Agrégation Externe de Mathématiques

Analyse Numérique

Franck Boyer

e-mail : franck.boyer@univ-amu.fr

Aix-Marseille Université

14 octobre 2013

Table des matières

Avant-Propos

Ce document a pour vocation de rassembler les éléments d'analyse numérique qui peuvent être utiles pour préparer les oraux de l'agrégation externe de Mathématiques. Les étudiants préparant l'option B sont les premiers concernés mais pas seulement. Un certain nombre de chapitres font partie intégrante du programme général du concours et font l'objet de leçons. Par ailleurs, les éléments contenus dans ces notes fournissent de nombreux exemples et potentiels développements pour enrichir vos leçons (d'analyse ET d'algèbre).

Le premier chapitre est purement indicatif pour vous sensibiliser aux problèmes naturellement liés à toute méthode de calcul sur un ordinateur en précision finie (arithmétique flottante).

Le chapitre 2 traite des méthodes de résolution d'équations non-linéaires du type $f(x) = 0$.

Les chapitres 3 et 4 traitent des problèmes de l'interpolation et de l'approximation polynomiale et de l'intégration numérique.

Ces notes seront modifiées et mises à jour durant l'année. N'hésitez surtout pas à me transmettre des commentaires, des questions ou la liste de toutes les coquilles, erreurs qui émaillent certainement ce texte.

Chapitre I

Arithmétique de l'ordinateur. Erreurs numériques. Stabilité

1 L'ordinateur et les nombres

1.1 Les entiers

On commence par l'écriture des entiers en base b.

On se donne un entier $b > 0$ (la base). On rappelle que la division euclidienne consiste à écrire, de façon unique, tout entier $a \geq 0$ sous la forme

$$a = bq + r, \quad \text{avec } \ q \geq 0 \text{ et } 0 \leq r < b.$$

On déduit de cela que tout entier $d \neq 0$ s'écrit de façon unique sous la forme

$$d = d_n b^n + \cdots + d_0, \ \text{ avec } 0 \leq d_i \leq b-1, \text{ et } d_n > 0.$$

Le nombre $n+1$ est le nombre de chiffres de l'écriture de d dans la base b. En pratique on écrit

$$d = \overline{d_n d_{n-1} \cdots d_1 d_0}, \ \text{ ou même } \ d = d_n d_{n-1} \cdots d_1 d_0, \ \text{ s'il n'y a pas de confusion possible.}$$

Exemple : le nombre 25 s'écrit de la façon suivante en base 2 : 11001.

On peut alors effectuer les opérations usuelles de la même façon que pour l'écriture en base 10.

Exemples :

$$25 + 13 = 38 \Longrightarrow 11001 + 01101 = 100110,$$

$$25 * 13 = 325 \Longrightarrow 11001 * 01101 = 101000101.$$

Les ordinateurs travaillent naturellement en base 2 (en binaire) et les nombres entiers sont stockés sur un certain nombre de chiffres maximum (on parle de bits). Un exemple usuel est le stockage des entiers sur 4 octets = 32 bits.

Ceci permet de représenter les entiers non-signés de 0 à $2^{32} - 1$ et les entiers signés de -2^{31} à $2^{31} - 1$ (le premier bit représentant alors le signe).

1.2 Les réels

Dans un ordinateur, un réel est stocké sous la forme d'un entier et d'un nombre de rangs de décalage de la virugle nécessaire pour obtenir le nombre souhaité.

Virgule fixe. On fixe par avance le décalage de la virgule souhaité qui sera le même pour tous les nombres.

Ceci n'est pas souhaitable car la plage de nombres représentables n'est pas très adaptée au calcul numérique. Donc : pas de très petits nombres ou de très grands.

D'autre part, on perd de la précision

Exemple : en base 10 avec un décalage de 4

$$\begin{aligned} 1234,5678 \ &\longrightarrow 12345678, \\ 1,2345678 \ &\longrightarrow 00012345 \quad \Leftarrow \text{perte de précision.} \end{aligned}$$

Dans ces conditions, le plus grand nombre représentable est : 9999.9999 et le plus petit nombre strictement positif représentable est 0.0001.

Virgule flottante. Idée : on ne fixe pas le décalage de virgule par avance. Chaque nombre réel sera donc codé par :
- Un signe $s \in \{-1, 1\}$ qui nécessite un seul symbole.
- Un nombre entier m à n chiffres (n fixé) qui représente les chiffres significatifs du nombre. On l'appelle **la mantisse**. On impose également que le premier chiffre de la mantisse soit non nul (sinon on peut tout décaler d'un cran au moins).
- Un entier relatif e compris entre e_{min} et e_{max} qui code le décalage de la virgule. On l'appelle l'exposant. En général il est codé en t chiffres, t fixé.

Le nombre total de symboles utilisé pour coder un tel nombre est donné par

$$N = 1 + n + t,$$

et le nombre réel ainsi représenté est

$$x = s \times m \times p^{e-n}. \tag{I.1}$$

Pourquoi y-a-t'il $e - n$ en exposant et non pas seulement e ? Cela vient du fait que l'on a choisi des mantisses entières. Si on note $m = \overline{m_1 \cdots m_n}$ l'écriture en base p de m, on peut encore écrire le nombre x sous la forme

$$x = s \times \overline{0, m_1 \cdots m_n} \times p^e. \tag{I.2}$$

Cette dernière écriture montre que l'exposant e caractérise bien l'ordre de grandeur du nombre x. Attention, certaines références appellent *mantisse* le nombre fractionnaire $\overline{0, m_1 \cdots m_n}$.

Reprenons un exemple avec un codage en 10 symboles en base 10. On garde 1 symbole pour le signe, 7 symboles pour les chiffres significatifs du nombre, 1 symbole pour le signe de l'exposant et 1 symbole pour la valeur absolue de l'exposant. Dans ces conditions le plus grand nombre représentable est $+9999999\,10^{9-7} = 999999900$ et le plus petit nombre positif représentable 10^{-9}.

Remarques
- La répartition des réels exactement représentables n'est absolument pas uniforme sur la droite réelle !
- La question de savoir si un réel x est exactement représentable ou pas dépend de la base ! ! Exemple : le nombre réel (en base 10) $0.2 = \frac{1}{5}$ s'écrit en base 2 :

$$0.0001100011\ldots$$

- Prenons un exemple : en base 2, $n = 3$, $e_{min} = -1$, $e_{max} = 3$. L'ensemble des réels positifs représentables dans cet exemple est donné par

$$0, 0.25, 0.3125, 0.3750, 0.4375, 0.5, 0.625, 0.750, 0.875, 1.0, 1.25, 1.50, 1.75,$$
$$2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 6.0, 7.0$$

Les flottants double précision usuels : Dans les langages de programmation modernes les réels dits *double précision* sont codés, en base $p = 2$, sur 64 bits de la façon suivante
- Le signe du nombre est codé sur 1 bit.
- La mantisse est codée sur 53 bits : $2^{52} \leq m \leq 2^{53} - 1$. **MAIS** comme on impose le premier chiffre de m d'être non nul, ce chifre est forcément égal à 1 et donc il est inutile de le conserver en mémoire. C'est pourquoi une telle mantisse occupe en réalité 52 bits en mémoire.
- l'exposant (signé) est codé sur 11 bits : soit $e_{min} = -2^{10} + 2 = -1022 \leq e \leq 2^{10} - 1 = 1023 = e_{max}$.

Le plus grand réel positif représentable est donc à peu près $F_{max} = 1.798\,10^{308}$ et le plus petit $F_{min} = 2.225\,10^{-308}$.

Le *epsilon-machine*, c'est-à-dire la distance entre le flottant 1.0000 et le flottant immédiatement supérieur est donc donné par (attention $n = 53$ pour les double précision même s'il ne faut que 52 bits pour représenter la mantisse)

$$\varepsilon = p^{1-n} = 2^{-52} = 2.220.10^{-16}.$$

Les flottants simple précision : Ils sont moins utilisés en calcul scientifique car les erreurs d'arrondis sont beaucoup plus importantes. Ceci dit, ils permettent de gagner un facteur 2 en stockage mémoire et d'accélérer également les calculs. Ce codage sur 32 bits est effectué de la façon suivante :
- Signe codé sur 1 bit.
- Mantisse codée sur 24 bits mais ici aussi l'un des bits est inutile donc le codage est en réalité effectué sur 23 bits.
- l'exposant signé est codé sur 8 bits, soit $e_{min} = -126 \leq e \leq 127 = e_{max}$.

On trouve

$$F_{min} = 1.175.10^{-38}, \quad F_{max} = 3.403.10^{38},$$
$$\varepsilon = 2^{-23} = 1.1920.10^{-7}.$$

1.3 De l'importance de la précision dans les calculs

Deux exemples célèbres :

- **Le premier vol d'Ariane 5 :** L'échec du premier vol d'Ariane 5 est du à une stupide erreur de conversion de nombres. Ainsi l'ordinateur de bord de la fusée calculait la vitesse horizontale de l'engin en codant le résultat comme un réel double précision.
  A un moment de l'algorithme de l'ordinateur de bord, cette valeur devait être arrondie à un nombre entier et stockée dans une variable entière sur 16 bits. Cette partie du programme datait d'Ariane 4 et ne posait aucun problème jusqu'a présent. Sauf que .... Ariane 5 est beaucoup plus puissante et rapide qu'Ariane 4. La valeur de la vitesse horizontale dépassait donc le nombre entier maximum représentable sur 16 bits, ce qui a fait planter l'ordinateur de bord et provoqué la déviation de la trajectoire du lanceur et donc l'ordre de destruction donnée par la base de commande au sol.
  Ironie du sort : le calcul de cette valeur entière arrondie ne servait essentiellement à rien et aurait pu être supprimée du logiciel de bord ...
- **Echec du missile anti-skud patriot en 1991 :** Bilan de l'opération = 28 morts et 100 blessés.
  Le processeur interne du missile comptabilisait le temps en nombre entier de dixièmes de secondes. Pour convertir ce nombre en un nombre de secondes cet entier était ensuite multiplié par le nombre $1/10$ ou plus exactement par une approximation sur 24 bits de celui-ci : 209715×2^{-21} qui approche $1/10$ à 10^{-7} près. Le processeur ayant été démarré une centaine d'heures avant le tir, l'horloge interne du missile était donc décalée de

$$\underbrace{100 \times 3600 \times 10}_{\text{nb de dixièmes de seconde dans 100h}} \times 9.53 \times 10^{-8} = 0.34\text{s}.$$

  La vitesse du missile skud est environ de 1676m.s^{-1}, ainsi le décalage de l'horloge interne du missile intercepteur était de

$$1676 * 0.34 \sim 575\text{m}.$$

  C'est pourquoi l'interception a échoué et le missile a frappé des forces alliées.
  Si la représentation flottante du nombre $1/10$ avait été l'arrondi au plus proche sur 24 bits, c'est-à-dire le nombre 13421773×2^{-27}, l'erreur finale commise aurait été inférieure à 10m et le missile aurait probablement été détruit quand même.

1.4 Opérations sur les flottants

Arrondi Travailler sur les nombres flottants nécessite la définition d'une notion d'arrondi. On note F l'ensemble de tous les flottants représentables dans le système choisi. On note F_{max} le plus grand flottant positif représentable et F_{min} le plus petit.

Définition et proposition I.1.1

On appelle opération d'arrondi l'application notée $fl : [-F_{max}, -F_{min}] \cup [F_{min}, F_{max}] \mapsto F$ qui à tout réel associe un nombre flottant de la façon suivante : $fl(x)$ est le flottant le plus proche de x. S'il y a deux flottants à égale distance de x, on décide de trancher en choisissant celui des deux dont la valeur absolue est la plus grande.
On vérifie les propriétés suivantes :

- *Propriété de projection : Si $x \in F$, alors $fl(x) = x$.*
- *Monotonie : Si $x \leq y$, $fl(x) \leq fl(y)$.*
- *Pour tout $x \in [-F_{max}, -F_{min}] \cup [F_{min}, F_{max}]$ on a*

$$\frac{|fl(x) - x|}{|x|} \leq \frac{\varepsilon}{2},$$

la quantité $\frac{\varepsilon}{2}$ est souvent notée u et appelée, l'unité d'arrondi.

Les réels plus grands que F_{max} en valeur absolue ne peuvent pas être arrondis et représentés, on est dans une situation d'*overflow*. Les réels plus petits que F_{min} en valeur absolue sont dans le même cas, on dit qu'on est en situation d'*underflow*.

Preuve :

Les propriétés de projection et de monotonie sont triviales, montrons la propriété d'erreur relative de l'arrondi. On suppose $x > 0$ et que x n'est pas un flottant (sinon $fl(x) = x$ et on a gagné). On écrit

$$x = \mu p^{e-n},$$

où μ est un réel vérifiant $p^{n-1} < \mu < p^n - 1$. On voit que x est compris entre les deux flottants consécutifs

$$\underline{x} = E(\mu)p^{e-n} < x < (E(\mu)+1)p^{e-n} = \overline{x},$$

où $E(\mu)$ désigne la partie entière de μ. Donc $fl(x)$ est celui des deux nombres $\underline{x}$, $\overline{x}$ qui est le plus proche de x. On a

$$|fl(x) - x| \leq \frac{|\overline{x} - \underline{x}|}{2} = \frac{p^{e-n}}{2}.$$

On trouve donc

$$\frac{|fl(x) - x|}{|x|} \leq \frac{p^{e-n}}{2\mu p^{e-n}} = \frac{1}{2\mu} \leq \frac{p^{1-n}}{2} = \frac{\varepsilon}{2}.$$

■

Opérations élémentaires Comment se passent les opérations dans l'ordinateur :

- Addition (la soustraction est similaire) :
  - On repère l'exposant le plus grand des deux nombres.
  - On décale la mantisse du flottant le plus petit pour se ramener au même exposant que le grand nombre.
  - On effecture l'addition des deux mantisses (addition d'entiers).
  - On décale à nouveau la mantisse et on corrige l'exposant pour obtenir un flottant normalisé.

  Si on ne rajoute pas des bits intermédiaires pour faire le calcul, le résultat peut être très mauvais. Exemple : soit à soustraire (base binaire $p = 2$, $n = 3$) $x = 0.100 \times 2^1$ (soit $x = 1$ en décimal) et $y = 0.111 \times 2^0$ (soit $y = 0.875$ en décimal).

  Si on ne prend pas garde, on décale les bits de y d'un cran vers la droite et on élimine le dernier bit (qui ne "rentre plus dans les cases") on obtient donc l'opération

  $$x - y \sim 0.100 \times 2^1 - 0.011 \times 2^1 = 0.001 \times 2^1 = 0.1 \times 2^{-1}, \text{ soit } 0.25 \text{ en décimal},$$

  le résultat exact étant égal à 0.125 (en décimal), on voit que l'on a commis une erreur de 100% sur le résultat.

  Pour régler, en partie, ce problème les ordinateurs effectuent le calcul en rajoutant un chiffre significatif dans l'étape intermédiaire.
- Multiplication (la division est similaire) :
  - On additionne les exposants.
  - On effecture la multiplication (entière) des mantisses.
  - On tronque le résultat pour ne garder que les n premiers chiffres significatifs.
  - On ajuste l'exposant.

L'addition se passe donc très mal dès qu'elle met en jeu des nombres dont l'ordre de grandeur est très différent, car si y est très petit devant x, on a $x = x + y$. On dit que x a absorbé y.

Les problèmes de l'arithmétique des flottants :

- L'addition sur ordinateur n'est pas associative :

$$(2^{-53} + 2^{-53}) + 1.0 \longrightarrow 1 + 2^{-52}.$$
$$(2^{-53} + 1.0) + 2^{-53} \longrightarrow 1.$$

- La multiplication n'est pas associative non plus !
- La multiplication n'est pas distributive par rapport à l'addition.
- Il existe des flottants non nuls x tels que

$$x \times \frac{1}{x} \neq 1.$$

Règle d'or 1 : Quand on doit additionner plusieurs nombres il faut commencer par additionner les plus petits pour limiter la casse !

Règle d'or 2 : Quand on travaille avec l'arithmétique flottante, il faut se méfier des tests d'égalité. Le résultat d'une opération qui devrait renvoyer 0 peut renvoyer en fait un résultat non nul (certes très petit mais non nul). Donc le test `variable==0.0` devrait être proscrit !

Formalisation : Dans la norme la plus répandue qui régit le comportement des microprocesseurs d'ordinateurs (la norme IEEE 754), il est imposé que les quatre opérations de base , addition, soustraction, multiplication et division soient **correctes** au sens suivant :

Définition I.1.2

On dit qu'une opération binaire $\diamond$ est correcte *en arithmétique flottante si, pour tous flottants x et y convenables, le résultat R (qui est un flottant !) renvoyé par l'ordinateur pour l'opération $x \diamond y$ est exactement l'arrondi du calcul exact. Autrement dit*

$$R = fl(x \diamond y).$$

2 Analyse de la propagation des erreurs

Comme on l'a déjà vu, la bonne notion d'erreur est la notion d'erreur relative. Si x est un nombre réel approché (à cause de sa représentation machine, ou à cause de l'algorithme numérique utilisé) par une valeur $\tilde{x}$, on définit l'erreur absolue comme étant égale à

$$\Delta(x) = |x - \tilde{x}|,$$

et l'erreur relative par

$$\mathrm{Err}(x) = \frac{\Delta(x)}{|x|} = \frac{|x - \tilde{x}|}{|x|}.$$

2.1 Addition

Que se passe-t'il quand on additionne deux réels approchés ?

$$\begin{aligned}
\mathrm{Err}(x+y) &= \frac{|x+y-\tilde{x}-\tilde{y}|}{|x+y|} \\
&\leq \frac{|x-\tilde{x}|+|y-\tilde{y}|}{|x+y|} \\
&\leq \frac{\mathrm{Err}(x)|x|+\mathrm{Err}(y)|y|}{|x+y|} \\
&\leq \max(\mathrm{Err}(x),\mathrm{Err}(y))\frac{|x|+|y|}{|x+y|}.
\end{aligned}$$

On voit que si x et y ont les mêmes signes, on trouve

$$\mathrm{Err}(x+y) \leq \max(\mathrm{Err}(x),\mathrm{Err}(y)),$$

Alors que s'ils ont des signes différents (autrement dit si on considère une soustraction) on trouve

$$\mathrm{Err}(x+y) \leq \max(\mathrm{Err}(x),\mathrm{Err}(y))\frac{|x|+|y|}{||x|-|y||},$$

ce qui peut être catastrophique si $|x|$ et $|y|$ sont proches.

A cela, il faut ajouter l'erreur de troncature (ou d'arrondi ou de représentation) de la somme. Dont on a supposé que les opérations étaient *correctes sur les flottants*, ce terme d'erreur est donc égal au plus égal en valeur absolue à $u|\tilde{x}+\tilde{y}|$.

2.2 Multiplication

Effectuons un calcul similaire pour la multiplication :

$$\begin{aligned}
\mathrm{Err}(xy) &= \frac{|xy-\tilde{x}\tilde{y}|}{|xy|} \\
&= \frac{|x-\tilde{x}||y|+|y-\tilde{y}||\tilde{x}|}{|xy|} \\
&\leq \frac{|x-\tilde{x}||y|+|y-\tilde{y}||x|+|y-\tilde{y}||x-\tilde{x}|}{|xy|} \\
&\leq \mathrm{Err}(x)+\mathrm{Err}(y)+\mathrm{Err}(x)\mathrm{Err}(y) \approx \mathrm{Err}(x)+\mathrm{Err}(y).
\end{aligned}$$

Dans ce cas les erreurs relatives s'additionnent quelles que soient les valeurs de x et de y !

Exemple complet : la multiplication de deux réels après arrondi. Soit à effectuer la multiplication de deux réels x et y par un ordinateur. On commencer par les rentrer dans l'ordinateur (qui les arrondis) puis il effectue l'opération correcte. Quelle confiance peut-on avoir dans le résultat ?

On sait que pour tout nombre réel, dans les limites de l'underflow et de l'overflow, on a

$$\mathrm{Err}(x) = \frac{|x - fl(x)|}{|x|} \leq u.$$

$$\begin{aligned}|x\times y - fl(fl(x)\times fl(y))| &\le |x\times y - fl(x)\times fl(y)| + |fl(x)\times fl(y) - fl(fl(x)\times fl(y))|\\ &\le (2u+u^2)|x\times y| + u\times |fl(x)\times fl(y)|\\ &\le (2u+u^2)[x\times y| + u\times (|x\times y| + |x\times y - fl(x)\times fl(y)|)\\ &\le (2u+u^2)|x\times y| + u\times (1+2u+u^2)|x\times y|\\ &\le u(3+3u+u^2)|x\times y|.\end{aligned}$$

L'erreur relative que l'on commet est donc de l'ordre de $3u$ car u est très petit devant 1.

2.3 Formules stables et instables

Exemple 1 Soit à programmer le calcul de $\frac{1}{x} - \frac{1}{(x+1)}$ pour x grand.

$$\text{Formule 1 : } f_1(x) = \frac{1}{x} - \frac{1}{x+1},$$

$$\text{Formule 2 : } f_2(x) = \frac{1}{x(x+1)}.$$

En double précision, on obtient les résultats suivants

x	$f_1(x)$	$f_2(x)$
$1.0000000000000E+14$	$9.9396474057847E-29$	$9.9999999999999E-29$
$2.0000000000000E+14$	$2.5243548967072E-29$	$2.5000000000000E-29$
$4.0000000000000E+14$	$6.3108872417681E-30$	$6.2500000000000E-30$
$8.0000000000000E+14$	$1.5777218104420E-30$	$1.5625000000000E-30$
$1.6000000000000E+15$	$3.9443045261051E-31$	$3.9062500000000E-31$
$3.2000000000000E+15$	$9.8607613152626E-32$	$9.7656250000000E-32$
$6.4000000000000E+15$	$2.4651903288157E-32$	$2.4414062500000E-32$
$1.2800000000000E+16$	$0.0000000000000E+00$	$6.1035156250000E-33$
$2.5600000000000E+16$	$0.0000000000000E+00$	$1.5258789062500E-33$
$5.1200000000000E+16$	$0.0000000000000E+00$	$3.8146972656250E-34$

Etudions cet exemple plus en détail. Supposons que $1/x$ et $1/(x+1)$ soient calculés avec une erreur relative qui est l'erreur d'arrondi u. L'erreur relative commise sur la soustraction de ces deux flottants est donc majorée par

$$u\frac{\frac{1}{x}+\frac{1}{x+1}}{\frac{1}{x}-\frac{1}{x+1}} = u\frac{2x+1}{1} \sim 2ux.$$

On a vu qu'en double précision $u \sim 10^{-16}$ et donc pour $x \sim 10^{15}$, l'erreur relative maximale commise sur le résultat est de

$$2ux \sim 0.2.$$

En réalité on observe une relative plus faible, de l'ordre de 0.01 mais qui reste très supérieure à l'erreur due aux arrondis.

A la limite, 1 absorbe x et on trouve un résultat nul à l'évaluation de f_1.

Avec la seconde formule, l'erreur relative commise sur le produit de $1/x$ avec $1/(x+1)$ est la somme des erreurs relatives, c'est-à-dire au pire $2u$, ce qui est bien sûr bien meilleur et de plus, indépendant de x !

Exemple 2. Soit à résoudre une équation du second degré en simple précision

$$x^2 - 1634x + 2 = 0,$$

les formules usuelles donnent

$$\Delta = 2669948,\ \ \sqrt{\Delta} \approx 1633,997552017750$$

$$x_1 = \frac{1634+\sqrt{\Delta}}{2},\ \ x_2 = \frac{1634-\sqrt{\Delta}}{2}.$$

On voit que x_1 va être calculé sans problème par cette formule alors que la soustraction de 1634 avec $\sqrt{\Delta}$ va introduire une erreur importante.

Ainsi, si on calcule avec 7 chiffres significatifs (c'est l'équivalent en décimal de la simple précision), on trouve

$$\sqrt{\Delta} = 1633.998,\ \ x_1 = 1633,999,\ \ x_2 = 0,001000.$$

Si on calcule en double précision, ou par la formule $x_2 = 2/x_1$ on va trouver

$$x_2 = 0.00122.$$

L'erreur relative qu'on commet sur x_2 est donc de l'ordre de 0.22, c'est-à-dire presque $1/4$!

Pour calculer x_2 dans ce cas, il vaut mieux utiliser la propriété que le produit des racines est $x_1 x_2 = 2$ et donc calculer

$$x_2 = \frac{2}{x_1}.$$

On remplace ainsi une soustraction instable par une division stable !

2.4 Evaluation d'une fonction

Soit à évaluer une fonction f en un point x qui n'est pas connu exactement, on notera $\tilde{x}$ la valeur approchée de x. Quelle est l'erreur commise sur la valeur de f en x ? On suppose ici que f est évaluée de façon exacte.

La réponse est donnée par le théorème des accroissements finis :

$$f(\tilde{x}) - f(x) = f'(\xi)(\tilde{x} - x),$$

où $\xi \in [x, \tilde{x}]$. Dans un premier temps on supposera f suffisament régulière pour que $f'(x)$ soit une boonne approximation de $f'(\xi)$ (on pourrait raffiner l'étude ici). Que trouve-t'on en terme d'erreur relative ?

$$\mathrm{Err}(f(x)) = \frac{|f(\tilde{x}) - f(x)|}{|f(x)|} \leq \frac{|f'(x)||x - \tilde{x}|}{|f(x)|} \leq \frac{|f'(x)||x|}{|f(x)|}\mathrm{Err}(x).$$

L'erreur relative est donc amplifiée par un facteur (qui dépend de f et de x), qu'on peut appeller le conditionnement de f en x et qui est donné par

$$\mathrm{Cond}(f, x) = \left|\frac{x f'(x)}{f(x)}\right|.$$

Si ce facteur est faible, l'erreur est peu propagée par l'évaluation de f. S'il est grand c'est l'inverse.

Prenons quelques exemples :

–
$$f(x) = x^n, \quad \mathrm{Cond}(f, x) = n.$$

–
$$f(x) = \sqrt{x}, \quad \mathrm{Cond}(f, x) = \frac{1}{2}.$$

–
$$f(x) = \exp(x), \quad \mathrm{Cond}(f, x) = |x|.$$

f est donc mal conditionnée pour des grandes valeurs de x.

–
$$f(x) = \ln(x), \quad \mathrm{Cond}(f, x) = \frac{1}{|\ln(x)|}.$$

f est donc mal conditionnée près de $x = 1$.

Exercice I.2.1

Calculer les trente premiers termes de la suite $u_{n+1} = |ln\ (u_n)|$ avec $u_0 = 2$. Recommencer avec des valeurs de u_0 très proches de 2 (à 10^{-7} près par exemple). Commentaires ?

Le calcul du conditionnement d'une fonction f en un point est un élément important mais c'est loin d'être le seul. Car c'est un peu l'arbre qui cache la forêt. En effet, bien souvent la fonction f n'est pas évaluée de façon exacte. Elle est obtenue par composition d'opérations élémentaires ! !

Retour sur l'exemple 1 page 3

On a vu que pour calculer $\frac{1}{x} - \frac{1}{x+1}$ il valait mieux mettre la fonction sous la forme factorisée $\frac{1}{x(x+1)}$. Est-ce que ce problème est dû au mauvais conditionnement de la fonction f en question ?

Calculons son conditionnement pour x grand

$$\mathrm{Cond}(f, x) = \frac{\frac{1}{x} - \frac{x}{(x+1)^2}}{\frac{1}{x} - \frac{1}{x+1}} = 1 + \frac{x}{x+1} \underset{\infty}{\sim} 2.$$

Le conditionnement est donc très bon au voisinage de l'infini. Il ne s'agit pas ici d'un problème de conditionnement ...

Autre exemple :

$$f(x) = \sqrt{1+x} - \sqrt{x}.$$

Tous calculs faits on trouve, pour $x \geq 0$ grand

$$\mathrm{Cond}(f,x) = \frac{1}{2}\frac{x}{\sqrt{x}\sqrt{1+x}} \sim \frac{1}{2}.$$

Cela semble très bon mais en pratique on peut obtenir de grandes erreurs (pour des grandes valeurs de x) car, comme on l'a vu la soustraction de grands nombres de même ordre de grandeur est une opération très mauvaise !

En revanche, si on utilise la formule équivalente suivante

$$f(x) = \frac{1}{\sqrt{x}+\sqrt{x+1}},$$

tout se passe bien car on a remplacé une soustraction par une addition !

Moralité : l'algorithme de calcul de la valeur d'une fonction en un point doit être examiné avec soin pour éviter la propagation excessives d'erreurs. Il faut donc bien distinguer la notion de **fonction** et la notion d'**expression** que l'on programme en pratique et qui est censée calculer la fonction f.

2.5 Evaluation d'un polynôme

On souhaite calculer la valeur d'un polynôme P en un point x donnée par

$$P(x) = a_0 + a_1 x + \cdots + a_n x^n.$$

On suppose les $(a_i)_i$ et x connus exactement (i.e. représentés de façon exacte par les flottants). D'après ce qui précède, la façon dont l'ordinateur va mener les calculs pour obtenir la valeur de $P(x)$ influe sur la précision du résultat.

1. Méthode naïve : on pose $x^0 = 1$, $s_0 = a_0$ et on calcule par récurrence

$$\begin{cases} x^k = x^{k-1} * x, \\ s_k = s_{k-1} + a_k x^k. \end{cases}$$

Cette méthode coûte 2 multiplications et une addition à chaque itération.

Calculons les erreurs absolues, notées par le symbole Δ. Quand on fait un produit, les erreurs relatives s'ajoutent donc nous avons

$$\Delta(a_k x^k) \leq ku|a_k x^k|.$$

Quand on fait une somme, les erreurs **absolues** s'ajoutent, ainsi que l'erreur d'arrondi

$$\begin{aligned} \Delta(s_k) &\leq \Delta(s_{k-1}) + ku|a_k x^k| + u(|s_{k-1}| + |a_k||x|^k), \\ &\leq \Delta(s_{k-1}) + ku|a_k||x|^k + u(|a_0| + \cdots + |a_k||x|^k). \end{aligned}$$

On somme tout ceci pour $k = 0, ..., n$ (et on utilise $\Delta s_0 = 0$. Il vient

$$\Delta(P(x)) \leq (n+1)u \sum_{k=0}^{n} |a_k||x|^k.$$

2. Méthode de Hörner : elle consiste à écrire le polynôme sous la forme suivante

$$P(x) = a_0 + x * (a_1 + x * (a_2 + \cdots + x * (a_{n-1} + x a_n) \cdots)),$$

et à effectuer le calcul en commençant par la contribution des termes les plus élevés.

La récurrence descendante s'écrit de la façon suivante, avec $p_n = a_n$

$$p_{k-1} = a_{k-1} + x p_k,$$

ce qui économise une multiplication à chaque itération.

A chaque itération, on vérifie que

$$p_k = a_k + a_{k+1} x + \cdots + a_n x^{n-k}.$$

Effectuons l'analyse de propagation des erreurs :

$$\begin{aligned}\Delta(p_{k-1}) &\leq \Delta(xp_k) + u(|a_{k-1}| + |xp_k|)\\ &\leq |x|\Delta(p_k) + u|x||p_k| + u(|a_{k-1}| + |x||p_k|)\\ &\leq |x|\Delta(p_k) + u(|a_{k-1}| + 2|x||p_k|).\end{aligned}$$

On trouve donc, in fine, puisque $\Delta p_n = 0$

$$\Delta P(x) = \Delta p_0 \leq u\Big[(|a_0| + 2|x||p_1|) + |x|\big(|a_1| + 2|x||p_2| + |x|(|a_2| + \cdots)\big)\Big].$$

Il vient

$$\Delta P(x) \leq u\sum_{k=0}^{n-1}|a_k||x|^k + 2u\sum_{k=1}^{n}|x|^k|p_k|.$$

En utilisant la formule qui donne p_k ci-dessus on trouve

$$\Delta P(x) \leq u\sum_{k=0}^{n-1}|a_k||x|^k + 2u\sum_{k=1}^{n}(|a_k||x|^k + \cdots + |a_n||x|^n) \leq u\sum_{k=0}^{n}(2k+1)|a_k||x|^k,$$

ce qui est sensiblement équivalent à la méthode précédente, à ceci près que les gros coefficients $(2k+1)$ sont portés par les termes d'ordre élevé. Si $P(x)$ est en fait une série entière convergente tronquée, on voit que ces termes sont les plus petits et contribuent donc moins à l'erreur.

3. On peut aussi centrer le polynôme en des points différents.

Exercice I.2.2

On considère le polynôme

$$P(x) = (1-x)^6 = 1 - 6x + 15x^2 - 20x^3 + 15x^4 - 6x^5 + x^6.$$

Tracer, à l'aide de Matlab ou de scilab, la courbe représentative de ce polynôme dans l'intervalle $[0.995, 1.005]$ *en utilisant respectivement les deux expressions algérbiquement équivalentes de* P.

3 Sommation de séries

3.1 Sommation des séries à termes positifs

Rappels : Soit $\sum_{n=0}^{+\infty} a_n$ une série à termes positifs.

- S'il existe $C < 1$ telle que $a_{n+1} \leq Ca_n$ pour tout n, alors la série est convergente et on a

$$R_N = \sum_{n=N+1}^{\infty} a_n \leq a_{N+1}\sum_{k\geq 0} C^k = \frac{a_{N+1}}{1-C}.$$

- **Critère de d'Alembert :** Si $\limsup \frac{a_{n+1}}{a_n} < 1$ alors la série converge. Si la limite supérieure est > 1, la série diverge. Si elle est égale à 1, on ne peut pas conclure directement.
- Si une série à termes positifs convergente relève du critère de d'Alembert, alors on peut majorer le reste pour n assez grand. En effet soit $C = \limsup \frac{a_{n+1}}{a_n} < 1$, et $\varepsilon > 0$ assez petit pour que $C + \varepsilon < 1$. Il existe un rang n_0 à partir duquel on a $a_{n+1} \leq (C+\varepsilon)a_n$. A partir de ce rang le reste s'écrit donc

$$R_N = \sum_{n=N+1}^{+\infty} a_n \leq a_{N+1}\sum_{k=0}^{\infty}(C+\varepsilon)^k = \frac{a_{N+1}}{1-(C+\varepsilon)},\ \ \forall N \geq n_0.$$

- Si une série à termes positifs converge, alors n'importe quelle somme partielle est inférieure à la somme de la série. La valeur approchée obtenue en arrêtant la sommation à n'importe quel rang est donc toujours une valeur approchée par défaut.

Exemple : Pour $x > 0$, on a

$$\exp(x) = \sum_{n=0}^{+\infty}\frac{x^n}{n!}.$$

Dans ce cas, on peut estimer le reste de la façon suivante

$$R_N = \sum_{n=N+1}^{\infty} \frac{x^n}{n!} = x^{N+1} \sum_{k=0}^{\infty} \frac{x^k}{(N+1+k)!} \leq \frac{x^{N+1}}{(N+1)!} \sum_{k=0}^{\infty} \frac{x^k}{k!},$$

car

$$(N+1+k)! \geq (N+1)!k!.$$

On en déduit une majoration du reste

$$R_N \leq \frac{x^{N+1}}{(N+1)!} e^x.$$

Pour x proche de 0, les deux facteurs x^{N+1} et $\frac{1}{(N+1)!}$ tendent vers 0 rapidement et donc on voit que des petites valeurs de N suffiront à assurer une précision ε souhaitée. En revanche pour de grandes valeurs de x, le terme x^{N+1} tend vers l'infini et e^x peut être grand, ce qui rend l'estimation du reste très mauvaise et il faudra donc de grandes valeurs de N pour assurer une précision donnée.

Par exemple, prenons $\varepsilon = 10^{-6}$. Voici les valeurs de N que l'on trouve

$$x = 0.5 \Rightarrow N = 8$$

$$x = 1 \Rightarrow N = 10$$

$$x = 10 \Rightarrow N = 44.$$

$$x = 20 \Rightarrow N = 80.$$

En pratique, que font les ordinateurs pour calculer e^x avec x de façon efficace ? Il y a plusieurs méthodes, l'une d'elles est d'utiliser les propriétés de l'exponentielle. On suppose pour cela que $\log(2)$ est connu avec grande précision (en général sa valeur est stockée **en dur** dans la mémoire du micro-processeur). Pour tout $x > 0$ donné, on calcule la division euclidienne de x par $\log(2)$ avec reste symétrique, c'est-à-dire qu'on cherche un entier $a \in \mathbb{Z}$ et $r \in \mathbb{R}$ tel que $|r[\leq \frac{\log(2)}{2}$ tels que

$$x = a \log(2) + r.$$

Pour cela, on utilise l'algorithme de division usuel suivi d'un arrondi entier par exemple. On utilise alors la formule

$$e^x = e^{a \log(2) + r} = 2^a \times e^r.$$

Le calcul de l'exponentielle de x est donc ramené au calcul de l'exponentielle de r (que l'on effectue grâce au développement en série entière de l'exponentielle) suivi d'une multiplication par 2^a ce qui est très facile en arithmétique flottante (il s'agit seulement d'un décalage d'exposant).

Comparaison avec des intégrales : Soit la série à termes positifs $\sum_{n\geq 0} a_n$. On suppose que a_n s'écrit sous la forme $a_n = f(n)$ où f est une fonction continue sur $\mathbb{R}$, positive et décroissante dont l'intégrale généralisée $\int_0^{+\infty} f(x)\,dx$ est convergente.

Théorème I.3.3

Sois les hypothèses ci-dessus, la série $\sum_{n\geq 0} a_n$ est convergente et on a une estimation du reste

$$\int_{N+1}^{+\infty} f(x)\,dx \leq R_N \leq \int_N^{+\infty} f(x)\,dx.$$

Ainsi, si on sait calculer une primitive de la fonction f, on peut avoir une estimation précise du reste.

Exemple : On considère la série $\sum_{n\geq 1} \frac{1}{n^2}$. On peut appliquer le théorème précédent avec $f(x) = \frac{1}{x^2}$. On obtient donc que la série st convergente. On peut démontrer (mais c'est assez difficile) que la some de la série vaut

$$\sum_{n\geq 1} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

On a deplus l'estimation du reste

$$\frac{1}{N+1} \leq R_N \leq \frac{1}{N}.$$

Ainsi pour estimer la valeur de $\frac{\pi^2}{6}$ (et par exemple en déduire une valeur approchée de π) à 10^{-6} près, il faut sommer la série précédente jusqu'à

$$N \sim 10^6.$$

3.2 Théorème des séries alternées

Théorème I.3.4

Soit $(a_n)_n$ une suite de réels **positifs**. *On considère la série, dite* alternée*, suivante*

$$\sum_{n=0}^{+\infty}(-1)^n a_n. \tag{I.3}$$

On suppose que la suite $(a_n)_n$ est **décroissante** *et tend vers 0, alors*

1. *La série* (I.3) *converge.*
2. *Le reste $R_N = \sum_{n=N+1}^{\infty}(-1)^n a_n$ de la série au rang N est majoré par le dernier terme négligé :*
$$|R_N| \leq a_{N+1}.$$
3. *Le signe de R_N est celui du dernier terme négligé i.e. $(-1)^{N+1}$.*

Exemple 1 : pour $x < 0$

$$\exp(x) = \sum_{n=0}^{+\infty}(-1)^n \frac{|x|^n}{n!}.$$

Il s'agit d'une série alternée *à partir d'un certain rang* (à partir de $n_0 + 1 \geq x$). Donc en théorie on peut contrôler le reste. Le problème est que les termes de la série peuvent être grands en module avant ce certain rang et on tombe alors sur des soustractions instables.

Ainsi le calcul de $e^{-15} \sim 3\,10^{-7}$ par cette formule donne des résultats tout à fait incorrects en double précision (on fait une erreur d'un facteur 2 !). En calculant la taille des termes que l'on somme, on trouve que le plus grand terme en valeur absolue est celui pour $n = 15$ et vaut environ 334864 ! En réalité le terme de rang 14 et de rang 15 sont exactement opposés ! Leur prise en compte est donc inutile.

Solution du problème pour cet exemple : utiliser la formule $e^x = \frac{1}{e^{-x}}$. Comme $-x$ est positif on peut utiliser le paragraphe précédent puis ensuite prendre l'inverse.

Exemple 2 : pour tout $x \in \mathbb{R}$, on a

$$\sin(x) = \sum_{n=0}^{+\infty}(-1)^n \frac{x^{2n+1}}{(2n+1)!}.$$

Pour calculer $\sin(1000)$, on n'a pas intérêt à utiliser directement cette formule car les erreurs d'arrondi vont être très grandes et de plus, il faudra sommer beaucoup de termes pour avoir une précision acceptable.

L'astuce est donc d'utiliser les propriétés de 2π-périodicité de la fonction sinus pour se ramener au calcul de $\sin(x)$ pour $x \in]-\pi, \pi[$. On peut même se ramener à un réel dans $]-\frac{\pi}{2}, \frac{\pi}{2}|$ si on utilise les formules de trigonométrie et la formule suivante pour le cosinus :

$$\cos(x) = \sum_{n=0}^{+\infty}(-1)^n \frac{x^{2n}}{(2n)!}.$$

Chapitre II

Méthodes itératives de résolution d'équations

Le but de ce chapitre est d'étudier des méthodes itératives qui permettent de calculer certaines quantités. En général, il s'agit de construire des suites qui convergent vers le résultat cherché et d'estimer l'erreur commise en s'arrêtant à la n-ième itération du calcul.

Remarquons dès à présent que la résolution d'une équation $f(x) = b$ est bien entendu un problème de même nature que celui de la résolution de $f(x) = 0$, en changeant f en $f - b$.

Les références utiles pour ce chapitre sont typiquement [QSG10, HH06].

Notons que les questions d'existence et unicité de solution(s) de $f(x) = 0$ sont loin d'être triviales en général (sauf dans le cas linéaire ...) et seront rapidement discutées.

1 Les théorèmes de base en analyse réelle

La plupart des résultats de ce chapitre et des suivants reposent de façon fondamentale sur des théorèmes d'analyse élémentaires qu'il faut connaître et comprendre sur le bout des doigts. Le point clé de toute l'histoire est la présence, dans $\mathbb{R}$, d'une relation d'ordre compatible avec les structures algébriques et avec la topologie de $\mathbb{R}$.

Notation : pour $\alpha, \beta \in \mathbb{R}$, on note $S(\alpha, \beta) \subset \mathbb{R}$ le segment joignant α à β c'est-à-dire

$$S(\alpha, \beta) = \begin{cases} [\alpha, \beta], & \text{si } \alpha \leq \beta \\ [\beta, \alpha], & \text{si } \beta \leq \alpha. \end{cases}$$

Théorème II.1.1 (Valeurs intermédiaires)

Soit $f : [a, b] \to \mathbb{R}$ une fonction continue. Pour tout $y \in S(f(a), f(b))$, il existe $c \in [a, b]$ tel que

$$f(c) = y.$$

En général, le c ainsi obtenu n'est pas unique.

Equivalent en dimension supérieure : l'image d'un connexe par une application continue est un connexe.

Théorème II.1.2 (Rolle)

Soit $f : [a, b] \to \mathbb{R}$ une fonction continue, dérivable dans $]a, b[$ telle que $f(a) = f(b)$, alors il existe $c \in]a, b[$ tel que

$$f'(c) = 0.$$

En général, le c ainsi obtenu n'est pas unique.

Preuve :

Si f est constante, n'importe quel élément c de $]a, b[$ convient.

Sinon, on peut supposer par exemple que $\sup_{[a,b]} f > f(a)$ (quitte à changer f en $-f$). Comme f est continue et $[a, b]$ compact, il existe $c \in [a, b]$ tel que $\sup_{[a,b]} f = f(c)$ et de plus, $c \in]a, b[$.

Comme $f(c) \geq f(c + h)$ pour tout $h > 0$, on obtient que $f'(c) \leq 0$ et comme $f(c) \geq f(c - h)$ pour tout $h > 0$, on

obtient aussi que $f'(c) \geq 0$, d'où le résultat. ■

Corollaire II.1.3 (Théorème des accroissements finis)

Soit $f : [a, b] \to \mathbb{R}$ une fonction continue, dérivable dans $]a, b[$. Il existe $c \in]a, b[$ tel que

$$f(b) - f(a) = f'(c)(b - a).$$

Preuve :

Il suffit d'appliquer le théorème de Rolle à $g(x) = f(x) - (x - a)\frac{f(b)-f(a)}{b-a}$. ■

Ces théorèmes permettent ensuite de montrer par exemple toutes les formules de Taylor. Le théorème des accroissements finis n'étant qu'une formule de Taylor à l'ordre 0.

2 Résolution d'une equation $f(x) = 0$, $f : \mathbb{R} \to \mathbb{R}$

On cherche ici à calculer une approximation d'une solution de l'équation $f(x) = 0$ où f est une fonction de $\mathbb{R}$ dans $\mathbb{R}$, qu'on supposera au moins continue. Nous alons voir plusieurs méthodes pour arriver à ce résultat que nous essaierons de comparer.

2.1 Existence et unicité de solutions

En dimension 1 d'espace, le moyen le plus simple de prouver l'existence d'une solution c'est de montrer qu'il existe deux points a et b tels que $f(a)$ et $f(b)$ sont de signe contraires et de conclure par le théorème des valeurs intermédiaires. L'unicité provient le plus souvent d'une hypothèse de monotonie.

2.2 La méthode graphique

Malgré son côté *rustique*, la lecture graphique d'une solution est souvent extrêmement simple à mettre en place et permet d'obtenir à moindre frais une valeur approchée d'une solution à l'aide d'outils très simples (une calculatrice graphique, un logiciel de type Scilab, ...)

Il suffit bien entendu de tracer la courbe de la fonction et la droite $\{y = 0\}$ puis de chercher leurs intersections, en utilisant le zoom si besoin.

2.3 La méthode de dichotomie ou de bissection

On suppose donnés deux réels $a < b$ tels que $f(a)f(b) < 0$, c'est-à-dire que $f(a)$ et $f(b)$ sont de signes opposés (et aucun des deux n'est une racine !). D'après le théorème des valeurs intermédiaires, nous savons qu'il existe un $x^* \in]a, b[$ tel que $f(x^*) = 0$.

La méthode de dichotomie consiste à diviser à chaque itération l'intervalle en deux-sous intervalles et de déterminer dans lequel des deux sous-intervalles se trouve x^*.

- On pose $a_0 = a$, $b_0 = b$.
- Pour tout $n \geq 0$, calculer $\xi_n = \frac{a_n+b_n}{2}$, $f(a_n)$, $f(b_n)$ et $f(\xi_n)$.
- Si $f(\xi_n) = 0$, on a trouvé une racine x^*. Fin de l'algorithme.
- Si $f(a_n) \times f(\xi_n) > 0$, poser $a_{n+1} = \xi_n$ et $b_{n+1} = b_n$.
- Si $f(a_n) \times f(\xi_n) < 0$, poser $a_{n+1} = a_n$ et $b_{n+1} = \xi_n$.

Théorème II.2.4

La suite $(a_n)_n$ ainsi obtenue est croissante. La suite $(b_n)_n$ est décroissante. Les deux suites convergent vers un réel x^* vérifiant $f(x^*) = 0$.
De plus, on a l'estimation

$$|a_n - x^*| \leq |a - b|2^{-n},$$
$$|b_n - x^*| \leq |a - b|2^{-n}.$$

Cette méthode a plusieurs avantages :

- Elle n'utilise que des évaluations de f. Il n'y a besoin en réalité que du calcul de $f(\xi_n)$ à chaque itération.
- On a une estimation précise de l'erreur. Si l'intervalle initial est de taille $|b - a| = 1$, le nombre d'itérations nécessaires pour obtenir la solution avec une précision absolue ε donnée se détermine comme suit

$$2^{-n} \leq \varepsilon \Longrightarrow n \leq \frac{|\log(\varepsilon)|}{\log 2},$$

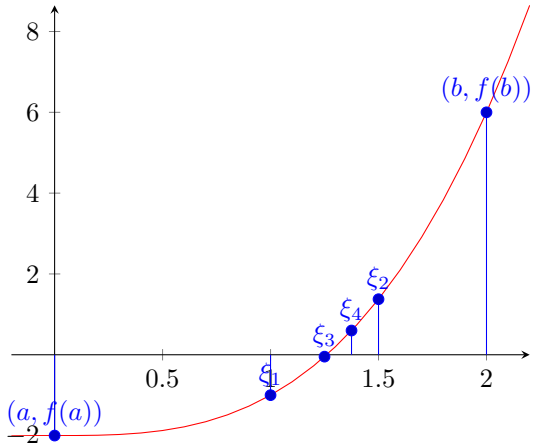

FIGURE II.1 – La méthode de dichotomie

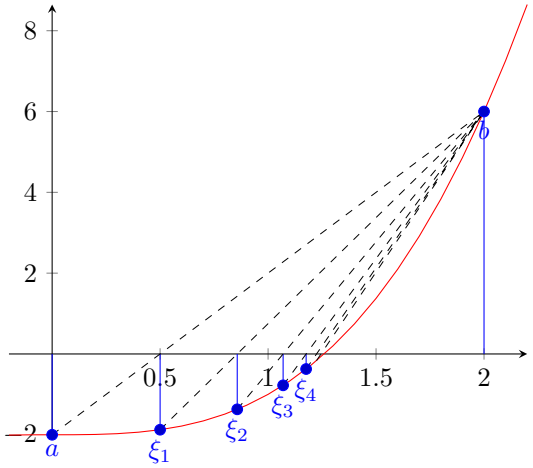

FIGURE II.2 – La méthode de la fausse position

si bien sûr on néglige ici les erreurs d'arrondis pourtant inévitables comme on l'a vu précédemment.
Son inconvénient majeur est une convergence assez lente. A chaque itération on gagne un chiffre juste en écriture binaire. On verra que certaines méthodes font beaucoup mieux.

En pratique, la méthode de dichotomie peut être utilisée pour se rapprocher raisonnablement d'une racine et ainsi assurer que les méthodes que nous allons voir dans la suite vont converger.

2.4 La méthode de la fausse position : *regula falsi*

Principe : L'idée de la méthode est similaire à celle de la dichotomie sauf qu'on choisit un autre point ξ_n que le milieu de a_n et b_n comme nouveau point de comparaison. Plus précisément, on prend pour ξ_n le point d'intersection de la sécante définie par les points $(a_n, f(a_n))$ et $(b_n, f(b_n))$ avec l'axe des ordonnées.

La formule explicite est

$$\xi_n = \frac{f(a_n)b_n - f(b_n)a_n}{f(a_n) - f(b_n)}.$$

Comme $f(a_n)$ et $f(b_n)$ sont toujours de signe contraire au cours de la méthode, la division ne pose pas de problème et ξ_n est bien une combinaison convexe de a_n et b_n.

Si f est une fonction affine, la méthode va donc converger en 1 itération. Dans le cas général, on espère améliorer la convergence, néanmoins, on n'est pas assuré que l'intervalle $[a_n, b_n]$ tende vers 0.

Exemple : Si la fonction f est convexe et si $f(a) < 0$ et $f(b) > 0$, alors la suite $(b_n)_n$ construite par cette méthode va être constante et la suite (a_n) va croitre vers l'unique racine de la fonction. Ceci vient du fait que la corde d'une fonction convexe se situe toujours au-dessus de la courbe. Ainsi l'intersection de la corde avec l'axe des ordonnées $(\xi_n, 0)$ est au-dessus du point $(\xi_n, f(\xi_n))$ ce qui prouve que $f(\xi_n) \leq 0$.

Convergence :

Théorème II.2.5

La suite $(a_n)_n$ construite par la méthode est croissante, la suite $(b_n)_n$ est décroissante. Donc elles sont toutes les deux convergentes, on notera α et β leurs racines respectives. On a alors $f(\alpha) = 0$ ou bien $f(\beta) = 0$.

Preuve :

La monotonie des deux suites est triviale. Les deux suites sont donc convergentes (car bornées). Comme la fonction f est continue, les suites $(f(a_n))_n$ et $(f(b_n))_n$ convergent respectivement vers $f(\alpha)$ et $f(\beta)$. De plus nous avons $f(a_n) \leq 0$ pour tout n et $f(b_n) \geq 0$ pour tout n. On en déduit que $f(\alpha) \leq 0$ et $f(\beta) \geq 0$.

On veut montrer que $f(\alpha)$ ou $f(\beta)$ est nul. Supposons le contraire. On a alors $f(\alpha) < 0$ et $f(\beta) > 0$. On en déduit que la suite $(\xi_n)_n$ converge vers

$$\xi = \frac{f(\alpha)\beta - f(\beta)\alpha}{f(\alpha) - f(\beta)}.$$

Or $(\xi_n)_n$ est une suite extraite ou bien de la suite $(a_n)_n$ ou bien de la suite $(b_n)_n$. Donc on a $\xi = \alpha$ ou $\xi = \beta$. Ceci est impossible car ξ est une combinaison strictement convexe entre α et β (qui sont distincts ...). ■

2.5 Méthodes de point fixe

On cherche à construire une suite définie par récurrence de la façon suivante

$$x_{n+1} = g(x_n), \tag{II.1}$$

où g est une fonction explicite à déterminer de sorte que la suite $(x_n)_n$ converge vers x^* solution de l'équation $f(x) = 0$.

Exemples de telles suites :

$$\begin{cases} \text{Ex 1} : x_{n+1} = x_n + 1, \\ \text{Ex 2} : x_{n+1} = -x_n, \\ \text{Ex 3} : x_{n+1} = x_n^2, \\ \text{Ex 4} : x_{n+1} = x_n^2 - 2, \\ \text{Ex 5} : x_{n+1} = \sqrt{2 + x_n}, \\ \text{Ex 6} : x_{n+1} = 1 + \dfrac{2}{x_n}. \end{cases} \tag{II.2}$$

Limites potentielles :

Théorème II.2.6

Soit $g : \mathbb{R} \mapsto \mathbb{R}$ une fonction continue. Si la suite $(x_n)_n$ définie par (II.1) *est bien définie et converge vers une limite $l \in \mathbb{R}$ alors l est un point fixe de g c'est-à-dire que l vérifie*

$$g(l) = l.$$

Preuve :

Si $x_n \to l$, alors $x_{n+1} \to l$. De plus comme g est supposée continue on a aussi $g(x_n) \to l$. On peut donc passer à la limite dans la définition de la suite et obtenir

$$g(l) = l.$$

■

Ce théorème *ne dit pas* que la suite $(x_n)_n$ converge. Il permet seulement de déterminer les éventuelles limites possibles de la suite.

Retour sur les exemples :

- Exemple 1 : Si la suite converge, la limite doit vérifier $l = l + 1$ ce qui est bien sûr impossible. On en conclut que la suite x_n ne peut pas converger. En fait on peut calculer explicitement $x_n = x_0 + n$ et on voit que $x_n \to +\infty$.
- Exemple 2 : La limite éventuelle doit vérifier $l = -l$, ca ne peut donc être que 0. Dans ce cas précis, on voit que la suite $(x_n)_n$ ne peut converger que si $x_0 = 0$.

– Exemple 3 : Si une limite existe on trouve $l = l^2$, et donc $l = 0$ ou bien $l = 1$. Dans cet exemple on peut calculer les valeurs de x_n explicitement. On trouve

$$x_n = (x_0)^{2n}.$$

Le comportement de la suite est donc le suivant :
  – Si $|x_0| < 1$, x_n tend vers 0.
  – Si $x_0 \in \{-1, 1\}$, x_n tend vers 1 (la suite est même stationnaire à partir d'un certain rang).
  – Si $|x_0| > 1$, x_n tend vers $+\infty$.
– Exemple 4 : La limite potentielle l vérifie

$$l = l^2 - 2, \quad \text{ou} \quad l^2 - l - 2 = 0,$$

– Exemple 5 : La suite n'est bien définie que si $x_0 > 2$ la limite potentielle l vérifie

$$l = \sqrt{2 + l},$$

elle est donc nécessairement positive et vérifie

$$l^2 - l - 2 = 0.$$

– Exemple 6 : On est sûr que la suite est bien définie si $x_0 > 0$, la limite potentielle l vérifie

$$l = 1 + \frac{2}{l}, \quad \text{ou} \quad l^2 - l - 2 = 0.$$

Les trois derniers exemples sont tous différents, les suites $(x_n)_n$ obtenues sont différentes mais néanmoins, **si elles convergent** elles convergent toutes vers une solution de l'équation

$$f(l) = 0, \ \text{où } f(x) = x^2 - x - 2.$$

Peut-on maintenant essayer de comprendre si l'une de ces suites va converger ou pas ?

Proposition II.2.7

On suppose que g est une fonction continue qui envoie un intervalle borné $[a, b]$ **dans lui-même**. *Alors g admet un point fixe dans $[a, b]$ c'est-à-dire :*

$$\exists l \in [a, b], \ \ g(l) = l.$$

Preuve :

On considère la fonction $\psi(x) = g(x) - x$ et on vérifie que $\psi(a) \leq 0$ et $\psi(b) \geq 0$. L'existence du point fixe découle donc du théorème des valeurs intermédiaires. ■

Quelques remarques s'imposent :

– Sous les conditions de la proposition précédente, la suite $(x_n)_n$ est bien définie par (II.1) dès que $x_0 \in [a, b]$ et toutes les itérées appartiennent alors à $[a, b]$.
– La fonction g peut avoir plusieurs points fixes (exemple $g(x) = x$).

Malheureusement ces hypothèses ne suffisent pas à assurer la convergence de la suite vers un point fixe. Ansi l'exemple 2 ci-dessus vérifie les hypothèses de la proposition avec n'importe quel intervalle $[-a, a]$ symétrique autour de 0. On a déjà vu que la fonction g admet un **unique** point fixe ($l = 0$) dans ces intervalles mais que la suite $(x_n)_n$ ne converge que si $x_0 = 0$.

Définition II.2.8

On dit qu'une fonction $g : [a, b] \mapsto \mathbb{R}$ est **contractante** *s'il existe un nombre positif k vérifiant $k < 1$ et tel que*

$$|g(x) - g(y)| \leq k|x - y|, \ \ \forall x, y \in [a, b].$$

On dit aussi que g est k-contractante.

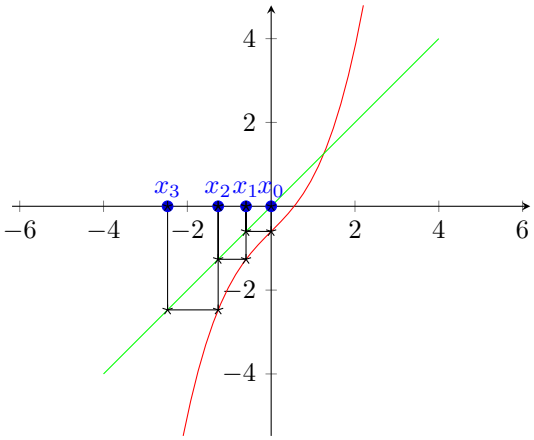

FIGURE II.3 – Itération de point fixe pour $g(x) = x + 0.3(x^3 - 2)$

La condition $k < 1$ est tout à fait cruciale ! ! La défition ci-dessus exprime que la distance entre les images de x et y par g est plus petite d'un facteur (au moins) k que la distance entre x et y.

Théorème II.2.9 (du point fixe de Banach)

Si g est une fonction continue de $[a, b]$ dans lui-même et que g est **contractante***, alors*
- *g admet un unique point fixe l dans $[a, b]$.*
- *Pour toute donnée initiale $x_0 \in [a, b]$, la suite $(x_n)_n$ définie par* (II.1) *converge vers l.*
- *On a l'estimation de l'erreur suivante*

$$|x_n - l| \leq |x_0 - l|k^n, \ \ \forall n \geq 0.$$

Remarquons que la convergence est géométrique (on dit que la convergence est linéaire).

Ce théorème est en fait valable en toute généralité sur un espace métrique complet avec une démonstration en tout point similaire. L'hypothèse de complétude est essentielle car la preuve repose de façon fondamentale sur la notion de suite de Cauchy.

Preuve :
- L'unicité du point fixe ne fait aucun doute en utilisant la contractivité ...
- On va montrer l'existence en montrant que la suite des itérées converge (et ce pour toute donnée initiale), ce qui montrera les points 2 et 3 du théorème.
  Pour cela on va montrer qua la suite $(x_n)_n$ des itérées est de Cauchy. En effet, en utilisant la k-contractivité on montre aisément, par récurrence, que l'on a

$$|x_{n+p} - x_n| \leq k^n|x_p - x_0| \leq k^n|b - a|. \tag{II.3}$$

  Or, comme $k < 1$ (c'est évidemment crucial ici : l'exemple $g(x) = -x$ est parlant de ce point de vue) la suite k^n tend vers 0 quand n tend vers l'infini, ce qui montre bien que la suite est de Cauchy.
  Comme $\mathbb{R}$ est complet, la suite $(x_n)_n$ est convergente et on a alors vu que sa limite l était nécessairement un point fixe de g. Ce dernier étant unique de surcroît, comme on l'a vu précédemment.
- Si on reprend (II.3) et que l'on fait tendre p vers l'infini, on obtient

$$|l - x_n| \leq k^n|l - x_0| \leq k^n|b - a|,$$

  ce qui fournit une estimation de la vitesse de convergence de la suite $(x_n)_n$.

■

Une généralisation assez utile de ce théorème, et elle aussi valable sur n'importe quel espace métrique complet, est la suivante

Théorème II.2.10

On suppose qu'il existe $p \geq 1$ tel que la p-ième itérée de g notée $g^p = g \circ \cdots \circ g$ est contractante sur $[a, b]$ alors :
- *g admet un unique point fixe l dans $[a, b]$.*
- *Pour toute donnée initiale $x_0 \in [a, b]$, la suite itérée $(x_n)_n$ définie par* (II.1) *converge vers l.*

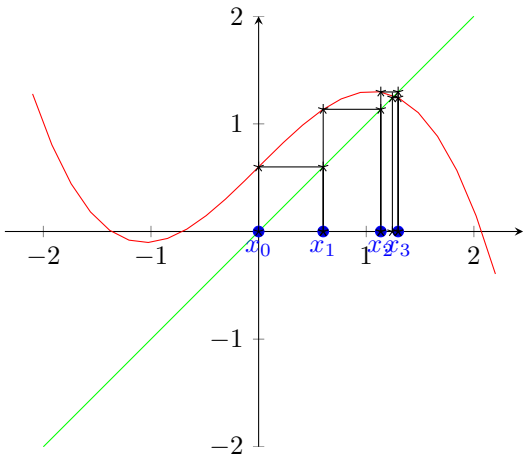

FIGURE II.4 – Itération de point fixe pour $g(x) = x - 0.3(x^3 - 2)$

Il est remarquable ici que ce soit la suite itérée par g qui converge et pas seulement celle itérée par g^p !

Preuve :

On commence par appliquer le théorème du point fixe de Banach à g^p. Il existe donc un point fixe l de g^p. Celui-ci vérifie

$$g(g^p(l)) = g(l) = g^p(g(l)),$$

donc $g(l)$ est aussi un point fixe de g^p. Ce dernier étant unique, on en déduit que $g(l) = l$ et donc que l est un point fixe de g. Comme tout point fixe de g est aussi point fixe de g^p, l'unicité est claire.

On constate maintenant que, pour tout $0 \leq k \leq p-1$, la suite $(x_{np+k})_n$ vérifie la relation de récurrence

$$x_{(n+1)p+k} = g^p(x_{np+k}),\ \forall n \geq 0$$

et donc elle converge vers l d'après le théorème précédent (car la convergence de la suite d'itérées dans le théorème de Banach ne dépend pas de la donnée initiale !).

Comme toutes les suites extraites $(x_{np+k})_n$ convergent vers la même limite l, on en déduit que toute la suite $(x_n)_n$ converge vers l. **Sauriez-vous le justifier complètement ?** ■

A titre d'exemple, on peut considérer la suite

$$x_{n+1} = \cos(x_n),$$

dans l'intervalle $[-1, 1]$. La fonction $x \mapsto g(x) = \cos(x)$ n'est pas contractante mais son itérée $g^2 = \cos \circ \cos$ l'est.

Fonctions contractantes et dérivées : Pour déterminer si une fonction donnée est contractante et si on peut appliquer les résultats précédents, la méthode la plus simple est de calculer la dérivée de g, si bien sûr elle est dérivable.

Proposition II.2.11

Soit $g : [a, b] \mapsto [a, b]$ une fonction dérivable. Soit $0 \leq k < 1$. Les propositions suivantes sont équivalentes :

- *On a*
$$|g'(\xi)| \leq k, \quad \forall \xi \in [a, b],$$
- *g est k-contractante sur $[a, b]$.*

La démonstration découle, dans un sens, de la définition de la dérivée, dans l'autre sens du théorème des accroissements finis.

Théorème II.2.12

Soit g une fonction de classe $\mathcal{C}^1$ de $\mathbb{R}$ dans $\mathbb{R}$. On suppose que g a un point fixe $l \in \mathbb{R}$, et que de plus on a $|g'(l)| < 1$.
Alors, il existe un intervalle $[a, b]$ contenant l, et $0 \leq k < 1$ tels que g soit k-contractante sur $[a, b]$ et envoie $[a, b]$ dans lui-même. En conséquence pour toute donnée initiale dans cet intervalle $[a, b]$, la suite des itérées $(x_n)_n$ converge vers l et on a

$$|x_n - l| \leq |x_0 - l|k^n.$$

Si de plus, $g'(l) = 0$ et g est de classe $\mathcal{C}^2$, alors la convergence est quadratique, c'est-à-dire qu'il existe $C > 0$ tel que

$$|x_{n+1} - l| \leq C|x_n - l|^2, \ \ \forall n \geq 0.$$

Preuve :

Choisissons k tel que $|g'(l)| < k < 1$. Comme g est de classe $\mathcal{C}^1$, il existe un intervalle symétrique autour de l noté $I = [l + \delta, l + \delta]$, qu'on peut même choisir maximal tel que $\sup_I |g'| \leq k$.

Pour tout $x \in I$ nous avons

$$|g(x) - l| = |g(x) - g(l)| \leq (\sup_I |g'|)|x - l| \leq k|x - l| \leq \delta,$$

ce qui prouve que $g(x) \in [l - \delta, l + \delta] = I$. D'où le résultat.

La preuve de la convergence quadratique repose sur la formule de Taylor

$$x_{n+1} - l = g(x_n) - g(l) = \underbrace{g'(l)(x_n - l)}_{=0} + \frac{1}{2}(x_n - l)^2 g''(\xi_n),$$

d'où le résultat avec $C = \frac{1}{2} \sup_{[l-\delta, l+\delta]} |g''|$. ■

Définition II.2.13

Un point fixe l d'une fonction g de classe $\mathcal{C}^1$ est dit :
- **attractif** *: si $|g'(l)| < 1$.*
- **répulsif** *: si $|g'(l)| > 1$.*

Le théorème ci-dessus montre que près d'un point fixe attractif, la suite des itérées va toujours converger.

Reprenons les trois exemples ci-dessus. Le point fixe cherché est la solution positive de l'équation $l^2 - l - 2 = 0$, celle-ci n'est autre que $l = 2$.

- Exemple 4 : $g(x) = x^2 - 2$. On calcule $g'(2) = 4$, le point fixe est répulsif, la fonction g n'est pas contractante au voisinage du point fixe. Il est probable que cette suite ne converge pas vers le point fixe. En tout état de cause, il faut faire une analyse plus détaillée de cet exemple pour en comprendre le comportement.
- Exemple 5 : $g(x) = \sqrt{2 + x}$. On calcule $g'(2) = 1/4$, donc le point fixe est attractif et la fonction g est contractante au voisinage du point fixe. Pour toute donnée initiale assez proche de celui-ci, la méthode va converger. La vitesse de convergence sera, *grosso modo* de $1/4^n$, ce qui est bien mieux que la méthode de dichotomie.
  Précisons un intervalle de données initiales admissibles. Il faut trouver les valeurs de x autour du point fixe pour lesquelles $|g'| < 1$. On obtient la convergence de la suite des itérées pour toute donnée dans $]-\frac{7}{4}, +\infty[$.
- Exemple 6 : $g(x) = 1 + \frac{2}{x}$. On calcule $g'(2) = -1/2$, ici encore le point fixe est attractif et la fonction g est contractante et donc la convergence est assurée. Un intervalle de convergence de la suite est donc $]\sqrt{2}, +\infty[$.

Choix de g connaissant f En général, on connaît la fonction f dont on cherche les zéros, il faut donc trouver une fonction g dont les points fixes sont les zéros de f. On a vu plus haut qu'il y a de nombres façons non-équivalentes de le faire. On va étudier un premier choix possible

$$g(x) = x - \alpha f(x), \tag{II.4}$$

où α est un coefficient réel à déterminer. Il est clair que les points fixes de g sont les zéros de f.

Comment choisir le paramètre α pour pouvoir appliquer le théorème du point fixe ? Calculons la dérivée de g en un point fixe l :

$$g'(l) = 1 - \alpha f'(l).$$

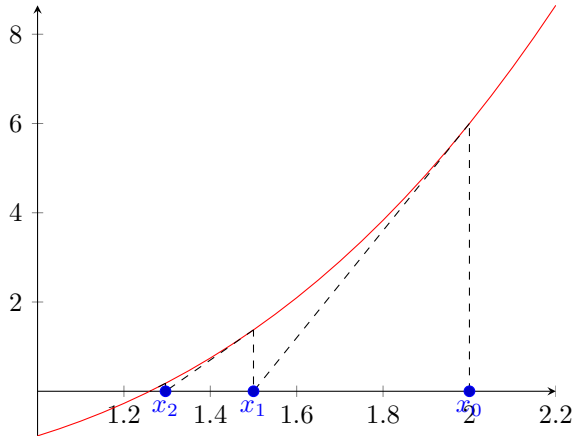

FIGURE II.5 – Méthode de Newton pour $(x^3 - 2)$

On a vu que pour assurer la convergence, il faut que cette quantité soit la plus petite possible, et même nulle pour obtenir une convergence quadratique. Pour cela, il faut prendre

$$\alpha = \frac{1}{f'(l)}.$$

Le problème est qu'en général, on ne connaît évidemment pas l ni donc $f'(l)$. En pratique, en prenant un coefficient α proche de $1/f'(l)$ on espère de bonnes propriétés de convergence. C'est l'une des façons d'appréhender la méthode de Newton que l'on décrit ci-dessous.

2.6 Méthode de Newton

La méthode de Newton est basée sur le travail effectué précédemment. On a vu que la convergence de la méthode itérative définie par (II.4) serait la plus efficace si α était proche de $1/f'(l)$. Cela donne l'idée d'utiliser le processus itératif sur la fonction suivante

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

La méthode de Newton consiste donc à construire la suite itérative suivante

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

On remarque immédiatement que cette suite n'est bien définie que si $f'(x_n)$ est non nul (et même pas trop petit), ce qui sera le cas par exemple si $f'(l) \neq 0$.

Théorème II.2.14

On suppose que $f'(l) \neq 0$ et que g est une fonction de classe $\mathcal{C}^2$. Il existe un intervalle centré en l qui laisse fixe la fonction g tel que pour toute donnée initiale dans cet intervalle, la suite des itérées $(x_n)_n$ converge vers l de façon quadratique.

Preuve :

Par construction, nous avons $g'(l) = 0$. Calculons

$$|x_{n+1} - l| = |g(x_n) - g(l)| \leq \left(\sup_{S(x_n,l)} |g'| \right) |x_n - l|,$$

Or, pour $x \in S(x_n, l)$ (on note $S(a, b)$ le segment entre a et b) on a

$$|g'(x)| = |g'(x) - g'(l)| \leq \left(\sup_{S(x_n,l)} |g''| \right) |x_n - l|.$$

On en déduit que

$$|x_{n+1} - l| \leq \left(\sup_{S(x_n,l)} |g''| \right) |x_n - l|^2.$$

Si on suppose que g'' est bornée par M, alors on voit que

$$|x_n - l| \leq \frac{1}{M} \Longrightarrow |x_{n+1} - l| \leq |x_n - l| \leq \frac{1}{M}.$$

L'intervalle centré en l et de rayon $\frac{1}{M}$ est donc invariant par g. Dans cet intervalle, nous avons la majoration

$$|x_{n+1} - l| \leq M|x_n - l|^2.$$

On dit que la convergence est quadratique. On obtient l'estimation finale

$$|x_n - l| \leq \frac{1}{M}(M|x_0 - l|)^{2^n}.$$

■

Remarque II.2.15

- *La convergence quadratique implique que le nombre de chiffres justes dans le résultat double à chaque itération.*
- *Le fait que la convergence doit quadratique implique que*

$$\lim_{n\to\infty} \frac{|x_{n+1} - l|}{|x_n - l|} = 0,$$

ce qui implique enfin que

$$\lim_{n\to\infty} \frac{x_n - x_{n+1}}{x_n - l} = 1.$$

Donc, pour n assez grand nous avons l'équivalent

$$|x_{n+1} - x_n| \sim |x_n - l|,$$

et donc $|x_{n+1} - x_n|$ est un bon indicateur de l'erreur. En général, pour obtenir une erreur absolue de l'erreur sur le point fixe l d'ordre $\varepsilon > 0$, on arrête l'algorithme de Newton dès que

$$|x_{n+1} - x_n| \leq \varepsilon.$$

Remarquons également que nous savons que pour n grand, $x_n - x_{n+1}$ et $x_n - l$ sont du même signe, ce qui permet de déterminer si l'approximation x_n obtenue à la fin de l'algorithme est une approximation par excès ou par défaut de la valeur de l.

Exemple : Soit à calculer numériquement la racine carrée de 2. On cherche donc la solution positive de l'équation $0 = f(x) = x^2 - 2$.

La méthode de Newton pour cette équation donne

$$x_{n+1} = x_n - \frac{x_n^2 - 2}{2x_n} = \frac{x_n^2 + 2}{2x_n} = \frac{1}{2}\left(x_n + \frac{2}{x_n}\right).$$

Cherchons un intervalle centré en $\sqrt{2}$ dans lequel on est assurés de la convergence. Pour cela il faut trouver $\delta > 0$ tel que $|g'| \leq 1$ sur $]\sqrt{2} - \delta, \sqrt{2} + \delta[$.

On trouve

$$g'(x) = \frac{1}{2}\left(1 - \frac{2}{x^2}\right).$$

On constate que, pour $x > 0$, g' est toujours plus petit que $\frac{1}{2}$, il faut trouver l'ensemble des x pour lesquels $g'(x) \geq -1$. On obtient l'intervalle de convergence

$$\left]\sqrt{2}\left(1 - \frac{1}{\sqrt{3}}\right), \sqrt{2}\left(1 + \frac{1}{\sqrt{3}}\right)\right[.$$

Dans cet exemple, on peut montrer que cet intervalle de convergence n'est pas optimal et qu'en fait la suite des itérées converge pour toute donnée initiale strictement positive comme on va le voir ci-dessous.

Méthode de Newton pour les fonctions convexes : Si on applique la méthode de Newton à une fonction convexe, les choses se passent encore mieux que dans le cas général.

Théorème II.2.16

Soit f est de classe $\mathcal{C}^2$. Soit l un zéro de f tel que $f'(l) > 0$ et soit $b > l$ tel que f est convexe sur $[l, b]$. Pour toute donnée initiale x_0 dans $[l, b]$, la suite des itérées de la méthode de Newton $(x_n)_n$ est décroissante, reste dans l'intervalle $[l, b]$ et converge vers l.

Preuve :

Montrons tout d'abord les propriétés suivantes :

- f est strictement croissante sur $[l, b]$.
- f est strictement positive sur $]l, b]$.

D'après la convexité de f, la tangente de f au point x_n est située sous la courbe représentative de f, en particulier la valeur de cette tangente au point l est strictement négative et donc l'intersection x_{n+1} de la tangente avec l'axe des x est forcément comprise entre l et x_n. ■

Si on revient à l'exemple du calcul de la racine carrée de 2 vu précédemment, on peut voir aisément que la fonction $f(x) = x^2 - 2$ est convexe, que $f'(\sqrt{2}) = 2\sqrt{2} > 0$ et que donc la méthode de Newton converge pour toute donnée initiale $x_0 \geq \sqrt{2}$.

En fait, on peut faire encore mieux car on constate que si $x_0 \in]0, \sqrt{2}[$, alors la première itérée x_1 vérifie $x_1 > \sqrt{2}$ et on est donc ramenés à l'étude précédente. On a donc convergence de la méthode de Newton pour toute donnée initiale positive.

2.7 Méthode de Newton pour les polynômes

Comme on l'a vu précédemment pour la fonction $f(x) = x^2 - 2$, la méthode de Newton s'applique en général avec bonheur à la résolution d'équations polynomiales.

Soit P un polynôme de degré k ayant k racines distinctes, on les note $r_1 < \cdots < r_k$. On suppose, sans perte de généralité, que P est unitaire (i.e. le coeff de plus haut degré est x^k).

Pour tout $x \geq r_k$ on a $P'(x) > 0$ et $P''(x) > 0$. On peut donc appliquer le théorème précédent au polynôme P et donc la méthode de Newton va converger de façon monotone vers r_k pour toute donnée initiale plus grande que r_k.

Pour trouver les autres racines, une fois connue r_k, on aimerait factoriser P sous la forme

$$P(x) = (x - r_k)Q(x),$$

et appliquer la méthode de Newton à Q. En pratique le calcul de Q peut poser problème, d'autant que r_k n'est connue que de façon approchée. L'idée est donc de remarquer que

$$P'(x) = Q(x) + (x - r_k)Q'(x),$$

et donc

$$\frac{P'(x)}{P(x)} = \frac{1}{x - r_k} + \frac{Q'(x)}{Q(x)}.$$

Ainsi on peut écrire l'itération de la méthode de Newton pour le polynôme Q de la façon suivante :

$$x_{n+1} = x_n - \frac{1}{\frac{P(x_n)}{P'(x_n)} - \frac{1}{x_n - r_k}}.$$

L'intérêt de cette méthode est qu'elle ne nécessite que le calcul de P et de P' et de plus on peut seulement utiliser la valeur approchée de r_k obtenue précédemment.

2.8 Divers exemples et remarques

1. Pour une racine multiple, la méthode de Newton converge seulement linéairement. En revanche, si on connaît par avance l'ordre de multiplicité m de la racine recherchée, on peut modifier la méthode pour récupérer une convergence quadratique

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

On pourra observer cela sur la fonction $f(x) = x^2$.

2. Calcul de l'inverse d'un nombre avec une calculatrice dont la touche de division est en panne :
   Soit $a > 0$. On souhaite trouver x tel que $ax = 1$. Pour cela on observe que cela revient à trouver un point fixe non nul de la fonction $g(x) = 2x - ax^2$. La dérivée de g au point solution $x = 1/a$ est donnée par
   $$g'(1/a) = 2 - 2a(1/a) = 0.$$
   La méthode de point fixe va donc converger de façon quadratique.
   L'algorithme s'écrit
   $$x_{n+1} = 2x_n - ax_n^2.$$
   On voit bien que le calcul des itérées ne requiert que des multiplications et des soustractions.
   On pourra, à titre d'exercice, étudier les propriétés de convergence de la suite $(x_n)_n$ en fonction de la donnée initiale.
3. Calcul du taux d'intérêt souhaitable pour un prêt.
   Si on en emprunte à la banque un capital C, à un taux t, pendant n mensualités, la mensualité que l'on devra régler se calcule par la formule
   $$M = \frac{C\frac{t}{12}}{1 - \left(1 + \frac{t}{12}\right)^{-n}}.$$
   Etant donné un capital C, une durée souhaitée n et une mensualité maximale que l'on souhaite payer M, quel est le taux d'intérêt que je dois essayer de négocier auprès de ma banque ?

3 Résolution de $f(x) = 0$ en dimension quelconque (finie)

Pour résoudre les équations en dimension quelconque on ne peut évidemment pas utiliser des méthodes reposant sur la relation d'ordre dans $\mathbb{R}$ comme la dichtomie, la méthode de la sécante, etc ... En revanche, la méthode de point fixe ou celle de Newton continuent à s'appliquer. Dans ce dernier cas, on l'appelle la méthode de Newton-Raphson. C'est bien sûr la Jacobienne (ou la différentielle) qui joue le rôle de la dérivée.

3.1 Méthode de point-fixe

Comme on l'a vu plus haut, le théorème du point fixe de Banach est valable dans n'importe quel espace métrique complet. On peut donc appliquer toute la théorie précédente au cas de fonctions de $\mathbb{R}^d$ dans $\mathbb{R}^d$ (voire même en dimension infinie). Quelques remarques sont nécessaires :

- La propriété de contraction est une propriété **métrique**, ce qui signifie que le choix d'une bonne norme est essentiel pour ne pas "rater" une application contractante. Voir les exemples linéaires ci-dessous.
- La condition sur la dérivée assurant la contraction d'une application devient (sur un domaine K, compact, convexe et invariant par g)
  $$\sup_K \|Dg\| < 1,$$
  où $\|.\|$ est la norme matricielle sur $M_d(\mathbb{R})$ associée à la norme choisie sur $\mathbb{R}^d$.

Lemme II.3.17

On note $\mathcal{N}$ l'ensemble des normes sur $M_d(\mathbb{C})$ subordonnées à une norme sur $\mathbb{C}^d$, c'est-à-dire l'ensemble des normes N sur $M_d(\mathbb{C})$ qui s'écrivent sous la forme

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

pour une certaine norme $\|.\|$ sur $\mathbb{C}^d$.
On a alors le résultat suivant, pour toute matrice $A \in M_d(\mathbb{C})$

$$\rho(A) = \inf_{N \in \mathcal{N}} N(A).$$

Preuve :

- Soit $N \in \mathcal{N}$ et $\|.\|$ la norme sur $\mathbb{C}^d$ correspondante.
  Pour toute valeur propre $\lambda \in \mathbb{C}$, de vecteur propre $x \in \mathbb{C}^d$, on a $Ax = \lambda x$ et donc
  $$|\lambda|\|x\| = \|Ax\| \leq N(A)\|x\|,$$

d'où

$$|\lambda| \leq N(A).$$

Ceci étant vrai pour toute valeur propre, on en déduit $\rho(A) \leq N(A)$ et donc

$$\rho(A) \leq \inf_{N \in \mathcal{N}} N(A).$$

– L'inégalité inverse repose sur le théorème de la base adaptée que l'on va rappeller ici. L'histoire commence par le théorème de trigonalisation qui dit que toute matrice A (à valeurs complexes) il existe une matrice de passage P et une matrice triangulaire T telles que

$$A = P^{-1}TP.$$

Remarquons que les valeurs propres de A sont les mêmes que celles de T (qui sont ses coefficients diagonaux).
On pose maintenant $D_\delta = \mathrm{diag}(1, \delta, \ldots, \delta^{d-1})$ et on observe, par un petit calcul, que la matrice $D_\delta^{-1}TD_\delta$ converge vers la diagonale de T quand $\delta \to 0$. On en déduit que

$$\|D_\delta^{-1}TD_\delta\|_\infty \xrightarrow[\delta \to 0]{} \rho(T) = \rho(A).$$

On pose alors

$$N_\delta(M) = \|D_\delta^{-1}PMP^{-1}D_\delta\|_\infty,$$

qui est bien une norme sur l'ensemble des matrices, par ailleurs, elle est bien dans $\mathcal{N}$ car

$$N_\delta(M) = \sup_{x \neq 0} \frac{\|D_\delta^{-1}PMP^{-1}D_\delta x\|_\infty}{\|x\|_\infty} = \sup_{y \neq 0} \frac{\|D_\delta^{-1}PMy\|_\infty}{\|D_\delta^{-1}Py\|_\infty},$$

et donc c'est la norme associée à $y \mapsto \|D_\delta^{-1}Py\|_\infty$.
Ainsi, à $\varepsilon > 0$ fixé, il existe $\delta > 0$ tel que on a $N_\delta(A) \leq \rho(A) + \varepsilon$. On a donc

$$\rho(A) + \varepsilon \leq \inf_{N \in \mathcal{N}} N(A),$$

et donc

$$\rho(A) \leq \inf_{N \in \mathcal{N}} N(A).$$

■

On peut maintenant en déduire le lemme suivant.

Lemme II.3.18

Soit $A \in M_d(\mathbb{C})$, on a l'équivalence

$$\left[(A^n)_n \xrightarrow[n \to \infty]{} 0\right] \Longleftrightarrow \rho(A) < 1.$$

Preuve :

$\Leftarrow$ Si $\rho(A) < 1$, il existe une norme $N \in \mathcal{N}$ telle que $N(A) < 1$. On a alors

$$N(A^n) \leq (N(A))^n \xrightarrow[n \to \infty]{} 0.$$

$\Rightarrow$ Soit $\lambda \in \mathbb{C}$ valeur propre de A de plus grand module associée à un vecteur propre x. On a donc en particulier

$$A^n x = \lambda^n x.$$

Par hypothèse, on a donc $(\lambda^n)_n$ qui doit tendre vers 0 ce qui impose $|\lambda| < 1$ et donc $\rho(A) < 1$.

■

On déduit maintenant aisément le résultat suivant qui nous donne une condition nécessaire et suffisante pour que la méthode de point fixe fonctionne pour une application affine.

Théorème II.3.19

Soit $B \in M_d(\mathbb{C})$ et $c \in \mathbb{C}^d$. Les propositions suivantes sont équivalentes :

1. *Il existe une norme sur $\mathbb{C}^d$ telle que l'application $x \mapsto Bx + c$ est contractante.*
2. *Le rayon spectral de B vérifie $\rho(B) < 1$.*

Comme vous l'avez déjà vu, on peut utiliser ce résultat pour construire des méthodes de résolution de systèmes linéaires. Soit à résoudre

$$Ax = b,$$

on suppose qu'on dispose d'une décomposition de A sous la forme $A = M - N$ avec M inversible de sorte que l'équation précédente est équivalente à

$$x = M^{-1}Nx + M^{-1}b,$$

c'est-à-dire à un problème de point fixe pour l'application $x \mapsto M^{-1}Nx + M^{-1}b$.

La méthode itérative associée va donc converger dès lors que le rayon spectral de $M^{-1}N$ vérifie

$$\rho(M^{-1}N) < 1.$$

Par ailleurs, pour que la méthode soit utilisable en pratique il faut être capable de calculer aisément la matrice M^{-1} ou, à défaut, de savoir résoudre les systèmes linéaires associés à M de façon simple.

Ce sera par exemple le cas si M est diagonale ou triangulaire supérieure. On tombe alors que les méthodes de Jacobi, de Gauss-Seidel, etc ...

Dans le cadre non-linéaire général, on dispose alors d'un théorème équivalent au Théorème II.2.12, qui s'énonce comme suit et se démontre comme dans le cas $d = 1$, une fois qu'on a choisi une norme convenable sur $\mathbb{R}^d$ (pour laquelle la norme induite de la Jacobienne $Dg(l)$ est strictement plus petite que 1).

Théorème II.3.20

Soit g une fonction de classe $\mathcal{C}^1$ de $\mathbb{R}^d$ dans $\mathbb{R}^d$. On suppose que g a un point fixe $l \in \mathbb{R}$, et que de plus on a $\rho(Dg(l)) < 1$ (on dit alors que le point fixe est attractif).
Alors, il existe un compact K tel que $l \in \circ K$, et $0 \leq k < 1$ tels que g envoie K dans lui-même et soit k-contractante sur cet ensemble pour une certaine norme N. En conséquence pour toute donnée initiale dans K, la suite des itérées $(x_n)_n$ converge vers l et on a

$$N(x_n - l) \leq N(x_0 - l)k^n.$$

Comme toutes les normes sur $\mathbb{R}^d$ sont équivalentes ; on peut également écrire pour toute norme $\|.\|$

$$\|x_n - l\| \leq Ck^n.$$

Si de plus, $Dg(l) = 0$ et g est de classe $\mathcal{C}^2$, alors la convergence est quadratique, c'est-à-dire qu'il existe $C > 0$ tel que

$$N(x_{n+1} - l) \leq CN(x_n - l)^2, \ \ \forall n \geq 0.$$

3.2 Méthode de Newton-Raphson

Par des raisonnements analogues à ceux qui ont conduit à la construction de la méthode de Newton en 1D (ou bien de type géométrique, ou de type analytique ...), nous pouvons considérer la méthode itérative suivante pour une fonction $f : \mathbb{R}^d \to \mathbb{R}^d$ de classe $\mathcal{C}^1$ par exemple (en réalité, la méthode peut également fonctionner en dimension infinie ...)

$$x_{n+1} = x_n - (Df(x_n))^{-1} f(x_n), \ \ \forall n \geq 0.$$

On dispose alors de deux théorèmes généraux de convergence. Le premier est simple à prouver mais relativement inutile puisque ces hypothèses font intervenir la solution du problème (que l'on ne connaît pas ...) alors que le second est

plus fin et ne fait intervenir que des hypothèses sur le comportement de la fonction f près de la donnée initiale choisie x_0.

Théorème II.3.21

On se donne une application $f : U \mapsto \mathbb{R}^d$ de classe $\mathcal{C}^1$ définie sur un ouvert U de $\mathbb{R}^d$. On suppose qu'il existe $x^* \in U$, $R > 0$, $M > 0$ et $L > 0$ tels que
- *La boule $B(x^*, R)$ est incluse dans U et $f(x^*) = 0$.*
- *La différentielle de f est inversible en tout point de $B(x^*, R)$ et on a*

$$\|(Df(x))^{-1}\| \leq M, \ \ \forall x \in B(x^*, R).$$

- *La différentielle de f est Lipschitzienne sur $B(x^*, R)$ de constante de Lipschitz L, i.e.*

$$\|Df(x) - Df(y)\| \leq L\|x - y\|, \ \ \forall x, y \in B(x^*, R).$$

On pose alors $r = \min(R, \frac{2}{ML})$.
Alors, pour toute donnée initiale x_0 dans $B(x^*, r)$, la relation de récurrence

$$x_{n+1} = x_n - (Df(x_n))^{-1} f(x_n),$$

définit une suite dont tous les termes sont dans $B(x^*, r)$ et qui converge vers x^*. De plus, la convergence est quadratique.

Preuve :

- Pour tout $x \in B(x^*, r) \subset B(x^*, R)$ on a bien $Df(x)$ inversible et comme $f(x^*) = 0$, on a

$$\begin{aligned}\big(x - (Df(x))^{-1} f(x)\big) - x^* &= (x - x^*) - (Df(x))^{-1}(f(x) - f(x^*)) \\ &= -(Df(x))^{-1}\left(f(x) - f(x^*) - Df(x)(x - x^*)\right) \\ &= -(Df(x))^{-1} \int_0^1 (Df(x^* + t(x - x^*)) - Df(x)).(x - x^*)\, dt,\end{aligned}$$

  de sorte qu'en prenant la norme et en utilisant les hypothèses, on obtient

$$\big\|\big(x - (Df(x))^{-1} f(x)\big) - x^*\big\| \leq ML\|x - x^*\|^2 \int_0^1 t dt \leq \frac{ML}{2} r^2,$$

  et par définition de r, on a bien que

$$x - (Df(x))^{-1} f(x) \in B(x^*, r).$$

  Ainsi, la suite $(x_n)_n$ est bien définie et contenue dans la boule.
- Par ailleurs, l'inégalité précédente montre aussi que

$$\|x_{n+1} - x^*\| \leq \frac{ML}{2}\|x_n - x^*\|^2,$$

  ce qui donne bien la convergence voulue dès lors que

$$\frac{ML}{2}\|x_0 - x^*\| < 1,$$

  ce qui est bien le cas sous les hypothèses énoncées.

■

Théorème II.3.22 (Théorème de Newton-Kantorovich)

On se donne $f : U \mapsto \mathbb{R}^d$ de classe $\mathcal{C}^1$ et $x_0 \in U$. On suppose que
- *$Df(x_0)$ est inversible.*
- *Si on pose $h_0 = -(Df(x_0))^{-1}f(x_0)$ et $x_1 = x_0 + h_0$ alors on a*

$$B(x_1, \|h_0\|) \subset U.$$

- *La différentielle de f est Lipschitzienne sur la boule $B(x_1, \|h_0\|)$ de constante de Lipschitz L.*
- *La condition de petitesse suivante est vérifiée*

$$\|f(x_0)\|\|(Df(x_0))^{-1}\|^2 L \leq \frac{1}{2}.$$

Alors la fonction f admet un unique zéro, noté x^*, dans la boule $B(x_1, \|h_0\|)$ et la suite de Newton partant de x_0 est bien définie et converge vers x^*.

Remarquons que, rien ne dit que la différentielle de f au point x^* est inversible et donc rien ne garantit la convergence quadratique de la méthode. De fait, la preuve du théorème (voir par exemple [HH06]), montre que Df est inversible en x_n pour tout n (ce qui est indispensable pour définir la suite) mais elle ne donne que l'estimation $\|Df(x_{n+1})^{-1}\| \leq 2\|(Df(x_n))^{-1}\|$.

3.3 Exemples

On donne ici quelques exemples non standard d'applications de la méthode de Newton.

3.3.1 Calcul approché de l'inverse d'une matrice

Pour $U = GL_d(\mathbb{R})$ ouvert de $M_d(\mathbb{R})$, et A une matrice inversible fixée.
Pour tout $M \in U$, on pose $f(M) = M^{-1} - A$. La différentielle de f s'écrit

$$df(M).H = -M^{-1}.H.M^{-1},$$

ce qui montre que la méthode de Newton-Raphson pour ce problème s'écrit

$$M_{n+1} = M_n + M_n.(M_n^{-1} - A).M_n = 2M_n - M_n.A.M_n.$$

Elle converge vers l'inverse de A dès lors que $\rho(\mathrm{Id} - A.M_0) < 1$.

3.3.2 Un modèle simplifié de positionnement GPS

Le principe du récepteur GPS est le suivant : on cherche à déterminer la position (x, y, z) d'un point (dans le repère géostationnaire par exemple) en utilisant les informations envoyées par 4 satelittes différents. Chaque d'entre eux envoie sa position (x_i, y_i, z_i) au moment de l'envoi du signal ainsi que l'heure exacte t_i à laquelle le signal a été envoyé.

Le récepteur recoit les 4 signaux à un instant T_i sachant que tous les satellites sont parfaitement synchronisés entre eux mais que l'horloge interne du récepteur est très certainement déphasée d'une durée inconnue Δ.

Sachant que les signaux se propagent à la vitesse de la lumière c (connue), les équations liant les différentes grandeurs sont

$$\sqrt{(x-x_1)^2 + (y-y_1)^2 + (z-z_1)^2} = c(T_1 - t_1 + \Delta),$$

$$\sqrt{(x-x_2)^2 + (y-y_2)^2 + (z-z_2)^2} = c(T_2 - t_2 + \Delta),$$

$$\sqrt{(x-x_3)^2 + (y-y_3)^2 + (z-z_3)^2} = c(T_3 - t_3 + \Delta),$$

$$\sqrt{(x-x_4)^2 + (y-y_4)^2 + (z-z_4)^2} = c(T_4 - t_4 + \Delta).$$

Il s'agit d'un système de quatre équations à quatre inconnues (x, y, z, Δ) que l'on peut résoudre à l'aide de la méthode de Newton.

3.3.3 Calcul de valeurs propres d'une matrice symétrique (voir [Ded06])

Soit A une matrice symétrique. Chercher une valeur propre λ de A associée à un vecteur propre normalisé v revient à résoudre l'équation

$$0 = F(v, \lambda) \stackrel{\text{def}}{=} \begin{pmatrix} Av - \lambda v \\ (1 - \|x\|^2)/2 \end{pmatrix}.$$

Pour vérifier les hypothèses du théorème de Newton-Raphson, on va calculer la différentielle de F en un point solution (v, λ) et montrer que, si la valeur propre est simple, alors la méthode de Newton va converger pour une donnée initiale assez proche de la solution cherchée.

Calculons

$$DF(v, \lambda).(w, \mu) = \begin{pmatrix} Aw - \lambda w - \mu v \\ -(v, w) \end{pmatrix}.$$

Ainsi si (w, μ) annulle la différentielle, la deuxième équation nous dit que v et w sont orthogonaux et comme A est symétrique, Aw et v sont aussi orthogonaux. Donc, en prenant le produit scalaire de la première équation par v, on trouve que $\mu = 0$ et il reste finalement

$$Aw = \lambda w, \text{ et } (v, w) = 0.$$

Il existe une solution non triviale à ce problème si et seulement si λ est valeur propre double.

Ainsi, si λ est simple, la Jacobienne est inversible et le théorème de Newton-Raphson peut s'appliquer.

3.3.4 Décomposition de Dunford et méthode de Newton (Cf. [BR06])

Il s'agit de montrer que pour toute matrice $A \in M_d(K)$ dont le polynôme caractéristique est scindé (ce qui est toujours le cas sur $\mathbb{C}$ par exemple), alors il existe un unique couple de matrices (D, N) telles que

1. $A = D + N$.
2. D et N commuttent.
3. D est diagonalisable et N est nilpotente.

Par ailleurs, D et N sont des polynômes en A.

Plusieurs démonstrations de ce résultat existent (à l'aide du lemme chinois et de la relation de Bezout notamment) mais il se trouve que l'on peut montrer le résultat suivant :

Soit Q un polynôme annulateur de A à racines simples (i.e. le polynôme minimal dans les conditions présentes) que l'on peut obtenir à l'aide de la formule $Q = P/(P \wedge P')$, où P est le polynôme caractéristique de A. Alors la suite définie par récurrence par $A_0 = A$ et

$$A_{n+1} = A_n - (P'(A_n))^{-1} P(A_n), \quad \forall n \geq 0,$$

est bien définie et stationne à partir d'un certain rang sur une matrice D diagonalisable, qui est égale à un polynôme en A et telle que $A - D$ est nillpotente.

Chapitre III

Interpolation / Approximation

Beaucoup de notions abordées dans ce chapitre peuvent servir à illustrer vos leçons. C'est en particulier le cas des leçons d'analyse portant sur les suites et séries de fonctions, les espaces L^p, les espaces de Hilbert, ...

Les références possibles sont nombreuses, je propose les ouvrages suivants [Sch91, Chap VII],[Dem91, Chap II], [CM84, Chap I].

Introduction

La problématique de l'interpolation est la suivante : on cherche une fonction, la plus simple possible (polynomes, polynomes par morceaux, ...) qui vérifie certaines contraintes, typiquement :

1. On impose la valeur y_i de la fonction sur une famille de points $(x_i)_{0\leq i\leq n}$.
2. On impose la valeur g_i de la dérivée de la fonction en ces mêmes points.

1 Interpolation de Lagrange

On cherche à résoudre le problème de l'interpolation dans la classe des polynômes avec des contraintes du premier type, c'est-à-dire sur les valeurs de la fonction.

1.1 Existence et unicité du polynôme de Lagrange

Théorème III.1.1

Etant donnés $n+1$ points distincts $(x_i)_{0\leq i\leq n}$ et $n+1$ valeurs réelles $(y_i)_{0\leq i\leq n}$, il **existe** *un* **unique** *polynôme p de* **degré inférieur ou égal à** *n qui vérifie*

$$p(x_i) = y_i, \ \ \forall 0 \leq i \leq n.$$

Ce polynôme est appelé le polynôme d'interpolation de Lagrange des points du plan $(x_i, y_i)_{0\leq i\leq n}$.

Remarque III.1.2

- *n est le degré minimal pour lequel on peut assurer l'existence d'un tel polynôme. Il peut arriver que le degré réel du polynôme de Lagrange soit plus faible.*
- *Si on autorise les polynômes de degré supérieur ou égal à $n+1$ alors il existe une infinité de polynômes vérifiant les contraintes, en effet il suffit d'interpoler sur un $n+2$-ième point pour voir que c'est possible.*

Commençons par étudier des exemples :

- $n = 0$: on interpole un seul point, le polynôme de Lagrange est le polynôme constant égal à y_0.
- $n = 1$: on interpole deux points, le polynôme de Lagrange est l'équation de la droite reliant les deux points

$$p(x) = y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0}.$$

- $n = 2$: on interpole trois points, le polynôme de Lagrange est l'équation d'une parabole que l'on va essayer de trouver. Commençons par remarquer que le polynome que l'on cherche doit passer par les deux premiers points,

une idée est donc de le chercher sous la forme suivante

$$p(x) = y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0} + q(x),$$

où q est un polynôme à déterminer. Celui-ci doit vérifier

$$q(x_0) = 0, q(x_1) = 0, \text{ et } q(x_2) = y_2.$$

A cause des deux premières conditions, on doit avoir

$$q(x) = a(x - x_0)(x - x_1),$$

où a est une constante car on cherche un polynôme de degré au plus 2. Il vient

$$y_0 + (x_2 - x_0)\frac{y_1 - y_0}{x_1 - x_0} + a(x_2 - x_0)(x_2 - x_1) = y_2,$$

$$a = \frac{y_2 - y_0}{(x_2 - x_0)(x_2 - x_1)} - \frac{y_1 - y_0}{(x_1 - x_0)(x_2 - x_1)},$$

ce qui fournit le polynôme q et donc le polynôme p.

Démontrons maintenant le théorème.

Preuve :

- **Unicité :** Supposons qu'il existe deux polynômes p_1 et p_2 de degré inférieur ou égal à n répondant à la question et considérons le polynôme $p = p_1 - p_2$. Celui vérifie :
  - p est de degré au plus n.
  - $p(x_i) = p_1(x_i) - p_2(x_i) = y_i - y_i = 0$ pour tout $0 \leq i \leq n$.

  p est donc un polynôme de degré au plus n qui a *au moins* $n+1$ racines. Ceci n'est possible que si p est le polynôme nul, ce qui implique $p_1 = p_2$.
- **Existence :** On introduit les polynômes de Lagrange élementaires

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

Ces polynômes, sont de degré exactement n et vérifient

$$L_i(x_k) = \delta_{ik}.$$

On vérifie alors aisément que le polynôme

$$p(x) = \sum_{i=0}^{n} y_i L_i(x),$$

est bien un polynôme d'interpolation comme défini ci-dessus.

■

On peut donner une démonstration plus algébrique de ce résultat.

Preuve (via l'algèbre linéaire):

Si on note $\mathbb{R}_n[X]$ l'ensemble des polynômes réels de degré inférieur ou égal à n, montrer le théorème revient à démontrer que l'application

$$\Phi_n : p \in \mathbb{R}_n[X] \mapsto \begin{pmatrix} p(x_0) \\ p(x_1) \\ \vdots \\ p(x_n) \end{pmatrix} \in \mathbb{R}^{n+1}$$

est bijective. Or, cette application est clairement linéaire entre deux espaces de même dimension. En conséquence de quoi il suffit de montrer que Φ_n est injective.

Ceci est vrai car si $\Phi_n(p) = 0$, cela implique que p s'annulle en au moins $n + 1$ points distincts et donc que le polynôme p est nul. ■

N.B. : La matrice de l'application Φ_n dans les bases canoniques de $\mathbb{R}_n[X]$ (1, X, ..., X^n) et de $\mathbb{R}^{n+1}$ est donnée par

$$V_n = \begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & & & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix}.$$

Cette matrice est connue sous le nom de **matrice de Vandermonde** et le calcul de son déterminant est un grand classique.

Lemme III.1.3

Le déterminant de V_n vaut

$$\det(V_n) = \Pi_{0\leq j<i\leq n}(x_i - x_j),$$

en particulier il est nul si et seulement si les $(x_i)_i$ sont tous distincts.

Preuve :

De nombreuses démonstrations de ce résultat existent. Nous en proposons une mais vous pouvez aussi consulter par exemple [Dem91, page 22]. L'idée est de considérer V_n comme une fonction de la variable x_0 (les autres variables étant fixées). On constate immédiatement qu'il s'agit d'un polynôme de degré n (car le déterminant est une application multilinéaire) et que ce polynôme s'annule en les n points distincts x_1, x_2, ..., x_n (car alors deux lignes de la matrice sont égales !). Il existe donc un nombre α_n ne dépendant que de $x_1, ..., x_n$ tel que

$$V_n = \alpha_n(x_0 - x_1)(x_0 - x_2)\dots(x_0 - x_n).$$

Le nombre α_n est le coefficient dominant du polynôme V_n. On voit donc en développant le déterminant suivant la première ligne que ce coefficient vaut

$$\begin{vmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & & & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{vmatrix},$$

qui n'est rien d'autre que le déterminant de Vandermonde V_{n-1} associé aux n points $x_1, x_2, \dots, x_n$. Le résultat suit par une récurrence immédiate. ■

1.2 Calcul du polynôme de Lagrange

La formule explicite obtenue dans la démonstration du théorème permet, en théorie, de calculer le polynôme de lagrange. Ceci étant dit, ces formules coutent $O(n^2)$ opérations à chaque évaluation du polynôme de Lagrange. C'est la raison pour laquelle on préfère utiliser d'autres formules, que l'on appelle les formules de Newton, ou formules des différences divisées.

Définition III.1.4 (Différences divisées)

Soit $(x_i, y_i)_{0\leq i\leq n}$ une famille de points. On définit alors les différences divisées, par récurrence, de la façon suivante

$$\delta^0 y[x_i] = y_i, \ \ \forall 0 \leq i \leq n,$$

$$\delta^1 y[x_i, x_{i+1}] = \frac{\delta^0 y[x_{i+1}] - \delta^0 y[x_i]}{x_{i+1} - x_i}, \ \ \forall 0 \leq i \leq n-1,$$

$$\delta^2 y[x_i, x_{i+1}, x_{i+2}] = \frac{\delta^1 y[x_{i+1}, x_{i+2}] - \delta^1 y[x_{i+1}, x_i]}{x_{i+2} - x_i}, \ \ \forall 0 \leq i \leq n-2,$$

$$\delta^k y[x_i, \dots, x_{i+k}] = \frac{\delta^{k-1} y[x_{i+1}, \dots, x_{i+k}] - \delta^{k-1} y[x_i, \dots, x_{i+k-1}]}{x_{i+k} - x_i}, \ \ \forall 0 \leq i \leq n-k.$$

Bien que ces formules aient l'air compliqué, elles sont faciles à mettre en oeuvre comme on le verra dans un exemple un peu plus loin.

Théorème III.1.5

Le polynôme d'interpolation de Lagrange est donné par

$$p(x) = \delta^0 y[x_0] + (x - x_0)\delta^1 y[x_0, x_1] + (x - x_0)(x - x_1)\delta^2 y[x_0, x_1, x_2] + \dots + (x - x_0)\dots(x - x_n)\delta^n y[x_0, \dots, x_n].$$

Preuve :

On procède par récurrence. Les cas $n \in \{0, 1, 2\}$ ont déjà été traités en exemple (**exercice**).

On suppose que le résultat est vrai au rang $n-1$, de sorte que le polynôme de Lagrange pour les n premiers points $(x_i, y_i)_{0\leq i n-1}$ est donné par

$$\tilde{p}(x) = \delta^0 y[x_0] + (x-x_0)\delta^1 y[x_0,x_1] + (x-x_0)(x-x_1)\delta^2 y[x_0,x_1,x_2] + \cdots \\ + (x-x_0)\dots(x-x_{n-1})\delta^{n-1} y[x_0,\dots,x_{n-1}].$$

On cherche le polynôme p sous la forme

$$p(x) = \tilde{p}(x) + \tilde{q}(x),$$

avec $\tilde{q}$ de degré au plus n. D'après les propriétés de $\tilde{p}$, on voit que $\tilde{q}$ s'annulle forcément aux points $x_0,\dots,x_{n-1}$, il est donc de la forme

$$\tilde{q}(x) = \tilde{a}(x-x_0)\cdots(x-x_{n-1}),$$

où $\tilde{a}$ est une constante pour cause de degré. Il nous reste à déterminer cette constante.

Pour cela, on utilise à nouveau l'hypothèse de récurrence et on appelle $\bar{p}$ le polynôme d'interpolation de Lagrange sur les n derniers points $(x_i, y_i)_{1\leq i\leq n}$ qui est donné par

$$\bar{p}(x) = \delta^0 y[x_1] + (x-x_1)\delta^1 y[x_1,x_2] + (x-x_1)(x-x_2)\delta^2 y[x_1,x_2,x_3] + \cdots \\ + (x-x_1)\dots(x-x_n)\delta^{n-1} y[x_1,\dots,x_n].$$

Pour les mêmes raisons que tout à l'heure on a

$$p(x) = \bar{p}(x) + \bar{q}(x),$$

où maintenant on a

$$\bar{q}(x) = \bar{a}(x-x_1)\cdots(x-x_n).$$

Au final on a montré

$$p(x) = \tilde{p}(x) + \tilde{a}(x-x_0)\cdots(x-x_{n-1}) = \bar{p}(x) + \bar{a}(x-x_1)\cdots(x-x_n).$$

Si on identifie les termes de degré n dans ces formules, on trouve que $\tilde{a} = \bar{a}$, qu'on notera a dorénavant.

On soustrait les deux membres de l'égalité et on trouve

$$\tilde{p}(x) - \bar{p}(x) = a(x-x_1)\cdots(x-x_{n-1})\big[(x-x_n) - (x-x_0)\big],$$

d'où

$$\frac{\tilde{p}(x) - \bar{p}(x)}{x_0 - x_n} = a(x-x_1)\cdots(x-x_{n-1}).$$

En identifiant les coefficients dominants on trouve

$$a = \frac{\delta^{n-1}[y](x_0,\dots,x_{n-1}) - \delta^{n-1}[y](x_1,\dots,x_n)}{x_0 - x_n},$$

ceci donne le résultat. ■

Remarques

- Si on souhaite rajouter un point d'interpolation : on doit seulement calculer n coefficients supplémentaires (i.e. rajouter une "ligne") au tableau. Ceci va nous couter $O(n)$ opérations.
- Si on applique bêtement la formule donnée par le théorème précédent, en supposant avoir déjà calculé les différences divisées, on voit qu'il faudra effectuer de l'ordre de $O(n^2)$ opérations.
  Une façon plus économique de faire le calcul est d'utiliser l'algorithme de Hörner qui consiste à récrire la formule sous la forme

  $$p(x) = \delta^0 y[x_0] + (x-x_0)\times(\delta^1 y[x_0,x_1] + (x-x_1)\times(\delta^2 y[x_0,x_1,x_2] \\ + \cdots + (x-x_n)\times\delta^n y[x_0,\cdots,x_n])),$$

  qui ne nécessite que $O(n)$ opérations et seulement n multiplications.
- L'ordre dans lequel on prend les points (x_i, y_i) n'a aucune influence sur le polynôme de Lagrange, en conséquence la valeur de la différence divisée $\delta^n y[x_0,\cdots,x_n]$ ne dépend pas de l'ordre des variables dans le calcul.

Exemple Nous allons calculer le polynôme de Lagrange dans l'exemple suivant

x_i	$y_i = \delta^0 y$	$\delta^1 y$	$\delta^2 y$	$\delta^3 y$	$\delta^4 y$	$\delta^5 y$
0	-1					
		1				
2	1		$\frac{3}{8}$			
		$\frac{5}{2}$		$-\frac{77}{120}$		
4	6		$-\frac{17}{6}$		$\frac{167}{960}$	
		-6		$\frac{3}{4}$		$-\frac{287}{9600}$
5	0		$\frac{5}{3}$		$-\frac{1}{8}$	
		$\frac{2}{3}$		$-\frac{1}{4}$		
8	2		$\frac{1}{6}$			
		$\frac{3}{2}$				
10	5					

Le polynôme obtenu est alors

$$p(x) = -1 + x + x(x-2)\frac{3}{8} - x(x-2)(x-4)\frac{77}{120} + x(x-2)(x-4)(x-5)\frac{167}{960} - x(x-2)(x-4)(x-5)(x-8)\frac{287}{9600}.$$

En formulation de Hörner on trouve

$$p(x) = -1 + x\left(1 + (x-2)\left(\frac{3}{8} + (x-4)\left(-\frac{77}{120} + (x-5)\left(\frac{167}{960} - (x-8)\frac{287}{9600}\right)\right)\right)\right).$$

Remarque III.1.6

La méthode des différences divisées de Newton peut-être vue comme un solveur en O(n^2) opérations pour la matrice de Vandermonde (à comparer aux O(n^3) itérations nécessaires pour la méthode du pivot de Gauss (i.e. pour la méthode LU).
Plus précisément, le calcul des différences divisées elles-mêmes prend O(n^2) opérations. Ensuite, il faut trouver les coefficients du polynôme p à partir de la formule de Hörner.

1.3 Estimation de l'erreur d'approximation

Etant donné une famille de points $(x_i, y_i)_i$ on a vu comment construire le polynôme de plus bas degré passant par ces points.

Supposons maintenant que les points y_i soient en fait les valeurs d'une fonction f (connue ou inconnue) aux points x_i : $y_i = f(x_i)$. La question est : si on utilise le polynôme d'interpolation de Lagrange pour obtenir des valeurs approchées de la fonction f en d'autres points que les poins x_i, quelle confiance peut-on avoir dans les résultats ?

Dans toute la suite, on va supposer (juste pour simplifier) que les x_i sont ordonnés $x_0 < \cdots < x_n$.

Commençons par un théorème classique du calcul différentiel :

Théorème III.1.7 (de Rolle)

Soit f une fonction continue sur $[a,b]$ et dérivable dans $]a,b[$. Si $f(a) = f(b)$ alors il existe un $\xi \in]a,b[$ tel que $f'(\xi) = 0$.

Preuve :

Si f est une fonction constante, le résultat est acquis. Sinon, il existe un point dans $]a,b[$ où f atteint son maximum global ou son minimum global. En ce point, sa dérivée est nécessairement nulle (savoir le démontrer !). ■

On peut déduire de ce théorème une version plus générale qui est la suivante :

Théorème III.1.8 (Rolle généralisé)

Soit $a = x_1 < ... < x_k = b$, $k \geq 2$ points distincts formant une subdivision de $[a,b]$ et $\alpha_1, ..., \alpha_k$ des entiers naturels non nuls. On pose $n = \left(\sum_{i=1}^{k} \alpha_i\right) - 1$.
Soit f une fonction de classe $\mathcal{C}^n$ sur $[a,b]$ qui s'annule à l'ordre α_i au point x_i pour tout $i \in \{1, ..., k\}$, alors il existe $\xi \in]a,b[$ tel que

$$f^{(n)}(\xi) = 0.$$

Preuve :

On procède par récurrence sur n.

- Pour $n = 1$: on a nécessairement $k = 2$ et $\alpha_1 = \alpha_2 = 1$.
- Supposons le résultat vrai au rang n et montrons-le au rang $n+1$. D'après le théorème de Rolle appliqué à f entre chaque couple de points x_i, x_{i+1}, il existe $k-1$ points y_i, $i = 1, ..., k-1$, vérifiant

$$x_i < y_i < x_{i+1}, \ \ \forall i = 1, ..., k-1,$$

et tels que $f'(y_i) = 0$. Ainsi la fonction f' s'annule à l'ordre $\alpha_i - 1$ en tous les points x_i (si $\alpha_i - 1 = 0$ on retire ce point de la subdivision et elle s'annule à l'ordre 1 (au moins !) en les points y_i. Le total des ordres d'annulation de f' est donc :

$$\sum_{i=1}^{k}(\alpha_i - 1) + (k-1) = \sum_{i=1}^{k}\alpha_i - 1 = n,$$

ce qui montre que f' vérifie l'hypothèse de récurrence au rang précédent et donc qu'il existe un $\xi \in]a, b[$ tel que $(f')^{(n)}(\xi) = 0$. Il est clair que ce ξ répond à la question.

■

Noter que ce résultat implique immédiatement la formule de Taylor-Lagrange en utilisant une preuve similaire à la preuve suivante.

Théorème III.1.9 (Erreur d'approximation du polynôme de Lagrange)

Si $f : [a, b] \mapsto \mathbb{R}$ est $n+1$ fois dérivable et p est le polynôme d'interpolation de Lagrange de f aux points $(x_i)_{0\leq i\leq n} \subset [a, b]$, alors pour tout $y \in [a, b]$, il existe $\xi \in]a, b[$ tel que

$$f(y) - p(y) = (y - x_0)\cdots(y - x_n)\frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Preuve :

Fixons $y \in [a, b]$ et supposons que y n'est pas l'un des x_i, sinon il n'y a rien à démontrer.

Considérons la fonction

$$x \mapsto \Psi(x) = p(x) - f(x) + M(x - x_0)\cdots(x - x_n),$$

où M est un réel à déterminer.

Justement, choisissons M pour que la fonction Ψ s'annulle au point y. Pour cela, il faut

$$p(y) - f(y) + M(y - x_0)\cdots(y - x_n) = 0, \tag{III.1}$$

ce qui donne

$$M = \frac{f(y) - p(y)}{(y - x_0)\cdots(y - x_n)},$$

la division étant permise car y n'est pas égal à l'un de x_i.

Ce réel M étant fixé, regardons la fonction Ψ. Il s'agit d'une fonction qui s'annulle en $n+2$ points distincts : les (x_i) et y. Par applications successives du théorème de Rolle, on en déduit que la dérivée $n+1$-ième de Ψ doit nécessairement s'annuller dans l'intervalle $[a, b]$.

Il existe donc $\xi \in]a, b[$ tel que $\Psi^{(n+1)}(\xi) = 0$. Comme le polynôme p est de degré au plus n, sa dérivée $n+1$-ième est nulle, il reste

$$-f^{(n+1)}(\xi) + M(n+1)! = 0.$$

En remplaçant cette nouvelle valeur de M dans la formule (III.1), on obtient le résultat. ■

Approximation de la dérivée

Théorème III.1.10

Avec les mêmes notations que dans le théorème III.1.9, nous avons

$$\sup_{y\in[a,b]} |f'(y) - p_n'(y)| \leq \frac{(b-a)^n}{n!}\sup_{[a,b]} |f^{(n+1)}|.$$

Preuve :

D'après le théorème de Rolle, il existe n points distincts $\tilde{x}_0, \ldots, \tilde{x}_{n-1}$ tels que $f'(\tilde{x}_i) = p_n'(\tilde{x}_i)$, ce qui prouve que p_n' n'est autre que le polynôme d'interpolation de Lagrange de la fonction f' aux points $\tilde{x}_i$. En appliquant le théorème précédent, on obtient le résultat. ■

Applications

Corollaire III.1.11

Soit $[a,b]$ comme dans le théorème III.1.9 et f de classe $\mathcal{C}^\infty$. Supposons qu'il existe $M > 0$ telle que $\sup_{\mathbb{R}} |f^{(k)}| \leq M$ pour tout k.
Pour tout $n \in \mathbb{N}$, on choisit une famille de $n+1$ points d'interpolation $(x_i^n)_{0\leq i\leq n}$ dans $[a,b]$ et on note p_n le polynôme d'interpolation de Lagrange de f en ces points.
Alors, la suite de fonctions $(p_n)_n$ converge uniformément vers f sur $[a,b]$.

Preuve :
D'après le théorème III.1.9, on a

$$\sup_{y\in[a,b]} |f(y) - p_n(y)| \leq \frac{M}{(n+1)!} \sup_{y\in[a,b]} |(y-x_0^n)\cdots(y-x_n^n)| \leq \frac{M(b-a)^{n+1}}{(n+1)!},$$

et ceci tend vers 0. ■

Corollaire III.1.12

Si dans le corollaire précédent on suppose plutôt que f est une fonction développable en série entière autour du centre $c = (a+b)/2$ de l'intervalle et de rayon de convergence $R > 3(b-a)/2$, alors le résultat précédent demeure.

Ce résultat nécessite donc un rayon de convergence très supérieur au diamètre de l'intervalle considéré.
Preuve ([Dem91, p.31], [HH06]):
On se donne un r tel que $(b-a)/2 < r < R$ et on se donne le DSE de f sous la forme

$$f(x) = \sum_{k\geq 0} a_k (x-c)^k.$$

Par choix de r, on sait que la suite $(a_k r^k)$ est bornée. On note M une borne de cette suite de sorte que l'on a $|a_k| \leq C/r^k$ pour tout $k \geq 0$.
De plus, on peut dériver f terme à terme à l'intérieur du disque de convergence, donc on trouve

$$f^{(n)}(x) = \sum_{k\geq 0} a_k \frac{d^n}{dx^n}((x-c)^k) = \sum_{k\geq n} a_k \frac{k!}{(k-n)!}(x-c)^{k-n}.$$

Comme pour $x \in [a,b]$ on a $|x-c| \leq (b-a)/2$, on en déduit

$$\|f^{(n)}\|_{L^\infty([a,b])} \leq Cr^{-n} \sum_{k\geq n} \frac{k!}{(k-n)!} \left(\frac{b-a}{2r}\right)^{k-n} = \frac{Cn!r^{-n}}{\left(1-\frac{b-a}{2r}\right)^n}.$$

L'erreur d'interpolation est donc majorée par

$$\|f-p_n\|_{L^\infty} \leq \frac{(b-a)^n}{(n+1)!}\|f^{(n+1)}\|_{L^\infty} \leq \frac{C(b-a)^n r^{-n}}{\left(1-\frac{b-a}{2r}\right)^{n+1}} = \frac{Cr(b-a)^n}{\left(r-\frac{b-a}{2}\right)^{n+1}}.$$

Cette quantité tend vers 0 dès lors que $r - \frac{b-a}{2} > b-a$. Par hypothèse sur R, on peut choisir un tel r au début de la démonstration et le résultat est donc démontré. ■

Phénomène de Runge : Les exemples précédents ne sont en réalité pas génériques car on peut montrer que, en général, la suite des interpolés ne converge pas vers f uniformément sur l'intervalle de travail considéré. L'exemple classique est celui de la fonction $x \mapsto 1/(x^2+\alpha^2)$ sur $[-1,1]$ qui est étudié en grand détail dans [Dem91, p. 36]. Le rayon de convergence du DSE de cette fonction autour de 0 est exactement égal à α. Le théorème précédent dit que si $\alpha > 3$, alors on a convergence uniforme de $(p_n)_n$ vers cette fonction sur $[-1,1]$. Dans la référence ci-dessus, il est démontré que si α est petit, alors la suite des (p_n) ne converge pas, même simplement, sur l'intervalle $[-1,1]$.

1.4 L'opérateur d'interpolation. Choix des points d'interpolation

Etant donné un n et un choix de points d'interpolation dans $[a,b]$, on considère l'opérateur $I_n : f \in C^0([a,b]) \mapsto I_n f = p_n \in \mathbb{R}_n[X]$ qui à une fonction fait correspondre son polynôme d'interpolation de Lagrange. **Attention :** malgré la notation cet opérateur dépend non seulement de n mais aussi du choix des points d'interpolation.

Dans la suite, on identifie un polynôme à sa fonction-polynôme associée et on munit $C^0([a,b])$ et $\mathbb{R}_n[X]$ de la norme uniforme $\|.\|_{L^\infty([a,b])}$.

Résumons les propriétés de base de cet opérateur : I_n est linéaire, continu et sa norme (pour les topologies considérées ci-dessus) vaut

$$\Lambda_n \stackrel{\text{def}}{=} \|I_n\| = \|\lambda_n\|_\infty, \tag{III.2}$$

où $\lambda_n(x) = \sum_{i=0}^n |L_i(x)|$. Le nombre Λ_n est appelé la constante de Lebesgue associée au choix de points de discrétisation effectué au début.

Exercice III.1.1

Démontrer la formule (III.2).

On va voir que ce nombre n'est jamais borné quand n tend vers l'infini. Je donne ici quelques exemples de résultats sans démonstration :

– Si on choisit les points d'interpolation équi-distants

$$x_i^n = a + \frac{(b-a)i}{n}, \ \ 0 \leq i \leq n,$$

on a l'estimation

$$\Lambda_n \underset{+\infty}{\sim} \frac{2^{n+1}}{en\log(n)}.$$

Ce résultat est difficile mais vous pouvez trouver dans [Dem91, p. 47] la preuve d'une minoration moins précise mais suffisante pour nos besoins

$$\Lambda_n \geq \frac{2^n}{4n^2}.$$

– Si on choisit les points d'interpolation de Tchebychev

$$x_i^n = \cos\left(\frac{2i+1}{2n+2}\pi\right), \ \ 0 \leq i \leq n,$$

alors on a

$$\Lambda_n \underset{+\infty}{\sim} \frac{2}{\pi}\log(n).$$

La preuve de la borne inférieure est facile [Dem91, p. 50], celle de la borne supérieure est plus difficile [Dem91, p. 48]. Voir également [CLF95, Exo 22-6, p. 172]

Si par ailleurs, pour tout $n \geq 1$, on appelle $\overline{\Lambda}_n$ la meilleure constante de Lebesgue parmi tous les choix possibles de points d'interpolation (c'est-à-dire l'infimum de Λ_n pour tout choix de ces points) on peut montrer (voir par exemple [CM84, Exercice 1.6, p 31], [CM86, Exercice 1.4.3, p 17 et son corrigé p. 37] pour un résultat analogue bien qu'un peu plus faible) que

$$\overline{\Lambda}_n \underset{+\infty}{\sim} \frac{2}{\pi}\log(n).$$

Ainsi les points de Tchebychev sont essentiellement optimaux du point de vue de la constante de Lebesgue.

En tout état de cause, pout tout choix des points d'interpolation, on a $\lim_{n\to\infty}\Lambda_n = +\infty$, ce qui implique par le **théorème de Banach-Steinhaus** qu'il existe toujours une fonction continue f telle que $I_n(f)$ ne soit pas bornée (en particulier ne converge pas vers f !).

2 L'interpolation de Hermite

On se demande si, dans le même esprit que ce qui précède, on peut trouver un polynôme de degré le plus faible possible qui satisfasse des conditions du type $p(x_i) = y_i$ mais aussi des conditions sur la dérivée $p'(x_i) = z_i$.

La réponse est oui en toute généralité, c'est l'interpolation de Hermite, mais on ne va traiter ici qu'un exemple simple.

Théorème III.2.13

Soient x_0, x_1 deux réels distincts. Soient y_0, y_1, z_0, z_1 quatre réels quelconques.
Il existe un unique polynôme p de degré 3 qui vérifie

$$p(x_0) = y_0, \;\; p(x_1) = y_1,$$

$$p'(x_0) = z_0, \;\; p'(x_1) = z_1.$$

De plus, on a l'estimation

$$\sup_{x\in[x_0,x_1]} |p(x)| \leq 3(|y_0| + |y_1|) + |x_1 - x_0|(|z_0| + |z_1|). \tag{III.3}$$

Preuve :

Considérons pour commencer le polynôme d'interpolation de Lagrange π (de degré 1) associé aux points (x_0, y_0) et (x_1, y_1). Celui-ci s'écrit

$$\pi(x) = y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0}.$$

Le polynôme p que l'on cherche doit donc s'écrire sous la forme

$$p(x) = \pi(x) + (x - x_0)(x - x_1)q(x),$$

où q est degré au plus 1. On va chercher q sous la forme

$$q(x) = a(x - x_0) + b(x - x_1).$$

Le calcul de p' se déroule ainsi :

$$p'(x) = \pi'(x) + (2x - x_0 - x_1)q(x) + (x - x_0)(x - x_1)q'(x),$$

$$p'(x) = \underbrace{\frac{y_1 - y_0}{x_1 - x_0}}_{=\delta^1 y[x_0,x_1]} + (2x - x_0 - x_1)\big[a(x - x_0) + b(x - x_1)\big] + (a + b)(x - x_0)(x - x_1).$$

On obtient

$$p'(x_0) = \delta^1 y[x_0, x_1] + b(x_1 - x_0)^2,$$

$$p'(x_1) = \delta^1 y[x_0, x_1] + a(x_1 - x_0)^2.$$

Les conditions cherchées sont donc équivalentes à

$$z_0 = \delta^1 y[x_0, x_1] + b(x_1 - x_0)^2,$$

$$z_1 = \delta^1 y[x_0, x_1] + a(x_1 - x_0)^2.$$

Dont la solution est donnée par

$$a = \frac{1}{(x_1 - x_0)^2}\left(z_1 - \delta^1 y[x_0, x_1]\right),$$

$$b = \frac{1}{(x_1 - x_0)^2}\left(z_0 - \delta^1 y[x_0, x_1]\right).$$

Ceci définit donc de façon unique le polynôme p. L'estimation proposée suit immédiatement par majoration des différents

termes. ■

Remarque III.2.14

Comme on le verra par la suite, ce qui est important dans l'estimation (III.3)*, c'est la dépendance du second membre par rapport à la taille de l'intervalle. Il existe une méthode assez générale (dite "par homogénéité") pour établir ce type d'estimations et qui évite de recourir au calcul explicite du polynôme p.*

– *On commence par traiter le cas de l'intervalle $[0,1]$: on sait que l'application qui à p associe $(p(0),p(1),p'(0),p'(1))$ est linéaire, bijective. Son inverse est donc également linéaire et continue, ce qui prouve l'existence d'un nombre universel $C>0$ tel que*

$$\sup_{x\in[0,1]}|p(x)|\leq C(|p(0)|+|p(1)|+|p'(0)|+|p'(1)|),\ \ \forall p\in\mathbb{R}_3[X]. \tag{III.4}$$

– *Si maintenant on considère l'intervalle $[x_0,x_1]$ et un polynôme $q\in\mathbb{R}_3[X]$ quelconque, on applique* (III.4) *au polynôme $t\mapsto p(t)=q(x_0+t(x_1-x_0))$, ce qui donne après calcul*

$$\sup_{x\in[x_0,x_1]}|q(x)|\leq C(|q(x_0)|+|q(x_1)|)+C|x_1-x_0|(|q'(x_0)|+|q'(x_1)|).$$

Calculons pour un usage ultérieur, la valeur de la dérivée seconde de p aux deux points d'interpolation

$$p''(x)=2q(x)+2(2x-x_0-x_1)(a+b),$$

d'où

$$p''(x_0)=2b(x_0-x_1)+2(x_0-x_1)(a+b)=2(x_0-x_1)(a+2b)=\frac{-2}{x_1-x_0}\left(z_1+2z_0-3\delta^1y[x_0,x_1]\right),$$

$$p''(x_1)=2a(x_1-x_0)+2(x_1-x_0)(a+b)=2(x_1-x_0)(2a+b)=\frac{2}{x_1-x_0}\left(2z_1+z_0-3\delta^1y[x_0,x_1]\right).$$

Estimation de l'erreur

Théorème III.2.15

Soit f une fonction de classe $\mathcal{C}^4$ sur un intervalle $[a,b]$ et p le polynôme de Hermite (de degré inférieur ou égal à 3) vérifiant

$$p(a)=f(a),p'(a)=f'(a),p(b)=f(b),\ \text{et}\ p'(b)=f'(b).$$

On a alors

$$\forall x\in[a,b],\exists\xi\in]a,b[,\ \ f(x)-p(x)=(x-a)^2(x-b)^2\frac{f^{(4)}(\xi)}{4!}.$$

En particulier,

$$\|p-f\|_\infty\leq\frac{|b-a|^4}{384}\|f^{(4)}\|_\infty.$$

Preuve :

Si $x=a$ ou $x=b$ il n'y a rien à démontrer. Pour $x\in]a,b[$, on introduit le nombre M tel que

$$f(x)-p(x)=M(x-a)^2(x-b)^2.$$

On introduit alors la fonction $\psi:t\mapsto\psi(t)=f(t)-p(t)-M(t-a)^2(t-b)^2$. Par définition de p et de M, on a

$$\psi(a)=\psi'(a)=\psi(b)=\psi'(b)=\psi(x)=0.$$

Par application répétée du théorème de Rolle, on montre qu'il existe un point $\xi\in]a,b[$ tel que $\psi^{(4)}(\xi)=0$. Comme $p^{(4)}(\xi)=0$, on trouve que

$$f^{(4)}(\xi)=4!M,$$

d'où le résultat. ■

Cas général de l'interpolation de Hermite Dans le cas général, l'existence et l'unicité d'un polynôme d'interpolation se démontre par un argument d'algèbre linéaire.

Théorème III.2.16

Soient $n \geq 1, k \geq 1$ et $x_1 < \ldots < x_k$ des points distincts de $\mathbb{R}$ et $\alpha_1, \ldots, \alpha_k \in \mathbb{N}^*$ des entiers strictement positifs tels que $\sum_{i=1}^{k} \alpha_i = n+1$.
Pour tout $i \in \{1, \ldots, k\}$ et tout $l \in \{1, \ldots, \alpha_i\}$, on se donne des nombres $y_i^l \in \mathbb{R}$. Il existe alors un unique polynôme de degré au plus égal à n, $p \in \mathbb{R}_n[X]$, tel que

$$\forall i \in \{1, \ldots, k\}, \ \ \forall l \in \{1, \ldots, \alpha_i\}, \ \ p^{(l-1)}(x_i) = y_i^l.$$

Preuve :

Il suffit de considérer l'application linéaire $\Phi : p \in \mathbb{R}_n[X] \mapsto \mathbb{R}^{\alpha_1} \times \cdots \times \mathbb{R}^{\alpha_k}$ qui a tout polynôme p associe l'ensemble des valeurs de p et de ses dérivées successives jusqu'à l'ordre $\alpha_i - 1$ au point x_i.

Comme les deux espaces vectoriels mis en jeu sont de même dimension (par hypothèse sur les α_i), il suffit de vérifier que Φ est injective, ce qui est immédiat en comptant les racines et en utilisant le théorème de Rolle. ■

Remarque III.2.17

Si $k = 1$, i.e. si nous avons un seul point d'interpolation, alors le polynôme de Hermite d'ordre n coïncide avec le polynôme de Taylor de la fonction considérée en ce point.

Pour une méthode de calcul du polynôme d'interpolation de Hermite basée sur la division euclidienne vous pouvez par exemple consulter [Dem91, Exercice 6.3, p. 57].

En général, le calcul d'une base de polynômes d'interpolation de Hermite (de la même façon que les l_i pour l'interpolation de Lagrange) est assez complexe. Dans le cas où tous les α_i sont égaux à 2, on peut montrer que le polynôme de Hermite s'écrit

$$p(x) = \sum_{i=1}^{k} y_i^1 H_i^1(x) + \sum_{i=1}^{k} y_i^2 H_i^2(x),$$

où

$$H_i^1(x) = (1 - 2l_i'(x_i)(x - x_i))(l_i(x))^2,$$

$$H_i^2(x) = (x - x_i)(l_i(x))^2.$$

Voir par exemple [Rom96, Pb. 33, p151].

Remarque Le choix des conditions d'interpolation ne peut pas être fait au hasard. Par exemple, on peut montrer (voir [QSS07, p. 289]) qu'il n'existe pas de polynôme de degré inférieur ou égal à 3 vérifiant les conditions suivantes

$$p(-1) = 1, p'(-1) = 1, p'(1) = 2, p(2) = 1.$$

En effet, si on cherche $p(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$, on voit que les coefficients de p doivent vérifier les équations suivantes

$$\begin{pmatrix} 1 & -1 & 1 & -1 \\ 0 & 1 & -2 & 3 \\ 0 & 1 & 2 & 3 \\ 1 & 2 & 4 & 8 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 2 \\ 1 \end{pmatrix}.$$

La matrice de ce système est singulière, et on peut vérifier qu'il n'a pas de solution.

3 Interpolation polynômiale par morceaux. Splines cubiques

On a vu que l'interpolation sur un nombre de plus en plus important de points ne permettait pas nécessairement d'obtenir une meilleure approximation. De plus, le résultat obtenu est très sensible aux erreurs d'arrondis, i.e. très instable.

On se propose ici, plutôt que d'interpoler par une fonction globalement polynômiale, d'interpoler par une fonction polynômiale par morceaux.

3.1 Interpolation constante par morceaux

Théorème III.3.18

Pour toute famille de points $(x_i, y_i)_i$, il existe une unique fonction π_n, constante sur chaque intervalle $[x_{i-1}, x_i[$ et telle que $\pi_n(x_i) = y_i$ pour tout i.
Si les valeurs y_i sont telles que $y_i = f(x_i)$, où f est une fonction de classe $\mathcal{C}^1$ donnée, alors on a l'estimation

$$\sup_{x\in[x_0,x_n]} |\pi_n(x) - f(x)| \leq h \sup_{[x_0,x_n]} |f'|,$$

où $h = \max_i |x_{i+1} - x_i|$

Preuve :
Il s'agit d'une simple utilisation de l'inégalité des accroissements finis sur chaque intervalle $[x_i, x_{i+1}]$. ■

3.2 Interpolation affine par morceaux

Théorème III.3.19

Pour toute famille de points $(x_i, y_i)_i$, il existe une unique fonction $\tilde{\pi}_n$, affine sur chaque intervalle $[x_{i-1}, x_i]$ et telle que $\tilde{\pi}_n(x_i) = y_i$ pour tout i.
Si les valeurs y_i sont telles que $y_i = f(x_i)$, où f est une fonction de classe $\mathcal{C}^2$ donnée, alors on a l'estimation

$$\sup_{x\in[x_0,x_n]} |\tilde{\pi}_n(x) - f(x)| \leq \frac{h^2}{8} \sup_{[x_0,x_n]} |f''|,$$

où $h = \max_i |x_{i+1} - x_i|$

Preuve :
Pour x donné, on détermine un indice i tel que $x \in [x_i, x_{i+1}]$, on applique alors le théorème III.1.9 d'estimation de l'erreur d'approximation de l'interpolation de Lagrange sur l'intervalle $[x_i, x_{i+1}]$ donc avec $n = 1$. ■

Cette estimation nous dit que l'erreur d'approximation par une fonction affine par morceaux est proportionnelle au carré du pas de la discrétisation et à la norme infinie de la dérivée seconde de f. On gagne donc un ordre de convergence par rapport à l'approximation affine par morceaux.

L'avantage de cette formule c'est que quand n augmente, i.e. quand h diminue, la borne obtenue ne fait pas intervenir des dérivées de plus en plus élevées de la fonction. Ceci assure la **convergence** de l'approximation vers la fonction f.

L'inconvénient de cette approche est double :
- La qualité de l'approximation n'est pas excellente : la convergence est d'ordre 2 seulement.
- Même si la fonction f de départ à approcher est très régulière, la fonction approchante π n'est plus régulière. Elle n'est même pas dérivable aux points x_i.

3.3 Les splines cubiques

Une amélioration possible de la méthode est de chercher non plus une approximation affine par morceaux mais polynômiale de degré 3 par morceaux. Ceci nous donne plus de libertés sur la fonction interpolante et permet donc de lui imposer plus de contraintes.

Théorème III.3.20

Soit une famille de points $(x_i, y_i)_{0\leq i\leq n}$, et deux valeurs $z_0, z_n \in \mathbb{R}$. Il existe alors une unique fonction $\pi : [x_0, x_n] \mapsto \mathbb{R}$ vérifiant les propriétés suivantes :
- *Interpolation : Pour tout $0 \leq i \leq n$, $\pi(x_i) = y_i$.*
- *Pour tout $0 \leq i \leq n-1$, la restriction de π à l'intervalle $[x_i, x_{i+1}]$ est une fonction polynômiale de degré au plus 3.*
- *Régularité : π est une fonction de classe $\mathcal{C}^2$ sur $[x_0, x_n]$.*
- *Les dérivées de π aux extrémités de l'intervalle sont fixées : $\pi'(x_0) = z_0$, $\pi'(x_n) = z_n$.*

La fonction π est appellée spline cubique qui interpole les points (x_i, y_i) avec dérivées aux bornes z_0 et z_n fixées.

Preuve :

On cherche une fonction π de classe $\mathcal{C}^2$, en particulier elle doit être dérivable en tous les points x_i. On va chercher les valeurs que doivent prendre ces dérivées. Notons z_i pour $0 \leq i \leq n$ la dérivée cherchée de la fonction π au point x_i. Les valeurs z_0 et z_n sont des données du problème, il nous faut déterminer les z_i pour $1 \leq i \leq n-1$.

Comme on veut que π soit polynômiale de degré au plus 3 sur chacun des intervalles, on voit d'après le théorème III.2.13 que π est définie de façon unique par les valeurs y_i et les valeurs des dérivées z_i.

Soit $1 \leq i \leq n-1$. On peut calculer la dérivée seconde de π à droite et la dérivée seconde de π à gauche. On trouve respectivement, d'après le paragraphe précédent :

$$\pi''(x_i^+) = \frac{-2}{x_{i+1}-x_i}\left(z_{i+1} + 2z_i - 3\delta^1 y[x_i, x_{i+1}]\right),$$

$$\pi''(x_i^-) = \frac{2}{x_i - x_{i-1}}\left(2z_i + z_{i-1} - 3\delta^1 y[x_{i-1}, x_i]\right).$$

La question est donc : peut-on trouver des valeurs des dérivées $(z_i)_{1\leq i\leq n-1}$ pour lesquelles la dérivée seconde de π est continue, autrement dit

$$\frac{2}{x_i - x_{i-1}}(2z_i + z_{i-1}) + \frac{2}{x_{i+1}-x_i}(2z_i + z_{i+1}) = G_i, \quad \forall 1 \leq i \leq n-1, \tag{III.5}$$

où G_i est un terme qui ne dépend que des données et qui s'écrit

$$G_i = \frac{6}{x_i - x_{i-1}}\delta^1 y[x_{i-1}, x_i] + \frac{6}{x_{i+1}-x_i}\delta^1 y[x_i, x_{i+1}].$$

Le but est de trouver les pentes z_i, on vient de montrer qu'elle sont solutions d'un certain système linéaire. En effet les équations (III.5) peuvent s'écrire sous la forme $Az = G$ où A est une matrice carrée de taille $n-1$ tridiagonale et G un second membre connu qui s'écrit

$$G = \begin{pmatrix} G_1 - \frac{2}{x_1 - x_0} z_0 \\ G_2 \\ \cdots \\ G_{n-2} \\ G_{n-1} - \frac{2}{x_n - x_{n-1}} z_n \end{pmatrix}.$$

Notez la forme particulière du premier et du dernier terme à cause du fait que z_0 et z_n sont des données du problème.

Dans le cas d'une subdivision uniforme de l'intervalle de pas h, la matrice A s'écrit

$$A = \frac{1}{h}\begin{pmatrix} 8 & 2 & 0 & \cdots & \cdots & \\ 2 & 8 & 2 & 0 & \cdots & 0 \\ 0 & 2 & 8 & 2 & 0 & \vdots \\ \vdots & & & \vdots & & \\ 0 & \cdots & 0 & 2 & 8 & 2 \\ & & & & 2 & 8 \end{pmatrix}.$$

Démontrons que ce système linéaire est inversible. Plus précisément on va montrer le :

Lemme III.3.21

Soit $F \in \mathbb{R}^{n-1}$ donné et soit $Z \in \mathbb{R}^{n-1}$ vérifiant le système $AZ = F$ alors nous avons l'inégalité

$$(\max_i |z_i|) \leq h(\max_i |F_i|),$$

où $h = \max_j |x_{j+1} - x_j|$.
Autrement dit, nous avons $\|A^{-1}\|_\infty \leq h$.

Preuve (du lemme):

L'inversibilité de la matrice découle immédiatement du fait qu'elle soit à diagonale strictement dominante (Cf. [LT00]) Ceci étant dit, la preuve de l'estimation de l'inverse va également impliquer qu'elle est inversible.

Soit l l'indice pour lequel $|z_l| = \max_i |z_i|$.

– Si $l \in \{2, \ldots, n-2\}$, on utilise l'équation (III.5), et on trouve

$$\frac{2}{x_l - x_{l-1}}(2z_l + z_{l-1}) + \frac{2}{x_{l+1}-x_l}(2z_l + z_{l+1}) = F_l.$$

Il vient

$$\frac{2}{x_l - x_{l-1}}(2z_l) + \frac{2}{x_{l+1} - x_l}(2z_l) = F_l - \frac{2}{x_l - x_{l-1}}(z_{l-1}) - \frac{2}{x_{l+1} - x_l}(z_{l+1}),$$

et donc

$$\frac{2}{x_l - x_{l-1}}2|z_l| + \frac{2}{x_{l+1} - x_l}2|z_l| \leq |F_l| + \frac{2}{x_l - x_{l-1}}\underbrace{|z_{l-1}|}_{\leq |z_l|} + \frac{2}{x_{l+1} - x_l}\underbrace{|z_{l+1}|}_{\leq |z_l|},$$

soit

$$2|z_l|\left(\frac{1}{x_l - x_{l-1}} + \frac{1}{x_{l+1} - x_l}\right) \leq |F_l|.$$

Il reste

$$|z_l| \leq \frac{1}{2}\left(\frac{(x_{l+1} - x_l)(x_l - x_{l-1})}{x_{l+1} - x_{l-1}}\right)|F_l| \leq \frac{h}{4}|F_l|,$$

d'où le résultat.

– Si $l = 1$ ou $l = n - 1$, le calcul est similaire avec un terme en moins dans la ligne correspondante de la matrice.

■

Si on applique le lemme avec le second membre F nul, on trouve que $Z = 0$, ce qui montre que la matrice (carrée) du système est injective donc bijective, ce qui prouve l'existence et l'unicité de la spline cubique recherchée. ■

On peut alors maintenant le résultat suivant en utilisant nos connaissances sur l'interpolation de Hermite.

Théorème III.3.22

Si f est une fonction de classe $\mathcal{C}^4$, $(x_i)_{0\leq i\leq n}$ une discrétisation de pas $h = \max_i |x_{i+1} - x_i|$, π la spline cubique naturelle interpolant f aux points $(x_i)_i$ et dont les dérivées aux bornes sont données par

$$\pi'(x_0) = f'(x_0),\ \pi'(x_n) = f'(x_n).$$

Alors on a l'estimation d'erreur suivante :

$$\sup_{x\in[x_0,x_n]} |\pi(x) - f(x)| \leq Ch^4 \sup_{[x_0,x_n]} |f^{(4)}|,$$

où C est une constante universelle.

Preuve :

L'idée est la suivante : on sait déjà que π prend les mêmes valeurs que f aux noeuds, par contre *a priori* il n'y a aucune raison que π' et f' coincident aux noeuds. Si c'était le cas, il suffirait d'appliquer l'estimation d'erreur de l'interpolation de Hermite.

On va donc commencer par estimer la différence entre $\pi'(x_i)$ et $f'(x_i)$. Avec les notations du théorèmre précédent, on veut donc estimer les quantités $|z_i - f'(x_i)|$. Il s'agit d'une estimation de consistance très semblable à ce que l'on fait pour les schémas pour les EDO ou les EDP.

On pose $\bar{z}_i = f'(x_i)$ et on introduit le vecteur $E = (z_i - \bar{z}_i)_{1\leq i\leq n-1}$. On calcule alors l'image par la matrice A de ce vecteur, notée $R = AE$. Pour $1 \leq i \leq n - 1$, on obtient

$$R_i = G_i - \frac{2}{x_i - x_{i-1}}(2f'(x_i) + f'(x_{i-1})) - \frac{2}{x_{i+1} - x_i}(2f'(x_i) + f'(x_{i+1})),$$

où on rappelle que G_i ne dépend que des $y_i = f(x_i)$. Pour les cas $i = 1$ et $i = n - 1$, on a également utilisé le fait que $z_0 = f'(x_0) = \bar{z}_0$ et que $z_n = f'(x_n) = \bar{z}_n$.

In fine, la quantité R_i ne dépend que des valeurs de f et de sa dérivée aux noeuds d'interpolation. On peut donc utiliser des formules de Taylor pour estimer ces quantités.

On obtient ainsi le résultat intermédiaire suivant

$$\forall x \in \mathbb{R}, \forall h \neq 0, \exists \xi_1, \xi_2 \in S(x, x + h),$$
$$\frac{6}{h}\frac{f(x + h) - f(x)}{h} - \frac{2}{h}(2f'(x) + f'(x + h)) = f''(x) + \frac{h^2}{4}f^{(4)}(\xi_1) - \frac{h^2}{3}f^{(4)}(\xi_2), \quad \text{(III.6)}$$

où on a noté $S(x, x + h)$ le segment d'extrémités x et $x + h$ (h peut être négatif !).

Si on utilise (III.6) tout d'abord avec $x = x_i$, $h = x_{i+1} - x_i$ puis avec $x = x_i$ et $h = x_{i-1} - x_i$ et que l'on soustrait les deux résultats, on observe que le terme en $f''(x_i)$ se simplifie et il reste

$$|R_i| \leq \frac{7}{6}h^2\|f^{(4)}\|_\infty.$$

On a donc estimé R, et grâce au lemme III.3.21, on obtient que

$$\forall i \in \{1, \dots, n-1\}, \ |z_i - f'(x_i)| \leq \frac{7}{6} h^3 \|f^{(4)}\|_\infty, \tag{III.7}$$

cette inégalité étant également vraie pour $i = 0$ et $i = n$, car par construction $z_0 = f'(x_0)$ et $z_n = f'(x_n)$.

On est maintenant en position pour démontrer le résultat final. On fixe un $i \in \{0, \dots, n-1\}$ et on se place sur l'intervalle $[x_i, x_{i+1}]$. On note $\tilde{\pi}$ le polynôme de Hermite de f sur cet intervalle, c'est-à-dire l'unique polynôme de degré inférieur ou égal à 3 tel que

$$\tilde{\pi}(x_i) = f(x_i), \ \tilde{\pi}(x_{i+1}) = f(x_{i+1}), \ \tilde{\pi}'(x_i) = f'(x_i), \ \tilde{\pi}'(x_{i+1}) = f'(x_{i+1}).$$

D'après le théorème III.2.15 appliqué à l'intervalle $[x_i, x_{i+1}]$, on sait que l'erreur entre $\tilde{\pi}$ et f peut être majorée par

$$\sup_{[x_i, x_{i+1}]} |\tilde{\pi} - f| \leq \frac{h^4}{384} \|f^{(4)}\|_\infty.$$

Il nous reste donc à montrer que la différence entre π et $\tilde{\pi}$ est également d'ordre h^4 sur cet intervalle. Pour cela on observe que $q = \pi - \tilde{\pi}$ est un polynôme de Hermite sur cet intervalle vérifiant

$$q(x_i) = q(x_{i+1}) = 0, \ \ q'(x_i) = z_i - f'(x_i), \ q'(x_{i+1}) = z_{i+1} - f'(x_{i+1}).$$

Ainsi, l'estimation de stabilité (III.3) nous donne

$$\sup_{[x_i, x_{i+1}]} |q| \leq h(|z_i - f'(x_i)| + |z_{i+1} - f'(x_{i+1})|),$$

et ainsi grâce à l'estimation (III.7) on trouve que

$$\sup_{[x_i, x_{i+1}]} |q| \leq \frac{7}{3} h^4 \|f^{(4)}\|_\infty.$$

Le résultat vient en rassemblant les deux inégalités obtenues ci-dessus. ■

Remarque III.3.23

La construction de la spline π nécessite la connaissance de la dérivée de la fonction aux extrémités. Si on ne dispose pas de cette information (si on accède à f par des mesures par exemple), on peut remplacer les conditions de dérivées aux bornes par des formules d'approximation d'ordre au moins 3 (pourquoi ?) n'utilisant que les valeurs de f aux noeuds d'interpolation. Par exemple, dans le cas d'une discrétisation uniforme

$$\pi'(x_0) = \frac{\alpha_0 f(x_0) + \alpha_1 f(x_1) + \alpha_2 f(x_2) + \alpha_3 f(x_3)}{h},$$

où les α_i sont solutions du système

$$\begin{aligned} \alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 &= 0, \\ \alpha_1 + 2\alpha_2 + 3\alpha_3 &= 1, \\ \alpha_1 + 4\alpha_2 + 9\alpha_3 &= 0, \\ \alpha_1 + 8\alpha_2 + 27\alpha_3 &= 0. \end{aligned}$$

Avec une idée similaire pour définir $\pi'(x_n)$ on construit une nouvelle spline cubique. On peut montrer que cette spline vérifie la même estimation d'erreur que précédemment pour les fonctions de classe $\mathcal{C}^4$.

Exercice III.3.2

Justifier les affirmations de la remarque précédente. En particulier, la raison pour laquelle on choisit les α_i de cette manière et pourquoi le système obtenu admet-il bien une unique solution.

Noter que l'on peut modifier un peu cette définition des splines cubiques en imposant des conditions différentes aux deux extrémités de l'intervalle. Les propriétés obtenues sont similaires (voir [Rom96, p. 171]).

4 Approximation polynômiale

On a vu ci-dessus que les polynômes d'interpolation de Lagrange associés à une fonction f ne convergent pas toujours vers la fonction f, et ce quelque soit le choix des points d'interpolation. Cela dépend beaucoup de la fonction f.

Le but de ce paragraphe est donc d'étudier le problème de l'approximation polynômiale sans imposer de contraintes (très fortes) d'interpolation.

4.1 Les fonctions développables en séries entières

C'est là le premier exemple d'approximation polynômiale. Si f est une fonction DSE au voisinage de 0 de rayon de convergence $R > 0$, alors la suite de polynômes

$$P_n(x) = \sum_{k=0}^{n-1} \frac{f^{(k)}(0)}{k!} x^k,$$

converge localement uniformément vers f (ainsi que toutes leurs dérivées) sur le disque ouvert de convergence.

Le piège est que cette suite de polynômes peut converger sans toutefois que la limite ne soit égale à f comme par exemple le cas de $f(x) = e^{-1/x^2}$.

4.2 Théorème de Weierstrass. Polynômes de Bernstein

Le résultat essentiel est le suivant.

Théorème III.4.24 (Weierstrass)

Soit $[a, b]$ un intervalle fermé borné de $\mathbb{R}$. Pour toute fonction continue $f : [a, b] \mapsto \mathbb{R}$, il existe une suite de fonctions **polynomiales** *$(f_n)_n$ qui converge uniformément vers f sur $[a, b]$.*

Remarque III.4.25

- *Vous pouvez trouver de nombreuses preuves différentes de ce résultat (qui par ailleurs admet un certain nombre de généralisation (Th. de Stone-Weierstrass), en particulier aux dimensions supérieures à 1). C'est un grand classique à connaître (éventuellement en connaître plusieurs preuves et applications).*
- *Le sujet d'Analyse-Probabilités de l'agreg 1991 traite de questions connexes (en particulier sur la densité de l'ensemble des polynômes dans des espaces fonctionnels bien choisis).*
- *La preuve (constructive) proposée ici repose sur les polynômes de Bernstein (elle admet une interprétation probabiliste en terme de loi binomiale, qu'il est bon de connaître [HH06, Remarque 11.2.4, p. 204], [CLF95, Exo 21-3]).*
- *Ces polynômes de Bernstein ont également des applications en CAO (les fameuses courbes de Bézier par exemple) voir également [HH06].* **ATTENTION :** *les polynômes de Bernstein n'interpolent pas la fonction aux noeuds où f est évaluée.*

Ce résultat a de nombreuses applications, on peut noter la suivante par exemple :

Corollaire III.4.26

Soit $f : [a, b] \mapsto \mathbb{R}$ une fonction continue. On suppose que

$$\forall n \geq 0, \quad \int_a^b f(x) x^n \, dx = 0,$$

alors f est la fonction nulle.

Preuve :

On considère une suite $(f_n)_n$ de fonctions polynomiales qui converge uniformément vers f, d'après le théorème Weierstrass. L'hypothèse sur f, implique que

$$\forall n \geq 0, \quad \int_a^b f(x) f_n(x) \, dx = 0.$$

Mais la suite de fonctions $(ff_n)_n$ converge uniformément vers f^2 sur $[a,b]$. En effet, on a

$$\sup_{[a,b]} |f^2 - ff_n| \leq \left(\sup_{[a,b]} |f|\right)\left(\sup_{[a,b]} |f - f_n|\right),$$

ce dernier terme tendant vers 0 par hypothèse.

D'après le théorème de passage à la limite uniforme sous l'intégrale, on a

$$\int_a^b f^2(x)\,dx = \lim_{n\to\infty} \int_a^b f(x)f_n(x)\,dx = 0.$$

Ainsi f^2 est une fonction continue, positive et dont l'intégrale est nulle, elle est donc nécessairement nulle. ■

On verra également plus loin (voir Théorème III.4.36) le corollaire suivant.

Corollaire III.4.27

L'ensemble $\mathbb{R}[X]$ (des fonctions polynômes), est dense dans $L^2(]a,b[)$. Comme cet ensemble admet une base dénombrable, il exite une base hilbertienne de $L^2(]a,b[)$ formée de polynômes.

On va maintenant démontrer le théorème de Weierstrass, par une preuve constructive qui donne explicitement une telle suite de fonctions polynomiales, que l'on appelle *polynômes de Bernstein*.

Pour cela, pour tout $\delta > 0$ on pose

$$\omega(f,\delta) = \sup_{\substack{x,y\in[a,b]\\ |x-y|\leq\delta}} |f(x) - f(y)|.$$

Remarquons que cette fonction est bien définie par un argument de compacité. La fonction $\omega(f,.)$ ainsi définie, s'appelle le module de continuité de f. D'après le théorème de Heine, toute fonction continue f sur $[a,b]$ est uniformément continue. Ceci implique (est équivalent !) à la propriété

$$\lim_{\delta\to 0} \omega(f,\delta) = 0. \tag{III.8}$$

Preuve (du Théorème III.4.24):

– On observe tout d'abord qu'on peut se ramener au cas où $[a,b] = [0,1]$, en posant

$$\tilde{f}(t) = f(a + t(b-a)),\ \ \forall t \in [0,1].$$

Supposons donc maintenant que $[a,b] = [0,1]$.

– On introduit la suite $(f_n)_n$ définie de la façon suivante

$$f_n(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) C_n^k x^k (1-x)^{n-k},\ \ \forall x \in [0,1]. \tag{III.9}$$

La fonction f_n est bien un polynôme de degré au plus n.

Lemme III.4.28

On a les formules suivantes

$$\sum_{k=0}^{n} C_n^k x^k (1-x)^{n-k} = 1,\ \ \forall x \in \mathbb{R}, \tag{III.10}$$

$$\sum_{k=0}^{n} \frac{k}{n} C_n^k x^k (1-x)^{n-k} = x,\ \ \forall x \in \mathbb{R}, \tag{III.11}$$

$$\sum_{k=0}^{n} \left(\frac{k}{n}\right)^2 C_n^k x^k (1-x)^{n-k} = x^2 + \frac{x(1-x)}{n},\ \ \forall x \in \mathbb{R}. \tag{III.12}$$

Preuve :

– La formule (III.10) est simplement la formule du Binôme de Newton appliqué à $1 = (x + (1-x))^n$.

– Ecrivons :

$$\begin{aligned}\sum_{k=0}^{n}\frac{k}{n}C_n^k x^k(1-x)^{n-k} &= \sum_{k=1}^{n}\frac{k}{n}C_n^k x^k(1-x)^{n-k}\\ &= \sum_{k=1}^{n}C_{n-1}^{k-1}x^k(1-x)^{(n-1)-(k-1)}\\ &= x\sum_{k=1}^{n}C_{n-1}^{k-1}x^{k-1}(1-x)^{(n-1)-(k-1)}\\ &= x\sum_{k=0}^{n-1}C_{n-1}^{k}x^{k}(1-x)^{n-1-k} = x,\end{aligned}$$

en utilisant (III.10) au rang $n-1$.

– Cette fois-ci, on écrit (pour $n\geq 2$ sinon le calcul est immédiat)

$$\begin{aligned}\frac{n}{n-1}\sum_{k=0}^{n}\left(\frac{k}{n}\right)^2 C_n^k x^k(1-x)^{n-k} &= \frac{n}{n-1}\sum_{k=1}^{n}\left(\frac{k}{n}\right)^2 C_n^k x^k(1-x)^{n-k}\\ &= \sum_{k=1}^{n}\left(\frac{k^2}{n(n-1)}\right)C_n^k x^k(1-x)^{n-k}\\ &= \sum_{k=2}^{n}\left(\frac{k(k-1)}{n(n-1)}\right)C_n^k x^k(1-x)^{n-k} + \frac{1}{n-1}\sum_{k=1}^{n}\left(\frac{k}{n}\right)C_n^k x^k(1-x)^{n-k}\\ &= x^2\sum_{k=2}^{n}C_{n-2}^{k-2}x^{k-2}(1-x)^{(n-2)-(k-2)} + \frac{1}{n-1}x\\ &= x^2+\frac{1}{n-1}x.\end{aligned}$$

En multipliant la formule ainsi démontrée par $\frac{n-1}{n}$, on trouve bien le résultat annoncé.

■

Si on multiplie (III.10) par $f(x)$, on trouve

$$f(x)=\sum_{k=0}^{n}f(x)C_n^k x^k(1-x)^{n-k},\ \ \forall x\in[0,1]. \tag{III.13}$$

Reprenons maintenant la démonstration de la convergence uniforme de $(f_n)_n$ vers f sur $[0,1]$. Prenons $\delta>0$. Cette valeur de δ étant fixée, on choisit un $x\in[0,1]$, on utilise (III.9) et (III.13), puis on effectue la majoration suivante, en séparant dans la somme, l'ensemble des indices k tels que $|x-k/n|\leq\delta$ et ceux tels que $|x-k/n|>\delta$:

$$\begin{aligned}|f_n(x)-f(x)| &\leq \sum_{\substack{0\leq k\leq n\\ |x-k/n|\leq\delta}}|f(x)-f(k/n)|C_n^k x^k(1-x)^{n-k} + \sum_{\substack{0\leq k\leq n\\ |x-k/n|>\delta}}|f(x)-f(k/n)|C_n^k x^k(1-x)^{n-k}\\ &\leq \omega(f,\delta)\sum_{\substack{0\leq k\leq n\\ |x-k/n|\leq\delta}}C_n^k x^k(1-x)^{n-k} + \omega(f,\delta)\sum_{\substack{0\leq k\leq n\\ |x-k/n|>\delta}}\left(1+\frac{|x-k/n|}{\delta}\right)C_n^k x^k(1-x)^{n-k}\\ &\leq 2\omega(f,\delta)\underbrace{\sum_{k=0}^{n}C_n^k x^k(1-x)^{n-k}}_{=1} + \frac{\omega(f,\delta)}{\delta^2}\sum_{\substack{0\leq k\leq n\\ |x-k/n|>\delta}}(x-k/n)^2C_n^k x^k(1-x)^{n-k}\\ &\leq 2\omega(f,\delta)+\frac{\omega(f,\delta)}{\delta^2}\sum_{k=0}^{n}(x-k/n)^2C_n^k x^k(1-x)^{n-k}\\ &= 2\omega(f,\delta)+\frac{\omega(f,\delta)}{\delta^2}\left(x^2-2x^2+x^2+\frac{x(1-x)}{n}\right)\\ &= \omega(f,\delta)\left(2+\frac{1}{4n\delta^2}\right).\end{aligned}$$

Ainsi, si on choisit maintenant $\delta=\frac{1}{\sqrt{n}}$, on trouve l'estimation

$$\|f_n-f\|_\infty\leq\frac{9}{4}\omega\left(f,\frac{1}{\sqrt{n}}\right).$$

Cette dernière quantité tend bien vers 0 quand $n \to +\infty$ d'après (III.8).

■

La démonstration précédente donne un résultat relativement précis sur la convergence des polynômes de Bernstein en fonction du module de continuité de f. Il est possible d'estimer ce module de continuité en fonction de la régularité de la fonction. Ainsi, on peut montrer

- Si f est Lipschitzienne (en particulier si elle est $\mathcal{C}^1$), on a

$$\omega(f, \delta) \leq K\delta,$$

où K est la constante de Lipschitz de f.
- Si f est α-Hölderienne avec $\alpha \in]0, 1[$, on trouve

$$\omega(f, \delta) \leq K\delta^\alpha.$$

Attention : On ne peut pas espérer beaucoup mieux que les estimations ci-dessus même pour des fonctions très régulières. En effet, on peut montrer que

$$\lim_{\delta \to 0} \frac{\omega(f, \delta)}{\delta} = 0 \implies f \equiv 0.$$

Compléments :

- Dans [Rom96, Pb 35, p. 160], vous trouverez des propriétés supplémentaires de ces polynômes de Bernstein :
  - Si f est de classe $\mathcal{C}^p$, $p \geq 1$, alors la suite des dérivées $(f_n^{(p)})_n$ converge uniformément sur $[a, b]$ vers $f^{(p)}$.
  - Si f est de classe $\mathcal{C}^2$ et $[a, b] = [0, 1]$, la suite $(n(f_n - f))_n$ converge uniformément vers $\frac{x(1-x)}{2} f''(x)$. En particulier, dans ce cas précis, on déduit que, sauf si f est affine,

$$\|f_n - f\|_\infty \sim_\infty \frac{C}{n},$$

ce qui améliore sensiblement l'estimation par le module de continuité donné dans le théorème et fournit par ailleurs un équivalent et non juste une borne supérieure.
- **Polynômes de Jackson. Théorème de Jackson.** Les polynômes de Bernstein ne sont pas les *meilleurs* polynômes approximants pour des fonctions continues, ils sont cependant faciles à construire et à calculer dans beaucoup de cas pratiques. On peut néanmoins construire explicitement des polynômes (dits de Jackson), pour lesquels l'approximation est de meilleure qualité.
  Plus précisément, pour toute fonction continue, on peut trouver des polynômes (g_n) de degré n tels que

$$\|f - g_n\|_\infty \leq M\omega\left(f, \frac{b-a}{n}\right).$$

Voir par exemple [CM84, Th. 1.7],[Dem91, p. 43], [CLF95, Exo 21-5].
- **Densité de certains sous-ensembles de polynômes.** En utilisant le théorème de Weierstrass on peut montrer par exemple que l'ensemble des polynômes pairs (i.e. ne contenant que des termes de degré pair) est également dense dans $\mathcal{C}^0([a, b])$. Idem avec les polynômes impairs. C'est un bon exercice ...

4.3 Le problème de la meilleure approximation

Etant donné une topologie sur $\mathcal{C}^0([a, b])$ (on considérera seulement celle de la convergence uniforme, i.e. la norme infinie, ou bien la topologie L^2), on se demande maintenant s'il existe un meilleur polynôme de degré fixé qui approche une fonction donnée.

On va commencer par voir que l'existence d'un tel polynôme est toujours vraie et facile à prouver.

Théorème III.4.29

Soit N une norme sur $\mathcal{C}^0([a, b])$. Il existe toujours au moins un polynôme p_n de degré au plus n vérifiant

$$N(f - p_n) = \inf_{q \in \mathbb{R}_n[X]} N(f - q).$$

Preuve :

Le point-clé de la démonstration c'est le fait que $\mathbb{R}_n[X]$ est un espace de dimension finie. Dans cet espace on considère l'ensemble

$$K = \{q \in \mathbb{R}_n[X], \ \ N(f - q) \leq N(f)\}.$$

Cet ensemble est :

- non vide (il contient 0),
- borné (car par inégalité triangulaire on a $q \in K \Rightarrow N(q) \leq 2N(f)$),
- fermé car c'est l'image réciproque du fermé $[0, N(f)]$ par l'application continue $\varphi : q \in \mathbb{R}_n[X] \mapsto N(f-q)$ (êtes-vous capables de justifier la continuité de φ ?).

Comme nous sommes en dimension finie, l'ensemble K est donc un compact non vide et la fonction φ atteint donc son infimum en au moins un point de K. Il existe donc $p \in K$ tel que

$$N(f-p) = \inf_{q \in K} N(f-q).$$

Cet élément p répond à la question car tous les éléments de $\mathbb{R}_n[X] \setminus K$ ne peuvent en aucun cas être les infimums. ■

La question de l'unicité et de la caractétisation éventuelle de ces polynômes est plus délicate et dépend des normes choisies (Remarque : on ne considère ici que des intervalles bornés !).

4.3.1 Le cas de la norme uniforme

On considère ici l'espace $\mathcal{C}^0([a,b])$ muni de la norme infinie.

Théorème III.4.30 (de Tchebychev, [Dem91, p. 40], [CM84, p. 25])

Pour toute $f \in \mathcal{C}^0([a,b])$, il existe un **unique** *polynôme p_n de degré au plus n vérifiant*

$$\|f - p_n\|_\infty = \inf_{q \in \mathbb{R}_n[X]} \|f-q\|_\infty.$$

De plus, celui-ci est l'unique élément de $\mathbb{R}_n[X]$ tel que $f - p_n$ équi-oscille sur au moins $n+2$ points de $[a,b]$.

Dans ce théorème, qu'on admet ici, on a utilisé la définition suivante.

Définition III.4.31

On dit qu'une fonction $g \in \mathcal{C}^0([a,b])$ équi-oscille sur $k+1$ points de $[a,b]$, s'il existe des points distincts $x_0 < \ldots < x_k$ de $[a,b]$ tels que

$$\forall i \in \{0,...,k\}, \ \ |g(x_i)| = \|g\|_\infty,$$
$$\forall i \in \{0,...,k-1\}, \ \ g(x_{i+1}) = -g(x_i).$$

Autrement dit, la fonction doit atteindre son sup en valeur absolue au moins $k+1$ fois avec changement de signe !

Ce qu'il faut retenir ici c'est que le fait d'approcher au mieux une fonction par un polynôme en norme infinie, induit automatiquement une oscillation de l'erreur.

Convergence. Liens avec l'interpolation de Lagrange

- D'après le théorème de Weiestrass (et même, plus précisément le théorème de Bernstein), on sait que le polynôme de meilleure approximation p_n vérifie

$$\|p_n - f\|_\infty \leq C\omega\left(f, \frac{1}{\sqrt{n}}\right),$$

  et donc bien la suite $(p_n)_n$ converge uniformément vers f.
- La caractérisation de Tchebycheff, montre que la fonction $f - p_n$ doit nécessairement équi-osciller en au moins $n+2$ points. D'après le théorème des valeurs intermédiaires, cette fonction va donc s'annuler au moins $n+1$ fois. Il existe donc des points $x_{0,n} < ... < x_{n,n}$ tels que

$$f(x_{i,n}) = p_n(x_{i,n}), \ \ \forall i \in \{0,...,n\}.$$

  Ainsi, il existe un choix de $n+1$ points dans $[a,b]$, tel que le polynôme de meilleure approximation uniforme de f sur $[a,b]$ soit exactement le polynôme interpolé de Lagrange en ces points.
  Pour ce choix de points d'interpolation, on a donc la convergence uniforme de la suite des polynômes interpolés.

On peut résumer les propriétés génériques de convergence de l'interpolation de Lagrange comme suit.

Proposition III.4.32

Soit $[a,b]$ un intervalle compact de $\mathbb{R}$.
- *On suppose données pour tout $n \geq 0$, une famille de $n+1$ points distincts de $[a,b]$, alors il existe une fonction continue sur $[a,b]$ telle que la suite des polynômes d'interpolation de Lagrange en ces points soit non bornée.*
- *A contrario, pour toute fonction f continue sur $[a,b]$, il existe au moins un choix des points d'interpolation pour lesquels la suite des polynômes d'interpolation de Lagrange de f converge uniformément vers f.*

Application : pourquoi les polynômes de Tchebychev interviennent dans la théorie de l'interpolation de Lagrange ?
On peut maintenant comprendre pourquoi les polynômes de Tchebychev jouent un rôle particulier dans la théorie de l'interpolation de Lagrange.

D'après le théorème III.1.9, on a vu que l'erreur d'interpolation était majorée par

$$\|f-p_n\|_\infty \leq \|(x-x_0)\dots(x-x_n)\|_\infty \frac{\|f^{(n+1)}\|_\infty}{(n+1)!}.$$

Ainsi, une façon de choisir les points d'interpolation est de minimiser la norme infinie du polynôme $\pi_n(x) = (x-x_0)\dots(x-x_n)$ sur l'intervalle $[a,b]$. Ce polynôme est de la forme $\pi_n(x) = x^{n+1} - q_n(x)$ où q_n est de degré au plus n. On peut donc s'intéresser à trouver $q_n \in \mathbb{R}_n[X]$ tel que

$$\|x^{n+1} - q_n(x)\|_\infty = \inf_{p\in\mathbb{R}_n[X]} \|x^{n+1} - p(x)\|_\infty.$$

Autrement dit, il s'agit de trouver la meilleure approximation polynômiale de degré n de la fonction x^{n+1}.

D'après le théorème III.4.30, ce polynôme q_n est caractérisé par le fait que $x^{n+1}-q_n(x)$ équi-oscille entre $n+2$ points dans $[a,b]$. En particulier, par le théorème des valeurs intermédiaires, le polynôme $\pi_n = x^{n+1}-q_n(x)$ a exactement $n+1$ racines distinctes dans $[a,b]$, il est donc bien de la forme attendue $\pi_n(x) = (x-x_0)...(x-x_n)$.

Prenons $[a,b] = [-1,1]$ pour simplifier les calculs, on pose alors

$$T_n(x) = \cos(n\text{Arccos } x).$$

On a les propriétés suivantes
- $T_0(x) = 1, T_1(x) = x$.
- Pour tout $n \geq 2$, on a $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$.
- T_n est un polynôme de degré n et de coefficient dominant 2^{n-1}.
- T_n équi-oscille sur $n+1$ points (les points $y_k = \cos(k\pi/n)$ pour $k = 0,...,n$).

En conséquence de quoi, le polynôme $q_n(x) = x^{n+1} - 2^{-n}T_{n+1}(x)$ est de degré au plus n (le coefficient de x^{n+1} est nul) et il vérifie $x^{n+1} - q_n(x) = 2^{-n}T_{n+1}(x)$ qui est une fonction équi-oscillante sur $n+2$ points.

D'après le théorème de Tchebychev, le polynôme q_n est bien le polynôme de meilleure approximation uniforme dans $\mathbb{R}_n[X]$ de x^{n+1}. Ainsi $\pi_n(x) = x^{n+1} - q_n(x) = 2^{-n}T_{n+1}(x)$ est un bon candidat pour être le polynôme recherché.

On peut vérifier explicitement que T_{n+1} a bien $n+1$ racines distinctes dans $]-1,1[$ qui sont les $x_k = \cos\left(\pi\frac{2k+1}{2n+2}\right)$ pour $k = 0,\dots,n$.

Les points de Tchebychev ainsi définis sont donc les "meilleurs" points d'interpolation possibles, comme nous l'avons déjà vu. Le cas d'un intervalle $[a,b]$ se traite de la même façon.

Retour à la convergence des interpolations de Lagrange

Théorème III.4.33

Si on note p_n le polynôme de Lagrange de f associé à un choix de points d'interpolation, alors on a l'estimation

$$\|f-p_n\|_\infty \leq (1+\Lambda_n)\left(\inf_{\mathbb{R}_n[X]} \|f-q_n\|_\infty\right).$$

Il y a donc une compétition entre Λ_n qui tend vers l'infini et le fait que le second facteur tende vers 0 (Théorème de Weierstrass).

Par exemple, si f est Höldérienne, l'interpolée de Lagrange sur les noeuds de Tchebychev converge uniformément vers f car Λ_n croit comme $\log(n)$ alors que $\inf_{\mathbb{R}_n[X]} \|f-q_n\|_\infty$ se comporte au pire en $\omega(f, \frac{1}{\sqrt{n}}) \sim \frac{C}{n^{\alpha/2}}$, où α est l'exposant de Hölder de f. On a ici utilisé le théorème d'approximation de Bernstein.

4.3.2 Le cas de la norme L^p, $1 < p < +\infty$

L'unicité dans ce cas est liée à la propriété de stricte convexité de la norme L^p qui s'énonce de la façon suivante

$$\forall f, g \in L^p(]a,b[), \forall t \in]0,1[, \|tf + (1-t)g\|_{L^p} < t\|f\|_{L^p} + (1-t)\|g\|_{L^p}.$$

Ceci se démontre en reprenant la démonstration de l'inégalité de Minkowski et en analysant les cas d'égalités dans l'inégalité triangulaire.

4.3.3 Le cas de la norme L^1

On a également unicité du polynôme de meilleure approximation dans ce cas mais la démonstration est plus délicate. On renvoie par exemple à [GT96, Exo 11, p. 139].

4.3.4 Le cas de la norme L^2. Polynômes orthogonaux

L'espace L^2 étant un espace de Hilbert, la propriété d'unicité du polynôme de meilleure approximation p_n est immédiate (inégalité du parallèogramme) et celui-ci n'est autre que la projection orthogonale de la fonction considérée sur l'espace $\mathbb{R}_n[X]$.

Voir par exemple [Sch91, p157 et suivantes].

Théorème III.4.34

Pour tout $n \geq 0$, on note $p_n \in \mathbb{R}_n[X]$ le polynôme de meilleure approximation L^2 d'une fonction continue $f : [a,b] \to \mathbb{R}$. On a la convergence $\|p_n - f\|_2 \to 0$ quand $n \to \infty$.

Preuve :

Si on note $\tilde{p}_n$ le polynôme de meilleure approximation uniforme de f, on a

$$\|p_n - f\|_2 \leq \|\tilde{p}_n - f\|_2 \leq \sqrt{b-a}\|\tilde{p}_n - f\|_\infty \to 0.$$

■

Remarque III.4.35

- *Ce théorème se généralise aux fonctions moins régulières (i.e. dans L^2) grâce à la densité de l'ensemble $\mathcal{C}^0([a,b])$ dans L^2.*
- *On peut montrer (voir par exemple [AH09, p. 160]) que, si f est continue, le polynôme de meilleure approximation L^2 converge* **uniformément** *vers f.*

Calcul du polynôme de meilleure approximation L^2 Prenons $[a,b] = [0,1]$ pour simplifier les calculs. Si on veut calculer ce polynôme sous la forme

$$p_n(x) = a_0 + a_1 x + ... + a_n x^n,$$

les équations vérifiées par les $(a_k)_k$ sont données par

$$\int_0^1 x^i p_n(x)\,dx = \int_0^1 x^i f(x)\,dx, \ \ \forall i \in \{0,...,n\},$$

c'est-à-dire

$$\sum_{k=0}^{n} a_k \int_0^1 x^{k+i}\,dx = \int_0^1 x^i f(x)\,dx, \ \ \forall i \in \{0,...,n\}.$$

Il s'agit donc d'inverser la matrice *dite de Hilbert* dont le terme général est donné par $\left(\frac{1}{k+i+1}\right)_{i,k}$. Malgré ses airs innofencifs, cette matrice est très mal conditionée et sa résolution est numériquement très difficile et instable (pour $n = 7$ par exemple, la norme de l'inverse de la matrice de hilbert est de l'ordre 10^{10}).

L'idée de base pour simplifier ces calculs est la suivante : si on connaît une base orthogonale de l'espace $\mathbb{R}_n[X]$, alors la matrice du système est diagonale et donc très facile à inverser. Il se trouve qu'il existe un procédé *automatique* pour produire une base orthogonale d'une espace euclidient à partir de n'importe quelle autre base. Il s'agit de la méthode d'orthonormalisation de Gram-Schmidt.

Dans le cas qui nous occupe, il se trouve que l'algorithme est particulièrement simple à mettre en place comme on le verra par la suite.

Théorème III.4.36

Il existe une unique famille de polynômes **unitaires** $(P_n)_{n\geq 0}$ *tels que :*
– *le degré de P_n est égal à n pour tout n.*
– *P_n est orthogonal à $\mathbb{R}_{n-1}[X]$ pour $n \geq 1$.*

On peut en réalité appliquer tout ce qui précède en remplaçant l'espace $L^2(]a,b[)$ usuel par l'espace

$$L^2_w(]a,b[) = \left\{ f :]a,b[\to \mathbb{R} \text{ mesurables telles que } \int_a^b w|f|^2\,dx < +\infty \right\} \Big/ \underset{\text{p.p.}}{\sim},$$

muni du produit scalaire

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

où w est une fonction strictement positive et intégrable sur $]a,b[$. [1]

Le point intéressant est que cette base peut être calculée par une récurrence assez simple, si l'on pose $P_{-1} = 0$ par convention.

Proposition III.4.37

Quelque soit le poids w comme ci-dessus, la suite de polynômes orthogonaux associée vérifie

$$P_0 = 1,$$

et la relation de récurrence

$$P_{n+1}(x) = (x - a_n)P_n(x) - b_nP_{n-1}(x),\ \forall n \geq 0,$$

avec

$$a_n = \frac{(xP_n, P_n)_w}{\|P_n\|^2_w},$$

$$b_n = \frac{(xP_n, P_{n-1})_w}{\|P_{n-1}\|^2_w} = \frac{\|P_n\|^2_w}{\|P_{n-1}\|^2_w}.$$

Preuve :

On pose $Q_{n+1} = P_{n+1} - xP_n(x)$ qui est un polynôme de degré inférieur ou égal à n. De plus, si q est un polynôme de $\mathbb{R}_{n-2}[X]$, on a

$$(Q_{n+1}, q)_w = (P_{n+1}, q)_w - (P_n, xq)_w = 0,$$

car $xq \in \mathbb{R}_{n-1}[X]$ et donc est orthogonal à P_n.

Ainsi, le polynôme Q_{n+1} appartient à l'espace engendré par P_n et P_{n-1} ce qui justifie la forme donnée dans le théorème.

On identifie les coefficients en écrivant

$$0 = (P_{n+1}, P_{n-1})_w = (xP_n, P_{n-1})_w - b_n\|P_{n-1}\|^2_w,$$

$$0 = (P_{n+1}, P_n)_w = (xP_n, P_n)_w - a_n\|P_n\|^2_w.$$

Par ailleurs, la formule de récurrence au rend précédent s'écrit

$$P_n(x) = (x - a_{n-1})P_{n-1}(x) - b_{n-1}P_{n-2}(x),$$

ce qui donne, en prenant le produit scalaire avec le polynôme P_n,

$$\|P_n\|^2_w = (xP_{n-1}, P_n)_w,$$

car tous les autres termes sont nuls par orthogonalité. On trouve la seconde formule du coefficient b_n. ■

1. On peut même considérer le cas où $]a,b[$ est non borné, mais il faut alors supposer que les fonctions polynômes sont intégrables contre w.

Exemples

- Cas $w(x) = 1$ et $[a, b] = [-1, 1]$: les polynômes P_n sont appelés les polynômes de Legendre et sont donnés (à un facteur de normalisation près) par
$$P_n(x) = \frac{d^n}{dx^n}\left((x^2-1)^n\right).$$
- Cas $w(x) = \frac{1}{\sqrt{1-x^2}}$ et $[a, b] = [-1, 1]$: les polynômes P_n sont les polynômes de Tchebychev (que l'on retrouve donc encore ici !).
On remarque que dans ce cas le produit scalaire (xp_n, p_n) est toujours nul car c'est l'intégrale d'une fonction impaire sur un intervalle symétrique. Ceci implique que le coefficient a_n dans la récurrence est toujours nul. Il reste à déterminer le coefficient b_n. Après calcul on trouve que $b_n = 1/4$, ce qui donne la récurrence
$$P_{n+1} = xP_n - \frac{1}{4}P_{n-1}.$$
On a plus souvent l'habitude de travailler avec le polynôme $T_n = 2^{n-1}P_n$ qui vérifie la relation de récurrence
$$T_{n+1} = 2xT_n - T_{n-1},$$
et dont on vérifie que c'est bien l'unique polynôme tel que
$$\cos(nx) = T_n(\cos x), \quad \forall x \in \mathbb{R}.$$

La propriété suivante est très importante en vue des applications à l'intégration numérique, comme on le verra plus loin.

Théorème III.4.38

Quelque soit le poids w positif et intégrable sur $[a, b]$, le n-ième polynôme orthogonal (avec $n \geq 1$) a exactement n racines simples dans $]a, b[$.

Preuve :

Notons $x_1 < ... < x_k$ avec $k \leq n$ les k racines réelles de P dans $]a, b[$ qui sont de multiplicité impaire. Si $k = n$, le résultat est démontré (car les multiplicités sont nécessairement égales à 1).

On suppose que $k < n$, on pose alors $q(x) = (x - x_1)...(x - x_k)$ et d'après la propriété d'orthogonalité on a
$$0 = (P_n, q)_w = \int_a^b w(x)P_n(x)q(x)\,dx,$$
or P_nq est un polynôme qui ne change pas de signe dans $]a, b[$ car toutes ses racines dans cet intervalle sont de multiplicité paire. Ainsi wP_nq est de signe constant et d'intégrale nulle donc c'est une fonction nulle. Le poids ne s'annulant pas c'est que la fonction est identiquement nulle, ce qui est une contradiction. ■

Remarque III.4.39

On peut également montrer que les racines des polynômes orthogonaux sont entrelacées, c'est-à-dire qu'entre deux racines de P_{n+1} il y a une et une seule racine de P_n (voir par exemple [CLF95, Exo 21-6, p154]).
C'est une jolie propriété mais un peu plus anecdotique pour ce qui nous intéresse ici.

Si on rassemble les divers éléments obtenus ci-dessus, on arrive à la conclusion suivante :

Théorème III.4.40

Il existe une base hilbertienne $(\tilde{P}_n)_n$ de l'espace $L^2_w(]a, b[)$ formée de polynômes telle que pour tout $n \geq 0$, $\deg \tilde{P}_n = n$.
En conséquence, pour toute fonction $f \in L^2_w(]a, b[)$, on a
$$\sum_{n=0}^{+\infty}(f, \tilde{P}_n)^2_w = \|f\|^2_w, \quad \textit{Identité de Bessel-Parseval},$$
et on a, au sens de L^2_w, la formule
$$\sum_{n=0}^{+\infty}(f, \tilde{P}_n)_w\tilde{P}_n = f. \tag{III.14}$$

On rappelle que la formule (III.14) signifie que la suite des sommes partielles

$$S_N = \sum_{n=0}^{N} (f, \tilde{P}_n)_w \tilde{P}_n,$$

converge vers f dans L^2_w.

Preuve :

La famille $(P_n)_n$ construite précédemment est déjà orthogonale, il suffit donc de normaliser ses éléments en posant

$$\tilde{P}_n = \frac{P_n}{\|P_n\|_w}.$$

On montre maintenant l'inégalité de Bessel. Soit $N \geq 0$; par construction la somme partielle S_N n'est rien d'autre que la projection orthogonale de f sur $\mathbb{R}_N[X]$. Par le théorème de Pythagore on a donc

$$\|S_N\|_w^2 + \|f - S_N\|_w^2 = \|f\|_w^2,$$

et donc

$$\|S_N\|_w^2 \leq \|f\|_w^2.$$

Par ailleurs, par orthonormalité de la famille choisie, on a

$$\|S_N\|_w^2 = \sum_{n=0}^{N} |(f, \tilde{P}_n)_w|^2.$$

Ainsi la convergence de la série numérique est établie ainsi que l'inégalité

$$\sum_{n=0}^{\infty} |(f, \tilde{P}_n)_w|^2 \leq \|f\|_w^2.$$

On peut donc définir l'opérateur

$$T : f \in L^2_w \mapsto Tf = \sum_{n=0}^{+\infty} (f, \tilde{P}_n)_w \tilde{P}_n \in L^2_w.$$

D'après ce qui précède, cet opérateur est continu et de norme 1. Par ailleurs, par construction, on a $Tf = f$ pour tout $f \in \mathbb{R}[X]$.

Par ailleurs, pour toutes fonctions $f, g \in \mathcal{C}^0([a,b])$, nous avons

$$\|Tf - Tg\|_{L^2_w} \leq \|f - g\|_{L^2_w} \leq C_w \|f - g\|_\infty,$$

où $C_w = \left(\int_a^b w(x)\,dx\right)^{\frac{1}{2}}$. D'après le théorème de Weierstrass, on en déduit que l'égalité $Tf = f$ se prolonge à toutes les fonctions continues.

On peut enfin conclure à l'égalité $Tf = f$ pour tout $f \in L^2_w$ en invoquant la densité de l'ensemble des fonctions continues dans L^2_w. Ce dernier point est un peu délicat, je propose donc de le traiter à part dans la proposition qui va suivre. ■

Il reste donc à démontrer le résultat suivant. La preuve est un peu plus technique et nécessite des outils d'analyse fonctionnelle avancés.

Proposition III.4.41

- *L'espace $\mathcal{C}^0([a,b])$ est dense dans $L^\infty(]a,b[)$ faible-$\star$.*
- *L'espace $\mathcal{C}^0([a,b])$ est dense dans $L^2_w(]a,b[)$.*

Preuve :

- Soit $f \in L^\infty(]a,b[) \subset L^1(]a,b[)$. Par densité de $\mathcal{C}^0([a,b])$ dans $L^1(]a,b[)$, il existe [2] une suite $(\varphi_n)_n$ de fonctions continues telles que $\varphi_n \to f$ dans L^1. Quitte à en extraire une sous-suite, on peut également supposer que $\varphi_n \to f$ presque partout (pourquoi ?). Enfin, si on pose $\tilde{\varphi}_n = \max(-\|f\|_\infty, \min(\varphi_n, \|f\|_\infty))$, on voit que les fonctions $\tilde{\varphi}_n$ sont continues, bornées par $\|f\|_\infty$ et convergent presque partout et dans L^1 vers f.

2. Ce résultat est fondamental en analyse fonctionnelle, il repose notamment sur la régularité de la mesure de Lebesgue et sur le lemme de Urysohn, voir par exemple [Rud95]

Pour montrer la convergence faible-$*$, il faut maintenant montrer que pour tout $g \in L^1(]a,b[)$, on a

$$\int_a^b g(x)\tilde{\varphi}_n(x)\,dx \xrightarrow{n\to\infty} \int_a^b g(x)f(x)\,dx.$$

Ceci est vrai par une simple application du théorème de convergence dominée. En effet, on a la majoration uniforme

$$|g(x)\tilde{\varphi}_n(x)| \leq \|f\|_\infty |g(x)|,$$

avec $g \in L^1$ et la convergence presque partout est claire.

– Soit $f \in L^2_w$ telle que

$$0 = (f,\varphi)_w = \int_a^b wf\varphi\,dx, \ \ \forall \varphi \in \mathcal{C}^0([a,b]), \tag{III.15}$$

et montrons que $f = 0$. Pour tout $k \geq 0$, on considère la fonction $f_k = 1_{\{|f|\leq k\}} f$. Par construction on a $f_k \in L^\infty(]a,b[)$ et comme $\mathcal{C}^0([a,b])$ est dense dans $L^\infty(]a,b[)$ faible-$*$, il existe une suite $(\varphi_\varepsilon)_\varepsilon$ de fonctions continues telles que $\varphi_\varepsilon \rightharpoonup f_k$. On applique (III.15) à φ_ε et on passe à la limite quand $\varepsilon \to 0$ grâce à la propriété de convergence faible-$\star$. On déduit que

$$\int_a^b wf f_k\,dx = 0, \ \forall k \geq 0.$$

On peut maintenant faire tendre k vers l'infini par convergence dominée. On trouve que

$$\int_a^b w|f|^2\,dx = 0,$$

et donc que f est nulle presque partout.

Désuisons-en le résultat de densité qui nous intéresse. Notons E l'adhérence de $\mathcal{C}^0([a,b])$ dans L^2_w, de sorte que E est un sous-espace (en particulier convexe) fermé de L^2_w. On veut montrer que $E = L^2_w$.

Soit donc $g \in L^2_w$ et posons $f = g - P_E g$ où $P_E g$ désigne la projection orthogonale de g sur E. Par construction f est orthogonale à E et donc, en particulier, à $\mathcal{C}^0([a,b])$, ce qui signifie que $(f,\varphi)_w = 0$ pour toute fonction $\varphi \in \mathcal{C}^0([a,b])$. D'après ce qu'on a montré plus haut, cela impose $f = 0$ et donc $g = P_E g$, ce qui montre que $g \in E$. Ceci étant vrai pour tout $g \in L^2_w$ on a bien montré l'égalité des espaces et donc la propriété de densité annoncée.

■

4.4 Meilleure approximation L^2 au sens discret. Moindres carrés

On va aborder ici un problème que l'on relier à tout ce qui précède de deux façons différentes.

– **Interpolation sur-déterminée** : on se donne $n+1$ points d'interpolation $(x_i,y_i)_{0\leq i\leq n}$ que l'on cherche à interpoler à l'aide d'un polynôme de degré au plus m avec $m < n$. Bien entendu, ceci n'est strictement pas possible en général. C'est pourquoi on va relaxer les conditions d'interpolation en demandant seulement que le polynôme recherché approche le moins mal possible les valeurs y_i aux points x_i.

Plusieurs définitions de la notion de "proche" peuvent être choisies. On considèrera seulement ici la notion de "moindres carrés" qui consiste à rechercher $p \in \mathbb{R}_m[X]$ qui minimise sur $\mathbb{R}_m[X]$ la fonctionnelle

$$q \in \mathbb{R}_m[X] \mapsto J(q) = \sum_{i=0}^n |q(x_i) - y_i|^2.$$

– Supposons que l'on cherche à calculer le polynôme de degré m de meilleure approximation L^2 d'une fonction f sur un intervalle $[a,b]$. Ce problème est théoriquement résoluble mais nécessite le calcul d'un certain nombre d'intégrales, qu'en général on ne peut calculer explicitement. Une idée peut alors être de remplacer la norme L^2 pondérée de $f - p$ par une approximation de celle-ci (on verra les méthodes d'approximation dans le prochain chapitre). On considère ici en toute généralité une méthode d'intégration numérique qui fait intervenir des points $x_0 < ... < x_n$ de $[a,b]$ et des poids $\omega_i \in \mathbb{R}$.

Ainsi, on se retrouve à rechercher le polynôme $p \in \mathbb{R}_m[X]$ qui minimise la fonctionnelle

$$q \in \mathbb{R}_m[X] \mapsto J(q) = \sum_{i=0}^n \omega_i |q(x_i) - f(x_i)|^2.$$

On voit bien que les deux problèmes ont la même structure que l'on peut résumer de la façon suivante : étant donnés des points $x_0 < ... < x_n$, des valeurs $y_0, ..., y_n$ et des poids $\omega_i \in \mathbb{R}$, peut-on trouver un polynôme $p \in \mathbb{R}_m[X]$ qui minimise la fonctionnelle suivante

$$J(q) = \sum_{i=0}^n \omega_i |q(x_i) - y_i|^2. \tag{III.16}$$

Dans toute la suite on supposera que les poids ω_i sont strictement positifs (ceci n'est pas vrai pour toutes les formules d'intégration numérique ! !). Dans le cas contraire, le problème devient déraisonnablement instable. En conséquence la fonctionnelle J est positive.

Remarque III.4.42

Si $m \geq n$, ce problème n'a que peu d'intérêt car on sait qu'alors on peut exactement interpoler les valeurs y_i aux points x_i et que donc le polynôme de Lagrange annule (et donc minimise) la fonctionnelle J.

Théorème III.4.43

Sous les hypothèses précédentes, il existe un unique polynôme $p \in \mathbb{R}_m[X]$ qui minimise J. Les coefficients de ce polynôme peuvent être obtenus en résolvant un système linéaire de taille $m+1$.

Preuve :

On vérifie tout d'abord que

$$\|q\| = \left(\sum_{i=0}^{n} \omega_i |q(x_i)|^2 \right)^{\frac{1}{2}},$$

est une norme euclidienne sur $\mathbb{R}_m[X]$ (on utilise ici de façon cruciale l'hypothèse $m < n$). On établit ensuite, par l'inégalité de Cauchy-Schwarz que l'on a

$$J(q) \geq \frac{1}{2}\|q\|^2 - C, \ \ \forall q \in \mathbb{R}_m[X],$$

où $C \in \mathbb{R}$ ne dépend que des données (y_i, ω_i).

En conséquence, la fonction J est coercive et strictement convexe. Elle admet donc un unique minimiseur p dans $\mathbb{R}_m[X]$. Celui-ci est caractérisé par le fait que la différentielle de J est nulle en ce point. On doit donc avoir $DJ(p).q = 0$, pour tout $q \in \mathbb{R}_m[X]$. La fonctionnelle étant quadratique, on calcule immédiatement

$$DJ(p).q = 2\sum_{i=0}^{n} \omega_i (p(x_i) - y_i).q(x_i).$$

Cette quantité doit donc être nulle pour tout $q \in \mathbb{R}_m[X]$. Si on cherche p sous la forme

$$p(x) = \sum_{k=0}^{m} a_k x^k,$$

et que l'on applique la condition d'Euler-Lagrange à tous les monômes $q = 1$, $q = x$, ..., $q = x^m$. On obtient les équations suivantes

$$\sum_{i=0}^{n} \omega_i \left(\sum_{k=0}^{m} a_k (x_i)^k - y_i \right).(x_i)^j = 0, \ \ \forall j \in \{0, ..., m\}.$$

Si on introduit la matrice **rectangle** de Vandermonde $V = ((x_i)^j)_{\substack{0 \leq i \leq n \\ 0 \leq j \leq m}}$, alors le problème se met sous la forme

$${}^tVD_\omega Va = {}^tVD_\omega y,$$

où a est le vecteur des coefficients inconnus $a_0, ..., a_n$, y le vecteur des données $y_0, ..., y_n$ et D_ω la matrice diagonale dont les coefficients diagonaux sont les ω_i.

Dans le cas où tous les poids sont égaux à 1 on trouve l'équation ${}^tVVa = {}^tVy$. Cette équation est parfois appelée l'équation normale associée à A.

On voit immédiatement (pourquoi ?) que la matrice V est de rang maximal, c'est-à-dire en l'occurence (comme $m < n$) que ses colonnes sont linéairement indépendantes. En conséquence, A est injective et la matrice ${}^tAD_\omega A$ est une matrice symétrique définie positive (en particulier inversible).

En effet, pour tout vecteur $a \in \mathbb{R}^{m+1}$, on a

$${}^ta{}^tAD_\omega Aa = (D_\omega Aa, Aa) = \sum_{i=0}^{n} \omega_i (Aa)_i^2 \geq 0,$$

de plus, cette quantité est nulle si et seulement si le vecteur Aa est nul mais par injectivité cela ne peut arriver que pour $a = 0$. ■

Remarque III.4.44

L'exemple le plus classique de cette méthode est celle de la régression linéaire qui correspond au cas $m = 1$. Il s'agit d'approcher au mieux une famille de points par une droite.

5 Compléments et remarques

Algorithmes en MATLAB : Le livre [QSS07] contient le listing de petits programmes en Matlab correspondant à la plupart des méthodes présentées ici. La syntaxe étant très proche de celle de Scilab, cela peut être fort utile.

Une jolie extension du théorème de Weierstrass : Il s'agit d'une possibilité très intéressante de développement. Cela fait intervenir un calcul de déterminant, la notion de déterminant de Gram, etc ... En particulier, la démonstration est de nature essentiellement Hilbertienne alors même que le résultat s'énonce dans l'espace des fonctions continues, qui n'est pas un Hilbert ! Voir [Gou94, p. 286],[CLF95, Exo 21-1,p 139].

Théorème III.5.45 (Muntz)

Soit $(\lambda_n)_n$ une suite de réels positifs et $[a, b]$ un intervalle compact de $\mathbb{R}$.
L'espace vectoriel engendré par les fonctions $f_n : x \mapsto x^{\lambda_n}$ est dense dans $\mathcal{C}^0([a, b])$ si et seulement si la série $\sum_n \frac{1}{\lambda_n}$ est divergente.

Chapitre IV

Intégration numérique

Le but de ce chapitre est d'étudier quelques méthodes de calcul approché de l'intégrale d'une fonction f entre a et b. La plus grande partie du chapitre est consacrée au cas où a et b sont finis mais on parlera aussi un peu du cas où a et/ou b sont infinis.

Une méthode d'intégration numérique est souvent appelée **méthode de quadrature**.

1 Méthodes de quadrature élémentaires

1.1 Généralités

L'idée de base d'une méthode de quadrature élémentaire est d'introduire une approximation $\tilde{f}$ de la fonction f sur l'intervalle $[a, b]$ dont on sache aisément calculer l'intégrale. Si on sait contrôler l'erreur (en norme uniforme) que l'on commet en approchant f par $\tilde{f}$, on aura

$$\left| \int_a^b f(x)\,dx - \int_a^b \tilde{f}(x)\,dx \right| \leq |b-a| \|f - \tilde{f}\|_\infty.$$

La première idée qui vient à l'esprit est d'utiliser les méthodes d'interpolation polynômiales pour construire la fonction approchante $\tilde{f}$ [1].

Ainsi, si on considère $n+1$ points distincts $x_0 < ... < x_n$ dans $[a, b]$ et prenons pour $\tilde{f}$ l'interpolée de f en ces points. On introduit alors la formule de quadrature

$$I_n(f) = \int_a^b \tilde{f}\,dx = \sum_{i=0}^{n} f(x_i) \left(\int_a^b l_i(x)\,dx \right),$$

où l_i est le i-ième polynôme élémentaire de Lagrange associé à la subdivision de $[a, b]$ choisie.

On met cette formule sous la forme

$$I_n(f) = \sum_{i=0}^{n} \omega_i f(x_i).$$

On a vu que lorsque n tend vers l'infini, la convergence uniforme de l'interpolée de Lagrange n'était pas vraie en général, ce qui laisse supposer que la convergence de $I_n(f)$ vers l'intégrale de f n'est pas vraie en général.

S'ajoutent à cela des problèmes de stabilité de la méthode que l'on peut résumer de la façon suivante :

- Les coefficients ω_i ne sont pas positifs en général, ce qui implique que la formule de quadrature appliquée à une fonction positive, ne sera pas toujours positive.
- La norme de l'opérateur I_n vaut $\sum_{i=0}^{n} |\omega_i|$ et peut tendre vers l'infini.

Le résultat général associé à ces formules est le suivant

Proposition IV.1.1

La formule I_n est exacte pour les polynômes de degré inférieur ou égal à n

$$I_n(f) = \int_a^b f(x)\,dx, \quad \forall f \in \mathbb{R}_n[X].$$

Si de plus n est pair et que les points de la subdivision sont symétriques par rapport au centre de l'intervalle $[a, b]$, alors la formule est également exacte sur les polynômes de degré $n + 1$.

1. D'autres idées sont possibles : fonctions trigonométriques, fractions rationnellesn etc ...

Preuve :

Si f est un polynôme de $\mathbb{R}_n[X]$, alors l'interpolée de Lagrange $\tilde{f}$ est égal à f et il est alors clair que $I_n(f) = \int_a^b \tilde{f}\,dx = \int_a^b f(x)\,dx$.

Supposons maintenant que n est pair et considérons un polynôme p de degré exactement $n+1$. On peut écrire ce polynôme sous la forme

$$p(x) = \alpha \underbrace{\left(x - \frac{a+b}{2}\right)^{n+1}}_{\stackrel{\text{def}}{=}\pi(x)} + q(x),$$

avec $q \in \mathbb{R}_n[X]$. D'après ce qui précède on a $I_n(q) = \int_a^b q\,dx$. Il reste donc à montrer l'exactitude de la formule sur le polynôme π. Grâce à l'hypothèse de symétrie, on vérifie que l'on a

$$l_i(a+b-x) = l_{n-i}(x), \ \ \forall i \in \{0,...,n\}, \ \forall x \in [a,b],$$

et donc que

$$\omega_i = \omega_{n-i}, \ \ \forall i \in \{0,...,n\}.$$

Par ailleurs, comme $n+1$ est impair, on a

$$\pi(x_i) = -\pi(x_{n-i}), \ \ \forall i \in \{0,...,n\},$$

et donc

$$I_n(\pi) = \sum_{i=0}^{n} \omega_i f(x_i) = 0.$$

Comme par ailleurs, on a toujours par imparité de $n+1$

$$\int_a^b \left(x - \frac{a+b}{2}\right)^{n+1} dx = 0,$$

le résultat est démontré. ■

Définition IV.1.2

Le degré maximal des polynômes pour lesquels une formule de quadrature est exacte, s'appelle le **degré d'exactitude** *de la formule.*

ATTENTION : Certains livres (même parmi les meilleurs !) appellent ce degré d'exactitude : *l'ordre de la méthode*. On verra par la suite que cette dénomination est **TRES** trompeuse et je propose de ne pas l'utiliser.

Remarque IV.1.3

Une des contraintes fondamentales que l'on a en construisant des méthodes de quadrature, c'est que toutes les quantités (les points, les coefficients, etc ...) doivent être calculables sans trop d'efforts !
Par ailleurs, la performance de la méthode sera évaluée en comparant la précision obtenue par rapport au coût du calcul (typiquement : le nombre d'évaluations de la fonction nécessaires pour obtenir le résultat).

Exercice IV.1.1

Quelle formule de quadrature obtient-on si on choisit pour $\tilde{f}$ le polynôme de Bernstein de degré n associé à f ?

1.2 Rectangles à gauche et à droite

Si on prend un seul point d'interpolation (i.e. $n = 0$), on retrouve différents choix classiques :

$$I_0^{RG}(f) = (b-a)f(a), \ \text{ rectangle à gauche,}$$
$$I_0^{RD}(f) = (b-a)f(b), \ \text{ rectangle à droite,}$$
$$I_0^{PM}(f) = (b-a)f\left(\frac{a+b}{2}\right), \ \text{ point-milieu.}$$

D'après ce qui précède, les deux premières ont un degré d'exactitude de 0, alors que la dernière a un degré d'exactitude de 1.

D'un certain point de vue, on voit que la méthode du point-milieu est le meilleur choix parmi les méthodes à un point d'interpolation.

1.3 Formules de Newton-Cotes

Il s'agit de mettre en oeuvre la formule générale dans le cas d'une subdivision uniforme de l'intervalle.
- Cas $n = 0$: on retrouve la formule du point-milieu.
- Cas $n = 1$: $x_0 = a, x_1 = b$, on obtient la méthode du trapèze

$$I_1^T(f) = (b-a)\left(\frac{1}{2}f(a) + \frac{1}{2}f(b)\right),$$

dont le degré d'exactitude est 1.
- Cas $n = 2$: c'est la méthode de Simpson

$$I_2^S(f) = (b-a)\left(\frac{1}{6}f(a) + \frac{4}{6}f\left(\frac{a+b}{2}\right) + \frac{1}{6}f(b)\right),$$

dont le degré d'exactitude est 3.
- Cas $n = 3$: c'est la méthode de Newton (dite méthode des $3/8$)

$$I_3^N(f) = (b-a)\left(\frac{1}{8}f(a) + \frac{3}{8}f\left(\frac{2a+b}{3}\right) + \frac{3}{8}f\left(\frac{a+2b}{3}\right) + \frac{1}{8}f(b)\right),$$

qui a également un degré d'exactitude de 3.

En pratique, les formules de Newton-Cotes n'ont d'intérêt que pour n pair.

Par ailleurs, les coefficients ω_i ne sont positifs que pour $n \leq 7$, ce qui fait que les formules deviennent instables pour de grandes valeurs de n. On peut en effet montrer que la somme $\sum_{i=0}^{n} |\omega_{i,n}|$ tend vers l'infini quand n tend vers l'infini.

1.4 Formules de Gauss

Nous avons vu que, pour tout choix de $n+1$ points d'interpolation, on peut construire une méthode qui est exacte sur les polynômes de degré au plus n (et même $n+1$ dans certains cas). On peut maintenant se poser la question de savoir s'il existe un *meilleur* choix de ses points.

Commençons par un exemple avec $n = 1$ (i.e. 2 points d'interpolation) dans l'intervalle $]-1, 1[$. Etant donné $\alpha \in]0, 1]$, on considère les 2 points d'interpolation $-\alpha, \alpha$ symétriques par rapport au centre de l'intervalle. La formule de quadrature associée est donnée par

$$I_1(f) = \underbrace{2}_{=(b-a)} \frac{1}{2}\left(f(-\alpha) + f(\alpha)\right) = f(-\alpha) + f(\alpha).$$

Par construction cette formule est exacte pour tout $f \in \mathbb{R}_1[X]$. Essayons de voir si on peut choisir α pour qu'elle soit aussi exacte pour $f(x) = x^2$. On calcule

$$\int_{-1}^{1} x^2\, dx = \frac{2}{3}, \;\; I(f) = 2\alpha^2,$$

ce qui nous donne

$$\alpha = \frac{1}{\sqrt{3}}.$$

Ainsi, la formule

$$I_1(f) = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right),$$

est exacte pour les polynômes de degré 2. Par l'argument de symétrie déjà utilisé, on voit qu'en réalité la formule est également exacte pour les polynômes de degré 3 !

Ainsi, par le simple choix judicieux des points d'interpolation (sans en changer le nombre !) on a obtenu une formule à 2 points dont le degré d'exactitude est 3 au lieu de 1 pour la formule usuelle de Newton-Cotes (i.e. la formule du trapèze).

Nous venons de construire la méthode de Gauss à 2 points. On va montrer que cette même idée peut se mettre en place pour tout choix de n.

Théorème IV.1.4

Soit $[a, b]$ un intervalle compact de $\mathbb{R}$ et $n \geq 0$. Il existe un unique choix de $n+1$ points d'interpolation $x_0 < ... < x_n$ tel que la formule de quadrature basée sur l'interpolation en ces points a un degré d'exactitude exactement de $2n+1$. C'est la formule de Gauss à $n+1$ points notée I_n^G.
Par ailleurs, il n'existe aucune formule à $n+1$ points dont le degré d'exactitude est strictement plus grand que $2n+1$.

Preuve :

– Commençons par la dernière propriété. Soit $x_0 < ... < x_n$ des points d'interpolation dans $[a, b]$. On considère le polynôme $q(x) = (x - x_0)^2...(x - x_n)^2$. Celui-ci est de degré $2n + 2$, donc si la formule de quadrature associée est exacte sur $\mathbb{R}_{2n+2}[X]$, on aura

$$\int_a^b q(x)\,dx = I_n(q) = \sum_{i=0}^{n} \omega_i q(x_i).$$

Ceci est impossible car q est positif non identiquement nul, donc son intégrale est strictement positive alors que la somme de droite est nulle par construction de q.

– Supposons que la formule donnée dans l'énoncé existe et posons $\pi(x) = (x - x_0)...(x - x_n)$. C'est un polynôme de degré $n + 1$.
Soit maintenant p un polynôme de degré inférieur ou égal à n. Le produit $p\pi$ est de degré au plus $2n + 1$ et donc on doit avoir

$$\int_a^b p\pi\,dx = \sum_{i=0}^{n} \omega_i p(x_i)\pi(x_i) = 0.$$

Ainsi le polynôme π est unitaire et orthogonal à $\mathbb{R}_n[X]$, ce qui montre que c'est nécessairement le $(n + 1)$-ième polynôme de Legendre.

– Ce qui précède suggère de prendre comme points d'interpolation les racines du $n + 1$-ième polynôme de Legendre (on a déjà vu que ce polynôme admet bien $n + 1$ racines distinctes dans $]a, b[$).
Montrons maintenant que ce choix convient en montrant que la formule I_n est exacte pour tout polynôme p de degré inférieur ou égal à $2n + 1$. Pour cela, on effectue la division euclidienne de p par le polynôme π (le $n + 1$-ième polynôme de Legendre), ce qui donne

$$p = \pi q + r, \quad \text{avec } \deg(r) \leq n.$$

Par ailleurs, au vu des degrés de p et de π, on a $\deg(q) \leq n$. Par construction on a

$$\int_a^b p\,dx = \int_a^b (\pi q + r)\,dx = \int_a^b r\,dx,$$

car π est orthogonal à $\mathbb{R}_n[X]$. De plus, πq s'annule sur les points de la subdivision utilisée et donc $I_n(p) = I_n(\pi q + r) = I_n(r)$. Comme la formule est exacte sur les polynômes de degré inférieur ou égal à n, on a bien l'égalité annoncée.

■

En plus d'être exactes sur le plus de polynômes possibles, les méthodes de Gauss ont d'excellentes propriétés de stabilité et de convergence (même si cette notion a un intérêt pratique modéré comme on le verra ci-dessous).

Théorème IV.1.5

Soit $[a, b]$ un intervalle compact de $\mathbb{R}$. La formule de quadrature de Gauss à $n+1$ points n'a que des coefficients $\omega_{i,n}$ strictement positifs et de somme égale à $b - a$.
En conséquence, la méthode est convergente au sens où

$$I_n^G(f) \xrightarrow[n\to+\infty]{} \int_a^b f(x)\,dx, \quad \forall f \in \mathcal{C}^0([a, b]).$$

Preuve :

– Fixons n pour le moment et notons $l_i(x) = \Pi_{j\neq i}(x - x_j)$ le i-ième polynôme élémentaire de Lagrange associé aux points de Gauss. Le polynôme l_i^2 est de degré $2n$ et donc s'intègre exactement par la formule de Gauss

$$\int_a^b l_i^2(x)\,dx = \sum_{j=0}^{n} \omega_{j,n} \underbrace{l_i^2(x_j)}_{=\delta_{ij}} = \omega_{j,n},$$

ce qui montre bien que $\omega_{j,n} > 0$. Par ailleurs la somme des coefficients n'est autre que la méthode de Gauss appliquée au polynôme constant égal à 1, qui vaut donc la longueur de l'intervalle.

– Montrons maintenant la propriété de convergence. Pour cela, on introduit pour tout $n \geq 0$

$$T_n(f) \stackrel{\text{def}}{=} \int_a^b f(x)\,dx - I_n^G(f),$$

et on veut montrer que $T_n(f) \to 0$ quand $n \to \infty$ pour f continue. D'après les propriétés précédentes nous avons

$$|T_n(f)| \leq 2(b - a)\|f\|_\infty, \quad \forall f \in \mathcal{C}^0([a, b]), \forall n \geq 0.$$

Fixons $\varepsilon > 0$. D'après le théorème de Weierstrass, il existe un polynôme p_ε tel que $\|f - p_\varepsilon\|_\infty \leq \varepsilon$. On a alors

$$|T_n(f)| \leq |T_n(p_\varepsilon)| + |T_n(f - p_\varepsilon)| \leq |T_n(p_\varepsilon)| + 2(b-a)\|f - p_\varepsilon\|_\infty \leq |T_n(p_\varepsilon)| + 2(b-a)\varepsilon.$$

Mais, si on prend $n \geq n_\varepsilon$ le degré de p_ε, on sait que la formule de Gauss est exacte sur p_ε, c'est-à-dire que $T_n(p_\varepsilon) = 0$ et donc on trouve

$$|T_n(f)| \leq 2(b-a)\varepsilon, \ \ \forall n \geq n_\varepsilon,$$

ce qui montre bien le résultat.

■

En pratique, on n'utilise pas les formules de Gauss dans la limite $n \to +\infty$ car le calcul des points de Gauss et des coefficients associés n'est pas facile (il faut déterminer les zéros des polynômes orthogonaux !) et assez couteux. De plus, quand la fonction étudiée est très régulière, l'erreur $|T_n(f)|$ est contrôlée par des dérivées d'ordre élevé de la fonction, on n'est donc pas assuré, en général, d'une convergence très rapide.

Comme pour les formules de Newton-Cotes, les formules de Gauss sont quasi-systématiquement utilisées sous la forme composite comme on va le voir dans le paragraphe suivant. En effet, prendre n grand dans les méthodes de Gauss pose les problèmes suivants :

- Le calcul des coefficients de la méthode peut devenir couteux et instable quand n augmente (tout dépend du poids w bien sûr). Voir le paragaphe 3.1.
- La précision de la méthode dépend de la norme des dérivées d'ordre n de la fonction étudiée, ce qui peut devenir rédhibitoire pour beaucoup de fonctions dont les dérivées successives peuvent vite devenir grandes.

2 Formules composites

2.1 Généralités

L'idée est de décomposer l'intervalle $[a, b]$ initial en sous-intervalles via une subdivision $a = x_0 < \cdots < x_m = b$ et d'utiliser la relation de Chasles

$$\int_a^b f(x)\,dx = \sum_{i=0}^{m-1} \int_{x_i}^{x_{i+1}} f(x)\,dx.$$

On est ensuite ramenés à évaluer une approximation de chacune des intégrales élémentaires. Pour effectuer cette quadrature élémentaire, on utilise l'une des formules de quadrature étudiées plus haut (Newton-Cotes, Gauss, etc ...) avec un nombre de points fixés. Dans ce qui suit, on utilisera la même formule sur chaque sous-intervalle mais bien sûr rien ne nous y oblige en toute généralité.

Fixons maintenant les notations. On se donne une formule de quadrature notée I_0^1 pour approcher l'intégrale de fonctions continues sur l'intervalle de référence $[0, 1]$. On notera

$$I_0^1(f) = \sum_{k=0}^{p} \omega_k f(\theta_k),$$

avec $\theta_k \in [0, 1]$ et $\omega_k \in \mathbb{R}$. Dans toute la suite, on suppose que la formule I_0^1 a un degré d'exactitude de n, c'est-à-dire que n est le plus grand entier tel que

$$I_0^1(f) = \int_0^1 f(x)\,dx, \ \ \forall f \in \mathbb{R}_n[X].$$

On peut ensuite transposer cette formule à un intervalle compact arbitraire $[\alpha, \beta]$ en écrivant

$$I_\alpha^\beta(f) = (\beta - \alpha) \sum_{k=0}^{p} \omega_k f(\alpha + \theta_k(\beta - \alpha)).$$

Par un changement de variable immédiat, on voit que cette formule a également un degré d'exactitude de n sur l'intervalle $[\alpha, \beta]$.

Théorème IV.2.6

Soit I_0^1 une formule d'intégration élémentaire sur $[0, 1]$ de degré d'exactitude n et I_α^β la formule associée sur $[\alpha, \beta]$. Il existe un $C > 0$, qui ne dépend que de p, des ω_k et des θ_k qui définissent la méthode élémentaire I_0^1, telle que

$$\forall \alpha < \beta, \forall f \in \mathcal{C}^{n+1}([\alpha, \beta]), \ \ \left| \int_\alpha^\beta f(x)\,dx - I_\alpha^\beta(f) \right| \leq C|\beta - \alpha|^{n+2} \|f^{(n+1)}\|_{L^\infty([\alpha,\beta])}.$$

Preuve :

La preuve se déroule en deux temps. On montre tout d'abord le résultat sur l'intervalle élémentaire $[0,1]$, puis ensuite on en déduit le cas général par changement de variable affine.

– Supposons que $[\alpha,\beta]=[0,1]$. La formule de Taylor-Lagrange implique que

$$\left| f(x) - \underbrace{\left(f(0) + xf'(0) + \cdots + \frac{x^n}{n!} f^{(n)}(0) \right)}_{\stackrel{\text{def}}{=} P(x)} \right| \leq \frac{1}{(n+1)!} \|f^{(n+1)}\|_{L^\infty(]0,1[)}, \quad \forall x \in [0,1].$$

Le polynôme P est le polynôme de Taylor de f en 0. Comme il est de degré inférieur ou égal à n, la formule de quadrature est exacte, i.e. $I_0^1(P) = \int_0^1 P(x)\,dx$. En conséquence, on a

$$\begin{aligned} \left| \int_0^1 f(x)\,dx - I_0^1(f) \right| &= \left| \int_0^1 (f(x) - P(x))\,dx - I_0^1(f-P) \right| \\ &\leq \|f-P\|_\infty + \sum_{k=0}^p |\omega_k| \|f-P\|_\infty \leq \frac{1}{(n+1)!} \underbrace{\left(1 + \sum_{k=0}^p |\omega_k| \right)}_{=C} \|f^{(n+1)}\|_\infty. \end{aligned}$$

Ce qui donne bien le résultat attendu.

– Si on se place maintenant sur l'intervalle $[\alpha,\beta]$, on pose $\tilde{f}(t) = f(\alpha + (\beta-\alpha)t)$, de sorte que

$$\int_\alpha^\beta f(x)\,dx = (\beta-\alpha) \int_0^1 \tilde{f}(t)\,dt,$$

$$I_\alpha^\beta(f) = (\beta-\alpha) I_0^1(\tilde{f}).$$

Le résultat précédent donne donc bien

$$\left| \int_\alpha^\beta f(x)\,dx - I_\alpha^\beta(f) \right| \leq C|\beta-\alpha| \|\tilde{f}^{(n+1)}\|_{L^\infty(]0,1[)} \leq C|\beta-\alpha|^{n+2} \|f^{(n+1)}\|_{L^\infty(]0,1[)}.$$

■

Remarque IV.2.7

La constante C obtenue dans le théorème n'est pas optimale. La valeur optimale de la constante peut être déterminée à l'aide de notions un peu plus avancées comme le noyau de Peano par exemple (voir [Dem91, CM84]). Il ne me semble pas utile d'aller aussi loin dans le cadre de l'agrégation.

On peut maintenant définir la formule composite basée sur la quadrature élémentaire I_0^1 que l'on s'est fixée. Pour cela, étant donnée une subdivision $a = x_0 < x_1 < ... < x_{m-1} < x_m = b$ de l'intervalle $[a,b]$, on pose

$$S_m(f) = \sum_{i=0}^{m-1} I_{x_i}^{x_{i+1}}(f).$$

Si f est suffisament régulière, cette formule approche l'intégrale souhaitée et on peut préciser l'erreur commise.

Théorème IV.2.8

Si on note $h = \max_{0\leq i\leq x_{m-1}} |x_{i+1} - x_i|$ le pas de la subdivision, pour toute fonction $f \in \mathcal{C}^{n+1}([a,b])$, on a

$$\left| \int_a^b f(x)\,dx - S_m(f) \right| \leq C|b-a| \|f^{(n+1)}\|_\infty h^{n+1}.$$

Autrement dit, l'ordre de convergence de la méthode composite est $n+1$.

Preuve :

On utilise la relation de Chasles et le résultat précédent

$$\begin{aligned}\left|\int_a^b f(x)\,dx - S_m(f)\right| &\leq \sum_{i=0}^{m-1}\left|\int_{x_i}^{x_{i+1}} f(x)\,dx - I_{x_i}^{x_{i+1}}(f)\right| \\ &\leq C\left(\sum_{i=0}^{m-1}|x_{i+1}-x_i|^{n+2}\right)\|f^{(n+1)}\|_\infty \\ &\leq Ch^{n+1}\left(\sum_{i=0}^{m-1}|x_{i+1}-x_i|\right)\|f^{(n+1)}\|_\infty \\ &= C|b-a|\|f^{(n+1)}\|_\infty h^{n+1}.\end{aligned}$$

■

2.2 Exemples

2.2.1 Formules de Newton-Cotes composites

Il s'agit de mettre en oeuvre l'approche générale ci-dessus dans le cas où la méthode de quadrature élémentaire I_0^1 est obtenue à partir de l'interpolation de Lagrange sur une subdivision uniforme de l'intervalle de référence $[0,1]$.

Méthode des rectangles à gauche La formule s'écrit

$$S_m^{RG}(f) = \sum_{i=0}^{m-1}(x_{i+1}-x_i)f(x_i).$$

Elle est du premier ordre.

Méthode des rectangles à droite La formule s'écrit

$$S_m^{RD}(f) = \sum_{i=0}^{m-1}(x_{i+1}-x_i)f(x_{i+1}).$$

Elle est également d'ordre 1.

Méthode du point milieu

$$S_m^{PM}(f) = \sum_{i=0}^{m-1}(x_{i+1}-x_i)f\left(\frac{x_{i+1}+x_i}{2}\right).$$

Celle-ci est d'ordre 2.

Méthodes des trapèzes La formule s'écrit

$$S_m^{T}(f) = \sum_{i=0}^{m-1}\frac{1}{2}(f(x_i)+f(x_{i+1}))(x_{i+1}-x_i).$$

Cette méthode est également d'ordre 2.

Remarquons au passage que l'on a

$$S_m^{T}(f) = \frac{1}{2}(S_m^{RG}(f)+S_m^{RD}(f)).$$

Remarque IV.2.9

En termes de performance (rapport précision/coût), la méthode du point-milieu et celle des trapèzes sont très comparables. Elles ont quand même quelques propriétés différentes. Par exemple, si la fonction f à intégrer est convexe, on peut montrer que la formule du point-milieu sous-estime l'intégrale, c'est-à-dire que

$$S_m^{PM}(f) \leq \int_a^b f(x)\,dx,$$

alors que la méthode des trapèzes, sur-estime l'intégrale

$$S_m^{T}(f) \geq \int_a^b f(x)\,dx.$$

Sauriez-vous le démontrer en quelques lignes ?

Méthode de Simpson

$$S_m^S(f) = \frac{x_1 - x_0}{6} f(x_0) + \frac{1}{6}\sum_{i=1}^{m-1}(x_{i+1} - x_{i-1}) f(x_i) + \frac{4}{6}\sum_{i=1}^{m}(x_i - x_{i-1}) f(\frac{x_i + x_{i-1}}{2}) + \frac{x_n - x_{n-1}}{6} f(x_n),$$

que l'on peut aussi écrire

$$S_m^S(f) = \frac{1}{6} S_m^{RG}(f) + \frac{1}{6} S_m^{RD}(f) + \frac{4}{6} S_m^{PM}(f).$$

Cette méthode est d'ordre 4.

2.2.2 Formules de Gauss composites

On peut appliquer toute la stratégie précédente à des formules de Gauss composites. La différence essentielle avec les formules composites de Newton-Cotes, vient du fait que les bords de l'intervalle élémentaire $[0, 1]$ ne sont jamais utilisés dans les méthodes de Gauss traditionnelles que nous avons vues précédemment.

C'est un (petit) inconvénient pratique car du coup aucune des valeurs de f calculées dans l'intervalle $[x_i, x_{i+1}]$ ne peut être ré-utilisée pour les intervalles voisins. Par ailleurs, il arrive assez souvent que la subdivision de l'intervalle choisie soit telle que les valeurs de f soient aisément accessibles en les x_i. C'est l'une des raisons pour lesquelles on peut préférer, par exemple, baser le calcul sur les méthodes élémentaires de Gauss-Lobatto que nous allons évoquer maintenant.

Considérons l'intervalle $[0, 1]$ et choisissons une subdivision en $n + 1$ points en **imposant** que le premier et le dernier point soient 0 et 1 respectivement. La subdivision s'écrit donc $x_0 = 0 < x_1 < ... < x_{n-1} < x_n = 1$. Une telle méthode ne peut avoir un degré d'exactitude supérieur ou égal à $2n$ car sinon, elle serait exacte sur le polynôme $q(x) = x(1-x)(x-x_1)^2...(x-x_{n-1})^2$. Or, la valeur de l'intégrale approchée de ce polynôme est nulle alors même qu'il garde un signe constant sur $]0, 1[$.

On va maintenant montrer qu'il y a un (unique) choix des points pour lesquels la méthode de quadrature obtenue est exactement de degré d'exactitude $2n - 1$. On considère pour cela la suite $(P_n)_n$ des polynômes orthogonaux de Legendre (i.e. pour le poids $w = 1$) sur l'intervalle $[0, 1]$. On va choisir pour $x_1, ..., x_{n-1}$ les $n - 1$ racines (qui sont distinctes et dans $]0, 1[$) du polynôme dérivé P_n'.

La formule de quadrature obtenue s'écrit

$$I_n(f) = \omega_0 f(0) + \sum_{i=1}^{n-1} \omega_i f(x_i) + \omega_n f(1).$$

Par construction, elle est exacte jusqu'au degré n (comme toutes les méthodes basées sur l'interpolation polynômiale). Si maintenant on prend un polynôme de degré inférieur ou égal à $2n - 1$ et qu'on effectue sa division euclidienne par $x(1 - x)L_n'(x)$, on voit que le reste (qui est de degré inférieur ou égal à n) s'intègre de façon exacte par la formule de quadrature. Le quotient q de cette division est degré $n - 2$ au plus et on doit montrer que la formule est exacte sur le produit $x(1 - x)L_n'(x)q(x)$. Par choix des points la méthode fournit la valeur nulle sur un tel polynôme, il faut donc nous assurer que l'intégrale exacte est bien nulle. Il s'agit de calculer par intégration par parties

$$\int_0^1 x(1-x)q(x)L_n'(x)\,dx = \underbrace{[x(1-x)q(x)L_n(x)]_0^1}_{=0} - \int_0^1 \frac{d}{dx}(x(1-x)q(x))L_n(x)\,dx,$$

cette dernière intégrale est nulle car le polynôme contre lequel on intègre L_n est de degré au plus $n-1$ et il est donc orthogonal à L_n.

Le résultat est donc démontré. La formule de Gauss-Lobatto ainsi construite sert alors de base à la construction d'une formule composite utilisant les valeurs de la fonction aux points de la subdivision de l'intervalle $[a,b]$.

2.3 Cas particulier des fonctions périodiques ou à support compact dans $]a,b[$. Formule d'Euler-MacLaurin

Supposons que la fonction $f:[a,b]\to\mathbb{R}$ que l'on cherche à intégrer soit
- ou bien périodique : au sens où f est la restriction à $[a,b]$ d'une fonction $(b-a)$ périodique sur $\mathbb{R}$.
- ou bien à support compact dans $]a,b[$: c'est-à-dire que f est identiquement nulle au voisinage de a et de b.

Dans ces deux cas, et sous réserve de la régularité de f, la formule des trapèzes composites construite sur une subdivision uniforme de l'intervalle possède des propriétés de convergence bien meilleures dues à des compensations entre les premiers termes de l'erreur.

Pour comprendre cela, regardons la formule des trapèzes élémentaires sur l'intervalle de référence $[0,1]$ et introduisons d'abord une notion utile pour le calcul qui va suivre.

Définition IV.2.10

On appelle polynômes de Bernoulli, la famille de polynômes $(B_p)_{p\geq 0}$ définis par, $B_0=1$ et par les formules de récurrence

$$B'_{p+1}(x)=(p+1)B_p(x),\quad \int_0^1 B_{p+1}(x)\,dx=0.$$

Ces polynômes vérifient $B_p(0)=B_p(1)$, pour $p\geq 2$ et cette valeur est notée b_p et appelée p-ième nombre de Bernoulli.

L'égalité des valeurs de B_p en 0 et 1 pour $p\geq 2$, provient du calcul suivant

$$B_p(1)-B_p(0)=\int_0^1 B'_p(x)\,dx=p\int_0^1 B_{p-1}(x)\,dx=0.$$

On a par ailleurs $B_1(x)=x-1/2$ d'où $B_1(1)=1/2$ et $B_1(0)=-1/2$.

Proposition IV.2.11

Soit $f:[0,1]\to\mathbb{R}$ de classe $\mathcal{C}^n$, $n\geq 1$. Nous avons alors la formule

$$\frac{1}{2}f(0)+\frac{1}{2}f(1)=\int_0^1 f(x)\,dx + \sum_{p=1}^{n-1}(-1)^{p+1}\frac{b_{p+1}}{(p+1)!}\left(f^{(p)}(1)-f^{(p)}(0)\right)+\frac{(-1)^{n+1}}{n!}\int_0^1 B_n(x)f^{(n)}(x)\,dx.$$

Preuve :

Il s'agit juste d'intégrer par parties n fois. On commence par

$$\int_0^1 f(x)\,dx=\int_0^1 f(x)B_0(x)\,dx=\int_0^1 f(x)B'_1(x)\,dx$$
$$=[f(x)B_1(x)]_0^1-\int_0^1 f'(x)B_1(x)\,dx=\frac{1}{2}f(0)+\frac{1}{2}f(1)-\int_0^1 f'(x)B_1(x)\,dx,$$

ce qui est bien la formule attendue pour $n=1$. On continue ensuite d'intégrer par parties pour faire apparaître les dérivées successives de f. ∎

On en déduit, par un changement de variable affine, une formule similaire sur un intervalle $[x_{i-1},x_i]$ de longueur h qui s'écrit

$$\frac{h}{2}f(x_{i-1})+\frac{h}{2}f(x_i)=\int_{x_{i-1}}^{x_i} f(x)\,dx+\sum_{p=1}^{n-1}(-1)^{p+1}h^{p+1}\frac{b_{p+1}}{(p+1)!}\left(f^{(p)}(x_i)-f^{(p)}(x_{i-1})\right)$$
$$+h^n\frac{(-1)^{n+1}}{n!}\int_{x_{i-1}}^{x_i} B_n\left(\frac{x-x_{i-1}}{x_i-x_{i-1}}\right)f^{(n)}(x)\,dx.$$

En sommant ces égalités, sur une subdivision uniforme de $[a,b]$, on remarque que les termes de dérivées de f aux noeuds de la discrétisation se compensent exactement. On obtient ainsi la formule d'Euler-Mac Laurin qui s'écrit

$$S_m^T(f) = \int_a^b f(x)\,dx + \sum_{p=1}^{n-1}(-1)^{p+1}h^{p+1}\frac{b_{p+1}}{(p+1)!}\left(f^{(p)}(b) - f^{(p)}(a)\right) + h^n\frac{(-1)^{n+1}}{n!}\int_a^b \tilde{B}_{n,m}(x)f^{(n)}(x)\,dx, \quad \text{(IV.1)}$$

où $\tilde{B}_{n,m}$ est une fonction bornée (définie sur chacun des m sous-intervalles à partir de B_n par changement de variable affine).

On en conclut le résultat suivant.

Théorème IV.2.12

Si $f : [a,b] \to \mathbb{R}$ est de classe $\mathcal{C}^n$ est périodique, ou à support compact dans $]a,b[$, la formule des trapèzes sur une subdivision uniforme de l'intervalle vérifie

$$\left|S_m^T(f) - \int_a^b f(x)\,dx\right| \leq C_n(b-a)\|f^{(n)}\|_\infty h^n,$$

où C_n ne dépend que de n.

Preuve :

Si f est à support compact, tous les termes de dérivées de f en a et b sont nuls. Si f est périodique, on a $f^{(p)}(a) = f^{(p)}(b)$ pour tout p et donc cette somme s'annule aussi.

Dans les deux cas, il reste

$$S_m^T(f) = \int_a^b f(x)\,dx + h^n\frac{(-1)^{n+1}}{n!}\int_a^b \tilde{B}_{n,m}(x)f^{(n)}(x)\,dx.$$

■

Remarque IV.2.13

Dans le cas de fonctions périodiques, la méthode des rectangles à gauche ou à droite a exactement la même propriété.

Dans le cas général : La formule d'Euler-MacLaurin permet d'utiliser la méthode des trapèzes de façon astucieuse pour accéder à une convergence d'ordre supérieurs (pour peu que f soit régulière).

Pour fixer les idées, on fixe $n = 3$ dans (IV.1). De sorte que l'on a

$$\left|S_m^T(f) - \int_a^b f(x)\,dx - \alpha h^2\right| \leq Ch^3\|f^{(3)}\|_\infty, \ \ \forall m \geq 0,$$

où le coefficient α est fixé et ne dépend que des dérivées de f en a et b et des nombres de Bernoulli.

L'idée de le méthode d'accélération, consiste à combiner astucieusement les valeurs approchées obtenues avec m et $2m$ points d'intégrations (ce qui divise le pas h par 2). On obtient

$$\left|S_m^T(f) - \int_a^b f(x)\,dx - \alpha h^2\right| \leq Ch^3\|f^{(3)}\|_\infty,$$

$$\left|S_{2m}^T(f) - \int_a^b f(x)\,dx - \alpha\frac{h^2}{4}\right| \leq \frac{C}{8}h^3\|f^{(3)}\|_\infty,$$

de sorte que

$$\left|\frac{4S_{2m}^T(f) - S_m^T(f)}{4-1} - \int_a^b f(x)\,dx\right| \leq C'h^3\|f^{(3)}\|_\infty.$$

Ainsi la formule

$$\frac{4S_{2m}^T(f) - S_m^T(f)}{4-1}$$

est une approximation à l'ordre 3 de l'intégrale recherchée. Par ailleurs, l'évaluation de S_{2m}^T n'utilise que m évaluations de f supplémentaires par rapport à S_m^T, ce qui fait que la méthode ainsi obtenue a le même coût de calcul que S_{2m}^T.

Cette approche peut se généraliser pour obtenir une convergence à un ordre arbitraire (dépendant de la régularité de la fonction à intégrer). Il s'agit de la méthode de Romberg, que vous pourrez trouver dans [Dem91, p. 83]

3 Quelques commentaires et compléments

3.1 Sur le calcul des points et des poids des méthodes de Gauss

Il s'agit de calculer les racines du n-ième polynôme orthogonal associé au poids de quadrature ainsi que les poids. Nous verrons dans le chapitre suivant une méthode basée sur la bissection/dichtomie pour résoudre se problème.

On va montrer ici comment ramener le problème à un calcul de valeurs propres. C'est la méthode connue sous le nom de méthode de Golub-Welsh. Elle est basée sur la relation de récurrence linéaire vérifiée par les polynômes orthogonaux.

$$P_{n+1}(x) = (x - a_n)P_n(x) - b_n P_{n-1}(x),$$

Si x est une racine de P_{n+1}, on a les relations

$$0 = (x - a_n)P_n(x) - b_n P_{n-1}(x),$$

$$P_n(x) = (x - a_{n-1})P_{n-1}(x) - b_{n-1} P_{n-2}(x),$$

et ainsi de suite, jusqu'à

$$P_1(x) = (x - a_0)P_0(x), \text{ et } P_0(x) = 1.$$

Ainsi, on observe que l'on a

$$\underbrace{\begin{pmatrix} a_0 & 1 & 0 & & \\ b_1 & a_1 & 1 & 0 & \\ 0 & b_2 & a_2 & 1 & \\ & \ddots & \ddots & \ddots & \ddots \\ & & 0 & b_n & a_n \end{pmatrix}}_{=J} \underbrace{\begin{pmatrix} P_0(x) \\ P_1(x) \\ \\ P_n(x) \end{pmatrix}}_{=\varphi_x} = x \begin{pmatrix} P_0(x) \\ P_1(x) \\ \\ P_n(x) \end{pmatrix}.$$

Autrement dit x est une valeur propre de J pour le vecteur propre φ_x.

Ainsi, les $n+1$ racines de P_{n+1} sont exactement les $n+1$ valeurs propres de J. De plus, φ_x est l'unique vecteur propre associé dont la première composante vaut 1 (par convention $P_0(x) = 1$).

Notons Φ la matrice dont les colonnes sont les vecteurs $\varphi_{x_0}, ..., \varphi_{x_n}$ correspondant aux $n+1$ valeurs propres. On a donc

$$J\Phi = \Phi D,$$

où D est la matrice diagonale contenant les x_i sur la diagonale.

On sait que la formule de quadrature est exacte sur tous les polynômes de degré inférieur ou égal à $2n+1$ elle est donc en particulier exacte sur tous les produits $p_j p_k$ avec $j, k \in \{0, ..., n\}$, ce qui donne pour tout j et k

$$\|p_j\|_w^2 \delta_{jk} = (p_j, p_k)_w = \sum_{i=0}^{n} \omega_i p_j(x_i) p_k(x_i).$$

Si on note Ω la matrice diagonale contenant les poids ω_i que l'on cherche on a donc obtenu

$$\underbrace{\operatorname{diag}(\|p_j\|_w^2)}_{=D'} = \Phi \Omega {}^t\Phi,$$

ce qui donne

$$\Phi^{-1} D' {}^t\Phi^{-1} = \Omega,$$

puis

$$\Omega^{-1} = {}^t\Phi (D')^{-1} \Phi.$$

Ceci donne

$$\frac{1}{\omega_i} = \sum_{j=0}^{n} |p_j(x_i)|^2 \frac{1}{\|p_j\|_w^2}.$$

Les normes $\|p_j\|_w$ sont calculables à partir des coefficients b_j par les formules

$$\|p_0\|_w = \left(\int_a^b w(x)\,dx\right)^{\frac{1}{2}},$$

$$b_j = \frac{\|P_j\|_w^2}{\|P_{j-1}\|_w^2}.$$

3.2 Calcul approché d'intégrales généralisées

Par coupure de l'intervalle d'intégration Soit à évaluer l'intégrale $\int_a^{+\infty} f(x)\,dx$ que l'on suppose convergente. Pour obtenir cette valeur avec une précision ε souhaitée, on procède en deux temps :

- On détermine un réel $b > a$ tel que
$$\int_b^{+\infty} f(x)\,dx \leq \frac{\varepsilon}{2},$$
celui-ci existant toujours par définition de la convergence de l'intégrale. Il y a plusieurs façons possibles de déterminer un tel réel b.
- Une fois déterminé ce réel b, on évalue l'intégrale définie $\int_a^b f$ par l'une des formules de quadratures étudiées précédemment (ou par n'importe quel autre méthode du même type). On détermine alors le nombre de points de la discrétisation nécessaire pour que la formule de quadrature $S_m(f)$ approche $\int_a^b f$ à $\frac{\varepsilon}{2}$ près.

In fine on a donc

$$\left|\int_a^{+\infty} f - S_m(f)\right| \leq \left|\int_a^b f + \int_b^{+\infty} f - S_m(f)\right| \leq \varepsilon.$$

Par les méthodes de Gauss-Laguerre On reprend toute la théorie précédente sur l'intervalle $[0, +\infty[$ avec le poids e^{-x}. Les polynômes orthogonaux correspondants sont appelés polynômes de Laguerre et vérifient

$$L_n(x) = (-1)^n e^x \frac{d}{dx^n}\left(e^{-x}x^n\right),$$

et la relation de récurrence ($L_{-1} = 0$, $L_0 = 1$)

$$L_{n+1}(x) = (x - 2n - 1)L_n(x) - n^2 L_{n-1}(x).$$

Si on fixe n et qu'on considère $0 < x_0 < ... < x_n$ les $n+1$ racines de L_{n+1}, alors la formule de Gauss-Laguerre pour approcher l'intégrale $\int_0^{+\infty} e^{-x} f(x)\,dx$ est donnée par

$$I_n(f) = \sum_{i=0}^{n} \omega_i f(x_i),$$

avec les ω_i adéquats.

Si f est suffisamment régulière et que ses dérivées ne croissent pas trop vite à l'infini, la formule est très précise.

Exemple : calcul d'une transformée de Laplace : Soit f convenable et F sa transformée de Laplace sur $\mathbb{R}^+$ définie par

$$F(t) = \int_0^{+\infty} e^{-tx} f(x)\,dx, \quad \forall t > 0,$$

alors on peut effectuer le changement de variable $y = tx$ et obtenir

$$F(t) = \frac{1}{t}\int_0^{+\infty} e^{-y} f(y/t)\,dy,$$

puis appliquer la formule de Gauss-Laguerre à cette intégrale

$$F(t) \sim \frac{1}{t}\sum_{i=0}^{n} \omega_i f\left(\frac{x_i}{t}\right).$$

Par les méthodes de Gauss-Hermite Pour calculer des intégrales du type

$$\int_{-\infty}^{+\infty} e^{-x^2} f(x)\,dx,$$

qui apparaissent souvent en théorie des probabilités notamment. On peut utiliser toute la théorie précédente avec l'espace $L^2_w(]-\infty, +\infty[)$ où le poids est la fonction gaussienne $w(x) = e^{-x^2}$.

Les polynômes orthogonaux ainsi engendrés s'appelle les polynômes de Hermite et la formule de quadrature de Gauss correspondante se construit de manière similaire à celle de Gauss-Legendre évoquée ci-dessus.

3.3 Intégrales 2D/3D

Le calcul d'intégrales multidimensionnelles est un problème très important et beaucoup plus délicat que celui des intégrales 1D. La plupart des stratégies consistent à découper le domaine en sous-domaines de géométrie simple (des triangles, des quadrangles, des tétraèdres, etc ...). L'intégrale complète est alors la somme d'intégrales élémentaires.

Pour chacun d'entre eux, on effectue un changement de variable simple (affine ou quadratique) qui ramène le calcul à celui de l'intégrale d'une nouvelle fonction (attention au Jacobien du changement de variable !) sur un domaine de référence (triangle simplexe unité, carré unité, etc ...). Sur ces domaines de référence, on peut mettre en place des formules de quadrature Gaussiennes de tout ordre, dont les principales sont tabulées et disponibles dans la littérature.

La qualité de l'approximation dépend alors de la régularité de la fonction à intégrer, de la taille des sous-domaines (pas du maillage), du degré d'exactitude de la formule choisie.

Une autre méthode (dans les cas simples) consiste à utiliser le théorème de Fubini pour se ramener au cas d'intégrales monodimensionnelles.

Ainsi, si l'on doit calculer l'intégrale de f sur un domaine de la forme

$$\Omega = \big\{(x,y) \in \mathbb{R}^2,\ \ a \leq x \leq b,\ \ \alpha(x) \leq y \leq \beta(x)\big\},$$

on peut écrire

$$\int_\Omega f\,dx\,dy = \int_a^b dx \left(\int_{\alpha(x)}^{\beta(x)} f(x,y)\,dy \right),$$

puis appliquer une méthode de quadrature sur l'intervalle $[a,b]$ puis sur chaque intervalle $[\alpha(x_i), \beta(x_i)]$.

Chapitre V

Exercices

Exercice 1

1. *Soit $a > 0$. Ecrire la méthode de Newton pour la fonction $f_1(x) = ax - 1$. Commentaire ?*
2. *Ecrire maintenant la méthode de Newton pour la fonction $f_2(x) = (1/x) - a$. Voyez-vous un intérêt à cette méthode ? Etudier le comportement de la suite $(x_n)_n$ en fonction de x_0.*
3. *Soit maintenant $A \in GL_d(\mathbb{R})$ une matrice inversible.*
   (a) *Montrer $GL_d(\mathbb{R})$ est un ouvert de $M_d(\mathbb{R})$.*
   (b) *Montrer que $M \in GL_d(\mathbb{R}) \mapsto M^{-1}$ est différentiable et calculer sa différentielle en tout point.*
   (c) *Ecrire la méthode de Newton pour résoudre l'équation*

$$X^{-1} = A,$$

et montrer que la suite de matrices ainsi définies converge vers A^{-1} si et seulement si $\rho(\mathrm{Id} - AX_0) < 1$.

Solution :

1. La méthode de Newton dans ce cas, s'écrit simplement

$$x_{n+1} = x_n - \frac{f_1(x_n)}{f_1'(x_n)} = \frac{1}{a},$$

autrement dit, la méthode converge en une seule itération vers $1/a$. Ca n'a donc bien évidemment aucun intérêt.

2. La fonction f_2 a le même zéro que f_1 mais les itérées de Newton sont maintenant données par

$$x_{n+1} = x_n - \frac{f_2(x_n)}{f_2'(x_n)} = x_n - \frac{1/x_n - a}{-1/x_n^2} = 2x_n - ax_n^2.$$

Le calcul des itérées successives ne nécessite que des produits et additions (et pas de divisions).
Pour étudier le comportement global de cette suite, on pose $y_n = x_n - 1/a$ et on obtient la récurrence

$$y_{n+1} = -ay_n^2.$$

Ceci prouve déjà que, quelle que soit la donnée initiale, pour tout $n \geq 1$, on a $y_n \leq 0$, c'est-à-dire $x_n \leq 1/a$. Si maintenant on pose $z_n = ay_n$, on obtient

$$z_{n+1} = -z_n^2,$$

et donc

$$z_n = -(z_0)^{2^n}.$$

Conclusions :
- Si $z_0 = \pm 1$, i.e. $x_0 = 0$ ou $x_0 = 2/a$, alors la suite $(z_n)_n$ est constante égale à -1 pour tout $n \geq 1$, ce qui signifie que l'on a
$$x_n = 0, \ \forall n \geq 0.$$
- Si $|z_0| < 1$, i.e. si $0 < x_0 < 2/a$, alors la suite $(z_n)_n$ tend vers 0 quadratiquement, ce qui montre que $(x_n)_n$ tend vers $1/a$ quadratiquement.
- Si $|z_0| > 1$, i.e. si $x_0 < 0$ ou $x_0 > 2/a$, alors la suite $(z_n)_n$ tend vers $-\infty$ et il en est donc de même de la suite $(x_n)_n$.

3. (a) $GL_d(\mathbb{R})$ est l'image réciproque de $\mathbb{R}^*$ par le déterminant (qui est une application continue car polynomiale), c'est donc un ouvert de $M_d(\mathbb{R})$.
   (b) Soit $M \in GL_d(\mathbb{R})$ et $H \in M_d(\mathbb{R})$. Supposons que $\|HM^{-1}\| < 1$ (pour une certaine norme matricielle fixée), alors on sait que $I + HM^{-1}$ est inversible (Lemme de Von Neumann) et que l'on a le développement en série
$$(I + HM^{-1})^{-1} = \sum_{k \geq 0} (-HM^{-1})^k.$$
On peut donc écrire
$$(M + H)^{-1} = M^{-1}(I + HM^{-1})^{-1} = M^{-1}\left(\sum_{k \geq 0} (-HM^{-1})^k\right).$$

En isolant les termes d'ordre 0 et 1 dans cette écriture, on obtient immédiatement

$$(M+H)^{-1} = M^{-1} - M^{-1}HM^{-1} + O(\|H\|^2).$$

Ceci prouve que $\varphi : M \mapsto M^{-1}$ est différentiable en tout point de $GL_d(\mathbb{R})$ et que sa différentielle en un point M est donnée par

$$D\varphi(M).H = -M^{-1}HM^{-1}.$$

On peut même calculer l'inverse de cette application :

$$(D\varphi(M))^{-1}.Y = -MYM, \ \ \forall Y \in M_d(\mathbb{R}).$$

(c) La méthode de Newton pour la fonction $X \mapsto \varphi(X) - A$ s'écrit sous la forme

$$X_{n+1} = X_n - (D\varphi(X_n))^{-1}(\varphi(X_n) - A),$$

ce qui donne ici

$$X_{n+1} = X_n + X_n(X_n^{-1} - A)X_n = 2X_n - X_nAX_n. \tag{1}$$

On trouve bien une généralisation du cas monodimensionel étudié plus haut.

Noter qu'il est *a priori* nécessaire que X_n soit inversible pour pouvoir écrire l'itération suivante (pour que $\varphi(X_n)$ soit bien défini !) mais, en réalité, on voit que la formule (1) est bien définie pour tout n, sans condition.

De manière similaire à ce que nous avons fait plus haut, on pose $Z_n = I - AX_n$ de sorte que l'on obtienne l'équation

$$Z_{n+1} = Z_n^2,$$

et donc

$$Z_n = (Z_0)^{2^n}.$$

Un théorème du cours nous montre que la suite Z_n tend vers 0 si et seulement si le rayon spectal de Z_0 est strictement inférieur à 1.

■

Exercice 2

Soit $A \in M_d(\mathbb{R})$ une matrice symétrique définie positive et $b \in \mathbb{R}^d$. On introduit la fonctionnelle J définie par

$$J(x) = \frac{1}{2}(Ax, x) - (b, x), \quad \forall x \in \mathbb{R}^d.$$

1. *Montrer que J admet un unique minimiseur x^* sur $\mathbb{R}^d$ et que celui-ci est caractérisé par l'équation $Ax^* = b$.*
2. *On considère la méthode de gradient à pas constant pour calculer ce minimiseur*

$$x_{n+1} = x_n - \alpha \nabla J(x_n).$$

Montrer que la suite $(x_n)_n$ converge vers x^* si et seulement si $0 < \alpha < 2/\rho(A)$.

Solution :

1. On remarque déjà que $J(0) = 0$ et donc que l'infimum de J est négatif ou nul (éventuellement égal à $-\infty$).
Comme A est définie positive on a l'inégalité suivante

$$(Ax, x) \geq \lambda_1 \|x\|^2, \ \ \forall x \in \mathbb{R}^d,$$

où $\lambda_1 > 0$ est la plus petite valeur propre de A.
Ainsi, on obtient la minoration suivante

$$J(x) \geq \frac{\lambda_1}{2}\|x\|^2 - (b, x) \geq \frac{\lambda_1}{2}\|x\|^2 - \|b\|\|x\|.$$

On en déduit que

$$J(x) \leq 0 \Longrightarrow \|x\| \leq \|b\|.$$

Il s'ensuit que

$$\inf_{\mathbb{R}^d} J = \inf_{\overline{B}(0,\|b\|)} J.$$

On est ramenés à la minimisation de la fonctionnelle continue J sur le compact $\overline{B}(0, \|b\|)$. Par compacité, ce problème admet au moins une solution notée x^*.
Comme la fonctionnelle J est de classe $\mathcal{C}^\infty$ (elle est polynômiale), sa différentielle soit s'annuler en tout point de minimum de x^*. On calcule la différentielle de la façon suivante

$$J(x+h) = \frac{1}{2}(A(x+h), x+h) - (b, x+h) = J(x) + (Ax - b, h) + \frac{1}{2}(Ah, h),$$

ce qui montre que

$$DJ(x).h = (Ax - b, h),$$

ou encore

$$\nabla J(x) = Ax - b.$$

Tout minimiseur doit donc vérifier $Ax^* = b$ et comme A est inversible, on déduit qu'il en existe un unique.

2. Ayant calculé le gradient de J à la précédente question, on voit que les itérations s'écrivent

$$x_{n+1} = x_n - \alpha(Ax_n - b),$$

et comme $Ax^* = b$, on pose $y_n = x_n - x^*$ et on obtient

$$y_{n+1} = (\mathrm{Id} - \alpha A)y_n.$$

Comme la matrice $\mathrm{Id} - \alpha A$ est diagonalisable à valeurs propres réelles, on voit que la suite $(y_n)_n$ tend vers 0 (pour tout choix de la donnée initiale) si et seulement si

$$\rho(\mathrm{Id} - \alpha A) < 1.$$

Comme les valeurs propres de A sont positives, cette condition n'est certainement pas vérifiée si $\alpha \leq 0$. Dans le cas $\alpha > 0$, il suffit de s'assurer que la plus petite valeur propre de $\mathrm{Id} - \alpha A$ est strictement plus grande que -1, ce qui donne la condition $\alpha < 2/\rho(A)$.

■

Exercice 3 (Assez difficile ...)

La méthode de la sécante pour la résolution de $f(x) = 0$ consiste à itérer, à partir de données initiales x_0 et x_1 distinctes et telles que $f(x_0) \neq f(x_1)$, de la façon suivante

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

1. *Illustrez sur un dessin la construction des itérées successives. Commentez les différences par rapport à la méthode de Newton.*
2. *Quels peuvent être les avantages, en pratique, de cette méthode par rapport à celle de Newton ?*
3. *On suppose que f a un zéro simple noté x^*. On a donc $f(x^*) = 0$ et $f'(x^*) \neq 0$.*
   (a) *Montrer qu'il existe $\delta > 0$ tel que*
   $$\left| \frac{f'(x)}{f'(y)} - 1 \right| \leq \frac{1}{2}, \ \ \forall x, y \in [x^* - \delta, x^* + \delta].$$
   Dans la suite, une telle valeur de δ est fixée et on note $I_\delta = [x^* - \delta, x^* + \delta]$.
   (b) *Montrer que si $x_0, x_1 \in I_\delta$, alors, la suite $(x_n)_n$ est bien définie, contenue dans I_δ et vérifie*
   $$|x_{n+1} - x^*| \leq \frac{1}{2}|x_n - x^*|.$$
   (c) *En déduire que la suite $(x_n)_n$ converge vers x^*.*
4. *On va montrer maintenant que la convergence est sur-linéaire, mais pas quadratique. On suppose pour cela que $f''(x^*) \neq 0$.*
   (a) *Pour tout $n \geq 0$, on pose $e_n = x_n - x^*$ et on note $C = \frac{f''(x^*)}{(2f'(x^*))}$. Montrer que, pour tout n, il existe $\alpha_n, \beta_n, \gamma_n \in I_\delta$ tels que*
   $$e_{n+1} = e_n - \frac{f'(\alpha_n) e_n}{f'(x^*) + e_{n-1} f''(\gamma_n) + \frac{f''(\beta_n)}{2}(e_n - e_{n-1})},$$
   et telles que $\alpha_n \to x^*$, $\beta_n \to x^*$, $\gamma_n \to x^*$, avec $|\alpha_n - x^*| \leq |e_n|$.
   (b) *Montrer que $e_n / e_{n-1} \to 0$.*
   (c) *Montrer qu'on a*
   $$e_{n+1} \sim C e_n e_{n-1}.$$
   (d) *Soit $p = \frac{1+\sqrt{5}}{2} > 1$ le nombre d'or. On rappelle qu'il vérifie $p(p-1) = 1$. On pose*
   $$y_n = \frac{|e_{n+1}|}{|C|^{1/p} |e_n|^p}.$$
   Montrer que
   $$y_{n+1} \sim y_n^{1-p}.$$
   (e) *En déduire qu'il existe une suite $(\varepsilon_n)_n$ tendant vers 0 telle que*
   $$|\log y_{n+1}| \leq (p-1)|\log y_n| + |\varepsilon_n|, \ \ \forall n \geq 1.$$
   (f) *Montrer que $(\log y_n)_n$ tend vers 0.*
   (g) *Conclure à l'ordre de convergence de la méthode.*

Solution :

1. Graphiquement, au lieu de tirer la tangente à la courbe représentative de f au point x_n, on tire la corde construite sur $(x_{n-1}, f(x_{n-1}))$ et $(x_n, f(x_n))$.
2. Cette méthode ne nécessite pas le calcul de la dérivée de f, de plus elle n'utilise que des valeurs de f déjà calculée précédemment. Si f est une fonction coûteuse à évaluer la méthode est donc assez économe en terme de temps de calcul.

3. (a) Comme $f'(x^*)$ est non nul, on a

$$\lim_{\delta\to 0}\inf_{I_\delta}|f'| = \lim_{\delta\to 0}\sup_{I\delta}|f'| = |f'(x^*)| > 0,$$

et donc il existe un $\delta > 0$ assez petit, tel que

$$\sup_{I_\delta}|f'| \leq \frac{3}{2}\inf_{I_\delta}|f'|.$$

En particulier on a donc

$$|f'(x)| \leq \frac{3}{2}|f'(y)|, \ \ \forall x, y \in I_\delta,$$

mais aussi

$$|f'(y)| \leq 2|f'(x)|, \ \ \forall x, y \in I_\delta.$$

Ceci montre que

$$\frac{1}{2} \leq \frac{|f'(x)|}{|f'(y)|} \leq \frac{3}{2}, \ \ \forall x, y \in I_\delta.$$

Comm f' ne change pas de signe sur I_δ (sinon on aurait $\inf_{I_\delta}|f'| = 0$ ce qui est exclu d'après l'inégalité ci-dessus), cela donne le résultat.

(b) Il suffit de montrer par récurrence que si x_{n-1} et x_n sont dans I_δ, alors x_{n+1} est aussi dans I_δ. En effet, s'il arrive que x_{n+1} et x_n sont égaux (alors la méthode ne serait pas bien définie au cran suivant) cela signifie que $f(x_n)$ était nul et donc que la solution avait été trouvée de façon exacte et la méthode arrêtée à l'itération en question.
Pour cela, on écrit

$$\begin{aligned} x_{n+1} - x^* &= x_n - x^* - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n) \\ &= (x_n - x^*)\left(1 - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}\frac{f(x_n) - f(x^*)}{x_n - x^*}\right) \\ &= (x_n - x^*)\left(1 - \frac{f'(\xi_n)}{f'(\zeta_n)}\right), \end{aligned}$$

où ξ_n et ζ_n sont dans I_δ et donnés par le théorème des accroissements finis. D'après la question précédente, la seconde parenthèse est bien plus petite que $1/2$ en valeur absolue. On en déduit que

$$|x_{n+1} - x^*| \leq \frac{1}{2}|x_n - x^*|,$$

et en particulier que x_{n+1} est bien dans I_δ.

(c) D'après l'estimation précédente, on a bien convergence de la suite $(x_n)_n$ vers x^* avec une estimation, pour l'instant, linéaire de l'erreur.

4. (a) On écrit à nouveau (en utilisant que $f(x^*) = 0$)

$$e_{n+1} = x_{n+1} - x^* = e_n - \frac{e_n - e_{n-1}}{f(x_n) - f(x_{n-1})}(f(x_n) - f(x^*))$$

La formule de Taylor à l'ordre 2 donne

$$f(x_n) = f(x_{n-1}) + (e_n - e_{n-1})f'(x_{n-1}) + \frac{1}{2}(e_n - e_{n-1})^2 f''(\beta_n),$$

où β_n est compris entre x_{n-1} et x_n (en particulier cette suite tend vers x^* par le théorème des gendarmes).
De la même façon, on a

$$f(x_n) = f(x^*) + e_n f'(\alpha_n),$$

où α_n tend vers x^* et même $|\alpha_n - x^*| \leq |e_n|$. Enfin, toujours par le théorème des accroissements finis, on peut écrire

$$f'(x_{n-1}) = f'(x^*) + e_{n-1}f''(\gamma_n),$$

avec $\gamma_n \to x^*$.

En reportant tout cela dans la formule ci-dessus, on trouve

$$\begin{aligned} e_{n+1} &= e_n - \frac{(e_n - e_{n-1})e_n f'(\alpha_n)}{(e_n - e_{n-1})f'(x_{n-1}) + \frac{1}{2}(e_n - e_{n-1})^2 f''(\beta_n)} \\ &= e_n - \frac{e_n f'(\alpha_n)}{f'(x_{n-1}) + \frac{1}{2}(e_n - e_{n-1})f''(\beta_n)} \\ &= e_n - \frac{e_n f'(\alpha_n)}{f'(x^*) + e_{n-1}f''(\gamma_n) + \frac{1}{2}(e_n - e_{n-1})f''(\beta_n)} \end{aligned}$$

(b) On suppose que e_n n'est jamais nul (sinon cela signifie qu'on a atteint la solution exacte et l'analyse asymptotique n'a plus lieu d'être). En divisant la précédent formule par e_n, on trouve

$$\frac{e_{n+1}}{e_n} = 1 - \frac{f'(\alpha_n)}{f'(x^*) + e_{n-1}f''(\gamma_n) + \frac{1}{2}(e_n - e_{n-1})f''(\beta_n)},$$

et on peut faire tendre n vers l'infini en utilisant la convergence des suites $(\alpha_n)_n$, $(\beta_n)_n$ et $(\gamma_n)_n$ ainsi que le fait que $(e_n)_n$ tend vers 0.
Il vient immédiatement

$$\lim_{n\to+\infty} \frac{e_{n+1}}{e_n} = 1 - \frac{f'(x^*)}{f'(x^*)} = 0.$$

(c) On reprend la formule ci-dessous que l'on divise par e_{n-1}, il vient

$$\frac{e_{n+1}}{e_n e_{n-1}} = \frac{\frac{f'(x^*)-f'(\alpha_n)}{e_{n-1}} + f''(\gamma_n) + \frac{1}{2}(\frac{e_n}{e_{n-1}} - 1)f''(\beta_n)}{f'(x_{n-1}) + e_{n-1}f''(\gamma_n) + \frac{1}{2}(e_n - e_{n-1})f''(\beta_n)}.$$

En utilisant le résultat de la question précédente, on peut clairement passer à la limite dans tous les termes. Le plus délicat étant le premier terme du numérateur que l'on majore par

$$\left|\frac{f'(x^*) - f'(\alpha_n)}{e_{n-1}}\right| \leq (\sup|f''|)\frac{|x^* - \alpha_n|}{|e_{n-1}|} \leq (\sup|f''|)\frac{|e_n|}{|e_{n-1}|} \to 0.$$

In fine, on obtient la limite

$$\lim_{n\to\infty} \frac{e_{n+1}}{e_n e_{n-1}} = \frac{f''(x^*)}{2f'(x^*)},$$

qui est le nombre C recherché.

(d) On effectue le calcul suivant en utilisant le fait que $p(p-1) = 1$,

$$y_{n+1}y_n^{p-1} = \frac{|e_{n+2}|}{|C|^{1/p}|e_{n+1}|^p}\frac{|e_{n+1}|^{p-1}}{|C|^{(p-1)/p}|e_n|^{p(p-1)}} = \frac{|e_{n+2}|}{|C||e_{n+1}||e_n|}$$

et ceci converge bien vers 1 d'après la question précédente.

(e) On écrit

$$\log y_{n+1} = \log\left(\frac{y_{n+1}}{y_n^{1-p}}\right) + \log\left(y_n^{1-p}\right),$$

et on trouve le résultat en passant à la valeur absolue et en posant

$$\varepsilon_n = \log\left(\frac{y_{n+1}}{y_n^{1-p}}\right),$$

qui tend bien vers 0 d'après la question précédente.

(f) Le point clé ici est de constater que $0 < p - 1 < 1$ (sinon il n'y a aucune chance que le résultat fonctionne). On voit que si ε_n était nul, on aurait simplement affaire à une suite géométrique tendant vers 0. Pour simplifier les notations, on pose $z_n = |\log y_n|$.
Soit $\gamma > 0$ fixé. Il existe $N \geq 0$ tel que $|\varepsilon_n| \leq \gamma$ pour $n \geq N$. Pour de telles valeurs de n, on a

$$z_{n+1} \leq (p-1)z_n + \delta,$$

ce qui donne

$$\left(z_{n+1} - \frac{\delta}{2-p}\right) \leq (p-1)\left(z_n - \frac{\delta}{2-p}\right).$$

Ainsi, on obtient une majoration

$$\left(z_n - \frac{\delta}{2-p}\right) \leq (p-1)^{n-N}\left(z_N - \frac{\delta}{2-p}\right),$$

et comme $p-1<1$, on peut passer à la limite supérieure et obtenir que

$$\limsup_{n\to\infty}\left(z_n - \frac{\delta}{2-p}\right) \leq 0,$$

ou encore que

$$\limsup_{n\to\infty} z_n \leq \frac{\delta}{2-p}.$$

Comme ceci est vrai pour tout $\delta>0$, on déduit que $\limsup_{n\to\infty} z_n = 0$ ce qui montre que $(z_n)_n$ tend vers 0 car c'est une suite à termes positifs.

(g) La question précédente montre que $(\log y_n)_n$ tend vers 0, ce qui montre que $(y_n)_n$ tend vers 1 et on obtient donc l'équivalent

$$|e_{n+1}| \sim |C|^{1/p}|e_n|^p.$$

L'ordre de la méthode est donc exactement le nombre d'or p. C'est un ordre intermédiaire entre le cas linéaire (méthode de point fixe standard) et le cas quadratique (Newton).

■

Exercice 4

Soit $\varphi : \mathbb{R}^d \mapsto \mathbb{R}^d$ une application (non nécessairement linéaire) vérifiant
– φ est Lipschitzienne sur $\mathbb{R}^d$.
– φ est monotone au sens suivant : il existe $\alpha > 0$ tel que

$$(\varphi(x) - \varphi(y), x - y) \geq \alpha \|x - y\|^2, \ \ \forall x, y \in \mathbb{R}^d,$$

où $(\cdot, \cdot)$ est le produit scalaire euclidien sur $\mathbb{R}^d$.
Montrer que pour tout $b \in \mathbb{R}^d$, il existe un unique $x^* \in \mathbb{R}^d$ solution de

$$\varphi(x^*) = b,$$

et proposer une méthode itérative permettant de calculer cette solution.

Solution :

On va chercher une solution de $\varphi(x) = b$ sous la forme d'un point fixe de l'application

$$\Phi(x) = x - \rho(\varphi(x) - b),$$

pour un choix de $\rho > 0$ convenable.

On va montrer que pour un ρ bien choisi, l'application Φ est contractante. Calculons

$$\|\Phi(x) - \Phi(y)\|^2 = \|(x - y) - \rho(\varphi(x) - \varphi(y))\|^2 = \|x - y\|^2 - 2\rho(x - y, \varphi(x) - \varphi(y)) + \rho^2 \|\varphi(x) - \varphi(y)\|^2.$$

Si on note L la constante de Lipschitz de φ et l'hypothèse de monotonie, on trouve

$$\|\Phi(x) - \Phi(y)\|^2 \leq \|x - y\|^2 - 2\rho\alpha\|x - y\|^2 + \rho^2 L^2 \|x - y\|^2 = (1 - 2\rho\alpha + L^2\rho^2)\|x - y\|^2.$$

Comme $\alpha > 0$, il existe bien un nombre $\rho > 0$ assez petit tel que

$$(1 - 2\rho\alpha + L^2\rho^2) < 1,$$

et donc tel que l'application Φ est contractante et admet donc un unique point fixe.

De plus, les itérées de Picard définies par

$$x_{n+1} = x_n - \rho(\varphi(x_n) - b),$$

pour un choix quelconque de x_0 vont converger vers la solution. ■

Exercice 5 (Interpolation de Hermite)

On se donne deux points distincts $x_0 < x_1$ ainsi que quatre réels y_0, y_1, z_0 et z_1.

1. *Montrer qu'il existe un unique polynôme de degré 3 vérifiant*

$$p(x_0) = y_0, \;\; p(x_1) = y_1,$$

$$p'(x_0) = z_0, \;\; p'(x_1) = z_1.$$

On explicitera l'expression de p en fonction des données.

2. *Montrer l'estimation*

$$\sup_{x\in[x_0,x_1]} |p(x)| \leq 3(|y_0| + |y_1|) + |x_1 - x_0|(|z_0| + |z_1|).$$

3. *Calculer $p''(x_0)$ et $p''(x_1)$ en fonction de $x_0, x_1, y_0, y_1, z_0, z_1$.*
4. *Soit maintenant f une fonction de classe $\mathcal{C}^4$. On suppose que*

$$y_0 = f(x_0), y_1 = f(x_1), z_0 = f'(x_0), \text{ et } z_1 = f'(x_1).$$

Montrer que dans ces conditions, le polynôme de Hermite p vérifie l'estimation d'erreur suivante

$$\forall x \in [x_0, x_1], \exists \xi \in]x_0, x_1[, \text{ tel que } f(x) - p(x) = (x - x_0)^2(x - x_1)^2 \frac{f^{(4)}(\xi)}{4!}.$$

5. *En déduire que*

$$\sup_{[x_0,x_1]} |f - p| \leq |x_1 - x_0|^4 \frac{\|f^{(4)}\|_\infty}{384}.$$

Solution :

Voir la page 38.

■

Exercice 6 (Calcul de la racine carrée d'une matrice SDP)

1. *Soit A une matrice symétrique définie positive. Montrer qu'il existe une unique matrice symétrique définie positive B telle que $B^2 = A$.*
2. *Calculer la differentielle en tout point de l'application $\varphi : X \in M_d(\mathbb{R}) \mapsto X^2 - A$.*
3. *Montrer que si X est symétrique définie positive, alors $D\varphi(X)$ est bijective.*
4. *Montrer que si X est SDP et M est symétrique, alors $(D\varphi(X))^{-1}M$ est également symétrique.*
5. *Ecrire la méthode de Newton appliquée à la résolution de l'équation $B^2 = A$ à partir d'une donnée initiale X_0 symétrique définie positive. Que peut-il se produire au cours des itérations ?*
6. *Montrer que si on prend $X_0 = \mathrm{Id}$, alors la méthode proposée converge vers l'unique solution du problème.*
7. **Question bonus :** *Montrer que si C et D sont deux matrices, l'application*

$$X \in M_d(\mathbb{R}) \mapsto CX + XD \in M_d(\mathbb{R}),$$

est bijective si et seulement si $0 \notin \mathrm{sp}(C) + \mathrm{sp}(D)$.

Solution :

1. Comme A est symétrique réelle, elle est diagonalisable en base orthonormée et à valeurs propres $(\lambda_i)_i$ réelles. De plus, ces valeurs propres sont strictement positives car A est SDP.
   Il existe donc une matrice orthogonale P telle que

$$A = {}^tPDP,$$

   avec $D = \mathrm{diag}(\lambda_1, ..., \lambda_d)$. Cela signifie que la i-ième colonne de P, notée p_i vérifie

$$Ap_i = \lambda_i p_i.$$

   Si une matrice SDP B telle que $B^2 = A$ existe, alors elle commute avec A car

$$AB = B^2B = B^3 = BB^2 = BA.$$

   Donc, pour tout i, on a

$$ABp_i = BAp_i = \lambda_i Bp_i,$$

   et donc Bp_i (qui est non nul car B est SDP donc inversible) est également un vecteur propre de A pour la valeur propre λ_i.
   – Si λ_i est une valeur propre simple, cela impliquerait que Bp_i est proportionnel à p_i et donc que p_i est vecteur propre de B.
   – Si λ_i est valeur propre multiple (de multiplicité k), le raisonnement précédent doit être précisé. Si on rassemble les colonnes de P correspondant à la valeur propre en question dans une matrice rectangle $\tilde{P} \in M_{d,k}(\mathbb{R})$, on a obtenu l'existence d'une matrice carrée $C \in M_k(\mathbb{R})$ telle que

$$B\tilde{P} = \tilde{P}C.$$

     Comme les colonnes de $\tilde{P}$ sont orthonormales, on obtient

$$C = {}^t\tilde{P}B\tilde{P},$$

     et donc C est SDP. Par ailleurs, on a

$$\lambda_i\tilde{P} = A\tilde{P} = B^2\tilde{P} = B(\tilde{P}C) = \tilde{P}C^2,$$

     et donc

$$C^2 = \lambda_i I. \tag{2}$$

     Comme C est SDP, elle est diagonalisable à valeurs propres strictement positives. D'après (2), la seule possibilité c'est que C a une unique valeur propre égale à $\sqrt{\lambda_i}$ et donc nécessairement $C = \sqrt{\lambda_i}I$.

Dans tous les cas, on a donc montré que ${}^tPBP$ est nécessairement la matrice diagonale dont les coefficients diagonaux sont les racines carrées des valeurs propres de A.

2. On développe le carré (en prenant garde à la non-commutativité des matrices) pour écrire

$$\varphi(X+H) = (X+H)^2 - A = \underbrace{X^2 - A}_{=\varphi(X)} + XH + HX + H^2,$$

ce qui montre que φ est différentiable en X et que

$$D\varphi(X).H = XH + HX$$

On prendra garde au fait que, pour tout $X \in M_d(\mathbb{R})$, on a

$$D\varphi(X) \in L(M_d(\mathbb{R}), M_d(\mathbb{R})).$$

3. On suppose que X est SDP. Soit H dans le noyau de $D\varphi(X)$, on a donc

$$XH + HX = 0.$$

Soit v un vecteur propre de X pour la valeur propre $\lambda > 0$. L'équation précédente montre que

$$XHv = -HXv = -\lambda Hv.$$

Si Hv est non nul, c'est un vecteur propre de X pour la v.p. $-\lambda < 0$ ce qui n'est pas possible car X n'a que des valeurs propres positives. On a donc nécessairement $Hv = 0$.
Comme ceci est vrai pour tout vecteur propre de X, et qu'il existe une base de $\mathbb{R}^d$ formée de tels vecteurs propres, on déduit que $H = 0$ et donc que $D\varphi(X)$ est injective et donc bijective.

4. On pose $Y = (D\varphi(X))^{-1}M$. Cette matrice vérifie donc

$$XY + YX = M.$$

Si on transpose l'équation et qu'on utilise la symétrie de M et X, on trouve

$$M = X({}^tY) + ({}^tY)X.$$

Ainsi, ${}^tY$ et Y vérifie la même équation matricielle dont on a vu plus haut qu'elle admettait une unique solution. On a donc ${}^tY = Y$ ce qui est bien le résultat attendu.

5. Usuellement, la méthode s'écrit
$$X_{n+1} = X_n - (D\varphi(X_n))^{-1}(X_n^2 - A).$$
Il pourrait arriver que, pour un certain n, la matrice X_n ainsi obtenue ne soit plus symétrique définie positive (on est certain qu'elle sera symétrique mais rien ne dit qu'elle sera définie positive). Dans ce cas, on est pas assuré que l'itération suivante soit bien définie, car $D\varphi(X_n)$ n'est plus nécessairement inversible.

6. Soit P orthogonale qui diagonalise A : ${}^tPAP = D$. Posons $Y_n = {}^tPX_nP$. On peut alors vérifier par récurrence que Y_n va être diagonale pour tout n (on utilise ici le fait que A et I sont simultanément diagonalisables dans la base donnée par P).
De plus, à i fixé, le i-ième coefficient diagonal de Y_n, noté $y_{n,i}$ vérifie la relation de récurrence

$$y_{n+1,i} = y_{n,i} - \frac{1}{2y_{n,i}}((y_{n,i}^2 - \lambda_i) = \frac{1}{2}\left(y_{n,i} + \frac{\lambda_i}{y_{n,i}}\right).$$

On reconnaît la méthode de Newton (scalaire) appliquée à la fonction $y \mapsto y^2 - \lambda_i$. Celle-ci converge dès lors que la donnée initiale est positive, ce qui est le cas ici car $y_{0,i} = 1$.

7. Appelons φ l'application concernée, $\varphi : X \mapsto CX + XD$.
   – Supposons d'abord que $0 \in \mathrm{Sp}(C) + \mathrm{Sp}(D)$. Cela signifie qu'il existe une valeur propre λ de C telle que $-\lambda$ soit valeur propre de ${}^tD$. On note v et w des vecteurs propres respectifs

$$Cv = \lambda v, \quad {}^tDw = -\lambda w.$$

La matrice $X = v \otimes w = v{}^tw$ vérifie

$$CX + XD = Cv{}^tw + v{}^t({}^tDw) = \lambda v{}^tw - \lambda v{}^tw = 0,$$

et donc φ n'est pas bijective car elle a un noyau non-trivial.

– Supposons maintenant que $0 \notin \mathrm{Sp}(C) + \mathrm{Sp}(D)$. Soit X dans le noyau de φ. On a donc

$$CX + XD = 0.$$

On voit immédiatement que pour tout polynôme p, on a

$$p(C)X = Xp(-D).$$

Si on prend pour p le polynôme caractéristique de C

$$p(x) = \prod_i (x - \lambda_i)^{m_i},$$

avec $\lambda_i \in \mathcal{C}$ et $m_i \geq 1$, on a $p(C) = 0$ d'après le théorème de Cayley-Hamilton, ce qui donne

$$Xp(-D) = 0.$$

Par ailleurs, $p(-D)$ est inversible car chacun des facteurs $(-D - \lambda_i)$ l'est d'après l'hypothèse (sinon $-\lambda_i$ serait valeur propre de D et 0 serait dans la somme des deux spectres).
Il reste donc $X = 0$, et le résultat est prouvé.

■

Exercice 7

Soit $f : \mathbb{R} \to \mathbb{R}$ de classe $\mathcal{C}^3$ qui admet une racine x^* avec $f'(x^*) \neq 0$. On se donne par ailleurs une fonction $\varphi : \mathbb{R} \to \mathbb{R}$ de classe $\mathcal{C}^1$ vérifiant $\varphi(x^*) = x^*$.
Pour une donnée initiale x_0 fixée, on considère la suite définie par récurrence par la formule

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

1. *Montrer qu'il existe $\delta > 0$ tel que, si x_0 appartient à l'intervalle $I_\delta = [x^* - \delta, x^* + \delta]$, alors la suite $(x_n)_n$ est bien définie et reste dans I_δ.*
2. *Montrer que, dans ce cas, la suite $(x_n)_n$ converge vers x^* et que la convergence est quadratique.*
3. *Montrer que si $\varphi'(x^*) = 1/2$ et si φ est de classe $\mathcal{C}^2$, alors la convergence de $(x_n)_n$ vers x^* est cubique.*
4. *Montrer que la fonction $\varphi(x) = x - \frac{f(x)}{2f'(x)}$ peut convenir.*

Solution :

1. On note $g(x) = x - \frac{f(x)}{f'(\varphi(x))}$ la fonction que l'on itère dans cette méthode. On voit immédiatement que x^* est un point-fixe de g.
   Si on montre que $|g'(x^*)| < 1$ alors, l'existence d'un intervalle I_δ stable par g sera garantie par le théorème des accroissements finis et la continuité de g' (on pourra en effet trouver un $\delta > 0$ tel que $\sup_{I_\delta} |g'| < 1$).
   Un simple calcul donne
   $$g'(x) = 1 - \frac{f'(x)}{f'(\varphi(x))} + \frac{f(x)f''(\varphi(x))\varphi'(x)}{(f'(\varphi(x)))^2},$$
   et donc
   $$g'(x^*) = 0.$$
2. D'après le choix de δ dans la question précédente, la fonction g est contractante sur I_δ, donc le théorème du point fixe de Banach nous dit que g admet un unique point fixe dans I_δ (qui est bien sûr x^*) et que la suite $(x_n)_n$ converge vers x^* pour toute donnée initiale $x_0 \in I_\delta$.
   Par ailleurs, comme $g'(x^*) = 0$, on peut utiliser la formule de Taylor suivante
   $$\forall x \in I_\delta, \exists \xi \in I_\delta, \text{ tel que } g(x) = \underbrace{g(x^*)}_{=x^*} + \underbrace{g'(x^*)(x - x^*)}_{=0} + \frac{1}{2}g''(\xi)(x - x^*)^2,$$
   D'où l'on tire
   $$|x_{n+1} - x^*| \leq M|x_n - x^*|^2,$$
   où $M = \frac{1}{2}\sup_{I_\delta} |g''|$. La convergence est donc quadratique.
3. L'idée est d'utiliser la formule de Taylor à l'ordre suivant. Pour obtenir l'ordre cubique, il faudra que $g''(x^*)$ soit également nul. C'est là où va servir l'hypothèse sur $\varphi'(x^*)$. On calcule donc $g''(x^*)$ (en n'oubliant pas que tous les termes contenant $f(x^*)$ seront nuls et que $\varphi(x^*) = x^*$). Il vient
   $$g''(x^*) = -\frac{f''(x^*)}{f'(x^*)} + \frac{f''(x^*)\varphi'(x^*)}{f'(x^*)} + \frac{f''(x^*)\varphi'(x^*)}{f'(x^*)} = \frac{f''(x^*)}{f'(x^*)}(-1 + 2\varphi'(x^*)) = 0.$$
   Ainsi, la formule de Taylor à l'ordre 3 donne
   $$|x_{n+1} - x^*| \leq M|x_n - x^*|^3,$$
   avec $M = \frac{1}{6}\sup_{I_\delta} |g'''|$.
4. Le problème de la construction précédente, c'est de trouver une fonction φ convenable sans bien sûr supposer connu le point x^* que l'on cherche (sinon la fonction $\varphi(x) = x^* + (x - x^*)/2$ conviendrait ...).
   On va montrer que la fonction $\varphi(x) = x - \frac{f(x)}{2f'(x)}$ proposée dans l'énoncé convient. Il est clair que $\varphi(x^*) = x^*$ et de plus $\varphi'(x^*) = 1/2$. Cette fonction ne nécessite pas la connaisance *a priori* du point recherché ce qui était l'objectif recherché.
   In fine, la méthode itérative cubique que l'on vient de construire s'écrit
   $$x_{n+1} = x_n - \frac{f(x_n)}{f'\left(x_n - \frac{f(x_n)}{2f'(x_n)}\right)}.$$

■

Exercice 8 (Méthodes itératives pour les matrices à diagonale strictement dominante)

On dit qu'une matrice carrée $A \in M_d(\mathbb{C})$ est à diagonale strictement dominante si on a

$$\text{Pour tout } 1 \leq i \leq d, \quad |a_{ii}| > \sum_{j \neq i} |a_{ij}|.$$

1. *Montrer que toute matrice à diagonale strictement dominante est inversible.*
2. **Théorème de Gershgoring :** *Montrer que pour toute matrice $M \in M_d(\mathbb{C})$, on a*

$$\mathrm{sp}(M) \subset \bigcup_{i=1}^{d} \overline{D}\left(m_{ii}, \sum_{j \neq i} |m_{ij}|\right),$$

les disques apparaissant dans le membre de droite sont appelés, les disques de Gershgoring de la matrice M.

3. *Soit A une matrice à diagonale strictement dominante et $b \in \mathbb{C}^d$. On écrit*

$$A = L + D + U,$$

où L est triangulaire inférieure stricte, U est triangulaire supérieure stricte et D est diagonale.

   (a) *Montrer que, pour tout $x_0 \in \mathbb{C}^d$, la méthode itérative définie par*

$$x_{n+1} = D^{-1}(-(L+U)x_n + b),$$

   converge vers l'unique solution de $Ax = b$. Cette méthode s'appelle la méthode de Jacobi.

   (b) *Montrer que, pour tout $x_0 \in \mathbb{C}^d$, la méthode itérative définie par*

$$x_{n+1} = (D+L)^{-1}(-Ux_n + b),$$

   converge vers l'unique solution de $Ax = b$. Cette méthode s'appelle la méthode de Gauss-Seidel.

Solution :

Les deux premières questions de cet exercice sont très classiques et méritent d'être connues.

1. Soit A à diagonale strictement dominante. Supposons que A ne soit pas inversible. Il existe donc un $x \in \mathbb{C}^d$ non trivial tel que $Ax = 0$. Choisissons un indice i tel que $|x_i| = \|x\|_\infty$ et écrivons la i-ième ligne de l'équation $Ax = 0$

$$\sum_{j=1}^{d} a_{ij} x_j = 0,$$

   soit encore

$$a_{ii} x_i = -\sum_{j \neq i} a_{ij} x_j.$$

   En prenant le module, il vient

$$|a_{ii}||x_i| \leq \left(\sum_{j \neq i} |a_{ij}|\right) \|x\|_\infty.$$

   Par choix de i, on a $|x_i| = \|x\|_\infty > 0$ et donc on trouve

$$|a_{ii}| \leq \sum_{j \neq i} |a_{ij}|,$$

   ce qui contredit l'hypothèse.

2. Soit $\lambda \in \mathrm{sp}(M)$. Cela signifie que $M - \lambda I$ n'est pas inversible. En particulier, d'après le point précédent, cette matrice n'est pas à diagonale strictement dominante. Cela signifie qu'il existe i tel que

$$|m_{ii} - \lambda| \leq \sum_{j \neq i} |m_{ij}|,$$

   c'est-à-dire que λ appartient au disque centré en m_{ii} et de rayon $\sum_{j \neq i} |m_{ij}|$. C'est bien le résultat attendu.

3. (a) On a vu en cours que la convergence de cette méthode est assurée dès lors que le rayon spectral de la matrice de l'itération $M = D^{-1}(L+U)$ est strictement inférieur à 1. Pour montrer celà, on va utiliser le théorème de Gershgoring. Il nous suffit de montrer que tous les disques de Gershgoring de cette matrice sont contenus dans le disque ouvert unité $D(0,1)$.

   On constate que, par construction, tous les coefficients diagonaux de M sont nuls, les disques de Gershgoring sont donc centrés en 0, il suffit de montrer que leurs rayons sont tous strictement plus petits que 1.

   On calcule pour cela

   $$\forall i \neq j, \;\; m_{ij} = \frac{1}{a_{ii}} a_{ij},$$

   de sorte que

   $$\sum_{j \neq i} |m_{ij}| = \frac{1}{|a_{ii}|} \left(\sum_{j \neq i} |a_{ij}| \right) < 1,$$

   d'après la condition de stricte domination de la diagonale de A. Le résultat est prouvé.

   (b) Même méthode ici pour la matrice

   $$M = (D+L)^{-1} U.$$

   Soit λ valeur propre de M. Cela signifie que $M - \lambda I$ n'est pas inversible, ce qui s'écrit encore

   $$U - \lambda(D+L), \;\; \text{n'est pas inversible.}$$

   Montrons que cela ne peut pas se produire si $|\lambda| \geq 1$. Supposons donc $|\lambda| \geq 1$ et montrons que la matrice $N = U - \lambda(D+L)$ est à diagonale strictement dominante.

   Soit i quelconque. On a

   $$n_{ii} = -\lambda a_{ii},$$
   $$n_{ij} = a_{ij}, \;\; \text{si } j > i,$$
   $$n_{ij} = -\lambda a_{ij}, \;\; \text{si } j < i.$$

   Ainsi, comme $|\lambda| \geq 1$, on a

   $$\sum_{j \neq i} |n_{ij}| = \sum_{j > i} |a_{ij}| + |\lambda| \sum_{j < i} |a_{ij}| \leq |\lambda| \left(\sum_{j \neq i} |a_{ij}| \right).$$

   Comme A est à diagonale strictement dominante, on en déduit que

   $$\sum_{j \neq i} |n_{ij}| < |\lambda| |a_{ii}| = |n_{ii}|.$$

   Ainsi N est elle-même à diagonale strictement dominante et donc inversible. Cela prouve que λ ne peut pas être valeur propre de M et le résultat attendu est prouvé.

   ■

Exercice 9 (Calcul de valeurs propres par la méthode de Newton)

Soit A une matrice symétrique. Proposez une méthode de type Newton pour calculer une valeur propre de A et un vecteur propre associé.
Montrer que cette méthode converge dès lors qu'on prend une donnée initiale proche d'une solution associée à une valeur propre simple.

Solution :

Voir la Section 3.3.3 pour un très bref aperçu de la méthode. ■

Exercice 10 (Méthodes d'accélération de convergence)

Soit $g : [a,b] \to [a,b]$ une application contractante de classe $\mathcal{C}^3$. On note l son unique point fixe dans $[a,b]$.

1. *Montrer que $|g'(l)| < 1$. On pose $k = g'(l)$ et on suppose dorénavant que $k \neq 0$.*
2. *Soit $x_0 \in [a,b]$ et $(x_n)_n$ la suite définie par*

$$x_{n+1} = g(x_n).$$

On suppose dans la suite que $x_n \neq l$ pour tout n (sans quoi la méthode converge en un nombre fini d'itérations).
Montrer qu'il existe $C \in \mathbb{R}^*$ telle que

$$x_n - l \underset{+\infty}{\sim} Ck^n.$$

3. *On suppose que $g''(l) \neq 0$. Montrer qu'il existe $D \in \mathbb{R}^*$ telle que*

$$x_n = l + Ck^n + Dk^{2n} + o(k^{2n}).$$

4. *Pour tout $n \geq 0$, on pose*

$$y_n = \frac{x_{n+1} - kx_n}{1-k}.$$

Montrer qu'il existe $D' \in \mathbb{R}$ tel que

$$y_n - l \underset{+\infty}{\sim} D'k^{2n}.$$

Qu'en concluez-vous ? Quel est l'inconvénient de cette méthode ?

5. *Pour tout $n \geq 0$ (assez grand), on pose*

$$k_n = \frac{x_{n+1} - x_n}{x_n - x_{n-1}},$$

et

$$z_n = \frac{x_{n+1} - k_n x_n}{1-k_n}.$$

Montrer que, pour une certaine valeur de D'', on a

$$z_n - l \underset{+\infty}{\sim} D''k^{2n}.$$

Commentaires ?

Solution :

1. Si $L < 1$ est la constante de Lipschitz de g sur $[a,b]$, on a

$$\left| \frac{g(x) - g(l)}{x - l} \right| \leq L, \ \ \forall x \in [a,b] \setminus \{l\}.$$

En passant à la limite, on trouve $|g'(l)| \leq L$ et donc en particulier $|g'(l)| < 1$.

2. Le théorème du cours montre que

$$|x_n - l| \leq |x_0 - l| L^n, \ \ \forall n \geq 0. \tag{3}$$

On soustrait l'équation $l = g(l)$ à la relation de réccurence qui définit la suite et on pose $e_n = x_n - l$. Cela s'écrit

$$e_{n+1} = g(x_n) - g(l),$$

et donc avec la formule de Taylor

$$e_{n+1} = ke_n + \frac{1}{2} e_n^2 g''(\zeta_n),$$

où $\zeta_n \in]a,b[$. Il s'en suit que

$$\frac{e_{n+1}}{k^{n+1}} = \frac{e_n}{k^n} + \frac{1}{2k} \left(\frac{e_n}{k^n} \right) e_n g''(\zeta_n) = \frac{e_n}{k^n} \left(1 + \frac{1}{2k} e_n g''(\zeta_n) \right).$$

Comme $g''(\zeta_n)$ est bornée (par la norme infinie de g'') et que e_n tend vers 0, la seconde parenthèse est toujours positive pour n assez grand, ce qui prouve dans un premier temps que e_n/k^n garde un signe constant à partir d'un

certain rang et par hypothèse cette quantité est non nulle. Sans perte de généralité, on suppose donc que $e_n/k^n > 0$ pour tout n.
On peut donc désormais poser $z_n = \log(e_n/k^n)$ et il s'agit désormais de montrer que cette suite converge. Pour cela on prend le logarithme de l'équation précédente qui donne

$$z_{n+1} = z_n + \log\left(1 + \frac{1}{2k} e_n g''(\zeta_n)\right),$$

La convergence de la suite $(z_n)_n$ est donc équivalente à la convergence de la série

$$\sum_{n\geq 0} \log\left(1 + \frac{1}{2k} e_n g''(\zeta_n)\right).$$

On va montrer, plus précisément que cette série est absolument convergente. Pour cela, on observe que $e_n g''(\zeta_n)$ tend vers 0 et donc, on a l'équivalent

$$\left|\log\left(1 + \frac{1}{2k} e_n g''(\zeta_n)\right)\right| \underset{+\infty}{\sim} \frac{1}{2|k|}|e_n||g''(\zeta_n)|,$$

et comme la série de terme général e_n est convergente (voir (3)), on a bien convergence de la série étudiée et le résultat est démontré.

3. On pose maintenant
$$y_n = \frac{e_n}{k^n} - C.$$
D'après la question précédente, cette suite tend vers 0 et la relation de récurrence donne
$$y_{n+1} = y_n + \frac{1}{2} k^{n-1} \left(\frac{e_n}{k^n}\right)^2 g''(\zeta_n).$$
Comme $(y_n)_n$ tend vers 0, on peut sommer la série de terme général $y_{n+1} - y_n$ ce qui onne pour tout $N \geq 0$
$$y_N = -\sum_{n=N}^{+\infty} \frac{1}{2} k^{n-1} \underbrace{\left(\frac{e_n}{k^n}\right)^2 g''(\zeta_n)}_{=z_n}.$$
On observe que ζ_n tend vers l et donc $g''(\zeta_n)$ tend vers $g''(l) \neq 0$ et ainsi
$$z_n \xrightarrow[n\to\infty]{} C^2 g''(l),$$
et donc
$$y_N = -\frac{C^2 g''(l)}{2} \sum_{n=N}^{+\infty} k^{n-1} - \sum_{n=N}^{+\infty} \frac{1}{2} k^{n-1}(z_n - C^2 g''(l)).$$
S'agissant de séries géométriques, le premier terme se calcule explicitement et le second se majore ce qui donne
$$\left|y_N + \frac{C^2 g''(l)}{2(1-k)} k^{N-1}\right| \leq \frac{1}{2(1-|k|)} \left(\sup_{n\geq N} |z_n - C^2 g''(l)|\right) |k|^{N-1}.$$
Divisons le tout par $|k|^N$, ce qui donne
$$\left|\frac{y_N}{k^N} + \frac{C^2 g''(l)}{2k(1-k)}\right| \leq \frac{1}{2|k|(1-|k|)} \left(\sup_{n\geq N} |z_n - C^2 g''(l)|\right).$$
Par définition de la convergence de $(z_n)_n$ le terme de droite tend vers 0 quand $N \to +\infty$. On a donc bien montré que
$$y_N \underset{+\infty}{\sim} Dk^k,$$
avec $D = -\frac{C^2 g''(l)}{2k(1-k)}$, ce qui est bien le résultat attendu.

4. D'après les résultats précédents, on a
$$\begin{aligned} y_n - l = \frac{e_{n+1} - k e_n}{1-k} &= \frac{Ck^{n+1} + Dk^{2n+2} + o(k^{2n}) - k(Ck^n + Dk^{2n} + o(k^{2n}))}{1-k}, \\ &= \frac{-Dk^{2n}}{k} + o(k^{2n}). \end{aligned}$$

La conclusion est que nous avons accéléré la convergence de la méthode. Initialement, la suite $(x_n)_n$ converge vers l de façon géométrique de rapport k alors que la nouvelle suite $(y_n)_n$ converge vers l avec un rapport k^2 qui est bien plus petit que k.

L'inconvénient de cette méthode d'accélération est qu'elle nécessite la connaissance de k, qui est n'est pas disponible *a priori*.

5. Etudions la suite k_n :

$$k_n = \frac{Ck^{n+1} - Ck^n + o(k^n)}{Ck^n - Ck^{n-1} + o(k^n)} = \frac{Ck - C + o(1)}{C - C/k + o(1)} = \frac{Ck - C}{C - C/k} + o(1) = k + o(1).$$

Ainsi la suite $(k_n)_n$ est calculable à partir des itérées précédentes et converge vers k. L'idée de la construction de z_n est donc de remplacer dans la définition de y_n la valeur (inconnue) de k par les valeurs approchées k_n.

On peut en réalité être plus précis et obtenir

$$\begin{aligned} k_n - k &= \frac{x_{n+1} - x_n - kx_n + kx_{n-1}}{x_n - x_{n-1}} \\ &= \frac{Ck^{n+1} - (1+k)Ck^n + Ck^n + Dk^{2n+2} - D(1+k)k^{2n} + Dk^{2n-1} + o(k^{2n})}{Ck^n - Ck^{n-1} + o(k^n)} \\ &\underset{\infty}{\sim} \frac{D}{C}\frac{k^3 + k^2 - k + 1}{k - 1}k^n, \end{aligned}$$

ce qui montre que $k_n - k \sim Ek^n$ pour un certain $E \neq 0$.

On peut alors exprimer z_n en fonction de y_n et obtenir

$$z_n - l = y_n - l + \frac{k_n - k}{1 - k_n}\big((y_n - l) - (x_n - l)\big).$$

et donc $z_n - l$ est bien équivalent à un terme en k^{2n}.

■

Exercice 11

Soit $A \in M_d(\mathbb{R})$ une matrice carrée symétrique définie positive et $B \in M_{k,d}(\mathbb{R})$ une matrice rectangle **surjective**.

1. *Montrer que la matrice*
$$C = \begin{pmatrix} A & {}^tB \\ B & 0 \end{pmatrix},$$
est inversible, symétrique, mais pas définie positive.

2. *On se donne $f \in \mathbb{R}^d$ et $g \in \mathbb{R}^k$ et on cherche à trouver l'unique solution $(u,p) \in \mathbb{R}^d \times \mathbb{R}^k$ du problème*
$$\begin{cases} Au + {}^tBp = f, \\ \qquad Bu = g. \end{cases} \tag{4}$$

   (a) *Pour $\alpha > 0$, on propose la méthode itérative suivante : on se donne $p_0 \in \mathbb{R}^k$, puis on résout pour tout $n \geq 0$*
$$\begin{cases} Au_{n+1} + {}^tBp_n = f, \\ \qquad p_{n+1} = p_n + \alpha(Bu_{n+1} - g). \end{cases}$$
   Montrer que les suites $(u_n)_n$ et $(p_n)_n$ sont bien définies et convergent vers u et p respectivement, dès lors que
$$0 < \alpha < \frac{2}{\rho(M)}, \quad \text{où } M = BA^{-1}{}^tB.$$

   (b) *Que gagne-t'on, en pratique, à essayer d'utiliser cette méthode plutôt que de tenter de résoudre le système initial par une méthode traditionnelle ?*

   (c) *On se donne maintenant un nouveau paramètre $r > 0$. Montrer que la solution (u,p) de* (4) *vérifie aussi*
$$\begin{cases} Au + {}^tB(p + rBu) = f + r{}^tBBg, \\ \qquad\qquad Bu = g. \end{cases} \tag{5}$$
   Ceci nous amène à considérer la méthode itérative suivante
$$(A + r{}^tBB)u_{n+1} + {}^tBp_n = f + r{}^tBBg,$$
$$p_{n+1} = p_n + \alpha(Bu_{n+1} - g).$$

   (d) *Montrer que $B(A + r{}^tBB)^{-1}{}^tB = M(\mathrm{Id} + rM)^{-1}$.*

   (e) *En déduire que la nouvelle méthode itérative proposée ci-dessus converge dès que*
$$\alpha < 2\left(r + \frac{1}{\rho(M)}\right).$$

   (f) *Quel est l'avantage de cette méthode par rapport à la précédente ?*

Solution :

1. Si un vecteur $X = \begin{pmatrix} u \\ p \end{pmatrix}$ avec $u \in \mathbb{R}^d$, $p \in \mathbb{R}^k$ est dans le noyau de C alors on a
$$\begin{cases} Au + {}^tBp = 0, \\ \qquad Bu = 0. \end{cases}$$
On prend le produit scalaire (dans $\mathbb{R}^d$) de la première équation par u et on trouve
$$(Au, u) + ({}^tBp, u) = 0,$$
alors que le produit scalaire (dans $\mathbb{R}^k$) de la seconde équation par p donne
$$0 = (Bu, p) = (u, {}^tBp).$$
En combinant les deux résultats obtenus, on trouve finalement que
$$(Au, u) = 0.$$

Comme A est définie positive, ceci implique que $u = 0$. La première équation du système donne maintenant que ${}^tBp = 0$.

Comme B est surjective, ${}^tB$ est injective et donc $p = 0$. On a donc prouvé que $u = 0$ et $p = 0$ donc $X = 0$. Ceci montre bien que C est bijective.

La symétrie de C ne fait aucun doute. Montrons qu'elle n'est pas définie positive. Pour cela, il suffit de constater que

$$C\begin{pmatrix}0\\p\end{pmatrix} = \begin{pmatrix}{}^tBp\\0\end{pmatrix},$$

et donc que

$$\left(C\begin{pmatrix}0\\p\end{pmatrix}, \begin{pmatrix}0\\p\end{pmatrix}\right) = 0.$$

Ceci étant vrai pour tout choix de $p \in \mathbb{R}^k$, il est bien clair que C n'est pas définie positive.

2. (a) Pour tout $n \geq 0$, si p_n est donné, u_{n+1} se calcule aisément en inversant A (qui est bien inversible) par

$$u_{n+1} = A^{-1}(f - {}^tBp_n). \tag{6}$$

Ensuite, p_{n+1} est directement défini par la formule donnée dans l'énoncé. L'existence des suites $(u_n)_n$ et $(p_n)_n$ est donc claire. De plus, on peut éliminer u_n de l'histoire et obtenir la relation de récurrence sur $(p_n)_n$

$$p_{n+1} = p_n + \alpha(BA^{-1}(f - {}^tBp_n) - g) = (I - \alpha BA^{-1}{}^tB)p_n + \alpha(BA^{-1}f - g).$$

La matrice d'itération est donné par $I - \alpha M$, où $M = BA^{-1}{}^tB$ est définie dans l'énoncé.

Le théorème du cours montre que la convergence de la suite $(p_n)_n$ (et donc de la suite $(u_n)_n$ d'après (6)) est assurée dès que le rayon spectral de cette matrice est strictement inférieur à 1.

Or, on voit immédiatement que M est SDP (on note $\lambda_i > 0$ ses valeurs propres) et donc que les valeurs propres de $I - \alpha M$ sont les $1 - \alpha\lambda_i$. Ce sont des nombres réels strictement inférieurs à 1, la condition sur le rayon spectral devient donc

$$1 - \alpha\lambda_i > -1, \ \ \forall i,$$

ce qui est équivalent à la condition de l'énoncé $\alpha < 2/\rho(M)$.

(b) Résoudre un système linéaire pour la matrice SDP A est "relativement" facile (on a des méthodes assez performantes pour le faire comme le gradient conjugué par exemple) alors que résoudre un système pour la matrice C qui n'a pas une structure SDP n'est pas chose facile par les méthodes traditionnelles.

La méthode proposée consiste donc à résoudre, à chaque itération, un système associée à la matrice A (plus petite que C), ce qui est relativement plus simple. Bien entendu, il faut que le nombre d'itérations nécessaire soit modéré pour y gagner quelquechose.

(c) L'équivalence du nouveau système avec le précédent est claire. Comme par ailleurs, la matrice $A + r{}^tBB$ est également SDP, tout ce qui a été fait précédemment s'applique.

(d) Calculons

$$\begin{aligned} B(A + r{}^tBB)^{-1}{}^tB(I + rM) &= B(A + r{}^tBB)^{-1}({}^tB + r{}^tBBA^{-1}{}^tB) \\ &= B(A + r{}^tBB)^{-1}(A + r{}^tBB)A^{-1}{}^tB \\ &= BA^{-1}{}^tB = M, \end{aligned}$$

et le résultat est prouvé.

(e) En reprenant la démarche de la question a) on obtient que la méthode va converger dès que

$$\alpha < \frac{2}{\rho(B(A + r{}^tBB)^{-1}{}^tB)},$$

or d'après la question précédente, les valeurs propres de $B(A + r{}^tBB)^{-1}{}^tB$ sont les

$$\lambda_i(1 + r\lambda_i)^{-1}.$$

Le rayon spectral recherché est donc $\rho(M)/(1 + r\rho(M))$ et la condition sur α s'écrit

$$\alpha < 2(1 + r\rho(M))/\rho(M) = 2\left(r + \frac{1}{\rho(M)}\right).$$

(f) Dans la première approche, la condition sur α dépend du rayon spectral de M que l'on ne peut que difficilement calculer ... Dans le second cas, il suffit de prendre $\alpha = 2r$ par exemple pour assurer la convergence indépendamment du rayon spectral de M.

■

Bibliographie

[AH09] Kendall ATKINSON et Weimin HAN : *Theoretical numerical analysis*, volume 39 de *Texts in Applied Mathematics*. Springer, Dordrecht, third édition, 2009. A functional analysis framework.

[BR06] P. BOYER et J.J. RISLER : *Algèbre pour la licence 3*. Dunod, 2006.

[CLF95] Antoine CHAMBERT-LOIR et Stéphane FERMIGIER : *Exercices de Mathématiques pour l'agrégation. Analyse 2*. Masson, 1995.

[CM84] Michel CROUZEIX et Alain L. MIGNOT : *Analyse numérique des équations différentielles*. Collection Mathématiques Appliquées pour la Maîtrise. Masson, Paris, 1984.

[CM86] Michel CROUZEIX et Alain L. MIGNOT : *Exercices d'analyse numérique des équations différentielles*. Collection Mathématiques Appliquées pour la Maîtrise. Masson, Paris, 1986.

[Ded06] J.P. DEDIEU : *Points fixes, Zéros et la méthode de Newton*. Springer, 2006.

[Dem91] Jean-Pierre DEMAILLY : *Analyse Numérique et équations différentielles*. Collection Grenoble Sciences. Presses Universitaires de Grenoble, 1991.

[Gou94] Xavier GOURDON : *Les maths en tête, mathématiques pour M', Analyse*. Ellipses, 1994. 416 pages.

[GT96] S. GONNORD et N. TOSEL : *Thèmes d'Analyse pour l'Agrégation, Tome 1 : Topologie et Analyse fonctionnelle*. Ellipses, 1996.

[HH06] John HUBBARD et Florence HUBERT : *CALCUL SCIENTIFIQUE : De la théorie à la pratique. Premier volume : Equations algébriques, traitement du signal et géométrie effective.* Vuibert, 2006.

[LT00] P. LASCAUX et R. THÉODOR : *Analyse numérique matricielle appliquée à l'art de l'ingénieur - Tome 1*. Dunod, 2000.

[QSG10] Alfio QUARTERONI, Fausto SALERI et Paola GERVASIO : *Calcul Scientifique*. Springer, 2010. Cours, exercices corrigés et illustrations en MATLAB et Octave. 2nde édition.

[QSS07] Alfio QUARTERONI, Riccardo SACCO et Fausto SALERI : *Méthodes numériques*. Springer-Verlag Italia, Milan, 2007. Algorithmes, analyse et applications.

[Rom96] Jean-Etienne ROMBALDI : *Problèmes corrigés d'analyse numérique*. Collection Enseignement des Mathématiques. Masson, Paris, 1996.

[Rud95] Walter RUDIN : *Analyse fonctionnelle*. Édiscience international, 1995.

[Sch91] Michelle SCHATZMAN : *Analyse numérique*. InterEditions, Paris, 1991. Cours et exercices pour la licence.