ANALYSE NUMÉRIQUE

JEAN-PAUL CALVI

1R2

UPS
Université de Toulouse

ISNB 978-2-9546609-0-5

R2 Décembre 2014

L'ouvrage est disponible en ligne sur les pages suivantes :

— http://univ.jeanpaulcalvi.com

— http://www.math.univ-toulouse.fr/~calvi

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

Bashô

Préface

J'attribue mon intérêt - tardif - pour l'analyse numérique à deux accidents non entièrement indépendants et conjointement à l'origine de ce livre. Le premier, qui permit le second, remonte à ma rencontre avec Len Bos, aujourd'hui professeur à l'université de Vérone, un remarquable numéricien - et très passable fermier - tandis que le second se produisit un peu plus tard lorsque le Département de Mathématiques de l'Université Paul Sabatier me confia, il y a déjà une dizaine d'années, un cours d'analyse numérique destiné à des étudiants de la filière informatique. Si la Providence seule explique le premier, le cours d'informatique en question m'échut par une méthode intéressante que j'appelle celle du dernier reste : le département me fit plusieurs fois l'honneur de m'attribuer un enseignement dont aucun de mes collègues ne voulait, attention que j'ai toujours tenté de m'expliquer, même difficilement, par l'évidence de mes qualités pédagogiques. J'abordai la préparation de ce cours avec une intention hérétique, celle de m'adresser au public auquel il était destiné, en suivant un syllabus à la fois laconique et banal, dans l'idée de procurer une culture numérique rudimentaire mais cohérente à qui ne se risquerait pas à lire une seule des démonstrations qu'il contiendrait. En réalité, je rencontrai au début plusieurs promotions d'étudiants remarquables qui me dissuadèrent de supprimer les démonstrations qu'ils n'étaient pas contraints de lire et firent en sorte que mon cours ne prenne jamais la direction d'une introduction purement informelle. Si bien qu'un public plus expert s'intéressa au texte et je fus par la suite conduit à développer le contenu jusqu'à inclure quelques éléments surtout destinés à des lecteurs qui se spécialisaient en mathématiques, en avertissant les autres par le symbole $*$. Pour rester cohérent avec mon objectif, je me suis appuyé sur des bases mathématiques 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 une certaine familiarité avec le calcul matriciel. J'ai inséré d'assez nombreux exercices, souvent élémentaires, y compris dans le cours du texte, dans l'espoir d'aider à la compréhension. J'ai aussi inclus les codes de fonctions SCILAB qui implémentent les algorithmes fondamentaux ; ces codes ne sont pas nécessairement les plus efficaces ni les plus élégants. Mes goût personnels se sont insinués au travers de quelques codes de calculs formels adaptés au logiciel MAXIMA [*].

Au final, le texte contient 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, les quatrième et cinquième chapitre, consacrés à l'analyse numérique matricielle ne sont sans doute que des esquisses. Les exercices donneront aux lecteurs intéressés une approche plus riche du sujet.

Les questions de complexité et de stabilité des procédés numériques sont introduites de manière concrète et informelle, et sont abordées chaque fois que c'est possible sans être excessivement technique. J'ai toujours jugé qu'il n'y avait pas de plus décourageante manière de commencer un cours d'analyse numérique que par un chapitre sur l'étude des erreurs.

Des versions préliminaires ont été progressivement mises en ligne depuis 2008 et y ont été souvent téléchargées, plus de vingt mille fois dans la dernière année et, dans les deux tiers des cas, hors de France. Il est possible qu'une future seconde édition élargie prenne la forme d'une publication classique si elle existe encore au moment où je l'aurai complétée.

Foix, Juin 2013
Jean-Paul Calvi [†]

∗. Les logiciels SCILAB et MAXIMA sont des logiciels libres téléchargeables sur internet et adaptés à tous les systèmes d'exploitations

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

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.

Table des matières

Table des codes SCILAB

Table des codes MAXIMA

I

Interpolation

§ 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 Écrire 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,

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

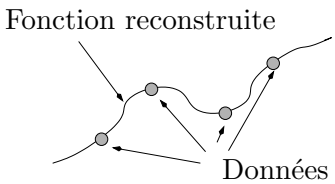
E 2 Peut-on retrouver un polynôme de degré m quand on sait que $x_1,\dots,x_m$ sont ses racines ?

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 conduits à *reconstruire* une fonction globale f_r à partir d'un nombre fini de données (a_i,f_i).



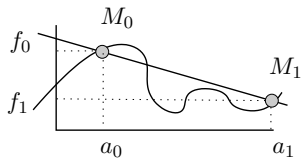
Sauf cas très simple, la fonction f_r ne coïncidera pas avec la fonction idéale f mais il faut faire en sorte qu'elle n'en soit pas trop éloignée.

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, à la section 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 prendre 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

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

Le problème est à peine plus compliqué lorsque nous disposons 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.

∗. Dans le cas complexe, c'est-à-dire, lorsqu'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.

†. La seule exception à cette convention se trouve au paragraphe 2.6 consacré à l'algorithme de Neville-Aitken.

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 $p(a_i) = f_i$, $i = 0,\ldots,d$. De manière précise, posant $p(x) = \sum_{i=0}^{m} c_i x^i$, nous devons déterminer 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 est proposé à l'exercice 43) 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 $i \in \{0,\ldots,d\}$ 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 $A = \{a_0, \dots, a_d\}$ un ensemble de $d+1$ nombres réels (deux à deux) distincts. Quelles que soient les valeurs $f_0, f_1, \dots, f_d$, il existe un et un seul polynôme $p \in \mathscr{P}_d$ tel que $p(a_i) = f_i$, $i = 0, 1, \dots, d$. Ce polynôme, est donné par la formule*

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

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

Démonstration. La seule affirmation que nous n'avons 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 $i = 0, \dots, d$, $q(a_i) = f_i - f_i = 0$. 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 $p(a_i) = f(a_i)$ $(i = 0, 1, \dots, d)$ s'appelle alors le polynôme d'**interpolation de Lagrange** de f aux points a_i. Il est noté $\mathbf{L}[a_0, \dots, a_d; f]$ 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 $A = \{a_0, \dots, a_d\}$ et non du $(d+1)$-uplet $(a_0, \dots, a_{d+1})$ *. 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 $0, 1, \dots, d$ alors

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

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

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

Cette expression de $\mathbf{L}[A; f] = \mathbf{L}[a_0, \dots, a_d; f]$ est connue sous le nom de **formule d'interpolation de Lagrange**.

∗. Rappelons que la différence entre un ensemble de $d+1$ éléments deux à deux distincts et un $(d+1)$-uplet est que, dans ce dernier, l'ordre dans lequel les éléments sont écrits à toute son importance. Avec un ensemble de $d+1$ éléments deux à deux distincts, nous pouvons former $(d+1)!$ différents $(d+1)$-uplet.

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

1.5 Simplification de l'expression de la formule d'interpolation de Lagrange

Il est encore possible de simplifier l'écriture des fondamentaux de Lagrange ℓ_i en introduisant le polynôme w défini par

$$w(x) = (x-a_0)(x-a_1)\cdots(x-a_d) = \prod_{i=0}^{d}(x-a_i).$$

Ensuite, nous appelons w_i le polynôme obtenu en retirant le facteur $(x-a_i)$ dans w de sorte que nous avons à la fois

$$w(x) = w_i(x)\cdot(x-a_i) \quad \text{et} \quad w_i(x) = \prod_{j=0, j\neq i}^{d}(x-a_j).$$

En dérivant la première de ces expressions, nous obtenons

$$w'(x) = w_i'(x)\cdot(x-a_i) + w_i(x) \text{ et en prenant } x = a_i \text{ il vient } w'(a_i) = w_i(a_i) = \prod_{j=0, j\neq i}^{d}(a_i-a_j).$$

Le dernier terme est exactement le dénominateur dans l'expression de ℓ_i donnée à la relation (1.15) si bien que

$$\ell_i(x) = \frac{w(x)}{w'(a_i)(x-a_i)}$$

et la formule d'interpolation de Lagrange devient

$$\mathbf{L}[a_0,\ldots,a_d;f](x) = \sum_{i=0}^{d} f(a_i)\frac{w(x)}{w'(a_i)(x-a_i)}. \tag{1.17}$$

Cette nouvelle expression est surtout intéressante lorsque le polynôme w a une expression assez simple mais elle est aussi utile pour programmer le calcul formel de $\mathbf{L}[a_0,\ldots,a_d;f](x)$, voir 2.7.

1.6 Propriétés algébriques et linéarité

Il est essentiel de retenir l'équivalence suivante

$$\left.\begin{array}{c} p\in\mathscr{P}_d \\ p(a_i) = f(a_i) \quad i=0,\ldots,d \end{array}\right\} \Leftrightarrow p = \mathbf{L}[a_0,\ldots,a_d;f]. \tag{1.18}$$

En particulier,

$$\text{si } p\in\mathscr{P}_d \text{ alors } \mathbf{L}[a_0,\ldots,a_d;p] = p. \tag{1.19}$$

Il faut prendre garde que cette propriété n'est valable que lorsque le degré de p est inférieur où égal à d *. La relation (1.18) signifie que pour établir qu'un polynôme donné p est égal au polynôme d'interpolation de Lagrange d'une fonction f aux points $a_0,\ldots,a_d$, il suffit de vérifier que le degré de q est inférieur ou égal à d et que $q(a_i) = f(a_i)$ pour $i = 0,\ldots,d$.

La relation (1.19) 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

$$\sum_{i=0}^{d}\ell_i = 1. \tag{1.20}$$

∗. Pour la calcul de $\mathbf{L}[a_0,\ldots,a_d;p]$ lorsque le degré de p est strictement supérieur à d, voir l'exercice 26.

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 $\mathbf{L}[A,\cdot]$ qui à toute fonction f définie (au moins) sur $A = \{a_0,\dots,a_d\}$ fait correspondre son polynôme d'interpolation $\mathbf{L}[A;f] \in \mathscr{P}_d$,*

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

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

$$\text{(1.21)} \quad \begin{cases} \mathbf{L}[A;f+g] &= \mathbf{L}[A;f]+\mathbf{L}[A;g], \quad f,g \in \mathscr{F}(A) \\ \mathbf{L}[A;\lambda f] &= \lambda \mathbf{L}[A;f], \quad f \in \mathscr{F}(A), \quad \lambda \in \mathbb{R} \end{cases}.$$

E 7 Montrer les propriétés (1.21).

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.

Algorithme 3. *Les données de l'algorithme sont (i) le vecteur $a = (a_0,\dots,a_d)$ formé des points d'interpolation, (ii) le vecteur $f = (f_0,\dots,f_d)$ formé des valeurs d'interpolations (iii) 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) $P := P + L \times f_i$.

Voici une traduction en code **scilab** de l'algorithme ci-dessus

Code SCILAB 1 (Formule d'interpolation de Lagrange). Dans le code suivant

(a) N est un vecteur formé des points d'interpolation, La longueur de N est n.

(b) Y est un vecteur de même dimension que N contenant les valeurs d'interpolation,

(c) X est un vecteur contenant les points en lesquels le polynôme doit être calculé.

```
function [P]=lagrange(N, V, X)
n=length(N);
P=0;
for i=1:n,
  L=1;
  for j=[1:i-1,i+1:n]
    L=L.*(X-N(j))./(N(i)-N(j));
  end
  P=P+L*V(i);
end
endfunction
```

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** $a_i = -1+2i/d$, $i = 0, 1, \ldots, d$ 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. 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 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.

*. 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.

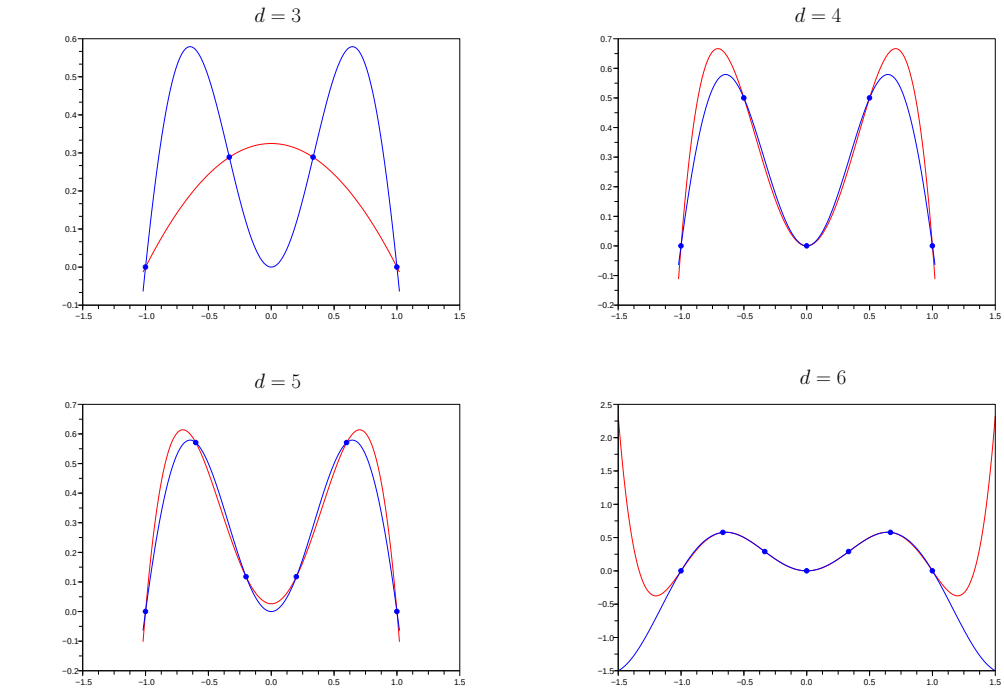


TABLE 1 – Quelques polynômes d'interpolation de la fonction $f(x) = x \sin \pi x$ sur l'intervalle $[-1, 1]$.

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*[a].

a. Les valeurs de la fonctions $f(a_i)$ sont données ; nous ne devons pas les calculer.

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 $u_d = v_d(1+\varepsilon_d)$. 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. ∎

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 $\mathbf{L}[0,\dots,a_6;M_3](x)=x^3$ calculés par l'algorithme 3.

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 (1.19) montre que

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

Nous avons modifié l'algorithme 3, pour qu'il produise un polynôme. C'est une bien mauvaise idée, du point de vue de la précision des calculs, mais elle nous servira à montrer comment un algorithme mathématiquement correct peut donner des résultats très médiocres. Les coefficients du polynôme obtenu sont indiqués 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

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

du résultat. Un calculateur ne travaille qu'avec une famille finie (très étendue) de nombres F et le résultat 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 $\mathrm{F}(a*b)$ 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 $\mathrm{F}(a*b)$ 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 induisent des erreurs au niveau du résultat comparables à (la somme de) 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

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

et des propriétés techniques des calculateurs. Cette analyse, en général est délicate et, dans ce cours, nous n'aurons l'occasion d'en considérer 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 $\mathbf{L}[0,\dots,a_{29};p]$ 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 $a_0\dots,a_d$ – 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 $f(x+\varepsilon)$ 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'habileté 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 $A=\{a_0,a_1,\dots,a_d\}$ un ensemble de $d+1$ nombres réels distincts et f une fonction définie (au moins) sur A. Alors*

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

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

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

et calculons ses valeurs en les points a_i. Nous utilisons un point d'interrogation pour indiquer une valeur inconnue mais sans influence. Pour le point a_0, nous avons

$$Q(a_0)=\frac{1}{a_0-a_d}\{f(a_0)(a_0-a_d)-[?]\times 0\}=f(a_0)$$

et le même calcul vaut le point a_d,

$$Q(a_d)=\frac{1}{a_0-a_d}\{[?]\times 0-f(a_d)(a_d-a_0)\}=f(a_d).$$

Quant aux points a_i pour $1\le i\le d$, nous avons

$$Q(a_i)=\frac{1}{a_0-a_d}\{f(a_i)(a_i-a_d)-f(a_i)(a_i-a_0)\}=f(a_i)\ (1\le i\le d-1)$$

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

Corollaire 6. *Sous les mêmes hypothèses, pour tout couple d'indice (i,j) dans $\{0,\dots,d\}$ avec $i\neq j$,*

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

Démonstration. Cela provient immédiatement du théorème précédent, en tenant compte du fait que les points pouvant être permutés, nous pouvons mettre a_i à la place de a_0 et a_j à la place de a_d. ■

2.6 L'algorithme de Neville-Aitken

Posons $A = \{x_1, x_2, \ldots, x_{d+1}\}$ 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 $m \in \{0, 1, \ldots, d\}$ comme suit

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

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

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

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

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

Remarquons que le cas $m = d$ dans la relation (2.7) nous donne

$$p_{d+1,d} = \mathbf{L}[x_1, x_2, \ldots, x_{d+1}; f].$$

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

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

de sorte que

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

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

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

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

Algorithme 8. *Les données de l'algorithme sont (a) le vecteur $x = (x_1, \ldots, x_{d+1})$ formé des points d'interpolation, (b) le vecteur $f = (f_1, \ldots, f_{d+1})$ 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 $j \in [1 : d+1]$, $P(j,1) = f_j$.

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

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

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

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 – Coefficients instables obtenus par une version de l'algorithme de Neville-Aitken.

2.7 Algorithme de calcul formel

Le programme suivant donne une expression du polynôme d'interpolation correspond à la formule d'interpolation de Lagrange (1.17). Nous l'utiliserons dans les chapitres ultérieurs.

Code MAXIMA 1 (Formule d'interpolation de Lagrange). Dans le code suivant

(a) f est une fonction.

(b) N est la liste (le vecteur) des points d'interpolation.

(c) La fonction évalue le polynôme d'interpolation au point x. Dans le programme ci-dessous, d désigne la longueur de N, c'est-à-dire le nombre de points d'interpolation de sorte que le polynôme obtenu est de degré au plus $d-1$.

```
lagrange(f,N,x):=block(d:length(N),
                        w:prod(y-N[i],i,1,d),
                        dw:diff(w,y,1),
                        sum(f(N[i])*ratsimp(ev(w,y=x)/((x-N[i])
                            *ev(dw,y=N[i]))),i,1,d)
                        );
```

Exemple 1. Nous pouvons obtenir la formule générale du polynôme d'interpolation de degré 2 aux points a, $(a+b)/2$ et b comme suit.

Entrée > $LG : lagrange(G,[a,(a+b)/2,b],x)$

(2.9) **Sortie >**
$$-\frac{G\left(\frac{b+a}{2}\right)\left(4x^2+(-4b-4a)x+4ab\right)}{b^2-2ab+a^2}+\frac{G(b)\left(2x^2+(-b-3a)x+ab+a^2\right)}{b^2-2ab+a^2}+\frac{G(a)\left(2x^2+(-3b-a)x+b^2+ab\right)}{b^2-2ab+a^2}$$

§ 3. ÉTUDE DE L'ERREUR

3.1 L'énoncé du théorème

Comme le polynôme d'interpolation $\mathbf{L}[a_0,\ldots,a_d;f]$ est égal à la fonction f en tous les points a_i, $i=0,\ldots,d$, il est naturel d'espérer que la différence entre f et ce polynôme aux autres points sera petite c'est-à-dire, que $\mathbf{L}[a_0,\ldots,a_d;f]$ 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 son polynôme d'interpolation $\mathbf{L}[a_0,\ldots,a_d;f](x)$, c'est-à-dire trouver une majoration de la valeur absolue de E_x. La figure fait apparaître cette erreur dans le cas $d=1$. Cette erreur est une distance,

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

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

Théorème 9. *Soient $f \in \mathscr{C}^{d+1}([a,b])$ et $A = \{a_0, a_1, \dots, a_d\} \subset [a,b]$. 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 9 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 9 est donc une extension.

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 10.

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

En particulier,

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

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

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

L'exercice 31 plus bas est du même type.

(Sol. 1 p. 123.)

La démonstration du théorème 9, 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 de la forme plus générale suivante.

Théorème 11 (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, $i=0,\dots,k$, 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 $c_i\in]x_i,x_{i+1}[$, $i=0,\dots,k-1$, tels que $u'(c_i)=0$.

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 11.

3.3 Démonstration du théorème 9

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

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

Il faut prendre garde ici que x est un paramètre fixé et t est la variable. Puisque $f\in\mathscr{C}^{d+1}[a,b]$, nous avons aussi $u\in\mathscr{C}^{d+1}[a,b]$. 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 11 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 9 montre que si 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, $i = 0,\dots,d$ 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 découvert 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.

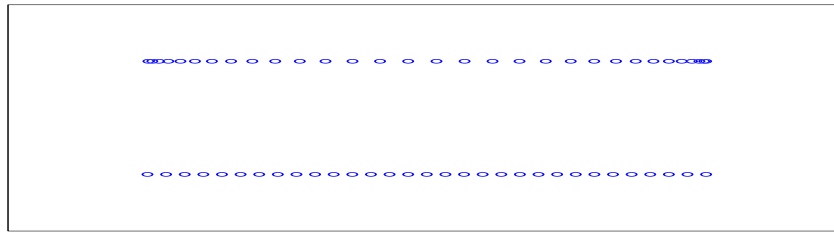


FIGURE 1 – Répartition des points de chebyshev et des points equidistants ($d = 30$).

E 14 La cas où $[a,b]$ est un intervalle quelconque (avec $a < b$) se déduit facilement du cas $a = -1$ et $b = 1$. Comment ?

*. Le choix de points d'interpolation peut être limité lorsque les valeurs $f(a_i)$ proviennent de mesures obtenues expérimentalement.

†. P. L. Chebyshev (1821-1894) est un des pères de la théorie de l'approximation des fonctions.

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 d'interpolation de Lagrange. Cette intuition est renforcée par les exemples de la table 1 (p. 8). 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 allemand Runge (1856-1927) 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 présente 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.

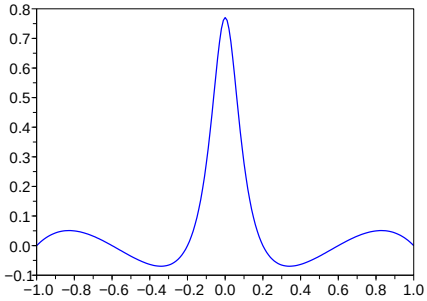
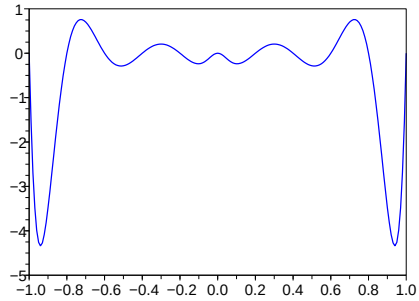
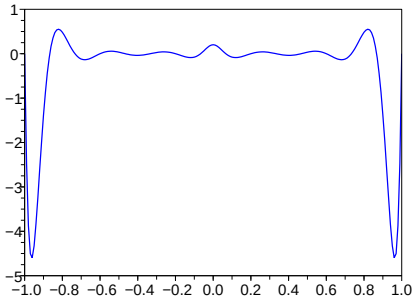
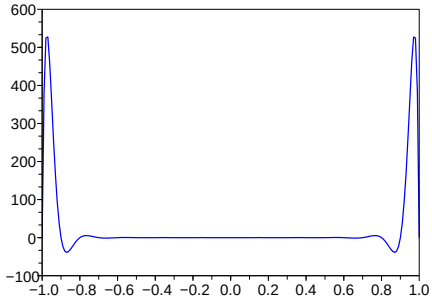
$d = 5$	$d = 10$
	
$d = 15$	$d = 20$
	

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

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

∗. 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.

fonction de Runge et donne la fonction d'erreur entre cette fonction et le polynôme d'interpolation aux points de Chebyshev.

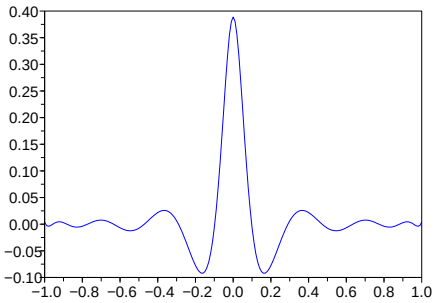
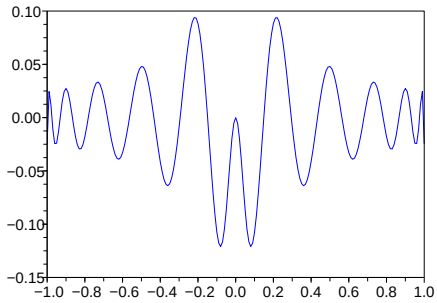
$d = 15$	$d = 20$
	

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

§ 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,

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

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

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

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,

(4.3) $h = \max\limits_{i=0,\ldots,d} h_i = \max\limits_{i=0,\ldots,d-1} (a_{i+1} - a_i).$

Ces définitions sont mises en évidence sur la figure 2.

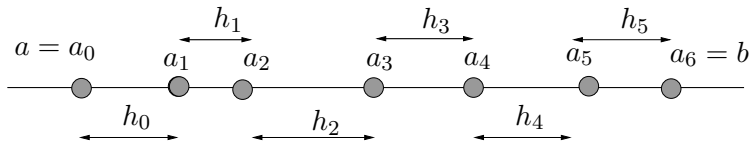


FIGURE 2 – Subdivision et écart d'une subdivision.

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

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

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 15 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

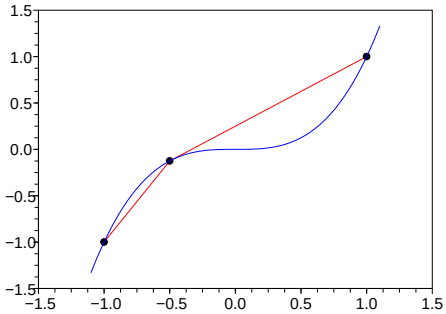
$$\text{(4.4)} \quad \begin{cases} \mathbf{PL}[\sigma\,,\, f](x) = \mathbf{L}[a_i\,,\, a_{i+1}\,;\, f_i\,,\, f_{i+1}](x), & x \in [a_i\,,\, a_{i+1}[, \\ \mathbf{PL}[\sigma\,,\, f](b) = f_d. \end{cases}$$

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

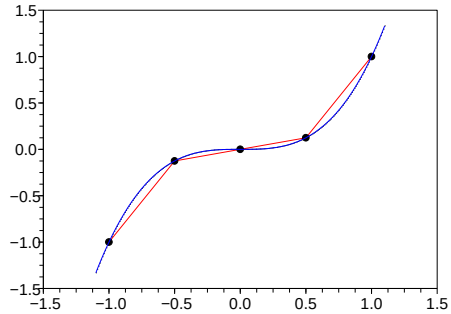
$$y_i = f(a_i), \quad i = 0\,, \dots, d,$$

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 (7 : 1) $\sigma = (-1, -0.5, 1)$ puis (7 : 2) $\sigma = (-1, -0.5, 0, 0.5, 1)$.



(1)



(2)

TABLE 7 – Deux exemples de polylignes.

Théorème 12. $\mathbf{PL}[\sigma, f]$ *est une fonction affine par morceaux continue satisfaisant*

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

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

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

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 la subdivision, ici, aux points a_i, $i = 1, \ldots, d$. Commençons par prendre un point a_i avec $1 \leq i \leq d-1$ 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 $f(a_i)$ et la fonction est donc 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 13. *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 16 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 17 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 18 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 à 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. Une autre estimation, concernant les fonctions deux fois continûment dérivables est proposée à l'exercice 39.

Théorème 14. *Soit f une fonction continûment dérivable sur $[a,b]$ et σ une subdivision de $[a,b]$ d'écart h. Pour tout $x \in [a,b]$,*

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

Démonstration. L'inégalité a démontrer est évidente lorsque x est l'un des points de la subdivision $\sigma = (a_0, \dots, a_d)$ car alors $f(x) = \mathbf{PL}[\sigma;f](x)$. Nous supposerons que $x \neq a_i$, $i = 0, \dots, d+1$. 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 (1.20) – 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

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

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

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. ■

∗. 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.

Corollaire 15. *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*

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

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,

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

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

Corollaire 16. *Lorsque σ^d est la subdivision formée des points équidistants*

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

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 19 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, $i = 0, \ldots, d$ 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 $a_{d+1} > a_d = b$.

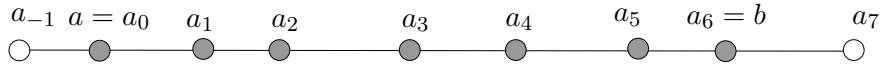


FIGURE 3 – Sudvision complétée des points a_{-1} et a_{d+1}.

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

En formule, la fonction b_i est définie par

x	$]-\infty, a_{i-1}]$	$[a_{i-1}, a_i]$	$[a_i, a_{i+1}]$	$[a_i, \infty]$
$b_i(x)$	0	$(x-a_{i-1})/(a_i-a_{i-1})$	$(x-a_{i+1})/(a_i-a_{i+1})$	0

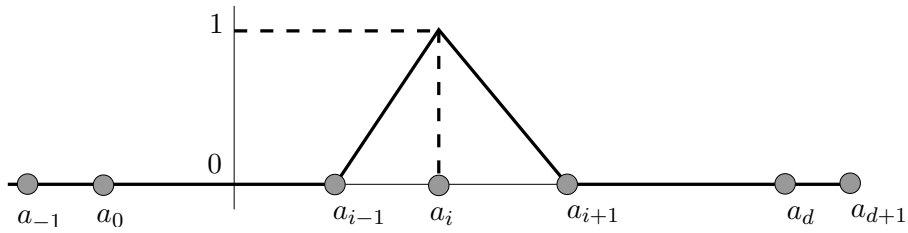


FIGURE 4 – Graphe de la fonction b_i.

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

$$0 \leq b_i(x) \leq 1, \quad x \in [a,b], \tag{4.16}$$

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

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 (**??**) 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 18. $\sum_{i=0}^{d} b_i = 1$.

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 13. ■

4.5 ∗ Approximation des fonctions continues par des fonctions polylignes

Théorème 19. *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,*

$$\lim_{d \to \infty} \max_{x \in [a,b]} |f(x) - \mathbf{PL}[\sigma^d; f](x)| = 0. \tag{4.17}$$

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 $|f(x) - f(y)| \leq \varepsilon$ 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 $|f(y)-f(x_0)| \leq \varepsilon$ 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 (du théorème 19). Nous devrons montrer que pour tout $\varepsilon > 0$ fixé à l'avance, il existe $d_0 \in \mathbb{N}$, d_0 dépendant de ε, tel que

$$d \geq d_0 \implies \max_{x \in [a,b]} |f(x) - \mathbf{PL}[\sigma^d; f](x)| \leq \varepsilon. \tag{4.18}$$

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),

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

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

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

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

$$\max_{x \in [a,b]} |f(x) - \mathbf{PL}[\boldsymbol{\sigma}^{\mathbf{d}}; f](x)| \leq 2\omega_f(h_d). \tag{4.19}$$

Puisque f est continue sur $[a,b]$, elle y 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 (4.19)

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

ce qu'il fallait établir. ∎

E 20 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 21 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 de degré au plus 1, $\mathbf{L}[a_i, a_{i+1}; f]$ par des polynômes d'interpolation de degré au plus d, $\mathbf{L}[a_i, a_{i,1}, \ldots, a_{i,d-1}, a_{i+1}; f]$ 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 ne présente aucune difficulté. Les résultats de cette section s'étendent directement au cas général. Les lecteurs intéressés peuvent s'entraîner à rédiger les démonstrations. Il est plus fécond de se concentrer sur l'insuffisance majeure des polylignes, à savoir d'être non différentiables. 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.

HISTOIRE. — L'extrait suivant est tiré du mémoire « Nouveau moyen de déterminer les longitudes de Jupiter et de Saturne au moyen, d'une table à simple entrée »de J. L. **Lagrange** (1781).

> « Lorsque sur une courbe on a déterminé différents points, on peut trouver les points intermédiaires en regardant comme rectilignes les portions de la courbe qu'ils comprennent ; on substitue ainsi un polygone à la courbe, et il est clair que cette supposition est d'autant moins inexacte que les points déterminés sont plus voisins. Mais on s'approchera encore plus de la vérité en regardant chaque portion de la courbe comme l'arc d'une courbe parabolique qui passerait par les points déterminés, et l'approximation sera d'autant plus grande qu'il y aura un plus grand nombre de points sur l'arc parabolique. Tel est le fondement des méthodes ordinaires d'interpolation, méthodes dont on se sert pour trouver, dans les Tables à simple entrée, les valeurs intermédiaires, au moyen des différences entre les termes consécutifs. »

Les oeuvres de Lagrange sont disponibles en ligne à la Bibliothèque Nationale de France (Gallica).

§ 5. EXERCICES ET PROBLÈMES

22 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 $p(a) = \alpha$, $p(b) = \beta$, $p(a) + p'(b) = \gamma$.

23 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. 124)

24 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 $\mathbf{L}[0, 3/4, 1; f](1/2)$.

25 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 $d = 2$.

26 Interpolation et division euclidienne. Rappelons 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.

27 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).

28 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.

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

$$p(x) = (x-x_1)(x-x_2)\dots(x-x_n), \quad \text{puis,}$$

$$q(x) = (x-y_1)(x-y_2)\cdots(x-y_m).$$

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. 124.)

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 autre 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. 123.)

32 Polynômes d'interpolation de certaines fractions rationnelles. On considère $d+1$ nombres réels $a_0, a_1, \ldots, a_d$ deux à deux distincts et λ un paramètre réel différent de chacun des a_i c'est-à-dire $\lambda \neq a_i$ pour $i = 0, 1, \ldots, a_d$. 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. 124.)

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 :

$$\cos(a+b) = \cos a\cos b - \sin a\sin b \quad \text{et} \quad \cos(a-b) = \cos a\cos b + \sin a\sin b.$$

(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, 30 . 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 $(d+1)$ fois continûment dérivable sur $[-1,1]$ alors pour tout $x \in [-1,1]$, on a

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

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

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

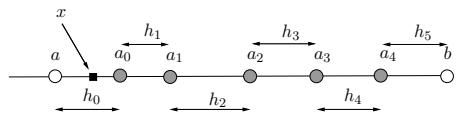
CAS OÙ x EST COMPRIS ENTRE a ET a_0.

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. 124)

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 :

(a) 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).

(b) 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 $M_{f\cdot g}$ et $r_{f\cdot g}$ 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

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

Or pour tout $\nu\in\mathbb{R}$, la suite $\nu^{d+1}/(d+1)!$ 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

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

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. Étudier les propriétés démontrées à l'exercice 28 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 $\mathscr{C}^2$. Montrer que si f est une fonction de classe $\mathscr{C}^2$ 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

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

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

COMMENTAIRE. — ∗ Supposons que $\sigma^d = (a, a_1^d, \dots, a_{d-1}^d, b)$ 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 9 sur l'erreur la fonction et son polynôme d'interpolation dans le cas $d = 1$ donnent

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

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'annule 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 $\sigma = (a = a_0, a_1, \dots, a_{d-1}, a_d = b)$ 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, $i = 0, \dots, d$ 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. Étant 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 que (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 $\varepsilon(x) = f(x) - \mathbf{L}[X,f](x)$ 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 $\varepsilon(x) = w(x)g(x)$ 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 $|f'(x_i) - Q_{x_i}(f)|$.

COMPLÉMENT

43 Déterminant de Vandermonde et interpolation de Lagrange. Nous avons vu que si $A = \{a_0,\dots,a_d\}$ et $p(x) = \sum_{i=0}^{d} c_i x^i$, pour que la condition $p(a_i) = f(a_i)$, $i = 0,\dots,d$ soit vérifiée, il faut et il suffit que les coefficients c_i satisfassent le système linéaire (à $d+1$ équations et $d+1$ inconnues)

$$\sum_{i=0}^{d} c_i a_j^i = f(a_j), \quad j = 0,\dots,d. \tag{5.10}$$

Nous noterons $\mathrm{VDM}(a_0,\dots,a_d)$ le déterminant de (la matrice associée à) ce système,

$$\mathrm{VDM}(a_0,\dots,a_d) = \begin{vmatrix} 1 & a_0 & a_0^2 & \cdots & a_0^d \\ 1 & a_1 & a_1^2 & \cdots & a_1^d \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & a_d & a_d^2 & \cdots & a_d^d \end{vmatrix}.$$

Pour que le système (5.10) admette une solution unique, son déterminant $\mathrm{VDM}(a_0,\dots,a_d)$ doit être non nul, voir le chapitre IV. Le but de cet exercice est de calculer ce déterminant et de montrer qu'il est non nul dès lors que les points a_j sont deux à deux distincts.

(a) Montrer que si les points a_j ne sont pas deux à deux distincts alors $\mathrm{VDM}(a_0,\dots,a_d)$ est nul.

(b) Nous supposons maintenant que les a_j sont deux à deux distincts. Considérons le polynôme $R(t)$ défini pour $t \in \mathbb{R}$ par

$$R(t) = \mathrm{VDM}(a_0,\dots,a_{d-1},t) = \begin{vmatrix} 1 & a_0 & a_0^2 & \cdots & a_0^d \\ 1 & a_1 & a_1^2 & \cdots & a_1^d \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & a_{d-1} & a_{d-1}^2 & \cdots & a_{d-1}^d \\ 1 & t & t^2 & \cdots & t^d \end{vmatrix}.$$

Montrer que R est un polynôme de degré d qui s'annule pour $t = a_j$, $j = 0,\dots,d-1$, et dont le coefficient de plus haut degré est $\mathrm{VDM}(a_0,\dots,a_{d-1})$. En déduire que

$$\mathrm{VDM}(a_0,\dots,a_d) = \mathrm{VDM}(a_0,\dots,a_{d-1})\prod_{j=0}^{d-1}(a_d - a_j).$$

(c) Montrer que

$$\text{VDM}(a_0,\dots,a_d)=\prod_{0\le i<j\le d}(a_j-a_i).$$

(d) En déduire que $\text{VDM}(a_0,\dots,a_d)$ est non nul à la seule condition que les points soient deux à deux distincts.

(e) Retrouver l'expression des polynômes fondamentaux de Lagrange à l'aide des formules de Cramer.

44 Soient $a\le a_0<\cdots<a_d\le b$ $d+1$ points d'interpolation deux à deux distincts dans l'intervalle $[a,b]$. Montrer, par exemple en faisant un schéma, que quel que soit M il existe une fonction f continue sur $[a,b]$ telle que

$$\max_{x\in[a,b]}|f(x)-\mathbf{L}[a_0,\dots,a_d;f](x)|\ge M.$$

Cela signifie que l'erreur commise entre une fonction continue et son polynôme d'interpolation peut être arbitrairement grand. En est-il de même si l'on considère l'erreur relative

$$\max_{x\in[a,b]}\frac{|f(x)-\mathbf{L}[a_0,\dots,a_d;f](x)|}{|f(x)|};$$

supposant que f ne s'annule pas sur $[a,b]$? En d'autres termes, est-il vrai que quel que soit M il existe une fonction f continue sur $[a,b]$ telle que

$$\max_{x\in[a,b]}\frac{|f(x)-\mathbf{L}[a_0,\dots,a_d;f](x)|}{|f(x)|}\ge M.$$

§ 6. NOTES ET COMMENTAIRES

Sur le contenu

J'ai plusieurs fois lu ou entendu dire que l'interpolation de Lagrange était un procédé d'approximation inefficace ou désuet qui serait délaissé par l'analyse numérique moderne et il m'est arrivé, par le passé, d'être impressionné par ces jugements expéditifs, spécialement pour la raison que l'interpolation polynomiale est mon sujet de recherche favori. Un tel jugement qui s'appuie généralement sur une interprétation trop hâtive du phénomène de Runge (3.5) n'a pas de fondement. L'interpolation de Lagrange, explicitement ou non, intervient dans la plupart des algorithmes fondamentaux de l'analyse numérique. Les analyses sont condamnés à travailler avec un nombre des données finies (et en précision finie) qui les contraignent à utiliser des procédés de reconstruction dont l'interpolation de Lagrange est le prototype. Le traitement des deux principales alternatives : l'usage de fonctions splines ou l'utilisation de la technique des moindres carrés se trouve d'ailleurs facilité par une connaissance solide de l'interpolation de Lagrange.

Une des utilisations récentes les plus spectaculaires de l'interpolation de Lagrange est le programme CHEBFUN constitué de programmes utilisables sur le logiciel MATLAB (semblable à SCILAB), lancé par Nick Trefethen et Zachary Battles en 2002 qui consiste à obtenir la plupart des informations usuelles sur les fonctions réelles sur $[-1,1]$ (il est toujours possible de se ramener à ce cas) à partir de ses polynômes d'interpolation de Lagrange calculés en un grand nombre de points de Chebyshev, dont les valeurs sont obtenues par la formule de Lagrange barycentrique (Cf. exercices 30 et 34) qui est optimale du point de vue de la stabilité.

Le première version de mon texte présentait une introduction à l'interpolation d'Hermite (qui fait intervenir des conditions sur les dérivées), aux différences divisées, à l'approche de Newton. Ces questions présentaient des difficultés pour l'auditoire auquel je voulais m'adresser et j'ai considéré qu'il fallait les réserver pour un second cours d'analyse numérique et les substituer dans un premier temps par une étude des polylignes.

Sur les exercices

L'idée de l'exercice 29 sur les groupements de points d'interpolation vient de Criscuolo et al. (1990). L'exercice 12 est tiré de Démidovitch and Maron (1979). Le problème 42 sur l'approximation des dérivées à l'aide des polynômes d'interpolation est inspiré de Ralston and Rabinowitz (2001, 4.1). J'ai appris les estimations de l'exercice 35 dans un cours en ligne d'A. Bellen (Trieste). Je fais partie des analystes qui ignoraient l'intérêt de la formule barycentrique de Lagrange avant la lecture de l'article de Berrut and Trefethen (2004).

Sur les difficultés

La formule d'interpolation de Lagrange est généralement assez facilement maniée par les étudiants, comme aussi la formule de Neville-Aitken et la formule d'erreur. Les difficultés principales sont les quelques points qui s'appuient sur un formalisme mathématique qui n'est plus facilement maîtrisé : la perception de l'application qui à une fonction f fait correspondre son polynôme d'interpolation, la réalisation et l'exploitation du fait que cette application est linéaire, la compréhension des propriétés caractéristiques des polynômes d'interpolation (avec le degré et les valeurs aux points d'interpolation) et l'avantage d'établir les propriétés à partir de ces propriétés caractéristiques. Les étudiants, presque systématiquement, essayent d'établir les propriétés des polynômes d'interpolation à partir d'un calcul sur la formule explicite de Lagrange.

II

Intégration

§ 1. Formules de quadratures élémentaires

1.1 L'énoncé du 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

$$\int_a^b f(x)dx = F(b) - F(a), \quad \text{où } F \text{ est une primitive de } f.$$

Dire que F est une primitive de F signifie que $F' = f$ et il est connu que toute fonction continue possède une telle primitive, et à une constante multiplicative près, une seule. Pour appliquer le théorème fondamental, c'est-à-dire essentiellement pour obtenir une primitive F de f, nous disposons de divers outils théoriques dont les plus fondamentaux sont le *théorème de changement de variable* et le *théorème d'intégration par partie*. Il reste qu'il n'est possible de déterminer explicitement une primitive F, c'est-à-dire l'écrire comme une combinaison de fonction élémentaires, que pour une classe relativement restreinte de fonctions f. L'exemple le plus important est celui des fonctions gaussiennes de la forme

$$f(x) = \exp\left(\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

qui jouent un rôle fondamental en statistiques. Même lorsque une expression de la primitive F peut être acquise, celle-ci est souvent si compliquée qu'elle n'aide guère dans l'évaluation de la quantité $F(b) - F(a)$. Par exemple, des techniques de changements de variables classiques connues de Maxima (avec lequel est effectué ici le calcul) permettent d'obtenir qu'une primitive de

$$f(x) = \frac{\left(x^3+1\right)^{\frac{1}{3}}}{x^2}$$

est donnée par

$$F(x) = \frac{\log\left(\frac{\left(x^3+1\right)^{\frac{2}{3}}}{x^2} + \frac{\left(x^3+1\right)^{\frac{1}{3}}}{x} + 1\right)}{6} + \frac{\arctan\left(\frac{\frac{2\left(x^3+1\right)^{\frac{1}{3}}}{x}+1}{\sqrt{3}}\right)}{\sqrt{3}} - \frac{\log\left(\frac{\left(x^3+1\right)^{\frac{1}{3}}}{x} - 1\right)}{3} - \frac{\left(x^3+1\right)^{\frac{1}{3}}}{x}.$$

Le calcul de $F(b) - F(a)$ nécessite l'emploi d'un processus d'approximation (pour effectuer le calcul du logarithme, de l'arc tangente et des racines cubiques. Dans ce cas, il est tout aussi naturel et souvent moins coûteux de chercher **directement** une approximation de l'intégrale. Ce commentaire ne signifie qu'il faille délaisser les méthodes théoriques de calculs et ne pas maîtriser, pour les mathématiciens, la technique de changements de variables. Une analyse théorique du problème numérique à résoudre conduit souvent à une stratégie de calcul de numérique et elle sert toujours de garde-fou pour détecter les résultats numériques incohérents.

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,

$$g = \mathbf{L}[x_0,\dots,x_d;f], \quad \text{avec } A = \{x_0,\dots,x_d\} \subset [a,b]$$

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

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

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 $\mathscr{C}[a,b]$, autrement dit, elle vérifie

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

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

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

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

$$p \in \mathscr{P}_d \Rightarrow p = \mathbf{L}[x_0,\dots,x_d;p] \Rightarrow \mathrm{Q}(p) = \int_a^b \mathbf{L}[x_0,\dots,x_d;p](x)dx = \int_a^b p(x)dx.$$

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 $\ell_j(a_j) = 1$, et $\sum_{i=0}^d \lambda_i \ell_j(a_i) = \lambda_j$. ∎

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 $i = 0,1,\dots,d$.

E 45 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.

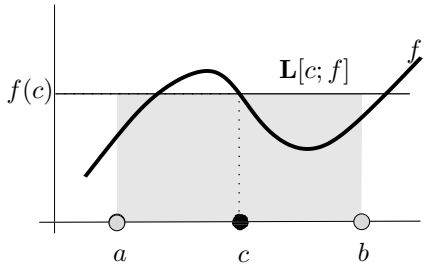
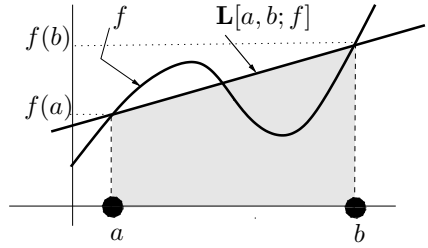
A) POINT MILIEU	B) TRAPÈZE
	
Nous avons posé $c=(a+b)/2$. L'aire de la partie grisée est égale à $\mathrm{Q}(f)$.	L'aire de la partie grisée est égale à $\mathrm{Q}(f)$.

TABLE 1 – Méthode du point milieu et du trapèze.

Dans la pratique, grâce au procédé de composition que nous verrons plus bas et qui consiste à appliquer ces approximations sur des intervalles très petit, des résultats 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 $\mathrm{Q}(f)=(b-a)f(\frac{a+b}{2})$s'appelle la **formule du point milieu**. Lorsque $f(c)>0$, $\mathrm{Q}(f)$ 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 $f\in\mathscr{C}([a,b])$. Nous prenons $d=1$ et $A=\{a,b\}$. L'approximation

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

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 $\mathrm{Q}(f) = \frac{(b-a)}{2}(f(a)+f(b))$ 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

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

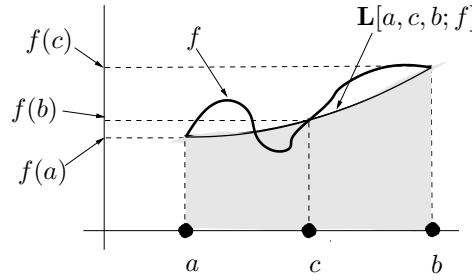
devient

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

Le calcul est proposé à l'exercice I.27. L'expression

$$\mathrm{Q}(f) = \frac{(b-a)}{6}\left(f(a)+4f(c)+f(b)\right)$$

s'appelle la **formule de Simpson**. L'aire de la partie grisée sur la figure ci-contre est égale à $\mathrm{Q}(f)$



Le calcul est facilement traité par un logiciel de calcul formel.

Code MAXIMA 2 (Obtention de la formule de Simpson). Dans le code suivant nous définissons, comme une expression, le polynôme L d'interpolation de Lagrange aux trois points a,b et c, calculons son intégrale et simplifions avec *ratsimp* et *factor* l'expression du résultat. Nous aurions pu utiliser la fonction définie au I.2.7.

```
1    c:(a+b)/2;
     L:f(a)*((x-c)*(x-b))/((a-c)*(a-b))+
3    f(c)*((x-a)*(x-b))/((c-a)*(c-b))+
     f(b)*((x-a)*(x-c))/((b-a)*(b-c));
5    factor(ratsimp(integrate(L,x,a,b)));
```

Le résultat de la dernière instruction produit :

$$\frac{(b-a)\left(4\mathrm{f}\left(\frac{b+a}{2}\right)+\mathrm{f}(b)+\mathrm{f}(a)\right)}{6}.$$

§ 3. ÉTUDE DE L'ERREUR

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

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

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

En particulier,

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

Comme de nombreux résultats d'analyse numérique, la démonstration de ce théorème est basée sur la **formule de Taylor** qui est présentée en appendice, voir le théorème A.2 .

Démonstration du Théorème 2. Posons $c = \frac{a+b}{2}$. Pour tout $x \in [a,b]$, une application de la formule de Taylor à l'ordre deux 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

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

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

$$\int_a^b f(x)dx \geq \int_a^b \left\{ f(c) + f'(c)(x-c) + m_2(x-c)^2 \right\} dx \tag{3.1}$$

$$\geq \mathrm{Q}(f) + \frac{1}{2}[(x-c)^2]_a^b + \frac{m_2}{2\cdot 3}[(x-c)^3]_a^b \tag{3.2}$$

$$\geq \mathrm{Q}(f) + 0 + \frac{m_2}{2\cdot 3}\left(\frac{(b-a)^3}{2^3}\right), \tag{3.3}$$

et il suit

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

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

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

Regroupant les deux estimations, nous tirons

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

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

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

3.2 Estimation de l'erreur dans la formule du trapèze

Théorème 3. *Soit $f \in \mathscr{C}^2([a,b])$. Il existe $\theta \in [a,b]$ tel que*

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

En particulier,

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

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.9, 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

$$\int_a^b (x-a)(x-b)dx = -(b-a)^3/6, \tag{3.4}$$

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 démonstration 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 46 Démontrer la relation (3.4).

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 60.

Théorème 4. *Soit $f \in \mathscr{C}^4([a,b])$. Nous avons*

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

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

E 47 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 $\mathbf{L}[x_0,\ldots,x_d;f]$ 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 $\sigma = (a = a_0, a_1, \ldots, a_n = b)$ de $[a,b]$ et, dans chaque intervalle $[a_i, a_{i+1}]$, $d+1$ points distincts $X^i = \{x_0^i, x_1^i, \ldots, x_d^i\}$ pour construire la formule d'approximation

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

avec

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

La localisation des points x_j^i dans chaque sous-intervalle est illustrée dans la figure 1. La relation de

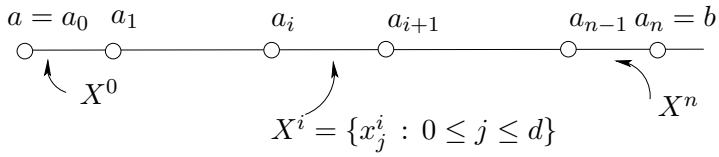


FIGURE 1 – Localisation des points dans une formule de quadrature composée

Chasles pour les intégrales nous donne

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

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

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

Toute expression Qc de la forme

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

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

Théorème 5.

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

Démonstration. Avec les notations précédentes, nous avons

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

4.2 Exemples fondamentaux de formules composées

Soit $n \in \mathbb{N}^\star$, $I = [a,b]$ et $f \in \mathscr{C}[a,b]$. Écrivons $h(n) = (b-a)/n$ 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 48 Montrer que si $\mathrm{Qc}(f)$ désigne la formule des trapèzes composées avec $n+1$ points équidistants $a(i,n) = a + i(b-a)/n$ alors

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

où σ est la subdivision $a = a(0,n) < a(1,n) < \cdots < a(n,n) = b$. Nous renvoyons à 4.I pour les informations nécessaires sur les polylignes.

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.

4.3 Codes Scilab

Nous donnons les codes correspondant aux méthodes du point milieu, des trapèzes et de Simpson pour évaluer l'intégrale Le code calcule une valeurs approchée de

$$\int_a^b \mathrm{fonc}(x)dx. \tag{4.5}$$

En particulier, dans tous les codes ci-dessus,

(a) a, b sont donc les bornes de l'intervalle,

(b) *fonc* désigne la fonction à intégrer,

Code SCILAB 2 (Méthode du point milieu). Dans le code suivant,

(a) n est le nombre sous-intervalles utilisés,

(b) *NodesPM* est le vecteur contenant les milieux de ces sous-intervalles.

```
1  function [y]=PM(a,b,n,fonc);
   h=(b-a)/n;
3  NodesPM=a+((2*[1:n]-1)*(h/2));
   TBS=feval(NodesPM, fonc); // ou TBS=fonc(NodesPM);
5  y=(b-a)*sum(TBS)/n;
   endfunction;
```

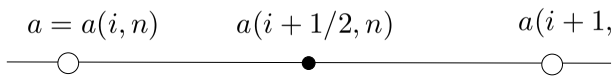
Code SCILAB 3 (Méthode du trapèze). Dans le code suivant,

(a) n est le nombre de sous-intervalles utilisés,

(b) *NodesTrap* est le vecteur contenant les points intervenant dans la méthode.

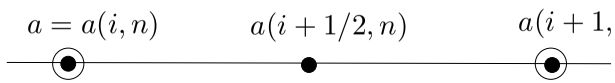
Principales formules de quadratures composées

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



Point milieu

Trapèze



Simpson

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

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 – Comparaison des diverses méthodes pour $\pi = \int_0^1 4/(1+x^2)dx$.

```
function [y]=TRAP(a,b,n,fonc);
  h=(b-a)/n;
  NodesTrap=a+h*[1:n-1];
  ValuesTrap=fonc(NodesTrap);
  y=h*(sum(ValuesTrap)+0.5*fonc(a)+0.5*fonc(b));
endfunction;
```

Code SCILAB 4 (Méthode de Simpson). Dans le code suivant,

(a) *n* est le nombre de sous-intervalles utilisés,

(b) *extrem* est le vecteur formé des extrémités des sous-intervalles.

(c) *milieux* est le vecteur formé des milieux des sous-intervalles.

```
function [y]=SIMPSON(a,b,n,fonc);
  h=(b-a)/n;
  extrem=a+h*[1:n-1];
  valextrem=fonc(extrem);
  milieux=a+(h/2)*(2*[1:n]-1);
  valmilieu=fonc(milieux);
  y=(h/6)*(fonc(a)+fonc(b)+2*sum(valextrem)+4*sum(valmilieu));
endfunction;
```

En principe, les formules d'erreur de la table 2 permettent de déterminer le nombre de sous-intervalles nécessaires pour obtenir la précision voulue. C'est le nombre *n* qui est demandé dans de nombreux exercices académiques comme d'ailleurs ici aux problèmes 56 et 57. Dans la pratique, un tel exercice est rarement réalisé. Il y a pour cela au moins deux raisons. La première c'est que l'emploi de ces formules nécessite la connaissance d'une borne de la valeur absolue de la dérivée seconde (dans le cas de la formule du point milieu et de celle des trapèzes) ou de la dérivée quatrième (dans le cas de la formule de Simpson) et ces bornes peuvent difficiles à obtenir surtout. La seconde raison, c'est que même si l'emploi des formules d'erreur est possible elles donnent très souvent un valeur de *n* exagérément grande. Si l'on ne connaît pas *a priori* le paramètre *n* employé dans les codes précédent, on utilise en général ce qui est appelé un test d'arrêt, pour fixer la valeur *n*. Un test d'arrêt communément employé est d'arrêter l'augmentation de *n* lorsque les valeurs approchées successives obtenues ne diffèrent pas plus que d'un epsilon fixé par l'utilisateur. Voici un tel code appliqué à la formule du point milieu.

Code SCILAB 5 (Un test d'arrêt sur la méthode du point milieu). Dans le code suivant,

(a) *myeps* est l'epsilon fixé par l'utilisateur, ici 10^{-8}.

(b) *PM* est la fonction définie au code SCILAB 2.

(c) Le résultat affiche le couple A=(nombres de sous-intervalles, valeur approchée).

```
    myeps=10^(-8);
2   k=1;
    while abs(PM(0,1,k,fonc)-PM(0,1,k+1,fonc)) >
    myeps; k=k+1; end
4   A=[k,PM(0,1,k,fonc)]
```

Ce test peut s'avérer médiocre lorsque la convergence est lente. En prenant $\varepsilon = 10^{-6}$ et $fonc(x) = 4/(1+x^2)$ sur $[0,1]$, le code précédent donne $k = 55$ et l'intégrale (qui vaut π) est approchée avec une erreur de $2.75\,10^{-5}$. Pour $\varepsilon = 10^{-8}$, $= 255$ est l'erreur est $1.28\,10^{-6}$.

Signalons en outre que l'usage d'une instruction *while* peut être dangereux. Elle peut donner lieu à une exécution infinie si la condition d'arrêt n'est jamais vérifiée. Il est judicieux d'imposer un nombre maximal d'itérations après lequel la boucle s'interrompt, que la condition du *while* soit satisfaite ou non.

E 49 Rajouter une instruction d'échappement comme indiqué ci-dessus dans le code SCILAB 5. On pourra consulter le détail du fonctionnement de la boucle *while* sur les fichiers d'aide scilab ainsi que de l'instruction *break* *.

§ 5. EXERCICES ET PROBLÈMES

50 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.

51 Une expression de l'erreur dans la formule des trapèzes. Soit f une fonction de classe $\mathscr{C}^2$ 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.

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

$$\int_a^b \mathbf{L}[x_1,x_2;f](x)dx = \frac{b-a}{2}\,[f(x_1)+f(x_2)] \tag{5.1}$$

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

53 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

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

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

*. Voir aussi l'exercice III.63 à la suite du code SCILAB 7.

(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).

54 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)

55 Quelle est la méthode utilisée dans l'approximation

$$\int_1^2 e^{-x}dx \approx h\{e^{-(1+h/2)}+e^{-(1+3h/2)}+\cdots+e^{-(1+(2n-1)h/2)}\},$$

où $h = 1/n$? Calculer le terme de droite et estimer l'erreur commise.

56 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

(5.2) $f(ta+(1-t)b) \leq tf(a)+(1-t)f(b)$.

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 (5.2) 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. 126.)

NOTE. — Voir l'exercice I.11.

57 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.

58 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 $|I - A(n,f)| \leq 10^{-3}$. 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. 127.)

59 Méthode des paraboles 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. 128.)

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

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

$$\left|\int_a^b f(t)dt - (b-a)f\left(\frac{a+b}{2}\right)\right| \leq \frac{(b-a)^3}{24} \cdot \max_{[a,b]} |f^{(2)}| \tag{5.3}$$

$$\left|\int_a^b f(t)dt - \frac{(b-a)}{2}[f(a)+f(b)]\right| \leq \frac{(b-a)^3}{12} \cdot \max_{[a,b]} |f^{(2)}|. \tag{5.4}$$

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

$$\left|\int_a^b f(t)dt - \frac{(b-a)}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]\right| \leq \frac{(b-a)^5}{2880} \cdot \max_{[a,b]} |f^{(4)}|. \tag{5.5}$$

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

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

puis

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

(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 $\bar{t} \in [0,h]$ tel que $\Phi^{(3)}(\bar{t}) = 0$.
(c) En déduire l'estimation (5.5).
(d) Démontrer (5.4) en s'inspirant de la démonstration précédente. On considérera

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

puis

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

(e) Démontrer (5.3) toujours en suivant la même technique : quelle est la fonction Ψ_{milieu} appropriée ?

§ 6. NOTES ET COMMENTAIRES

Sur le contenu

Je m'en suis tenu au strict minimum, correspondant d'ailleurs au contenu de la plupart des introductions classiques, en essayant d'unifier utilement la construction des trois quadratures élémentaires. Il serait souhaitable de traiter l'accélération de Romberg mais cela ne m'a paru réalisable sans étendre le cours plus qu'il n'était possible. De mon point de vue, l'approche la plus naturelle, sinon la plus simple, dans l'étude des erreurs est celle qui passe par l'analyse du noyau de Peano, mais cette approche ne peut être envisagée que dans un second cours d'analyse numérique. Les démonstrations de ce chapitre sur les estimations des erreurs présentent au moins l'intérêt de mettre en application des théorèmes classique d'analyse élémentaire.

Sur les exercices

Les quadratures pour lesquelles toutes les valeurs $f(t_i)$ sont affectées du même coefficient s'appellent des **quadratures de Chebyshev** ; l'exercice 45 donne le seul exemple simple, il est tiré de Démidovitch and Maron (1979). J'ai emprunté l'exercice 55 à Paterson (1991). La démonstration des formules d'erreur esquissée à l'exercice 60 est assez populaire, spécialement, m-a-t-il semblé, chez les non-spécialistes d'analyse numérique. J'ai lu ces démonstrations pour la première fois dans le traité d'analyse de Hardy (1952) d'où je les ai tirées. L'exercice 59 sur les paraboles à chevauchement (anecdotique) provient de Davis (1975) qui est l'un des deux grands traités classiques sur le calcul approché des intégrales, l'autre étant celui de Krylov (1962). Les exercices 52 et 53 peuvent être traités avec Maxima comme dans le code 2 et cette manière est recommandée.

Sur les difficultés

L'exercice sur lequel j'interrogeais presque toujours les étudiants reprenait les questions formulées au début de l'exercice 57 ou encore, typiquement, était formulé par un énoncé comme celui-ci

> On veut déterminer une approximation J de
>
> $$I = \int_1^2 \sqrt{x} e^x dx,$$
>
> à l'aide de la méthode des trapèzes (composée) de telle sorte que $|I - J| < 10^{-3}$. Donner une estimation du nombre d'opérations $(+, -, \times, \div,$ calcul d'une racine carrée, calcul d'une exponentielle) nécessaire.

Il s'agit dans un premier temps de déterminer un nombre de sous-intervalles suffisant à garantir que le terme de droite dans la formule d'erreur soit plus petit que la borne indiquée. Pour procéder ainsi, il est nécessaire de disposer d'une borne supérieure, ici, de la valeur absolue de la dérivée seconde de la fonction. Ce travail présente une difficulté importante pour beaucoup d'étudiants. Malheureusement ce type d'exercice — outre que, s'il reproduit une démarche rationnelle, elle est d'un intérêt limité du point de vue du numéricien qui trouvera généralement plus efficient de se faire une idée de la précision des résultats en analysant des séries de valeurs obtenues en faisant croitre le nombre de sous-intervalles — focalise indûment l'esprit des étudiants sur les formules d'erreur, leur fait perdre de vue que ces estimations, basées sur des inégalités triangulaires répétées, produisent des estimations pessimistes, et, ce qui est plus grave, induisent l'idée que le nombre suffisant de sous-intervalles est le nombre nécessaire ou encore des jugements absurdes comme « la méthode du point du point milieu est meilleure que celle des trapèzes car... » [*]. Un travail expérimental intéressant à mener avec des étudiants motivés serait précisément de mettre en évidence une relation entre l'erreur réelle et l'erreur maximale prédite pour une quadrature donnée en considérant un grand échantillon de fonctions.

∗. Voir aussi le commentaire suivant le Code Scilab 4.

III

Équations numériques

§ 1. Introduction

Le problème de construire une suite convergente vers la solution d'une équation numérique est certainement à l'origine des plus anciens algorithmes mathématiques. L'exemple le plus célèbre est l'algorithme attribué à Héron d'Alexandrie mais qui était vraisemblablement connu des mathématiciens babyloniens. Cet algorithme fournit une approximation rapide de $\sqrt{a}$, $a > 0$, qui est l'unique solution positive de l'équation $x^2 - a = 0$. Les équations numériques interviennent fréquemment en mathématiques. Des questions fondamentales qui apparaissent dans de nombreux contextes scientifiques comme la détermination de points d'intersection de graphes ou celle des extremums de fonctions numériques (qui passe par le calcul des points d'annulation de la dérivée) conduisent à la résolution d'équations. Les méthodes avancées d'analyse numérique comportent aussi souvent des résolutions d'équations numériques, notamment dans l'étude des équations différentielles *.

De manière précise, ce chapitre est consacré au problème suivant. Étant donnée une fonction continue $f : [a,b] \to \mathbb{R}$, nous cherchons les réels x dans $[a,b]$ satisfaisant $f(x) = 0$. Un tel réel s'appelle une solution de l'équation $f(x) = 0$. On dit aussi, surtout lorsque f est un polynôme, que x est un zéro de f ou encore une racine de f. En principe, trois questions se posent.

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

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

(c) 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.

L'étude des points (a) et (b) ne sera abordée que dans la dernière partie de ce chapitre. Ailleurs, nous supposerons que l'équation admet une et une seule solution dans $[a,b]$, ou bien cela découlera immédiatement de nos hypothèses. Rappelons que, dans les cas simples, une étude élémentaire de fonction (avec tableau de variation) permet de s'assurer si cette hypothèse est satisfaite ou non. Parmi le grand nombre de méthodes disponibles pour répondre au troisième point, nous étudierons quatre techniques 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**.

Signalons que les méthodes que nous allons exposer sont si naturelles, elles ont une signification géométrique si simple, qu'il est pratiquement impossible d'en tracer l'origine. D'ailleurs, il n'y a pas de terminologie fixe. Ce qui dans un texte est appelé *méthode de la sécante* portera dans l'autre le nom de *méthode des cordes*, celle de *dichotomie* ou *bissection* s'appelle parfois, notamment dans les textes en langue anglaise, la *méthode d'encadrement*.

*. Une équation différentielle est une équation dont l'inconnue est une fonction et la relation déterminant l'inconnue fait intervenir la dérivée de cette inconnue.

§ 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 ;

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

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

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 a) et b) 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 $n\geq 1$,

(b1) $c_n=\frac{a_n+b_n}{2}$,

(b2) 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

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

Voici le code scilab pour la méthode de dichotomie.

Code SCILAB 6 (Algorithme de dichotomie). La fonction suivante donne une valeur approchée d'une solution de l'équation $fonc(x) = 0$ sur l'intervalle $[exg\,,\, exd]$ en appliquant la méthode de dichotomie avec n itérations. Le code suppose que $fonc(exg) \cdot fonc(exd) < 0$.

```
function [y]=DICHO(exg,exd,fonc,n);
  a=exd;
  b=exg;
  for i=1:n;
    y=(a+b)*0.5;
    s=sign(fonc(y)*fonc(b));
    if s==-1 then;
      a=y;
      else;
      b=y;
    end;
  end;
endfunction;
```

E 61 Quelle est le résultat donné par la fonction DICHO si l'hypothèse $fonc(exg) \cdot fonc(exd) < 0$ n'est pas satisfaite ? Rajouter un test au début du code pour s'assurer que la condition est vérifée et imprimant 'Attention : la fonction ne satisfait la condition des signes' dans le cas contraire. Pour l'instruction d'impression utiliser *disp('Attention : la fonction ne satisfait la condition des signes.')*.

Grâce au théorème 2, il suffit de rajouter une ligne de code pour obtenir une fonction *DICHEPS* (*exg*, *exd*, *fonc*, *myeps*) qui donne une approximation avec une précision *myeps* fixée par l'utilisateur.

E 62 Ecrire cette fonction.

Le code suivant adopte une construction un peu différente en faisant à un instruction du type 'tant que' (while).

Code SCILAB 7 (Algorithme de dichotomie à précision fixée). La fonction suivante donne une valeur approchée par la méthode de dichotomie d'une solution de l'équation $fonc(x) = 0$ sur l'intervalle $[exg\,,\, exd]$ avec une erreur au plus égale à *myeps*. Le code suppose que $fonc(exg) \cdot fonc(exd) < 0$.

```
function [y]=DICHOEPS(exg, exd, fonc, myeps);
  a=exg;
  b=exd;
  k=0;
    while abs((exd-exg)/2^{k+1})> myeps;
        y=(a+b)*0.5;
        s=sign(fonc(y)*fonc(b));
        if s==-1 then;
            a=y;
            else;
            b=y;
         end;
    k=k+1;
    end;
endfunction
```

E 63 A la suite du code SCILAB II.5, nous avons signalé un danger éventuel lié à l'utilisation des boucles *while*, voir l'exercice II.49. Pourquoi cette mise en garde n'a-t-elle pas lieu d'être ici ?

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 premiers termes de la suite et la seconde colonne en reporte les valeurs.

(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 64 Montrer que l'algorithme 1 fonctionne encore si l'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.

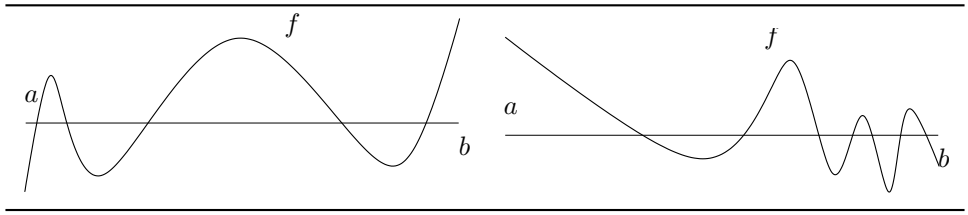


FIGURE 1 – Vers laquelle des racines convergera l'algorithme de dichotomie ?

§ 3. MÉTHODE DE NEWTON

3.1 Construction

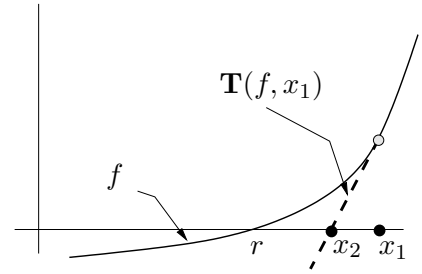
Supposons que $f\in\mathscr{C}^1([a,b])$ et que l'équation $f(x)=0$ admette une et une seule racine, notée r, dans $[a,b]$. L'idée de la **méthode de Newton** consiste à remplacer l'équation $f(x)=0$ par l'équation $T_1(x)=0$ où T_1 est le polynôme de Taylor de f de degré 1 en un point x_1. L'idée est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

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

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



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 que 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 en résolvant 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 s'appelle le schéma de Newton. Le mot *schéma* est un synonyme ancien du mot *algorithme*.

E 65 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 **??**.

Théorème 3. *Soit f une fonction de classe $\mathscr{C}^2$ 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.

Les valeurs de l'exemple précédent sont calculées avec le code suivant.

Code SCILAB 8 (Méthode de Newton). La fonction suivante applique le schéma de Newton à l'équation $fonc(x)=0$.

(a) *dfonc* est la dérivée de *fonc*.

(b) *a* est le point de départ ($a=x_0$).

(c) *n* est le nombre d'itération (l'algorithme produit x_n).

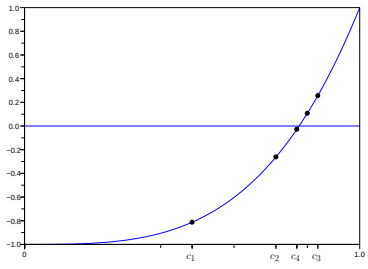
Aucune condition n'étant testée, il n'y a pas garantie de convergence.

```
function [y]=NEWTON(a,fonc,dfonc,n);
    y=a;
    for i=1:n;
    y=y-fonc(y)/dfonc(y);
    end
endfunction
```

E 66 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

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



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

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

$x^4+x^3-1=0$	$x-\sin x-1/4=0$
$x\in[0,1]$	$x\in[0,\pi/2]$

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

$3x^5-x^4-1=0$			$f(x)=x-\sin(x)-1/4$		
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

Ensuite, en utilisant la formule de Taylor de f à l'ordre 2 en x_0 (théorème A.2), 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 propriété P_{n+1},

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

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 (théorème II.2), 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

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

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

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

établie 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 est comparable à 10^{-3}, au rang $n+1$, elle sera au pire comparable à 10^{-6}, au rang $n+2$, à 10^{-12} et ainsi de suite.

E 67 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. Le lecteur écrira la définition de la suite de Newton adaptée à chaque cas.

3.4 Calcul formel

Les logiciels de calcul formels sont bien adaptés au calcul des suites de Newton puisqu'ils peuvent, dans la plupart des cas classiques, produire les dérivées des fonctions considérées. Il permet aussi, comme nous allons le voir, de mettre en évidence l'aspect fonctionnel (le caractère global) de l'approximation fournie.

Code MAXIMA 3 (Formule d'interpolation de Lagrange). Dans le code suivant

(a) f est une expression (et non une fonction) de la variable x.

(b) *init* est le point de départ (x_0).

(c) La fonction retourne le n-ème élément de la suite de Newton.

```
newton(f,init,n):=block(df:diff(f,x),new:init,
    for i:1 thru n do
                    new:ratsimp(new-ev(f,[x=new])/ev
                        (df,[x=new])),
    new);
```

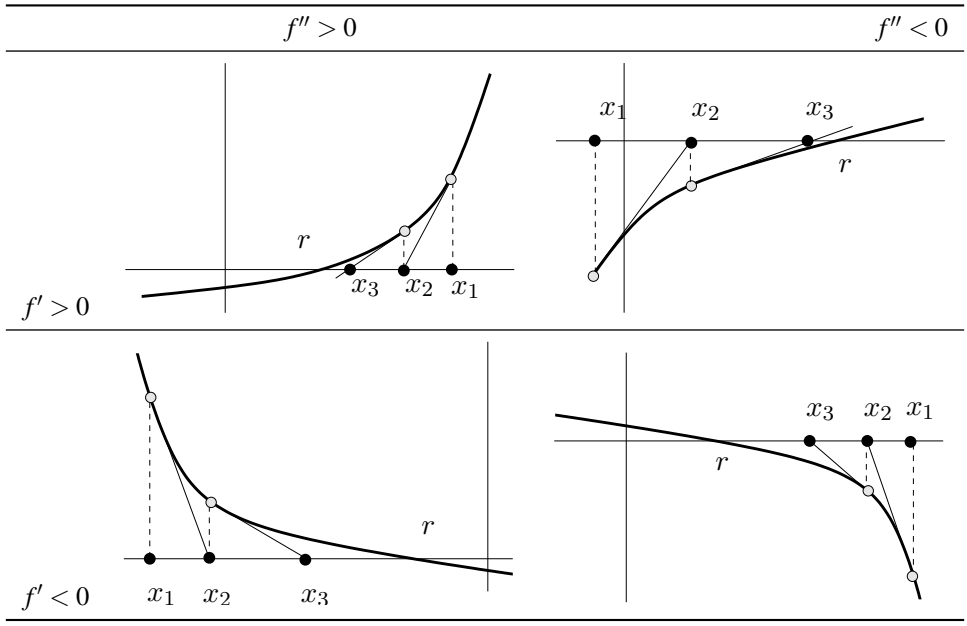
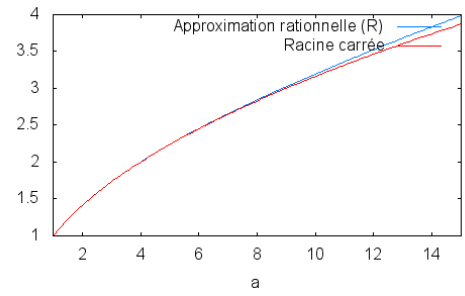


TABLE 3 – Quatre schémas de Newton convergents.

En appliquant l'algorithme avec $g(x) = x^2 - a$ avec $x_0 = a$ et $a > 0$, nous obtenons une approximation de $\sqrt{a}$, voir l'exercice 80. Il est facile de montrer, par exemple en utilisant une démonstration par récurrence, que le n-ième terme de la suite de Newton sera une fraction rationnelle en $x_0 = a$. Le code ci-dessus explicite cette fraction rationnelle. Par exemple, le calcul de $newton(g,a,3)$ donne

$$R(a) = \frac{a^4 + 28a^3 + 70a^2 + 28a + 1}{8a^3 + 56a^2 + 56a + 8}.$$

De fait la fonction R, obtenue simplement avec trois itérations de la suite de Newton, fournit une excellente approximation de la fonction racine carrée au moins sur l'intervalle $[1,6]$ comme le suggère le graphe ci-contre.



§ 4. MÉTHODE DE LA SÉCANTE

4.1 Construction

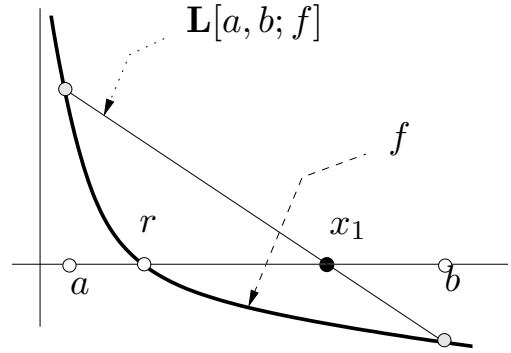
Supposons à nouveau que $f \in \mathscr{C}[a,b]$, que l'équation $f(x) = 0$ admette une et une seule solution r dans $[a,b]$ et enfin que $f(a) < 0$, $f(b) > 0$. Comme dans la méthode de Newton, la **méthode de la sécante** consiste à remplacer l'équation $f(x) = 0$ par une équation polynomiale du premier 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

$$\mathbf{L}[a,b;f](x) = f(a)\frac{x-b}{a-b} + f(b)\frac{x-a}{b-a} = f(a) + \frac{f(b)-f(a)}{b-a}(x-a), \tag{4.1}$$

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

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

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 pour l'instant à considérer $\mathbf{L}[x_1,b;f](x)=0$ dont 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 68 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 65.

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 **??**.

Théorème 4. *Soit f une fonction de classe $\mathscr{C}^2$ 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

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

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 82 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

$$a < x_n \leq x_{n+1} < r < b, \quad n \in \mathbb{N}^\star. \tag{4.2}$$

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

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

Il reste à utiliser le théorème sur l'erreur dans l'interpolation de Lagrange (th. I.9) qui nous donne

$$f(x_n) - L[x_{n-1}, b; f](x_n) = \frac{1}{2}\,(x_n - x_{n-1})(x_n - b) f''(\xi), \tag{4.4}$$

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.

Code SCILAB 9 (Méthode de la sécante). La fonction suivante applique le schéma de la sécante (avec extrémité droite fixe) à l'équation $fonc(x) = 0$.

(a) *exd* est le point fixe de la sécante (ici, l'extrémité droitre).

(b) a est le point de départ ($a = x_0$).

(c) n est le nombre d'itération (l'algorithme produit x_n).

Aucune condition n'étant testée, il n'y a pas garantie de convergence.

```
function [y]=SECANTE(a,exd,fonc,n);
    y=a;
    b=exd
    for i=1:n;
    y=(y*fonc(b)-b*fonc(y))/(fonc(b)-fonc(y));
    end
endfunction
```

E 69 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 70 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. Il est de tous les procédés considérés dans ce chapitre, le plus fondamental. Il permet d'unifier leur traitement, cela sera expliqué dans le complèment proposé plus bas.

5.2 Énoncé 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*

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

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

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

(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

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

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

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

avec $0 \leq k < 1$, 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 $\forall \varepsilon > 0$, $\exists \eta > 0$ tel que les conditions ($|x-x_0| \leq \eta$ et $x \in I$) impliquent $|g(x)-g(x_0)| \leq \varepsilon$. Or ε était fixé, il suffit de prendre $\eta = \varepsilon/k$.

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

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

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

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

Bien entendu, il existe des fonctions fonctions contractantes non dérivables. La fonction $g(x) = |x|/2$, $x \in \mathbb{R}$ en est un exemple très simple.

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 mais sans doute plus instructive. Elle est traitée en complément à la section 6.

5.3 Illustration graphique

La figure 2 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 $(r, f(r)) = (r,r)$.

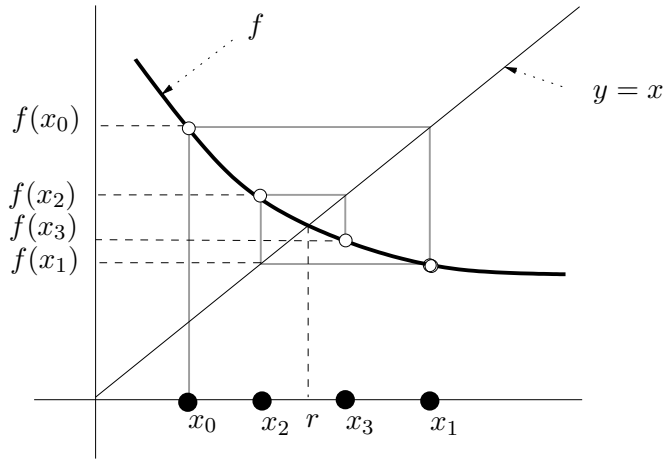


FIGURE 2 – Construction des quatre premiers termes d'un suite d'approximations successives.

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.

Code SCILAB 10 (Approximations successives). La fonction suivante applique le schéma des approximations successives à l'équation $fonc(x) = x$.

$f(x)=x-\sin(x)-1/4$			$f(x)=x-\sin(x)-1/4$		
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

(a) a est le point de départ ($a=x_0$).

(b) n est le nombre d'itération (l'algorithme produit x_n).

Aucune condition n'étant testée, il n'y a pas garantie de convergence.

```
1   function [y]=AS(a,fonc,n);
    y=a;
3   for i=1:n;
        y=fonc(y);
5   end;
    endfunction
```

E 71 Montrer que la fonction f définie par $f(x)=\sin(x)+1/4$ satisfait aux 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 $|s_1-s_2|=|g(s_2)-g(s_2)|\leq k|s_1-s_2|$ 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) $g(a)\in[a,b]\implies g(a)\geq a\implies f(a)\geq 0$.

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

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

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

■

Lemme 7. $\forall q > p \geq 0$,

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

Démonstration.

$$\begin{aligned}\left|x_q - x_p\right| &= \left|x_q - x_{q-1} + x_{q-1} - x_{q-2} + \cdots + x_{p+1} - x_p\right| \\ &\leq \left|x_q - x_{q-1}\right| + \left|x_{q-1} - x_{q-2}\right| + \cdots + \left|x_{p+1} - x_p\right| \\ &\leq (k^{q-1} + k^{q-2} + \cdots + k^p)\left|x_1 - x_0\right| \quad \text{(d'après le Lemme 6)} \\ &\leq k^p(k^{q-1-p} + k^{q-2-p} + \cdots + k^1 + 1)\left|x_1 - x_0\right| \\ &\leq k^p \frac{1-k^{q-p}}{1-k}\left|x_1 - x_0\right| \leq \frac{k^p - k^q}{1-k}\left|x_1 - x_0\right|. \quad ■\end{aligned}$$

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. Elle est donnée dans le paragraphe suivant.

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

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

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

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

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

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

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 $\varepsilon > 0$, il existe un entier $N = N_\varepsilon$ tels que $q > p > N \implies |x_p - x_q| \leq \varepsilon$. 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 $\varepsilon > 0$ et choisissons N de telle sorte que $\frac{k^N}{1-k}\left|x_1 - x_0\right| \leq \varepsilon$ 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 $\frac{k^N}{1-k}\left|x_1 - x_0\right|$ 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 7, nous avons

$$(5.2)\quad \left|x_q - x_p\right| \leq \frac{k^p - k^q}{1-k}\left|x_1 - x_0\right| \leq \frac{k^p}{1-k}\left|x_1 - x_0\right| \leq \frac{k^N}{1-k}\left|x_1 - x_0\right| \leq \varepsilon.$$

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.

5.6 Le problème de la stabilité dans les approximations successives

Dans le théorème 5 du point fixe, nous devons calculer des suites à récurrence simple $x_{n+1} = g(x_n)$. Naturellement, dans le calcul sur machine, il ne sera pas possible d'obtenir exactement $g(x_n)$ et donc x_{n+1} mais seulement une *approximation* de $g(x_{n+1}$; disons que le calcul produira $g(x_n) + \varepsilon_n$ où ε_n est l'erreur de calcul due à la machine. En général, cette erreur ε_n, très petite, dépend non seulement du calculateur mais aussi de la fonction g elle-même. Pour simplifier l'étude, nous supposerons que ε_n se trouve dans l'intervalle $]-\varepsilon,\varepsilon]$. La suite effectivement calculée par la machine est alors de la forme

$$X_{n+1} = g(X_n) + \varepsilon_n, \quad n \in \mathbb{N}.$$

Même en nous plaçant dans les hypothèses du théorème 5, nous ne pouvons plus espérer que cette suite converge vers la solution s de l'équation $g(x) = x$. Le seul objectif que nous pouvons nous fixer, c'est que pour n grand, la valeur de X_n ne s'éloignera de la solution cherchée s que par une distance comparable, en un certain sens, à la marge d'erreur ε. Le théorème suivant montre que, sous certaines hypothèses, cet objectif est atteint.

Théorème 8. *Nous supposons que les hypothèses du théorème 5 sont satisfaites. Nous supposons en outre que l'intervalle fermé I sur lequel est définie la fonction g contient l'intervalle $J_\rho = [s-\rho, s+\rho]$ où s est la solution de l'équation $g(x) = x$ et que, de plus,*

$$|X_0 - x| \leq \rho_0 \leq \rho - \frac{\varepsilon}{1-k}. \tag{5.3}$$

Alors tous les éléments de la suite

$$X_{n+1} = g(X_n) + \varepsilon_n, \quad n \in \mathbb{N}, \tag{5.4}$$

avec $-\varepsilon \leq \varepsilon_n \leq \varepsilon$, sont bien définis et satisfont la relation

$$|X_n - s| \leq \frac{\varepsilon}{1-k} + k^n \left(\rho_0 - \frac{\varepsilon}{1-k}\right). \tag{5.5}$$

Remarquons que l'hypothèse sur ρ exige de connaitre une première approximation de s ; celle-ci étant connue, l'hypothèse (5.3) requiert de l'améliorer seulement légèrement puisque le terme $\frac{\varepsilon}{1-k}$ est généralement petit devant ρ.

Démonstration. Observons qu'il suffit d'établir l'inégalité 5.5 puisque, k étant plus petit que 1, nous avons

$$\frac{\varepsilon}{1-k} + k^n \left(\rho_0 - \frac{\varepsilon}{1-k}\right) < \frac{\varepsilon}{1-k} + \left(\rho_0 - \frac{\varepsilon}{1-k}\right) = \rho,$$

ce qui implique que X_{n+1} est bien défini. Pour établir l'inégalité 5.5, nous procédons de la manière suivante. Supposant que l'inégalité est établie jusqu'au rang n, nous avons

$$|X_{n+1} - s| = |g(X_n) - g(s) + \varepsilon_n| \leq k|X_n - s| + \varepsilon.$$

En appliquant la même technique de majoration à $|X_n - s|$, nous arrivons à

$$|X_{n+1} - s| \leq k^2|X_{n-1} - s| + k\varepsilon + \varepsilon;$$

et, en continuant, autant qu'il est possible, nous tirons finalement l'inégalité

$$\begin{aligned} |X_{n+1} - s| \leq k^{n+1}|X_0 - s| + \varepsilon(1 + k + k^2 + \cdots + k^n) &\leq k^{n+1}\rho_0 + \varepsilon\frac{1-k^{n+1}}{1-k} \\ &\leq k^{n+1}\left(\rho_0 - \frac{\varepsilon}{1-k}\right) + \frac{\varepsilon}{1-k}, \end{aligned} \tag{5.6}$$

qui est l'inégalité recherchée au rang $n+1$. ∎

E 72 Le théorème ci-dessus montre qu'il est raisonnable d'espérer construire une approximation X_n de s pour laquelle nous aurons $|X_n - s| < 2\varepsilon/(1-k)$. En combien d'itération ?

§ 6. ∗ DAVANTAGE SUR LE THÉORÈME DU POINT FIXE ET SES APPLICATIONS

Dans cette partie nous montrons comment, correctement modifié, avec des hypothèses à la fois précisées et affaiblies, le théorème du point fixe permet d'étudier des techniques importantes de résolution d'équations numériques de la forme $f(x) = 0$ comme les méthodes de Newton ou de la sécante que nous avons abordées plus haut d'un autre point de vue.

6.1 Sur la rapidité de convergence

La première critique que suscite le théorème 5, c'est que la vitesse de convergence de la suite (x_n) fournie par les approximations successives est décevante. En se basant sur cet énoncé, pour avoir un résultat théoriquement meilleur que celui donné par la méthode élémentaire de dichotomie (théorème 2), nous devrions nous assurer que $\sup_{x\in I} |g'(x)| < 1/2$, ce qui est certainement une hypothèse trop forte. Le théorème suivant montre qu'en réalité l'estimation donnée dans le théorème 5 est pessimiste.

Théorème 9. *Supposons que la suite x_n définie par la relation de récurrence $x_{n+1} = g(x_n)$ avec $x_0 = z \in I$ converge vers s telle que $g(s) = s$. Supposons en outre que la fonction g soit dérivable au point s. Nous avons*

$$\lim_{n\to\infty} \frac{|x_{n+1} - s|}{|x_n - s|} = |g'(s)|. \tag{6.1}$$

Cet énoncé suppose que $x_n \neq s$, $n \in \mathbb{N}$.

Démonstration. Observant que $x_{n+1} - s = g(x_n) - g(s)$, il vient

$$\frac{x_{n+1} - s}{x_n - s} = \frac{g(x_n) - g(s)}{x_n - s},$$

et, puisque nous savons que $x_n \to s$, la conclusion résulte immédiatement de la définition de la dérivée de g en s et de la continuité de l'application $\|\cdot\|$. ■

Corollaire 10. *Sous les hypothèses du théorème 9, pour tout $\mu > |g'(s)|$, nous avons $|x_n - s| = 0(\mu^n)$.*

Rappelons que nous disons qu'un suite u_n est égale à $0(\mu^n)$ s'il existe une constante M telle que $u_n \leq M\mu^n$ pour tout $n \in \mathbb{N}$ ou, ce qui revient au même, pour tour n à partir d'un certain rang [*]

Démonstration. Puisque $\mu > |g'(s)|$, il existe n_0 tel que

$$\frac{|x_{n+1} - s|}{|x_n - s|} \leq \mu, \quad n \geq n_0.$$

En appliquant autant de fois que possible cette inégalité, nous obtenons,

$$[x_{n+1} - s| \leq \mu|x_n - s| \leq \mu^2|x_{n-1} - s| \leq \cdots \leq \mu^{n-n_0+1}|x_{n_0} - s|, \quad n \geq n_0.$$

Nous avons établi

$$|x_{n+1} - s| \leq M\mu^N, \quad n \geq n_0,$$

avec $M = |x_{n_0} - s|/\mu^{n-n_0+1}$, ce qu'il fallait établir. ■

∗. Si nous savons que $u_n \leq M\mu^n$ pour $n \geq n_0$ alors pour tout $n \in \mathbb{N}$, $u_n \leq M'\mu^n$ avec $M' = \max\{M, u_0/\mu^0, u_1 : \mu^1, \ldots, u_{n_0-1}/\mu^{n_0-1}\}$.

Ce que le théorème 9 met en évidence, c'est que ce n'est pas la quantité $\sup_{x\in I}|g'(x)|$ (ou la constante de contraction) qui détermine la rapidité de convergence de la suite x_n du théorème 5 vers s mais la seule valeur de la dérivée de la fonction g au point s. Naturellement, cette valeur, le plus souvent, sera inférieure (en valeur absolue) à $\sup_{x\in I}|g'(x)|$ et la rapidité de convergence meilleure que celle garantie par le théorème 5. L'énoncé du théorème 9 suggère aussi que dans le cas, que nous pouvons supposer exceptionnel, où $g'(s)$ serait égal à zéro, la convergence pourrait être encore plus rapide. Il en est bien ainsi comme le montre le théorème suivant.

Théorème 11. *Supposons que la suite x_n définie par la relation de récurrence $x_{n+1}=g(x_n)$ avec $x_0=z\in I$ converge vers s telle que $g(s)=s$. Supposons en outre que la fonction g soit $d+1$ fois continûment dérivable ($d\geq 0$) sur un voisinage*[a] *$\mathscr{V}$ de s avec*

$$g'(s)=g''(s)=\cdots=g^{(d)}(s)=0.$$

Nous avons

$$\lim_{n\to\infty}\frac{|x_{n+1}-s|}{|x_n-s|^{d+1}}=\frac{|g^{(d+1)}(s)|}{(d+1)!}. \tag{6.2}$$

a. Un voisinage $\mathscr{V}$ de s ici n'est autre qu'un intervalle contenant s dans son intérieur.

Démonstration. Nous utilisons la formule de Taylor (théorème II.2) pour $g(x_n)$ au point s à l'ordre d,

$$g(x_n)=g(s)+g'(s)(x_n-s)+\cdots+\frac{g^{(d)}(s)}{d!}(x_n-s)^d+\frac{g^{(d+1)}(\xi_n)}{d!}(x_n-s)^{d+1}, \tag{6.3}$$

où ξ_n est un réel compris entre x_n et s. En utilisant l'annulation des dérivés de g en s, le fait que $g(x_n)=x_{n+1}$ et $g(s)=s$, nous arrivons à $x_{n+1}-s=\frac{g^{(d+1)}(\xi-n)}{d!}(x_n-s)^{d+1}$, puis

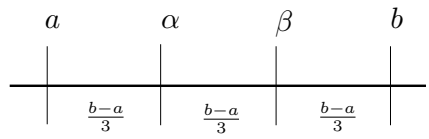
$$\frac{x_{n+1}-s}{(x_n-s)^{d+1}}=\frac{g^{(d+1)}(\xi_n)}{d!}.$$

La conclusion du théorème s'en déduit en faisant $n\to\infty$ dans cette expression. En effet puisque ξ_n est toujours compris entre s et x_n qui tend vers s, ξ_n lui-même tend vers s et la continuité de $g^{(d+1)}$ en s assure alors $g^{(d+1)}(\xi_n)\to g^{(d+1)}(s)$ lorsque $n\to\infty$. ■

Nous pourrions penser que puisque la valeur de s n'est pas connue — il s'agit précisément d'en fournir une approximation — le théorème précédent ne pourrait servir qu'à expliquer *a posteriori* une convergence particulièrement rapide de la suite (x_n) fournie par le théorème du point fixe. Nous verrons cependant plus avant qu'il peut aussi être employé, sans connaître précisément la valeur de s, dans l'étude théorique des suites de Newton et de la sécante. Nous nous occuperons d'abord d'une autre difficulté liée à l'application du théorème du point fixe.

6.2 Sur l'hypothèse de stabilité de l'intervalle

Dans la pratique, plus encore que le fait de devoir montrer qu'elle est contractante c'est-à-dire k-lipschitzienne avec $k<1$, il est difficile de s'assurer que la fonction g laisse stable l'intervalle I ($g(I)\subset I$). Nous allons voir comment il est possible de s'affranchir de cette hypothèse.



Théorème 12. *Soit g une fonction dérivable sur l'intervalle $[a,b]$ avec $|g'| \leq q < 1$ sur $[a,b]$. Décomposons $[a,b]$ en trois intervalles égaux de longueurs $(b-a)/3$ et appelons $[\alpha,\beta]$ l'intervalle intermédiaire (voir le schéma ci-après). Nous supposons en outre que g admet un point fixe s dans $[\alpha,\beta]$. Alors*

(a) La suite (x_n) définie par $x_{n+1} = g(x_n)$ et $x_0 = z \in [\alpha,\beta]$ est bien définie.

(b) La suite (x_n) converge vers s qui est l'unique solution de l'équation $g(x) = x$ sur $[a,b]$.

(c) Les estimations des théorèmes 5 (avec $k = q$), 9 et 11, avec leurs hypothèses respectives, sont valables.

Ce résultat ne suppose plus que l'intervalle de départ de la fonction g soit stable. Le prix à payer pour la suppression de cette hypothèse est que l'existence de la racine s ne découle plus du théorème mais en devient une hypothèse. Nous devons même supposer que cette racine se trouve suffisamment à l'intérieur de l'intervalle. En analyse numérique, l'ajout de cette hypothèse n'est pas dramatique. Nous pouvons souvent établir l'existence d'une racine par une simple application du théorème des valeurs intermédiaires, par exemple en mettant en évidence une valeur positive et une valeur négative de la fonction $x \mapsto g(x) - x$.

Démonstration. Nous montrons que la suite (x_n) est bien définie, c'est-à-dire que sous les hypothèses données, x_n se trouve toujours dans l'intervalle $[a,b]$. Les démonstrations des autres assertions sont en tous points identiques à celles données pour le théorème 5. Pour établir que la suite est bien définie, nous montrons par récurrence que $|x_n - s| \leq (b-a)/3$ pour tout n. Pour $n = 0$, l'affirmation fait partie des hypothèses. En effet, puisque x_0 et s sont dans $[\alpha,\beta]$, $|x_0 - s| \leq \beta - \alpha = (b-a)/3$. Supposons que $|x_n - s| \leq (b-a)/3$, nous avons $|x_{n+1} - s| = |g(x_n) - g(s)| \leq q|x_n - s| \leq |x_n - s| \leq (b-a)/3$. Enfin, la relation $|x_n - s| \leq (b-a)/3$ assure que $x_n \in [a,b]$ puisque, par construction, l'intervalle de centre s et de rayon $(b-a)/3$ est inclus dans $[a,b]$. ■

Théorème 13. *Soit g une fonction continûment dérivable sur un intervalle $[a,b]$ admettant un point fixe s dans l'intérieur de $[a,b]$. Si $|g'(s)| < 1$ alors il existe un intervalle J dans $[a,b]$ tel que pour tout point $z \in J$, la suite (x_n) définie par $x_0 = z$ et $x_{n+1} = g(x_n)$ soit bien définie et converge vers s. De plus, si g est $(d+1)$-fois continûment dérivable et $0 = g'(s) = \cdots = g^{(d)}(s)$ la converge de (x_n) vers s est d'ordre $d+1$ en ce sens que*

$$|x_{n+1} - s[= O(|x_n - s|^{d+1}).$$

Ici nous avons encore affaibli l'hypothèse sur la fonction g. Nous ne supposons plus que la dérivée de la fonction g soit bornée par $q < 1$ sur la totalité de l'intervalle mais en un unique point, le point fixe s dont nous supposons encore l'existence. Là encore, comme il est naturel, une diminution des hypothèses induit une diminution de la précision de la conclusion. Ainsi nous avons simplement l'existence de l'intervalle J mais ne nous ne savons rien a priori de l'endroit où il se trouve.

Démonstration. Il suffit de trouver un intervalle $I = [a',b']$ de centre s pour lequel $\max_{t\in I} |g'(t)| \leq q < 1$. Dans ce cas, nous pouvons appliquer le théorème précédent qui nous donnera $J = [\alpha,\beta]$ avec $\alpha = a' + (b'-a')/3$ et $\beta = b' - (b'-a')/3$, l'intervalle J ainsi construit contenant bien s. L'estimation sur la rapidité de convergence se démontrerait ensuite comme dans la démonstration du théorème 11. Maintenant, l'existence de l'intervalle I découle de la continuité * de g' en s. ■

∗. Prenons q entre $|g'(s)|$ et 1, il existe $\eta > 0$ tel que $x \in [s-\eta, s+\eta]$ entraîne $|g'(s)| \leq q$.

6.3 Application à l'étude de la méthode de la sécante

Rappelons que la suite de la sécante définie au paragraphe 4.1 pour la résolution de l'équation $f(x)=0$ est donnée, sous réserve que les quantités soient bien définies, par $x_0=\alpha$ et

$$x_{n+1}=\frac{x_nf(b)-bf(x_n)}{f(b)-f(x_n)}.$$

Théorème 14. *Soit f une fonction deux fois continûment dérivable sur $[a,b]$ avec $f(a)<0<f(b)$ et admettant une unique racine s dans $[a,b]$. Si le diamètre $(b-a)$ de l'intervalle $[a,b]$ est suffisamment petit, il existe un intervalle J contenant s tel que pour tout $z\in J$, la suite (x_n) ci-dessus soit bien définie et converge vers s.*

Démonstration. Nous remarquons que la suite(x_n) est de la forme $x_{n+1}=g(x_n)$ avec $g(x)=xf(b)-bf(x)/(f(b)-f(x))$ et $g(x)=x$ équivaut à $f(x)=0$. En vertu du théorème 13, il suffit d'établir que $|g'(s)|<1$. Or

$$g'(x)=\frac{(f(b)-bf'(s))f(b)+(sf(b)f'(s))}{(f(b)-f(s))^2}.$$

En faisant $x=s$, compte tenu de $f(s)=0$, nous avons

$$g'(s)=\frac{f(b)-(b-s)f'(s)}{f(b)}.$$

Ici, la formule de Taylor nous dit que le numérateur est $(b-a)^2f''(\xi)/2$ de sorte que

$$|g'(s)|\leq\frac{(b-a)^2}{2|f(b)|}\max_{[a,b]}|f^{(2)}|,$$

qui montre que $|g'(s)|<1$ si $b-a$ est assez petit. ■

E 73 Donner un résultat similaire pour la suite définie par $x_0=z$ et

$$x_{n+1}=x_n-\frac{b-a}{f(b)-f(a)}f(x_n),\quad n\geq 0.$$

On commencera par expliciter sur un graphique la construction de la suite (x_n). Cette suite s'appelle parfois la **méthode des cordes**. On pourra remarquer que x_{n+1} est solution de l'équation

$$0=f(x_n)+\frac{f(b)-f(a)}{b-a}(x-x_n).$$

6.4 Application à la méthode de Newton

Nous revenons maintenant à la méthode de Newton étudiée à la section 3. Toujours sous réserve que les quantités soient bien définies, la suite de Newton est définie par par $x_0=\alpha$ et

$$x_{n+1}=\frac{x_nf(b)-bf(x_n)}{f(b)-f(x_n)}.$$

par $x_0=z$ et

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)},\quad n\geq 0.$$

La même technique que celle utilisée ci-dessus permet d'établir le théorème suivant.

Théorème 15. *Il existe un intervalle J contenant s tel que pour tout $z \in J$, la suite (x_n) ci-dessus soit bien définie et converge vers s. De plus si s est une racine simple ($f'(s) \neq 0$) alors la convergence est d'ordre 2,*

$$|x_{n+1} - s| = 0(|x_n - s|^2).$$

Démonstration. Le principe est identique à celui de la démonstration du théorème 14. Nous considérons la fonction g définie par $g(x) = x - f(x)/f'(x)$. Le calcul des deux première dérivées montre que $g'(s) = 0$ puis $g''(s) \neq 0$. ■

E 74 Dans le cas où s est une racine de multiplicité $m > 1$, étudier la convergence de la suite (x_n) définie par $x_0 = z$ et

$$x_{n+1} = x_n - m\frac{f(x_n)}{f'(x_n)}, \quad n \geq 0.$$

§ 7. ∗ INTERPOLATION DE LAGRANGE ET SECONDE MÉTHODE DE LA SÉCANTE

Nous étudions ici une modification de la méthode de la sécante qui exprime x_{n+1} à l'aide de x_n et x_{n-1} de sorte que le théorème du point fixe, du moins sous sa forme élémentaire, ne s'applique pas ici ∗.

Soit $f : [a,b] \to \mathbb{R}$ une fonction deux fois continûment dérivable avec $f(a) < 0$ et $f(b) > 0$. Nous supposerons que f est strictement croissante, avec $f' > 0$ et appellerons s l'unique racine de l'équation $f(x) = 0$ dans $[a,b]$. Nous posons $x_0 = a$, $x_1 = b$ et supposant que $f(x_i)$ est définie pour $i = 2, \ldots, n$, c'est-à-dire $x_i \in [a,b]$ nous définissons x_{n+2} comme la racine du polynôme (de degré 1) $\mathbf{L}[x_n, x_{n+1}; f]$. Autrement dit x_{n+2} est déterminé par la condition

$$\mathbf{L}[x_n, x_{n+1}; f](x_{n+2}) = 0.$$

Nous appellerons cette suite, lorsqu'elle est bien définie, la **seconde méthode de la sécante**. La construction est illustrée sur la figure 7.

Le même calcul que celui qui a été fait au paragraphe 4.1 donne

$$x_{n+2} = \frac{x_n f(x_{n+1}) - f(x_n)x_{n+1}}{f(x_{n+1}) - f(x_n)} = x_n - f(x_n)\frac{x_n - x_{n+1}}{f(x_{n+1}) - f(x_n)}. \tag{7.1}$$

Puisque f est strictement croissante deux fois continûment dérivable, elle est bijective de $[a,b]$ sur $[f(a), f(b)]$ et, puisque $f' > 0$, sa bijection réciproque f^{-1} est elle-même deux fois continûment dérivable.

Lemme 16. $x_{n+2} = \mathbf{L}[f(x_n); f(x_{n+1}); f^{-1}](0)$.

Démonstration. En utilisant la formule (4.1) pour le polynôme d'interpolation de Lagrange aux points $f(x_n)$ et $f(x_{n+1})$, nous obtenons

$$\mathbf{L}[f(x_n), f(x_{n+1}); f^{-1}](0) = f^{-1}(f(x_n)) + \frac{f^{-1}(f(x_{n+1})) - f^{-1}(f(x_n))}{f(x_{n+1}) - f(x_n)} \tag{7.2}$$

$$= (0 - f(x_n)) = x_n - \frac{x_{n+1} - x_n}{f(x_{n+1}) - f(x_n)} f(x_n) = x_{n+2}. \tag{7.3}$$

■

∗. Une suite définie par une relation de récurrence de la forme $x_{n+2} = h(x_{n+1}, x_n)$ peut encore s'interpréter comme une suite de la forme $X_{n+1} = H(X_n)$ en posant $X_n = (x_n, x_{n+1})$ et $H(x,y) = (y, f(y,x))$. Il faudrait donc disposer d'un théorème du point fixe pour les fonctions de deux variables. Un tel théorème existe et même dans un cadre beaucoup plus général mais il n'est pas assez élémentaire pour être traité ici.

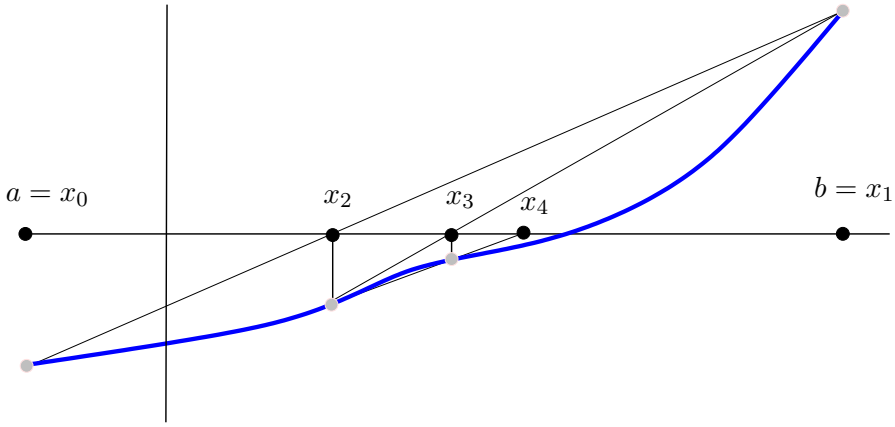


FIGURE 3 – Seconde méthode de la sécante.

Lemme 17. *Sous les hypothèses sur f indiquées plus haut, si la seconde suite de la sécante est définie jusqu'au rang $n+1$, nous avons*

$$|x_{n+2} - s| \leq K|x_{n+1} - s|\,|x_n - s|,$$

avec

$$K = K(a,b) = \frac{M_2 M_1^2}{m_1^3},$$

où

$$M_1 = \max_{x\in[a,b]} |f'(x)|, \quad M_2 = \max_{x\in[a,b]} |f''(x)|, \quad \textit{et} \quad m_1 = \min_{x\in[a,b]} |f'(x)|.$$

Démonstration. Nous tirons immédiatement du lemme 16

$$s - x_{n+2} = f^{-1}(0) - \mathbf{L}[f(x_n), f(x_{n+1}); f^{-1}](0).$$

Le terme de droite peut être estimé en utilisant le théorème sur l'erreur dans l'interpolation de Lagrange (théorème I.9) avec f^{-1} à la place de f . Nous obtenons

$$|s - x_{n+2}| \leq \max_{[f(a),f(b)]} |(f^{-1})^{(2)}|\,|0 - f(x_n)|\,|0 - f(x_{n+1}|. \tag{7.4}$$

Maintenant de la relation

$$(f^{-1})^{(2)} = \frac{f''(f^{-1})}{(f'(f^{-1}))^3}$$

nous tirons

$$\max_{[f(a),f(b)]} |(f^{-1})^{(2)}| \leq \frac{M_2}{m_1^3}. \tag{7.5}$$

D'autre part, en utilisant le théorème des accroissements finis,

$$\begin{aligned} |0 - f(x_n)|\,|0 - f(x_{n+1}| &= |f(s) - f(x_n)|\,|f(s) - f(x_{n+1}| \\ &= |s - x_{n+1}||f'(c_{n+1}|\,|s - x_n||f'(c_n| \leq M_1^2 |s - x_{n+1}||s - x_n|. \end{aligned} \tag{7.6}$$

Finalement, en utilisant (7.5) et (7.6) dans (7.4), nous arrivons à l'inégalité annoncée. ■

Corollaire 18. *Si l'intervalle $[a,b]$ est suffisamment petit alors x_n est bien défini pour toute valeur de n.*

Démonstration. Nous montrons que si a et b sont suffisamment proche pour que $K(a,b)(b-a) \leq 1$ alors x_n est bien définie pour tout $n \in \mathbb{N}$. Supposant que $x_j \in [a,b]$, $0 \leq j \leq n+1$ avec $n+1$, $n \geq 0$, nous montrons que $x_{n+2} \in [a,b]$. En effet, en vertu du lemme précédent et en utilisant d'abord le fait que x_n et s sont dans $[a,b]$ puis le fait que $K(a,b)(b-a) \leq 1$,

$$|x_{n+2}-s| \leq K|x_{n+1}-s|\,|x_n-s| \leq K|b-a| \leq K(b-a)|x_{n+1}-s| \leq |x_{n+1}-s|,$$

ce qui montre que x_{n+1} est plus proche de s que de x_n et en particulier se trouve dans $[a,b]$. ∎

Supposons que $[a,b]$ soit assez petit que la seconde suite de la constante soit bien définie. Si nous posons $u_n = \log(K|x_n - s))$, le lemme 17 se traduit par

$$u_{n+2} \leq u_{n+1} + u_n, \quad n \geq 0. \tag{7.7}$$

Lemme 19. *La suite u_n est majorée par la suite de Fibonacci définie par $v_{n+2} = v_{n+1} + v_n$, $n \geq 0$, avec $v_0 = \log(K|s-a|)$ et $v_1 = \log(K|s-b|)$.*

Démonstration. La démonstration de $u_n \leq v_n$ se fait immédiatement par récurrence sur n à partir de (7.7). ∎

Théorème 20. *Si $K(b-a) < 1$ alors il existe une constante positive $\rho < 1$ telle que*

$$|s-x_n| = 0\left(\rho^{((1+\sqrt{5})/2)^n}\right).$$

Démonstration. On démontre par récurrence qu'il existe deux λ et β tels que

$$v_n = \lambda((1-\sqrt{5})/2)^n + \beta((1+\sqrt{5})/2)^n, \quad n \in \mathbb{N}.$$

Ces deux réels sont déterminés en fonction de $v_0 = \log(K|s-a|)$ et $v_1 = \log(K|s-b|)$. Un calcul fastidieux donne

$$\alpha = \frac{\left(\sqrt{5}+5\right)v_0 - 2\sqrt{5}\,v_1}{10}, \quad \beta = -\frac{\left(\sqrt{5}-5\right)v_0 - 2\sqrt{5}\,v_1}{10}.$$

Ici il est important de noter que, grâce à l'hypothèse $K(b-a) < 1$, la seconde constante est négative. La première, α ne joue aucun rôle car $((1-\sqrt{5})/2)^n \to 0$ lorsque $n \to \infty$. Finalement, nous avons

$$\frac{u_n}{((1+\sqrt{5})/2)^n} \to \beta < 0, \quad n \to \infty.$$

Choisissons β' entre β et 0, nous avons $\frac{u_n}{((1+\sqrt{5})/2)^n} \leq \beta' < 0$ pour n assez grand, disons $n \geq n_0$. En prenant l'exponentielle nous obtenons

$$K|s-x_n| \leq \exp\left(\beta'((1+\sqrt{5})/2)^n\right), \quad n \geq n_0$$

qui donne encore, en posant $\rho = \exp(\beta') < 1$,

$$|s-x_n| \leq \frac{1}{K}\rho^{((1+\sqrt{5})/2)^n}, \quad n \geq n_0,$$

et ceci achève la démonstration du théorème. ∎

§ 8. EXERCICES ET PROBLÈMES

75 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.

76 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)

77 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 $\bar{x}_n$. Faut-il choisir $\bar{x}_0=0,5$ ou $\bar{x}_0=0,75$? Expliquer votre choix puis calculer les deux premières valeurs $(\bar{x}_1,\bar{x}_2)$.

(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. 128.)

78 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. 128.)

79 Un exemple. On considère l'équation

$$x^5-7x+4=0. \tag{8.1}$$

(a) Montrer que l'équation (8.1) 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 ?

80 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 (premier siècle après J.-C.).

81 Utiliser la méthode de la sécante pour obtenir une approximation de la fonction racine carrée comme il a été fait avec la méthode de Newton au 3.4.

82 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 $f(x) = 0$.

Le schéma ci-après montre la construction des valeurs x_1 et x_2 de la suite. Rappelons que f strictement

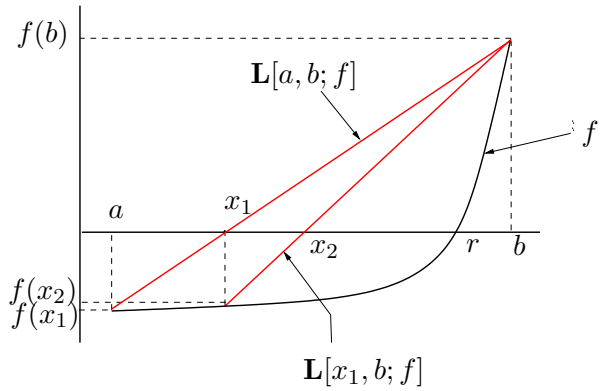


FIGURE 4 – Schéma de la sécante.

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

$$P(n) : \quad a \leq x_n < x_{n+1} < r < b.$$

(c) En déduire la démonstration du théorème.

83 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 $k \geq 2$ x_{k+1} est l'abscisse 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. Écrire un algorithme calculant les n premières valeurs de la suite x_k.

84 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 abscisses. 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 premières 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 ?

85 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}).

86 Un exemple. On considère l'équation

$$x^3 - 3x - 1 = 0. \tag{8.2}$$

L'équation (8.2) 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. 129.)

87 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.

§ 9. NOTES ET COMMENTAIRES

Sur le contenu

Ce chapitre contient deux objets fondamentaux des mathématiques en général, le principe de bissection et, plus encore, celui des approximations successives. Le second est un thème évidemment très riche qui se prête à des développements assez profonds sans quitter le domaine de l'analyse élémentaire et c'est ce qui explique la présence des sections 6 et 7 dans ce chapitre, sections dont l'étude pourra surtout intéresser le lecteur spécialisé en mathématiques. Le livre de Ralston and Rabinowitz (2001) est très riche. Il y a un chapitre notable sur ces questions dans Dieudonné (1968). Le sujet se prête bien à un analyse qui ne sera pas excessivement technique de la stabilité des algorithmes comme au 5.6 pour laquelle je me suis inspiré de Isaacson and Keller (1994), référence que je tiens, et beaucoup d'autres avec moi, comme un des meilleurs traités d'analyse numérique. J'espère compléter cette étude dans une future édition. Pour être pleinement satisfait, j'aurais aussi voulu pouvoir analyser, au moins de manière informelle, une routine de calcul de toutes les solutions d'une équation polynomiale.

J'ai essayé d'éliminer les traces des raisonnements classiques qui subsistaient dans les premières versions de ce cours et que l'on trouve encore couramment dans des introductions à l'analyse numérique, selon lesquels la méthode de dichotomie ne serait qu'une méthode de première approche, ne servant qu'à repairer le point initial d'une méthode de Newton, où encore que la méthode de la sécante doit être utilisée en parallèle avec celle de Newton pour obtenir deux suites adjacentes convergeant vers la racine cherchée. La puissance de calcul des ordinateurs modernes rend essentiellement caduque ces conseils.

Sur les exercices

La majorité des exercices consistent à demander de produire une solution approchée d'une équation donnée, après avoir vérifié que des conditions d'application figurant dans le cours étaient satisfaites. Il y a fort peu de chances qu'un numéricien procède de cette manière — la vérification numérique d'une hypothèse de convexité sur un intervalle étant d'ailleurs elle-même essentiellement aussi complexe que la résolution d'une équation numérique. Comme d'habitude, dans la pratique, la plupart des numériciens préfèreront procéder par une batterie de tests en variant des points de départ pour mettre en évidence des suites convergentes.

IV

Systèmes linéaires

§ 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, $i = 1,\dots,5$, 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 nœuds $n_{i,j}$ représentés par des disques de couleur noire, sachant que la température en un nœud donné est égale à la moyenne des températures des autres nœuds 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 nœuds n_{ij} et 25 inconnues $t(n_{ij})$, $1 \leq i, j \leq 5$, et, pour déterminer la température en ces 25 nœuds, 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, elle était déjà connue des mathématiciens chinois de l'antiquité. Au reste, la technique d'élimination successive des inconnues est très naturelle et aisément compréhensible par des mathématiciens débutants.

1.2 Le formalisme

La résolution des systèmes minéaires forment un chapitre fondamental de la partie des mathématiques connue sous le nom d'algèbre linéaire. 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 [*]. Nous considérons le système de n équations à n inconnues suivant

$$(1.1)\qquad \begin{array}{c} \mathbf{L_1} \\ \mathbf{L_2} \\ \vdots \\ \mathbf{L_i} \\ \vdots \\ \mathbf{L_n} \end{array} \left\{ \begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1n}x_n & = & c_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \dots & + & a_{2n}x_n & = & c_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{i1}x_1 & + & a_{i2}x_2 & + & \dots & + & a_{in}x_n & = & c_i \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \dots & + & a_{nn}x_n & = & c_n \end{array} \right. .$$

*. * Elles seraient d'ailleurs également valables pour des systèmes dont les coefficients a_{ij} seraient des éléments d'un corps commutatif quelconque.

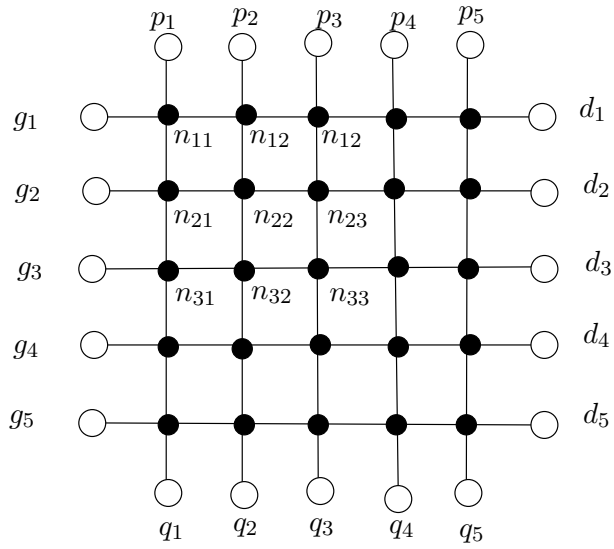


FIGURE 1 – Le problème de la grille

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

(1.2) $\mathbf{L_i}$: $a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = c_i$

s'appelle la i-ième **ligne** du système. Le système (1.1) se représente aussi sous la forme compacte

(1.3) $\displaystyle\sum_{j=1}^{n} a_{ij}x_j = c_i, \qquad i = 1,2,\ldots,n.$

Au système linéaire (1.1) est associée l'**équation matricielle**

(1.4) $AX = C,$

où la matrice A et les vecteurs X et C sont définis par

(1.5) $$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \ldots & a_{in} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}, \quad X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix}, \quad C = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_i \\ \vdots \\ c_n \end{pmatrix}.$$

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 $x = (x_1, x_2, \ldots, x_n)$ 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

$$(1.6)\quad \mathscr{A}(x) = \big(\sum_{j=1}^{n} a_{1,j}x_j, \ldots, \sum_{j=1}^{n} a_{i,j}x_j, \ldots, \sum_{j=1}^{n} a_{n,j}x_j\big), \quad x = (x_1, \ldots, x_n).$$

E 88 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 89 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 90 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

$$(1.7)\quad \det A \neq 0 \iff \mathscr{A}\text{bijective} \iff \mathscr{A}\text{surjective} \iff \mathscr{A}\text{injective} \iff \ker \mathscr{A} = \{0\}.$$

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 91 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. *Lorsque* $\det A \neq 0$ *la coordonnée* x_j *de la solution* x *du système (1.1) est donnée par la formule*

$$x_j = \frac{1}{\det A}\begin{vmatrix} a_{11} & \ldots & a_{1j-1} & c_1 & a_{1j+1} & \ldots & a_{1n} \\ a_{21} & \ldots & a_{2j-1} & c_2 & a_{2j+1} & \ldots & a_{2n} \\ \vdots & & & & & & \vdots \\ a_{n1} & \ldots & a_{nj-1} & c_n & a_{nj+1} & \ldots & a_{nn} \end{vmatrix}, \quad j = 1, \ldots, n.$$

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 95).

§ 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 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 de trois équations à trois inconnues.

$$(S_1) \qquad \begin{cases} l_{11}x_1 & & & & & = & c_1 \\ l_{21}x_1 & + & l_{22}x_2 & & & = & c_2 \\ l_{31}x_1 & + & l_{32}x_2 & + & l_{33}x_3 & = & c_3 \end{cases},$$

(Système triangulaire inférieur)

$$(S_2) \qquad \begin{cases} u_{11}x_1 & + & u_{12}x_2 & + & u_{13}x_3 & = & c_1 \\ & & u_{22}x_2 & + & u_{23}x_3 & = & c_2 \\ & & & & u_{33}x_3 & = & c_3 \end{cases}.$$

(Système triangulaire supérieur)

Chacun des deux systèmes se résout facilement par **substitutions successives** (à condition que les éléments diagonaux soient non nuls),

$$(S_1) \qquad \begin{cases} x_1 = \frac{c_1}{l_{11}} \\ x_2 = (c_2 - l_{21}x_1)/l_{22} \\ x_3 = (c_3 - l_{31}x_1 - l_{32}x_2)/l_{33} \end{cases},$$

$$(S_2) \qquad \begin{cases} x_3 = \frac{c_3}{u_{33}} \\ x_2 = (c_2 - u_{23}x_3)/u_{22} \\ x_1 = (c_1 - u_{12}x_2 - u_{13}x_3)/u_{33} \end{cases}.$$

La technique de substitutions successives s'applique de la même manière aux systèmes triangulaires de n équations à n inconnues. Nous présentons maintenant ces algorithmes.

2.2 Les algorithmes de substitution successives

Algorithme 4. *Les solutions du système triangulaire* inférieur

$$LX = C \qquad \textit{avec} \quad \begin{cases} l_{ij} = 0 & \textit{si} \quad i < j, \\ l_{ii} \neq 0, \end{cases}$$

sont données par les relations

$$\begin{cases} x_1 = \frac{c_1}{l_{11}}, \\ x_i = \frac{1}{l_{ii}}\left(c_i - \sum_{j=1}^{i-1} l_{ij}x_j\right), \qquad i = 2,3,\dots,n. \end{cases}$$

Algorithme 5. *Les solutions du système triangulaire* supérieur

$$UX = C \qquad \text{avec} \quad \begin{cases} u_{ij} = 0 & \text{si} \quad i > j, \\ u_{ii} \neq 0, \end{cases}$$

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 équivalentes 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 ($\times$ ou $\div$) de sorte que le nombre total d'opérations N_n est donné par

$$N_n = 1 + \sum_{i=2}^{n} (2i-1) = 1 + 2\sum_{i=2}^{n} i - (n-1) = 1 + 2\frac{n(n+1)}{2} - 2 - (n-1) = n^2. \qquad \blacksquare$$

§ 3. L'ALGORITHME DE GAUSS

Nous indiquons maintenant comme réduire n'importe quel système régulière à un système triangulaire supérieur. Nous décrivons le procédé dans le cas de dimension 3 pour lequel nous pouvons visualiser toutes les étapes.

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.

$$S^{(0)} \quad \begin{matrix} \mathbf{L}_1^{(0)} \\ \mathbf{L}_2^{(0)} \\ \mathbf{L}_3^{(0)} \end{matrix} \quad \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = c_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = c_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = c_3 \end{cases}.$$

Étape 1. Élimination de x_1 dans $\mathbf{L}_2^{(0)}$ et $\mathbf{L}_3^{(0)}$.

$$\left\| \begin{matrix} \mathbf{L}_2^{(0)} \longleftarrow \mathbf{L}_2^{(0)} - \frac{a_{21}}{a_{11}} \mathbf{L}_1^{(0)} \\ \mathbf{L}_3^{(0)} \longleftarrow \mathbf{L}_3^{(0)} - \frac{a_{31}}{a_{11}} \mathbf{L}_1^{(0)} \end{matrix} \right. .$$

Attention, nous divisons par a_{11}. Cela n'est possible que si a_{11} est non nul. Nous arrivons à

$$S^{(1)} \quad \begin{matrix} \mathbf{L}_1^{(0)} \\ \mathbf{L}_2^{(1)} \\ \mathbf{L}_3^{(1)} \end{matrix} \quad \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = c_1 \\ 0 + (a_{22} - \frac{a_{21}}{a_{11}}a_{12})x_2 + (a_{23} - \frac{a_{21}}{a_{11}}a_{13})x_3 = c_2 - \frac{a_{21}}{a_{11}}c_1 \\ 0 + (a_{32} - \frac{a_{31}}{a_{11}}a_{12})x_2 + (a_{33} - \frac{a_{31}}{a_{11}}a_{13})x_3 = c_3 - \frac{a_{31}}{a_{11}}c_1 \end{cases}.$$

que nous écrivons encore sous la forme

$$S^{(1)} \quad \begin{matrix}\mathbf{L}_1^{(0)}\\ \mathbf{L}_2^{(1)}\\ \mathbf{L}_3^{(1)}\end{matrix} \quad \begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & c_1 \\ 0 & + & a_{22}^{(1)}x_2 & + & a_{23}^{(1)}x_3 & = & c_2^{(1)} \\ 0 & + & a_{32}^{(1)}x_2 & + & a_{33}^{(1)}x_3 & = & c_3^{(1)} \end{cases}.$$

Étape 2. Élimination 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

$$S^{(2)} \quad \begin{matrix}\mathbf{L}_1^{(0)}\\ \mathbf{L}_2^{(1)}\\ \mathbf{L}_3^{(2)}\end{matrix} \quad \begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & c_1 \\ 0 & + & a_{22}^{(1)}x_2 & + & a_{23}^{(1)}x_3 & = & c_2^{(1)} \\ 0 & + & 0 & + & (a_{33}^{(1)} - \frac{a_{32}^{(1)}}{a_{22}^{(1)}}a_{23}^{(1)})x_3 & = & c_3^{(1)} - \frac{a_{32}^{(1)}}{a_{22}^{(1)}}c_2^{(1)} \end{cases},$$

que nous écrivons

$$S^{(2)} \quad \begin{matrix}\mathbf{L}_1^{(0)}\\ \mathbf{L}_2^{(1)}\\ \mathbf{L}_3^{(1)}\end{matrix} \quad \begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & c_1 \\ 0 & + & a_{22}^{(1)}x_2 & + & a_{23}^{(1)}x_3 & = & c_2^{(1)} \\ 0 & + & 0 & + & a_{33}^{(2)}x_3 & = & c_3^{(2)} \end{cases}.$$

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 par exemple 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

$$\begin{cases} a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & = & c_2 \\ 0 & + & a_{12}x_2 & + & a_{13}x_3 & = & c_1 \\ a_{31}x_1 & + & a_{32}x_2 & + & a_{33}x_3 & = & c_3 \end{cases}.$$

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'ils ne soient pas proches de 0, voir 3.4.

E 92 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 stratégie de pivot)

Nous décrivons maintenant l'algorithme de Gauss ci-dessus dans le cas d'un système à n équations et n inconnues. L'énoncé ne suppose pas que le système soit régulier. L'algorithme est muni d'une instruction d'arrêt pour le cas où $\det A = 0$.

Algorithme 7 (Notation Ligne). *Nous considérons 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). *Nous considérons 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*

$$a_{jl} \longleftarrow a_{i_0 l} \quad \text{pour} \quad l \geq j$$
$$a_{i_0 l} \longleftarrow a_{jl} \quad \text{pour} \quad l \geq j$$
$$c_j \longleftarrow c_{i_0}$$
$$c_{i_0} \longleftarrow c_j$$

1.2 *Pour $i > j$ faire*

1.2.1

$$m_i = \frac{a_{ij}}{a_{jj}}$$
$$c_i \longleftarrow c_i - m_i c_j.$$

1.2.2 *pour $k > j$ faire*

$$a_{ik} \longleftarrow a_{ik} - m_i a_{jk}.$$

2 *Résoudre le système (triangulaire)*

$$\sum_{i=k}^{n} a_{ki}x_i = b_k, \qquad k = 1,2,\ldots,n$$

E 93 **Méthode de Gauss avec stratégie de pivot.** Pour que l'algorithme de Gauss fonctionne, nous avons vu que les pivots (les nombres par lesquels nous divisions 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). Pour des raisons qui seront expliquées au 3.4, c'est généralement une bonne idée de choisir comme pivot non le premier nombre non nuls mais celui dont la valeur absolue est la plus grande.

Modifier l'algorithme de Gauss de telle sorte que le j-ème pivot soit choisi comme ci-dessus.

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

$$N_n = \underbrace{\sum_{j=1}^{n-1}\left(\sum_{i=j+1}^{n}\big(3+\sum_{k=j+1}^{n}2\big)\right)}_{\text{coût de } \boxed{1}} + \underbrace{n^2}_{\text{coût de } \boxed{2}},$$

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 $\boxed{?}$ 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

$$\lim_{n\to\infty}\frac{N_n}{\frac{2}{3}n^3} = 1.$$

■

3.4 Les sources d'erreurs

La méthode de Gauss en principe devrait fournir la solution exacte du système. Dans la pratique, il n'en est pas ainsi dès lors que l'ordre n du système n'est plus très petit. Comme nous l'avons expliqué au I.2.4, les calculateurs travaillent (généralement) avec un ensemble fini de nombres F et la plupart des opérations sont effectuées avec une erreur très petite mais non nulle et lorsque le nombre d'opérations est grand, il peut arriver que ces petites erreurs cessent de se cumuler raisonnablement (disons, comme leur somme) pour conduire à des résultats inexploitables. Dans l'algorithme de Gauss 8, les erreurs peuvent se produire dans les calculs

$$m_i = \frac{a_{ij}}{a_{jj}} \quad \text{et} \quad a_{ik} - m_i a_{jk}.$$

Les autres erreurs apparaissent au niveau de la résolution du système triangulaire final. Il y a deux raisons au moins qui font désirer un coefficient m_i assez petit. Comme l'ensemble F utilisé par les calculateurs est formé de nombres qui ne sont pas équidistribués mais sont moins dense pour les très grands nombres, si m_i est très grand il sera représenté par un nombre $\boxed{m_i}$ qui pourra être éloigné de m_i. La seconde raison c'est si m_i est très grand tandis que, ce qui est raisonnable au moins dans les première étapes, les coefficients a_{ik} et a_{jk} sont comparable dans le calcul de $a_{ik}-m_ia_{jk}$, l'apport de a_{ik} tendra à être négligé et, dans le pire des cas, $a_{ik}-m_ia_{jk}$ donnera le même résultat que $-m_ia_{jk}$ ce qui revient à supposer que le coefficient a_{ik} est nul. Par exemple si $a_{ik}=1$, $a_{jk}=2$ et $m_i=10^{10}$ alors le calculateur (ici Scilab) effectuant $1-10^{10}*2$ retournera $-2.000\,D+10$. En choisissant comme pivot le coefficient de plus grande valeur absolue sur la ligne, comme dans le code 11 ci-après et comme suggéré à l'exercice 93 nous nous assurons que les multiplicateurs m_i sont dans $[-1,1]$ ce qui permet éviter le phénomène ci-dessus. Cette stratégie ne constitue certainement pas une panacée et n'empêche pas l'algorithme de Gauss de faillir dans certains cas, ne serait-ce que parce que la supposition que les coefficients sont de grandeurs comparables n'est pas généralement vérifiée. Il existe d'autres stratégies de choix. Nous pouvons par exemple chercher le pivot parmi tous les éléments de la matrice carrée restante (ce qui conduira à une permutation de ligne et une permutation de colonne) mais cette technique augmente considérablement le nombre de tests à effectuer. Certains ont proposé de comme le plus grand des nombres disponibles sur la ligne et sur la colonne.

3.5 Code et commentaires

Dans le code Scilab ci-dessous, l'algorithme est traduit en choisissant comme pivot, le plus grand, en valeur absolue, des nombres disponibles (sur la ligne)

Code SCILAB 11 (Algorithme dee Gauss pour la résolution des systèmes linéaires). Dans le code suivant, nous définissons la fonction *gauss* qui possède deux argument, A et b de la forme indiquée ci-dessous.

(a) A est une matrice carrée de dimension n et b est un vecteur colonne de dimension n. L'ordre n est recherché par l'algorithme.

(b) L'algorithme retourne, lorsque c'est possible, une vecteur colonne $X=gauss(A,B)$ de dimension n vérifiant $AX=b$ (plus exactement une valeur approchée d'un tel X).

(c) L'algorithme prévoit deux sorties d'échec. La première dans où les données fournies sont incompatibles, la seconde lorsque la matrice A est non inversible ou lorsque les pivots, même s'ils sont encore non nuls, deviennent si petits qu'il est possible de suspecter que la matrice soit non inversible et qui, qu'elle le soit ou non, risqueraient de conduire à des résultats imprécis. La barrière, haute est ici fixée à 10^{-6}.

La construction est commentée plus bas.

```
1  function X=gauss(A,b)
2  [n,m]=size(A);
3  [nb,mb]=size(b);
4  TS=0;
5  if (or([~(n==m), ~(mb==1), ~(n==nb)])==%T) then
6  TS=1;
7  else
8     L=[A,b];
9     for j=[1:n-1];
10      [M,k]=max(abs(L(j:n,j)));
11        if (M < 10^(-6)) then  TS=2;
12        else
13           S=L(k+j-1,:);
14           SS=L(j,:);
15           L(j,:)=S;
16           L(k+j-1,:)=SS;
17           for i=j+1:n;
18              L(i,:)=clean(L(i,:)-(L(i,j)/L(j,j)).*L(j,:));
19           end;
20        end;
21    end;
22 end
23 if TS==1 then X='EXIT 1 (dimensions !)';
24 elseif TS==2 then X='EXIT 2 (matrice non inversible ou instable)'
25 else
26    X=zeros(n,1);
27    X(n)=L(n,n+1)/L(n,n);
28    for s=n-1:-1:1;
29    X(s)=(1/L(s,s))*(L(s,n+1)-sum(L(s,s:n)*X(s:n)));
30    end
31 end
32 endfunction
```

Commentaire.

(a) La variable TS est utilisée pour gérer les sorties du calcul. Lorsque TS prend la valeur 1, les données sont incompatibles, c'est l'EXIT 1, tandis-que lors TS prend la valeur 2, c'est que tous les coefficients de la ligne sont trop petits.

(b) Les variables S et SS sont utilisées pour gérer la permutation des lignes.

(c) A la ligne 10, M est le maximum et k est l'indice de l'élément maximal. Attention, comme nous cherchons dans une matrice de longueur $n-j+1$, il faut prendre garde de récupérer l'élément correct dans les lignes suivantes, c'est ce qui explique l'emploi de l'indice $k+j-1$.

(d) L'instruction *clean* égale à 0 des nombres très petits.

Avec une matrice A (et un vecteur b correspondant) d'ordre 200 choisis aléatoirement (avec des éléments dans $[-1,1]$) l'algorithme a retourné en une seconde un vecteur X tel que $AX = b+e$ avec tous les coefficients de e plus petits que 10^{-13} en valeurs absolue.

L'EXIT 2 se produit par exemple lorsque A est choisie comme la **matrice de Hilbert** d'ordre 20. Nous sommes ici dans le cas où les coefficients des matrices intermédiaires deviennent de grandeurs très divergentes et la condition $|m_i| \leq 1$ n'interdit pas l'apparition du phénomène de fausse annulation décrit au 3.4. Les matrices de Hilbert ont pour coefficient $h_{ij} = 1/(-i+j-1)$. Ce sont des matrices (inversibles) qui sont connues pour être numériquement difficiles à traiter. Elles sont souvent utilisées comme matrices de test pour mettre à l'épreuve les algorithmes d'algèbre linéaire.

§ 4. EXERCICES ET PROBLÈMES

94 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 ?

95 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 ?

96 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 formée 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 $n \to \infty$.

97 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 ?

98 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.

99 Matrices triangulaires creuses. On considère un système linéaire à n équations et n inconnues, $n \geq 4$, triangulaire supérieur, de la forme

$$(4.1) \quad \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = c_1 \\ a_{22}x_2 + a_{23}x_3 + a_{24}x_4 = c_2 \\ \cdots \\ a_{ii}x_i + a_{i,i+1}x_{i+1} + a_{i,i+2}x_{i+2} = c_i \\ \cdots \\ a_{n-2,n-2}x_{n-2} + a_{n-2,n-1}x_{n-1} + a_{n-2,n}x_n = c_{n-2} \\ a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n = c_{n-1} \\ a_{nn}x_n = c_n \end{cases},$$

où les coefficients a_{ii}, $i = 1, \ldots, n$ sont supposés non nuls.

(a) Écrire 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. 130.)

100 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

$$S^{(0)} \quad \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = c_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = c_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = c_3 \end{cases}.$$

On rappelle que les systèmes $S^{(1)}$ (resp. $S^{(2)}$) obtenus après la première (resp. la seconde) étape de l'algorithme sont donnés par

$$S^{(1)} \quad \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = c_1 \\ 0 + (a_{22}-\frac{a_{21}}{a_{11}}a_{12})x_2 + (a_{23}-\frac{a_{21}}{a_{11}}a_{13})x_3 = c_2-\frac{a_{21}}{a_{11}}c_1 \\ 0 + (a_{32}-\frac{a_{31}}{a_{11}}a_{12})x_2 + (a_{33}-\frac{a_{31}}{a_{11}}a_{13})x_3 = c_3-\frac{a_{31}}{a_{11}}c_1 \end{cases},$$

abrégé en

$$S^{(1)} \quad \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = c_1 \\ 0 + a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 = c_2^{(1)} \\ 0 + a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 = c_3^{(1)} \end{cases},$$

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 $S^{(2)}$. Écrire les trois matrices $A^{(0)}, A^{(1)}, A^{(2)}$.

(b) Vérifier que

$$A^{(1)}=L_1A^{(0)} \quad \text{et} \quad A^{(2)}=L_2A^{(1)}$$

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. 130.)

101 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.

(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 :

$$(4.2)\quad L = \begin{pmatrix} 1 & & & 0 \\ \beta_2 & 1 & & \\ & \ddots & \ddots & \\ 0 & & \beta_n & 1 \end{pmatrix} \quad \text{et} \quad U = \begin{pmatrix} \alpha_1 & c_1 & & 0 \\ & \alpha_2 & \ddots & \\ & & \ddots & c_{n-1} \\ 0 & & & \alpha_n \end{pmatrix}.$$

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 $A = L \cdot U$. 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 dite **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. 131.)

102 Méthode de Cholesky pour les matrices tridiagonales symétriques. On considère une matrice carrée A d'ordre $p \geq 3$ à coefficients réels de la forme

$$(4.3)\quad A = \begin{pmatrix} b_1 & c_1 & & & & 0 \\ c_1 & b_2 & c_2 & & & \\ & c_2 & b_3 & c_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & c_{p-2} & b_{p-1} & c_{p-1} \\ 0 & & & & c_{p-1} & b_p \end{pmatrix}.$$

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

$$(4.4)\qquad d_1 = \sqrt{b_1}, \quad f_1 = c_1/d_1;$$

$$(4.5)\quad d_j = \sqrt{b_j - f_{j-1}^2}, \quad j = 2, \ldots, p; \quad f_j = c_j/d_j, \quad j = 2, \ldots, p-1.$$

Ici, on suppose que les nombres dont on prend les racines carrées sont positifs.

(a) Étude 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.

103 Méthode de Jordan. On considère le système

$$S^{(0)} \qquad \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = c_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = c_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = c_3 \end{cases}.$$

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) Écrire un algorithme (en notation ligne) effectuant le travail ci-dessus dans le cas d'un système linéaire de n équations et n inconnues.

104 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

$$(4.6) \quad A = L \cdot T(L) \quad \text{et} \quad l_{kk} > 0, \quad k = 1, \dots, n.$$

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 détermine le nombre d'opérations élémentaires $+, -, \times, \div$ et aussi la racine carrée $\sqrt{}$ employées 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. 132.)

§ 5. NOTES ET COMMENTAIRES

Sur le contenu

Ici, je devais m'adresser à un public qui pouvait n'avoir qu'une connaissance minimale d'algèbre linéaire (opérations sur les matrices, calculs de déterminant) sans pouvoir m'appuyer sur aucun formalisme vectoriel abstrait. Il était hors de question de parler explicitement de factorisation de matrices. Il en résulte un traitement très élémentaire de l'algorithme de Gauss — mais, selon moi, qui le connaît peut bien se dire en possession de la moitié de l'essentiel de la résolution numérique des systèmes linéaires. J'ai toujours été agacé par une présentation générale de l'algorithme de Gauss dans lequel le lecteur ne peut que se perdre, ou tout au moins s'assommer, dans les jeux d'indices. C'est la raison pour laquelle j'ai décidé d'expliciter l'algorithme dans le cas d'une matrice d'ordre 3, qui permet visualiser toutes les étapes. L'écriture de l'algorithme général va ensuite de soi, comme j'ai pu le vérifier avec mes étudiants.

Sur les exercices

J'ai proposé quelques développements sous forme d'exercices, le plus important du point de vue des applications concerne celui des matrices tridiagonales.

V

Analyse matricielle

§ 1. Introduction et avertissement

Pour quantifier les erreurs et analyser la convergence des méthodes d'approximation utilisées dans la résolution des systèmes linéaires, il est nécessaire de se munir des outils adéquates pour mesurer la grandeur des matrices et des vecteurs. Il s'agit, en somme, de choisir l'outil qui prendra le rôle que tient la valeur absolue dans l'analyse des fonctions d'une variable. Cet outil prend le nom de **norme**. Nous l'appliquerons en particulier à la construction de ce que l'on appelle des **méthodes itératives** de résolution d'un système linéaire. Elles consistent à construire des suites de vecteurs $x^{(k)}$, définies par des relations de récurrences simples, ne faisant en général intervenir que des produits de matrices, qui convergent en un sens que le notion de norme permettra justement de préciser vers l'unique solution du système. Ces méthodes sont employées lorsque les méthodes directes ne peuvent l'être, soit qu'elles nécessitent trop d'opérations, soit qu'elles sont trop instables, c'est-à-dire faussées par les erreurs d'arrondis – comme nous avons vu qu'il peut arriver, par exemple, dans l'algorithme de Gauss – ou bien en complément des méthodes directes pour affiner le résultat obtenu. Elles ne sont en tous les cas utiles que pour le traitement des grands systèmes. Le lecteur trouvera dans ce chapitre des illustrations et des exercices qui sont, pour cette raison, tout à fait irréalistes, avec des systèmes à 2 ou 3 inconnues. Ils servent uniquement à rendre plus concret le fonctionnement des algorithmes présentés.

Nous chercherons à traiter les notions nouvelles de la manière la plus élémentaire possible, manière qui est malheureusement rarement la plus courte. Il est cependant nécessaire que le lecteur soit familier après le produit matriciel pour lire valablement le contenu de ce chapitre.

§ 2. Normes vectorielles

2.1 Définitions

Nous travaillons avec l'espace vectoriel $\mathbb{R}^n$. Nous disons qu'une application $N : \mathbb{R}^n \to \mathbb{R}^+$ est une **norme** si elle vérifie les trois conditions suivantes.

(a) $N(x) = 0 \iff x = 0, x \in \mathbb{R}$.

(b) $N(\lambda x) = |\lambda| N(x), \lambda \in \mathbb{R}, x \in \mathbb{R}^n$. En particulier, $N(-x) = N(x), x \in \mathbb{R}^n$.

(c) $N(x+y) \leq N(x) + N(y), x, y \in \mathbb{R}^n$. Cette inégalité s'appelle l'**inégalité triangulaire**.

Observons que des applications répétées de l'inégalité triangulaire donnent

$$N\left(\sum_{i=1}^{j} v_i\right) \leq \sum_{i=1}^{j} N(v_i), \quad v_i \in \mathbb{R}^n, \quad i = 1, \ldots, j. \tag{2.1}$$

Plus généralement, en combinant l'inégalité triangulaire avec (b), nous avons

$$N\left(\sum_{i=1}^{j} \lambda_i v_i\right) \leq \sum_{i=1}^{j} |\lambda_i| N(v_i), \quad v_i \in \mathbb{R}^n, \ \lambda_i \in \mathbb{R}, \quad i = 1, \ldots, j. \tag{2.2}$$

Nous déduisons facilement de l'inégalité triangulaire une autre inégalité très utile,

$$|N(x) - N(y)| \leq N(x-y), \quad x, y \in \mathbb{R}^n. \tag{2.3}$$

Elle s'établit en remarquant que $x=(x-y)+y$ d'où $N(x)=N((x-y)+y)\leq N(x-y)+N(y)$ et $N(x)-N(y)\leq N(x-y)$. En permutant les rôles de x et de y, nous obtenons $N(y)-N(x)\leq N(y-x)=N(x-y)$. La réunion des deux inégalité donnent (2.3).

Nous emploierons souvent une notation de la forme $N(x)=\|x\|$.

2.2 Exemples fondamentaux

Pour l'utilisation simple que nous en ferons, les trois exemples de normes ci-dessus suffiront.

Théorème 1. *Les applications $\|.\|_\infty$, $\|.\|_1$ et $\|.\|_2$ définies par les relations ci-dessous sont des normes*[a] *sur $\mathbb{R}^n$. (a) $\|x\|_\infty=\max_{1\leq i\leq n}|x_i|$, (b) $\|x\|_1=\sum_{i=1}^n|x_i|$, (c) $\|x\|_2=\sqrt{\sum_{i=1}^n x_i^2}$ où nous avons posé $x=(x_1,x_2,\ldots,x_n)$,*

a. L'utilisation de l'indice ∞ pour $\|x\|_\infty$ est certainement malheureux puisque ∞ ne joue aucun rôle ici mais elle est assez fermement encrée dans la littérature et comme toujours dans ce cas, il est inutile d'essayer de s'éloigner de l'usage.

La norme $\|\cdot\|_\infty$ est souvent appelée la **norme sup** tandis que la norme $\|\cdot\|_2$ est reconnue sous le terme de **norme euclidienne**.

Démonstration. Nous démontrons seulement que $\|\cdot\|_2$ est une norme. Nous devons d'abord établir que $\|x\|_2=0\iff x=0$. Mais $\|x\|_2=0\iff\|x\|_2^2=0$. Or $\|x\|_2^2$ est une somme de nombres positifs donc il est nul si et seulement si chacun des termes de la somme est nul, $\|x\|_2=0$ si et seulement si $x_i^2=0$ ou encore $x_i=0$ pour $i=1,\ldots,n$. Montrons le second point de la définition. Soient $\lambda\in\mathbb{R}$ et $x\in\mathbb{R}^n$, nous avons

$$\|\lambda x\|_2=\sqrt{\sum_{i=1}^n(\lambda x_i)^2}=\sqrt{\sum_{i=1}^n\lambda^2x_i^2}=\sqrt{\lambda^2\sum_{i=1}^n x_i^2}=|\lambda|\sqrt{\sum_{i=1}^n x_i^2}=|\lambda|.\|x\|_2.$$

Le troisième point, l'inégalité triangulaire, se traduit en passant au carré par $\|x+y\|_2^2\leq\|x\|_2^2+\|y\|_2^2+2\|x\|_2\|y\|_2$, ou encore,

$$\sum_{i=1}^n(x_i+y_i)^2\leq\sum_{i=1}^n x_i^2+\sum_{i=1}^n y_i^2+2\sqrt{\sum_{i=1}^n x_i^2\sum_{i=1}^n y_i^2}$$

Et, puisque

$$\sum_{i=1}^n(x_i+y_i)^2=\sum_{i=1}^n x_i^2+\sum_{i=1}^n y_i^2+2\sum_{i=1}^n x_iy_i,$$

la dernière relation est encore équivalente à

$$\sum_{i=1}^n x_iy_i\leq\sqrt{\sum_{i=1}^n x_i^2\sum_{i=1}^n y_i^2}.$$

Or cette dernière inégalité est vraie, c'est la fameuse **inégalité de Cauchy-Schwarz** que nous établissons ci-après. ■

Lemme 2 (Inégalité de Cauchy-Schwarz). *Pour tout couple de vecteurs (x,y) dans $\mathbb{R}^n$, nous avons*

$$\left|\sum_{i=1}^n x_iy_i\right|\leq\sqrt{\sum_{i=1}^n x_i^2}\cdot\sqrt{\sum_{i=1}^n y_i^2}.$$

Démonstration. Considérons l'application f définie pour $t \in \mathbb{R}$ par $f(t) = \sum_{i=1}^{n}(x_i + ty_i)^2$. Puisque $f(t)$ est une somme de carrés nous avons $f(t) \geq 0$, $t \in \mathbb{R}$. Mais f est un trinôme du second degré. En effet,

$$f(t) = t^2 \sum_{i=1}^{n} y_i^2 + t \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} x_i^2.$$

Or nous savons que les trinômes du second degré qui gardent un signe constant ont un discriminant (Δ) négatif ou nul. Ici

$$\Delta = 4\left(\sum_{i=1}^{n} x_i y_i\right)^2 - 4\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right).$$

La condition $\Delta \leq 0$ conduit à la relation

$$\left(\sum_{i=1}^{n} x_i y_i\right)^2 \leq \left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right),$$

dont nous tirons l'inégalité de Cauchy-Schwarz. ∎

E 105 Montrer que les applications $\|.\|_\infty$ et $\|.\|_1$ définies pour $x = (x_1, x_2, \ldots, x_n)$ par les relations $\|x\|_\infty = \max_{1\leq i\leq n} |x_i|$ et $\|x\|_1 = \sum_{i=1}^{n} |x_i|$ sont des normes sur $\mathbb{R}^n$.

E 106 Lorsque $n = 2$, déterminer et représenter pour chacune des deux normes l'ensemble des $x = (x_1, x_2)$ dont la norme est ≤ 1 (respectivement < 1, $= 1$).

2.3 Équivalence

Il est facile de comparer les trois normes définies dans le théorème 1. Nous avons, pour tout $x \in \mathbb{R}^n$,

$$\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty, \tag{2.4}$$

mais aussi

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty, \quad \text{et} \quad \|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2. \tag{2.5}$$

E 107 Établir les inégalités ci-dessus.

En réalité, des inégalités comme (2.5) et (2.4) existent pour chaque couple de normes sur $\mathbb{R}^n$. C'est ce qu'indique le théorème suivant. Ce résultat est utile pour la suite mais sa démonstration fait appel à quelques connaissances de topologie et sa lecture peut être omise par le lecteur non mathématicien.

Théorème 3. *Si N_1 et N_2 sont deux normes quelconques de $\mathbb{R}^n$ alors il existent des constantes réelles (positives) c_1 et c_2 telles que*

$$c_1 N_1(x) \leq N_2(x) \leq c_2 N_1(x), \quad x \in \mathbb{R}^n.$$

On dit que toutes les normes de $\mathbb{R}^n$ sont équivalentes.

∗ *Démonstration.* Il nous suffit d'établir le résultat lorsque $N_2 = \|\cdot\|_\infty$. En effet, si N_1 et N_2 sont deux normes quelconques et si nous savons que, pour certaines constantes c_1, c_2, γ_1 et γ_2, nous avons

$$c_1 N_1(x) \leq \|\cdot\|_\infty \leq c_2 N_1(x), \quad \text{et} \quad \gamma_1 N_2(x) \leq \|\cdot\|_\infty \leq \gamma_2 N_2(x),$$

alors nous tirerons immédiatement

$$\frac{c_1}{\gamma_2} N_1(x) \leq N_2(x) \leq \frac{c_2}{\gamma_1} N_1(x).$$

Soit donc N une norme quelconque. Nous cherchons à établir l'existence de deux constantes c et C telles que

$$c\|x\|_\infty \leq N(x) \leq C\|x\|_\infty, \quad x \in \mathbb{R}^n. \tag{2.6}$$

L'existence de C s'établit facilement. Pour $x = (x_1, \ldots, x_n) = \sum_{i=1}^n x_i e_i$ avec $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, la relation (2.2) nous donne

$$N(x) = N\left(\sum_{i=1}^n x_i e_i\right) \leq \sum_{i=1}^n N(x_i e_i) = \sum_{i=1}^n |x_i| N(e_i) \leq \left(\sum_{i=1}^n N(e_i)\right) \cdot \|x\|_\infty, \tag{2.7}$$

de sorte qu'il suffit de choisir $C = \sum_{i=1}^n N(e_i)$ dans (2.6).

L'existence de la seconde constante est plus délicate à établir. Notons B l'ensemble des vecteurs x de $\mathbb{R}^n$ pour lesquels $N(x) = 1$,

$$B = \{x \in \mathbb{R}^n \ : \ N(x) = 1\}.$$

Nous affirmons qu'il existe $K > 0$ tel que pour tout $x \in B$, $\|x\|_\infty \leq K$. Nous établissons l'existence de cette constante K en supposant le contraire et en recherchant une contradiction. Supposons donc que ce nombre K n'existe pas. Dans ce cas, pour tout $k \geq 2$, nous pouvons trouver $x^{(k)} \in B$ tel que $\|x^{(k)}\|_\infty \geq k$ – sans quoi, il serait permis de choisir $K = k$. Posons alors $v^{(k)} = \|x^{(k)}\|_\infty^{-1} \cdot x^{(k)}$. Nous avons d'une part

(a) $\|v^{(k)}\|_\infty = 1$, $k \in \mathbb{N}^\star$ et, d'autre part,

(b) $N(v^{(k)}) = 1/\|x^{(k)}\|_\infty$, $k \in \mathbb{N}^\star$, de sorte que $N(v^{(k)}) \leq 1/k$ et $\lim_{k\to\infty} N(v^{(k)}) = 0$.

La première information entraîne que chacune des coordonnées du vecteur $v^{(k)}$ se trouve dans l'intervalle $[-1, 1]$,

$$|v_i^{(k)}| \leq 1, \quad i = 1, \ldots, n, \quad k \in \mathbb{N}^\star.$$

Puisque la suite des premières coordonnées, $v_1^{(k)}$, est une suite bornée, elle admet * une sous-suite convergente, disons $v_1^{(\phi(k))}$. La suite des secondes coordonnées de $v^{(\phi(k))}$, $v_2^{(\phi(k))}$, étant également bornée, possède à son tour une sous-suite convergente et, continuant ainsi, nous obtiendrons finalement une sous-suite de $v^{(k)}$ dont *toutes* les coordonnées seront des suites convergentes. Pour simplifier les écritures, nous noterons cette suite $w^{(k)}$. Puisque $w^{(k)}$ et une sous-suite de $v^{(k)}$, les propriétés (a) et (b) demeurent de sorte que les deux premières propriétés ci-dessous sont héritées de $v^{(k)}$ tandis que que la troisième résulte du choix de $w^{(k)}$.

(c) $\|w^{(k)}\|_\infty = 1$,

(d) $\lim_{k\to\infty} N(w^{(k)}) = 0$

(e) $\lim_{k\to\infty} w_i^{(k)} = l_i$, $i = 1, \ldots, n$.

De la troisième relation, nous tirons

$$\lim_{k\to\infty} |w_i^{(k)} - l_i| = 0, \quad i = 1, \ldots, n,$$

puis $\lim_{k\to\infty} \max_{i=1,\ldots,d} |w_i^{(k)} - l_i| = 0$ ou encore

$$\lim_{k\to\infty} \|w^{(k)} - l\|_\infty = 0, \quad l = (l_1, \ldots, l_n).$$

L'inégalité (2.3) appliquée avec la norme $\|\cdot\|_\infty$ nous donne, compte tenu de (2.7),

$$|N(W^{(k)}) - N(l)| \leq N(W^{(k)} - l) \leq C\|W^{(k)} - l\|_\infty.$$

*. C'est ici que nous utilisons un argument topologique : toute suite bornée de $\mathbb{R}$ admet une sous-suite convergente. Ce résultat se trouve dans toutes les introduction à l'analyse des fonctions d'une variables réelle.

En faisant $k \to \infty$, nous obtenons $N(l) = 0$ et, puisque N est une norme, $l = 0$ ce qui contredit $\|l\|_\infty = 1$. Cela signifie que le rejet de l'existence de K a conduit à une contradiction. Il existe donc un nombre K tel que

$$N(x) = 1 \implies \|x\| \leq K.$$

Si x est un vecteur non nul quelconque, nous avons alors

$$N\left(\frac{1}{N(x)}x\right) = 1 \implies \left\|\frac{1}{N(x)}x\right\|_\infty \leq K \implies \|x\|_\infty \leq KN(x),$$

qui conduit directement à l'inégalité qui nous manquait en prenant $c = 1/K$. ■

2.4 Convergence d'une suite de vecteurs

Une définition très intuitive de la convergence d'une suite de vecteurs de $\mathbb{R}^n$ s'impose en considérant les n suites formées par chacune des coordonnées. De manière précise, nous dirons qu'une suite de vecteurs

$$x^{(k)} = (x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}), \quad k = 0, 1, 2, \ldots$$

converge vers le vecteur limite $l = (l_1, l_2, \ldots, l_n)$ lorsque $k \to \infty$ si chaque composante (coordonnée) du vecteur $x^{(k)}$ converge vers la composante correspondante du vecteur l, c'est-à-dire,

$$\lim_{k\to\infty} x_i^{(k)} = l_i, \quad 1 \leq i \leq n.$$

Le théorème suivant montre que cette propriété naturelle s'exprime à l'aide de n'importe quelle norme et, bien que cela puisse paraître dans un premier temps surprenant, l'usage de ces dernières facilitera l'étude des suites de vecteurs plutôt qu'elle ne la compliquera.

Théorème 4. *Soit N une norme vectorielle sur $\mathbb{R}^n$. Pour qu'une suite de vecteurs $x^{(k)}$ converge vers l lorsque $k \to \infty$ il faut et il suffit que $\lim_{k\to\infty} N(x^{(k)} - l) = 0$.*

Démonstration. Nous montrons d'abord le théorème dans le cas où $N = \|\cdot\|_1$.

Étape 1. La condition est nécessaire

Nous supposons que $x_i^{(k)} \to l_i$ pour $i = 1, \ldots, n$ et montrons que $\|x^{(k)} - l\|_1 \to 0$ lorsque $k \to \infty$. Or $\|x^{(k)} - l\|_1 = \sum_{i=1}^n |x_i^{(k)} - l_i|$ mais $x_i^{(k)} \to l_i \iff |x_i^{(k)} - l_i| \to 0$ donc $\|x^{(k)} - l\|_1$ tend vers 0 comme somme de n suites tendant vers 0.

Étape 2. La condition est suffisante

Nous supposons cette fois que $\|x^{(k)} - l\|_1 \to 0$, fixons j quelconque dans $\{1, \ldots, n\}$ et montrons que $x_j^{(k)} \to l_j$. Nous avons

$$0 \leq |x_j^{(k)} - l_j| \leq \sum_{i=1}^n |x_i^{(k)} - l_i| = \|x^{(k)} - l\|_1 \to 0 \ (k \to 0)$$

de sorte que $|x_j^{(k)} - l_j| \to 0$ et $x_j^{(k)} \to l_j$, $k \to \infty$.

Maintenant si N est une norme quelconque, d'après le théorème 3, il existe des constantes c_1 et c_2 telles que

$$c_1 N(x) \leq \|x\|_1 \leq c_2 N(x), \quad x \in \mathbb{R}^n.$$

De l'inégalité inférieure nous déduisons que la condition $\lim_{k\to\infty} \|x^{(k)} - l\|_1 = 0$ implique $\lim_{k\to\infty} N(x^{(k)} - l) = 0$ tandis que nous déduisons de l'inégalité supérieure que la condition $\lim_{k\to\infty} N(x^{(k)} - l) = 0$ implique $\lim_{k\to\infty} \|x^{(k)} - l\| = 0$. Les conditions $\lim_{k\to\infty} \|x^{(k)} - l\| = 0$ et $\lim_{k\to\infty} N(x^{(k)} - l) = 0$ sont donc équivalents et, la première étant aussi équivalente à la convergence de la suite de vecteurs vers l, il en est de même de la seconde. ■

§ 3. NORMES MATRICIELLES

Il y a peu de différences entre un vecteur et une matrice. Nous pouvons d'ailleurs très bien considérer une matrice d'ordre n comme un vecteur formé de n^2 coordonnées, c'est à dire un élément de $\mathbb{R}^{n^2}$. Aussi les normes définies dans la partie précédente peuvent être utilisées pour les matrices et nous pourrions par exemple considérer la norme $\|M\|_1$ définie par $\max_{i,j=1,\dots,n}|a_{ij}|$ pour $M=(a_{ij})$. Cependant, en procédant de cette manière, nous perdons la possibilité d'exploiter simplement les propriétés du produit matriciel dans le calcul des normes de matrices. C'est pour cette raison qu'a été définie la notion de **norme matricielle** définie ci-dessous.

3.1 Définition

Nous travaillons sur l'espace vectoriel $M_n = M_n(\mathbb{R})$ des matrices carrées à n lignes et n colonnes formées de coefficients réels. Une application V de M_n dans $\mathbb{R}^+$ s'appelle une **norme matricielle** si elle vérifie les quatre conditions suivantes.

(a) $V(A)=0$ si et seulement si $A=0$, $A\in M_n$.

(b) $V(\lambda A)=|\lambda|V(A)$, $A\in M_n$, $\lambda\in\mathbb{R}$.

(c) $V(A+B)\le V(A)+V(B)$, A, $B\in M_n$.

(d) $V(AB)\le V(A)V(B)$, A, $B\in M_n$.

C'est seulement la dernière propriété qui distingue les normes matricielles des normes vectorielles. Elle s'appelle la propriété de **sous-multiplicativité**.

Les normes matricielles sur M_n pouvant être vues comme des normes vectorielles sur $\mathbb{R}^{n^2}$, le théorème 3 d'équivalence des normes s'applique à elles : *toutes les normes matricielles sur M_n sont équivalentes entre elles*.

Les plus importantes des normes sont celles qui sont induites, dans un sens expliqué ci-dessous, par une norme vectorielle.

E 108 Montrer que pour toute norme matricielle V, on a $V(I)\ge 1$ où I désigne la matrice identité.

3.2 Normes induites

Théorème 5. *Soit N une norme vectorielle sur $\mathbb{R}^n$. L'application V, définie sur M_n par*

$$V(A)=\sup_{x\neq 0}\frac{N(Ax)}{N(x)},$$

est une norme matricielle[a]*. Nous dirons que V est la norme matricielle* ***induite*** *par N.*

a. Par la notation Ax, nous entendons la matrice A multipliée par le vecteur colonne déterminé par $x\in\mathbb{R}^n$. Nous utiliserons aussi, lorsque cela sera plus clair, la notation $A(x)$.

Il découle immédiatement de la définition que

$$N(Ax)\le V(A)N(x),\quad x\in\mathbb{R}^n. \tag{3.1}$$

qui est une inégalité fondamentale que nous utiliserons de manière répétée. Observons d'ailleurs que si nous avons une inégalité de la forme

$$N(ax)\le MN(x),\quad x\in\mathbb{R}^n,$$

alors nous pouvons déduire que $V(A)\le M$. D'ailleurs $V(A)$ n'est autre la plus petites des constantes M pour lesquelles l'inégalité ci-dessus est satisfaite.

L'élément le plus ennuyeux dans la définition de $V(A)$, c'est que pour le calculer, il faille obtenir une borne supérieure (sup *) sur l'ensemble de tous les vecteurs non nuls. En réalité, nous pouvons nous limiter à considérer le sup pour tous les x satisfaisant $N(x) = 1$. En effet, posons

$$V'(A) = \sup_{N(x)=1} \frac{N(Ax)}{N(x)} = \sup_{N(x)=1} N(Ax),$$

puisque nous prenons ici le sup sur un ensemble plus petit, nous avons immédiatement $V'(A) \leq V(A)$. Ensuite, pour un vecteur x quelconque mais non nul, nous avons $N(x) \neq 0$, et, puisque le vecteur $y = \frac{1}{N(x)}x$ vérifie $N(y) = 1$,

$$\frac{N(Ax)}{N(x)} = N\left(\frac{1}{N(x)}Ax\right) = N(Ay) \leq V'(A),$$

de sorte qu'en prenant le sup pour $x \neq 0$ sur la gauche, nous obtenons $V(A) \leq V'(A)$ et, par double inégalité, $V(a) = V'(A)$. Pourtant, si cette observation permet de réduire l'ensemble des x sur lequel il faut déterminer un sup, elle n'implique pas, du moins directement, que ce sup ne soit pas égal à l'infini, hypothèse qui invaliderait immédiatement la conclusion du théorème. En réalité le sup est toujours fini et il est même toujours atteint. Cela résulte d'un simple résultat de topologie †. Ici, pour rester pleinement élémentaire, nous pouvons observer que si le sup est fini pour une norme N alors il le sera pour toutes les autres normes. Cela provient encore une fois du théorème 3 sur l'équivalence des normes, puisque si N_1 et N_2 sont deux normes, l'inégalité $c_1 N_1(x) \leq N_2(x) \leq c_2 N_1(x)$ entrainera

$$\frac{N_2(Ax)}{N_2(x)} \leq \frac{c_2}{c_1}\frac{N_1(Ax)}{N_1(x)},$$

de sorte que si le sup est fini pour N_1, il le sera aussi pour N_2. Or, à l'aide du théorème 6 établi plus bas, nous voyons que le sup est fini pour $\|\cdot\|_1$; il l'est donc pour n'importe quelle norme.

Démonstration. La question de savoir si V est correctement définie étant réglée, nous nous concentrons sur la vérification des quatre axiomes de la définition. Les points (a) et (b) sont immédiats ; nous démontrerons seulement les points (c) et (d), c'est-à-dire la sous-multiplicativité. Soient $A, B \in M_n$ et $x \neq 0$. Nous avons d'abord, en utilisant l'inégalité triangulaire, pour N

$$N((A+B)(x)) = N(Ax+Bx) \leq N(Ax) + N(Bx),$$

d'où nous tirons

$$\frac{N((A+B)(x))}{N(x)} \leq \frac{N(Ax)}{N(x)} + \frac{N(Bx)}{N(x)} \leq V(A) + V(B).$$

Puis, en prenant le sup sur la gauche, il vient

$$\sup_{x\neq 0} \frac{N((A+B)(x))}{N(x)} \leq V(A) + V(B),$$

qui donne finalement

$$V(A+B) \leq V(A) + V(B).$$

Pour montrer la sous-multiplicativité de V, nous procédons comme suit.

Étape 1. $Bx = 0$.

∗. À partir de maintenant, lorsque cela permettra d'alléger le texte nous écrirons *sup* pour *borne supérieure*.

†. ∗ Le lecteur ayant suivi un cours de topologie élémentaire reconnaîtra que la fonction A est continue, que l'ensemble des x tel que $N(x) = 1$ est un ensemble fermé borné de $\mathbb{R}^n$, donc un compact, sur lequel toute fonction continue atteint ses bornes.

Dans ce cas $(AB)(x) = A(0) = 0$ de sorte que $\frac{N((AB)(x))}{N(x)} = 0 \leq V(A)\cdot V(B)$.

Étape 2. $Bx \neq 0$;

Dans ce cas, nous pouvons par $N(Bx) \neq 0$. Nous avons alors

$$\frac{N((AB)(x)}{N(x)} = \frac{N(B(Ax))}{N(Bx)}\frac{N(Bx)}{N(x)} \leq V(A)\cdot V(B).$$

Par conséquent *pour tout x non nul* nous avons

$$\frac{N((AB)(x))}{N(x)} \leq V(A)\cdot V(B).$$

En prenant le sup dans le terme de gauche nous obtenons $V(A\cdot B) \leq V(A)\cdot V(B)$ et cela termine la preuve que V est une norme matricielle. ■

E 109 Montrer que si V est la norme matricielle induite par une norme vectorielle N alors $V(I) = 1$. Comparer avec l'exercice 108.

3.3 Exemples fondamentaux

Nous montrons que pour deux normes très simples, le calcul de la norme induite est très simple, au contraire de ce que pouvait laisser entrevoir la définition. Nous notons $|||\cdot|||_1$ (resp. $|||\cdot|||_\infty$) la norme matricielle induite par $\|\cdot\|_1$ (resp. par $\|\cdot\|_\infty$),

$$|||A|||_1 = \sup_{x\neq 0}\frac{\|Ax\|_1}{\|x\|_1} \quad \text{et} \quad |||A|||_\infty = \sup_{x\neq 0}\frac{\|Ax\|_\infty}{\|x\|_\infty}.$$

Théorème 6. *Si $A = (a_{ij}) \in M_n$ alors*

$$|||A|||_1 = \max_{j=1,\dots,n}\sum_{i=1}^{n}|a_{ij}| \quad \textit{et} \quad |||A|||_\infty = \max_{i=1,\dots,n}\sum_{j=1}^{n}|a_{ij}|.$$

Par exemple, si

$$A = \begin{pmatrix} 1 & 0 & 3 \\ -2 & -1 & 0 \\ 3 & 2 & 0 \end{pmatrix}, \quad |||A|||_1 = \max(6,3,3) = 6 \quad \text{et} \quad |||A|||_\infty = \max(4,3,5) = 5.$$

Démonstration.

Étape 1. Pour la norme $|||\cdot|||_1$.
Pour $x = (x_1, x_2, \dots, x_n) \neq 0$ nous avons

$$Ax = (\sum_{j=1}^{n} a_{1j}x_j, \sum_{j=1}^{n} a_{2j}x_j, \dots, \sum_{j=1}^{n} a_{nj}x_j),$$

et, en utilisant en particulier les propriétés de la valeur absolue et une permutation de sommation,

$$\|Ax\|_1 = \sum_{i=1}^{n}\left|\sum_{j=1}^{n}a_{ij}x_j\right| \leq \sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}||x_j| \leq \sum_{j=1}^{n}|x_j|\left(\sum_{i=1}^{n}|a_{ij}|\right) \leq \sum_{j=1}^{n}|x_j| \max_{j=1,\dots,n}\left(\sum_{i=1}^{n}|a_{ij}|\right)$$
$$\leq \max_{j=1,\dots,n}\left(\sum_{i=1}^{n}|a_{ij}|\right) \times \sum_{j=1}^{n}|x_j| \leq \max_{j=1,\dots,n}\left(\sum_{i=1}^{n}|a_{ij}|\right) \times \|x\|_1.$$

Il suit que

$$\frac{\|Ax\|_1}{\|x\|_1} \leq \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij}|.$$

En prenant le sup de tous les termes sur la gauche gauche, nous arrivons à

$$|||A|||_1 \leq \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij}|.$$

Pour nous assurer que le terme de droite est bien un sup et non uniquement un majorant du sup, il suffit de mettre en évidence une valeur de x pour laquelle nous aurons l'égalité

$$\frac{\|Ax\|_1}{\|x\|_1} = \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij}|.$$

Prenons un indice j_0 tel que

$$\sum_{i=1}^{n} |a_{ij_0}| = \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij}|,$$

et posons ensuite $x_0 = (0,0,\dots,0,1,0\dots,0)$ où le 1 se trouve à la j_0-ième coordonnée. La calcul de $\|Ax_0\|_1$ donne $\sum_{i=1}^{n} |a_{ij_0}|$ et comme $\|x_0\|_1 = 1$, nous avons la propriété souhaitée.

Étape 2. Pour la norme $|||\cdot|||_\infty$.

Le principe est très similaire. Observons d'abord que

$$\|Ax\|_\infty = \max_{i=1,\dots,n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \leq \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}||x_j| \leq \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}| \|x\|_\infty \leq \|x\|_\infty \cdot \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}|,$$

d'où il résulte rapidement que

$$|||A|||_\infty \leq \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}|.$$

Pour s'assurer que l'égalité est réalisée, il suffit de trouver un vecteur x_0 tel que $\|x_0\|_\infty = 1$ et

$$\|Ax_0\|_\infty \geq \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}|;$$

puisque cela impliquera $V(A) \geq \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}|$ et par double inégalité, le résultat recherché. Pour cela, nous choisissons un indice i_0 pour lequel le max est atteint, puis définissons x_0 en fixant sa j-ème coordonnée égale au signe ($+1$ ou -1) du coefficient $a_{i_0 j}$. Cette définition entraine que $\|x_0\|_\infty = 1$ et

$$a_{i_0 j} x_{0j} = |a_{i_0 j}|, \quad j = 1,\dots n.$$

Il suit que

$$\max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}| = \sum_{j=1}^{n} |a_{i_0 j}| = |\sum_{j=1}^{n} a_{i_0 j} x_{0j}| = |(Ax_0)_{i_0}| \leq \|Ax_0\|_\infty,$$

et cela achève la démonstration du théorème. ■

E 110 Puisque les normes $|||.|||_1$ et $|||.|||_\infty$ sont équivalentes, il existe des constantes c et C telles que $c|||.|||_1 \leq |||.|||_\infty \leq C|||.|||_1$. Déterminer de telles constantes c et C.

3.4 Convergence d'une suite de matrices

Exactement comme dans le cas des suites de vecteurs, il est naturel de dire qu'une suite de matrices

$$A^{(k)} = (a_{ij}^{(k)}) \in M_n \quad k = 0, 1, 2, \ldots$$

converge vers une matrice $L = (l_{ij}) \in M_n$ lorsque $k \to \infty$ lorsque chacune des n^2 suites de coefficients $a_{ij}^{(k)}$ converge vers le coefficient correspondant l_{ij} de L, c'est-à-dire

$$\lim_{k\to\infty} A^{(k)} = L \quad \Longleftrightarrow \quad \lim_{k\to\infty} a_{ij}^{(k)} = l_{ij}, \quad i, j \in \{1, 2, \ldots, n\}.$$

Théorème 7. *Si $A^{(k)}$ est une suite de matrices de M_n convergeant vers L lorsque $k \to \infty$ alors quel que soit le vecteur $x \in \mathbb{R}^n$, la suite de vecteurs $A^{(k)}x$ converge vers le vecteur Lx.*

Démonstration. Soit i quelconque dans $\{1, \ldots, n\}$. Nous montrons que la i-ième coordonnée de $A^{(k)}x$ converge vers la i-ième coordonnée de Lx. Nous avons

$$\left(A^{(k)}x\right)_i = \sum_{j=1}^{n} a_{ij}^{(k)} x_j \overset{k\to\infty}{\longrightarrow} \sum_{j=1}^{n} \left\{\lim_{k\to\infty} a_{ij}^{(k)}\right\} x_j = \sum_{j=1}^{n} L_{ij} x_j = (Lx)_i. \qquad \blacksquare$$

Nous montrons maintenant que la convergence d'une suite de matrice s'exprime facilement en faisant intervenir une norme matricielle.

Théorème 8. *Soit V une norme matricielle sur $M_n(\mathbb{R})$. Pour qu'une suite de matrices $A^{(k)}$ converge vers L lorsque $k \to \infty$ il faut et il suffit que $\lim_{k\to\infty} V(A^{(k)} - L) = 0$.*

Démonstration. Nous effectuons la démonstration dans le cas où $V = |||\cdot|||_1$). Le cas général s'obtient en faisant appel au théorème 3 d'équivalence des normes comme nous l'avons fait dans la démonstration du théorème 4.

Supposons que $|||A^{(k)} - L|||_1 \to 0$ lorsque $k \to \infty$ et fixons i_0, j_0 dans $\{1, 2, \ldots, n\}$. Nous avons

$$0 \leq \left|a_{i_0,j_0}^{(k)} - l_{i_0,j_0}^{(k)}\right| \leq \sum_{i=1}^{n} \left|a_{i,j_0}^{(k)} - l_{i,j_0}^{(k)}\right| \leq \max_j \sum_{i=1}^{n} \left|a_{i,j}^{(k)} - l_{i,j}^{(k)}\right| = |||A^{(k)} - L|||_1.$$

Par conséquent, $\lim_{k\to\infty} |||A^{(k)} - L|||_1 = 0 \Rightarrow \lim_{k\to\infty} \left|a_{i_0,j_0}^{(k)} - l_{i_0,j_0}^{(k)}\right|$ i.e. $\lim_{k\to\infty} a_{i_0,j_0}^{(k)} = l_{i_0,j_0}^{(k)}$.

Réciproquement, si pour tous $i, j \in \{1, \ldots, n\}$ nous savons que $\lim_{k\to\infty} a_{i,j}^{(k)} = l_{i,j}^{(k)}$ alors

$$\lim_{k\to\infty} |||A^{(k)} - L|||_1 = 0.$$

En effet, la suite $x_j^{(k)}$ définie par

$$x_j^{(k)} = \sum_{i=1}^{n} \left|a_{i,j}^{(k)} - l_{i,j}^{(k)}\right|$$

converge vers 0 comme somme de n suites convergeant vers 0. Maintenant le maximum de n suites convergeant vers 0 converge aussi vers 0 de sorte que

$$|||A^{(k)} - L|||_1 = \max_j x_j^{(k)} \to 0, \quad k \to \infty. \qquad \blacksquare$$

3.5 Algèbre des limites

Théorème 9. *Soient $(A^{(k)})$ et $(B^{(k)})$ deux suites de matrices de $M_n(\mathbb{R})$ et $\lambda \in \mathbb{R}$. Supposons que*

$$\lim_{k\to\infty} A^{(k)} = L_1 \quad \text{et} \quad \lim_{k\to\infty} B^{(k)} = L_2.$$

Alors

(a) $\lim_{k\to\infty} A^{(k)} + B^{(k)} = L_1 + L_2$,

(b) $\lim_{k\to\infty} \lambda A^{(k)} = \lambda L_1$,

(c) $\lim_{k\to\infty} A^{(k)} B^{(k)} = L_1 L_2$.

(d) $\lim_{k\to\infty} V(A^{(k)}) = V(L)$ *où V est une norme matricielle quelconque.*

Démonstration. Prenons V une norme matricielle. D'après le théorème précédent, pour montrer (a), il suffit de vérifier que

$$V\big((A^{(k)} + B^{(k)}) - (L_1 + L_2)\big) \overset{k\to\infty}{\longrightarrow} 0.$$

Or, grâce à l'inégalité triangulaire pour V, nous pouvons écrire

$$0 \le V\big((A^{(k)} + B^{(k)}) - (L_1 + L_2)\big) \le \underbrace{V(A^{(k)} - L_1)}_{\text{tend vers 0}} + \underbrace{V(B^{(k)} - L_2)}_{\text{tend vers 0}},$$

qui implique que le terme intermédiaire tend vers 0.

Le point (b) se démontre de manière similaire. Démontrons le (c). On a

$$\begin{aligned} 0 &\le V(A^{(k)}B^{(k)} - L_1L_2) \\ &= V\big((A^{(k)} - L_1)B^{(k)} + L_1(B^{(k)} - L_2)\big) \\ &\le \underbrace{V(A^{(k)} - L_1)}_{\text{tend vers 0}}.V(B^{(k)}) + \underbrace{V(L_1)}_{\text{constante}}.\underbrace{V(B^{(k)} - L_2)}_{\text{tend vers 0}}, \end{aligned}$$

où nous avons utilisé, pour la dernière inégalité, les propriétés (c) et (d) des normes matricielles. Étudions le terme $V(B^{(k)})$. Nous avons

$$V(B^{(k)}) = V(B^{(k)} - L + L) \le \underbrace{V(B^{(k)} - L)}_{\text{tend vers 0}} + \underbrace{V(L)}_{\text{constante}}.$$

Nous en déduisons que la suite $V(B^{(k)})$ est *bornée* si bien que

$$V(A^{(k)} - L_1) \cdot V(B^{(k)})$$

converge vers 0 et, finalement, la suite positive $V(A^{(k)}B^{(k)} - L_1L_2)$ se trouve majorée par une suite qui converge vers 0. Elle converge donc elle-même vers 0.

Le dernier point est conséquence de l'inégalité

$$\Big|V(A^{(k)}) - V(L)\Big| \le V(A^{(k)} - L),$$

voir (2.3) en tenant compte du fait, souligné plus haut, que les normes matricielles sont des normes vectorielles. ∎

E 111 Enoncer et établir les propriétés (a), (b) et (d) dans le cas des suites de vecteurs.

3.6 Le critère de Cauchy

Pour qu'une suite (x^k) de nombres réels converge il faut et il suffit — c'est un théorème fondamental de l'analyse — qu'elle satisfasse le **critère de Cauchy** i.e.

> Quel que soit $\varepsilon > 0$, il existe $p_0 \in \mathbb{N}$ tel que les conditions $m \in \mathbb{N}$ et $p \geq p_0$ entraînent $|x^{p+m} - x^p| \leq \varepsilon$.

Ce critère a déjà été utilisé au III.5.5. Un critère similaire se déduit facilement (en replaçant la valeur absolue par une norme) pour les suites de vecteurs et les suites de matrices. Nous l'énonçons uniquement dans ce dernier cas.

Théorème 10 (Critère de Cauchy pour les matrices). *Soit V une norme matricielle sur $M_n(\mathbb{R})$. Pour qu'une suite de matrices $A^{(k)}$ soit convergente il faut et il suffit qu'elle vérifie le* critère de Cauchy *à savoir :*

> *Quel que soit $\varepsilon > 0$, il existe $p_0 \in \mathbb{N}$ tel que les conditions $m \in \mathbb{N}$ et $p \geq p_0$ entraînent $V(A^{(p+m)} - A^{(p)}) \leq \varepsilon$.*

Démonstration. Elle s'effectue en se ramenant au cas des suites de nombres réels comme dans la démonstration du théorème 8. ■

§ 4. SUITES ET SÉRIES GÉOMÉTRIQUES DE MATRICES

4.1 Suites géométriques

Soit $A \in M_n$. On appelle **suite géométrique** (matricielle) de **raison** A, la suite de matrices dont le k-ième terme est

$$A^k = \underbrace{A \cdot A \cdot A \cdots A}_{\text{k fois}}.$$

Nous conviendrons que A^0 désigne toujours la matrice identité, $A^0 = I$.

Théorème 11. *Soit V une norme matricielle sur M_n et $A \in M_n$. Si $V(A) < 1$ alors $\lim_{k\to\infty} A^k = 0$. Autrement dit, une suite géométrique converge vers 0 dès que la norme de sa raison est strictement plus petite que 1.*

Démonstration. C'est une conséquence de la propriété de sous-multiplicativité des normes matricielles qui justifie chacune des inégalités ci-dessous. En effet,

$$\begin{aligned} 0 \leq V(A^k - 0) = V(A^K) = V(A \cdot A^{k-1}) \leq V(A) \cdot V(A^{k-1}) \leq V(A) \cdot V(A \cdot A^{k-2}) \\ \leq V(A) \cdot V(A) \cdot V(A^{k-2}) \leq V(A)^2 V(A^{k-2}) \leq \cdots \leq V(A)^k V(A^0) = V(A)^k V(Id). \end{aligned}$$

Or, puisque $V(A) \in [0,1[$, nous avons $\lim_{k\to\infty} [V(A)]^k = 0$ donc aussi, puisque $V(Id)$ est une constante,

$$\lim_{k\to\infty} [V(A)]^k V(Id) = 0,$$

et, par encadrement, il suit que $\lim_{k\to\infty} V(A^k) = 0$, ce qui signifie, d'après le théorème 8, $\lim_{k\to\infty} A^k = 0$. ■

4.2 Séries géométriques

Soit $A \in M_n$. Nous notons $S_k(A)$ la somme des k premiers termes de la suite géométrique de raison A,

$$S_k(A) = I + A + A^2 + \cdots + A^k,$$

où I désigne la matrice identité. La suite $S_k(A)$ est appelée **série géométrique (matricielle)** de raison A. Lorsque cette suite est convergente, sa limite est notée $\sum_{k=0}^{\infty} A^k$. On dit alors souvent, par abus de langage, que la série $\sum_{k=0}^{\infty} A^k$ est convergente.

Théorème 12. *Soient V une norme matricielle sur M_n et A une matrice vérifiant $V(A) < 1$. La série géométrique de raison A est convergente et on a*

$$\sum_{k=0}^{\infty} A^k = (I-A)^{-1}.$$

En particulier la matrice $(I-A)$ est inversible.

Démonstration. Nous établissons la convergence de la suite $S_k(A)$ en vérifiant le critère de Cauchy ce qui est justifie par le théorème 10.

Prenant $\varepsilon > 0$ quelconque, nous devons trouver $p_0 \in \mathbb{N}$, dépendant de ε, tel que

$$(m \in \mathbb{N} \text{ et } p \geq p_0) \Rightarrow V(S_{p+m}(A) - S_p(A)) \leq \varepsilon.$$

Étudions la quantité $V(S_{p+m}(A) - S_p(A))$. Observons d'abord que

$$S_{p+m}(A) - S_p(A) = A^{p+1} + A^{p+2} + \cdots + A^{p+m}.$$

En utilisant l'inégalité triangulaire, voir (2.1), nous obtenons

$$V\left(S_{p+m}(A) - S_p(A)\right) = V\left(A^{p+1} + A^{p+2} + \cdots + A^{p+m}\right) \leq V(A^{p+1}) + V(A^{p+2}) + \cdots + V(A^{p+m}).$$

Ensuite, en utilisant la majoration $V(A^j) \leq C(V(A))^j$ où $c = V(I)$ que nous avons établie dans la démonstration du théorème 11 pour majorer le terme de droite dans l'inégalité ci-dessus, nous obtenons

$$\begin{aligned} V\left(S_{p+m}(A) - S_p(A)\right) &\leq CV(A)^{p+1} + CV(A)^{p+2} + \cdots + CV(A)^{p+m} \\ &\leq CV(A)^{p+1}\left(1 + V(A) + \cdots + V(A)^{m-1}\right) \leq CV(A)^{p+1}\frac{1-V(A)^m}{1-V(A)}.. \end{aligned}$$

Or, puisque $V(A) < 1$, le dernier terme sur la droite est encore majoré par $CV(A)^{p+1}/(1-V(A))$ si bien que finalement

$$V\left(S_{p+m}(A) - S_p(A)\right) \leq \frac{CV(A)^{p+1}}{1-V(A)}.$$

Maintenant $\varepsilon > 0$ étant fixé. Puisque $V(A) < 1$, nous avons $\lim_{p\to\infty} \frac{CV(A)^{p+1}}{1-V(A)} = 0$ donc il existe $p_0 \in \mathbb{N}$ tel que

$$p \geq p_0 \Rightarrow \frac{CV(A)^{p+1}}{1-V(A)} \leq \varepsilon,$$

et par conséquent

$$\left.\begin{matrix} m \in \mathbb{N} \\ p \geq p_0 \end{matrix}\right\} \Rightarrow V(S_{p+m}(A) - S_p(A)) \leq \frac{CV(A)^{p+1}}{1-V(A)} \leq \varepsilon,$$

donc la suite $S_k(A)$ vérifie le critère de Cauchy et elle est bien convergente.

Appelons L sa limite, i.e. $S_k(A) \to L$. Nous avons donc aussi, par le théorème 9, $(I-A)S_k(A) \to (I-A)L$ mais

$$(I-A)S_k(A) = (I-A)(I+A+A^2+\cdots+A^k) = I-A^{k+1}.$$

Il suit que

$$\lim_{k\to\infty}(I-A^{k+1}) = (I-A)L \tag{4.1}$$

mais puisque $V(A) < 1$, d'après le théorème 11, $\lim_{k\to\infty} A^{k+1} = 0$ donc aussi, en utilisant le thèorème 9, $\lim_{k\to\infty}(I-A^{k+1}) = I-0 = I$. Finalement, avec (4.1), nous arrivons à $I = (I-A)L$ de sorte que $(I-A)$ est inversible et son inverse est L, $L = (I-A)^{-1}$. Nous avons ainsi démontré que la série géométrique de raison A converge vers $(I-A)^{-1}$, c'est-à-dire

$$\sum_{j=0}^{\infty} A^j = (I-A)^{-1}.$$ ■

Corollaire 13. *Sous les hypothèses du théorème et en supposant de plus que $V(Id) = 1$ on a*

$$V\left((I-A)^{-1}\right) \leq \frac{1}{1-V(A)}. \tag{4.2}$$

Démonstration. En utilisant l'inégalité triangulaire et la sous-multiplicativité de V, nous obtenons

$$V(S_k(A)) \leq \sum_{j=0}^{k} V^k(A).$$

Maintenant le théorème précédent et le théorème 9 (iv) donne

$$V\left((I-A)^{-1}\right) = V\left(\lim_{k\to\infty} S_k(A)\right) = \lim_{k\to\infty} V\left(S_k(A)\right) \leq \lim_{k\to\infty} \sum_{j=0}^{k} V^k(A) = \sum_{j=0}^{\infty} V^k(A) = \frac{1}{1-V(A)},$$

ce qu'il fallait démontrer. ■

E 112 Montrer que sous les hypothèses du corollaire 13, on a aussi

$$\frac{1}{1+V(A)} \leq V((I-A)^{-1}).$$

On pourra utiliser la relation $I = (I-A)(I-A)^{-1}$.

§ 5. APPLICATIONS

5.1 Équations matricielles de la forme $x = b + Ax$

Considérons une équation de la forme $x = b + Ax$ où $b \in \mathbb{R}^n$, $A = (a_{ij}) \in M_n(\mathbb{R})$ et $x \in \mathbb{R}^n$ est le vecteur inconnu. Cette équation est la forme matricielle du système

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = x_1 - b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = x_2 - b_2 \\ \vdots \quad\quad \vdots \quad\quad\quad\quad \vdots \quad\quad \vdots \\ a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n = x_i - b_i \\ \vdots \quad\quad \vdots \quad\quad\quad\quad \vdots \quad\quad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = x_n - b_n \end{cases}.$$

L'équation $x = b + Ax$ est équivalente à $(I-A)x = b$ et admet donc une solution unique si et seulement si $I-A$ est inversible ce qui est toujours le cas, d'après le théorème 12, lorsqu'il existe une norme matricielle V pour laquelle $V(A) < 1$. Dans ce cas, nous construisons la suite de vecteurs $(x^{(k)})$ définie par la relation de récurrence suivante

$$\begin{cases} x^{(0)} &= \omega \\ x^{(k+1)} &= b + Ax^{(k)}, \quad k \geq 0, \end{cases}$$

Théorème 14. *S'il existe une norme matricielle V telle que $V(A) < 1$ alors, quel que soit le vecteur de départ ω, la suite $x^{(k)}$ définie par la relation de récurrence ci-dessus converge vers l'unique solution du système linéaire $x = b + Ax$.*

Démonstration. Nous pouvons écrire une formule explicite pour le k-ème terme de la suite. En effet,

$$\begin{aligned} x^{(0)} &= w \\ x^{(1)} &= b + Aw \\ x^{(2)} &= b + Ax^{(1)} = b + A(b + Aw) = b + Ab + A^2 w \\ x^{(3)} &= b + Ax^{(2)} = b + A(b + Aw + A^2 w) = b + Ab + A^2 w + A^3 w. \end{aligned}$$

De manière générale, nous établissons facilement par récurrence que

$$x^{(k)} = b + Ab + \cdots + A^{k-1}b + A^k w = S_{k-1}(A)b + A^k w, \quad k \in \mathbb{N} \tag{5.1}$$

où $S_{k-1}(A) = I + A + A^2 + \cdots + A^{k-1}$. D'après le théorème 12) ci-dessus, puisque $V(A) < 1$, la suite $S_{k-1}(A)$ converge vers $(I-A)^{-1}$ et donc, comte tenu du théorème 7, la suite de vecteurs $S_{k-1}(A)b$ converge vers le vecteur $(I-A)^{-1}(b)$. D'autre part, toujours puisque $V(A) < 1$, nous avons, voir le théorème (11, $A^k w \to 0w = 0$. Passant à la limite dans (5.1), nous obtenons finalement

$$\lim_{k\to\infty} x^{(k)} = \lim_{k\to\infty} S_{k-1}(A)b + \lim_{k\to\infty} A^k w = (I-A)^{-1}(b) + 0 = (I-A)^{-1}(b),$$

qui est bien la solution du système considéré. ∎

Nous pouvons d'ailleurs facilement estimer la rapidité de convergence de la suite $(x^{(k)})$ vers la solution x dans le cas où la norme V est **subordonnée** à la norme vectorielle N, c'est-à-dire vérifie

$$N(Cx) \leq V(C)N(x), \quad C \in M_n(\mathbb{R}), x \in \mathbb{R}^n,$$

ce qui est toujours le cas lorsque V est la norme matricielle induite par N, voir (3.1).

En effet si une telle inégalité est vérifiée, à partir de (5.1), nous avons

$$\begin{aligned} x^{(k)} - x &= S_{k-1}(A)b + A^k w - (I-A)^{-1}b \\ &= (I + A + \cdots + A^{k-1})b - \left(\sum_{j=0}^{\infty} A^j\right)(b) + A^k w \\ &= -(A^k + A^{k+1} + \ldots)(b) + A^k w \\ &= -\left(\sum_{j=k}^{\infty} A^j\right)(b) + A^k w \\ &= -A^k(I + A + A^2 + \ldots)(b) + A^k w \\ &= -A^k\left(\sum_{j=0}^{\infty} A^j\right)(b) + A^k w \\ &= -A^k(I-A)^{-1}(b) + A^k w. \end{aligned}$$

En utilisant le fait que V est subordonnée à N, et l'inégalité $V(A^k) \leq V^k(A)V(Id)$, l'estimation ci-dessus donne

$$N(x^{(k)} - x) \leq V(Id)V^k(A) \cdot V\left((I-A)^{-1}\right) N(b) + V^k(A) \cdot N(w).$$

Si nous supposons en outre que $V(Id) = 1$, à l'aide de (4.2), nous arrivons

$$N(x^{(k)} - x) \leq V^k(A) \underbrace{\left(\frac{N(b)}{1 - V(A)} + N(w) \right)}_{\text{indépendant de } k}.$$

Plus $V(a)$ sera petit, plus $V^k(A)$ convergera vite vers 0 et donc plus la convergence de $x^{(k)}$ vers la solution x sera rapide.

5.2 Effet sur la solution d'une perturbation des coefficients de la matrice

Il arrive très souvent que l'on doive résoudre un système linéaire $Ax = c$, où A est inversible, pour lequel les valeurs exactes des coefficients de $A \in M_n$ et (ou) de $c \in \mathbb{R}^n$ ne sont pas exactement disponibles, ce qui est le cas lorsqu'ils proviennent de mesures, nécessairement imparfaites, ou encore sont disponibles mais pas directement exploitables sur un calculateur (travaillant en précision finie) ce qui est toujours le cas lorsque ces coefficients ne sont pas des décimaux. On est alors contraint de résoudre un **système approché** $A'x = c'$ où A' est, en un certain sens, proche de A et c' proche de c. Il n'est pas difficile de voir que si A' est suffisamment proche de A alors A' sera inversible et le système approché admettra bien une solution unique. La question fondamentale qui se pose alors est de savoir si l'unique solution de $A'x = c'$ sera proche de l'unique solution de $Ax = b$ et, surtout, de mesurer précisément cette proximité. les outils nécessaires pour étudier le problème sont les suivants.

(a) Nous utiliserons V une norme matricielle sur M_n induite par la norme vectorielle N sur $\mathbb{R}^n$.

(b) La différence entre la matrice théorique A et la matrice de travail A' est notée $\Delta(A) = A' - A$. . De même nous notons $\delta(c) = c' - c$ la différence entre le vecteur théorique c et le vecteur de travail c'.

(c) La proximité de A et A' et de c et c' sera mesurée, non par $V(\Delta(A))$ et $N(\delta(c)$, mais par les erreurs relatives

$$\frac{V(\Delta(A))}{V(A)}, \quad \frac{N(\delta(c))}{N(c)}.$$

(d) Appelons enfin r la solution de $Ax = c$ et r' la solution de $A'x = c$, puis $\delta(r) = r' - r$.

Le problème consiste donc à estimer la valeur de $\frac{N(\delta(r))}{N(r)}$ en fonction de $\frac{V(\Delta(A))}{V(A)}$ et de $\frac{N(\delta(c))}{N(c)}$.

Théorème 15. *On utilise les notations ci-dessus. Supposons que*

$$V\left(\Delta(A)\right) < \frac{1}{V(A^{-1})}. \tag{5.2}$$

Alors, posant $K(A) = V(A)V(A^{-1})$, on a

$$\frac{N\left(\delta(r)\right)}{N(r)} \leq \frac{K(A)}{1 - K(A)\frac{V\left(\Delta(A)\right)}{V(A)}} \left(\frac{N(\delta(c))}{N(c)} + \frac{V\left(\Delta(A)\right)}{V(A)} \right). \tag{5.3}$$

Le nombre $K(A)$ s'appelle le **nombre de conditionnement**, ou simplement le **conditionnement** de la matrice A. Plus ce nombre sera petit, moins la solution du système $Ax = c$ sera sensible aux perturbations des coefficients de A et de c.

Démonstration. Nous établissons quelques assertions qui conduisent à l'inégalité demandée.

Étape 1. $V\big(A^{-1}\Delta(A)\big) < 1$.

En effet, en vue de (5.2), la sous-multiplicativité de V donne

$$V\big(A^{-1}\Delta(A)\big) \leq V(A^{-1})V(\Delta(A)) < 1.$$

En vertu du théorème 12 cette première étape implique que $I + A^{-1}\Delta(A)$ est inversible.

Étape 2.

$$V\big((I - A^{-1}\Delta(A))^{-1}\big) \leq \frac{1}{1 - V(A^{-1})V\big(\Delta(A)\big)}.$$

C'est une simple application du corollaire 13.

Étape 3.

$$\delta(r) = \big(I + A^{-1}\Delta(A)\big)^{-1} A^{-1}\big(\delta(c) - \Delta(A)r\big). \tag{5.4}$$

Observons d'abord que

$$\Delta(A)(r) = (A - A')(r) = A(r) - A'(r' + \delta(r)) = c - c' - A'(\delta(r)) = \delta(c) - A'(\delta(r));$$

d'où nous tirons

$$\delta(c) - \Delta(A)(r) = A'(\delta(c)).$$

D'autre part,

$$A'(\delta(r)) = (A + \delta(A))(\delta(r)) = A(I - A^{-1}\Delta(A))(\delta(r)).$$

En regroupant les deux calculs, nous obtenons

$$\delta(c) - \Delta(A)(r) = A(I - A^{-1}\Delta(A))(\delta(r));$$

une relation immédiatement équivalente à celle qu'il fallait établir. L'estimation recherchée s'obtient maintenant en prenant la norme N de $\delta(c)$ dans (5.4) en tenant compte du fait que la norme V satisfait l'inégalité (3.1) et en utilisant l'estimation établie à la première étape. ■

Lorsque $\Delta(A) = 0$ (les coefficients de A sont sûrs), l'inégalité (5.3) se réduit à

$$\frac{N\big(\delta(r)\big)}{N(r)} \leq K(A)\,\frac{N(\delta(c))}{N(c)}.$$

Plus généralement, nous avons

$$\frac{1}{K(A)}\,\frac{N(\delta(c))}{N(c)} \leq \frac{N\big(\delta(r)\big)}{N(r)} \leq K(A)\,\frac{N(\delta(c))}{N(c)}.$$

Pour montrer l'inégalité sur la gauche il suffit d'observer d'une part que

$$A\delta(r) = \delta(b) \implies N(\delta(b)) \leq V(A)N(\delta(r)),$$

et d'autre part

$$Ar = b \implies r = A^{-1}b \implies N(r) \leq V(A^{-1})N(b) \implies \frac{1}{N(b)} \leq V(A^{-1})\frac{1}{N(r)}.$$

En multipliant les deux inégalités sur la droite ci-dessus, il vient

$$\frac{N(\delta(b))}{N(b)} \leq V(A)V(A^{-1}))\frac{N(\delta(r)}{N(r)},$$

d'où découle la relation cherchée.

Remarquons que si le nombre de conditionnement d'une matrice donne des informations, il n'est pas facile à calculer à partir de sa définition puisqu'il faudrait disposer de A^{-1} pour en déterminer la valeur. Il existe certaines techniques, plus ou moins efficaces, qui permettent de se faire une idée de $K(A)$ en utilisant moins d'opérations que pour calculer A^{-1}, voir les notes et commentaires à la fin du chapitre.

§ 6. DÉCOMPOSITION ET SUITE DE JACOBI

6.1 Définition

Nous considérons le système $Ax = b, A \in M_n$, inversible, $b \in \mathbb{R}^n$. Si $A = (a_{ij})$, nous utiliserons la décomposition suivante en **partie triangulaire supérieure**, **partie diagonale** et **partie triangulaire inférieure**, précisément,

$$A = D - E - F, \quad \text{où}$$

(a) $D = (d_{ij})$ est la matrice diagonale reprenant les éléments diagonaux de A : $d_{ii} = a_{ii}$ et $d_{ij} = 0$ pour $i \neq j$.

(b) $E = (e_{ij})$ est la matrice triangulaire inférieure reprenant les *opposés* des éléments de A : $e_{ij} = -a_{ij}$ pour $i > j$ et $e_{ij} = 0$ pour $i \leq j$.

(c) $F = (f_{ij})$ est la matrice triangulaire supérieure reprenant les *opposés* des éléments de A : $f_{ij} = -a_{ij}$ pour $i < j$ et $f_{ij} = 0$ pour $j \leq i$.

Exemple 1.

$$\begin{array}{ccccccccc} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} & = & \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 9 \end{pmatrix} & - & \begin{pmatrix} 0 & 0 & 0 \\ -4 & 0 & 0 \\ -7 & -8 & 0 \end{pmatrix} & - & \begin{pmatrix} 0 & -2 & -3 \\ 0 & 0 & -6 \\ 0 & 0 & 0 \end{pmatrix}. \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \\ A & = & D & - & E & - & F & & \end{array}$$

Lorsque D est inversible, ce qui est le cas si et seulement si $a_{ii} \neq 0$ pour tout $i \in \{1, \ldots, n\}$, alors nous pouvons former la matrice la matrice

$$B_J = D^{-1}(E+F) = D^{-1}(D-A) = I - D^{-1}A,$$

et le vecteur

$$b_J = D^{-1}b.$$

La matrice B_J s'appelle la **matrice de Jacobi** de A (et du système $Ax = b$) et le vecteur b_J s'appelle le **vecteur de Jacobi**.

Théorème 16. *Si D est inversible alors x est solution de $Ax = b$ si et seulement si x est solution de $x = b_J + B_J x$.*

Démonstration. C'est un calcul simple.

$$\begin{aligned} Ax = b &\iff D^{-1}(Ax) = D^{-1}(b) \iff (D^{-1}A)(x) = D^{-1}(b) \iff 0 = D^{-1}(b) - (D^{-1}A)(x) \\ &\iff x = D^{-1}(b) + x - (D^{-1}A)(x) \iff x = D^{-1}(b) + (I - D^{-1}A)(x) \iff x = b_J + B_J x. \quad \blacksquare \end{aligned}$$

La matrice de Jacobi permet ainsi de transformer un système de la forme $Ax = b$ en un système de la forme $x = b' + A'x$. Or, nous a vu au théorème 14 que lorsque, pour une certaine norme matricielle, $V(B) < 1$, nous pouvons construire une suite qui converge rapidement vers la solution de $x = b' + A'x$.

On définit la **suite de Jacobi** du système $Ax = b$ par la relation suivante,

$$(6.1)\quad \begin{cases} x^{(0)} &= \omega \\ x^{(k+1)} &= b_J + B_J x^{(k)}, \quad k \geq 0, \end{cases} \qquad \text{(SCHÉMA DE JACOBI).}$$

Il est facile d'expliciter la relation donnant le $k+1$-ième élément $x^{(k+1)}$ de la suite de Jacobi. Notant

$$x^{(k+1)} = (x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_n^{(k+1)}),$$

nous avons

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left[b_i - \sum_{j \neq i} a_{ij} x_j^{(k)} \right], \quad i = 1, 2, \ldots, n.$$

E 113 Combien d'opérations sont-elles nécessaires pour calculer l'élément $x^{(k)}$?

6.2 Convergence

Nous dirons que la matrice $A \in M_n$ est à **diagonale dominante par lignes** si elle vérifie les inégalités

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|, \quad 1 \leq i \leq n.$$

Cela signifie que chaque terme de la diagonale est (beaucoup) plus grand que les autres éléments de la ligne sur laquelle il se trouve. De la même manière, nous dirons que A est à **diagonale dominante par colonnes** si

$$|a_{jj}| > \sum_{i \neq j} |a_{ij}|, \quad 1 \leq j \leq n.$$

Théorème 17. *Si A est à diagonale dominante – que ce soit par lignes ou par colonnes – alors, quel que soit le vecteur de départ ω, la suite de Jacobi $(x^{(k)})$ définie ci-dessus converge vers l'unique solution du système linéaire $Ax = b$.*

Démonstration. Comme nous l'avons déjà rappelé, en vertu du théorème 14, il suffit de mettre en évidence une norme matricielle V telle que $V(B_J) < 1$ puisque dans ce cas, quel que soit w, la suite définie par le schéma de Jacobi convergera vers l'unique solution du système $x = b_J + B_J x$ qui est aussi, d'après le théorème 16, l'unique solution de $Ax = b$. Écrivons explicitement la matrice B_J. Nous avons

$$\begin{aligned} B_J &= D^{-1}(E+F) \\ &= \begin{pmatrix} \frac{1}{a_{11}} & 0 & & 0 \\ 0 & \frac{1}{a_{22}} & & 0 \\ & \ddots & \ddots & \\ 0 & & 0 & \frac{1}{a_{nn}} \end{pmatrix} \cdot \begin{pmatrix} 0 & -a_{12} & \ldots & -a_{1n} \\ -a_{21} & 0 & & \vdots \\ \vdots & & & -a_{n-1,n} \\ -a_{n1} & \ldots & -a_{n,n-1} & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & \ldots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & & \vdots \\ \vdots & & & -\frac{a_{n-1,n}}{a_{n-1,n-1}} \\ -\frac{a_{n1}}{a_{nn}} & \ldots & -\frac{a_{n,n-1}}{a_{nn}} & 0 \end{pmatrix}. \end{aligned}$$

Prenons maintenant $V = |||\cdot|||_\infty$. Le Théorème 6 dit que

$$|||B_J|||_\infty = \max_{i=1,\dots,n} \sum_{j=1}^{n} |\text{coef}(i,j) \text{ de } B_j| = \max_{i=1,\dots,n} \sum_{j=1,\neq i}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| = \max_{i=1,\dots,n} \frac{1}{|a_{ii}|} \sum_{j=1,\neq i}^{n} |a_{ij}|.$$

Supposons que A soit à diagonale dominante par lignes. Cela signifie exactement que

$$\frac{1}{|a_{ii}|} \sum_{j=1, j\neq i}^{n} |a_{ij}| < 1.$$

Maintenant, si nous avons n nombres, tous < 1, le maximum de ces n nombres est aussi < 1. Nous avons donc $|||B_J|||_\infty < 1$ et le théorème 14 s'applique. Dans le cas où A est à diagonale dominante par colonnes, nous faisons le même raisonnement en faisant intervenir la norme $|||B_J|||_1$. ■

Exemple 2. Considérons le système $Ax = b$ avec

$$A = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix} et \quad b = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \text{dont la solution est} x = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$

La matrice A est à diagonale dominante $(1 > 1/2)$ et

$$B_J = (I - D^{-1}A) = \begin{pmatrix} 0 & -1/2 \\ -1/2 & 0 \end{pmatrix} \quad \text{et} \quad b_j = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

Prenons $w = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. La suite de Jacobi définie par

$$\begin{cases} x^{(0)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ x^{(k+1)} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 0 & -1/2 \\ -1/2 & 0 \end{pmatrix} x^{(k)}, \quad k \geq 0, \end{cases}$$

converge vers l'unique solution du système. On vérifie facilement que

$$x^{(k)} = \begin{pmatrix} 1/2^k \\ \frac{2^{k+1}-1}{2^k} \end{pmatrix} \longrightarrow \begin{pmatrix} 0 \\ 2 \end{pmatrix} = x$$

§ 7. EXERCICES ET PROBLÈMES

114 Rappelons qu'un réel λ est appelé valeur propre de $A \in M_n$, s'il existe un vecteur non nul x tel que $Ax = \lambda x$. On note $\mathscr{V}(A)$ l'ensemble des valeurs propres réelles de A et $\rho(A) = \max\{|\lambda| s \,:\, \lambda \in \mathscr{V}(A)\}$. Montrer que pour toute matrice $A \in M_n$ et tout norme matricielle induite V, on a $V(A) \geq \rho(A)$.

115 La norme de Frobenius. Nous étudions une nouvelle norme matricielle, assez souvent utilisée, qui a la propriété de ne pas être une norme induite pas une norme vectorielle mais d'être pourtant subordonnée à une norme vectorielle. Pour $A = (a_{ij}) \in M_n$, nous posons

$$|||A|||_f = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2}.$$

Théorème 18. *L'application $|||\cdot|||_f$ définie ci-dessus est une norme matricielle. De plus elle vérifie*

(7.1) $\|Ax\|_2 \leq |||A|||_f \, \|x\|_2, \quad x \in \mathbb{R}^n.$

Elle s'appelle la ***norme de Frobenius****.*

L'inégalité (7.1) signifie précisément que la norme de Frobenius est subordonnées à la norme euclidienne.

A) Montrer que $|||\cdot|||_f$ est bien une norme matricielle. Pour le troisième point, l'inégalité triangulaire, on pourra utiliser l'inégalité de Cauchy-Schwarz (Lemme 2) appliquée avec des vecteurs de $\mathbb{R}^{n^2}$ plutôt qu'avec des vecteurs de $\mathbb{R}^n$. La démonstration du dernier point, la sous-multiplicativité, utilisera aussi une forme de l'inégalité de Cauchy-Schwarz.

B) Démontrer que $|||\cdot|||_f$ vérifie l'inégalité (7.1).

C) Montrer la norme de Frobenius n'est pas une norme induite, voir l'exercice 109.

116 Soit A une matrice inversible de $M_n(\mathbb{R})$ et V une norme matricielle sur $M_n(\mathbb{R})$. On suppose connue une matrice B telle $V(I-AB) \leq r$ avec $0 < r < 1$ où I désigne la matrice identité. On définit la suite de matrice B_k par $B_0 = B$ et $B_{k+1} = B_k(2I - AB_k)$.

A) Montrer que $A^{-1} - B_{k+1} = A^{-1}(I-AB)^{2^{k+1}}$.

B) En déduire que B_k converge vers A^{-1}.

C) On appelle N_k le nombre de multiplications nécessaire pour calculer B_k. Donner une estimation de N_{k+1} en fonction de N_k. En déduire une estimation de N_k lorsque $k \to \infty$.

117 Soit $A \in M_n$ une matrice *diagonalisable* (avec des valeurs propres réelles). Montrer que la série géométrique $\sum_{k=0}^{\infty} A^k$ converge si et seulement si toutes les valeurs propres de A sont dans $]-1,1[$.

NOTE. — Il est connu qu'une condition nécessaire et suffisante pour que $\sum_{k=0}^{\infty} A^k$ converge est que toutes ses valeurs propres, réelles ou complexes, soit dans le disque ouvert de centre l'origine et de rayon 1 (sans qu'il soit nécessaire de supposer que A est diagonalisable).

118 Soient A, C et R trois matrices de M_n avec $R = AC - 1$. On suppose que $V(R) < 1$ où V est une norme matricielle. Montrer que

$$V(A^{-1}) \leq \frac{V(C)}{1-V(R)};$$

puis que

$$\frac{V(R)}{V(A} \leq V(C - A^{-1}) \leq \frac{V(C)\,V(R)}{1-V(R)}.$$

119 Montrer que le schéma de Jacobi appliqué au système (S) ci-dessous fournit une suite qui converge vers l'unique solution. Calculer le vecteur $x^{(1)}$ obtenu en partant de $x^{(0)} = (1/2, -2/5, -2/5)$:

$$(S) \qquad \begin{cases} 10x + 5y = 5 \\ x - 5y + 3z = 2 \\ 5x + y - 10z = 4 \end{cases}$$

120 On considère le système linéaire

$$S: \quad \begin{cases} 3x_1 + x_2 = 1 \\ x_1 + 3x_2 + x_3 = 0 \\ x_2 + 3x_3 = 2 \end{cases}$$

On note $x^{(k)}$ la **suite de Jacobi** correspondant au système (S) où on choisit $x^{(0)} = (1,0,2)$.

A) Montrer en utilisant un théorème du cours que la suite $x^{(k)}$ converge vers l'unique solution du système (S).

B) Expliciter la relation de récurrence liant $x^{(k+1)}$ et $x^{(k)}$. (On calculera le vecteur b_J et la matrice B_J.)

C) Montrer que $|||B_J|||_\infty = 2/3$ puis que, pour $k \geq 0$ on a

$$\|x - x^{(k)}\|_\infty \leq (2/3)^k \|x - x^{(0)}\|_\infty$$

où x est la solution exacte du système (S). (On utilisera que $x = b_J + B_J x$)

121 On étudie une modification de la méthode de Jacobi pour résoudre les systèmes linéaires. Soit $A=(a_{ij})$ une matrice réelle à n lignes et n colonnes. On considère le système $Ax=b$ où b est un vecteur donné de $\mathbb{R}^n$. On appelle D la matrice diagonale formée des éléments diagonaux de A c'est-à-dire $D=(d_{ij})$ avec $d_{ii}=a_{ii}$ et $d_{ij}=0$ pour $i\neq j$.

Si D est inversible, pour $w\in\mathbb{R}$, $w>0$ on pose

$$B_{J,w}=I-wD^{-1}A \quad \text{et} \quad b_{J,w}=wD^{-1}b.$$

On remarquera que lorsque $w=1$ on retrouve la matrice de Jacobi habituelle B_J.

A) Montrer que si D est inversible alors x est solution de $Ax=b$ si et seulement si x est solution de $x=b_{J,w}+B_{J,w}x$.

B) Montrer que $B_{J,w}=wB_J+(1-w)I$.

C) Donner les coefficients de $B_{J,w}$ en fonctions des coefficients de A et de w.

D) On définit la suite de vecteurs $x^{(k)}$ par

$$\begin{cases} x^{(0)} &= & u \\ x^{(k+1)} &= & b_{J,w}+B_{J,w}x^{(k)} \quad (k\geq 0) \end{cases}$$

E) Montrer que si $x^{(k)}=\left(x_1^{(k)},x_2^{(k)},\ldots,x_n^{(k)}\right)$ alors

$$x_j^{(k+1)}=x_j^{(k)}+\frac{w}{a_{jj}}\left[b_j-\sum_{l=1}^{n}a_{jl}x_l^{(k)}\right],$$

et en déduire le nombre d'opérations nécessaires pour obtenir le vecteur $x^{(k)}$.

F) A partir de maintenant, on suppose que $n=3$ et que B_J est diagonalisable c'est-à-dire qu'il existe une matrice inversible P telle que

$$B_J=P^{-1}\begin{pmatrix}\lambda_1 & 0 & 0\\ 0 & \lambda_2 & 0\\ 0 & 0 & \lambda_3\end{pmatrix}P.$$

(a) Montrer que

$$B_{J,w}=P^{-1}\begin{pmatrix}(1-w)+w\lambda_1 & 0 & 0\\ 0 & (1-w)+w\lambda_2 & 0\\ 0 & 0 & (1-w)+w\lambda_3\end{pmatrix}P.$$

(b) On suppose que $\lambda_1=1/2$, $\lambda_2=2/3$, $\lambda_3=3/4$. Montrer qu'il existe un valeur de w telle que la plus grande valeur absolue des $(1-w)+w\lambda_i$ $(i=1,2,3)$ soit strictement plus petite que $1/2$.

(c) On se place dans les hypothèses de la question précédente et on choisit w satisfaisant la condition indiquée. Expliquer pourquoi la suite $x^{(k)}$ converge plus vite vers la solution du système $Ax=b$ que la suite de Jacobi habituelle.

122 On considère un système (S) de n équations à n inconnues dont la forme matricielle est $(S):Ax=b$ avec $A\in M_n$ et $b\in\mathbb{R}^n$. On suppose que A est inversible de sorte que le système (S) possède une et une seule solution.

On souhaite étudier la résolution de (S) avec la méthode itérative — dite du gradient — définie par les relations suivantes :

$$(7.2)\quad \begin{cases} x^{(0)} &= & \omega \\ x^{(k+1)} &= & x^{(k)}-\theta(Ax^{(k)}-b) \\ &= & (I-\theta A)x^{(k)}+\theta b, \quad k\geq 0 \end{cases} \qquad \text{(SCHÉMA DU GRADIENT).}$$

Ici, θ est un paramètre réel *non nul* ($\theta\in\mathbb{R}^\star$) et I désigne la matrice identité.

But de l'exercice. On se propose de montrer que si A est diagonalisable avec des valeurs propres (strictement) positives alors on peut trouver une valeur de θ pour laquelle la suite $x^{(k)}$ ci-dessus converge vers l'unique solution du système (S).

Rappel. On rappelle que A est diagonalisable s'il existe une matrice D diagonale (c'est-à-dire avec tous ses coefficients nuls *sauf* sur la diagonale principale) et une matrice inversible U telle que $A = UDU^{-1}$. Les éléments sur la diagonale de D sont appelés *valeurs propres* de A.

A) Montrer que si la suite $(x^{(k)})$ converge alors elle converge vers l'unique solution du système (S).

B) On pose $M_\theta = (I - \theta A)$. Montrer par récurrence que

$$x^{(k+1)} = M_\theta^{k+1}\omega + \theta \sum_{i=0}^{k} M_\theta^i b \quad (k \geq 0). \tag{7.3}$$

(On rappelle que, pour toute matrice M, on convient $M^0 = I$.)

C) Dans cette partie on suppose que A est diagonalisable.

(a) Montrer que si $A = UDU^{-1}$ avec D diagonale alors $M_\theta = UD_\theta U^{-1}$ avec une matrice diagonale D_θ que l'on précisera.

(b) Montrer que pour tout $k \geq 0$, on a $M_\theta^k = UD_\theta^k U^{-1}$.

D) Dans cette partie on suppose que A est diagonalisable avec des valeurs propres strictement positives.

(a) Montrer qu'il existe $\theta > 0$ tel que toutes les valeurs propres de D_θ sont situées dans l'intervalle $]-1,1[$.

(b) Montrer, en utilisant le cours et l'équation (7.3), que pour la valeur de θ trouvée à la question précédente, la suite $x^{(k)}$ converge vers l'unique solution de (S).

123 On considère le système linéaire d'ordre 3 suivant :

$$\begin{cases} a^2x + ay + 2z = 3 \\ x + ay + z = 1 \\ ax + 2y + 2az = 1 \end{cases}, \tag{7.4}$$

où a est un paramètre réel.

A) Trouver une condition sur le paramètre a, la moins contraignante possible, pour que la suite de Jacobi associée au système (7.4) ci-dessus construite avec $x^{(0)} = w$ quelconque converge vers l'unique solution de ce système.

Cette condition est supposée satisfaite dans les questions suivantes.

B) On note $x^{(k)}$ la suite de Jacobi construite à partir de $x^{(0)} = (3,1,1)$. Expliciter la relation de récurrence entre $x^{(k+1)}$ et $x^{(k)}$.

C) On rappelle que si r désigne la solution de (7.4) alors

$$\|r - x^{(k)}\|_\infty \leq |||B_J|||_\infty^k \|r - x^{(0)}\|_\infty$$

où B_J désigne la matrice de Jacobi du système. On suppose que $x^{(0)}$ vérifie $\|r - x^{(0)}\|_\infty < 1$. Donnez une estimation, en fonction de a, du nombre d'itérations nécessaires pour obtenir une approximation de r avec une erreur inférieure à 10^{-4}.

124 La suite de Gauss-Seidel. Pour $A \in M_n$, on utilise les notations $A = D - E - F$ introduites au 6.1. Si aucun des éléments diagonaux a_{ii} de A n'est nul la matrice $D - E$ est inversible (son déterminant est le produit des éléments diagonaux qui sont tous différents de 0). On peut donc construire la **matrice de Gauss-Seidel** de A (et du système $Ax = b$) comme suit

$$B_{GS} = (D - E)^{-1}F,$$

et le **vecteur de Gauss-Seidel**

$$b_{GS} = (D - E)^{-1}b.$$

Le calcul de B_{GS} nécessite (apparemment) le calcul de l'inverse de $D - E$ mais il s'agit de l'inverse d'une matrice triangulaire qui est toujours facile à obtenir.

A) Démontrer le théorème suivant.

Théorème 19. *Si les éléments diagonaux de A sont non nuls, c'est-à-dire si D est inversible, alors x est solution de $Ax = b$ si et seulement si x est solution de $x = b_{GS} + B_{GS}x$.*

On définit la **suite de Gauss-Seidel** par la relation de récurrence suivante.

$$\begin{cases} x^{(0)} &= \omega \\ x^{(k+1)} &= b_{GS} + B_{GS}x^{(k)}, \quad k \geq 0, \end{cases} \qquad \text{(SCHÉMA DE GAUSS-SEIDEL).}$$

B) Montrer que le $k+1$-ième élément $x^{(k+1)}$ de la suite de Gauss-Seidel est donné par

$$x^{(k+1)} = (x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_n^{(k+1)})$$

avec

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right], \quad i = 1, 2, \ldots, n.$$

C) Combien d'opérations sont-elles nécessaires pour calculer l'élément $x^{(k)}$?

D) Il est connu, nous admettrons ce résultat que *si A est à diagonale dominante alors quelle que soit le vecteur de départ ω la suite de Gauss-Seidel $x^{(k)}$ définie ci-dessus converge vers l'unique solution du système linéaire $Ax = b$.*

E) Montrer que lorsqu'on considère le système $Ax = b$ avec

$$A = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix} \text{ et } \quad b = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

Alors pour $w = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ on a

$$x^{(k+1)} = \begin{pmatrix} 2^{1-2(k+1)} \\ \frac{2^{2(k+1)+1}-1}{2^{2(k+1)}} \end{pmatrix} \longrightarrow \begin{pmatrix} 0 \\ 2 \end{pmatrix} = x.$$

125 On considère le système linéaire d'ordre 3 suivant :

$$(7.5) \quad \begin{cases} -4x + my + z = 3 \\ mx + 6y + 2z = 1 \\ 3x - y + 2mz = 1 \end{cases},$$

où m est un paramètre réel.

A) Montrer que si $2 < |m| < 3$ alors la suite de Jacobi ainsi que la suite de Gauss-Seidel associées au système (7.5) ci-dessus construites en prenant comme premier terme w quelconque converge vers l'unique solution de ce système.

B) On note $x_J^{(k)}$ la suite de Jacobi construite à partir de $x_J^{(k)}(0) = (3,1,1)$ et $x_{GS}^{(k)}$ la suite de Gauss-Seidel construite à partir de $x_{GS}^{(k)}(0) = (3,1,1)$. Expliciter les relations de récurrence entre $x_J^{(k+1)}$ et $x_J^{(k)}$ et entre $x_{GS}^{(k+1)}$ et $x_{GS}^{(k)}$.

§ 8. NOTES ET COMMENTAIRES

Sur le contenu

Dans un premier temps, suivant les indications du programme auquel je croyais devoir me tenir, j'avais conçu ce chapitre pour en faire une introduction aux méthodes itératives de résolution des systèmes linéaires. J'ai fait ensuite, rapidement, le constat que l'objectif poursuivi ne repayait pas l'effort de définition et de formalisme demandé au public auquel je voulais m'adresser, à quoi s'est ajouté, presque aussi rapidement, le constat que, si les innombrables méthodes itératives jouent sans doute un rôle important en algèbre linéaire numérique, les plus élémentaires d'entre elles n'ont plus leur place dans un cours d'introduction. J'avais alors décidé de supprimer ce chapitre de mon cours. Je le réintroduit dans cette version en en modifiant l'esprit et l'objectif. L'outil essentiel est celui de norme et ce chapitre est une introduction à son utilisation. C'est le seul chapitre de ce cours qui se consacre à un outil technique, un outil de travail, plutôt qu'à un problème mathématique fondamental. J'ai maintenu l'étude de la méthode de Jacobi dont l'élégance et la simplicité feront peut-être oublier la faible utilité. Je me suis efforcé de rester aussi facilement lisible que dans les autres parties du cours, ce qui m'a d'ailleurs conduit à donner une démonstration du théorème sur l'équivalence des normes que certains pourront trouver assez tortueuse. Le théorème du rayon spectral, quoique fondamental, ne pouvait être inclus. Le cas particulier des matrices diagonalisables ets proposé en exercice. La question du conditionnement d'une matrice est étudiée en détail dans Higham (2002).

Sur les exercices

J'ai converti en exercices, sans généralement rentrer dans le détail des questions de convergences, les présentations de quelques une des méthodes classiques d'itération dont j'ai dit qu'elles n'avaient plus leur place dans le cours.

A

Le théorème de Rolle, des accroissements finis et la formule de Taylor

Les démonstrations de nombreux résultats d'analyse numérique sont basées sur la **formule de Taylor**. Nous donnons la démonstration qui suffit au niveau de ce cours. Le théorème qui donne naissance à la formule de Taylor est le **théorème de Rolle**.

Théorème 1. *Soit f une fonction continue sur l'intervalle fermé $[\alpha,\beta]$ et dérivable sur l'intervalle ouvert $]\alpha,\beta[$. Si $f(\alpha)=f(\beta)$ alors il existe $c\in]\alpha,\beta[$ tel que $f'(c)=0$.*

La démonstration de ce théorème à son tour requiert l'emploi d'un autre théorème fondamental de l'analyse : toute fonction continue f sur un intervalle fermé borné $[\alpha,\beta]$ atteint ses bornes (sur $[\alpha,\beta]$). La démonstration de ce résultat, qui s'appuie sur les propriétés fondamentales de l'ensemble des réels, se trouve dans les traités d'introduction à l'analyse.

Démonstration. Notons $M:=\max\{f(x)\ :\ x\in[\alpha,\beta]\}$ et $m:=\min\{f(x)\ :\ x\in[\alpha,\beta]\}$. Si $M=m$ alors la fonction f est constante, sa dérivée nulle et n'importe quel c convient. Dans la cas contraire nous avons $M>f(\alpha)$ ou bien $m<f(\alpha)$. Nous supposerons que la première alternative se réalise. La démonstration est similaire dans le second cas. D'après le rappel ci-dessus, la valeur M est atteinte en un point au moins, disons, au point x_0, et à cause de l'hypothèse que $M>f(\alpha)$, le point $x-0$ est distinct aussi bien de α que de β. Nous allons établir que $f'(x_0)=0$ ce qui achèvera la démonstration du théorème. Pour cela nous observons le taux d'accroissement et remarquons que

$$\frac{f(x_0+h)-f(x_0)}{h}>0,\quad h<0\quad\text{et}\quad\frac{f(x_0+h)-f(x_0)}{h}<0,\quad h<0.$$

En prenant la limite lorsque h tend vers 0 par valeurs négatives dans la première inégalité, nous obtenons que $f'(x_0)\leq 0$ tandis qu' en prenant la limite lorsque h tend vers 0 par valeurs positives dans la première inégalité, nous obtenons que $f'(x_0)\geq 0$. Nous en tirons $f'(x_0)=0$. ■

Théorème 2 (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*

$$f(v)=f(u_0)+f'(u_0)(v-u_0)+\cdots+\frac{f^{(d)}(u_0)}{d!}(v-u_0)^d+\frac{f^{(d+1)}(\xi)}{d!}(v-u_0)^{d+1}. \tag{0.1}$$

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 $\mathscr{C}^{d+1}$ sur un intervalle contenant α et β de sorte que les conditions du théorème seront largement satisfaites.

Le cas particulier $d=0$ s'appelle le théorème (ou formule) des accroissements finis et s'écrit en remplaçant u_0 par u,

$$f(v)-f(u)=f'(\xi)(v-u).$$

Démonstration. La formule de Taylor est une conséquence du théorème de Rolle. De manière précise, elle s'obtient en appliquant d fois le théorème de rôle comme suit. Nous considérons la fonction

$$g(t) = f(v) - T^d(f,t) - K(t-u_0)^{d+1},$$

en choisissant la constante K en sorte que $g(v) = 0$. À cause de la forme particulière du polynôme de Taylor, nous avons

$$0 = g(u_0) = g'(u_0) = \cdots = g^{(d)}(u_0).$$

Puisque $g(v) = g(u_0) = 0$, le théorème de Rolle nous assure de l'existence de c_1 entre u_0 et v tel que $g'(c_1) = 0$. Maintenant, puisque $g'(c_1) = g'(u_0) = 0$, le théorème de Rolle nous assure de l'existence de c_2 entre c_1 et u_0 tel que $g''(c_2) = 0$. Nous continuons et après $d+1$ applications du théorème de Rolle, nous arrivons à l'existence de c_{d+1} tel que $g^{(d+1)}(c_{d+1}) = 0$ mais, en revenant à l'expression de g nous voyons de

$$g^{(d+1)}(c_{d+1}) = f^{(d+1)}(c_{d+1}) - (d+1)! \times K.$$

Cette formule pour K donne la relation cherchée. ■

B

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 $f(x)=\sqrt{x}\Longrightarrow f'(x)=\frac{1}{2}x^{-1/2}\Longrightarrow f''(x)=\frac{1}{2}\frac{-1}{2}x^{-3/2}\Longrightarrow f'''(x)=\frac{1}{2}\frac{-1}{2}\frac{-3}{2}x^{-5/2}$. Il suit, en utilisant que $|f'''|$ est décroissante sur $[100,144]$ que

$$\left|f'''(\zeta)\right|\leq\frac{3}{2^3}100^{-5/2}=\frac{3}{8\cdot10^5}.$$

En reportant dans l'inégalité précédente on arrive à

$$\left|\sqrt{115}-P(115)\right|\leq\frac{3}{3!\cdot8\cdot10^5}15\cdot6\cdot29=\frac{15\cdot6\cdot29}{16\cdot10^5}\approx1,63\cdot10^{-3}<1,8\cdot10^{-3}.$$

2 (← 31)

(a) Une application de la formule d'interpolation de Lagrange donne

$$\begin{array}{rl} \alpha=\mathbf{L}[0,1/6,1/4;f](1/5)= & f(0)\dfrac{(1/5-1/6)(1/5-1/4)}{(-1/6)(-1/4)} \\ & +f(1/6)\dfrac{(1/5-0)(1/5-1/4)}{(1/6-0)(1/6-1/4)}+f(1/4)\dfrac{(1/5-0)(1/5-1/6)}{(1/4-0)(1/4-1/6)}. \end{array}$$

Après simplification,

$$\alpha=-24/600+\sqrt{3}\times72/200+\sqrt{2}\times48/300=0,8098...$$

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

$$|\cos(\pi/5)-\alpha|\leq\pi^3\sqrt{2}/(2\cdot3!)\times(1/5)(1/20)(1/30)\approx1,22\cdot10^{-3}.$$

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 (← 23). 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 génerelament 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

$$\begin{aligned} |f(x) - \mathbf{L}[a_0,a_1,a_2,a_3,a_4;f](x)| &\leq \frac{\sup_{x\in[a,b]} |f^{(5)}|}{5!} \cdot |w_A(x)| \\ &\leq \frac{\sup_{x\in[a,b]} |f^{(5)}|}{5!} \cdot 5! h^5 = \sup_{x\in[a,b]} |f^{(5)}| \cdot h^5. \end{aligned}$$

E) Soit $i \in \{0,\ldots,3\}$. 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 $A = \{a_0,\ldots,a_n\} \subset [a,b]$ avec $a_i < a_{i+1}$ pour $i = 0,\ldots,n-1$. 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

$$|x-a_i| \leq |x-a_0| + |a_1-a_0| + \cdots + |a_{i-1}-a_i| \leq \sum_{j=0}^{i} h_i \leq (i+1)h, \quad i = 1,\ldots,n.$$

On en déduit $|w_A(x)| \leq h(2h)(3h)\ldots((n+1)h) = (n+1)!h^{n+1}$. En examinant les autres intervalles on se rend compte que cette majoration est la plus grossière et $|w_A(x)| \leq (n+1)!h^{n+1}$ sur $[a,b]$ d'où encore pour toute fonction de classe C^{n+1},

$$|f(x) - \mathbf{L}[a_0, a_1, \ldots, a_n; f](x)| \leq \sup_{x\in[a,b]} |f^{(n+1)}| \cdot h^{n+1}.$$

Les autres questions se généralisent suivant les mêmes lignes.

§ 2. CALCUL APPROCHÉ DES INTÉGRALES

7 (← 56.)

(a) L'inégalité à démontrer est

$$\frac{1}{ta+(1-t)b} \leq \frac{t}{a} + \frac{1-t}{b}.$$

En réduisant au même dénominateur, on montre que cette inégalité est équivalente à

$$\begin{aligned}
&\iff \frac{1}{ta+(1-t)b} \leq \frac{t}{a} + \frac{1-t}{b} \\
&\iff ab \leq (tb+(1-t)a)(ta+(1-t)b) \\
&\iff 0 \leq [t^2+(1-t)^2-1]ab + t(1-t)(b^2+a^2)
\end{aligned}$$

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 $a \leq x \leq b$ on a $b-x \geq 0 \Rightarrow \frac{b-x}{b-a} \geq 0$. D'autre part,

$$x \geq a \Rightarrow -x \leq -a \Rightarrow b-x \leq b-a \Rightarrow \frac{b-x}{b-a} \leq 1.$$

Prenant $t = \frac{b-x}{b-a}$ dans (5.2) on obtient en tenant compte que $(1-t) = 1 - \frac{b-x}{b-a} = \frac{x-a}{b-a}$,

$$f\left(a.\frac{b-x}{b-a} + b.\frac{x-a}{b-a}\right) \leq f(a)\frac{b-x}{b-a} + f(b)\frac{x-a}{b-a}.$$

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

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

(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 (← 58)

(a) D'après le cours, on a

$$|I-A(n,f)|\leq\frac{1}{2880n^4}\sup_{[0,1]}\left|f^{(4)}\right|.$$

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

$$A(n,f)-A_n(\tilde f)=$$
$$\tfrac{h_n}{6}\left[(f-\tilde f)(a)+(f-\tilde f)(b)+2\sum_{i=1}^{n-1}(f-\tilde f)(a+ih_n)+4\sum_{i=0}^{n-1}(f-\tilde f)(a+\tfrac{2i+1}{2}h_n)\right]$$

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

$$\left|I-A(\nu,\tilde f)\right|\leq|I-A(\nu,f)|+\left|A(\nu,f)-A(\nu,\tilde f)\right|\leq 10^{-3}+\varepsilon.$$

(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 (← 59.)

(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

$$\begin{aligned}
&\left|\textstyle\int_{x_i}^{x_{i+1}} f(x)dx - Q_i(f)\right| \\
&= \left|\frac{1}{2}\left\{\int_{x_i}^{x_{i+1}} f(x) - \mathbf{L}[x_{i-1},x_i,x_{i+1};f](x)dx\right\} + \frac{1}{2}\left\{\int_{x_i}^{x_{i+1}} f(x) - \mathbf{L}[x_i,x_{i+1},x_{i+2};f](x)dx\right\}\right| \\
&\leq \frac{1}{2}\int_{x_i}^{x_{i+1}} |f(x) - \mathbf{L}[x_{i-1},x_i,x_{i+1};f](x)|\,dx + \frac{1}{2}\int_{x_i}^{x_{i+1}} |f(x) - \mathbf{L}[x_i,x_{i+1},x_{i+2};f](x)|\,dx \\
&\leq \frac{1}{2}\sup_{[x_{i-1},x_{i+1}]}|f^{(3)}|\int_{x_i}^{x_{i+1}}|x-x_{i-1}||x-x_i|dx + \frac{1}{2}\sup_{[x_i,x_{i+2}]}|f^{(3)}|\int_{x_i}^{x_{i+1}}|x-x_i||x-x_{i+1}|dx \\
&\quad \leq C_i \cdot \sup_{[x_{i-1},x_{i+2}]}|f^{(3)}|
\end{aligned}$$

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 (← 78.)

(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 (← 77.)

(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 $\bar{x}_0 = 0,75$. La suite de newton est donnée par la formule

$$\bar{x}_{n+1} = \bar{x}_n - \frac{f(\bar{x}_n)}{f'(\bar{x}_n)} = \bar{x}_n - \frac{\bar{x}_n^4 + 2\bar{x}_n - 1}{4\bar{x}_n^3 + 4\bar{x}_n}.$$

On trouve $\bar{x}_1 = 0,655...$ et $\bar{x}_2 = 0,6437....$

(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 $\underline{x}_0 \le \underline{x}_1 \le \underline{x}_2 \le \underline{x}_3 \le r \le \bar{x}_2 \le \bar{x}_1 \le \bar{x}_0$. En particulier $0,6409... \le r \le 0,647...$ de sorte que l'approximation de r avec deux décimales exactes est $0,64$.

12 (← 86)

(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

$$(3.1) \quad 1 \le x \le 2 \implies 4 \le 3x+1 \le 7$$
$$\implies 1/7 \le (3x+1)^{-1} \le 1/4 \implies (1/7)^{2/3} \le f'(x) \le (1/4)^{2/3}.$$

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 (← 99)

(a) Le système se résoud immédiatement par substitutions successives (de "bas en haut") comme suit

$$(4.1)\quad \begin{cases} x_1 = \frac{c_1 - a_{13}x_3 - a_{12}x_2}{a_{11}} \\ x_2 = \frac{c_2 - a_{24}x_4 - a_{23}x_3}{a_{22}} \\ \dots \\ x_i = \frac{c_i - a_{i,i+2}x_{i+2} - a_{i,i+1}x_{i+1}}{a_{ii}} \\ \dots \\ x_{n-2} = \frac{c_{n-2} - a_{n-2,n-1}x_{n-1} - a_{n-2,n}x_n}{a_{n-2,n-2}} \\ x_{n-1} = \frac{c_{n-1} - a_{n-1,n}x_n}{a_{n-1,n-1}} \\ x_n = \frac{c_n}{a_{nn}} \end{cases}$$

(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 (← 100.)

(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

$$\begin{aligned} L_1 A^{(0)} &= \begin{pmatrix} 1 & 0 & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & 0 \\ -\frac{a_{31}}{a_{11}} & 0 & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \\ &= \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ (\frac{-a_{21}}{a_{11}} a_{11} + a_{21}) & (\frac{-a_{21}}{a_{11}} a_{12} + a_{22}) & (\frac{-a_{21}}{a_{11}} a_{13} + a_{23} \\ (\frac{-a_{31}}{a_{11}} a_{11} + a_{31}) & (\frac{-a_{31}}{a_{11}} a_{12} + a_{32}) & (\frac{-a_{31}}{a_{11}} a_{13} + a_{33}) \end{pmatrix} = A^{(1)} \end{aligned}$$

Ensuite

$$\begin{aligned} L_2 A^{(1)} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{a_{32}^{(1)}}{a_{22}^{(1)}} & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22}^{(1)} & (a_{23}^{(1)} \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} \end{pmatrix} \\ &= \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} \\ 0 & (\frac{-a_{32}^{(1)}}{a_{22}^{(1)}} a_{22}^{(1)} + a_{32}^{(1)}) & (\frac{-a_{32}^{(1)}}{a_{22}^{(1)}} a_{23}^{(1)} + a_{33}^{(1)}) \end{pmatrix} = A^{(2)} \end{aligned}$$

(c) On déduit de la question précédente que $A^{(2)} = L_2 A^{(1)} = L_2 L_1 A^{(0)}$. Il suffit alors de prendre $R = A^{(2)}$ car $A^{(2)}$ est triangulaire supérieure et $L = (L_2 L_1)^{-1}$ 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

$$L_i = \begin{pmatrix} Id_{i-1} & 0 \\ 0 & \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \frac{a_{i+1,i}^{(i-1)}}{a_{i,i}^{(i-1)}} & 1 & 0 & \cdots & 0 \\ \frac{a_{i+2,i}^{(i-1)}}{a_{i,i}^{(i-1)}} & 0 & 1 & & \\ \vdots & \vdots & & \ddots & 0 \\ \frac{a_{n,i}^{(i-1)}}{a_{i,i}^{(i-1)}} & 0 & \cdots & & 1 \end{pmatrix} \end{pmatrix}$$

où Id_{i-1} désigne la matrice Identité de dimension $i-1$. Cette matrice est supprimée dans le cas $i = 1$.

15 (← 101.)

(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 $i=2,\dots,n$ et $\alpha_i=a_i-\beta_i c_{i-1}\Rightarrow a_i=\alpha_i+\beta_i c_{i-1}$ pour $i=2,\dots,n$ 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)

$$Ly=b\Leftrightarrow\begin{cases} x_1=d_1 \\ \beta_2x_1+x_2=d_2 \\ \vdots \\ \beta_nx_{n-1}+x_n=d_n \end{cases}\Leftrightarrow\begin{cases} x_1=d_1 \\ x_i=d_i-\beta_ix_{i-1} \quad i=2,\dots,n. \end{cases}$$

Le nombre N d'opérations est donné par $N=(n-1).(1(-)+1(\times))=2(n-1)$.

(e)

$$Ux=y\Leftrightarrow\begin{cases} \alpha_1x_1+c_1x_2=y_1 \\ \alpha_2x_2+c_2x_3=y_2 \\ \vdots \\ \alpha_{n-1}x_{n-1}+c_{n-1}x_n=y_{n-1} \\ \alpha_nx_n=y_n \end{cases}\Leftrightarrow\begin{cases} x_n=\frac{y_n}{\alpha_n} \\ x_{n-i}=\frac{(y_{n-i}-c_{n-i}x_{n-i+1})}{\alpha_{n-i}} \quad 1\le i\le n-1. \end{cases}$$

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 (← 104.)

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) $Ax = b \iff (LT(L))(x) = b \iff L(T(L)x) = b \iff Ly = b$ 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 (G) 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

Berrut, J.-P. and Trefethen, L. N. (2004), 'Barycentric Lagrange interpolation', *SIAM Rev.* **46**(3), 501–517 (electronic).
URL: *http ://dx.doi.org/10.1137/S0036144502417715* 33

Criscuolo, G., Mastroianni, G. and Occorsio, D. (1990), 'Convergence of extended Lagrange interpolation', *Math. Comp.* **55**(191), 197–212.
URL: *http ://dx.doi.org/10.2307/2008800* 33

Crouzeix, M. and Mignot, A. L. (1984), *Analyse numérique des équations différentielles*, Masson, Paris.

Davis, P. J. (1975), *Interpolation and approximation*, Dover Publications Inc., New York. Republication, with minor corrections, of the 1963 original, with a new preface and bibliography. 49

Démidovitch, B. and Maron, I. R. (1979), *Eléments de calcul numérique*, Mir, Moscou. Traduit du russe. 33, 49

Dieudonné, J. (1968), *Calcul infinitésimal*, Hermann, Paris. 77

Hardy, G. H. (1952), *A course of pure mathematics*, Cambridge University Press, Cambridge. Dixième édition (première édition, 1908). 49

Higham, N. J. (2002), *Accuracy and stability of numerical algorithms*, second edn, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.
URL: *http ://dx.doi.org/10.1137/1.9780898718027* 120

Isaacson, E. and Keller, H. B. (1994), *Analysis of numerical methods*, Dover Publications Inc., New York. Corrected reprint of the 1966 original [Wiley, New York ; MR0201039 (34 #924)]. 77

Krylov, V. I. (1962), *Approximate calculation of integrals*, Translated by Arthur H. Stroud, The Macmillan Co., New York. Reprinted by Dover. 49

Paterson, A. (1991), *Differential equations and numerical analysis*, Cambridge university press, Cambridge. 49

Quarteroni, A., Sacco, R. and Salai, F. (1998), *Matematica Numerica*, Springer-Verlag, Milano.

Ralston, A. and Rabinowitz, P. (2001), *A first course in numerical analysis*, second edn, Dover Publications Inc., Mineola, NY. 33, 77

Sibony, M. and R., J. C. M. (1982), *Analyse numérique (2 tomes)*, Hermann, Paris.