0000004
11.	0.8826123	0.0000061	11.	1.1712296	- 5.563D-08
12.	0.8826139	0.0000016	18.	1.1712297	7.039D-14
13.	0.8826143	0.0000004	19.	1.1712297	- 9.992D-15
15.	0.8826144	2.570D-08	20.	1.1712297	1.332D-15
18.	0.8826144	4.226D-10			

TABLE 4 – Premiers termes de deux suites de méthode de la sécante

E 62 Indiquer comment adapter la méthode de la sécante à la résolution d'équations $f(x) = 0$ lorsque la fonction f et sa dérivée sont strictement monotones. On donnera un tableau correspondant au tableau 3 pour la méthode de Newton.

E 63 Sous les hypothèses des deux théorèmes précédents (th. 3 et th. 4), on construit la suite $(\underline{x}_n)$ fournie par la méthode de la sécante et la suite $(\overline{x}_n)$ fournie par la méthode de Newton. Montrer que lorsque $\underline{x}_n$ et $\overline{x}_n$ ont les mêmes k premières décimales, ce sont aussi les k premières de r.

§ 5 LE THÉORÈME DU POINT FIXE

5.1 Introduction

Dans cette partie, nous considérons les équations de la forme $x = g(x)$. Nous étudierons un théorème qui, à la fois (a) garantit l'existence et l'unicité de la solution (b) fournit une suite qui converge rapidement vers la solution. Le procédé employé – les approximations successives – joue un rôle très important en mathématiques. Il peut être étendu à l'étude d'équations plus complexes dans lesquelles les inconnues sont des fonctions, par exemple, les équations différentielles.

5.2 Enoncé du théorème du point fixe

Théorème 5. *Soit I un intervalle fermé (non nécessairement borné) et g une fonction de I dans I. S'il existe un réel $k<1$ tel que*

$$|g(x)-g(y)| \le k|x-y| \quad x,y \in I$$

alors l'équation

$$g(x)=x$$

admet une et une seule solution dans I. Cette solution est limite de la suite (x_n) définie par

$$\begin{cases} x_0 &= a \in I \\ x_{n+1} &= g(x_n) \quad (n \ge 0) \end{cases}$$

(On est libre de choisir n'importe quel x_0 dans I). De plus, si s est la solution de l'équation $g(x)=x$ alors

$$|s-x_n| \le \frac{k^n}{1-k}|x_1-x_0|, \qquad n \ge 1.$$

L'intervalle I est de la forme $I=\mathbb{R}$ ou $I=]-\infty,a]$ ou $[a,+\infty[$ ou $[a,b]$. Il est essentiel que g prenne ses valeurs dans I c'est-à-dire que son ensemble image soit inclus dans son ensemble de définition, faute de quoi nous ne serions plus sûrs que la suite (x_n) soit bien définie.

Lorsqu'une fonction vérifie une inégalité

$$|g(x)-g(y)| \le k|x-y|, \quad x,y \in I$$

avec $0 \le k < 1$, nous disons que f est **contractante** ou bien que c'est une **contraction** de constante k. Les fonctions contractantes sont continues en tout point. Fixons $x_0 \in I$ et montrons la continuité en x_0. Nous devons établir que $\forall \epsilon > 0$, $\exists \eta > 0$ tel que les conditions ($|x-x_0| \le \eta$ et $x \in I$) impliquent $\left|g(x)-g(x_0)\right| \le \epsilon$. Or ϵ était fixé, il suffit de prendre $\eta = \epsilon/k$.

Lorsque g est dérivable, pour qu'elle soit contractante de constante k, il suffit que

$$\sup_I |g'| \le k.$$

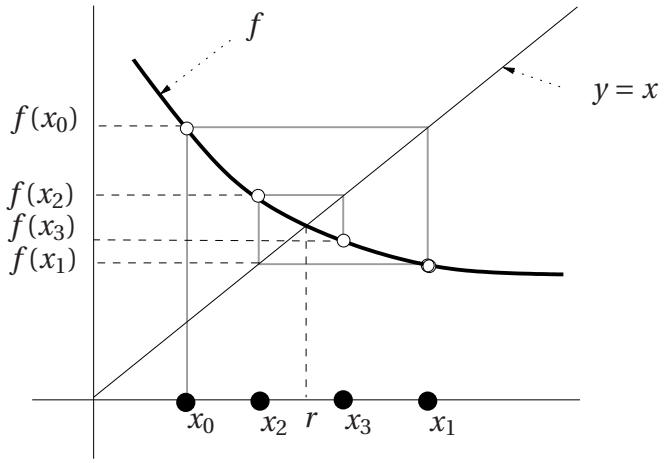
En effet, d'après le théorème des accroissements finis,

$$|g(x)-g(y)| = |g'(c)||x-y| \le k|x-y|.$$

Remarquons enfin que les suites définies par les méthode de Newton et de la sécante sont des cas particuliers de schémas d'approximations successives. Dans le premier cas, nous avons $g(x)=x-f(x)/f'(x)$ et dans le second $g(x)=(xf(b)-bf(x)/(f(b)-f(x))$. Cependant, l'étude de ces méthodes comme cas particuliers de la méthode du point fixe est moins élémentaire que celle qui a été donnée dans ce cours.

5.3 Illustration graphique

La figure 4 montre un exemple de construction des premiers termes de la suite des approximations successives. Remarquons que les points $M_n = (x_n, f(x_n)$ convergent en s'enroulant autour de $(r, f(r)) = (r, r)$.



Com. Construction des quatre premiers termes d'un suite d'approximations successives.

FIGURE 4 – La méthode du point fixe.

Exemple 4. La table 5 donne les résultats obtenus en appliquant la méthode du point fixe à l'équation $x = \sin(x) + 1/4$ en prenant deux points de départ différents.

E 64 Montrer que la fonction f définie par $f(x) = \sin(x) + 1/4$ vérifie bien les conditions du théorème du point fixe en prenant comme intervalle de départ $I = [0, \pi/2]$.

5.4 Démonstration du théorème du point fixe

Il est facile de voir que si l'équation $g(x) = x$ admet une solution alors cette solution est unique. En effet, si s_1 et s_2 sont deux solutions, nous avons $|s_1 - s_2| = \big|g(s_2) - g(s_2)\big| \leq k|s_1 - s_2|$ ce qui n'est possible que si $s_1 = s_2$ car $k < 1$.

Lorsque $I = [a, b]$ un argument très simple permet de montrer que l'équation $g(x) = x$ admet au moins une solution et donc, d'après la remarque précédente, une unique solution. Considérons en effet la fonction f définie sur $[a, b]$ par $f(x) = g(x) - x$. Nous avons

(a) $g(b) \in [a, b] \implies g(b) \leq b \implies f(b) \leq 0$ et

(b) $g(a) \in [a, b] \implies g(a) \geq a \implies f(a) \geq 0$.

$f(x) = x - \sin(x) - 1/4$			$f(x) = x - \sin(x) - 1/4$		
n	x_n	$x_n - x_{n-1}$	n	x_n	$x_n - x_{n-1}$
1.	1.		1.	0.5	
2.	1.091471	0.0914710	2.	0.7294255	0.2294255
3.	1.1373063	0.0458353	3.	0.9164415	0.1870159
4.	1.1575053	0.0201990	4.	1.0434407	0.1269993
5.	1.165804	0.0082987	5.	1.1141409	0.0707001
6.	1.1691054	0.0033014	6.	1.1475323	0.0333914
7.	1.1704012	0.0012958	7.	1.1617531	0.0142208
8.	1.1709071	0.0005058	8.	1.1675018	0.0057487
9.	1.1711041	0.0001971	9.	1.169773	0.0022713
10.	1.1711808	0.0000767	10.	1.170662	0.0008890
15.	1.1712292	0.0000007	15.	1.1712246	0.0000080
20.	1.1712296	6.090D-09	20.	1.1712296	7.084D-08
25.	1.1712297	5.426D-11	25.	1.1712297	6.312D-10
30.	1.1712297	4.834D-13	30.	1.1712297	5.624D-12

TABLE 5 – Premiers termes d'une suite d'approximations successives

De $f(b) \leq 0$ et $f(a) \geq 0$ nous déduisons à l'aide du théorème des valeurs intermédiaires que f admet une racine dans $[a,b]$ autrement que l'équation $g(x) = x$ admet une solution.

Nous nous replaçons maintenant dans le cas général où I n'est pa supposé de la forme $[a,b]$, nous admettrons pour le moment que la suite (x_n) converge mais montrerons que sa limite l satisfait la relation $g(l) = l$ ainsi que les inégalités annoncées par le théorème. Le premier point est immédiat. En effet, si $x_n \to l$ alors $x_{n+1} \to l$. Faisant $n \to \infty$ dans la relation $x_{n+1} = g(x_n)$, nous obtenons directement, grâce à la continuité de g, $l = g(l)$ de sorte que l est bien solution de l'équation $g(x) = x$ et, d'après ce qui précède, est l'unique solution. Le même raisonnement a été utilisée dans la démonstration du théorème 3.

Nous démontrerons les inégalités à l'aide de quelques lemmes.

Lemme. $\forall p \geq 0, \left|x_{p+1} - x_p\right| \leq k^p \left|x_1 - x_0\right|$.

Démonstration. D'après la définition de la suite et en utilisant que g est une contraction. Nous avons

$$\begin{aligned} \left|x_{p+1} - x_p\right| = \left|g(x_p) - g(x_{p-1})\right| &\leq k\left|x_p - x_{p-1}\right| = k\left|g(x_{p-1}) - g(x_{p-2})\right| \\ &\leq k^2\left|x_{p-1} - x_{p-2}\right| \leq \cdots \leq k^p \left|x_1 - x_0\right|. \end{aligned} \tag{5.1}$$

■

Lemme. $\forall q > p \geq 0$,

$$\left|x_q - x_p\right| \leq \frac{k^p - k^q}{1-k} |x_1 - x_0|.$$

Démonstration.

$$\begin{aligned}
\left|x_q - x_p\right| &= \left|x_q - x_{q-1} + x_{q-1} - x_{q-2} + \cdots + x_{p+1} - x_p\right| \\
&\leq \left|x_q - x_{q-1}\right| + \left|x_{q-1} - x_{q-2}\right| + \cdots + \left|x_{p+1} - x_p\right| \\
&\leq (k^{q-1} + k^{q-2} + \cdots + k^p) |x_1 - x_0| \quad \text{(d'après le Lemme 5.4)} \\
&\leq k^p (k^{q-1-p} + k^{q-2-p} + \cdots + k^1 + 1) |x_1 - x_0| \\
&\leq k^p \frac{1-k^{q-p}}{1-k} |x_1 - x_0| \leq \frac{k^p - k^q}{1-k} |x_1 - x_0|. \quad \blacksquare
\end{aligned}$$

C'est ce lemme qui permet de démontrer la convergence de la suite x_n que nous avons admis. La démonstration utilise le critère de Cauchy pour la convergence des suites. Ce critère sera rappelé plus bas et le lecteur intéressé pourra alors compléter par lui-même la démonstration du théorème du point fixe.

Admettant donc que la suite (x_n) converge, nous obtenons en appliquant l'inégalité du lemme précédent avec $p = n$ et $q = p + n$

$$\left|x_{p+n} - x_n\right| \leq \frac{k^n}{1-k}(1-k^q) |x_1 - x_0| \quad (n, p \geq 0).$$

Faisons $p \to \infty$ dans l'inégalité. Puisque

$$x_{p+n} \to l = s = \text{solution de "} g(x) = x\text{"}$$

et $k^q \to 0$ (car $0 < k < 1$) nous obtenons

$$|s - x_n| \leq \frac{k^n}{1-k} |x_1 - x_0|.$$

5.5 Démonstration de la convergence de la suite x_n

Le critère de Cauchy, qui est la propriété fondatrice de l'ensemble des nombres réels, dit que pour qu'un suite de nombres réels u_n converge, il faut et il suffit qu'elle satisfasse la condition suivante : pour tout $\epsilon > 0$, il existe un entier $N = N_\epsilon$ tels que $q > p > N \implies |x_p - x_q| \leq \epsilon$. Seule la condition suffisante est non élémentaire et c'est celle que nous devons utiliser pour la démonstration de la convergence de la suite u_n. Fixons $\epsilon > 0$ et choisissons N de telle sorte que $\frac{k^N}{1-k}|x_1 - x_0| \leq \epsilon$ où $k \in]0,1[$ est la constante de contraction de la fonction f ci-dessus. L'existence de N vérifiant la

condition demandée est garantie par le fait que la suite $\frac{k^N}{1-k}|x_1 - x_0|$ tend vers 0 lorsque $N \to \infty$, propriété qui découle elle-même du fait que k est une constante positive plus petite que 1. Maintenant si $q > p > N$, grâce au lemme 5.4, nous avons

$$|x_q - x_p| \leq \frac{k^p - k^q}{1-k}|x_1 - x_0| \leq \frac{k^p}{1-k}|x_1 - x_0| \leq \frac{k^N}{1-k}|x_1 - x_0| \leq \epsilon. \tag{5.2}$$

Ceci montre que la suite x_n vérifie le critère de Cauchy, il s'agit donc d'une suite convergente et, du fait que l'intervalle I est supposé fermé, sa limite est nécessairement incluse dans I.

§ 6 EXERCICES ET PROBLÈMES

65 Un exemple. Montrer que l'équation $x^4 + x^3 - 1 = 0$ admet une et une seule solution dans $[0,1]$. Trouver une valeur approchée avec une décimale exacte en utilisant l'algorithme de dichotomie.

66 Un exemple Montrer que l'équation $x^3 + 2x - 1 = 0$ admet une et une seule solution r dans $[0,1]$ et déterminer les deux première décimales de r. Le candidat choisira une méthode différente de la dichotomie. Il devra clairement expliquer sa démarche. (UPS, L2, 2006)

67 Un exemple. On considère l'équation $x^4 + 2x^2 - 1 = 0$.

(a) Montrer que l'équation admet une et une seule racine r dans $[0,1]$.

(b) Montrer en utilisant la méthode de dichotomie (en partant de $[a,b] = [0,1]$) que $r \in]0,5;0,75[$.

(c) On souhaite maintenant affiner l'approximation en utilisant la **méthode de Newton**. La suite de Newton est notée $\overline{x}_n$. Faut-il choisir $\overline{x}_0 = 0,5$ ou $\overline{x}_0 = 0,75$? Expliquer votre choix puis calculer les deux premières valeurs $(\overline{x}_1, \overline{x}_2)$.

(d) Si on souhaite appliquer la méthode de la sécante dont la suite est notée $\underline{x}_n$, faut-il choisir $\underline{x}_0 = 0,5$ ou $\underline{x}_0 = 0,75$? Expliquer votre choix puis calculer les trois premières valeurs $(\underline{x}_1, \underline{x}_2, \underline{x}_3)$.

(e) Déterminer r avec 2 décimales exactes en expliquant votre raisonnement.

(UPS, L2, 2004, sol. 11 p. 103.)

68 Un exemple. On souhaite trouver une valeur approchée de l'équation

$$\frac{e^{-x}}{x} = 1.$$

On définit la fonction f sur $\mathbb{R}$ par $f(x) = x - e^{-x}$.

(a) Montrer que x est solution de (E) si et seulement si $f(x) = 0$.

(b) Montrer, en étudiant la fonction f, que l'équation $f(x) = 0$ admet une et une seule solution dans $\mathbb{R}$ et que celle-ci se trouve dans l'intervalle $]0,1[$.

(c) Montrer que f est concave sur $[0,1]$ et en déduire le schéma de Newton approprié à l'approximation de r, la solution de $f(x) = 0$. (On pourra faire un schéma expliquant et illustrant le choix du point de départ x_0.)

(d) Calculer x_1, x_2, x_3, où (x_n) désigne la suite de Newton de point de départ x_0.

(e) Montrer, en utilisant un argument de type "dichotomie" que les quatre premières décimales de x_3 sont aussi les quatre premières décimales de la racine r (autrement dit les quatre premières décimales de x_3 sont "correctes").

(UPS, L2, 2005, sol. 10, p. 103.)

69 Un exemple. On considère l'équation

$$x^5 - 7x + 4 = 0 \tag{6.1}$$

(a) Montrer que l'équation (6.1) admet une et une seule solution dans l'intervalle $[0,1]$. Cette solution sera notée r.

(b) Donner une approximation de cette racine avec 4 décimales exactes en utilisant la méthode de Newton. Les détails des calculs devront figurer explicitement sur la copie et on devra justifier clairement les points suivants : (a) Pourquoi est-il légitime d'employer la méthode de Newton dans ce cas ? (b) Sur quoi se fonde votre choix du point de départ ? (c) Comment vous assurez-vous que les quatre décimales données sont bien celles de r ?

70 Une famille de schémas de Newton. Pour les équations $F(x) = 0$ suivantes, étudiez l'unicité des solutions et étudiez s'il est possible d'employer la méthode de Newton pour obtenir des solutions approchées (choix de l'intervalle, vérification des hypothèses sur la fonction, choix du point de départ). Le nombre a désigne toujours un nombre réel strictement positif.

(a) $F(x) = 1/x - a$,

(b) $F(x) = x^2 - a$.

NOTE. — La suite obtenue en (b) est connue depuis l'antiquité. La tradition l'a attribuée à Héron d'Alexandrie (1er siècle après JC).

71 Un résultat de convergence pour la suite de la sécante. on souhaite établir le résultat suivant que l'on comparera avec le théorème 4.

Théorème. *soit f une fonction continue strictement croissante et strictement convexe sur l'intervalle $[a,b]$ telle que $f(a) < 0 < f(b)$. La suite de la sécante*

$$\begin{cases} x_0 &= a \\ x_{n+1} &= \frac{x_n f(b) - b f(x_n)}{f(b) - f(x_n)} \quad (n \geq 1) \end{cases} \qquad \text{(SCHÉMA DE LA SÉCANTE)}$$

est bien définie et converge en croissant vers l'unique racine r de l'équation $f(x) = 0$.

[TH 5]

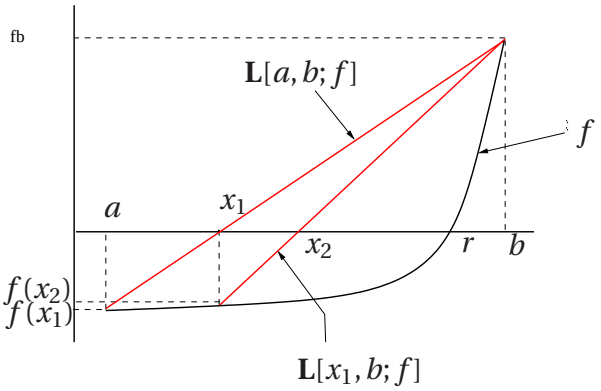


FIGURE 5 – Schéma de la sécante.

Le schéma ci-après montre la construction des valeurs x_1 et x_2 de la suite. On rappelle que f strictement convexe signifie que pour tous x, y dans $[a,b]$ on a $f(tx+(1-t)y) < tf(x)+(1-t)f(y)$ lorsque $t \in]0,1[$.

(a) Trouver $t \in]0,1[$ tel que $x_1 = ta+(1-t)b$. En déduire en utilisant la convexité de f que $f(x_1) < 0$ puis que $x_1 < r$.

(b) Montrer par récurrence sur n la propriété $P(n)$ définie par

$$P(n): \quad a \le x_n < x_{n+1} < r < b.$$

(c) En déduire la démonstration du théorème.

72 Une modification de la méthode de la sécante. Soit $f : [a,b] \to \mathbb{R}$ strictement croissante telle que $f(a) < 0$ et $f(b) > 0$. Pour approcher la racine $r \in]a,b[$ de l'équation $f(x) = 0$, on construit une suite x_k de la manière suivante $x_0 = a$, $x_1 = b$ et, pour $k \ge 2$ x_{k+1} est l'abcisse de l'intersection de la droite joignant les points $(x_k, f(x_k))$ et $(x_{k-1}, f(x_{k-1}))$ avec le droite $y = 0$.

(a) Construire sur une figure les quatre premiers points de la suite lorsque f est une fonction convexe. La construction vous parait-elle judicieuse lorsque f est décroissante convexe ?

(b) Donner l'équation exprimant x_{k+1} en fonction de x_k et x_{k-1}.

(c) Dans une autre variante, on construit x_{k+1} non à partir de x_k et x_{k-1} mais à partir de x_k et $x_{k'}$ où k' est le plus grand indice $(< k)$ tel que $f(x_k)$ et $f(x_{k'})$ soient de signes opposés. Donner un exemple pour lequel cette nouvelle suite ne coïncide pas avec la précédente. Ecrire un algorithme calculant les n premières valeurs de la suite x_k.

73 Une modification de la méthode de Newton. Dans la méthode de Newton, ayant à disposition les points $x_0, \dots, x_n$,on construit x_{n+1} en prenant l'intersection de la tangente au graphe de f en x_n avec l'axe des abcisses. Dans la méthode de Newton modifiée, ayant construit $x_0, \dots, x_n$, on construit x_{n+1} en prenant l'intersection avec l'axe des abscisses de la droite passant par x_n et parallèle à la tangente au graphe de f en x_0.

(a) On suppose que la fonction F est strictement croissante et strictement convexe sur $[a,b]$ avec une racine dans $]a,b[$. On prend $x_0 = b$. Faites un dessin faisant apparaître les quatre pre-

mières valeurs données par la méthode de Newton modifiée. Comparer avec le schéma correspondant pour la méthode de Newton ordinaire.

(b) Donner l'expression de x_{n+1} en fonction de x_n.

(c) Selon vous quels sont les avantages pratiques de cette modification ? Ses inconvénients ?

74 Monotonie des suites d'approximations successives. Soit f une fonction contractante de $[0,1]$ dans $[0,1]$ et $a \in [0,1]$. On construit la suite (x_n) définie par $x_0 = a$ et $x_{n+1} = f(x_n)$ dont on sait, d'après le cours, qu'elle converge vers l'unique solution de $f(x) = x$.

(a) On suppose que f est croissante, étudier la monotonie de la suite (x_n).

(b) On suppose que f est décroissante, étudier la monotonie des sous-suites (x_{2n}) et (x_{2n+1}).

75 Un exemple. On considère l'équation

$$x^3 - 3x - 1 = 0. \tag{6.2}$$

L'équation (6.2) possède une solution et une seule dans l'intervalle $[1,2]$. Cette propriété est admise, on ne demande pas de la démontrer. On appelle r cette solution. On cherche à obtenir une valeur approchée de r en utilisant le théorème du point fixe (la méthode des approximations successives). Pour cela on doit mettre l'équation sous la forme $x = f(x)$.

(a) Expliquer pourquoi le choix $f(x) = (x^3 - 1)/3$ n'est pas judicieux.

Dans la partie suivante on pose $f(x) = (3x+1)^{1/3}$.

(b) Montrer que la fonction $f : [1,2] \to \mathbb{R}$ vérifie toutes les hypothèses du théorème du point fixe (de telle sorte que toute suite x_n définie par $x_0 = a \in [1,2]$ et $x_{n+1} = f(x_n)$ $(n \geq 0)$ converge vers r).

(c) Montrer que la suite x_n est croissante ou décroissante suivant que $f(x_0) > x_0$ ou $f(x_0) < x_0$. Les deux cas peuvent-ils se produire ?

(d) Donner une approximation de r avec deux décimales exactes.

(e) Auriez-vous recommandé la méthode décrite dans cet exercice pour trouver une approximation de r ? Justifiez précisément votre réponse.

(Sol. 12 p. 104.)

76 Un exemple. Montrer que l'équation $\cos x + 1/10 = x$ admet une solution unique sur $[0, 3\pi/8]$ et donner — en justifiant mathématiquement votre réponse — une méthode (autre que la dichotomie) qui permettrait d'obtenir une approximation de cette solution. Donner une valeur approchée de la solution avec trois décimales exactes.

IV

RÉSOLUTION DES SYSTÈMES LINÉAIRES. MÉTHODES DIRECTES.

§ 1 RAPPEL SUR LES SYSTÈMES LINÉAIRES

1.1 Introduction

De nombreux phénomènes de physique ou d'économie se traduisent par des systèmes linéaires de plus ou moins grande dimension. Le problème de la grille illustré dans la figure 1 est un exemple typique. Les 20 extrémités de la grille que sont les points p_i, q_i, d_i et g_i, $i = 1,\ldots,5$, représentés par des disques de couleur blanche sur la figure, sont portés aux températures $t(p_i)$, $t(q_i)$, $t(d_i)$ et $t(g_i)$. Le problème est de déterminer la température aux noeuds $n_{i,j}$ représentés par des disques de couleur noire, sachant que la température en un noeud donné est égale à la moyenne des températures des autres noeuds auxquels il est connecté. Ainsi, par exemple, nous avons

$$t(n_{11}) = \frac{1}{4}(t(p_1)+t(g_1)+t(n_{12})+t(n_{21})) \quad \text{et} \quad t(n_{23}) = \frac{1}{4}(t(n_{12})+t(n_{22})+t(n_{33})+t(n_{24})).$$

Il y a 25 noeuds n_{ij} et 25 inconnues $t(n_{ij})$, $1 \le i, j \le 5$, et, pour déterminer la température en ces 25 noeuds, nous disposons de 25 équations. Nous avons donc ici autant d'inconnues que d'équations. Nous nous limiterons à l'étude de ce type de systèmes. Signalons encore, que les systèmes linéaires n'interviennent pas uniquement dans les

sciences appliquées. En mathématiques même, quantité de questions reposent *in fine* sur le résolution d'un système linéaire et pour ce qui est de l'analyse numérique, le phénomène est encore plus accentué car la plupart des techniques avancées nécessitent à un moment ou à un autre la résolution d'un système linéaire comportant un grand nombre d'inconnues. La résolution des systèmes linéaires est certainement un des rares problèmes fondamentaux des mathématiques. Malheureusement, ce cours voulant s'adresser à un auditoire n'ayant que des connaissances rudimentaires d'algèbre linéaire, nous ne pourrons que donner une très succincte introduction à la théorie à travers l'étude de la méthode de Gauss. Celle-ci est improprement attribuée à Gauss, la technique était déjà connue des mathématiciens chinois de l'antiquité.

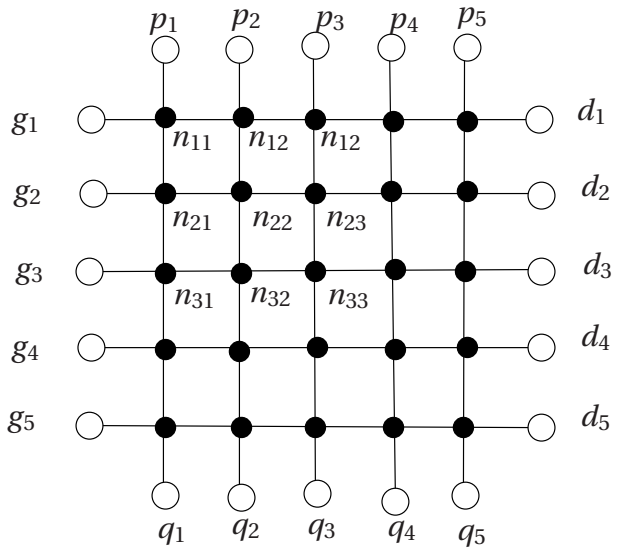


FIGURE 1 – Le problème de la grille

1.2 Le formalisme

Les nombres considérés sont des réels mais tout ce qui sera dit dans ce chapitre reste vrai avec des nombres complexes. Au reste, les techniques utilisées dans ce chapitre sont entièrement algébriques : elles ne font pas appel au concept de limite, et elles seraient également valables pour des systèmes dont les coefficients a_{ij} seraient des éléments d'un corps commutatif quelconque. Nous considérons le système de n

équations à n inconnues suivant

$$\begin{array}{c} \mathbf{L_1} \\ \mathbf{L_2} \\ \vdots \\ \mathbf{L_i} \\ \vdots \\ \mathbf{L_n} \end{array} \left\{ \begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1n}x_n & = & c_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \dots & + & a_{2n}x_n & = & c_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{i1}x_1 & + & a_{i2}x_2 & + & \dots & + & a_{in}x_n & = & c_i \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \dots & + & a_{nn}x_n & = & c_n \end{array} \right. . \tag{1.1}$$

Les a_{ij}'s sont appelés les **coefficients** du système (1.1), les x_i's sont les inconnues (ou les solutions) et les c_i's forment le **second membre**. L'expression

$$\mathbf{L_i} \ : \ a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = c_i \tag{1.2}$$

s'appelle la i-ième **ligne** du système. Le système (1.1) se représente aussi sous la forme compacte

$$\sum_{j=1}^{n} a_{ij}x_j = c_i, \qquad i = 1, 2, \dots, n. \tag{1.3}$$

Au système linéaire (1.1) est associée l'**équation matricielle**

$$AX = C, \tag{1.4}$$

où la matrice A et les vecteurs X et C sont définis par

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \dots & a_{in} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}, \qquad X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix}, \qquad C = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_i \\ \vdots \\ c_n \end{pmatrix}. \tag{1.5}$$

La matrice A ayant n lignes et n colonnes est dite matrice carrée d'**ordre** n. Notons que dans a_{ij}, l'indice i désigne la ligne tandis que j désigne la colonne. Le vecteur colonne X est appelé le **vecteur inconnu** (ou **vecteur solution**) et C est le **vecteur second membre**. Un vecteur $x \in \mathbb{R}^n$ est habituellement noté $x = (x_1, x_2, \dots, x_n)$ mais, pour des raisons propres au calcul matriciel, lorsque nous lui appliquons une matrice A, nous avons intérêt à le représenter comme une colonne X. Dans la suite de ce chapitre, nous ne distinguerons plus, au niveau de la notation, le vecteur ligne x du vecteur colonne

X. Indiquons toutefois que, à un niveau plus avancé, où tous les objets sont considérés comme des matrices, un vecteur x est une matrice à 1 ligne et n colonnes tandis que la colonne X est une matrice à n lignes et 1 colonne et il est alors essentiel de maintenir la distinction entre les deux.

Rappelons encore qu'une application linéaire $\mathscr{A}$ de $\mathbb{R}^n$ dans $\mathbb{R}^n$ est associée à la matrice A. Cette application est définie par la relation

$$\mathscr{A}(x) = \Big(\sum_{j=1}^{n} a_{1,j}x_j, \dots, \sum_{j=1}^{n} a_{i,j}x_j, \dots, \sum_{j=1}^{n} a_{n,j}x_j \Big), \quad x = (x_1, \dots, x_n). \tag{1.6}$$

E 77 Rappeler les liens entre les des images des éléments de la base canonique de $\mathbb{R}^n$ par l'application linéaire $\mathscr{A}$ et les coefficients de la matrice A.

E 78 La matrice A associée au système linéaire du problème de la grille (voir 1.1) est une matrice carrée d'ordre 25. Elle contient donc 625 coefficients. Combien parmi eux sont-ils non nuls ?

1.3 Rappels des résultats fondamentaux

Théorème 1 (†)**.** *On ne modifie pas les solutions d'un système linéaire si on ajoute à une ligne une combinaison linéaire des* autres *lignes. On écrit*

$$\mathbf{L_i} \longleftarrow \mathbf{L_i} + \sum_{j \neq i} \alpha_j \mathbf{L_j}.$$

Il faut prendre garde à ne pas oublier d'effectuer la combinaison linéaire au niveau du second membre.

Théorème 2 (†)**.** *Pour que le système (1.1) admette une et une seule solution il faut et il suffit que* $\det A \neq 0$. *Dans ce cas la matrice A est inversible et l'unique solution est donnée par* $X = A^{-1}(C)$.

Lorsque le système (1.1) admet une et une seule solution, nous disons que c'est un **système régulier**.

E 79 Rappeler les règles de calcul du déterminant d'une matrice.

La condition sur le déterminant de A est à son tour équivalente à des propriétés naturelles de l'application linéaire $\mathscr{A}$. De manière précise, si A est une matrice carrée d'ordre n d'application linéaire associée $\mathscr{A}$, nous avons

$$\det A \neq 0 \iff \mathscr{A} \text{bijective} \iff \mathscr{A} \text{surjective} \iff \mathscr{A} \text{injective} \iff \ker \mathscr{A} = \{0\}. \tag{1.7}$$

Rappelons qu'ici l'équivalence entre bijective, surjective et injective est vraie uniquement parce que $\mathscr{A}$ est une application linéaire entre deux espaces vectoriels de même dimension.

E 80 Donner un exemple d'application linéaire de $\mathbb{R}^2$ dans $\mathbb{R}^3$ qui soit injective (mais pas surjective). Donner un exemple d'application linéaire de $\mathbb{R}^3$ dans $\mathbb{R}^2$ surjective mais non injective.

Théorème 3 (Formules de Cramer , †)**.** *Lorsque* $\det A \neq 0$ *la coordonnée* x_j *de la solution* x *du système (1.1) est donnée par la formule*

$$x_j = \frac{1}{\det A}\begin{vmatrix} a_{11} & \dots & a_{1j-1} & c_1 & a_{1j+1} & \dots & a_{1n} \\ a_{21} & \dots & a_{2j-1} & c_2 & a_{2j+1} & \dots & a_{2n} \\ \vdots & & & & & & \vdots \\ a_{n1} & \dots & a_{nj-1} & c_n & a_{nj+1} & \dots & a_{nn} \end{vmatrix}, \quad j = 1, \dots, n.$$

Pour obtenir x_j *on doit donc calculer le déterminant obtenu à partir de* A *en substituant le vecteur* C *à la* j*-ième colonne de* A*.*

Malheureusement les formules de Cramer nécessitent un nombre d'opérations trop grand et elles sont en pratique inutilisables, cf. exercice 83).

§ 2 LE CAS DES SYSTÈMES TRIANGULAIRES

2.1 L'analyse du cas $n = 3$

Exception faite des systèmes diagonaux (ceux dont la matrice est une matrice diagonale) qui se réduisent à n équations du type $a_{ii}x_i = c_i$, $i = 1, \dots, n$, les systèmes les plus simples sont les **systèmes triangulaires** c'est-à-dire ceux dont les matrices sont triangulaires : tous les coefficients au dessus (matrice triangulaire inférieure) de la diagonale $(a_{11}, a_{22}, \dots, a_{nn})$ ou au dessous (matrice triangulaire supérieure) sont nuls. Dans cette partie, nous donnons les algorithmes élémentaires pour résoudre les systèmes linéaires triangulaires par substitutions successives et étudions la complexité de ces les algorithmes. Nous verrons ensuite comment tout système régulier peut se réduire à un système triangulaire. Cela signifie qu'étant donné un système régulier quelconque $AX = C$, il est toujours possible de construire une matrice triangulaire (supérieure) U et un vecteur C' tel que $UX = C'$ a la même solution que $AX = C$.

Considérons les deux systèmes linéaires suivants.

$$(S_1) \qquad\qquad (S_2)$$

$$\begin{cases} l_{11}x_1 & & & & & = & c_1 \\ l_{21}x_1 & + & l_{22}x_2 & & & = & c_2 \\ l_{31}x_1 & + & l_{32}x_2 & + & l_{33}x_3 & = & c_3 \end{cases} \qquad \begin{cases} u_{11}x_1 & + & u_{12}x_2 & + & u_{13}x_3 & = & c_1 \\ & & u_{22}x_2 & + & u_{23}x_3 & = & c_2 \\ & & & & u_{33}x_3 & = & c_3 \end{cases}$$

(Système triangulaire inférieur) (Système triangulaire supérieur)

Chacun des deux systèmes se résout facilement par **substitutions successives** (à condition que les éléments diagonaux soient non nuls)

$$(S_1) \qquad \begin{cases} x_1 &= & \frac{c_1}{l_{11}} \\ x_2 &= & (c_2 - l_{21}x_1)/l_{22} \\ x_3 &= & (c_3 - l_{31}x_1 - l_{32}x_2)/l_{33} \end{cases} \qquad (S_2) \qquad \begin{cases} x_3 &= & \frac{c_3}{u_{33}} \\ x_2 &= & (c_2 - u_{23}x_3)/u_{22} \\ x_1 &= & (c_1 - u_{12}x_2 - u_{13}x_3)/u_{33} \end{cases}$$

La technique de substitutions successives s'applique de la même manière aux systèmes triangulaires de n équations.

2.2 Les algorithme de substitution successives

Algorithme 4 (Substitutions successives L)**.** *Les solutions du système triangulaire* inférieur

$$LX = C \qquad avec \quad \begin{cases} l_{ij} = 0 \quad si \quad i < j, \\ l_{ii} \neq 0, \end{cases}$$

sont données par les relations

$$\begin{cases} x_1 = & \frac{c_1}{l_{11}} \\ x_i = & \frac{1}{l_{ii}}\left(c_i - \sum_{j=1}^{i-1} l_{ij}x_j\right), \qquad i = 2,3,\ldots,n. \end{cases}$$

Algorithme 5 (Substitutions successives U)**.** *Les solutions du système triangulaire* supérieur

$$UX = C \qquad avec \quad \begin{cases} u_{ij} = 0 \quad si \quad i > j, \\ u_{ii} \neq 0, \end{cases}$$

sont données par les relations

$$\begin{cases} x_n = & \frac{c_n}{u_{nn}} \\ x_i = & \frac{1}{u_{ii}}\left(c_i - \sum_{j=i+1}^{n} u_{ij}x_j\right), \qquad i = n-1, n-2,\ldots,1. \end{cases}$$

Dans les deux algorithmes, les conditions imposant $l_{ii} \neq 0$ et $u_{ii} \neq 0$, $i = 1,\ldots,n$ sont équivalents à la condition que le système est régulier. Cela provient du fait que le déterminant d'une matrice triangulaire est égal au produit des coefficients sur la diagonale de la matrice.

Théorème 6. *La résolution d'un système linéaire triangulaire (supérieur ou inférieur) de n équations à n inconnues par la méthode des substitutions successives nécessite n^2 opérations élémentaires* $(+,-,\times,\div)$.

Démonstration. Nous traitons seulement le cas d'un système triangulaire inférieur. Le calcul de x_1 requiert 1 opération tandis que pour x_i, $2 \leq i \leq n$, nous devons effectuer $i-1$ (+ ou −) et i (× ou ÷) de sorte que le nombre total d'opérations N_n est donné par

$$\begin{aligned} N_n &= 1+\sum_{i=2}^{n}(2i-1) \\ &= 1+2\sum_{i=2}^{n} i-(n-1) \\ &= 1+2\frac{n(n+1)}{2}-2-(n-1)=n^2. \quad \blacksquare \end{aligned}$$

§ 3 L'ALGORITHME DE GAUSS

3.1 Description de l'algorithme dans le cas d'un système de 3 équations à 3 inconnues. Notion de pivot

Soit à résoudre le système suivant que nous supposons régulier. Le déterminant de la matrice associée est donc supposé non nul.

$$S^{(0)} \quad \begin{matrix} \mathbf{L}_1^{(0)} \\ \mathbf{L}_2^{(0)} \\ \mathbf{L}_3^{(0)} \end{matrix} \quad \left\{ \begin{array}{lllllll} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & c_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & = & c_2 \\ a_{31}x_1 & + & a_{32}x_2 & + & a_{33}x_3 & = & c_3 \end{array} \right. .$$

Etape 1. Elimination de x_1 dans $\mathbf{L}_2^{(0)}$ et $\mathbf{L}_3^{(0)}$.

$$\left\| \begin{array}{l} \mathbf{L}_2^{(0)} \longleftarrow \mathbf{L}_2^{(0)} - \frac{a_{21}}{a_{11}}\mathbf{L}_1^{(0)} \\ \mathbf{L}_3^{(0)} \longleftarrow \mathbf{L}_3^{(0)} - \frac{a_{31}}{a_{11}}\mathbf{L}_1^{(0)} \end{array} \right. .$$

Attention, Nous divisons par a_{11}. Cela n'est possible que si a_{11} est non nul. Nous arrivons à

$$S^{(1)} \quad \begin{matrix} \mathbf{L}_1^{(0)} \\ \mathbf{L}_2^{(1)} \\ \mathbf{L}_3^{(1)} \end{matrix} \quad \left\{ \begin{array}{llll} a_{11}x_1 & +a_{12}x_2 & +a_{13}x_3 & = c_1 \\ 0 & +(a_{22}-\frac{a_{21}}{a_{11}}a_{12})x_2 & +(a_{23}-\frac{a_{21}}{a_{11}}a_{13})x_3 & = c_2-\frac{a_{21}}{a_{11}}c_1 \\ 0 & +(a_{32}-\frac{a_{31}}{a_{11}}a_{12})x_2 & +(a_{33}-\frac{a_{31}}{a_{11}}a_{13})x_3 & = c_3-\frac{a_{31}}{a_{11}}c_1 \end{array} \right. .$$

que nous écrivons encore sous la forme

$$S^{(1)} \quad \begin{matrix} \mathbf{L}_1^{(0)} \\ \mathbf{L}_2^{(1)} \\ \mathbf{L}_3^{(1)} \end{matrix} \quad \left\{ \begin{array}{lllllll} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & c_1 \\ 0 & + & a_{22}^{(1)}x_2 & + & a_{23}^{(1)}x_3 & = & c_2^{(1)} \\ 0 & + & a_{32}^{(1)}x_2 & + & a_{33}^{(1)}x_3 & = & c_3^{(1)} \end{array} \right. .$$

Etape 2. Elimination de x_2 dans $\mathbf{L}_3^{(2)}$.

$$\left\| \mathbf{L}_3^{(1)} \longleftarrow \mathbf{L}_3^{(1)} - \frac{a_{32}^{(1)}}{a_{22}^{(1)}} \mathbf{L}_2^{(1)} \right. .$$

Attention, nous divisons cette fois par $a_{22}^{(1)}$. La condition $a_{22}^{(1)} \neq$ est nécessaire. Il vient

$$S^{(2)} \quad \begin{matrix} \mathbf{L}_1^{(0)} \\ \mathbf{L}_2^{(1)} \\ \mathbf{L}_3^{(2)} \end{matrix} \left\{ \begin{array}{ccccccc} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & c_1 \\ 0 & + & a_{22}^{(1)}x_2 & + & a_{23}^{(1)}x_3 & = & c_2^{(1)} \\ 0 & + & 0 & + & (a_{33}^{(1)} - \frac{a_{32}^{(1)}}{a_{22}^{(1)}} a_{23}^{(1)})x_3 & = & c_3^{(1)} - \frac{a_{32}^{(1)}}{a_{22}^{(1)}} c_2^{(1)} \end{array} \right.$$

que nous écrivons

$$S^{(2)} \quad \begin{matrix} \mathbf{L}_1^{(0)} \\ \mathbf{L}_2^{(1)} \\ \mathbf{L}_3^{(1)} \end{matrix} \left\{ \begin{array}{ccccccc} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & c_1 \\ 0 & + & a_{22}^{(1)}x_2 & + & a_{23}^{(1)}x_3 & = & c_2^{(1)} \\ 0 & + & 0 & + & a_{33}^{(2)}x_3 & = & c_3^{(2)} \end{array} \right. .$$

Ce dernier système est triangulaire supérieur, nous pouvons donc le résoudre rapidement par substitutions successives comme expliqué dans la partie précédente. Il reste à examiner si, et comment, nous pouvons modifier la méthode dans le cas où un des nombres par lesquels nous devons diviser s'avère être égal à 0. Supposons que $a_{11} = 0$. Nous avons alors $a_{21} \neq 0$ ou $a_{31} \neq 0$ sinon la première colonne de la matrice du système serait nulle et son déterminant vaudrait 0 ce qui est contraire à l'hypothèse. Supposons pour fixer les idées que $a_{22} \neq 0$, nous permutons alors les lignes $\mathbf{L}_1^{(0)}$ et $\mathbf{L}_2^{(0)}$ et commençons la méthode décrite ci-dessus à partir du système

$$\left\{ \begin{array}{ccccccc} a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & = & c_2 \\ 0 & + & a_{12}x_2 & + & a_{13}x_3 & = & c_1 \\ a_{31}x_1 & + & a_{32}x_2 & + & a_{33}x_3 & = & c_3 \end{array} \right. .$$

Dans la deuxième étape, si nécessaire, c'est-à-dire si $a_{22}^{(1)} = 0$, nous pouvons permuter les lignes $\mathbf{L}_2^{(1)}$ et $\mathbf{L}_3^{(1)}$ de telle sorte que nous diviserons à nouveau par un nombre non nul.

Les nombres par lesquels nous effectuons les divisions dans les diverses étapes de l'algorithme s'appellent les **pivots de Gauss**. Pour que l'algorithme fonctionne, ces nombres doivent être non nuls, et, pour la précision des calculs, il est préférable qu'il ne soit pas proche de 0. Cette question ne sera pas abordée ici (cf. exercices).

Dans la partie suivante nous décrivons l'algorithme de Gauss ci-dessus dans le cas d'un système à n équations et n inconnues. L'énoncé est ne suppose pas que le système soit régulier. L'algorithme est muni d'une instruction d'arrêt pour le cas où $\det A = 0$.

E 81 Résoudre le système suivant en utilisant l'algorithme de Gauss

$$\begin{cases} x_1 + \frac{1}{2}x_2 + \frac{1}{3}x_3 = \frac{11}{6} \\ \frac{1}{2}x_1 + \frac{1}{3}x_2 + \frac{1}{4}x_3 = \frac{13}{12} \\ \frac{1}{3}x_1 + \frac{1}{4}x_2 + \frac{1}{5}x_3 = \frac{47}{60} \end{cases}$$

La solution est $(1,1,1)$.

3.2 Algorithme de Gauss (sans optimisation de pivot)

algorithme de Gauss

Algorithme 7 (Notation Ligne)**.** *On considère le système linéaire de matrice associée A*

$$\left(\mathbf{L_k}: \quad \sum_{j=1}^{n} a_{kj}x_j = c_k, \qquad k = 1,2,\ldots,n\right)$$

1 *Pour $j = 1,\ldots,n-1$ faire sauf ordre d'arrêt*

1.1 **Si** *$a_{ij} = 0$ pour tout $i \geq j$ alors ARRÊT. (A est non inversible.)*

Sinon *soit $i_o = \inf\{i, a_{ij} \neq 0\}$, faire*

$$\mathbf{L_j} \longleftarrow \mathbf{L_{i_0}}$$
$$\mathbf{L_{i_0}} \longleftarrow \mathbf{L_j}$$

1.2 *Pour $i > j$ faire*

$$\mathbf{L_i} \longleftarrow \mathbf{L_i} - \frac{a_{ij}}{a_{jj}}\mathbf{L_j}.$$

2 *Résoudre le système (triangulaire) formé des (nouvelles) lignes L_i par la méthode des substitutions successives.*

Algorithme 8 (Notation Coefficient)**.** *On considère le système linéaire de matrice associée A*

$$\left(\mathbf{L_k}: \quad \sum_{j=1}^{n} a_{kj}x_j = c_k, \qquad k = 1,2,\ldots,n\right)$$

1 *Pour $j = 1,\ldots,n-1$ faire sauf ordre d'arrêt*

1.1 **Si** *$a_{ij} = 0$ pour tout $i \geq j$ alors ARRÊT. (A est non inversible.)*

Sinon *soit* $i_o = \inf\{i, a_{ij} \neq 0\}$, *faire*

$$\begin{aligned} a_{jl} &\longleftarrow a_{i_0 l} \quad \text{pour} \quad l \geq j \\ a_{i_0 l} &\longleftarrow a_{jl} \quad \text{pour} \quad l \geq j \\ c_j &\longleftarrow c_{i_0} \\ c_{i_0} &\longleftarrow c_j \end{aligned}$$

1.2 *Pour* $i > j$ *faire*

1.2.1

$$\begin{aligned} m_i &= \frac{a_{ij}}{a_{jj}} \\ c_i &\longleftarrow c_i - m_i c_j. \end{aligned}$$

1.2.2 *pour* $k > j$ *faire*

$$a_{ik} \longleftarrow a_{ik} - m_i a_{jk}.$$

2 *Résoudre le système (triangulaire)*

$$\Big(\sum_{i=k}^{n} a_{ki} x_i = b_k \qquad k = 1, 2, \ldots, n\Big)$$

3.3 Coût de l'algorithme de Gauss

Théorème 9. *Le nombre* N_n *d'opérations élémentaires nécessaires pour résoudre un système linéaire à* n *équations et* n *inconnues (de déterminant non nul) par la méthode de Gauss est asymptotiquement égal à* $2n^3/3$. *On écrit* $N_n \sim 2n^3/3$ *et cela signifie* $\lim_{n\to\infty} \frac{N_n}{2n^3/3} = 1$.

Démonstration. Il découle de l'algorithme (version coefficients) que

$$N_n = \underbrace{\sum_{j=1}^{n-1}\left(\sum_{i=j+1}^{n}\big(3+\sum_{k=j+1}^{n}2\big)\right)}_{\text{coût de } \boxed{1}} + \underbrace{n^2}_{\text{coût de } \boxed{2}}$$

[TH 9]

où nous utilisons le théorème 6 pour déterminer le coût de $\boxed{2}$. Ensuite nous avons

$$\begin{aligned} N_n &= \sum_{j=1}^{n-1}\left(\sum_{i=j+1}^{n}\big(3+2(n-j)\big)\right)+n^2 \\ &= \sum_{j=1}^{n-1}\big(3+2(n-j)\big)(n-j)+n^2 \\ &= \sum_{j=1}^{n-1}\big(3+2j\big)(j)+n^2 \\ &= 3\frac{n(n-1)}{2}+2\sum_{j=1}^{n-1}j^2+n^2 \\ &= 3\frac{n(n-1)}{2}+2\frac{(n-1)(n)(2n-1)}{6}+n^2 \end{aligned}$$

où nous avons utilisé

$$\sum_{j=1}^{n-1}j^2=\frac{(n-1)(n)(2n-1)}{6}$$

qui se démontre aisément par récurrence sur n. Nous avons donc

$$N_n=\frac{2}{3}n^3+\boxed{?}n^2+\boxed{?}n+\boxed{?}$$

où les $\boxed{?}$'s désignent des constantes dont la valeur n'importe pas ici. IL suit que

$$\frac{N_n}{\frac{2}{3}n^3}=1+\boxed{?}\frac{1}{\frac{2}{3}n}+\boxed{?}\frac{1}{\frac{2}{3}n^2}+\boxed{?}\frac{1}{\frac{2}{3}n^3}$$

d'où il résulte immédiatement

$$\lim_{n\to\infty}\frac{N_n}{\frac{2}{3}n^3}=1.$$

■

§ 4 EXERCICES ET PROBLEMES

82 Calcul de la puissance k-ième d'une matrice. Toutes les matrices considérées dans cet exercice sont d'ordre n (i.e. dans $M_n(\mathbb{R})$.

(a) Calculer le nombre d'opérations nécessaires pour calculer le produit AB de deux matrices de $M_n(\mathbb{R})$.

(b) Quel sera le nombre d'opérations pour calculer A^k, $k \geq 2$, par récurrence à partir de $A^k = A \cdot A^{k-1}$?

(c) On reprend le calcul de A^k. Pour simplifier, on se limite au cas $k = 7$. Calculer le nombre d'opérations si on écrit $A^7 = (A^2)^2 \cdot A^2 \cdot A$. Comment généraliser cette méthode pour k quelconque ?

83 Impraticabilité de la méthode de Cramer. Calculer (en fonction de n) le nombre d'opérations élémentaires $(+, -, \times, \div)$ nécessaires pour résoudre un système linéaire de n équations à n inconnues en utilisant les formules de Cramer dans lesquelles on calcule les déterminants par la relation de récurrence

$$\det A = \sum_{i=1}^{n} (-1)^{i+n} a_{in} \det A_{in},$$

où A_{in} est la matrice obtenue en retirant de A la n-ième colonne et la i-ième ligne ?

Combien de temps prendrait la résolution d'un système linéaire à 50 inconnues et 50 équations si on utilisait un ordinateur capable d'effectuer 10^9 opérations à la seconde ?

84 Calcul de l'inverse d'une matrice triangulaire.

Soit U une matrice $n \times n$ triangulaire supérieure inversible : on a donc $u_{ij} = 0$ pour $i > j$ et $u_{ii} \neq 0$. On cherche un algorithme donnant la matrice U^{-1}.

(a) Montrer que la matrice U^{-1} est aussi triangulaire supérieure.

(b) On note v_{ij} le coefficient de U^{-1} à l'intersection de la i-ième ligne et j-ième colonne. D'après la première question on a $v_{ij} = 0$ dès que $i > j$. On note $v^{(j)}$ le vecteur $(v_{1j}, v_{2j}, \ldots, v_{jj}) \in \mathbb{R}^j$. On connaîtra donc U^{-1} dès qu'on connait les n vecteurs $v^{(1)} \in \mathbb{R}^1, v^{(2)} \in \mathbb{R}^2, \ldots, v^{(n)} \in \mathbb{R}^n$. On note enfin $U^{(j)}$ la matrices formées des j premières lignes et colonnes de U. Montrer que $v^{(j)}$ est solution du système triangulaire $U^{(j)} X = e^{(j)}$ où $e^{(j)} = (0, \ldots, 0, 1) \in \mathbb{R}^j$.

(c) En déduire qu'on peut calculer l'inverse d'une matrice triangulaire avec un nombre d'opérations N_n équivalent à $\frac{n^3}{3}$ lorsque $n \to \infty$.

85 Exemple d'instabilité de la méthode Gauss. On considère le système suivant

$$\begin{cases} 2,28101x + 1,61514y = 2,76255 \\ 1,61514x + 1,14365y = 1,95611 \end{cases}.$$

Résoudre ce système avec l'algorithme de Gauss en travaillant avec des nombres décimaux comportant un maximum de 8 décimales. La solution de ce système est $x = 9$ et $y = -11$. Que pensez vous de la précision du résultat ?

86 Un exemple. Estimer le temps nécessaire pour résoudre un système de 10^3 équations avec la méthode de Gauss sur un ordinateur capable d'effectuer 10^6 opérations à la seconde.

87 Méthode de Gauss avec optimisation de pivot. Pour que l'algorithme de Gauss fonctionne, on a vu que les pivots (les nombres par lesquels on divise au moment de l'élimination des inconnues) doivent être non nuls. L'algorithme donné dans le cours choisit le premier pivot possible

(il minimise le nombre de tests à effectuer). Or en analyse numérique, pour des questions d'erreurs d'arrondis (voir l'exercice précédent), on a intérêt à toujours diviser par les nombres les plus grands possibles (en valeur absolue).

Modifier l'algorithme de Gauss de telle sorte que le j-ème pivot soit choisi comme le plus grand nombre disponible sur la j-ième colonne.

88 Matrices triangulaires creuses On considère un système linéaire à n équations et n inconnues, $n \geq 4$, triangulaire supérieur, de la forme

$$\left\{\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & c_1 \\ a_{22}x_2 & + & a_{23}x_3 & + & a_{24}x_4 & = & c_2 \\ \cdots & & & & & & \\ a_{ii}x_i & + & a_{i,i+1}x_{i+1} & + & a_{i,i+2}x_{i+2} & = & c_i \\ \cdots & & & & & & \\ a_{n-2,n-2}x_{n-2} & + & a_{n-2,n-1}x_{n-1} & + & a_{n-2,n}x_n & = & c_{n-2} \\ & & a_{n-1,n-1}x_{n-1} & + & a_{n-1,n}x_n & = & c_{n-1} \\ & & & & a_{nn}x_n & = & c_n \end{array}\right. \tag{4.1}$$

où les coefficients a_{ii}, $i = 1, \ldots, n$ sont supposés non nuls.

(a) Ecrire l'algorithme de résolution par substitutions successives correspondant à ce cas particulier de matrice triangulaire.

(b) Déterminer, en fonction de n le nombre d'opérations élémentaires $(+, -, \times, \div)$ employées pour la résolution du système (4.1).

(Sol. 13 p. 105.)

89 Aspect matriciel de l'algorithme de Gauss. On considère le système 3×3 suivant dont on suppose qu'il admet une et une seule solution

$$S^{(0)} \quad \left\{\begin{array}{ccccccc} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & c_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & = & c_2 \\ a_{31}x_1 & + & a_{32}x_2 & + & a_{33}x_3 & = & c_3 \end{array}\right. .$$

On rappelle que les systèmes $S^{(1)}$ (resp. $S^{(2)}$) obtenus après la première (resp. la seconde) étape de l'algorithme sont donnés par

$$S^{(1)} \quad \left\{\begin{array}{llll} a_{11}x_1 & +a_{12}x_2 & +a_{13}x_3 & = c_1 \\ 0 & +(a_{22} - \frac{a_{21}}{a_{11}}a_{12})x_2 & +(a_{23} - \frac{a_{21}}{a_{11}}a_{13})x_3 & = c_2 - \frac{a_{21}}{a_{11}}c_1 \\ 0 & +(a_{32} - \frac{a_{31}}{a_{11}}a_{12})x_2 & +(a_{33} - \frac{a_{31}}{a_{11}}a_{13})x_3 & = c_3 - \frac{a_{31}}{a_{11}}c_1 \end{array}\right.$$

abrégé en

$$S^{(1)} \quad \left\{\begin{array}{ccccccc} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & c_1 \\ 0 & + & a_{22}^{(1)}x_2 & + & a_{23}^{(1)}x_3 & = & c_2^{(1)} \\ 0 & + & a_{32}^{(1)}x_2 & + & a_{33}^{(1)}x_3 & = & c_3^{(1)} \end{array}\right.$$

et

$$S^{(2)} \quad \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = c_1 \\ 0 + a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 = c_2^{(1)} \\ 0 + 0 + (a_{33}^{(1)} - \frac{a_{32}^{(1)}}{a_{22}^{(1)}}a_{23}^{(1)})x_3 = c_3^{(1)} - \frac{a_{32}^{(1)}}{a_{22}^{(1)}}c_2^{(1)} \end{cases}$$

abrégé en

$$S^{(2)} \quad \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = c_1 \\ 0 + a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 = c_2^{(1)} \\ 0 + 0 + a_{33}^{(2)}x_3 = c_3^{(2)} \end{cases}.$$

Hypothèse. On a supposé que les termes par lesquels on divise sont non nuls.

(a) On appelle $A = A^{(0)}$ la matrice du système $S^{(0)}$, $A^{(1)}$ la matrice du système $S^{(1)}$ et $A^{(2)}$ la matrice du système $S^{(2)}$. Ecrire les trois matrices $A^{(0)}$, $A^{(1)}$, $A^{(2)}$.

(b) Vérifier que

$$A^{(1)} = L_1 A^{(0)} \quad \text{et} \quad A^{(2)} = L_2 A^{(1)}$$

avec

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{a_{21}}{a_11} & 1 & 0 \\ -\frac{a_{31}}{a_{11}} & 0 & 1 \end{pmatrix} \quad \text{et} \quad L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{a_{32}^{(1)}}{a_{22}^{(1)}} & 1 \end{pmatrix}$$

(c) En déduire que l'algorithme de Gauss permet, sous réserve que l'hypothèse ci-dessus soit satisfaite, d'obtenir une factorisation de A de la forme $A = LR$ avec L une matrice triangulaire inférieure et R une matrice triangulaire supérieure.

(d) Expliquer comment un obtiendrait on résultat similaire en partant d'une matrice $n \times n$, $n \geq 2$. (On expliquera en particulier quelles matrices joueraient le rôle de L_1 et L_2 dans le cas n quelconque.)

(Sol. 14 p. 106.)

90 Systèmes tridiagonaux

Soient $n \in \mathbb{N}$, $n \geq 2$ et $a = (a_1, a_2, \ldots, a_n)$, $b = (b_2, b_3, \ldots, b_n)$ et $c = (c_1, c_2, \ldots, c_{n-1})$ trois suites finies de nombres réels.

Algorithme 10. *Il calcule les deux suites $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ et $\beta = (\beta_2, \beta_3, \ldots, \beta_n)$ comme suit :*

(a) $\alpha_1 = a_1$,

(b) Pour $i = 2, \ldots, n$ faire

  (a) $\beta_i = \frac{b_i}{\alpha_{i-1}}$

  (b) $\alpha_i = a_i - \beta_i c_{i-1}$

On supposera les suites a, b et c données de telle sorte que α_i ne s'annule jamais.

[TH 10]

(a) Déterminer en fonction de n le nombre d'opérations employées par l'algorithme ci-dessus pour obtenir les deux suites α et β.
On définit les matrices L et U à partir des suites α et β de la manière suivante :

$$L = \begin{pmatrix} 1 & & & 0 \\ \beta_2 & 1 & & \\ & \ddots & \ddots & \\ 0 & & \beta_n & 1 \end{pmatrix} \quad \text{et} \quad U = \begin{pmatrix} \alpha_1 & c_1 & & 0 \\ & \alpha_2 & \ddots & \\ & & \ddots & c_{n-1} \\ 0 & & & \alpha_n \end{pmatrix} \tag{4.2}$$

Tous les coefficients de L sont donc nuls excepté i) les coefficients diagonaux qui sont égaux à 1 et ii) les coefficients en dessous de la diagonale données par β. De la même manière, tous les coefficients de U sont nuls excepté i) les termes de la diagonales qui sont donnés par α et ii) les termes au dessus de la diagonales qui sont donnés par c.

(b) On note $A = L.U$. Démontrer que

$$A = \begin{pmatrix} a_1 & c_1 & & & & 0 \\ b_2 & a_2 & c_2 & & & \\ & b_3 & a_3 & c_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & b_{n-1} & a_{n-1} & c_{n-1} \\ 0 & & & & b_n & a_n \end{pmatrix}$$

La matrice A est appelée **matrice tridiagonale**. (Tous les coefficients sont nuls excepté sur la diagonale, et juste au dessus et juste au dessous de la diagonale.)

(c) Montrer que résoudre le système $Ax = d$ (d'inconnue x où d est un vecteur quelconque) est *équivalent* à résoudre les systèmes $Ly = d$ (d'inconnue y) *puis* $Ux = y$ (d'inconnue x).

(d) Montrer que la résolution du système $Ly = d$ par substitutions nécessite $2(n-1)$ opérations.

(e) Montrer que la résolution du système $Ux = y$ par substitutions nécessite $3n-2$ opérations.

(f) En combien d'opérations en tout (en fonction de n) peut-on résoudre un système $Ax = d$ où A est une matrice tridiagonale à n lignes et n colonnes ?

(sol. 15 p. 107.)

91 Méthode de Cholesky pour les matrices tridiagonales symétriques. On considère une ma-

trice carrée A d'ordre $p \geq 3$ à coefficients réels de la forme

$$A = \begin{pmatrix} b_1 & c_1 & & & & 0 \\ c_1 & b_2 & c_2 & & & \\ & c_2 & b_3 & c_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & c_{p-2} & b_{p-1} & c_{p-1} \\ 0 & & & & c_{p-1} & b_p \end{pmatrix} \tag{4.3}$$

Tous les coefficients sont nuls excepté sur la diagonale, et juste au dessus et juste au dessous de la diagonale. On définit ensuite les réels d_j $(j = 1, \ldots, p)$ et f_j $(j = 1, \ldots, p-1)$ par les relations

$$d_1 = \sqrt{b_1}, \quad f_1 = c_1/d_1; \tag{4.4}$$

$$d_j = \sqrt{b_j - f_{j-1}^2}, \quad j = 2, \ldots, p; \quad f_j = c_j/d_j, \quad j = 2, \ldots, p-1. \tag{4.5}$$

Ici, on suppose que les nombres dont on prend les racines carrées sont positifs.

(a) Etude d'un exemple. (a) Calculer les nombres d_j et f_j lorsque $p = 3$ et

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 5 & 4 \\ 0 & 4 & 5 \end{pmatrix}.$$

(b) Montrer que dans ce cas on a

$$A = \begin{pmatrix} d_1 & 0 & 0 \\ f_1 & d_2 & 0 \\ 0 & f_2 & d_3 \end{pmatrix} \begin{pmatrix} d_1 & f_1 & 0 \\ 0 & d_2 & f_2 \\ 0 & 0 & d_3 \end{pmatrix}.$$

(c) En déduire une résolution rapide du système $AX = C$ où $C = (1, -3, -5)$.

(b) On traite maintenant le cas général où p est quelconque et A est comme dans l'équation (4.3). (a) Déterminer le nombre d'opérations nécessaires pour calculer tous les nombres d_j et f_j. Les opérations sont $+$, $-$, $\times$, $\div$ et $\sqrt{\cdot}$. (b) Démontrer que que si S et ${}^tS$ sont les matrices

$$S = \begin{pmatrix} d_1 & f_1 & & & 0 \\ & d_2 & f_2 & & \\ & & \ddots & \ddots & \\ & & & d_{p-1} & f_{p-1} \\ 0 & & & & d_p \end{pmatrix} \quad \text{et} \quad {}^tS = \begin{pmatrix} d_1 & & & & 0 \\ f_1 & d_2 & & & \\ & \ddots & \ddots & & \\ & & f_{p-2} & d_{p-1} & \\ 0 & & & f_{p-1} & d_p \end{pmatrix}$$

alors

$$A = {}^tS \cdot S.$$

(c) En déduire une méthode simple pour résoudre les systèmes $AX = C$ et déterminer le nombre d'opérations utilisées par cette méthode.

92 Méthode de Jordan. On considère le système

$$S^{(0)} \quad \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = c_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = c_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = c_3 \end{cases}.$$

La première étape est identique à celle de la méthode de Gauss et, sous réserve que $a_{11} \neq 0$ elle conduit au système

$$S^{(1)} \quad \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = c_1 \\ 0 + a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 = c_2^{(1)} \\ 0 + a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 = c_3^{(1)} \end{cases}.$$

(a) Rappeler l'expression de $a_{32}^{(1)}$ en fonction des coefficients du système $S^{(0)}$.

(b) Montrer qu'en effectuant deux opérations sur les lignes et sous réserve que $a_{22}^{(1)} \neq 0$, le système $S^{(1)}$ est équivalent à un système $S^{(2)}$ de la forme

$$S^{(2)} \quad \begin{cases} a_{11}x_1 + 0 + a_{13}^{(2)}x_3 = c_1^{(2)} \\ 0 + a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 = c_2^{(1)} \\ 0 + 0 + a_{33}^{(2)}x_3 = c_3^{(2)} \end{cases}.$$

et donner l'expression de $a_{33}^{(2)}$.

(c) Montrer qu'en effectuant à nouveau deux opérations sur les lignes et sous réserve que $a_{33}^{(2)} \neq 0$, le système $S^{(2)}$ est équivalent à un système $S^{(3)}$ diagonal de la forme

$$S^{(3)} \quad \begin{cases} a_{11}x_1 + 0 + 0 = c_1^{(3)} \\ 0 + a_{22}^{(1)}x_2 + 0 = c_2^{(3)} \\ 0 + 0 + a_{33}^{(2)}x_3 = c_3^{(2)} \end{cases}.$$

et donner l'expression de $c_2^{(3)}$.

(d) Quel est l'intérêt de cet algorithme ? Comment le modifier pour traiter le cas où l'une des hypothèses de coefficients $\neq 0$ n'est pas vérifiée ?

(e) Ecrire un algorithme (en notation ligne) effectuant le travail ci-dessus dans le cas d'un système linéaire de n équations et n inconnues.

93 Algorithme de Cholesky. On étudie une méthode de résolution directe des systèmes linéaires $Ax = b$ lorsque la matrice A peut s'écrire comme le produit d'une matrice triangulaire par sa transposée. Toutes les définitions et propriétés de la transposée qui pourront être utiles sont indiquées dans l'énoncé.

Si A est la matrice dont le coefficient (i, j) est a_{ij}, la matrice transposée de A, notée $T(A)$, est la matrice dont le coefficient (i, j) est a_{ji}, autrement dit $(T(A))_{ij} = a_{ji}$: on permute donc le rôle des lignes et des colonnes et la i-ème ligne de A devient la i-ème colonne de $T(A)$. Par

exemple,

$$\text{si } A = \begin{pmatrix} 1 & 0 & 3 \\ 6 & 2 & 8 \\ -1 & -2 & 4 \end{pmatrix} \text{ alors } T(A) = \begin{pmatrix} 1 & 6 & -1 \\ 0 & 2 & -2 \\ 3 & 8 & 4 \end{pmatrix}.$$

On pourra librement utiliser les propriétés suivantes :

(a) $T(A \cdot B) = T(B) \cdot T(A)$

(b) $\det(T(A)) = \det(A)$

(c) $T(T(A)) = A$.

On remarquera en outre que si D est une matrice diagonale alors $T(D) = D$. Plus généralement A et $T(A)$ ont toujours la même diagonale.

A) Soit L une matrice triangulaire inférieure c'est-à-dire de la forme

$$L = \begin{pmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & & \\ \vdots & \ddots & \ddots & 0 \\ l_{n1} & \dots & l_{nn-1} & l_{nn} \end{pmatrix}$$

Tous les coefficients sont nuls sauf éventuellement les coefficients sur la diagonale et *en dessous* de la diagonale. Représenter la matrice $T(L)$. De quel type de matrice s'agit-il ? Que vaut le déterminant de L ?

$\boxed{H}$ *: A partir de maintenant, A désigne une matrice de $M_n(\mathbb{R})$ telle que (a) A est inversible et (b) $A = L \cdot T(L)$ où L est une matrice triangulaire inférieure.*

QUESTIONS PRÉLIMINAIRES D'ALGÈBRE LINÉAIRE.

B) Montrer que $T(A) = A$.

C) Montrer que, pour $k = 1, \dots, n$, on a $l_{kk} \neq 0$. (On rappelle que le déterminant d'un produit de matrices est égal au produit des déterminants des matrices).

D) Montrer que si D est une matrice diagonale avec uniquement des 1 ou des -1 sur la diagonale alors on a encore $A = L' \cdot T(L')$ avec $L' = L \cdot D$. En déduire que, sous l'hypothèse **H**, on peut toujours écrire $A = L' \cdot T(L')$ avec $L' = (l'_{ij})$ une matrice triangulaire inférieure telle que $l'_{kk} > 0$ pour $k = 1, \dots n$.

E) Montrer que résoudre le système $Ax = b$ est équivalent à résoudre les deux systèmes $L(y) = b$ et $T(L)x = y$. En combien d'opérations $(+, -, \times, \div)$ peut-on résoudre ces deux systèmes ?

F) Montrer que

$$\sum_{k=2}^{n} (n-k+1)(k-1) = \frac{n(n^2-1)}{6}.$$

On pourra librement utiliser le fait que $\sum_{j=1}^{n-1} j^2 = \frac{(n-1)n(2n-1)}{6}$.

L'ALGORITHME.

Dans cette partie, toujours sous l'hypothèse **H**, nous étudions une méthode, dite de Cholesky, pour déterminer L telle que

$$A = L \cdot T(L) \quad \text{et} \quad l_{kk} > 0 \ (k = 1, \dots, n). \tag{4.6}$$

Les colonnes de L seront déterminées par récurrence.

G) Montrer à l'aide de (4.6), que pour $j \leq i$ on a $a_{ij} = \sum_{s=1}^{j} l_{is} l_{js}$.

H) En déduire que $l_{11} = \sqrt{a_{11}}$ puis $l_{i1} = \frac{a_{i1}}{l_{11}}$ pour $i = 2, \dots, n$.

I) On suppose que l'on a construit les $k-1$ premières colonnes de L. Montrer, à l'aide de (G), que $l_{kk} = \sqrt{a_{kk} - \sum_{s=1}^{k-1} l_{ks}^2}$.

J) Montrer

$$l_{ik} = \frac{a_{ik} - \sum_{s=1}^{k-1} l_{is} l_{ks}}{l_{kk}} \quad (i = k+1, \dots, n).$$

K) Appliquer la méthode décrite dans les trois question précédentes pour trouver la matrice L dans le cas où

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix}.$$

Vérifier que pour la matrice L trouvée, on a bien $A = L \cdot T(L)$.

NOMBRE D'OPÉRATIONS.

On determine le nombre d'opérations élémentaires $+, -, \times, \div$ et aussi la racine carrée $\sqrt{\ }$ employees par l'algorithme de Cholesky.

L) Déterminer le nombre de racines carrées puis le nombre de divisions employées par l'algorithme de Cholesky.

M) Montrer que le nombre d'additions-soustractions employées par l'algorithme est égal à $\frac{n(n^2-1)}{6}$. Montrer ensuite que le nombre de multiplications employées par l'algorithme est égal à $\frac{n(n^2-1)}{6}$.

N) La méthode de Cholesky pour résoudre le système $Ax = b$ sous l'hypothèse **H** consiste à déterminer une matrice L puis à utiliser utiliser la propriété établie ci-dessus (E). Cette méthode est-elle plus performante que la méthode de Gauss ?

(Sol. 16 p. 108.)

V

SOLUTION DES EXERCICES

§ 1 Sur l'interpolation de Lagrange

1 (← 12.) Appelons P le polynôme d'interpolation $\mathbf{L}[a_0, a_1, a_2; f]$. D'après la formule d'interpolation de Lagrange, on a

$$\begin{array}{rcl} P(115) & = & 10\frac{115-121)(115-144)}{(100-121)(100-144)} + 11\frac{(115-100)(115-144)}{(121-100)(121-144)} + 12\frac{(115-100)(115-121)}{(144-100)(144-121)} \\ & \approx & 10,722753. \end{array}$$

D'après un théorème du cours, il existe $\zeta \in]100, 144[$ tel que

$$\sqrt{115} - P(115) = \frac{f^{(3)}(\zeta)}{3!}(115-100)(115-121)(115-144).$$

Or $f(x) = \sqrt{x} \implies f'(x) = \frac{1}{2}x^{-1/2} \implies f''(x) = \frac{1}{2}\frac{-1}{2}x^{-3/2} \implies f'''(x) = \frac{1}{2}\frac{-1}{2}\frac{-3}{2}x^{-5/2}$. Il suit, en utilisant que $\left|f'''\right|$ est décroissante sur $[100, 144]$ que

$$\left|f'''(\zeta)\right| \leq \frac{3}{2^3}100^{-5/2} = \frac{3}{8\cdot 10^5}.$$

En reportant dans l'inégalité précédente on arrive à

$$\left|\sqrt{115} - P(115)\right| \leq \frac{3}{3!\cdot 8\cdot 10^5}15\cdot 6\cdot 29 = \frac{15\cdot 6\cdot 29}{16\cdot 10^5} \approx 1,63\cdot 10^{-3} < 1,8\cdot 10^{-3}.$$

2 (← 31)

(a) Une application de la formule d'interpolation de Lagrange donne

$$\alpha = \mathbf{L}[0,1/6,1/4;f](1/5) = f(0)\frac{(1/5-1/6)(1/5-1/4)}{(-1/6)(-1/4)} + f(1/6)\frac{(1/5-0)(1/5-1/4)}{(1/6-0)(1/6-1/4)} + f(1/4)\frac{(1/5-0)(1/5-1/6)}{(1/4-0)(1/4-1/6)}.$$

Après simplification,

$$\alpha = -24/600 + \sqrt{3}\times 72/200 + \sqrt{2}\times 48/300 = 0,8098...$$

La valeur exacte de $f(1/5) = \cos(\pi/5)$ est $0,80901.....$

(b) En utilisant le théorème d'erreur du cours on a

$$|\cos(\pi/5) - \alpha| \leq \frac{\sup_{[0,1/4]}|f^{(3)}|}{3!}|1/5-0|\cdot|1/5-1/6|\cdot|1/5-1/4|.$$

Or, sur $[0,1/4]$, on a $|f^{(3)}(x)| = \pi^3\sin(\pi x) \leq \pi^3\sin(\pi/4) = \pi^3\sqrt{2}/2$. On en déduit que

$$|\cos(\pi/5) - \alpha| \leq \pi^3\sqrt{2}/(2\cdot 3!)\times(1/5)(1/20)(1/30) \approx 1,22\cdot 10^{-3}.$$

On remarquera que l'erreur réelle est sensiblement plus petite.

3 (← 32.)

(a) On a $q(\lambda) = 0$ donc λ est racine de q donc le polynôme $r(x)$ divise $q(x)$ c'est-à-dire $q(x) = r(x)T(x)$ avec T polynôme avec $\deg(T) = \deg(q) - 1 = \deg w - 1 = d$. Il suit que $f_\lambda = \frac{1}{w(\lambda)}T$ est un polynôme de degré d.

(b) Puisque a_i est racine de w on a

$$f(a_i) = \frac{w(\lambda) - w(a_i)}{w(\lambda)(\lambda - a_i)} = \frac{1}{\lambda - a_i}.$$

On a donc que f_λ est un polynôme de degré d qui prend les mêmes valeurs que $g_\lambda(t) = \frac{1}{\lambda - t}$ pour $t = a_i$, i=0,...,d. Il suit $f_\lambda = \mathbf{L}[a_0,\ldots,a_d;g_\lambda]$.

4 (← 22). On cherche $p(x) = c_0 + c_1x + c_2x^2$ (les inconnues sont les coefficients c_0, c_1 et c_2) tel que

$$\begin{cases} p(a) = \alpha \\ p'(b) = \beta \\ p''(c) = \gamma \end{cases} \Leftrightarrow \begin{cases} c_0 + c_1a + c_2a^2 = \alpha \\ c_1 + 2c_2b = \beta \\ 2c_2 = \gamma \end{cases}$$

Ce système admet une et une seule solution si son déterminant est non nul. Or ce déterminant est donné par

$$D = \begin{vmatrix} 1 & a & a^2 \\ 0 & 1 & 2b \\ 0 & 0 & 2 \end{vmatrix} = 2 \neq 0$$

5 (← 29.)

(a) Puisque X contient n points on a $\mathbf{L}[X; f/q] \in \mathscr{P}_{n-1}$ et on a aussi $q \in \mathscr{P}_m$. Il suit que $q.\mathbf{L}[X; f/q] \in \mathscr{P}_{m+n-1}$. On montre de la même manière que $p.\mathbf{L}[Y; f/p] \in \mathscr{P}_{m+n-1}$ d'où il résulte que $R_f \in \mathscr{P}_{m+n-1}$ comme somme de deux polynômes de $\mathscr{P}_{m+n-1}$.

(b) Puisque $x_i \in X$ on a $\mathbf{L}[X; f/q](x_i) = f(x_i)/q(x_i)$ et $p(x_i) = 0$ donc

$$R_f(x_i) = q(x_i).\mathbf{L}[X; f/q](x_i) + 0.\mathbf{L}[X; f/p](x_i) = q(x_i)\frac{f(x_i)}{q(x_i)} = f(x_i).$$

On montre de même que $R_f(y_j) = f(y_j)$.

(c) Les deux questions précédentes montrent que R_f satisfait les deux conditions caractéristiques de $\mathbf{L}[X \cup Y; f]$. On a donc $R_f = \mathbf{L}[X \cup Y; f]$.

6 (← 35.)

A) Puisque x est compris entre a et a_0, la distance entre x et a_0 est plus petite que $a_0 - a$ c'est-à-dire $|x - a_0| \leq h_0$. Ensuite $|x - a_i| \leq |x - a_0| + |a_0 - a_1| + \cdots + |a_{i-1} - a_i| \leq h_0 + h_1 + \ldots h_i$. On en déduit en majorant chacun des facteurs de $|w_A(x)| = |x - a_0||x - a_1||x - a_2||x - a_3||x - a_4|$ que

$$|w_A(x)| \leq h_0 \times (h_0 + h_1) \times (h_0 + h_1 + h_2) \times (h_0 + h_1 + h_2 + h_3) \times (h_0 + h_1 + h_2 + h_3 + h_4).$$

Il découle immédiatement

$$|w_A(x)| \leq 5!h^5$$

puisque $h_0 + \cdots + h_i \leq (i+1)h$.

B) Lorsque $x \in]a_0, a_1[$ on a à la fois $|x - a_0| \leq h_1$ et $|x - a_1| \leq h_1$ tandis que $|x - a_2| \leq |x - a_1| + |a_1 - a_2| \leq h_1 + h_2$ et plus généralement pour $i > 1$, $|x - a_i| \leq h_1 + \cdots + h_i$ de sorte que

$$|w_A(x)| \leq h_1 \times h_1 \times (h_1 + h_2) \times (h_1 + h_2 + h_3) \times (h_1 + h_2 + h_3 + h_4).$$

La déduction $|w_A(x)| \leq 4!h^5$ se déduit immédiatement comme dans la question précédente.

C) Les quatre inégalités demandées se traitent de manière similaire. Nous donnerons les détails de la démonstration de la troisième : si $x \in]a_3, a_4[$ alors $|w_A(x)| \leq 1! \cdot 4! \cdot h^5$. On remarque d'abord que, puisque $x \in]a_3, a_4[$, $|x - a_3| \leq h_4$ et $|x - a_4| \leq h_4$. Il reste à majorer $|x - a_0|$, $|x - a_1|$ et $|x - a_2|$. On a $|x - a_2| \leq |x - a_3| + |a_3 - a_2| \leq h_4 + h_3$. De même $|x - a_1| \leq h_2 + h_3 + h_4$ et $|x - a_0| \leq h_1 + h_2 + h_3 + h_4$. On en tire facilement l'inégalité demandée sur $w_A(x)$

D) Pour déduire

$$\max_{x \in [a,b]} |w_A(x)| \leq 5!h^5$$

il suffit de remarquer que les inégalités précédentes nous permettent de majorer $|w_A(x)|$ sur $[a, a_0] \cup]a_0, a_1| \cup \cdots \cup]a_4, b]$ donc sur $[a, b]$ — au points de passages a_i la fonction w_A s'annule — par le plus mauvais des majorants trouvés qui n'est autre que $5!h^5$.

La formule d'erreur pour l'interpolation de Lagrange donne pour toute fonction f de classe C^5 sur $[a,b]$ et tout $x \in [a,b]$, on a

$$|f(x) - \mathbf{L}[a_0,a_1,a_2,a_3,a_4;f](x)| \leq \frac{\sup_{x\in[a,b]} |f^{(5)}|}{5!} \cdot |w_A(x)| \leq \frac{\sup_{x\in[a,b]} |f^{(5)}|}{5!} \cdot 5! h^5 = \sup_{x\in[a,b]} |f^{(5)}| \cdot h^5.$$

E) Soit $i \in \{0,\dots,3\}$. La fonction $p(x) = (x-a_i)(x-a_{i+1})$ est négative sur $[a_i,a_{i+1}]$ et p' s'annule au point $m_i = (a_i+a_{i+1})/2$ qui est un minimum, avec

$$p(m_i) = -\big((a_{i+1}-a_i)/2\big)^2 = -h_{i+1}^2/4.$$

On en déduit

$$\begin{aligned} \sup_{x\in[a_i,a_{i+1}]} |(x-a_i)(x-a_{i+1})| &= \sup_{x\in[a_i,a_{i+1}]} -p(x) \\ &= -\inf_{x\in[a_i,a_{i+1}]} p(x) = -p(m_i) = h_{i+1}^2/4. \end{aligned}$$

F) Soit $x \in [a_0,a_1]$. Pour majorer $|w_A(x)|$ on majore d'abord le facteur $|(x-a_0)(x-a_1)|$ puis les facteurs $|x-a_2|$, $|x-a_3|$ et $|x-a_4|$. Pour le premier, on utilise l'inégalité obtenue à la question précédente $|(x-a_0)(x-a_1)| \leq h_1^2/4$ tandis que les trois autres termes sont majorés comme dans la première partie, par exemple $|x-a_2| \leq |x-a_1| + |a_1-a_2| \leq h_1+h_2$ puisque $x \in [a_0,a_1]$. Au total on a

$$|w_A(x)| \leq \frac{h_1^2}{4} \times (h_1+h_2) \times (h_1+h_2+h_3) \times (h_1+h_2+h_3+h_4) \leq 4! \frac{h^5}{4}.$$

G) Les inégalités obtenues sur les autres intervalles $[a_1,a_2]$, $[a_2,a_3]$ et $[a_3,a_4]$ utilisent la même technique. Il faut simplement faire attention au "facteur double" que l'on va garder : si $x \in [a_i,a_{i+1}]$ on majore le facteur double $|(x-a_i)(x-a_{i+1})|$ en utilisant un résultat ci-dessu. En réunissant les 4 inégalités on obtient une inégalité sur $[a,b]$ en prenant comme majorant le plus mauvais des quatre qui est cette fois $4!/4h^5$ de sorte que $\max_{x\in[a,b]} |w_A(x)| \leq 4! \frac{h^4}{4}$. En le portant dans la formule d'erreur pour l'interpolant de Lagrange on arrive à

$$|f(x) - \mathbf{L}[a_0,a_1,a_2,a_3,a_4;f](x)| \leq \sup_{x\in[a,b]} |f^{(5)}| \cdot \frac{h^5}{4\times 5}.$$

H) Toutes les majorations s'étendent facilement au cas où $A = \{a_0,\dots,a_n\} \subset [a,b]$ avec $a_i < a_{i+1}$ pour $i = 0,\dots,n-1$. On continue à définir $h_i = a_{i+1} - a_i$ avec les valeurs particulières h_0 et h_n. Par exemple si $x \in [a,a_0]$ alors on a $|x-a_0| \leq h_0$ et

$$|x-a_i| \leq |x-a_0| + |a_1-a_0| + \cdots + |a_{i-1}-a_i| \leq \sum_{j=0}^{i} h_i \leq (i+1)h, \quad i = 1,\dots,n.$$

On en déduit $|w_A(x)| \le h(2h)(3h)\dots((n+1)h) = (n+1)!h^{n+1}$. En examinant les autres intervalles on se rend compte que cette majoration est la plus grossière et $|w_A(x)| \le (n+1)!h^{n+1}$ sur $[a,b]$ d'où encore pour toute fonction de classe C^{n+1},

$$|f(x) - \mathbf{L}[a_0, a_1, \dots, a_n; f](x)| \le \sup_{x\in[a,b]} |f^{(n+1)}| \cdot h^{n+1}.$$

Les autres questions se généralisent suivant les mêmes lignes.

§ 2 CALCUL APPROCHÉ DES INTÉGRALES

7 (← 52.)

(a) L'inégalité à démontrer est

$$\frac{1}{ta+(1-t)b} \le \frac{t}{a} + \frac{1-t}{b}.$$

En réduisant au même dénominateur, on montre que cette inégalité est équivalente à

$$\begin{aligned}
&\iff \frac{1}{ta+(1-t)b} \le \frac{t}{a} + \frac{1-t}{b} \\
&\iff ab \le (tb+(1-t)a)(ta+(1-t)b) \\
&\iff 0 \le [t^2+(1-t)^2-1]ab + t(1-t)(b^2+a^2)
\end{aligned}$$

On vérifie ensuite aisément que le terme de droite n'est autre que $t(1-t)(a-b)^2$ qui est bien positif puisque les trois facteurs sont positifs.

(b) Puisque $a \le x \le b$ on a $b-x \ge 0 \Rightarrow \frac{b-x}{b-a} \ge 0$. D'autre part,

$$x \ge a \Rightarrow -x \le -a \Rightarrow b-x \le b-a \Rightarrow \frac{b-x}{b-a} \le 1.$$

Prenant $t = \frac{b-x}{b-a}$ dans (6.2) on obtient en tenant compte que $(1-t) = 1 - \frac{b-x}{b-a} = \frac{x-a}{b-a}$,

$$f\left(a.\frac{b-x}{b-a} + b.\frac{x-a}{b-a}\right) \le f(a)\frac{b-x}{b-a} + f(b)\frac{x-a}{b-a}.$$

Le terme de droite est la formule d'interpolation de Lagrange pour $\mathbf{L}[a,b;f](x)$ tandis que l'argument de f dans le premier membre est égal à x. On a donc montré

$$f(x) \le \mathbf{L}[a,b;f](x).$$

(c) On travaille avec les intervalles $[1,3/2]$ et $[3/2,2]$. L'approximation est donnée par

$$\frac{1}{4}(f(1)+2f(3/2))+f(2)) = \frac{1}{4}(1+4/3+1/2) = 17/24 \approx 0,708.$$

(d) On abrège valeur exacte en VE et valeur approchée en VA. Dans la méthode des trapèzes combinée on a en utilisant une inégalité prouvée ci-dessus

$$VE = \int_1^2 f(x)dx = \sum \int_{a_i}^{a_{i+1}} f(x)dx \leq \sum \int_{a_i}^{a_{i+1}} \mathbf{L}[a_i, a_{i+1}; f](x)dx = VA.$$

(e) L'erreur commise en subdivisant en n sous-intervalles est majorée par

$$\frac{(b-a)^5}{2880}.\sup_{[a,b]} f^{(4)}.$$

Ici, $b-a=1$ et, puisque $f(x)=1/x$, on a $f'(x)=-x^{-2}$, $f^{(2)}(x)=2x^{-3}$, $f^{(3)}(x)=-6x^{-4}$ et $f^{(4)}(x)=24x^{-5}$ d'où l'on déduit

$$\sup_{[a,b]} \left|f^{(4)}\right| = 24.$$

Pour commettre une erreur inférieure ou égale à 10^{-10} il suffit de choisir n tel que

$$\frac{24}{2880n^4} \leq 10^{-10}.$$

La plus petite valeur acceptable est $n=96$. (La valeur trouvée est en réalité très pessimiste.)

8 (← 54)

(a) D'après le cours, on a

$$\left|I - A(n,f)\right| \leq \frac{1}{2880n^4} \sup_{[0,1]} \left|f^{(4)}\right|.$$

Calculons les dérivées de $f(x)=e^{x^2}$. On a $f'(x)=2xf(x)$ puis $f''(x)=2f(x)+4x^2f(x)=2(1+2x^2)f(x)$. Ensuite $f^{(3)}(x)=2(4x)f(x)+2(1+2x^2)2xf(x)=(12x+8x^3)f(x)$. Finalement, $f^{(4)}=(12+24x^2)f(x)+(24x^2+16x^4)f(x)=(12+48x^2+16x^4)f(x)$. Il suit que

$$\sup_{[0,1]} \left|f^{(4)}\right| \leq (12+48+16)\cdot e = 76\cdot e.$$

Pour avoir la propriété demandée il suffit donc d'avoir

$$\frac{1}{2880n^4} 76\cdot e \leq 10^{-3}.$$

On vérifie immédiatement que $n=3$ est la plus petite valeur de n satisfaisant la condition.

(b) D'après le cours on

$$A(n,f) = \frac{h_n}{6}\left\{f(a)+f(b)+2\sum_{i=1}^{n-1} f(a+ih_n)+4\sum_{i=0}^{n-1} f(a+\frac{2i+1}{2}h_n)\right\}$$

où $a=0$, $b=1$ et $h_n=\frac{b-a}{n}$. De même,

$$A(n,\tilde{f})=\frac{h_n}{6}\left\{\tilde{f}(a)+\tilde{f}(b)+2\sum_{i=1}^{n-1}\tilde{f}(a+ih_n)+4\sum_{i=0}^{n-1}\tilde{f}(a+\frac{2i+1}{2}h_n)\right\}$$

d'où l'on déduit

$$\begin{aligned}&A(n,f)-A_n(\tilde{f})=\\&\tfrac{h_n}{6}\left[(f-\tilde{f})(a)+(f-\tilde{f})(b)+2\sum_{i=1}^{n-1}(f-\tilde{f})(a+ih_n)+4\sum_{i=0}^{n-1}(f-\tilde{f})(a+\tfrac{2i+1}{2}h_n)\right]\end{aligned}$$

puis, en utilisant que la valeur absolue d'une somme est majorée par la somme des valeurs absolues

$$\left|A(n,f)-A_n(\tilde{f})\right|\leq\frac{h_n}{6}\left[\varepsilon+\varepsilon+2\sum_{i=1}^{n-1}\varepsilon+4\sum_{i=0}^{n-1}\varepsilon\right]=(b-a)\cdot\varepsilon=\varepsilon$$

(c) On a

$$\left|I-A(\nu,\tilde{f})\right|\leq\left|I-A(\nu,f)\right|+\left|A(\nu,f)-A(\nu,\tilde{f})\right|\leq 10^{-3}+\varepsilon.$$

(d) Le résultat fourni par la calculatrice est $A(\nu,\tilde{f}$ avec $\varepsilon=10^{-12}$. Le résulat obtenu vérifie donc $\left|I-A(\nu,\tilde{f})\right|\leq 10^{-3}+10^{-12}$. La perte de précision de 10^{-12} due au calcul est négligeable devant l'erreur de 10^{-3} due à la méthode.

9 (← 55.)

(a) Immédiat : il suffit de calculer

$$\frac{1}{2}\left\{\int_{x_i}^{x_{i+1}}((a_i+a_{i+1})x^2+(b_i+b_{i+1})x+(c_i+c_{i+1}))dx\right\}.$$

(b) On a en utilisant à la troisième ligne sur le théorème sur l'erreur entre le polynôme d'interpolation et la fonction interpolée

$$\begin{aligned}&\left|\textstyle\int_{x_i}^{x_{i+1}}f(x)dx-Q_i(f)\right|\\&=\left|\tfrac{1}{2}\left\{\int_{x_i}^{x_{i+1}}f(x)-\mathbf{L}[x_{i-1},x_i,x_{i+1};f](x)dx\right\}+\tfrac{1}{2}\left\{\int_{x_i}^{x_{i+1}}f(x)-\mathbf{L}[x_i,x_{i+1},x_{i+2};f](x)dx\right\}\right|\\&\leq\tfrac{1}{2}\int_{x_i}^{x_{i+1}}\left|f(x)-\mathbf{L}[x_{i-1},x_i,x_{i+1};f](x)\right|dx+\tfrac{1}{2}\int_{x_i}^{x_{i+1}}\left|f(x)-\mathbf{L}[x_i,x_{i+1},x_{i+2};f](x)\right|dx\\&\leq\tfrac{1}{2}\sup_{[x_{i-1},x_{i+1}]}|f^{(3)}|\int_{x_i}^{x_{i+1}}|x-x_{i-1}||x-x_i|dx+\tfrac{1}{2}\sup_{[x_i,x_{i+2}]}|f^{(3)}|\int_{x_i}^{x_{i+1}}|x-x_i||x-x_{i+1}|dx\\&\quad\leq C_i\cdot\sup_{[x_{i-1},x_{i+2}]}|f^{(3)}|\end{aligned}$$

avec

$$2C_i=\int_{x_i}^{x_{i+1}}|x-x_{i-1}||x-x_i|dx+\int_{x_i}^{x_{i+1}}|x-x_i||x-x_{i+1}|dx.$$

(c) Si f est un polynôme de degré 2 alors $L[x_{i-1}, x_i, x_{i+1}; f] = f$. On en déduit facilement que $Q_i(f) = \frac{2}{2}\int_{x_i}^{x_{i+1}} f(x)dx$. La même relation vaut pour les premier et dernier termes de $Q(f)$ et l'égalité demandée est alors conséquence immédiate de la relation de Chasle pour les intégrales.

(d) Il suffit d'additionner les erreurs trouvées dans la partie A) en prenant soin que les premier et dernier termes de la somme sont différents (plus simples).

§ 3 SOLUTIONS APPROCHÉES DES ÉQUATIONS

10 (← 68.)

(a) On a $\frac{e^{-x}}{x} = 1 \iff x = e^{-x} \iff x - e^{-x} = 0$.

(b) La fonction f est (indéfiniment) dérivable sur $\mathbb{R}$. On a $f'(x) = 1 + e^{-x} > 1 > 0$ ($x \in \mathbb{R}$) donc f est strictement croissante sur $\mathbb{R}$. La fonction f est donc injective et l'équation $f(x) = 0$ admet au plus une solution. Comme $f(0) = -1 < 0$ et $f(1) = 1 - \frac{1}{e} > 0,5 > 0$, d'après le théorème des valeurs intermédiaires, f admet une racine (unique) dans $]0,1[$.

(c) On a $f''(x) = -e^{-x} < 0$ donc la fonction est concave (sur $\mathbb{R}$). Comme elle est aussi croissante, on prendra donc comme point de départ dans la suite de Newton, l'extrémité gauche de l'intervalle i.e. $x_0 = 0$ (cfr l'exercice 15 du dossier d'exercices).

(d) La suite de Newton est définie par $x_0 = 0$ et $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$, $n \geq 0$. On trouve $x_1 = 0,5$, $x_2 = 0,56631...$ et $x_3 = 0,56714....$

(e) On a $f(0,5671) = -6,78428E-05 < 0$ et $f(0,56719) = 7,32E-05 > 0$ dont d'après le théorème des valeurs intermédiaires, on a $r \in]0,5671; 0,56719[$. Il suit que les quatre première décimales de r sont bien 5671.

11 (← 67.)

(a) Considérons la fonction polynomiale f définie sur $\mathbb{R}$ par $f(x) = x^4 + 2x^2 - 1$. On a $f'(x) = 4x^3 + 4x > 0$ pour $x \in]0,1[$. Donc f est strictement croissante sur $[0,1]$ et définit une bijecton de $[0,1]$ sur $f([0,1]) = [-1,2]$. Puisque $0 \in [-1,2]$, il existe un et un seul $r \in [0,1]$ tel que $f(r) = 0$. Il est utile de noter pour la suite que $f''(x) = 12x^2 + 4 > 0$, donc f est convexe.

(b) $f(0,5) = 1/16 - 1/2 < 0$ et $f(1) = 2 > 0$ donc $r \in]0,5; 1[$. Ensuite $f(0,75) = (3/4)^4 + 18/16 - 1 > 0$ et $f(0,5) < 0$ donc $r \in]0,5; 0,75[$.

(c) La fonction étant strictement croissante convexe, on est directement dans le cas d'application du théorème du cours et le point de départ doit être pris à droite de la racine (faire un shéma.) On prendra donc $\overline{x}_0 = 0,75$. La suite de newton est donnée par la formule

$$\overline{x}_{n+1} = \overline{x}_n - \frac{f(\overline{x}_n)}{f'(\overline{x}_n)} = \overline{x}_n - \frac{\overline{x}_n^4 + 2\overline{x}_n - 1}{4\overline{x}_n^3 + 4\overline{x}_n}.$$

On trouve $\overline{x}_1 = 0,655...$ et $\overline{x}_2 = 0,6437....$

(d) Comme précédemment, on est directement dans le cas d'application du théorème du cours et le point de départ doit être pris à gauche de la racine (faire un schéma.) On prendra

donc $\underline{x}_0 = 0,5$. La suite de la sécante est donnée par la formule

$$\underline{x}_{n+1} = \frac{\underline{x}_n f(0,75) - 0,75 f(\underline{x}_n)}{f(0,75) - f(\underline{x}_n)}$$

On trouve $\underline{x}_1 = 0,624...$, $\underline{x}_2 = 0,636...$ et $\underline{x}_3 = 0,6409....$

(e) On sait (cours) que, f étant strictement croissante convexe, la suite de la sécante croît vers r tandis que la suite de Newton décroît vers r. On a donc $\underline{x}_0 \le \underline{x}_1 \le \underline{x}_2 \le \underline{x}_3 \le r \le \overline{x}_2 \le \overline{x}_1 \le \overline{x}_0$. En particulier $0,6409... \le r \le 0,647...$ de sorte que l'approximation de r avec deux décimales exactes est $0,64$.

12 (← 75)

(a) La fonction $f(x) = (x^3 - 1)/3$ ne laisse pas stable l'intervalle $[1,2]$ (c-a-d $f([1,2]) \not\subset [1,2]$) car $f(2) = 7/3 > 2$) donc on ne peut pas lui appliquer le théorème du point fixe. On peut aussi remarquer que $\max_{[1,2]} |f'(x)| = \max_{[1,2]} |3x^2 - 3| = 9 >> 1$.

(b) Pour s'assurer que f vérifie toutes les hypothèses du théorème du point fixe, nous devons montrer que $f([1,2]) \subset [1,2]$ puis que f est une contraction sur $[1,2]$ ce que nous ferons en montrant que sa dérivée est en valeur absolue bornée par un nombre $K < 1$ sur $[1,2]$.
Voyons le premier point. On a

$$f'(x) = (1/3) \cdot 3 \cdot (3x+1)^{-2/3} > 0 \quad \text{sur } [1,2]$$

donc f est strictement croissante et définit une bijection de $[1,2]$ sur $f([1,2]) = [f(1), f(2)] = [4^{1/3}; 7^{1/3}]$. Comme $4^{1/3} \approx 1,587 > 1$ et $7^{1/3} \approx 1,913 < 2$ on a bien $f([1,2]) \subset [1,2]$.
Pour le second point, on remarque que

$$1 \le x \le 2 \implies 4 \le 3x+1 \le 7$$
$$\implies 1/7 \le (3x+1)^{-1} \le 1/4 \implies (1/7)^{2/3} \le f'(x) \le (1/4)^{2/3}. \quad (3.1)$$

Comme f' est positive, on a

$$\max_{[1,2]} |f'(x)| = \max_{[1,2]} f'(x) \le (1/4)^{2/3} \approx 0,3968503 < 1.$$

Cela montre que f vérifie toutes les conditions du théorème du point fixe (de telle sorte que toute suite x_n définie par la $x_0 = a \in [1,2]$ et $x_{n+1} = f(x_n)$ $(n \ge 0)$ converge vers r).

(c) Montrons que si $f(x_0) > x_0$ alors (x_n) est croissante. Pour cela nous devons montrer que pour tout $n \ge 0$ on a $x_{n+1} \ge x_n$. Nous utilisons une démonstration par récurrence. Puisque $f(x_0) = x_1$, l'inégalité est vraie pour $n = 0$ par hypothèse et cela donne l'initialisation de la récurrence. Pour l'hérédité, supposant que $x_{n+1} > x_n$, nous montrons que $x_{n+2} > x_{n+1}$. La conclusion se déduit immédiatement de l'hypothèse de récurrence car f croissante et $x_{n-1} < x_n$ entraînent $f(x_{n+1}) < f(x_n)$ soit $x_{n+2} < x_{n+1}$. Le cas $f(x_0) < x_0$ se traite de manière similaire. Les deux cas peuvent évidemment se produire. Si $x_0 = 1$ alors $f(x_0) = A^{1/3} > 1 = x_0$ et si $x_0 = 2$ alors $f(x_0) = 7^{1/3} < 2 = x_0$.

(d) Le tableau ci-dessous donne les valeurs des suites x_n lorsque $x_0 = 1$ (suite croissante vers r) $x_0 = 2$ (suite décroissante vers r).

	$x_0 = 1$		$x_0 = 2$
$n = 1$	1,5874011	$< r <$	1,9129312
$n = 2$	1,7927904	$< r <$	1,8888351
$n = 3$	1,8545417	$< r <$	1,8820569
$n = 4$	1,8723251	$< r <$	1,8801413
$n = 5$	1,8773842	$< r <$	1,8795993

On en déduit que $r = 1,87$ avec deux décimales exactes. Naturellement, on aurait pu calculer une seule suite et s'assurer qu'on avait les bonnes décimales en utilisant la même idée que dans l'exercice précédent.

(e) La convergence des suites précédentes est très lente (d'après le cours, l'erreur à la k-ième itération est majorée par une constante multipliée par $(1/0,39)^k$ ce qui donne une convergence à peine plus rapide que la dichotomie. Ici, posant simplement, $g(x) = x^3 - 3x - 1$, on avait $g'(x) = 3x^2 - 3 > 0$ sur $]1,2]$ et $g''(x) = 6x > 0$ sur $[1,2]$, de sorte que la fonction était strictement croissante convexe et on pouvait appliquer la méthode de Newton avec $x_0 = 2$ (on prend l'extrémité supérieure : "schéma des 4 cas"). De manière précise, on trouve les valeurs suivantes :

x_0	2
x_1	1,8888889
x_2	1,8794516

La valeur x_2 est en réalité précise avec 3 décimales.

Deux itérations suffisent donc à obtenir le résultat obtenu en 5 itérations avec la méthode précédente.

§ 4 RÉSOLUTION DES SYSTÈMES LINÉAIRES. MÉTHODES DIRECTES

13 (← 88)

(a) Le système se résoud immédiatement par substitutions successives (de "bas en haut") comme suit

$$\left\{\begin{array}{lcl} x_1 & = & \frac{c_1 - a_{13}x_3 - a_{12}x_2}{a_{11}} \\ x_2 & = & \frac{c_2 - a_{24}x_4 - a_{23}x_3}{a_{22}} \\ \cdots & & \\ x_i & = & \frac{c_i - a_{i,i+2}x_{i+2} - a_{i,i+1}x_{i+1}}{a_{ii}} \\ \cdots & & \\ x_{n-2} & = & \frac{c_{n-2} - a_{n-2,n-1}x_{n-1} - a_{n-2,n}x_n}{a_{n-2,n-2}} \\ x_{n-1} & = & \frac{c_{n-1} - a_{n-1,n}x_n}{a_{n-1,n-1}} \\ x_n & = & \frac{c_n}{a_{nn}} \end{array}\right. \tag{4.1}$$

(b) Pour le nombre d'opérations

x_k	opérations
$k=n$	1 div.
$k=n-1$	1 soust., 1 mult., 1 div.
$1\le k\le n-2$	2 soust., 2 mult., 1 div

On a un total de

$$N = 1+3+\sum_{k=1}^{n-2} 5 = 4+5(n-2) = 5n-6.$$

14 (← 89.)

(a)

$$A^{(0)} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}, \quad A^{(1)} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & (a_{22}-\frac{a_{21}}{a_{11}}a_{12}) & (a_{23}-\frac{a_{21}}{a_{11}}a_{13}) \\ 0 & (a_{32}-\frac{a_{31}}{a_{11}}a_{12}) & (a_{33}-\frac{a_{31}}{a_{11}}a_{13}) \end{pmatrix}$$

et

$$A^{(2)} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} \\ 0 & 0 & (a_{33}^{(1)}-\frac{a_{32}^{(1)}}{a_{22}^{(1)}}a_{23}^{(1)}) \end{pmatrix}.$$

(b) On a

$$\begin{aligned} L_1A^{(0)} &= \begin{pmatrix} 1 & 0 & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & 0 \\ -\frac{a_{31}}{a_{11}} & 0 & 1 \end{pmatrix}\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \\ &= \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ (\frac{-a_{21}}{a_{11}}a_{11}+a_{21}) & (\frac{-a_{21}}{a_{11}}a_{12}+a_{22}) & (\frac{-a_{21}}{a_{11}}a_{13}+a_{23} \\ (\frac{-a_{31}}{a_{11}}a_{11}+a_{31}) & (\frac{-a_{31}}{a_{11}}a_{12}+a_{32}) & (\frac{-a_{31}}{a_{11}}a_{13}+a_{33}) \end{pmatrix} = A^{(1)} \end{aligned}$$

Ensuite

$$\begin{aligned} L_2A^{(1)} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{a_{32}^{(1)}}{a_{22}^{(1)}} & 1 \end{pmatrix}\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22}^{(1)} & (a_{23}^{(1)} \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} \end{pmatrix} \\ &= \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} \\ 0 & (\frac{-a_{32}^{(1)}}{a_{22}^{(1)}}a_{22}^{(1)}+a_{32}^{(1)}) & (\frac{-a_{32}^{(1)}}{a_{22}^{(1)}}a_{23}^{(1)}+a_{33}^{(1)}) \end{pmatrix} = A^{(2)} \end{aligned}$$

(c) On déduit de la question précédente que $A^{(2)} = L_2A^{(1)} = L_2L_1A^{(0)}$. Il suffit alors de prendre $R = A^{(2)}$ car $A^{(2)}$ est triangulaire supérieure et $L = (L_2L_1)^{-1}$ car L_2 et L_1 sont triangulaires inférieures et le produit de deux matrices triangulaires inférieures est encore une matrice triangulaire inférieure et enfin l'inverse d'une matrice triangulaire inférieure est encore triangulaire inférieure.

(d) Sous réserve que les divisions soient toujours possibles (ce n'est évidemment pas toujours le cas) le même procédé peut-être appliqué à une matrice $n \times n$. Notant $A^{(0)}$ la matrice du système de départ, l'algorithme de Gauss permet de construire une suite de matrices $A^{(i)}$, $i = 1 \ldots, n-1$ de telle sorte que $A^{(n-1)}$ soit triangulaire supérieure. On passe de $A^{(i-1)}$ à $A^{(i)}$ en multipliant par une matrice L_i, $A^{(i)} = L_i A^{(i-1)}$ avec L_i donnée par

$$L_i = \begin{pmatrix} Id_{i-1} & & 0 & & \\ & \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \frac{a_{i+1,i}^{(i-1)}}{a_{i,i}^{(i-1)}} & 1 & 0 & \cdots & 0 \\ \frac{a_{i+2,i}^{(i-1)}}{a_{i,i}^{(i-1)}} & 0 & 1 & & \\ \vdots & \vdots & & \ddots & 0 \\ \frac{a_{n,i}^{(i-1)}}{a_{i,i}^{(i-1)}} & 0 & \cdots & & 1 \end{pmatrix} \\ 0 & & & & \end{pmatrix}$$

où Id_{i-1} désigne la matrice Identité de dimension $i-1$. Cette matrice est supprimée dans le cas $i = 1$.

15 (← 90.)

(a) Les étapes (a) et (b) emploient $1(\div) + 1(-) + 1(\times)$ et elles sont répétées pour $i = 2$ jusqu'à $i = n$, soit $n-1$ fois. Au total on a $3(n-1)$ op.

(b) On obtient en effectuant le produit

$$A = \begin{pmatrix} \alpha_1 & c_1 & & & & 0 \\ \beta_2\alpha_1 & \beta_2 c_1 + \alpha_2 & c_2 & & & \\ & \beta_3\alpha_2 & \beta_3 c_2 + \alpha_3 & c_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \beta_{n-1}\alpha_{n-2} & \beta_{n-1}c_{n-2} + \alpha_{n-1} & c_{n-1} \\ 0 & & & & \beta_n\alpha_{n-1} & \beta_n c_{n-1} + \alpha_n \end{pmatrix}$$

On obtient la matrice demandée en remarquant que d'après la définition de l'algorithme on a $\beta_i\alpha_{i_1} = b_i$ pour $i = 2, \ldots, n$ et $\alpha_i = a_i - \beta_i c_{i-1} \Rightarrow a_i = \alpha_i + \beta_i c_{i-1}$ pour $i = 2, \ldots, n$ tandis que $a_1 = \alpha_1$.

(c) $Ax = b \Leftrightarrow (L.U)(x) = d \Leftrightarrow L(Ux) = d \Leftrightarrow (Ly = d \quad et \quad Ux = y)$.

(d)

$$Ly = b \Leftrightarrow \begin{cases} x_1 = d_1 \\ \beta_2 x_1 + x_2 = d_2 \\ \vdots \\ \beta_n x_{n-1} + x_n = d_n \end{cases} \Leftrightarrow \begin{cases} x_1 = d_1 \\ x_i = d_i - \beta_i x_{i-1} \quad i = 2, \ldots, n. \end{cases}$$

Le nombre N d'opérations est donné par $N = (n-1).(1(-) + 1(\times)) = 2(n-1)$.

(e)

$$Ux = y \Leftrightarrow \begin{cases} \alpha_1 x_1 + c_1 x_2 = y_1 \\ \alpha_2 x_2 + c_2 x_3 = y_2 \\ \vdots \\ \alpha_{n-1} x_{n-1} + c_{n-1} x_n = y_{n-1} \\ \alpha_n x_n = y_n \end{cases} \Leftrightarrow \begin{cases} x_n = \frac{y_n}{\alpha_n} \\ x_{n-i} = \frac{(y_{n-i} - c_{n-i} x_{n-i+1})}{\alpha_{n-i}} \end{cases} \quad 1 \le i \le n-1.$$

Le nombre N' d'opérations est donné par $N' = 1(\div) + (n-1).(1(\div) + 1(-) + 1(\times)) = 3n-2$.

(f) On commence à effectuer la décomposition $A = LU$ ce qui revient à utiliser l'algorithme et donc coûte $3n-3$ op. puis on résout $Ly = d$ pour $2n-2$ op. et finalement on résout $Ux = y$ pour $3n-2$ op. Au total le nombre d'op. est $8n-7$.

16 (← 93.)

A)

$$L = \begin{pmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & & \\ \vdots & \ddots & \ddots & 0 \\ l_{n1} & \dots & l_{nn-1} & l_{nn} \end{pmatrix} \Longrightarrow T(L) = \begin{pmatrix} l_{11} & l_{21} & \dots & l_{n1} \\ 0 & l_{22} & l_{32} & \\ \vdots & \ddots & \ddots & l_{nn-1} \\ 0 & \dots & 0 & l_{nn} \end{pmatrix}$$

La matrice $T(L)$ est triangulaire supérieure. Puisque L est triangulaire, son déterminant est égal au produit des coefficients sur la diagonale soit $\det(L) = l_{11} \cdots l_{nn}$.

B) En utilisant les propriétés de la transposée rappelées dans l'énoncé, on a $T(A) = T(LT(L)) = T(T(L))T(L) = LT(L) = A$.

C) $A = L \cdot T(L) \Longrightarrow \det(A) = \det(L)\det(T(L))$ mais puisque $\det(T(L)) = \det L$, on a $\det A = (\det L)^2 = l_{11}^2 \cdots l_{nn}^2$. Comme $\det A \neq 0$ les l_{kk} sont non nuls.

D) On a $L' \cdot T(L') = LDT(LD) = LDT(D)T(L) = LD^2T(L)$ car puisque D est diagonale, $T(D) = D$. D'autre part puisque D n'a que des 1 et des -1 sur sa diagonale, $D^2 = I$ et finalement $L' \cdot T(L') = LT(L) = A$. Sous l'hypothèse **H**, on peut toujours écrire $A = L' \cdot T(L')$ avec $L' = (l'_{ij})$ une matrice triangulaire inférieure telle que $l'_{kk} > 0$ pour $k = 1, \dots n$. En effet partant de la matrice L donnée par l'hypothèse **H**, d'après ce qui précède, il suffit de prendre $L' = LD$ avec $D = (d_{ij})$ la matrice diagonale telle que $d_{ii} = \text{signe}(l_{ii})$.

E) $Ax = b \Longleftrightarrow (LT(L))(x) = b \Longleftrightarrow L(T(L)x) = b \Longleftrightarrow Ly = b$ et $T(L)x = b$. Chacun des deux systèmes se résoud par substitutions successives en n^2 opérations (voir cours). Au total if faut donc $2n^2$ opérations.

F) On a

$$\begin{aligned}\sum_{k=2}^{n}(n-k+1)(k-1) &= n\sum_{k=2}^{n}(k-1)-\sum_{k=2}^{n}(k-1)^2\\ &= n\sum_{k=1}^{n-1}k-\sum_{k=1}^{n-1}k^2\\ &= n\frac{n(n-1)}{2}-\frac{n(n-1)(2n-1)}{6}\\ &= n(n-1)\frac{3n-(2n-1)}{6}=n(n-1)\frac{n+1}{6}=\frac{n(n^2-1)}{6}.\end{aligned}$$

G) D'après l'hypothèse, pour $j\le i$, on a $a_{ij}=\sum_{s=1}^{n}L_{is}T(L)_{sj}=\sum_{s=1}^{n}l_{is}l_{js}$. Mais, comme L est triangulaire inférieure, pour $s>j$ on a $l_{js}=0$, il reste donc $a_{ij}=\sum_{s=1}^{j}l_{is}l_{js}$.

H) En particulier on a $l_{11}^2=a_{11}$ puis $l_{11}l_{i1}=a_{i1}$ pour $i=2,\dots,n$ d'où l'on déduit immédiatement les formules demandées.

I) D'après (G), en prenant $i=j=k$, on obtient $a_{kk}=\sum_{s=1}^{k}l_{ks}^2$ d'où, en séparant l'indice $s=k$, $a_{kk}=l_{kk}^2+\sum_{s=1}^{k-1}l_{ks}^2$, d'où l'on déduit en utilisant la positivité de l_{kk} la relation $l_{kk}=\sqrt{a_{kk}-\sum_{s=1}^{k-1}l_{ks}^2}$.

J) Ici en employant (**??**) avec $j=k$ on obtient $a_{ik}=\sum_{s=1}^{k}l_{is}l_{ks}$. En séparant l'indice $s=k$, il vient $a_{ik}=l_{ik}l_{kk}+\sum_{s=1}^{k-1}l_{is}l_{ks}$ d'où l'on tire

$$l_{ik}=\frac{a_{ik}-\sum_{s=1}^{k-1}l_{is}l_{ks}}{l_{kk}}\quad(i=k+1,\dots,n).$$

On remarque que la division par l_{kk} est permise car il a été montré que l_{kk} est non nul.

K) Correspondant à la matrice

$$A=\begin{pmatrix}1&0&1\\0&1&0\\1&0&2\end{pmatrix}$$

On trouve

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

L) Il y a n racines carrées. Ensuite il y a $n-1$ divisions pour la première colonne ($k=1$) pouis $n-2$ pour la suivante ($k=2$) etc. Au total on trouve $\sum_{k=1}^{n}(n-k)=n(n-1)/2$ divisions.

M) Le calcul pour les additions-soustractions et les multiplications est presque identique. Nous nous limitons au cas des additions-soustractions. Les additions-soustractions apparaissent à partir de $k=2$. le calcul de l_{kk} en emploie $k-1$ et celui de l_{ik} pour $i\ge k+1$ aussi $k-1$ au total le nombre est $\sum_{k=2}^{n}(k-1)(n_k+1)$ le $n-k+1$ correspondant au nombre d'indices i tel que $i\ge k$. On trouve le résultat demandé grâce à la question préliminaire d'arithmétique.

N) En ajoutant les opérations nécessaires au calcul de L puis celles nécessaires à la résolution des systèmes triangulaires (voir 2.2 (4)) on arrive à un nombre d'opération asymptotiquement égal à $n^3/3$ contre $2n^3/3$ pour la méthode de Gauss. La méthode de Cholesky est par conséquent plus économique mais, bien sûr, elle ne s'applique qu'aux matrices A vérifiant l'hypothèse **H**.

INDEX

BIBLIOGRAPHIE

Crouzeix, M & A. L. Mignot [1984], *Analyse numérique des équations différentielles*, Masson, Paris.

Démidovitch, B. & I. R. Maron [1979], *Eléments de calcul numérique*, Mir, Moscou. Traduit du russe.

Hardy, G. H. [1952], *A course of pure mathematics*, Cambridge University Press, Cambridge. Dixième édition (première édition, 1908).

Paterson, A. [1991], *Differential equations and numerical analysis*, Cambridge university press, Cambridge.

Quarteroni, A., R. Sacco & F. Salai [1998], *Matematica Numerica*, Springer-Verlag, Milano.

Sibony, M. & J. C. Mardon R. [1982], *Analyse numérique (2 tomes)*, Hermann, Paris.