Grenoble INP – Pagora **Analyse numérique**
1ère année

Exercices corrigés

NB : Les exercices corrigés ici sont les exercices proposés durant les séances de cours. Les corrections données sont des corrections plus détaillées que celles fournies durant le cours (si le temps a permis de donner ces corrections). Si vous avez des questions concernant ces exercices, n'hésitez pas à envoyer un mail à votre enseignant d'analyse numérique pour lui poser une question. Si vous trouver des coquilles, des erreurs dans le présent document, n'hésitez pas à le signaler à votre enseignant par un mail.

Chapitre 1 : Introduction au calcul approché

Exercice 1 *Montrer que* 9325 *s'écrit bien* $(10010001101101)_2$ *en base* 2 *puis reconvertir* $(10010001101101)_2$ *en base* 10.

Pour convertir un entier de la base 10 à la base 2 (on verra que la méthode diffère légèrement pour un nombre décimal un peu plus tard), on divise l'entier par 2 (division euclidienne) et le reste correspond au dernier chiffre de l'entier en base 2. Pour 9325, cela donne

$$9325 = 2 \times 4662 + \mathbf{1}$$

et on itère le processus sur le quotient obtenu (jusqu'à ce qu'il vaille 1). Ainsi puisque

$$4662 = 2 \times 2331 + \mathbf{0}$$

On peut réécrire 9325 sous la forme

$$9325 = 2 \times (2 \times 2331 + 0) + 1 = 2^2 \times 2331 + 2^1 \times \mathbf{0} + 2^0 \times \mathbf{1}$$

et **01** sont les 2 derniers chiffres de 9325 écrit en base 2. Pour enfoncer le clou, on détaille encore l'itération suivante

$$2331 = 2 \times 1165 + \mathbf{1}$$

$$9325 = 2^2 \times (2 \times 1165 + 1) + 2^1 \times \mathbf{0} + 2^0 \times \mathbf{1} = 2^3 \times 1165 + 2^2 \times \mathbf{1} + 2^1 \times \mathbf{0} + 2^0 \times \mathbf{1}$$

et **101** sont les 3 derniers chiffres de 9325 écrit en base 2. On affiche ensuite le processus itératif dans son entier :

$$\begin{array}{rcccrcc}
9325 & = & 2 & \times & 4662 & + & \mathbf{1} \\
4662 & = & 2 & \times & 2331 & + & \mathbf{0} \\
2331 & = & 2 & \times & 1165 & + & \mathbf{1} \\
1165 & = & 2 & \times & 582 & + & \mathbf{1} \\
582 & = & 2 & \times & 291 & + & \mathbf{0} \\
291 & = & 2 & \times & 145 & + & \mathbf{1} \\
145 & = & 2 & \times & 72 & + & \mathbf{1} \\
72 & = & 2 & \times & 36 & + & \mathbf{0} \\
36 & = & 2 & \times & 18 & + & \mathbf{0} \\
18 & = & 2 & \times & 9 & + & \mathbf{0} \\
9 & = & 2 & \times & 4 & + & \mathbf{1} \\
4 & = & 2 & \times & 2 & + & \mathbf{0} \\
2 & = & 2 & \times & 1 & + & \mathbf{0} \\
1 & = & 2 & \times & 0 & + & \mathbf{1}
\end{array}$$

Donc, on peut décomposer 9325 de la manière suivante

$$\begin{aligned}9325 = 2^{13}\times\mathbf{1}+2^{12}\times\mathbf{0}+2^{11}\times\mathbf{0}+2^{10}\times\mathbf{1}+2^{9}\times\mathbf{0}+2^{8}\times\mathbf{0}+2^{7}\times\mathbf{0}\\ +2^{6}\times\mathbf{1}+2^{5}\times\mathbf{1}+2^{4}\times\mathbf{0}+2^{3}\times\mathbf{1}+2^{2}\times\mathbf{1}+2^{1}\times\mathbf{0}+2^{0}\times\mathbf{1}\end{aligned}$$

On a bien montré que $(9325)_{10} = (10010001101101)_2$. Il ne reste plus qu' ? reconvertir ce nombre binaire en base 10. Pour ce faire, on va procéder de manière itérative. On commence par cette première étape, **1** est le chiffre le plus fort (le plus à gauche) de $(10010001101101)_2$ et on construit le résultat intermédiaire de la manière suivante, on multiplie par 2 le résultat intermédiaire précédent (au départ 0) et on ajoute le chiffre le plus fort restant à traiter. On commence donc par

$$2\times 0+\mathbf{1} = 1 = 2^0\times\mathbf{1}$$

Le résultat intermédiaire est donc 1 et il ne reste plus qu'à traiter 0010001101101 du binaire $(10010001101101)_2$. On itère le processus. On multiplie par 2 le résultat intermédiaire (ici 1) puis on ajoute le chiffre le plus fort restant à traiter (soit ici **0**). D'où

$$2\times 1+\mathbf{0} = 2 = 2^1\times\mathbf{1}+2^0\times\mathbf{0}$$

Il ne reste plus qu'à traiter 010001101101 du binaire $(10010001101101)_2$. Si on détaille l'étape suivante, on a

$$2\times 2+\mathbf{0} = 4 = 2^2\times\mathbf{1}+2^1\times\mathbf{0}+2^0\times\mathbf{0}$$

Il ne reste plus qu'à traiter 10001101101 du binaire $(10010001101101)_2$. On affiche ensuite le processus itératif dans son entier :

$$\begin{array}{cccccccc}2&\times&0&+&\mathbf{1}&=&1\\2&\times&1&+&\mathbf{0}&=&2\\2&\times&2&+&\mathbf{0}&=&4\\2&\times&4&+&\mathbf{1}&=&9\\2&\times&9&+&\mathbf{0}&=&18\\2&\times&18&+&\mathbf{0}&=&36\\2&\times&36&+&\mathbf{0}&=&72\\2&\times&72&+&\mathbf{1}&=&145\\2&\times&145&+&\mathbf{1}&=&291\\2&\times&291&+&\mathbf{0}&=&582\\2&\times&582&+&\mathbf{1}&=&1165\\2&\times&1165&+&\mathbf{1}&=&2331\\2&\times&2331&+&\mathbf{0}&=&4662\\2&\times&4662&+&\mathbf{1}&=&9325\end{array}$$

On vérifie donc bien que $(9325)_{10} = (10010001101101)_2$. Notons bien que de processus décrit ici est juste le premier processus mais pris en sens inverse.

Exercice 2 *Écrire* $(34)_{10}$ *et* $(27)_{10}$ *en binaire puis effectuer l'opération en binaire* $(34)_{10}+(27)_{10}$ *et vérifier que le résultat obtenu soit le bon.*

Convertissons tout d'abord 34 en binaire. Cela donne

$$\begin{array}{rcrcrcl} 34 & = & 2 & \times & 17 & + & \mathbf{0} \\ 17 & = & 2 & \times & 8 & + & \mathbf{1} \\ 8 & = & 2 & \times & 4 & + & \mathbf{0} \\ 4 & = & 2 & \times & 2 & + & \mathbf{0} \\ 2 & = & 2 & \times & 1 & + & \mathbf{0} \\ 1 & = & 2 & \times & 0 & + & \mathbf{1} \end{array}$$

On a donc $(34)_{10} = (100010)_2$. Convertissons maintenant 27 en binaire. On a

$$\begin{array}{rcrcrcl} 27 & = & 2 & \times & 13 & + & \mathbf{1} \\ 13 & = & 2 & \times & 6 & + & \mathbf{1} \\ 6 & = & 2 & \times & 3 & + & \mathbf{0} \\ 2 & = & 2 & \times & 1 & + & \mathbf{1} \\ 1 & = & 2 & \times & 0 & + & \mathbf{1} \end{array}$$

et $(27)_{10} = (11011)_2$. On effectue maintenant l'addition de $(100010)_2$ et $(11011)_2$. Pour rappel, l'addition en binaire fonctionne de la manière suivante

+	0	1
0	0	1
1	1	10

D'où l'opération suivante

$$\begin{array}{cccccccc} & & 1 & 0 & 0 & 0 & 1 & 0 \\ + & & & 1 & 1 & 0 & 1 & 1 \\ \hline = & & 1 & 1 & 1 & 1 & {}^{1}0 & 1 \end{array}$$

On a $(100010)_2 + (11011)_2 = (111101)_2$. Or $(34)_{10} + (27)_{10} = (61)_{10}$, vérifions si $(61)_{10} = (111101)_2$.

$$\begin{array}{rcrcrcr} 2 & \times & 0 & + & \mathbf{1} & = & 1 \\ 2 & \times & 1 & + & \mathbf{1} & = & 3 \\ 2 & \times & 3 & + & \mathbf{1} & = & 7 \\ 2 & \times & 7 & + & \mathbf{1} & = & 15 \\ 2 & \times & 15 & + & \mathbf{0} & = & 30 \\ 2 & \times & 30 & + & \mathbf{1} & = & 61 \end{array}$$

On a bien $(61)_{10} = (111101)_2$, le résultat obtenu en binaire est bien conforme au résultat obtenu en base 10.

Exercice 3 *Écrire* $(90)_{10}$ *et* $(97)_{10}$ *en binaire puis effectuer l'opération en binaire* $(90)_{10} \times (97)_{10}$ *et vérifier que le résultat obtenu est le bon.*

Convertissons tout d'abord 90 en binaire. Cela donne

$$\begin{array}{rcrcrcl} 90 & = & 2 & \times & 45 & + & \mathbf{0} \\ 45 & = & 2 & \times & 22 & + & \mathbf{1} \\ 22 & = & 2 & \times & 11 & + & \mathbf{0} \\ 11 & = & 2 & \times & 5 & + & \mathbf{1} \\ 5 & = & 2 & \times & 2 & + & \mathbf{1} \\ 2 & = & 2 & \times & 1 & + & \mathbf{0} \\ 1 & = & 2 & \times & 0 & + & \mathbf{1} \end{array}$$

On a donc $(90)_{10} = (1011010)_2$. Convertissons maintenant 97 en binaire. On a

$$\begin{array}{rcrcrcl}
97 & = & 2 & \times & 48 & + & \mathbf{1} \\
48 & = & 2 & \times & 24 & + & \mathbf{0} \\
24 & = & 2 & \times & 12 & + & \mathbf{0} \\
12 & = & 2 & \times & 6 & + & \mathbf{0} \\
6 & = & 2 & \times & 3 & + & \mathbf{0} \\
3 & = & 2 & \times & 1 & + & \mathbf{1} \\
1 & = & 2 & \times & 0 & + & \mathbf{1}
\end{array}$$

et $(97)_{10} = (1100001)_2$. On effectue maintenant la multiplication de $(1011010)_2$ par $(1100001)_2$. Pour rappel, la multiplication en binaire fonctionne de la manière suivante

×	0	1
0	0	0
1	0	1

D'où l'opération suivante

$$\begin{array}{ccccccccccccccc}
 & & & & & & & & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\
\times & & & & & & & & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\
\hline
 & & & & & & & & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\
+ & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{0} \\
+ & & & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{0} & \mathbf{0} \\
+ & & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
+ & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
+ & & & 1 & 0 & 1 & 1 & 0 & 1 & 0 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
+ & & 1 & 0 & 1 & 1 & 0 & 1 & 0 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\hline
= & 1 & {}^{1}0 & {}^{1}0 & {}^{1}0 & {}^{1}1 & {}^{1}0 & {}^{1}0 & {}^{1}0 & 0 & 1 & 1 & 0 & 1 & 0
\end{array}$$

On a $(1011010)_2 \times (1100001)_2 = (10001000011010)_2$. Or $(90)_{10} \times (97)_{10} = (8730)_{10}$, vérifions si $(8730)_{10} = (10001000011010)_2$.

$$\begin{array}{ccrcccr}
2 & \times & 0 & + & \mathbf{1} & = & 1 \\
2 & \times & 1 & + & \mathbf{0} & = & 2 \\
2 & \times & 2 & + & \mathbf{0} & = & 4 \\
2 & \times & 4 & + & \mathbf{0} & = & 8 \\
2 & \times & 8 & + & \mathbf{1} & = & 17 \\
2 & \times & 17 & + & \mathbf{0} & = & 34 \\
2 & \times & 34 & + & \mathbf{0} & = & 68 \\
2 & \times & 68 & + & \mathbf{0} & = & 136 \\
2 & \times & 136 & + & \mathbf{0} & = & 272 \\
2 & \times & 272 & + & \mathbf{1} & = & 545 \\
2 & \times & 545 & + & \mathbf{1} & = & 1091 \\
2 & \times & 1090 & + & \mathbf{0} & = & 2182 \\
2 & \times & 2182 & + & \mathbf{1} & = & 4365 \\
2 & \times & 4365 & + & \mathbf{0} & = & 8730
\end{array}$$

On a bien $(8730)_{10} = (10001000011010)_2$, le résultat obtenu en binaire est bien conforme au résultat obtenu en base 10.

Exercice 4 *Si on dispose de 4 bits (bit de signe compris), quelles valeurs peuvent prendre les entiers codés sur ces 4 bits ?*

Si on dispose de 4 bits dont 1 de signe, il ne reste plus que 3 bits pour coder les entiers naturels (ceux plus grand que 0). Ils ne peuvent donc prendre que 2^3 valeurs distinctes dont la valeur 0. Les entiers naturels codés sont ainsi 0, 1, 2, 3, 4, 5, 6, et $7 = 2^3 - 1$. Maintenant, si on tient compte du bit de signe, les entiers codés devraient pouvoir varier entre -7 et 7.

Cependant deux combinaisons auraient la même valeur 0 : **1**000 et **0**000, le chiffre en gras désigne ici le bit de signe. Pour éviter cette redondance, on pose **1**000 = -8 (classiquement, le signe de bit lorsqu'il vaut 1 indique un nombre négatif).

Finalement, si on dispose de 4 bits (bit de signe compris), on peut coder les entiers de valeurs comprises entre $-8 = -2^3$ et $7 = 2^3 - 1$.

Exercice 5 *Vérifier l'égalité entre* $(9,90625)_{10}$ *et* $(1001,11101)_2$.

On distingue la partie entière et la partie décimale à traiter. On vérifier tout d'abord que $(9)_{10} = 1001_2$, en effet

$$\begin{array}{ccccccc} 9 & = & 2 & \times & 4 & + & \mathbf{1} \\ 4 & = & 2 & \times & 2 & + & \mathbf{0} \\ 2 & = & 2 & \times & 1 & + & \mathbf{0} \\ 1 & = & 2 & \times & 0 & + & \mathbf{1} \end{array}$$

On a donc

$$9,90625 = 2^3 \times \mathbf{1} + 2^2 \times \mathbf{0} + 2^1 \times \mathbf{0} + 2^0 \times \mathbf{1} + 0,90625$$

Mais on a aussi

$$9,90625 = 2^3 \times \mathbf{1} + 2^2 \times \mathbf{0} + 2^1 \times \mathbf{0} + 2^0 \times \mathbf{1} + 2^{-1} \times (2 \times 0,90625)$$

$$9,90625 = 2^3 \times \mathbf{1} + 2^2 \times \mathbf{0} + 2^1 \times \mathbf{0} + 2^0 \times \mathbf{1} + 2^{-1} \times 1,8125$$

$$9,90625 = 2^3 \times \mathbf{1} + 2^2 \times \mathbf{0} + 2^1 \times \mathbf{0} + 2^0 \times \mathbf{1} + 2^{-1} \times (\mathbf{1} + 0,8125)$$

On vient donc de calculer le premier chiffre après la virgule de 9,90625 en binaire (soit ici **1**). On réitère le même processus pour avoir le chiffre après la virgule suivant

$$9,90625 = 2^3 \times \mathbf{1} + 2^2 \times \mathbf{0} + 2^1 \times \mathbf{0} + 2^0 \times \mathbf{1} + 2^{-1} \times \mathbf{1} + 2^{-2} \times (2 \times 0,8125)$$

$$9,90625 = 2^3 \times \mathbf{1} + 2^2 \times \mathbf{0} + 2^1 \times \mathbf{0} + 2^0 \times \mathbf{1} + 2^{-1} \times \mathbf{1} + 2^{-2} \times 1,625$$

$$9,90625 = 2^3 \times \mathbf{1} + 2^2 \times \mathbf{0} + 2^1 \times \mathbf{0} + 2^0 \times \mathbf{1} + 2^{-1} \times \mathbf{1} + 2^{-2} \times (\mathbf{1} + 0,625)$$

Le deuxième chiffre après la virgule (en binaire) est donc **1**. Voici enfin directement, les traces des calculs pour obtenir tous les chiffres nécessaires après la virgule

$$9,90625 = 2^3 \times \mathbf{1} + 2^2 \times \mathbf{0} + 2^1 \times \mathbf{0} + 2^0 \times \mathbf{1} + 2^{-1} \times \mathbf{1} + 2^{-2} \times \mathbf{1} + 2^{-3} \times 1,25$$

$$9,90625 = 2^3 \times \mathbf{1} + 2^2 \times \mathbf{0} + 2^1 \times \mathbf{0} + 2^0 \times \mathbf{1} + 2^{-1} \times \mathbf{1} + 2^{-2} \times \mathbf{1} + 2^{-3} \times (\mathbf{1} + 0,25)$$

$$9,90625 = 2^3 \times \mathbf{1} + 2^2 \times \mathbf{0} + 2^1 \times \mathbf{0} + 2^0 \times \mathbf{1} + 2^{-1} \times \mathbf{1} + 2^{-2} \times \mathbf{1} + 2^{-3} \times \mathbf{1} + 2^{-4} \times 0,5$$

$$9,90625 = 2^3 \times \mathbf{1} + 2^2 \times \mathbf{0} + 2^1 \times \mathbf{0} + 2^0 \times \mathbf{1} + 2^{-1} \times \mathbf{1} + 2^{-2} \times \mathbf{1} + 2^{-3} \times \mathbf{1} + 2^{-4} \times \mathbf{0} + 2^{-5} \times (2 \times 0,5)$$

$$9,90625 = 2^3 \times \mathbf{1} + 2^2 \times \mathbf{0} + 2^1 \times \mathbf{0} + 2^0 \times \mathbf{1} + 2^{-1} \times \mathbf{1} + 2^{-2} \times \mathbf{1} + 2^{-3} \times \mathbf{1} + 2^{-4} \times \mathbf{0} + 2^{-5} \times \mathbf{1}$$

On vérifie donc bien que $(9,90625)_{10} = (1001,11101)_2$.

Chapitre 2 : Résolution d'équations non-linéaires

Exercice 6 *On définit la méthode du point fixe suivante*

$$\left\{\begin{array}{ll} x_0 & \text{fixé dans } [a,b] \\ x_{n+1} & = g(x_n) \end{array}\right.$$

On suppose que cette suite admet une limite sur $[a,b]$ notée l. Cette méthode est d'ordre p si $\dfrac{|x_{n+1}-l|}{|x_n-l|^p}$ admet une limite réelle strictement positive lorsque n tend vers l'infini.

On suppose g p fois dérivable sur $[a,b]$. En utilisant la formule de Taylor, montrer que la méthode est d'ordre p si et seulement si

$$g'(l) = g''(l) = \ldots = g^{(p-1)}(l) = 0 \qquad \text{et} \qquad g^{(p)}(l) \neq 0$$

On rappelle tout d'abord la formule de Taylor.

Soit $k \geq 1$ un entier et soit f une fonction de $\mathbb{R}$ dans $\mathbb{R}$ k fois dérivable en $a \in \mathbb{R}$, alors il existe une fonction ϵ_k de $\mathbb{R}$ dans $\mathbb{R}$ tel que

$$f(x) = f(a) + f'(a)(x-a) + f''(a)\frac{(x-a)^2}{2} + \ldots + f^{(k)}(a)\frac{(x-a)^k}{k!} + \epsilon_k(x)(x-a)^k$$

et

$$\lim_{x \longrightarrow a} \epsilon_k(x) = 0$$

Maintenant, si g est p fois dérivable sur $[a,b]$, on peut écrire que

$$g(x_n) = g(l) + g'(l)(x_n-l) + g''(l)\frac{(x_n-l)^2}{2} + \ldots + g^{(p)}(l)\frac{(x_n-l)^p}{p!} + \epsilon_p(x)(x_n-l)^p$$

avec

$$\lim_{x_n \longrightarrow l} \epsilon_p(x_n) = 0$$

Posons $e_n = x_n - l$, on a $x_{n+1} = g(x_n)$ et

$$\begin{aligned} e_{n+1} = g(x_n) - g(l) = g'(l)(x_n-l) + g''(l)\frac{(x_n-l)^2}{2} + \ldots + g^{(p)}(l)\frac{(x_n-l)^p}{p!} + \epsilon_p(x_n)(x_n-l)^p \\ = g'(l)e_n + g''(l)\frac{e_n^2}{2} + \ldots + g^{(p)}(l)\frac{e_n^p}{p!} + \epsilon_p(x)e_n^p \end{aligned}$$

– Si la méthode est d'ordre p alors le quotient $\dfrac{e_{n+1}}{e_n^p}$ tend vers un réel non-nul (car $\dfrac{|x_{n+1}-l|}{|x_n-l|^p}$ admet une limite réelle strictement positive et on enlève juste les valeurs absolues). Or d'après ce qui précède, on a l'égalité suivante

$$\frac{e_{n+1}}{e_n^p} = \frac{g'(l)}{e_n^{p-1}} + \frac{g''(l)}{2e_n^{p-2}} + \ldots + \frac{1}{e_n}\frac{g^{(p-1)}(l)}{(p-1)!} + \frac{g^{(p)}(l)}{p!} + \epsilon_p(x_n)$$

A gauche de l'égalité, le terme tend vers un réel non-nul. Concernant les termes à droite, on a que e_n tend vers 0 quand n tend vers l'infini donc

$$\lim_{n \longrightarrow +\infty} \frac{g'(l)}{e_n^{p-1}} = \left\{\begin{array}{ll} 0 & \text{si } g'(l) = 0 \\ \pm\infty & \text{sinon} \end{array}\right.$$

De même

$$\lim_{n\longrightarrow+\infty}\frac{g''(l)}{2e_n^{p-2}}=\begin{cases}0 & \text{si } g''(l)=0\\ \pm\infty & \text{sinon}\end{cases}$$

jusqu'à

$$\lim_{n\longrightarrow+\infty}\frac{1}{e_n}\frac{g^{(p-1)}(l)}{(p-1)!}=\begin{cases}0 & \text{si } g^{(p-1)}(l)=0\\ \pm\infty & \text{sinon}\end{cases}$$

Par contre $\dfrac{g^{(p)}(l)}{p!}$ est un réel et $\epsilon_p(x_n)$ tend vers 0. Par unicité de la limite des deux côtés de l'égalité, il faut que $g'(l)=g''(l)=\ldots=g^{(p-1)}(l)=0$ pour que le côté droit tende vers un réel, $\dfrac{g^{(p)}(l)}{p!}$ en l'occurence. La limite réelle est en plus différente de 0, donc on doit avoir $g^{(p)}(l)\neq 0$.

– Si on a

$$g'(l)=g''(l)=\ldots=g^{(p-1)}(l)=0 \qquad \text{et} \qquad g^{(p)}(l)\neq 0$$

alors

$$e_{n+1}=\frac{e_n^p}{p!}g^{(p)}(l)+e_n^p\epsilon_p(x_n)$$

et

$$\frac{e_{n+1}}{e_n^p}=\frac{1}{p!}g^{(p)}(l)+\epsilon_p(x_n)$$

d'où comme $\epsilon_p(x_n$ tend vers 0 quand n tend vers l'infini

$$\lim_{n\longrightarrow+\infty}\left|\frac{e_{n+1}}{e_n^p}\right|=\frac{1}{p!}|g^{(p)}(l)|\in\mathbb{R}_+^*$$

et la méthode est bien d'ordre p.

Exercice 7 (ordre de convergence de la méthode de Newton) *On rappelle ici la méthode de Newton, il s'agit d'une méthode du point fixe pour rśoudre $f(x)=0$ sur $[a,b]$ définissant la suite suivante*

$$\begin{cases}x_0 & \text{fixé dans } [a,b]\\ x_{n+1} & =g(x_n)=x_n-\dfrac{f(x_n)}{f'(x_n)}\end{cases}$$

On suppose que cette suite admet une limite sur $[a,b]$ notée l. Montrer que si f est 3 fois dérivable sur $[a,b]$ et que $f'(l)\neq 0$, alors la méthode de Newton est d'ordre 2 au moins.

La fonction g vaut ici

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

Sa dérivée vaut

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

et en l, cette dérivée vaut 0 car $f(l)=0$, l est la solution au problème étudié. Donc la méthode est d'ordre 2 au moins (d'ordre 2 si $g''(l)\neq 0$, d'ordre supérieur à 2 sinon). Si on va plus loin, calculons la dérivée seconde de g

$$g''(x)=\frac{(f'(x)f''(x)+f(x)f'''(x))\,(f'(x))^2-2f(x)f'(x)(f''(x))^2}{(f'(x))^4}$$

Elle vaut en l

$$g''(l) = \frac{f''(l)}{f'(l)}$$

Donc si $f''(l) \neq 0$ alors la méthode de Newton est d'ordre 2, sinon elle est d'ordre supérieur à 2.

Chapitre 3 : Approximation de fonctions

Méthode des moindres carrés

Exercice 8 *On dispose de n points (x_i, y_i), $i = 1, \ldots, n$ et on suppose que la fonction modèle est de la forme*

$$f(x, \beta_0) = \beta_0$$

Trouver β_0 minimisant la quantité

$$S(\beta_0) = \sum_{i=1}^{n} (y_i - f(x_i, \beta_0))^2$$

On cherche le minimum β_0^* tel que $S(\beta_0^*) = \min_{\beta_0 \in \mathbb{R}} S(\beta_0)$. Pour cela, on a besoin de calculer sa dérivée $S'(\beta_0)$. Celle-ci vaut

$$S'(\beta_0) = \sum_{i=1}^{n} [-2(y_i - f(x_i, \beta_0))] = -2 \sum_{i=1}^{n} (y_i - \beta_0)$$

Pour trouver un extremum local (minimum ou maximum local), il faut résoudre le problème $S'(\beta_0) = 0$. Dans notre cas, on en déduit que

$$\beta_0^* = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Il reste cependant à vérifier que β_0^* est bien un minimum et que celui-ci est global. Pour cela, étudions le comportement de $S(\beta_0)$. On a le tableau des variations suivant

β_0	$-\infty$		β_0^*		β_0
$S'(\beta_0)$		$-$	0	$+$	
$S(\beta_0)$	$+\infty$	$\searrow$	$S(\beta_0^*) \geq 0$	$\nearrow$	$+\infty$

Grâce à ce tableau, on vérifie que $\beta_0^* = \frac{1}{n} \sum_{i=1}^{n} y_i$ minimise bien la quantité $S(\beta_0)$.

Exercice 9 (Régression linéaire) *On dispose de n points (x_i, y_i), $i = 1, \ldots, n$ et on suppose que la fonction modèle est de la forme*

$$f(x, a, b) = ax + b$$

On cherche le minimum de

$$S(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Pour cela, il faut chercher où le gradient de S *vaut* 0*. On doit donc calculer les dérivées partielles de* S *par rapport à* a *et* b*. Celles-ci valent*

$$\begin{cases} \dfrac{\partial S}{\partial a} &= -2\displaystyle\sum_{i=1}^{n} x_i(y_i - ax_i - b) \\ \dfrac{\partial S}{\partial b} &= -2\displaystyle\sum_{i=1}^{n} (y_i - ax_i - b) \end{cases}$$

Le minimum de S *est donc atteint pour* (a,b) *solution du système suivant*

$$\begin{cases} \displaystyle\sum_{i=1}^{n} x_i(y_i - ax_i - b) &= 0 \\ \displaystyle\sum_{i=1}^{n} (y_i - ax_i - b) &= 0 \end{cases}$$

On note $\overline{x}$ *et* $\overline{y}$ *les moyennes des* x_i *et* y_i *:*

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Exprimer b *en fonction de* a*,* $\overline{x}$ *et* $\overline{y}$ *puis exprimer* a*.*

La seconde ligne du système à résoudre donne

$$\sum_{i=1}^{n} (y_i - ax_i - b) = 0 \Longleftrightarrow \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - b\underbrace{\sum_{i=1}^{n} 1}_{=n} = 0$$

D'où

$$b = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i\right) = \overline{y} - a\overline{x}$$

Maintenant réécrivons le système à résoudre :

$$\begin{cases} \displaystyle\sum_{i=1}^{n} x_i(y_i - ax_i - b) &= \displaystyle\sum_{i=1}^{n} x_i(y_i - \overline{y} - a(x_i - \overline{x})) &= 0 \\ \displaystyle\sum_{i=1}^{n} (y_i - ax_i - b) &= \displaystyle\sum_{i=1}^{n} (y_i - \overline{y} - a(x_i - \overline{x})) &= 0 \end{cases}$$

Si on multiplie la deuxième ligne par $\overline{x}$ et qu'on la soustrait à la première ligne, on obtient

$$\sum_{i=1}^{n} x_i(y_i - \overline{y} - a(x_i - \overline{x})) - \overline{x}\sum_{i=1}^{n} (y_i - \overline{y} - a(x_i - \overline{x})) = 0 - \overline{x} \times 0 = 0$$

Or

$$0 = \sum_{i=1}^{n} x_i(y_i - \overline{y} - a(x_i - \overline{x})) - \overline{x}\sum_{i=1}^{n}(y_i - \overline{y} - a(x_i - \overline{x})) = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y} - a(x_i - \overline{x}))$$
$$= \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) - a\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Et finalement, on obtient que

$$a = \frac{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Remarque : si vous obtenez le résultat suivant

$$a = \frac{\displaystyle\sum_{i=1}^{n}(x_i y_i - \overline{x}\,\overline{y})}{\displaystyle\sum_{i=1}^{n}(x_i^2 - \overline{x})} = \frac{\displaystyle\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\displaystyle\sum_{i=1}^{n} x_i^2 - n\overline{x}}$$

celui-ci est aussi valide. En effet, on l'égalité suivante pour le numérateur

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n}(x_i y_i - \overline{x} y_i - x_i\overline{y} + \overline{xy}) = \sum_{i=1}^{n} x_i y_i - \overline{x}\underbrace{\sum_{i=1}^{n} y_i}_{=n\overline{y}} - \overline{y}\underbrace{\sum_{i=1}^{n} x_i}_{=n\overline{x}} + n\overline{x}\,\overline{y} = \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}$$

et l'égalité suivante pour le dénominateur

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}(x_i^2 - 2\overline{x}x_i + \overline{x}^2) = \sum_{i=1}^{n} x_i^2 - 2\overline{x}\underbrace{\sum_{i=1}^{n} x_i}_{=n\overline{x}} + n\overline{x}^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

Exercice 10 (Moindres carrés linéaires) *On dispose de n points (x_i, y_i), $i = 1, \ldots, n$ et on suppose que la fonction modèle est de la forme*

$$f(x, \boldsymbol{\beta}) = \sum_{k=1}^{m} \beta_k \phi_k(x)$$

avec $\boldsymbol{\beta}$ le vecteur contenant tous les paramètres que nous allons chercher

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix}$$

On cherche le minimum de

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - f(x_i, \boldsymbol{\beta}))^2$$

Pour cela, il faut chercher où le gradient de S vaut 0. On doit donc calculer les dérivées partielles de S par rapport aux β_k. Celles-ci valent

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_k} = -2 \sum_{i=1}^{n} (y_i - f(x_i, \boldsymbol{\beta})) \frac{\partial f(x_i, \boldsymbol{\beta})}{\beta_k} = 0$$

Or on a pour tout $i = 1, \ldots, n$ et pour tout $k = 1, \ldots, m$

$$X_{ik} = \frac{\partial f(x_i, \boldsymbol{\beta})}{\beta_k} = \phi_k(x_i)$$

Posons la matrice et le vecteur

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1m} \\ X_{21} & X_{22} & \ldots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{nm} \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Montrer que $\boldsymbol{\beta}$ est minimum de S si le vecteur est solution du système suivant

$$\left(\mathbf{X}^T \mathbf{X}\right) \boldsymbol{\beta} = \mathbf{X}^T \mathbf{y}$$

Le minimum de S est atteint lorsque son gradient s'annule, ce qui veut dire que pour tout $k = 1, \ldots, m$

$$\sum_{i=1}^{n} (y_i - f(x_i, \boldsymbol{\beta})) \frac{\partial f(x_i, \boldsymbol{\beta})}{\beta_k} = 0$$

Cette égalité se réecrit de la manière suivante

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \beta_j X_{ij} \right) X_{ik} = 0$$

D'où pour tout $k = 1, \ldots, m$

$$\sum_{i=1}^{n} y_i X_{ik} = \sum_{i=1}^{n} \sum_{j=1}^{m} \beta_j X_{ij} X_{ik} = \sum_{j=1}^{m} \beta_j \left(\sum_{i=1}^{n} X_{ij} X_{ik} \right)$$

Maintenant étudions de plus près la relation matricielle

$$\left(\mathbf{X}^T \mathbf{X}\right) \boldsymbol{\beta} = \mathbf{X}^T \mathbf{y}$$

Avant cela, on fait un court rappel sur les matrices :

Soient $\mathbf{A}$ et $\mathbf{B}$ deux matrices, on note $[\mathbf{A}]_{ij}$ la valeur du coefficient de A à la ligne i, colonne j. Le produit $\mathbf{AB}$ n'est possible que s'il y a compatibilité des dimensions. Ainsi si $\mathbf{A}$ possède a_l lignes et a_c colonnes et si $\mathbf{B}$ possède b_1 lignes et b_2 colonnes, il faut que $a_c = b_l$ pour que le produit $\mathbf{AB}$ existe, cette matrice a d'ailleurs a_l lignes et b_c colonnes et pour tout $i = 1, \ldots, a_l$ et pour tout $k = 1, \ldots, b_c$

$$[\mathbf{AB}]_{ik} = \sum_{j=1}^{a_c} [\mathbf{A}]_{ij} [\mathbf{B}]_{jk}$$

Concernant ce qu'on appelle la matrice transposée de $\mathbf{A}$ notée $\mathbf{A}^T$, celle-ci a pour dimensions a_c lignes et a_l colonnes et

$$[\mathbf{A}^T]_{ij} = [\mathbf{A}]_{ji} \qquad i = 1, \ldots, a_c \qquad j = 1, \ldots, a_l$$

Revenons maintenant à l'exercice proprement dit. Étudions tout d'abors les compatibilité pour les dimensions, la matrice $\mathbf{X}$ possède n lignes et m colonnes donc la matrice $\mathbf{X}^T$ possède m lignes et n colonnes et $\mathbf{X}^T\mathbf{X}$ a m lignes et m colonnes. Le vecteur $\boldsymbol{\beta}$ possède lui m lignes et 1 colonne, donc le produit $(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta}$ est un vecteur (m lignes et bien s ?r 1 colonne). D'autre part, $\mathbf{y}$ est aussi un vecteur (n lignes, 1 colonne) donc $\mathbf{X}^T\mathbf{y}$ est aussi un vecteur (m lignes, 1 colonne). On a donc compatibilité pour les dimensions des deux côtés de l'égalité. Maintenant, on a

$$[\mathbf{X}^T\mathbf{X}]_{jk} = \sum_{i=1}^{n}[\mathbf{X}^T]_{ji}[\mathbf{X}]_{ik} = \sum_{i=1}^{n}[\mathbf{X}]_{ij}[\mathbf{X}]_{ik} = \sum_{i=1}^{n} X_{ij}X_{ik} \qquad j=1,\dots,m \qquad k=1,\dots,m$$

et la k ème composante du vecteur $(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta}$ vaut

$$[(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta}]_{k1} = \sum_{j=1}^{m}[\mathbf{X}^T\mathbf{X}]_{jk}[\boldsymbol{\beta}]_{j1} = \sum_{j=1}^{m}\beta_j\left(\sum_{i=1}^{n}X_{ij}X_{ik}\right) \qquad k=1,\dots,m$$

D'autre part, la k ème composante du vecteur $\mathbf{X}^T\mathbf{y}$ vaut

$$[\mathbf{X}^T\mathbf{y}]_{k1} = \sum_{i=1}^{n}[\mathbf{X}^T]_{ki}[\mathbf{y}]_{i1} = \sum_{i=1}^{n}[\mathbf{X}]_{ik}y_i = \sum_{i=1}^{n}X_{ik}y_i \qquad k=1,\dots,m$$

Or on a vu précédemment que

$$\sum_{i=1}^{n} y_i X_{ik} = \sum_{j=1}^{m}\beta_j\left(\sum_{i=1}^{n}X_{ij}X_{ik}\right)$$

donc pour tout $k=1,\dots,m$

$$[(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta}]_{k1} = [\mathbf{X}^T\mathbf{y}]_{k1}$$

On vérifie bien l'égalité vectorielle

$$(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}$$

Exercice 11 *On dispose de n points (x_i, y_i), $i=1,\dots,n$ et on suppose que la fonction modèle est de la forme*

$$f(x,\beta_0) = \beta_0$$

Chaque mesure y_i est entâchée d'une erreur dont l'écart-type est estimé à $\sigma_i = \frac{1}{\sqrt{w_i}}$. ($w_i \geq 0$). Trouver β_0^* minimum de

$$\chi^2(\beta_0) = \sum_{i=1}^{n} w_i(y_i-\beta_0)^2$$

On cherche le minimum β_0^* tel que $\chi^2(\beta_0^*) = \min_{\beta_0\in\mathbb{R}}\chi^2(\beta_0)$. Pour cela, on a besoin de calculer sa dérivée $\chi^{2'}(\beta_0)$. Celle-ci vaut

$$\chi^{2'}(\beta_0) = \sum_{i=1}^{n}[-2w_i(y_i - f(x_i,\beta_0))] = -2\sum_{i=1}^{n}w_i(y_i-\beta_0)$$

Pour trouver un extremum local (minimum ou maximum local), il faut résoudre le problème $\chi^{2'}(\beta_0) = 0$.

Dans notre cas, on en déduit que

$$\beta_0^* = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

β_0^* correspond à la moyenne pondérée des y_i. Il reste cependant à vérifier que β_0^* est bien un minimum et que celui-ci est global. Pour cela, étudions le comportement de $\chi^2(\beta_0)$. On a le tableau des variations suivant

β_0	$-\infty$		β_0^*		β_0
$\chi^{2'}(\beta_0)$		$-$	0	$+$	
$\chi^2(\beta_0)$	$+\infty$	$\searrow$	$\chi^2(\beta_0^*) \geq 0$	$\nearrow$	$+\infty$

Grâce à ce tableau, on vérifie que $\beta_0^* = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$ minimise bien la quantité $\chi^2(\beta_0)$.

Interpolation polynômiale

Exercice 12 *On dispose de $n+1$ points (x_i, y_i), $i = 0, \dots, n$. On suppose $x_0 < x_1 < x_n$. On construit f fonction d'interpolation entre x_0 et x_n tel que les points (x_i, y_i) soient reliés entre eux par des droites. Exprimer f sur $[x_0, x_n]$.*

Sur le tronçon $[x_i, x_{i+1}]$, f vaut une droite d'interpolation de la forme $ax+b$. Or $f(x_i) = y_i$ et $f(x_{i+1}) = y_{i+1}$. Il reste donc à trouver les coefficients a et b pour que la droite passe par les 2 points indiqués précédemment. Pour cela, il faut résoudre le système suivant

$$\left\{ \begin{array}{r} ax_i + b = y_i \\ ax_{i+1} + b = y_{i+1} \end{array} \right.$$

En soustrayant la première ligne à la seconde, on a

$$a(x_{i+1} - x_i) = (y_{i+1} - y_i)$$

On en déduit que

$$a = \frac{y_{i+1} - y_i}{x_{i+1} - x_i}$$

et b vaut

$$b = y_i - ax_i = y_i - \frac{y_{i+1} - y_i}{x_{i+1} - x_i} x_i$$

Donc sur le tronçon $[x_i, x_{i+1}]$, f vaut

$$f(x) = \frac{y_{i+1} - y_i}{x_{i+1} - x_i} x + y_i - \frac{y_{i+1} - y_i}{x_{i+1} - x_i} x_i = \frac{y_{i+1} - y_i}{x_{i+1} - x_i}(x - x_i) + y_i$$

pour $i = 0, \ldots, n-1$.

Exercice 13 *On dispose d'un ensemble de $n+1$ points (x_i, y_i), $i = 0, \ldots, n$ tel que si $i \neq j$, $x_i \neq x_j$. L'interpolation polynômiale sous **forme lagrangienne** noté L est une combinaison linéaire*

$$L(x) = \sum_{j=0}^{n} y_j \, \ell_j(x)$$

de polynômes de Lagrange

$$\ell_j(x) = \prod_{k=0, k\neq j}^{n} \frac{x - x_k}{x_j - x_k} = \frac{x - x_0}{x_j - x_0} \cdots \frac{x - x_{j-1}}{x_j - x_{j-1}} \frac{x - x_{j+1}}{x_j - x_{j+1}} \cdots \frac{x - x_n}{x_j - x_n}$$

Les polynômes de Lagrange sont de degré n donc L est au maximum de degré n.

Montrer que $L(x_i) = y_i$, pour $i = 0, \ldots, n$.

On veut montrer que $L(x_i) = y_i$ pour $i = 0, \ldots, n$. Pour cela, il faut tout d'abord calculer les $\ell_j(x_i)$. Pour $i \neq j$

$$\ell_j(x_i) = \prod_{k=0, k\neq j}^{n} \frac{x_i - x_k}{x_j - x_k}$$

Dans ce cas, l'indice k peut valoir i donc un des termes du produit est $\frac{x_i - x_i}{x_j - x_i}$ qui vaut 0 et le produit est forcément nul. On a ainsi que $\ell_j(x_i) = 0$ si $i \neq j$. Maintenant si $i = j$,

$$\ell_j(x_j) = \prod_{k=0, k\neq j}^{n} \frac{x_j - x_k}{x_j - x_k} = 1$$

On en déduit que

$$L(x_i) = \sum_{j=0}^{n} y_j \, \ell_j(x_i) = y_i$$

Exercice 14 *On considère l'ensemble de points suivants*

$$\left(-1, \frac{1}{26}\right) \ , \ (0, 1) \ , \ \left(1, \frac{1}{26}\right)$$

Calculer le polynôme d'interpolation L passant par ces trois points.

Tout d'abord, il faut calculer les $\ell_j(x)$, j valant 0, 1 ou 2. On a

$$\ell_0(x) = \prod_{k=0,k\neq 0}^{2} \frac{x_i - x_k}{x_j - x_k} = \frac{x - x_1}{x_0 - x_1}\frac{x - x_2}{x_0 - x_2} = \frac{x - 0}{-1 - 0}\frac{x - 1}{-1 - 1} = \frac{x(x-1)}{2} = \frac{1}{2}x^2 - \frac{1}{2}x$$

$$\ell_1(x) = \prod_{k=0,k\neq 1}^{2} \frac{x_i - x_k}{x_j - x_k} = \frac{x - x_0}{x_1 - x_0}\frac{x - x_2}{x_1 - x_2} = \frac{x - (-1)}{0 - (-1)}\frac{x - 1}{0 - 1} = -(x+1)(x-1) = -x^2 + 1$$

$$\ell_2(x) = \prod_{k=0,k\neq 2}^{2} \frac{x_i - x_k}{x_j - x_k} = \frac{x - x_0}{x_2 - x_0}\frac{x - x_2}{x_2 - x_1} = \frac{x - (-1)}{1 - (-1)}\frac{x - 0}{1 - 0} = \frac{x(x+1)}{2} = \frac{1}{2}x^2 + \frac{1}{2}x$$

D'où

$$L(x) = y_0\,\ell_0(x) + y_1\,\ell_1(x) + y_2\,\ell_2(x) = \frac{1}{26}\left(\frac{1}{2}x^2 - \frac{1}{2}x\right) - x^2 + 1 + \frac{1}{26}\left(\frac{1}{2}x^2 + \frac{1}{2}\right) = \frac{1}{26}x^2 - x^2 + 1 = -\frac{25}{26}x^2 + 1$$

Exercice 15 *On dispose d'un ensemble de $n+1$ points (x_i, y_i), $i = 0, \ldots, n$ tel que $x_0 < x_1 < \ldots x_n$. La spline cubique S interpolant ces points coïncide sur chaque intervalle $[x_i, x_{i+1}]$ avec un polynôme p_i de degré 3 de la forme :*

$$S(x) = p_i(x) = f_i + f_i'(x - x_i) + \frac{f_i''}{2!}(x - x_i)^2 + \frac{f_i'''}{3!}(x - x_i)^3$$

Les coefficients à déterminer sont les f_i, f_i', f_i'' et f_i'''.

On a $n+1$ points d'interpolations donc n intervalles $[x_i, x_{i+1}]$ et $4n$ inconnues à déterminer. On veut naturellement que

$$S(x_i) = p_i(x_i) = y_i \qquad i = 0, \ldots, n-1$$
$$S(x_{i+1}) = p_i(x_{i+1}) = y_i \qquad i = 0, \ldots, n-1$$

ce qui permet d'assurer la continuité de la spline cubique S. n a donc $4n$ inconnues et $2n$ relations utilisées pour les conditions usuelles. On se sert des $2n$ degrés de libertés restant pour imposer que la spline S soit $\mathcal{C}^2$. Pour que S soit $\mathcal{C}^2$ sur $[x_0, x_n]$, il faut que

$$S'(x_{i+1}) = p_i'(x_{i+1}) = p_{i+1}'(x_{i+1}) \qquad i = 0, \ldots, n-2$$
$$S''(x_{i+1}) = p_i''(x_{i+1}) = p_{i+1}''(x_{i+1}) \qquad i = 0, \ldots, n-2$$

On vient d'imposer $2 \times (n-1)$ conditions, il reste encore 2 ***degrés de liberté****. En général, pour terminer on prend $S''(x_0) = p_0''(x_0) = 0$ et $S''(x_n) = p_{n-1}''(x_n) = 0$ (splines naturelles) mais de nombreux autres choix sont possibles.*

On cherche maintenant à évaluer les coefficients f_i, f_i', f_i'' et f_i'''. Pour cela, il faut répondre aux questions suivantes

1. *Exprimer $p_i(x)$, $p_i'(x)$ et $p_i''(x)$ en fonction des f_i, f_i', f_i'' et f_i'''.*
2. *Montrer à partir de la condition $p_i(x_i) = y_i$ que*

$$f_i = y_i \qquad i = 0, \ldots, n-1$$

3. *Notons $h_i = x_{i+1} - x_i$ pour $i = 0, \ldots, n-1$ et $f[x_i, x_{i+1}] = \frac{y_{i+1} - y_i}{h_i}$. Montrer à partir de la condition $p_i(x_{i+1}) = y_{i+1}$ que*

$$f_i' = f[x_i, x_{i+1}] - \frac{f_i''}{2}h_i - \frac{f_i'''}{6}h_i^2 \qquad i = 0, \ldots, n-1$$

4. Montrer à partir de la condition $p_i''(x_{i+1}) = p_{i+1}''(x_{i+1})$ que

$$f_i''' = \frac{f_{i+1}'' - f_i''}{h_i} \qquad i = 0, \dots, n-2$$

5. On pose $f_n'' = p_{n-1}''(x_n)$. On a donc $f_{n-1}''' = \frac{f_n'' - f_{n-1}''}{h_{n-1}}$. Montrer que

$$f_i' = f[x_i, x_{i+1}] - \frac{1}{3}h_i f_i'' - \frac{1}{6}h_i f_{i+1}'' \qquad i = 0, \dots, n-1$$

6. Montrer à partir de la condition $p_i'(x_{i+1}) = p_{i+1}'(x_{i+1})$ que

$$h_i f_i'' + 2(h_i + h_{i+1})f_{i+1}'' + h_{i+1}f_{i+2}'' = 6(f[x_{i+1}, x_{i+2}] - f[x_i, x_{i+1}])$$

pour $i = 0, \dots, n-2$.

Il ne reste plus qu'à imposer f_0'' et f_n'' pour obtenir tous les coefficients.

1. On veut exprimer $p_i(x)$, $p_i'(x)$ et $p_i''(x)$ en fonction des f_i, f_i', f_i'' et f_i'''. Pour rappel, on a pour x dans $[x_i, x_{i+1}]$.

$$p_i(x) = f_i + f_i'(x - x_i) + \frac{f_i''}{2!}(x - x_i)^2 + \frac{f_i'''}{3!}(x - x_i)^3$$

Donc sur le même intervalle, on a aussi

$$p_i(x) = f_i + f_i'(x - x_i) + \frac{f_i''}{2}(x - x_i)^2 + \frac{f_i'''}{6}(x - x_i)^3$$

$$p_i'(x) = f_i' + f_i''(x - x_i) + \frac{f_i'''}{2}(x - x_i)^2$$

$$p_i''(x) = f_i'' + f_i'''(x - x_i)$$

2. On veut tout d'abord respecter la condition $p_i(x_i) = y_i$. Or

$$p_i(x_i) = f_i + f_i'(x_i - x_i) + \frac{f_i''}{2}(x_i - x_i)^2 + \frac{f_i'''}{6}(x_i - x_i)^3 = f_i$$

Pour respecter la condition voulue, il faut donc imposer que

$$f_i = y_i \qquad i = 0, \dots, n-1$$

3. On note $h_i = x_{i+1} - x_i$ pour $i = 0, \dots, n-1$ et $f[x_i, x_{i+1}] = \dfrac{y_{i+1} - y_i}{h_i}$. On veut respecter la condition $p_i(x_{i+1}) = y_{i+1}$, il faut donc que l'égalité suivante soit vraie

$$p_i(x_{i+1}) = \underbrace{f_i}_{=y_i} + f_i'h_i + \frac{f_i''}{2}h_i^2 + \frac{f_i'''}{6}h_i^3 = y_{i+1}$$

D'où

$$f_i'h_i = y_{i+1} - y_i - \frac{f_i''}{2}h_i^2 - \frac{f_i'''}{6}h_i^3$$

En divisant par h_i l'égalité précédente, on obtient la condition voulue

$$f_i' = \underbrace{f[x_i, x_{i+1}]}_{= \frac{y_{i+1} - y_i}{h_i}} - \frac{f_i''}{2}h_i - \frac{f_i'''}{6}h_i^2 \qquad i = 0, \dots, n-1$$

4. On veut maintenant satisfaire la condition $p_i''(x_{i+1}) = p_{i+1}''(x_{i+1})$. Or

$$p_i''(x_{i+1}) = f_i'' + f_i'''h_i$$

$$p_{i+1}''(x_{i+1}) = f_{i+1}''$$

Donc pour honorer la condition précédente, il faut imposer que

$$f_i'' + f_i'''h_i = f_{i+1}''$$

Et en divisant par h_i, il faut que

$$f_i''' = \frac{f_{i+1}'' - f_i''}{h_i} \qquad i = 0, \ldots, n-2$$

5. On pose $f_n'' = p_{n-1}''(x_n)$. On a donc $f_{n-1}''' = \dfrac{f_n'' - f_{n-1}''}{h_{n-1}}$ (car $f_n'' = p_{n-1}''(x_n) = f_{n-1}'' + h_{n-1}f_{n-1}'''$) et

$$f_i''' = \frac{f_{i+1}'' - f_i''}{h_i} \qquad i = 0, \ldots, n-1$$

En réinjectant ce résultat dans l'égalité trouvée à la question 3., on a

$$f_i' = f[x_i, x_{i+1}] - \frac{f_i''}{2}h_i - \frac{f_i'''}{6}h_i^2 = f[x_i, x_{i+1}] - \frac{f_i''}{2}h_i - \frac{h_i}{6}\frac{f_{i+1}'' - f_i''}{h_i}$$

D'où

$$f_i' = f[x_i, x_{i+1}] - \frac{h_i}{3}f_i'' - \frac{h_i}{6}f_{i+1}'' \qquad i = 0, \ldots, n-1$$

6. On veut enfin satisfaire la dernière condition $p_i'(x_{i+1}) = p_{i+1}'(x_{i+1})$. Or

$$\begin{aligned} p_i'(x_{i+1}) = f_i' + f_i''h_i + \frac{f_i'''}{2}h_i^2 &= f[x_i, x_{i+1}] - \frac{h_i}{3}f_i'' - \frac{h_i}{6}f_{i+1}'' + f_i''h_i + \frac{h_i^2}{2}\frac{f_{i+1}'' - f_i''}{h_i} \\ &= f[x_i, x_{i+1}] + \frac{1}{6}h_if_i'' + \frac{1}{3}h_if_{i+1}'' \end{aligned}$$

$$p_{i+1}'(x_{i+1}) = f_{i+1}' = f[x_{i+1}, x_{i+2}] - \frac{h_{i+1}}{3}f_{i+1}'' - \frac{h_{i+1}}{6}f_{i+2}''$$

Pour vérifier l'égalité voulue, il faut imposer que

$$f[x_i, x_{i+1}] + \frac{1}{6}h_if_i'' + \frac{1}{3}h_if_{i+1}'' = f[x_{i+1}, x_{i+2}] - \frac{h_{i+1}}{3}f_{i+1}'' - \frac{h_{i+1}}{6}f_{i+2}''$$

En multipliant par 6 l'égalité précedente et en réorganisant, on obtient

$$6(f[x_{i+1}, x_{i+2}] - f[x_i, x_{i+1}]) = h_if_i'' + 2(h_i + h_{i+1})f_{i+1}' + h_{i+1}f_{i+2}'' \qquad i = 0, \ldots, n-2$$

Chapitre 4 : Intégration numérique

Exercice 16 *1. Soit f une fonction intégrable sur le segment $[a, b]$ $(a < b)$. Quelle est la valeur moyenne de f sur ce segment ?*

2. *Dans le cas du traitement du signal, on peut vouloir connaitre la valeur moyenne $\tilde{f}(t)$ d'un signal f sur le segment $[0,t]$ ($t > 0$, variable dans le temps). Déduire de la question précédente l'expression de la fonction valeur moyenne $\tilde{f}(t)$.*

1. Notons f_{moy} la valeur moyenne de la fonction f sur $[a,b]$. f_{moy}, pour être une valeur moyenne, doit respecter la condition suivante

$$\int_a^b f_{moy}\,dt = \int_a^b f(t)\,dt$$

(on a le raisonnement analogue pour la moyenne arithmétique, $f_{moy} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} f_i$ donc f_{moy} doit respecter l'égalité $nf_{moy} = \displaystyle\sum_{i=1}^{n} f_i$). Donc

$$(b-a)f_{moy} = \int_a^b f(t)\,dt$$

et la valeur moyenne f_{moy} de f sur le segment $[a,b]$ vaut

$$f_{moy} = \frac{1}{b-a}\int_a^b f(t)\,dt$$

2. On veut maintenant exprimer la valeur moyenne $\tilde{f}(t)$ d'un signal (c'est une fonction) f sur le segment $[0,t]$ ($t > 0$). D'après la question précédente, on a

$$\tilde{f}(t) = \frac{1}{t}\int_0^t f(t)\,dt$$

Exercice 17 *Écrire un programme Scilab permettant d'estimer l'intégrale de $\dfrac{1}{1+x^2}$ sur $[0,1]$ par la méthode de Monte Carlo avec pour entrée N le nombre de valeurs aléatoires tirées pour calculer cette intégrale.*

Pour rappel, la fonction `rand(n,m)` retourne une matrice de taille $n \times m$ contenant des nombres aléatoires de loi uniforme compris entre 0 et 1.

On rappelle tout d'abord la méthode de Monte Carlo pour l'estimation d'une intégrale. On veut intégrer $f(x) = \dfrac{1}{1+x^2}$ sur $[0,1]$. Il faut tout d'abord calculer la quantité V définie par

$$V = \int_0^1 dx = 1$$

On peut maintenant rapeler la méthode de Monte Carlo. Celle-ci consiste donc à tirer aléatoirement de manière uniforme des valeurs $x_i \in [0,1]$, $i = 1,\ldots,N$ puis à approcher l'intégrale par

$$\int_0^1 f(x)\,dx \approx \frac{V}{N}\sum_{i=1}^{N} f(x_i) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+x_i^2}$$

Puisque la théorie est claire, passons à la mise en oeuvre pratique soit le code Scilab. Voici un exemple de solution au problème :

```
function QN = integraleMC(N)
   -- QN va contenir le resultat de l'integrale approchee
   QN = 0 ;
   -- boucle de calcul, on doit simuler N x_i puis ajouter f(x_i)/N a QN
   for k = 1:N
      -- k eme x_i simule
      u = rand(1,1) ;
      -- ajout de f(x_i)/N a QN
      QN = QN + (1./N).*(1./(1 + u.*u)) ;
   end
endfunction
```

On vient de donner un algorithme permettant de calculer

$$\int_0^1 \frac{1}{1+x^2}\,dx = \mathsf{arctan}(1)$$

Exercice 18 *On veut approcher l'intégrale*

$$I(f) = \int_{-1}^{1} f(x)\,dx$$

par la formule de quadrature suivante

$$A_0 f(-1) + A_1 f(1)$$

Calculer A_0 et A_1 pour que cette formule de quadrature soit de degré de précision au moins 1. Montrer que cette formule de quadrature a pour degré de précision 1.

On cherche une formule de quadrature à 2 points, elle doit donc être exacte sur l'ensemble des polynômes de degré au plus 1. Donc si $f(x) = 1$, $I(f) = A_0 + A_1$ et si $f(x) = x$, $I(f) = -A_0 + A_1$. Pour calculer A_0 et A_1, il faut donc résoudre le système suivant

$$\left\{ \begin{array}{rcccccc} A_0 & + & A_1 & = & I(1) & = & 2 \\ -A_0 & + & A_1 & = & I(x) & = & 0 \end{array} \right.$$

Si on ajoute la première à la seconde ligne, on a $2A_1 = 2$, d'où $A_1 = 1$. De plus, d'après la seconde ligne, $A_0 = A_1$. Finalement, la formule de quadrature est

$$f(-1) + f(1)$$

D'après ce qui précède, cette formule de quadrature est de degré de précision au moins 1. Étudions maintenant le cas où $f(x) = x^2$

$$f(x) = x^2 \qquad I(f) = \int_{-1}^{1} x^2 = \frac{2}{3} \neq f(-1) + f(1) = 2$$

La formule de quadrature n'est donc pas exacte pour $f(x) = x^2$. On en conclut qu'elle est bien de degré de précision 1.

Exercice 19 (Formule des trapèzes) *On cherche à approcher l'intégrale*

$$I(f) = \int_a^b f(x)\,dx \qquad a < b$$

par la formule des trapèzes (cas $n = 1$ des formules de Newton-Cotes fermé). On approxime ainsi cette intégrale par la formule de quadrature suivante

$$(b-a)\left(w_0^{(1)} f(a) + w_1^{(1)} f(b)\right)$$

avec

$$w_i^{(1)} = \frac{1}{b-a}\int_a^b \ell_i(x)\,dx$$

et $\ell_i(x)$, polynôme de Lagrange nécessaire pour l'interpolation.

Calculer les poids $w_0^{(1)}$ et $w_1^{(1)}$.

Pour cette formule de quadrature, les nœuds sont les points a et b donc pour calculer les poids $w_0^{(1)}$ et $w_1^{(1)}$, il faut connaitre les les polynôme de Lagrange $\ell_0(x)$ et $\ell_1(x)$ nécessaire pour le calcul du polynôme d'interpolation passant par les points $(x_0, y_0) = (a, f(a))$ et $(x_1, y_1) = (b, f(b))$. Pour rappel,

$$\ell_i(x) = \prod_{k=0, k\neq i}^{1} \frac{x - x_k}{x_j - x_k}$$

Dans notre cas, on a alors

$$\ell_0(x) = \frac{x - x_1}{x_0 - x_1} = \frac{x-b}{a-b}$$
$$\ell_1(x) = \frac{x - x_0}{x_1 - x_0} = \frac{x-a}{b-a}$$

D'où

$$w_0^{(1)} = \frac{1}{b-a}\int_a^b \ell_0(x)\,dx = \frac{1}{(b-a)(a-b)}\int_a^b (x-b)\,dx = \frac{1}{(b-a)(a-b)}\left[\frac{(x-b)^2}{2}\right]_a^b = -\frac{(a-b)^2}{2(b-a)(a-b)} = \frac{1}{2}$$

$$w_1^{(1)} = \frac{1}{b-a}\int_a^b \ell_1(x)\,dx = \frac{1}{(b-a)^2}\int_a^b (x-a)\,dx = \frac{1}{(b-a)^2}\left[\frac{(x-a)^2}{2}\right]_a^b = \frac{(b-a)^2}{2(b-a)^2} = \frac{1}{2}$$

La formule des trapèzes est donc

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2}\left(f(a) + f(b)\right)$$

Exercice 20 (Formule de Simpson) *On cherche à approcher l'intégrale*

$$I(f) = \int_a^b f(x)\,dx \qquad a < b$$

par la formule de Simpson (cas $n = 2$ des formules de Newton-Cotes fermé)

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

Montrer que la formule de Simpson est de degré de précision 3.

La formule de Simpson est une formule de quadrature de type interpolation à 3 points. D'après l'un des théorèmes du cours, elle est donc exacte pour tous les polynômes de degré au plus 2. Elle est donc de degré de précision au moins 2, ce que confirme les calculs suivants

$$f(x) = 1 \qquad I(f) = b - a \qquad \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right) = \frac{b-a}{6}(1+4+1) = b - a$$

La formule de Simpson a au moins un degré de précision supérieur à 0.

$$f(x) = x \qquad I(f) = \frac{b^2 - a^2}{2}$$

$$\frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right) = \frac{b-a}{6}(a + 2(a+b) + b) = \frac{3(b-a)}{6}(a+b) = \frac{b^2-a^2}{2}$$

La formule de Simpson a au moins un degré de précision supérieur à 1.

$$f(x) = x^2 \qquad I(f) = \frac{b^3 - a^3}{3}$$

$$\frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right) = \frac{b-a}{6}(a^2 + (a+b)^2 + b^2) = \frac{b-a}{6}(2a^2 + 2ab + 2b^2) = \frac{b^3-a^3}{3}$$

La formule de Simpson a au moins un degré de précision supérieur à 2. Maintenant pour le degrés supérieurs, on a

$$f(x) = x^3 \qquad I(f) = \frac{b^4 - a^4}{4}$$

$$\begin{aligned}
\frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right) &= \frac{b-a}{6}\left(a^3 + \frac{1}{2}(a+b)^3 + b^3\right) \\
&= \frac{b-a}{6}\left(a^3 + \frac{1}{2}a^3 + \frac{3}{2}a^2b + \frac{3}{2}ab^2 + \frac{1}{2}b^3 + b^3\right) = \frac{b-a}{4}(a^3 + a^2b + ab^2 + b^3) = \frac{b^4-a^4}{4}
\end{aligned}$$

La formule de Simpson a au moins un degré de précision supérieur à 3.

$$f(x) = x^4 \qquad I(f) = \frac{b^5 - a^5}{5}$$

$$\begin{aligned}
\frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right) &= \frac{b-a}{6}\left(a^4 + \frac{1}{4}(a+b)^4 + b^4\right) \\
&= \frac{b-a}{6}\left(a^4 + \frac{1}{4}a^4 + a^3b + \frac{3}{2}a^2b^2 + ab^3 + \frac{1}{4}b^4 + b^3\right) = \frac{b-a}{24}(5a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + 5b^4) \\
&= \frac{5}{24}b^5 - \frac{1}{24}ab^4 + \frac{1}{12}a^2b^3 - \frac{1}{12}a^3b^2 + \frac{1}{24}a^4b - \frac{5}{24}a^5
\end{aligned}$$

Le résultat obtenu par la formule de quadrature est différent de la valeur de l'intégrale pour $f(x) = x^4$. On en conclut que la formule de Simpson est de degré de précision 3.

Exercice 21 (Formule des rectangles) *Établir la formule des rectangles pour Newton-Cotes ouvert* $n = 0$ *puis montrer que cette formule est de degré de précision* 1.

Pour obtenir les formules de Newton-Cotes ouvert, on interpole f aux point suivants

$$x_i = a + (i+1)h \qquad i = 0, \ldots, n \qquad \text{avec} \qquad h = \frac{b-a}{n+2}$$

et on construit les formules de quadratures de la façon suivante :

$$\int_a^b f(x)\, dx \approx (b-a) \sum_{i=0}^{n} w_i^{(n)} f(x_i)$$

avec

$$w_i^{(n)} = \frac{1}{b-a} \int_a^b \ell_i(x)\, dx$$

et $\ell_i(x)$ est le polynôme de Lagrange défini de la manière suivante

$$\ell_i(x) = \prod_{k=0, k\neq i}^{n} \frac{x - x_k}{x_i - x_k}$$

avec $(x_i, f(x_i))$ les points d'interpolation utilisé.

Dans notre cas, $n = 0$ et on interpole en un seul point x_0. D'autre part, h vaut ici $\dfrac{b-a}{2}$ donc

$$x_0 = a + h = a + \frac{b-a}{2} = \frac{a+b}{2}$$

La formule des rectangle est donc de la forme

$$(b-a) w_0^{(0)} f\left(\frac{a+b}{2}\right)$$

Il ne reste plus qu'à évaluer $w_0^{(0)}$. On a

$$\ell_0(x) = \prod_{k=0, k\neq 0}^{0} \frac{x - x_k}{x_i - x_k} = 1$$

Pour s'en convaincre, on rappelle ici que le degré d'un polynôme de Lagrange est n (avec $n+1$ le nombre de points d'interpolation). Ici $n = 0$ donc $\ell_0(x)$ est un polynôme de degré 0 donc une constante (différente de 0). D'autre part, on a montré dans un exercice précédent que

$$\ell_i(x_j) = 0 \text{ si } i \neq j \qquad\qquad \ell_i(x_i) = 1$$

donc $\ell_0(x_0) = 1$ d'où $\ell_0(x) = 1$. Maintenant, on peut calculer $w_0^{(0)}$ et on a

$$w_0^{(0)} = \frac{1}{b-a} \int_a^b 1\, dx = \frac{b-a}{b-a} = 1$$

et la formule des rectangles (pour Newton-Cotes ouvert) est

$$(b-a) f\left(\frac{a+b}{2}\right)$$

Il nous faut maintenant montrer que cette formule est de degré de précision 1. On a déjà

$$f(x) = 1 \qquad I(f) = b - a \qquad (b-a) f\left(\frac{a+b}{2}\right) = b - a$$

La formule des rectangles a au moins un degré de précision supérieur à 0.

$$f(x) = x \qquad I(f) = \frac{b^2 - a^2}{2} \qquad (b-a) f\left(\frac{a+b}{2}\right) = (b-a)\frac{a+b}{2} = \frac{b^2 - a^2}{2}$$

La formule des rectangles a au moins un degré de précision supérieur à 1.

$$f(x) = x^2 \qquad I(f) = \frac{b^3 - a^3}{3}$$

$$(b-a) f\left(\frac{a+b}{2}\right) = (b-a)\left(\frac{a+b}{2}\right)^2 = \frac{b-a}{4}\left(a^2 + 2ab + b^2\right) = \frac{1}{4}(b^3 - a^2 b + ab^2 - a^3)$$

Le résultat obtenu par la formule de quadrature est différent de la valeur de l'intégrale pour $f(x) = x^2$. On en conclut que la formule des rectangles est de degré de précision 1.

Exercice 22 (Degré de précision des formules de Newton-Cotes) *Pour une formule de Newton-Cotes associée à une valeur impaire de n. Si $f \in \mathcal{C}^{n+1}([a,b])$, alors il existe un réel $K_n \neq 0$ et $\xi \in]a,b[$ tel que l'erreur R commise sur la valeur de l'intégrale soit*

$$R(f) = \frac{K_n}{(n+1)!}\,(b-a)^{n+2}\, f^{(n+1)}(\xi)$$

Pour une formule de Newton-Cotes associée à une valeur paire de n (sauf $n = 0$ pour Newton-Cotes fermé). Si $f \in \mathcal{C}^{n+2}([a,b])$, alors il existe un réel $M_n \neq 0$ et $\xi \in]a,b[$ tel que l'erreur R commise sur la valeur de l'intégrale soit

$$R(f) = \frac{M_n}{(n+2)!}\,(b-a)^{n+3}\, f^{(n+2)}(\xi)$$

Déduire le degré de précision des formules de Newton-Cotes à partir de ce qui précède.

On rappelle tout d'abord qu'une formule de quadrature est de degré de précision n si cette formule est exacte (donne la valeur exacte de l'intégrale approchée) pour toute fonction $f(x) = x^k$ avec $k \leq n$ et inexacte pour $f(x) = x^{n+1}$.

D'autre part, on rappelle que toute fonction de la forme x^k avec k entier positif est dérivable à l'infini sur $\mathbb{R}$ et ces dérivées successives sont aussi continues sur $\mathbb{R}$. On peut aussi montrer que si $f(x) = x^k$, alors la p ème dérivée ($p \leq k$) de f vaut

$$f^{(p)}(x) = p!\, x^{k-p}$$

et si $p > k$, $f^{(p)}(x) = 0$.

Maintenant revenons à l'exercice et étudions le cas n impair, on a

$$R(f) = \frac{K_n}{(n+1)!}\,(b-a)^{n+2}\, f^{(n+1)}(\xi)$$

Si $f(x) = x^k$ avec $k < n+1$, alors $f^{(n+1)}(x) = 0$ et $R(f) = 0$ (la formule de quadrature est donc exacte pour ces f choisis). On en déduit que le degré de précision des formules de Newton-Cotes avec n impair est au moins n. Si $f(x) = x^{n+1}$, alors $f^{(n+1)}(x) = (n+1)!$ et

$$R(f) = K_n\,(b-a)^{n+2} \neq 0$$

Les formules de Newton-Cotes ont un degré de précision n si n est impair.

On étudie ensuite le cas n pair (on exclue le cas Newton-Cotes fermé $n = 0$), on a

$$R(f) = \frac{M_n}{(n+2)!}\,(b-a)^{n+3}\,f^{(n+2)}(\xi)$$

Si $f(x) = x^k$ avec $k < n+2$, alors $f^{(n+2)}(x) = 0$ et $R(f) = 0$ (la formule de quadrature est donc exacte pour ces f choisis). On en déduit que le degré de précision des formules de Newton-Cotes avec n impair est au moins $n+1$. Si $f(x) = x^{n+2}$, alors $f^{(n+2)}(x) = (n+2)!$ et

$$R(f) = M_n\,(b-a)^{n+3} \neq 0$$

Les formules de Newton-Cotes ont un degré de précision $n+1$ si n est pair.

Exercice 23 (Formule composite des trapèzes) *La méthode pour obtenir une formule composite consiste à diviser l'intervalle $[a,b]$ en r sous-intervalles de longueur*

$$h = \frac{b-a}{r}$$

et d'introduire les points de subdivision

$$t_i = a + ih \qquad i = 0, \ldots, r$$

On forme

$$\int_a^b f(x)\,dx = \sum_{i=0}^{r-1} \int_{t_i}^{t_{i+1}} f(x)\,dx$$

et l'on applique sur chaque intervalle $[t_i, t_{i+1}]$ une des formules de Newton-Cotes.

On veut utiliser pour formule de Newton-Cotes la formule des trapèzes (Newton-Cotes fermé). Établir la formule de quadrature composite décrite précédemment. En combien de points est il nécessaire d'évaluer f pour pouvoir utiliser cette formule ?

On applique sur chaque intervalle $[t_i, t_{i+1}]$ la formule des trapèzes (Newton-Cotes fermé)

$$\int_{t_i}^{t_{i+1}} f(x)\,dx \approx \frac{h}{2}\,(f(t_i) + f(t_{i+1}))$$

et si on somme sur chaque intervalle, on obtient

$$\sum_{i=0}^{r-1} \int_{t_i}^{t_{i+1}} f(x)\,dx \approx \sum_{i=0}^{r-1} \frac{h}{2}\,(f(t_i) + f(t_{i+1}))$$

Après quelques simplifications, la formule composite des trapèzes s'écrit

$$h\left(\frac{f(a)}{2} + \sum_{i=1}^{r-1} f(t_i) + \frac{f(b)}{2}\right)$$

Pour la calculer, il est nécessaire d'évaluer f en les points t_i, $i = 0, \ldots, r$ ($a = t_0$ et $b = t_r$). Il faut donc évaluer f $r+1$ fois.

Exercice 24 (Un exemple de formule de Gauss) *Soit la formule de quadrature de Gauss suivante*

$$I(f) = \int_{-1}^{1} f(x)\,dx \approx I_n(f) = w_0 f(x_0) + w_1 f(x_1)$$

Le but est de choisir les nœuds x_0 et x_1 et les poids w_0 et w_1 (soit 4 variables) de telle manière à ce que le degré de précision de la formule de quadrature soit le plus élevé possible. Pour résoudre ce problème, on a besoin de 4 équations (afin d'avoir une une solution unique pour les nœuds et les poids). Pour les obtenir, on impose que

$$I(f) = I_n(f) \qquad \text{pour } f(x) = x^k \qquad k = 0, \ldots, 3$$

1. *Établir le système de 4 équations non-linéaires pour obtenir les nœuds et les poids.*
2. *Vérifier que*

$$w_0 = w_1 = 1 \qquad x_0 = -\frac{\sqrt{3}}{3} \quad x_1 = \frac{\sqrt{3}}{3}$$

   est solution du système précédent

3. *Vérifier que la formule obtenue*

$$\int_{-1}^{1} f(x)\,dx \approx f\left(-\frac{\sqrt{3}}{3}\right) + f\left(\frac{\sqrt{3}}{3}\right)$$

   est de degré de précision 3.

1. On impose que

$$I(f) = I_n(f) \qquad \text{pour } f(x) = x^k \qquad k = 0, \ldots, 3$$

donc si $f(x) = 1$, on doit imposer que

$$I(f) = \int_{-1}^{1} 1\,dx = 2 = w_0 + w_1 = I_n(f)$$

Si $f(x) = x$, on doit imposer que

$$I(f) = \int_{-1}^{1} x\,dx = 0 = w_0 x_0 + w_1 x_1 = I_n(f)$$

Si $f(x) = x^2$, on doit imposer que

$$I(f) = \int_{-1}^{1} x^2\,dx = \frac{2}{3} = w_0 x_0^2 + w_1 x_1^2 = I_n(f)$$

Si $f(x) = x^3$, on doit imposer que

$$I(f) = \int_{-1}^{1} x^3\,dx = 0 = w_0 x_0^3 + w_1 x_1^3 = I_n(f)$$

Le système de 4 équations non-linéaires pour obtenir les nœuds et les poids est donc

$$\left\{\begin{array}{rcl} w_0 + w_1 & = & 2 \\ w_0 x_0 + w_1 x_1 & = & 0 \\ w_0 x_0^2 + w_1 x_1^2 & = & \dfrac{2}{3} \\ w_0 x_0^3 + w_1 x_1^3 & = & 0 \end{array}\right.$$

2. On suppose que

$$w_0 = w_1 = 1 \qquad x_0 = -\frac{\sqrt{3}}{3} \quad x_1 = \frac{\sqrt{3}}{3}$$

et on vérifie

$$w_0 + w_1 = 1 + 1 = 2$$

$$w_0 x_0 + w_1 x_1 = -\frac{\sqrt{3}}{3} + \frac{\sqrt{3}}{3} = 0$$

$$w_0 x_0^2 + w_1 x_1^2 = \left(-\frac{\sqrt{3}}{3}\right)^2 + \left(\frac{\sqrt{3}}{3}\right)^2 = \frac{3}{9} + \frac{3}{9} = \frac{2}{3}$$

$$w_0 x_0^3 + w_1 x_1^3 = \left(-\frac{\sqrt{3}}{3}\right)^3 + \left(\frac{\sqrt{3}}{3}\right)^3 = -\frac{3\sqrt{3}}{27} + \frac{3\sqrt{3}}{27} = 0$$

On a bien que

$$w_0 = w_1 = 1 \qquad x_0 = -\frac{\sqrt{3}}{3} \quad x_1 = \frac{\sqrt{3}}{3}$$

est solution du système établi à la question précédente.

3. Par construction,

$$I(f) = I_n(f) \qquad \text{pour } f(x) = x^k \qquad k = 0, \ldots, 3$$

la formule de quadrature a un degré de précision au moins égal à 3. Maintenant, si $f(x) = x^4$, on a

$$I(f) = \int_{-1}^{1} x^4 \, dx = \frac{2}{5} \qquad I_n(f) = w_0 f(x_0) + w_1 f(x_1) = \left(-\frac{\sqrt{3}}{3}\right)^4 + \left(\frac{\sqrt{3}}{3}\right)^4 = \frac{1}{9} + \frac{1}{9} = \frac{2}{9}$$

La formule de quadrature n'est pas exacte pour $f(x) = x^4$, son degré de précision est donc 3.

Chapitre 5 : Résolution de systèmes linéaires

Exercice 25 *Les systèmes suivants ont-ils une solution unique, une infinité de solution ou pas de solution ? Justifier.*

$$\left\{ \begin{array}{rcrcr} 2x_1 & + & 6x_2 & = & 4 \\ -4x_1 & - & 12x_2 & = & -8 \end{array} \right. \tag{1}$$

$$\left\{ \begin{array}{rcrcr} x_1 & + & 3x_2 & = & 7 \\ 2x_1 & - & x_2 & = & 0 \end{array} \right. \tag{2}$$

$$\left\{ \begin{array}{rcrcr} 4x_1 & - & 3x_2 & = & 14 \\ 16x_1 & - & 12x_2 & = & 2 \end{array} \right. \tag{3}$$

Le système (1) admet une infinité de solution. La seconde ligne du système peut être obtenue en multipliant la première ligne du système par -2 (que ce soit du côté gauche que du côté droit du signe =). Il n'y a donc qu'une seule équation pour relier les variables x_1 et x_2 d'où l'infinité de solutions.

Le système (2) admet une unique solution ($x_1 = 1$, $x_2 = 2$). La seconde ligne du système ne peut pas être obtenue en multipliant la première par un coefficient.

Le système (3) n'admet aucune solution. En effet, si on multiplie la première ligne par 4 et qu'on la soustrait à la seconde ligne, on obtient du côté gauche de l'égalité

$$16x_1 - 12x_2 - 4(4x_1 - 3x_2) = 0$$

et $2 - 4 \times 14 = -54$ pour le côté droit de l'égalité. Il y a donc incompatibilité entre les deux équations du système.

Exercice 26 *Quel est le rang de la matrice suivante ?*

$$\mathbf{A} = \begin{pmatrix} 1 & 4 & 6 & 5 \\ 2 & 7 & 3 & 1 \\ 4 & 10 & 7 & 2 \\ 0 & 5 & 8 & 9 \end{pmatrix}$$

Comme la matrice $\mathbf{A}$ est de dimension 4×4, le rang de la matrice est compris entre 0 et 4. Notons l_i, $i = 1, \ldots, 4$ les lignes de la matrice $\mathbf{A}$. Au moins une des lignes est non nulle, donc $\mathsf{rg}(\mathbf{A}) \geq 1$.

l_1 et l_2 sont linéairement indépendants. En effet, il n'existe pas de réel λ tel que $l_2 = \lambda l_1$, donc $\mathsf{rg}(\mathbf{A}) \geq 2$.

l_1, l_2 et l_4 sont linéairement indépendants. En effet, il n'existe pas de réels λ et μ tel que $l_4 = \lambda l_1 + \mu l_2$, donc $\mathsf{rg}(\mathbf{A}) \geq 3$.

$l_3 = 2l_1 + l_2 - l_4$, l_3 n'est pas linéairement indépendante de l_1, l_2 et l_4 d'où

$$\mathsf{rg}(\mathbf{A}) = 3$$

Exercice 27 *Soit le système linéaire à résoudre suivant*

$$\left\{ \begin{array}{rcrcr} 3x_1 & + & 2x_2 & = & 12 \\ 2x_1 & + & 3x_2 & = & 13 \\ 2x_1 & - & 2x_2 & = & -2 \end{array} \right.$$

On a l'égalité suivante

$$\begin{pmatrix} 3 & 2 \\ 2 & 3 \\ 2 & -2 \end{pmatrix} = \begin{pmatrix} -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{6} & -\frac{2}{3} \\ -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{6} & \frac{2}{3} \\ 0 & \frac{2\sqrt{2}}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 5 & 0 \\ 0 & 3 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix}$$

Vérifier que l'égalité précédente est une décomposition en valeurs singulières et l'utiliser pour montrer que $\begin{pmatrix} 2 \\ 3 \end{pmatrix}$ *est l'unique solution du système décrit plus haut.*

On considère comme acquis la validité de l'égalité précédente (on peut s'amuser à vérifier si on veut ...). On note les matrices

$$\mathbf{U} = \begin{pmatrix} -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{6} & -\frac{2}{3} \\ -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{6} & \frac{2}{3} \\ 0 & \frac{2\sqrt{2}}{3} & \frac{1}{3} \end{pmatrix} \qquad \mathbf{\Sigma} = \begin{pmatrix} 5 & 0 \\ 0 & 3 \\ 0 & 0 \end{pmatrix} \qquad \mathbf{V} = \begin{pmatrix} -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix}$$

On remarque tout d'abord que

$$\mathbf{U}\mathbf{U}^T = \begin{pmatrix} -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{6} & -\frac{2}{3} \\ -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{6} & \frac{2}{3} \\ 0 & \frac{2\sqrt{2}}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{2}}{6} & -\frac{\sqrt{2}}{6} & \frac{2\sqrt{2}}{3} \\ -\frac{2}{3} & \frac{2}{3} & \frac{1}{3} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\sqrt{2}}{2}\frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{6}\frac{\sqrt{2}}{6} + \frac{2}{3}\frac{2}{3} & \frac{\sqrt{2}}{2}\frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{6}\frac{\sqrt{2}}{6} - \frac{2}{3}\frac{2}{3} & \frac{2\sqrt{2}}{3}\frac{\sqrt{2}}{6} - \frac{1}{3}\frac{2}{3} \\ \frac{\sqrt{2}}{2}\frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{6}\frac{\sqrt{2}}{6} - \frac{2}{3}\frac{2}{3} & \frac{\sqrt{2}}{2}\frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{6}\frac{\sqrt{2}}{6} + \frac{2}{3}\frac{2}{3} & -\frac{2\sqrt{2}}{3}\frac{\sqrt{2}}{6} + \frac{1}{3}\frac{2}{3} \\ \frac{2\sqrt{2}}{3}\frac{\sqrt{2}}{6} - \frac{1}{3}\frac{2}{3} & -\frac{2\sqrt{2}}{3}\frac{\sqrt{2}}{6} + \frac{1}{3}\frac{2}{3} & \frac{2\sqrt{2}}{3}\frac{2\sqrt{2}}{3} + \frac{1}{3}\frac{1}{3} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$\mathbf{U}$ est une matrice carré donc puisque $\mathbf{U}\mathbf{U}^T = \mathbf{I}_3$ l'identité, on a aussi $\mathbf{U}^T\mathbf{U} = \mathbf{I}_3$ (attention ce n'est vrai que parce que le résultat de la multiplication matricielle est l'identité). $\mathbf{U}$ est bien une matrice orthogonale. De même pour $\mathbf{V}$, on a

$$\mathbf{V}\mathbf{V}^T = \begin{pmatrix} -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{2}}{2}\frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2}\frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{2}\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2}\frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{2}\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2}\frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}\frac{\sqrt{2}}{2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$\mathbf{V}$ est bien une matrice orthogonale et comme $\mathbf{\Sigma}$ est une matrice diagonale dont les termes diagonaux sont positifs, on a bien la décomposition en valeurs singulières de la matrice $\mathbf{A}$ du système linéaire $\mathbf{A}\mathbf{x} = \mathbf{b}$.

D'autre part, $\mathsf{rg}(\mathbf{A}) = 2$ et on cherche 2 variables, le système admet donc une unique solution que l'on va calculer grâce à la formule

$$\mathbf{x} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{b}$$

On effectue ce calcul en plusieurs étapes

$$\mathbf{U}^T\mathbf{b} = \begin{pmatrix} -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{2}}{6} & -\frac{\sqrt{2}}{6} & \frac{2\sqrt{2}}{3} \\ -\frac{2}{3} & \frac{2}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 12 \\ 13 \\ -2 \end{pmatrix} = \begin{pmatrix} -\frac{12\sqrt{2}}{2} - \frac{13\sqrt{2}}{2} \\ 12\frac{\sqrt{2}}{6} - 13\frac{\sqrt{2}}{6} - \frac{4\sqrt{2}}{3} \\ -\frac{24}{3} + \frac{26}{3} - \frac{2}{3} \end{pmatrix} = \begin{pmatrix} -\frac{25\sqrt{2}}{2} \\ -\frac{3\sqrt{2}}{2} \\ 0 \end{pmatrix}$$

$$\mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{b} = \begin{pmatrix} \frac{1}{5} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \end{pmatrix} \begin{pmatrix} -\frac{25\sqrt{2}}{2} \\ -\frac{3\sqrt{2}}{2} \\ 0 \end{pmatrix} = \begin{pmatrix} -\frac{5\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{pmatrix}$$

$$\mathbf{x} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{b} = \begin{pmatrix} -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} -\frac{5\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{pmatrix} = \begin{pmatrix} \frac{5(\sqrt{2})^2}{2\times 2} - \frac{(\sqrt{2})^2}{2\times 2} \\ \frac{5(\sqrt{2})^2}{2\times 2} + \frac{(\sqrt{2})^2}{2\times 2} \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$