Introduction à l'Analyse Numérique

Notes du cours de la deuxième année de bachelier en sciences mathématiques

Jean-Pierre Schneiders

Année 2013-2014

Introduction

L'analyse numérique s'attache à fournir des moyens pour donner des solutions approchées à des problèmes mathématiques dont la résolution explicite est généralement impossible ou impraticable. Les solutions approchées sont le plus souvent calculées sur ordinateur au moyen d'un algorithme convenable. Idéalement, elles devraient s'accompagner d'une majoration de l'écart avec les solutions réelles.

A titre d'exemple, supposons que nous disposions d'une machine sachant effectuer les opérations arithmétiques en virgule flottante et que nous devions implémenter la fonction trigonométrique $\sin x$. L'analyse nous apprend que

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

pour tout $x \in \mathbb{R}$. On peut donc penser à approcher $\sin x$ par le polynôme

$$S_N(x) = \sum_{n=1}^{N} (-1)^{n-1} \frac{x^{2n-1}}{(2n-1)!}.$$

Comme la série est alternée, l'erreur absolue commise est

$$|\sin x - S_N(x)| = |R_N(x)| \leq \frac{|x|^{2N+1}}{(2N+1)!}.$$

On peut donc adopter l'algorithme suivant : on pose $u_1 = x$, $v_1 = x$ et on calcule u_k et v_k de proche en proche par les formules

$$\begin{aligned} u_{k+1} &:= -\frac{x^2}{(2k)(2k+1)} u_k \\ v_{k+1} &:= v_k + u_{k+1} \end{aligned}$$

et on s'arrête lorsque $|u_{n+1}|$ est inférieur à la précision absolue souhaitée. La valeur approchée de $\sin x$ est alors v_n.

Cette approche bien que correcte du point de vue mathématique est un peu naïve du point de vue numérique car on n'a pas tenu compte de toutes les sources d'erreur qui apparaissent dans sa mise en œuvre. En général, celles-ci peuvent se classer en trois catégories :

(a) Erreurs sur les données

(b) Erreurs d'arrondi

(c) Erreurs d'approximation

Ci-dessus, on n'a tenu compte que des erreurs du type (c) qui sont celles qui apparaissent par le fait que l'on remplace le problème traité par un problème approché plus simple. Les erreurs du type (b) proviennent de ce que les nombres ne sont pas représentés en machine avec une précision infinie et du fait que les opérations arithmétiques ne sont pas non plus effectuées avec une précision infinie. Les erreurs sur les données peuvent provenir par exemple d'erreurs de mesure. Dans notre exemple, il n'y en a pas. En général, les erreurs de ce type sont hors du contrôle du calcul mais il est néanmoins intéressant de voir en quoi la méthode de résolution y est sensible. C'est ce qu'on appelle l'étude de la stabilité de la méthode de résolution choisie.

1 Calculs sur ordinateur

Dans ce chapitre, β désignera un naturel supérieur ou égal à 2 qui servira de base de numération.

1.1 Représentation des entiers en base β

Soit a un entier naturel. Il est clair que a peut s'écrire de manière unique sous la forme

$$a = \sum_{k \in \mathbb{N}} d_k \beta^k$$

avec $d_k \in \{0, \ldots, \beta - 1\}$ pour tout $k \in \mathbb{N}$ et $d_k = 0$ pour k suffisamment grand.

Le coefficient d_k figurant dans l'égalité ci-dessus est appelé *chiffre de rang k de a en base β*. La *représentation de a en base β* est quant à elle la liste

$$d_q \ldots d_0$$

où q est le plus grand k tel que $d_k \neq 0$ si $a \neq 0$ ou la liste

$$0$$

si $a = 0$. Pour éviter tout risque de confusion sur la base utilisée, nous conviendrons de noter

$$(d_q \ldots d_0)_\beta$$

l'entier dont la représentation en base β est

$$d_q \ldots d_0.$$

En représentation décimale, nous laisserons tomber les parenthèses et l'indice β pour nous conformer à l'usage courant. Parmi les autres représentations fréquemment utilisées, citons

(a) la représentation binaire pour laquelle $\beta = 2$ et $d_k \in \{0, 1\}$;

(b) la représentation octale pour laquelle $\beta = 8$ et $d_k \in \{0, 1, \ldots, 7\}$;

(c) la représentation hexadécimale pour laquelle $\beta = 16$ et $d_k \in \{0, 1, \ldots, 15\}$.

Dans la suite, les chiffres binaires seront souvent appelés des *bits*[1] et pour éviter des ambiguïtés de notation, nous utiliserons les lettres A, B, …, F pour représenter les chiffres hexadécimaux 10, 11, …, 15.

1. Par contraction de l'expression anglaise *binary digits*.

Exemple 1.1.1. On a trivialement $0 = (0)_2 = (0)_8 = (0)_{16}$ et on vérifie aisément que

$$27 = (11011)_2 = (33)_8 = (1B)_{16}.$$

L'exemple précédent montre en particulier que le nombre des chiffres figurant dans la représentation d'un naturel a en base β dépend de cette base.

On vérifie de suite que la représentation en base β de $a \in \mathbb{N}_0$ comporte exactement n chiffres si et seulement si

$$(1)\beta^{n-1} + (0)\beta^{n-2} + \cdots + (0)1 \leq a \leq (\beta-1)\beta^{n-1} + (\beta-1)\beta^{n-2} + \cdots + (\beta-1)1$$

c'est à dire si $\beta^{n-1} \leq a \leq \beta^n - 1$. Il s'ensuit que les naturels dont la représentation en base β comporte au plus n chiffres sont ceux appartenant à l'intervalle $[0, \beta^n - 1]$.

Lorsque a est un entier strictement négatif, nous définirons les chiffres $(d_k)_{k\in\mathbb{N}}$ de a en base β comme étant ceux de $|a| = -a$ et la représentation de a en base β comme comme étant celle $|a|$ en base β précédée du signe $-$.

1.2 Codage des entiers en machine

Sur les ordinateurs récents, on code le plus souvent les entiers en utilisant leur représentation binaire.

Supposons disposer de n bits pour le stockage de nos entiers. Si nous ne souhaitons stocker que des entiers naturels, il résulte de ce qui précède que nous ne pouvons représenter exactement que les naturels inférieurs ou égaux à $2^n - 1$. Par exemple, avec un octet (resp. deux octets, quatre octets) de 8 bits, nous pouvons stocker les naturels inférieurs ou égaux à 255 (resp. 65535, 4294967295).

Si nous souhaitons utiliser nos n bits pour stocker des entiers relatifs, nous devons utiliser un procédé de codage un peu plus évolué. Les procédés les plus courants sont :

(a) le codage "signe, valeur absolue" ;

(b) le codage "complément à deux" ;

(c) le codage "biaisé".

Dans le cas (a), on réserve un bit pour coder le signe (le plus souvent 0 pour +, 1 pour $-$) et on utilise les $(n-1)$ autres bits pour stocker la représentation binaire de la valeur absolue. Les entiers représentables de la sorte sont donc ceux appartenant à l'intervalle $]-2^{n-1}, 2^{n-1}[$ et 0 a deux codages possibles.

Dans le cas (b), on stocke les n derniers chiffres binaires de a si $0 \leq a < 2^{n-1}$ et les n derniers chiffres binaires de $2^n + a$ si $-2^{n-1} \leq a < 0$. Par construction, le premier bit vaut alors 0 si $a \geq 0$ et 1 si $a < 0$.

Dans le cas (c), on choisit un biais entier et on représente un entier

$$a \in [-\text{biais}, 2^n - \text{biais}[$$

par les n derniers chiffres binaires de $a + \text{biais}$.

L'arithmétique entière étant particulièrement simple à implémenter dans le cas (b), c'est souvent ce codage qui est utilisé en pratique. Nous rencontrerons cependant les autres codages dans le stockage de l'exposant et de la mantisse des nombres réels.

Exemple 1.2.1. Prenons $n = 8$, $a = 10$ et biais $= 128$. On a alors la table suivante :

	$a_{\min}$	$-a$	a	$a_{\max}$
(a)	-127	10001010	00001010	127
(b)	-128	11110110	00001010	127
(c)	-128	01110110	10001010	127

où la colonne $a_{\min}$ (resp. $a_{\max}$) désigne l'entier codable minimum (resp. maximum) pour le procédé choisi et où les colonnes $-a$ et a donnent les codages de $-a$ et de a pour ce même procédé.

1.3 Les entiers machine en C

D'un point de vue pratique, ce qu'il importe surtout de savoir au sujet du format des entiers machine est la valeur de l'entier machine minimum et celle de l'entier machine maximum. En C, ces valeurs sont disponibles pour chacun des types entiers sous la forme de constantes prédéclarées dans le fichier `<limits.h>`. Les plus courantes sont précisées dans le tableau suivant :

Type	Min	Max
`char`	`CHAR_MIN`	`CHAR_MAX`
`short`	`SHRT_MIN`	`SHRT_MAX`
`int`	`INT_MIN`	`INT_MAX`
`long`	`LONG_MIN`	`LONG_MAX`

Le nombre de bits stockables dans une variable de type `char` étant quant à lui donné par la constante `CHAR_BIT` et valant le plus souvent 8.

En C, on peut se représenter la mémoire de l'ordinateur comme un ensemble de cases numérotées consécutivement, chaque case ayant juste la taille nécessaire pour stocker une variable de type `char`. Le numéro attribué à une case est son *adresse mémoire*. La valeur d'une variable x de type t est stockée en mémoire dans une suite de cases consécutives et le nombre de cases utilisées est donné par les expressions

`sizeof(x)` ou `sizeof(t)`.

En utilisant ces expressions, il est donc possible de déterminer combien de cases mémoire sont nécessaires pour stocker un entier de type `short`, `int` ou `long`. Le plus souvent les entiers de type `short` sont stockés sur deux cases et ceux de type `long` sur quatre cases. Les entiers de type `int` sont quant à eux les entiers machine les plus "naturels" et correspondent souvent soit aux entiers de type `short` soit aux entiers de type `long`.

1.4 Représentation des réels en base β

Soit a un réel positif ou nul. En généralisant ce qui a été fait pour les entiers, on peut montrer que a peut s'écrire de manière unique sous la forme

$$a = \sum_{k\in\mathbb{Z}} d_k \beta^k$$

où $(d_k)_{k\in\mathbb{Z}}$ est une famille d'entiers compris entre 0 et $\beta - 1$ pour laquelle

— il existe un $r \in \mathbb{Z}$ tel que $d_k = 0$ pour tout $k > r$,

— il n'existe pas de $s \in \mathbb{Z}$ tel que $d_k = \beta - 1$ pour tout $k < s$.

Le coefficient d_k figurant dans l'égalité ci-dessus est le *chiffre de rang k de a en base β*. Les chiffres de rang positif ou nul sont appelés *chiffres entiers*, ceux de rang négatif sont appelés *chiffres fractionnaires*.

La *représentation de a en base β* est quant à elle la liste

$$d_q \dots d_0 . d_{-1} \dots d_{-p} \dots$$

où q est le plus grand $k \geq 0$ tel que $d_k \neq 0$ ou 0 si un tel k n'existe pas.

Pour éviter tout risque de confusion sur la base utilisée, on conviendra de noter

$$(d_q \dots d_0 . d_{-1} \dots d_{-p} \dots)_\beta$$

le réel dont la représentation en base β est

$$d_q \dots d_0 . d_{-1} \dots d_{-p} \dots .$$

En représentation décimale, on laissera tomber les parenthèses et l'indice β pour se conformer à l'usage courant.

Exemple 1.4.1. On a

$$(1011.1)_2 = 11.5 \qquad (A0.F)_{16} = 160.9375.$$

Lorsque a est un réel strictement négatif, on définira les chiffres de a en base β comme étant ceux de $|a| = -a$ et la représentation de a en base β comme étant celle de $|a|$ précédée du signe $-$.

Soit a un réel non nul et soit d_k son chiffre de rang k en base β. Comme $a \neq 0$, il résulte de ce qui précède qu'il existe un plus petit entier entier e tel que $d_k = 0$ si $k \geq e$. Cet entier e est l'*exposant normalisé de a en base β* et les chiffres de rang $k < e$ sont les *chiffres significatifs de a en base β*. Comme e est aussi caractérisé par la double inégalité

$$\beta^{e-1} \leq |a| < \beta^e,$$

il est clair que

$$m = |a|\beta^{-e} \in \left[\beta^{-1}, 1\right[.$$

On dit que m est la *mantisse normalisée de a en base β*. Par construction, on a bien sûr

$$m = \sum_{k<e} d_k \beta^{k-e} = (0.d_{e-1}d_{e-2}\ldots)_\beta$$

et

$$a = \mathrm{sgn}(a) m \beta^e.$$

En particulier, le réel a est déterminé par son signe, sa mantisse normalisée et son exposant normalisé.

Vu ce qui a été fait pour les entiers, il est naturel d'essayer de stocker un réel en machine en mémorisant :

— soit sa représentation binaire ;

— soit son signe, ses chiffres significatifs binaires et son exposant normalisé binaire.

Malheureusement, dans un cas comme dans l'autre, les données à mémoriser peuvent requérir un nombre infini de bits. Il est donc indispensable de se limiter au stockage d'une valeur approchée codable en machine au moyen d'un nombre fini de bits.

1.5 Valeurs approchées d'un nombre réel

On appelle *valeur approchée d'un nombre réel a* tout nombre réel $\tilde{a}$ voisin de a.

L'*erreur absolue* associée à la valeur approchée $\tilde{a}$ du réel a est par définition le réel

$$\Delta a = \tilde{a} - a.$$

L'*erreur relative* associée à la valeur approchée $\tilde{a}$ du réel $a \neq 0$ est par définition le réel

$$\delta a = \frac{\tilde{a} - a}{a}.$$

Il résulte bien-sûr de ces définitions que l'on a alors

$$\tilde{a} = a + \Delta a \quad \text{et} \quad \tilde{a} = a(1 + \delta a).$$

Dans la suite, nous utiliserons parfois la notation abrégée

$$a \approx \tilde{a}$$

pour indiquer que $\tilde{a}$ est une valeur approchée de a et la notation abrégée

$$a = \tilde{a} \pm \varepsilon$$

pour indiquer que l'erreur absolue associée à la valeur approchée $\tilde{a}$ de a est de module inférieur à ε.

Exemple 1.5.1. Si $a = 1.25$ et si $\tilde{a} = 1.2$, alors $a \approx \tilde{a}$ et on a

$$\Delta a = -0.05 \quad \text{et} \quad \delta a = -0.04.$$

On en tire que

$$a = 1.2 \pm 510^{-2}$$

et que le module de l'erreur relative est de 4%.

1.6 Arrondis à p chiffres fractionnaires en base β

Soit a un nombre réel arbitraire. Appelons réel à p chiffres fractionnaires en base β tout réel dont seuls les p premiers chiffres fractionnaires en base β sont éventuellement non nuls et cherchons à associer à a une valeur approchée $\tilde{a}$ à p chiffres fractionnaires en base β qui soit en un certain sens optimale.

Les notions d'optimalité les plus naturelles correspondent à chercher $\tilde{a}$ tel que l'une des conditions suivantes ait lieu :

(i) Δa est positif ou nul et minimum ;

(ii) Δa est négatif ou nul et maximum ;

(iii) $|\Delta a|$ est minimum.

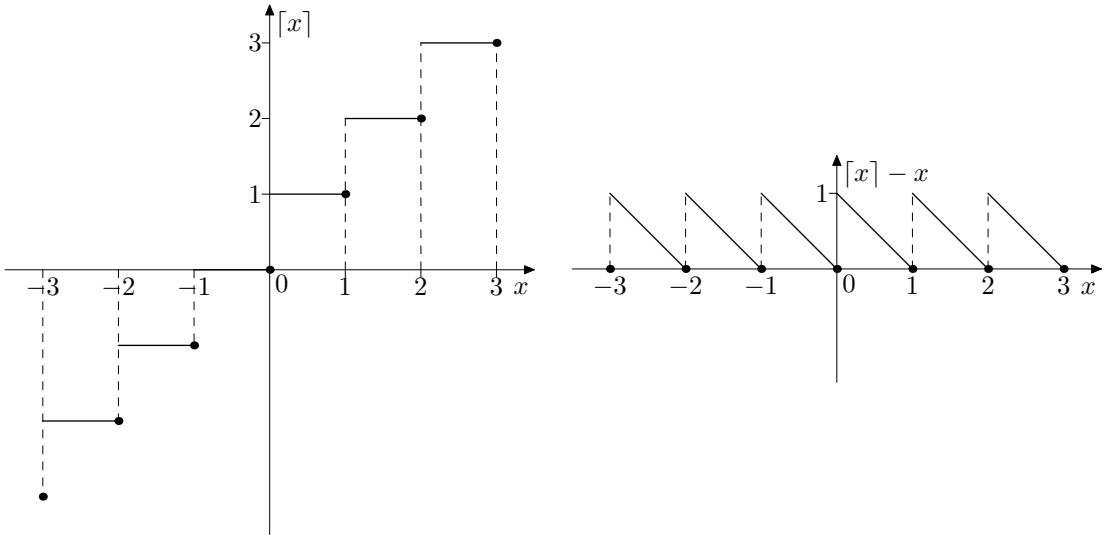
Dans le cas (i) on cherche donc un $\tilde{a} \geq a$ à p chiffres fractionnaires en base β pour lequel $\tilde{a}$ est minimum. Un tel $\tilde{a}$ est caractérisé par le fait que $\tilde{x} = \tilde{a}\beta^p$ est le plus petit entier qui majore $x = a\beta^p$. Il s'ensuit que $\tilde{a}$ doit être choisi de sorte que $\tilde{x}$ soit le plafond $\lceil x \rceil$ de x. En d'autres termes, on doit prendre

$$\tilde{a} = \lceil a\beta^p \rceil \beta^{-p}.$$

Cette valeur approchée de a est par définition l'*arrondi par excès de a à p chiffres fractionnaires en base β*. Vu ce qui précède, il est clair que l'on a

$$0 \leq \Delta a < \beta^{-p}.$$

De plus, les graphes



montrent que cet encadrement global est le meilleur possible.

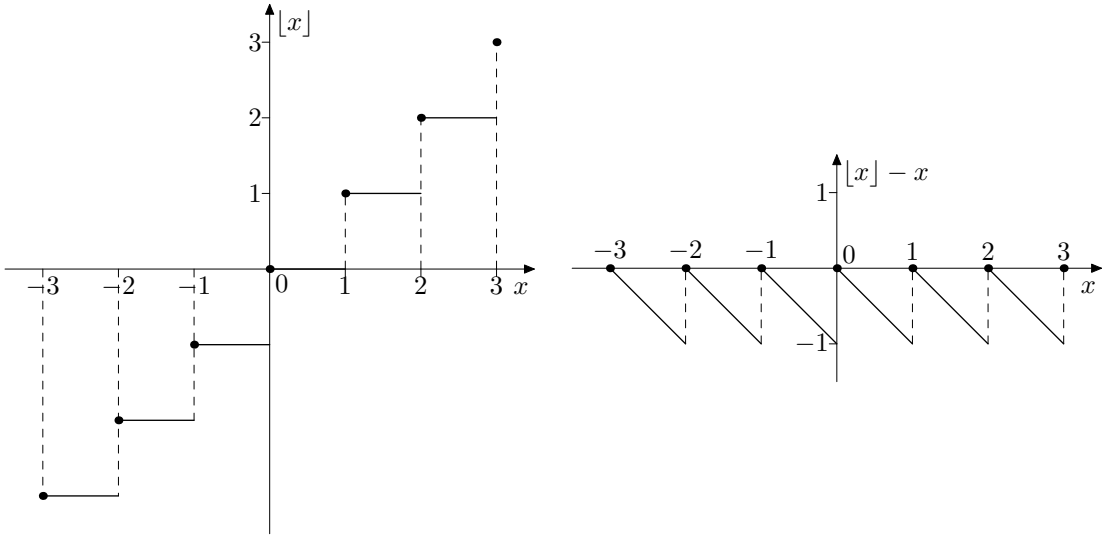
Dans le cas (ii) on cherche un $\tilde{a} \leq a$ à p chiffres fractionnaires en base β pour lequel $\tilde{a}$ est maximum. Un tel $\tilde{a}$ est caractérisé par le fait que $\tilde{x} = \tilde{a}\beta^p$ est le plus grand entier qui minore $x = a\beta^p$. Il s'ensuit que $\tilde{a}$ doit être choisi de sorte que $\tilde{x}$ soit le plancher $\lfloor x \rfloor$ de x. En d'autres termes, on doit prendre

$$\tilde{a} = \lfloor a\beta^p \rfloor \beta^{-p}.$$

Cette valeur approchée de a est par définition l'*arrondi par défaut de a à p chiffres fractionnaires en base β*. Vu ce qui précède, il est clair que l'on a

$$-\beta^{-p} < \Delta a \leq 0.$$

De plus, les graphes



montrent que cet encadrement global est le meilleur possible.

Passons à présent au cas (iii) et cherchons un $\tilde{a}$ à p chiffres fractionnaires pour lequel $|\tilde{a} - a|$ soit minimum. Comme un tel $\tilde{a}$ peut s'écrire de manière unique sous la forme

$$\tilde{a} = \tilde{x}\beta^{-p}$$

avec $\tilde{x} \in \mathbb{Z}$, tout revient à trouver un entier $\tilde{x}$ dont la distance au réel $x = a\beta^p$ soit minimum.

(a) Si un réel x n'est pas de la forme $n + \frac{1}{2}$ avec $n \in \mathbb{Z}$ il existe un et un seul entier $\tilde{x} \in \mathbb{Z}$ rendant $|\tilde{x} - x|$ minimum et cet entier est donné par

$$\tilde{x} = \begin{cases} \lfloor x \rfloor & \text{si } x - \lfloor x \rfloor < 1/2; \\ \lfloor x \rfloor + 1 & \text{si } x - \lfloor x \rfloor > 1/2. \end{cases}$$

Dans le premier cas, on a

$$-\frac{1}{2} < \tilde{x} - x \leq 0$$

et dans le second, on a

$$0 \leq \tilde{x} - x < \frac{1}{2}.$$

(b) Par contre, si le réel x est de la forme $n + 1/2$ avec $n \in \mathbb{Z}$, il existe deux $\tilde{x} \in \mathbb{Z}$ rendant $|\tilde{x} - x|$ minimum à savoir n et $n + 1$ et on a

$$\tilde{x} - x = -\frac{1}{2}$$

dans le premier cas et

$$\tilde{x} - x = \frac{1}{2}$$

dans le second.

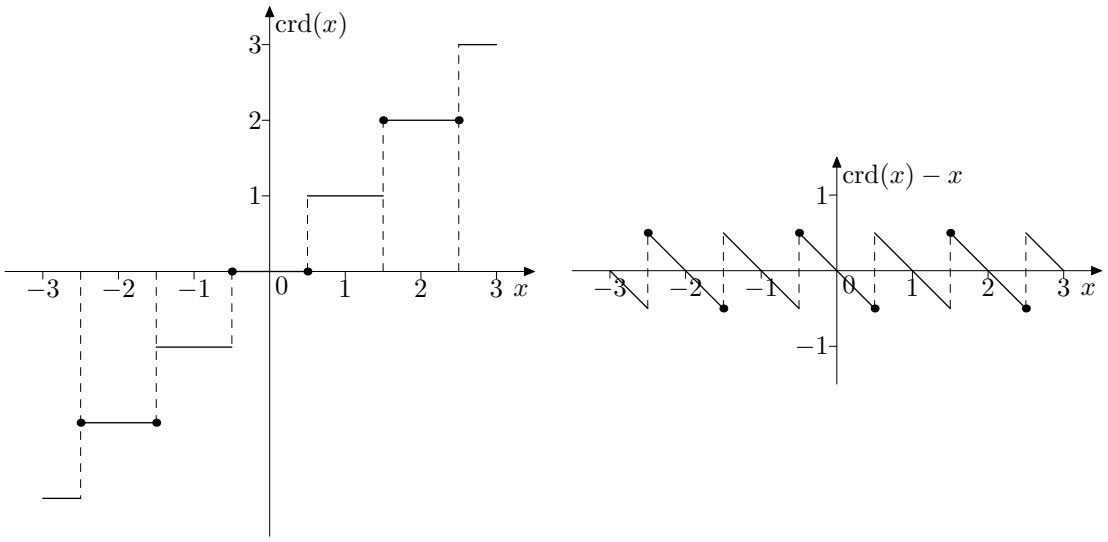
Pour que l'erreur d'arrondi soit aussi souvent positive que négative, nous conviendrons de définir l'*arrondi correct*[2] $\mathrm{crd}(x)$ de x comme étant égal à l'unique entier $\tilde{x}$ optimum dans le cas (a) et a l'unique entier $\tilde{x}$ optimum et pair dans le cas (b). Le $\tilde{a}$ correspondant est alors appelé *arrondi correct de a à p chiffres fractionnaires en base β* et est donné par

$$\tilde{a} = \mathrm{crd}(a\beta^p)\beta^{-p}.$$

Vu ce qui précède, on a bien sûr

$$0 \leq |\Delta a| \leq \frac{1}{2}\beta^{-p}.$$

De plus, les graphes



montrent que cet encadrement global et le meilleur possible.

En pratique, nous aurons surtout à calculer des arrondis corrects à p chiffres fractionnaires dans une base β divisible par deux. Dans un tel cas, il résulte de ce qui précède que l'arrondi correct de

$$a = (d_q \ldots d_0.d_{-1} \ldots d_{-p}d_{-p-1} \ldots)_\beta$$

est simplement

$$\tilde{a} = (d_q \ldots d_0.d_{-1} \ldots d_{-p})_\beta$$

2. Par opposition à l'arrondi simple $\mathrm{rd}(x)$ que l'on obtient en choisissant toujours le $\tilde{x}$ de plus grande valeur absolue dans le cas (b).

lorsque $d_{-p-1} < \beta/2$ ou lorsque d_{-p} est pair, $d_{-p-1} = \beta/2$, $d_{-p-2} = 0$, ... et

$$\tilde{a} = (d_q \dots d_0.d_{-1} \dots d_{-p})_\beta + \beta^{-p}$$

dans les autres cas.

Par construction, la fonction $\mathrm{crd}(x)$ est impaire et on a donc

$$\mathrm{crd}(x) = \mathrm{sgn}(x)\,\mathrm{crd}(|x|).$$

Il n'en est cependant pas de même des fonctions $\lceil x \rceil$ et $\lfloor x \rfloor$; c'est pourquoi il peut être également intéressant de disposer de la fonction de troncature

$$[x] = \mathrm{sgn}(x)\lfloor |x| \rfloor.$$

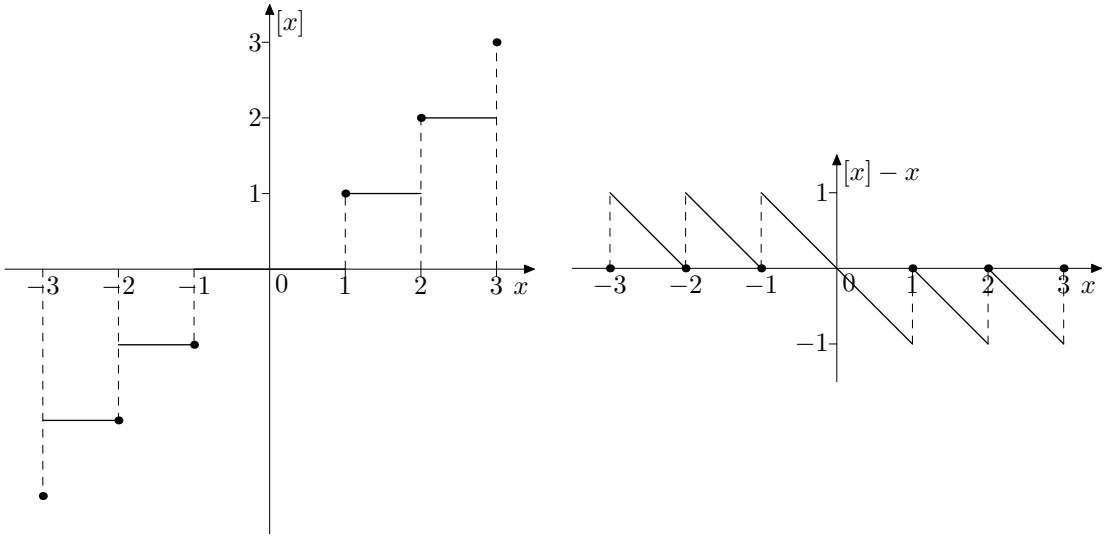
A tout réel a on peut donc aussi associer la valeur approchée

$$\tilde{a} = [a\beta^p]\beta^{-p}.$$

Cette valeur approchée est le *tronqué de a à p chiffres fractionnaires en base* β. Par construction, on a bien sûr

$$0 \leq |\Delta a| \leq \beta^{-p}$$

et les graphes



montrent que cette majoration globale est la meilleure possible. Il en résulte en particulier que le procédé de troncature est globalement moins précis que les divers procédés d'arrondi ; il est cependant plus simple à implémenter puisque le tronqué de

$$a = (d_q \dots d_0.d_{-1} \dots d_{-p}d_{-p-1} \dots)_\beta$$

à p chiffres fractionnaires en base β est simplement

$$\tilde{a} = (d_q \dots d_0.d_{-1} \dots d_{-p})_\beta.$$

1.7 Arrondis à p chiffres significatifs en base β

Soit a un réel non nul. Appelons réel à p chiffres significatifs en base β tout réel dont seuls les p premiers chiffres significatifs en base β sont éventuellement non nuls et cherchons à associer à a une valeur approchée à p chiffres significatifs en base β.

On sait que a peut s'écrire d'une et une seule manière sous la forme

$$a = \pm m\beta^e$$

avec $m \in [1/\beta, 1[$, $e \in \mathbb{Z}$. La valeur approchée

$$\tilde{a} = \pm \tilde{m}\beta^e$$

de a obtenue en remplaçant m par son arrondi correct (resp. son arrondi par excès, son arrondi par défaut, son tronqué) $\tilde{m}$ à p chiffres fractionnaires en base β sera appelé l'*arrondi correct (resp. l'arrondi par excès, l'arrondi par défaut, le tronqué) de a à p chiffres significatifs en base β*. Vu ce qui précède, on a respectivement

(i) $|\Delta a| \leq \frac{1}{2}\beta^{e-p}$;

(ii) $0 \leq \Delta a < \beta^{e-p}$;

(iii) $-\beta^{e-p} < \Delta a \leq 0$;

(iv) $|\Delta a| \leq \beta^{e-p}$.

Comme $m \geq 1/\beta$, on a donc aussi respectivement

(i) $|\delta a| \leq \dfrac{1}{2}\beta^{1-p}$;

(ii) $|\delta a| \leq \beta^{1-p}$ et $a\delta a \geq 0$;

(iii) $|\delta a| \leq \beta^{1-p}$ et $a\delta a \leq 0$;

(iv) $|\delta a| \leq \beta^{1-p}$.

Il résulte de ces majorations que l'erreur relative commise en remplaçant un réel par son arrondi correct (resp. par défaut, par excès) varie dans un intervalle de longueur β^{1-p}. Le nombre β^{1-p} est aussi l'écart entre 1 et le plus petit réel à p chiffres significatifs qui soit strictement plus grand que 1 et est souvent noté simplement eps lorsque β et p sont clairs par le contexte.

1.8 Chiffres exacts et précision d'une valeur approchée

Dans la suite, nous dirons que le chiffre de rang k d'une valeur approchée $\tilde{a}$ de a est *exact (resp. exact au sens large)* si on a

$$|\Delta a| \leq \frac{1}{2}\beta^k \qquad (\text{resp. } |\Delta a| \leq \beta^k \ ).$$

En particulier, nous dirons que $\tilde{a}$ est une valeur approchée de a dont les p premiers chiffres fractionnaires en base β sont exacts (resp. exacts au sens large) si on a

$$|\Delta a| \leq \frac{1}{2}\beta^{-p} \qquad (\text{resp. } |\Delta a| \leq \beta^{-p} \ ).$$

comme pour la valeur approchée de a à p chiffres fractionnaires en base β obtenue par arrondi correct (resp. par troncature).

Il est clair que la fonction qui associe à un réel x son arrondi correct (resp. son arrondi par défaut, son arrondi par excès, son tronqué) est discontinue en tout $x \in \mathbb{Z} + \frac{1}{2}$ (resp. $\mathbb{Z}$, $\mathbb{Z}$, $\mathbb{Z} \setminus \{0\}$). De plus, on vérifie aisément que si a, b sont des réels tels que $|a-b| < \beta^{-p}$ alors les arrondis corrects à p chiffres fractionnaires $\tilde{a}$ et $\tilde{b}$ de a et de b sont tels que $|\tilde{a} - \tilde{b}| \leq \beta^{-p}$ et que si $\varepsilon \in]0, \beta^{-p}[$ alors il n'existe aucun $\eta > 0$ tel que

$$|a-b| < \eta \Rightarrow |\tilde{a} - \tilde{b}| \leq \varepsilon.$$

Il est donc possible de déterminer une valeur approchée d'un réel a à p chiffres fractionnaires exacts au sens large à partir d'une suite a_m qui converge vers a. Mais il est en général impossible d'obtenir de la sorte une valeur approchée à p chiffres fractionnaires exacts [3].

Soit $\tilde{a}$ une valeur approchée de a dont les p premiers chiffres fractionnaires sont exacts mais dont le chiffre de rang $p+1$ n'est pas exact. On a alors

$$\frac{1}{2}\beta^{-p-1} < |\Delta a| \leq \frac{1}{2}\beta^{-p}$$

et

$$p \leq -\log_\beta 2|\Delta a| < p+1$$

d'où l'on tire que

$$p = \lfloor -\log_\beta 2|\Delta a| \rfloor.$$

Cette formule nous conduit à définir la *précision absolue de $\tilde{a}$ en base β* comme étant

$$p_a(a) = -\log_\beta 2|\Delta a|.$$

3. Ce phénomène est lié au *Tablemaker's Dilemma* de la littérature anglo-saxonne.

Par construction, on a

$$|\Delta a| = \frac{1}{2}\beta^{-p_a(a)}.$$

Par analogie, nous définirons la *précision relative de $\tilde{a}$ en base β* comme étant

$$p_r(a) = -\log_\beta 2|\delta a|.$$

De cette définition, on tire que

$$|\delta a| = \frac{1}{2}\beta^{-p_r(a)}$$

et que

$$\frac{1}{2}\beta^{e-p_r(a)-1} \leq |\Delta a| < \frac{1}{2}\beta^{e-p_r(a)}$$

si e est l'exposant normalisé de a en base β. En particulier, les $\lfloor p_r(a) \rfloor$ premiers chiffres significatifs de $\tilde{a}$ en base β sont exacts et le $(\lceil p_r(a) \rceil + 2)$-ème chiffre significatif de $\tilde{a}$ en base β est inexact.

1.9 Codage des réels en machine

Dans la majorité des ordinateurs récents, les réels sont représentés au moyen d'une valeur approchée en base $\beta = 2$ dont le stockage ne requiert qu'un nombre fixe de bits. Deux méthodes sont principalement utilisées.

1.9.1 Virgule fixe

Dans ce cas, on décide de ne conserver en machine qu'un nombre fixe p de bits fractionnaires. Comme le problème du codage d'un réel a ayant moins de p bits fractionnaires non nuls est équivalent à celui du codage de l'entier $a2^p$, on peut mettre en oeuvre un des procédés détaillés dans la section 1.2.

Pour le codage "signe, valeur absolue", les réels codables exactement en machine sont les nombres de la forme

$$\pm(d_q \dots d_0.d_{-1} \dots d_{-p})_2.$$

où $q = n - 1 - p$.

Cela étant, pour coder un réel quelconque a en machine, on doit donc lui associer une valeur approchée du type précédent. Si $|a| \geq 2^{q+1}$, ce n'est pas possible et on a un cas de dépassement de capacité vers le haut (overflow). Sinon, on peut représenter a par la valeur approchée que l'on obtient en arrondissant correctement a à p bits fractionnaires. L'erreur absolue est alors bornée par 2^{-p-1}.

Les principaux avantages de la représentation en virgule fixe sont de ramener les calculs sur des réels à des calculs sur des entiers et d'assurer une borne uniforme sur l'erreur absolue de codage.

Ses inconvénients essentiels sont de ne permettre la représentation approchée que d'un petit intervalle de réels et ce en n'assurant pas un bon contrôle de l'erreur relative.

1.9.2 Virgule flottante

Dans ce cas, on décide de ne conserver en machine qu'un nombre fixe p de bits significatifs.

Comme le problème du codage d'un réel a ayant moins de p bits significatifs non nuls est équivalent à celui du codage de l'exposant normalisé e et du *significande* $a2^{p-e}$, on est de nouveau ramené à un problème de codage de nombres entiers.

Si on choisit le codage "signe, valeur absolue" pour l'exposant normalisé et le significande, les nombres réels codables exactement en machine sont 0 et les nombres de la forme

$$\pm(0.a_1 \dots a_p)_2 2^e \qquad (a_1 = 1)$$

avec e de la forme

$$\pm(e_q \dots e_0)_2$$

et $q = n - 3 - p$.

Cela étant, pour coder un réel non-nul quelconque a en machine, on doit donc lui associer une valeur approchée du type précédent. Si l'exposant normalisé e de a est tel que $|e| \geq 2^{q+1}$, c'est impossible et on a un dépassement de capacité vers le haut si $e > 0$ ou vers les bas si $e < 0$. Sinon, on peut représenter a par la valeur approchée que l'on obtient en arrondissant correctement a à p chiffres significatifs. L'erreur relative est alors bornée par le nombre

$$u = 2^{-p}$$

souvent appelé *unité machine*.

L'inconvénient majeur du codage en virgule flottante est de nécessiter une implémentation spécifique des opérations arithmétiques. Si une telle implémentation doit être faite sous forme logicielle, elle conduira presque toujours à des calculs plus lents que ceux effectués en virgule fixe. Son avantage principal est de conduire à une borne uniforme pour l'erreur relative de codage.

Comme la plupart des ordinateurs récents implémentent l'arithmétique virgule flottante sous forme matérielle (coprocesseur arithmétique ou jeu d'instructions dédiées) il n'y a plus généralement de problème sensible au niveau de la vitesse.

C'est pourquoi c'est actuellement le codage en virgule flottante qui est presque universellement adopté.

1.10 Les réels machine en C

Le langage C dispose de plusieurs types de réels machine. Les plus courants étant les types `float` et `double`. Ces types correspondent à des codages en virgule flottante qui suivent les principes exposés ci-dessus. Pour utiliser correctement ces types en pratique, il est suffisant de connaître la base utilisée, le nombre p de chiffres significatifs utilisés pour le stockage de la mantisse normalisée et les valeurs extrêmes ($e_{\min}$,$e_{\max}$) que peut prendre l'exposant normalisé sans donner lieu à des problèmes de dépassement de capacité. Ces informations figurent dans le fichier `<float.h>` sous la forme de constantes prédéfinies. La base utilisée est donnée par la constante `FLT_RADIX` et les paramètres des types flottants usuels sont donnés par les constantes suivantes :

Type	p	$e_{\min}$	$e_{\max}$
`float`	`FLT_MANT_DIG`	`FLT_MIN_EXP`	`FLT_MAX_EXP`
`double`	`DBL_MANT_DIG`	`DBL_MIN_EXP`	`DBL_MAX_EXP`

Le codage réellement utilisé en machine pour les types flottants n'est pas précisé dans la spécification du C. Cependant, dans de nombreux cas le schéma suivi est celui préconisé par le standard IEEE 754 de 1985. Dans ce cas, un nombre machine normalisé de type `float` (resp. `double`) est codé sur 32 (resp. 64) bits avec 1 bit pour le signe, 8 (resp. 11) bits pour l'exposant et 23 bits pour stocker les bits de la mantisse à l'exception du premier (qui est toujours égal à 1). Le codage utilisé pour la mantisse est donc du type "signe, valeur absolue". Celui utilisé pour l'exposant fait quant à lui usage d'un biais égal à 126 (resp. 1022). Cependant, on a seulement $e_{\min} = -125$ et $e_{\max} = 128$ (resp. $e_{\min} = -1021$ et $e_{\max} = 1024$) car un code d'exposant égal à 0 est utilisé pour ± 0 et pour les réels dénormalisés alors qu'un code d'exposant égal à 255 (resp. 2047) est utilisé pour la représentation de $\pm\infty$ et des NaN [4] engendrés par certains types d'erreurs.

Exemple 1.10.1. Le codage de 12.25 en tant que `float` au format IEEE 754 est

$$0\ 10000010\ 10001000000000000000000$$

car

$$12.25 = (1100.01)_2 = (0.110001)_2\, 2^4$$

et

$$4 + 126 = 130 = (10000010)_2.$$

4. Contraction de l'expression anglaise *Not a Number*.

1.11 Arithmétique flottante correctement arrondie

Dans la suite, nous supposerons que la machine utilisée travaille avec un codage des réels en virgule flottante à p chiffres significatifs en base β du type discuté ci-dessus et nous noterons $\mathrm{fl}(x)$ le réel machine associé au réel x.

Vu ce qui précède, il est clair que l'ensemble des réels représentables exactement en machine n'est pas clos pour les opérations arithmétiques élémentaires et le mieux que l'on puisse exiger est que les opérations arithmétiques soient implémentées en machine de sorte que

$$x\,\mathrm{fl}(\mathrm{op})y = \mathrm{fl}(x \text{ op } y)$$

si x et y sont de réels représentables exactement en machine et si $\mathrm{fl}(\mathrm{op})$ est l'opération machine associée à $\mathrm{op} \in \{+, -, \cdot, /\}$ et si cette opération n'engendre pas un dépassement de capacité. Comme un tel dépassement signale généralement que les calculs ultérieurs ne donneront pas de résultat exploitable (et demande donc d'arrêter le travail en cours et de modifier l'approche utilisée), nous supposerons ici que nous sommes dans la situation idéale où $e_{\min} = -\infty$ et $e_{\max} = +\infty$. Dans ce cas, l'équation précédente entraîne alors que pour tout couple (x, y) de réels représentables exactement en machine, il existe un réel δ tel que

$$x\,\mathrm{fl}(\mathrm{op})y = (x \text{ op } y)(1+\delta), \qquad |\delta| \leq u.$$

Remarquons que même dans cette situation idéale les opérations machine ne jouissent pas des propriétés usuelles des opérations arithmétiques. En particulier, elles ne sont pas associatives et la multiplication n'est pas distributive par rapport à l'addition.

Exemple 1.11.1. On a

$$\beta^{-p}\,\mathrm{fl}(+)(1\,\mathrm{fl}(-)1) = \beta^{-p}$$

alors que

$$(\beta^{-p}\,\mathrm{fl}(+)1)\,\mathrm{fl}(-)1 = 0.$$

Les quelques résultats suivants montrent qu'il est assez facile d'implémenter une arithmétique flottante du type décrit ci-dessus à partir de l'arithmétique entière.

1.11.1 Addition et soustraction

Proposition 1.11.2. *Soient*

$$a = (.a_1 \cdots a_p)_\beta \beta^e \qquad et \qquad b = (.b_1 \cdots b_p)_\beta \beta^f$$

deux réels strictement positifs à p chiffres significatifs en base β écrits sous forme normalisée. Supposons que $a \geq b$ et posons

$$s = (a_1 \cdots a_p 0)_\beta + \lfloor (b_1 \cdots b_p 0)_\beta \beta^{f-e} \rfloor$$

et

$$\sigma = \begin{cases} 0 & \textit{si } (b_1 \cdots b_p 0)_\beta \beta^{f-e} \textit{ est entier ;} \\ 1 & \textit{sinon.} \end{cases}$$

Alors l'entier s a $p+1$ ou $p+2$ chiffres en base β et les arrondis corrects à p chiffres significatifs en base β de $c = a + b$ et de

$$\tilde{c} = (s\beta + \sigma)\beta^{e-p-2}$$

sont identiques.

Démonstration. Le cas $\sigma = 0$ étant trivial, nous pouvons supposer que $\sigma = 1$.

Dans ce cas, vu la définition de l'entier s, il est clair que

$$s = \lfloor c\beta^{p+1-e} \rfloor < c\beta^{p+1-e}.$$

Il est également clair que l'exposant normalisé de c est e ou $e+1$.

Dans le premier cas, nous devons montrer que

$$\mathrm{crd}(s\beta^{-1} + \sigma\beta^{-2}) = \mathrm{crd}(c\beta^{p-e}).$$

Vu ce qui précède, on sait que le premier chiffre fractionnaire de $c\beta^{p-e}$ coïncide avec le dernier chiffre de s. Si ce chiffre diffère de $\beta/2$, l'égalité précédente est donc clairement vérifiée. Si ce chiffre est égal à $\beta/2$, le fait que $c\beta^{p+1-e}$ et $s + \sigma\beta^{-1}$ possèdent des chiffres fractionnaires non nuls conduit à la même conclusion.

Dans le second cas, nous devons montrer que

$$\mathrm{crd}(s\beta^{-2} + \sigma\beta^{-3}) = \mathrm{crd}(c\beta^{p-e-1}).$$

Vu ce qui précède, on sait que le premier chiffre fractionnaire de $c\beta^{p-e-1}$ coïncide avec l'avant dernier chiffre de s. Si ce chiffre diffère de $\beta/2$, l'égalité précédente est donc clairement vérifiée. Si ce chiffre est égal à $\beta/2$, le fait que $c\beta^{p+1-e}$ et $s + \sigma\beta^{-1}$ possèdent des chiffres fractionnaires non nuls conduit à la même conclusion. □

Proposition 1.11.3. *Soient*

$$a = (.a_1 \cdots a_p)_\beta \beta^e \qquad et \qquad b = (.b_1 \cdots b_p)_\beta \beta^f$$

deux réels strictement positifs à p chiffres significatifs écrits sous forme normalisée. Supposons que $a > b$ et posons

$$s = (a_1 \cdots a_p 00)_\beta - \lceil (b_1 \cdots b_p 00)_\beta \beta^{f-e} \rceil$$

et

$$\sigma = \begin{cases} 0 & \textit{si } (b_1 \cdots b_p 00)_\beta \beta^{f-e} \textit{ est entier ;} \\ 1 & \textit{sinon.} \end{cases}$$

Alors l'entier s a $p+2$ ou $p+1$ chiffres en base β et les arrondis corrects à p chiffres significatifs en base β de $c = a - b$ et de

$$\tilde{c} = (s\beta + \sigma)\beta^{e-p-3}$$

sont identiques.

Démonstration. Le cas $\sigma = 0$ étant trivial, nous pouvons supposer que $\sigma = 1$.

Dans ce cas, vu la définition de l'entier s, il est clair que

$$s = \lfloor c\beta^{p+2-e} \rfloor < c\beta^{p+2-e}.$$

Comme $f < e - 2$, il est également clair que l'exposant normalisé de c est e ou $e - 1$.

Dans le premier cas, nous devons montrer que

$$\mathrm{crd}(s\beta^{-2} + \sigma\beta^{-3}) = \mathrm{crd}(c\beta^{p-e}).$$

Vu ce qui précède, on sait que

$$s\beta^{-2} < c\beta^{p-e} < (s+1)\beta^{-2}.$$

Si la partie fractionnaire de $s\beta^{-2}$ est strictement inférieure à $1/2$, elle vaut au plus $1/2 - \beta^{-2}$ et on a alors

$$\lfloor s\beta^{-2} \rfloor < c\beta^{p-e} < \lfloor s\beta^{-2} \rfloor + \frac{1}{2}.$$

Cela montre que

$$\mathrm{crd}(c\beta^{p-e}) = \lfloor s\beta^{-2} \rfloor = \mathrm{crd}(s\beta^{-2} + \sigma\beta^{-3}).$$

Si la partie fractionnaire de $s\beta^{-2}$ est supérieure ou égale à $1/2$, elle vaut au plus $1 - \beta^{-2}$ et on a alors

$$\lfloor s\beta^{-2} \rfloor + \frac{1}{2} < c\beta^{p-e} < \lfloor s\beta^{-2} \rfloor + 1.$$

Cela montre que

$$\mathrm{crd}(c\beta^{p-e}) = \lfloor s\beta^{-2} \rfloor + 1 = \mathrm{crd}(s\beta^{-2} + \sigma\beta^{-3}).$$

Dans le second cas, nous devons montrer que

$$\mathrm{crd}(s\beta^{-1} + \sigma\beta^{-2}) = \mathrm{crd}(c\beta^{p-e+1}).$$

Vu ce qui précède, on sait que

$$s\beta^{-1} < c\beta^{p-e+1} < (s+1)\beta^{-1}.$$

Si la partie fractionnaire de $s\beta^{-1}$ est strictement inférieure à $1/2$, elle vaut au plus $1/2 - \beta^{-1}$ et on a alors

$$\lfloor s\beta^{-1} \rfloor < c\beta^{p-e+1} < \lfloor s\beta^{-1} \rfloor + \frac{1}{2}.$$

Cela montre que

$$\mathrm{crd}(c\beta^{p-e+1}) = \lfloor s\beta^{-1} \rfloor = \mathrm{crd}(s\beta^{-1} + \sigma\beta^{-2}).$$

Si la partie fractionnaire de $s\beta^{-1}$ est supérieure ou égale à $1/2$, elle vaut au plus $1 - \beta^{-1}$ et on a alors

$$\lfloor s\beta^{-1} \rfloor + \frac{1}{2} < c\beta^{p-e+1} < \lfloor s\beta^{-1} \rfloor + 1.$$

Cela montre que

$$\mathrm{crd}(c\beta^{p-e+1}) = \lfloor s\beta^{-1} \rfloor + 1 = \mathrm{crd}(s\beta^{-1} + \sigma\beta^{-2}).$$

□

Exemple 1.11.4. Calculons $1.25 + 0.0124$ à trois chiffres significatifs exacts. Les calculs se disposent comme suit :

$$\begin{array}{cccc} 1 & 2 & 5 & 0 \\ & & 1 & 2 \\ \hline 1 & 2 & 6 & 2 \end{array}$$

et $\sigma = 1$. On en tire que la valeur approchée cherchée est 1.26.

Calculons $1.25 + 0.000501$ à quatre chiffres significatifs exacts. Les calculs se disposent comme suit :

$$\begin{array}{ccccc} 1 & 2 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5 \\ \hline 1 & 2 & 5 & 0 & 5 \end{array}$$

et $\sigma = 1$. On en tire que la valeur approchée cherchée est 1.251.

Calculons $1.25 - 0.0124$ à trois chiffres significatifs exacts. Les calculs se disposent comme suit :

$$\begin{array}{ccccc} 1 & 2 & 5 & 0 & 0 \\ 0 & 0 & 1 & 2 & 4 \\ \hline 1 & 2 & 3 & 7 & 6 \end{array}$$

et $\sigma = 0$. On en tire que la valeur approchée cherchée est 1.24.

Calculons $1.25 - 0.001491$ à quatre chiffres significatifs exacts. Les calculs se disposent comme suit :

$$\begin{array}{cccccc} 1 & 2 & 5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 5 & 0 \\ \hline 1 & 2 & 4 & 8 & 5 & 0 \end{array}$$

et $\sigma = 1$. On en tire que la valeur approchée cherchée est 1.249.

1.11.2 Multiplication

Proposition 1.11.5. *Soient*

$$a = (.a_1 \cdots a_p)_\beta \beta^e \qquad b = (.b_1 \cdots b_p)_\beta \beta^f$$

deux réels strictement positifs à p chiffres significatifs en base β écrits sous forme normalisée. Posons

$$m = (a_1 \cdots a_p)_\beta (b_1 \cdots b_p)_\beta.$$

Alors, l'entier m a $2p$ ou $2p-1$ chiffres en base β et $\tilde{c} = m\beta^{e+f-2p}$ est égal au produit $c = ab$. En particulier, les arrondis corrects à p chiffres significatifs en base β de $\tilde{c}$ et de c sont identiques.

Démonstration. Tout est évident. □

Exemple 1.11.6. Soient $\beta = 10$, $p = 3$, $a = .125$, $b = .213$. Disposons les calculs comme suit :

$$\begin{array}{ccccc} & & 1 & 2 & 5 \\ & & 2 & 1 & 3 \\ \hline & & 3 & 7 & 5 \\ & 1 & 2 & 5 & \\ 2 & 5 & 0 & & \\ \hline 2 & 6 & 6 & 2 & 5 \end{array}$$

d'où on tire que la valeur de ab arrondie à 3 chiffres significatifs est 0.0266 .

1.11.3 Division

Proposition 1.11.7. *Soient*

$$a = (.a_1 \cdots a_p)_\beta \beta^e \qquad b = (.b_1 \cdots b_p)_\beta \beta^f$$

deux réels strictement positifs à p chiffres significatifs en base β écrits sous forme normalisée. Notons q le quotient de la division euclidienne des entiers

$$(a_1 \cdots a_p)_\beta \beta^{p+1} \qquad et \qquad (b_1 \cdots b_p)_\beta$$

et posons

$$\sigma = \begin{cases} 0 & \text{si le reste de cette division est nul ;} \\ 1 & \text{sinon.} \end{cases}$$

Alors les arrondis corrects à p chiffres significatifs en base β du quotient $c = a/b$ et de

$$\tilde{c} = (q\beta + \sigma)\beta^{e-f-p-2}$$

sont identiques.

Exemple 1.11.8. Soient $\beta = 10$, $p = 3$, $a = 0.125$, $b = 0.315$. Les calculs se disposent comme suit :

1	2	5	0	0	0	0			3	1	5
	9	4	5					3	9	6	8
	2	0	5	0							
	1	8	3	5							
		2	1	5	0						
		1	8	9	0						
			2	6	0	0					
			2	5	2	0					
					8	0					

et on a $\sigma = 1$. On tire que la valeur de a/b arrondie à trois chiffres significatifs est 0.397.

Soient maintenant $a = 0.416$, $b = 0.126$. Les calculs se disposent comme suit :

$$\begin{array}{cccccccc|ccccc}
4 & 1 & 6 & 0 & 0 & 0 & 0 & & & & 1 & 2 & 6 \\
3 & 7 & 8 & & & & & & 3 & 3 & 0 & 1 & 5 \\
\cline{1-3}\cline{9-13}
 & 3 & 8 & 0 & & & & & & & & & \\
 & 3 & 7 & 8 & & & & & & & & & \\
\cline{2-4}
 & & & 2 & 0 & 0 & & & & & & & \\
 & & & 1 & 2 & 6 & & & & & & & \\
\cline{4-6}
 & & & & 7 & 4 & 0 & & & & & & \\
 & & & & 6 & 3 & 0 & & & & & & \\
\cline{5-7}
 & & & & 1 & 1 & 0 & & & & & &
\end{array}$$

et $\sigma = 1$. On en tire que la valeur de a/b arrondie à trois chiffres significatifs est 3.30.

1.12 Propagation des erreurs dans les calculs

Dans une étape de calcul sur machine, on part généralement de certains réels machine $(\tilde{x}_1, \ldots, \tilde{x}_n)$ représentant approximativement les données $x_1, \ldots, x_n \in \mathbb{R}$ et on leur associe des réels machine $\tilde{y}_1, \ldots, \tilde{y}_m$ représentant approximativement les résultats

$$\begin{aligned} y_1 &= f_1(x_1, \ldots, x_n) \\ &\vdots \\ y_m &= f_m(x_1, \ldots, x_n) \end{aligned}$$

En général, on a

$$\tilde{y}_j = \mathrm{fl}(f_j)(\tilde{x}_1, \ldots, \tilde{x}_n)$$

et pour contrôler la précision de l'approximation $\tilde{y}_j$, il faut contrôler à la fois l'écart entre $\mathrm{fl}(f_j)(\tilde{x}_1, \ldots, \tilde{x}_n)$ et $f_j(\tilde{x}_1, \ldots, \tilde{x}_n)$ et celui entre $f_j(\tilde{x}_1, \ldots, \tilde{x}_n)$ et $f_j(x_1, \ldots, x_n)$.

Nous allons d'abord considérer ce second problème en convenant de poser

$$\Delta f_j(x_1, \ldots, x_n) = f_j(\tilde{x}_1, \ldots, \tilde{x}_n) - f_j(x_1, \ldots, x_n)$$

et

$$\delta f_j(x_1, \ldots, x_n) = \Delta f_j(x_1, \ldots, x_n)/f_j(x_1, \ldots, x_n).$$

Pour les opérations arithmétiques élémentaires, on a le résultat suivant :

Proposition 1.12.1. *Soient $\tilde{x}$ et $\tilde{y}$ des approximations de x et y. Notons Δx, Δy et δx, δy les erreurs absolues et relatives associées. Alors,*

(a) $\Delta(x + y) = \Delta x + \Delta y$;

(b) $\Delta(xy) = y\Delta x + x\Delta y + \Delta x\Delta y$;

(c) $\Delta(x/y) = \dfrac{y\Delta x - x\Delta y}{y(y+\Delta y)}$ *si* $y,\ y+\Delta y \in \mathbb{R}_0$;

et en particulier, on a

(a') $\delta(x+y) = \dfrac{x}{x+y}\delta x + \dfrac{y}{x+y}\delta y$ *si* $x,\ y,\ x+y \in \mathbb{R}_0$;

(b') $\delta(xy) = \delta x + \delta y + \delta x\delta y$ *si* $x,\ y \in \mathbb{R}_0$;

(c') $\delta(x/y) = \dfrac{\delta x - \delta y}{1+\delta y}$ *si* $x,\ y,\ 1+\delta y \in \mathbb{R}_0$.

Démonstration. C'est immédiat. □

Remarque 1.12.2. La relation (a') montre que si x, y sont de même signe, alors

$$|\delta(x+y)| \leq |\delta x| + |\delta y|.$$

Par contre, si x, y sont de signes opposés et de modules voisins, alors $|\delta(x+y)|$ est généralement très grand. On traduit cela en disant que la soustraction de nombres voisins est mal conditionnée.

Une conséquence de ce phénomène est qu'il faut parfois modifier la formule utilisée pour calculer un résultat de manière à contourner cette neutralisation de termes. Par exemple, considérons l'équation

$$x^2 - 56x + 1 = 0.$$

Ses racines sont

$$x_1 = 28 - \sqrt{783}, \qquad x_2 = 28 + \sqrt{783}.$$

Or,

$$\sqrt{783} = 27.982 \pm \tfrac{1}{2}10^{-3}$$

est proche de 28. Si on effectue le calcul, on trouve

$$x_1 = 0.018 \pm \tfrac{1}{2}10^{-3}, \qquad x_2 = 55.982 \pm \tfrac{1}{2}10^{-3}.$$

Le module de l'erreur relative sur $\sqrt{783}$ est majoré par $2\ 10^{-5}$. Celui de l'erreur relative sur x_2 est majoré par 10^{-5} alors que celui de l'erreur relative sur sur x_1 est majoré par $3\ 10^{-2}$. Comme $x_1x_2 = 1$, on a aussi $x_1 = 1/x_2 = 0.0178629$ avec une erreur relative de module majoré par 10^{-5}, c'est-à-dire avec une erreur absolue de module majoré par $2\ 10^{-7}$. On en tire que

$$x_1 = 0.017863 \pm \tfrac{1}{2}10^{-6}.$$

Dans ce cas, on n'a pas seulement contourné le problème de la neutralisation des termes, mais on a aussi obtenu une précision absolue sur x_1 meilleure que celle sur $\sqrt{783}$. Cet exemple illustre bien l'importance du choix de l'algorithme utilisé pour résoudre un problème numérique.

Pour une opération $f(x_1, \ldots, x_n)$ plus compliquée que $+$, $\cdot$, $/$, on n'a pas de formule aussi simple que celles de la proposition précédente pour estimer l'erreur absolue et relative. On a cependant le résultat suivant :

Proposition 1.12.3. *Soit Ω un ouvert de $\mathbb{R}^n$, soient $\tilde{x}_1, \ldots, \tilde{x}_n$ des valeurs approchées de $x_1, \ldots, x_n$ et soient $\Delta x_1, \ldots, \Delta x_n$; $\delta x_1, \ldots, \delta x_n$ les erreurs absolues et relatives associées. Supposons que $f \in C_1(\Omega)$ et que le segment joignant $x = (x_1, \ldots, x_n)$ à $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_n)$ soit dans Ω, alors il existe $\theta \in]0, 1[$ tel que*

$$\Delta f(x_1, \ldots, x_n) = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(x + \theta \Delta x) \Delta x_j.$$

En particulier, si $x_1, \ldots, x_n, f(x_1, \ldots, x_n) \in \mathbb{R}_0$, on a

$$\delta f(x_1, \ldots, x_n) = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(x + \theta \Delta x) \frac{x_j}{f(x)} \delta x_j.$$

Démonstration. C'est une conséquence directe de la formule de Taylor avec reste de Lagrange. □

Corollaire 1.12.4. *Dans les conditions de la proposition précédente, on a*

$$\Delta f(x_1, \ldots, x_n) = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j} \Delta x_j + o(\Delta x)$$

$$\delta f(x_1, \ldots, x_n) = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j} \frac{x_j}{f} \delta x_j + o(\delta x).$$

Définition 1.12.5. Il résulte du corollaire précédent que

$$|\delta f(x_1, \ldots, x_n)| \leq \sum_{j=1}^{n} \left| \frac{\partial f}{\partial x_j} \frac{x_j}{f} \right| |\delta x_j| + o(\delta x).$$

Les nombres

$$\eta_j(f) = \left| \frac{\partial f}{\partial x_j} \frac{x_j}{f} \right|$$

donnent donc une estimation de la sensibilité de $\delta f(x_1, \ldots, x_n)$ aux erreurs relatives δx_j des arguments de f. On dit que $\eta_j(f)$ est le *nombre de conditionnement* de f *par rapport* à x_j. Le *nombre de conditionnement* de f est par définition

$$\eta(f) = \sum_{j=1}^{n} \left| \frac{\partial f}{\partial x_j} \frac{x_j}{f} \right|.$$

La fonction f sera dite *bien* (resp. *mal*) *conditionnée* si $\eta(f)$ est petit (resp. grand) par rapport au contexte.

Remarque 1.12.6. Il résulte de la définition précédente que si toutes les données sont entachées d'erreurs relatives de module majoré par δ alors f est entaché d'une erreur relative de module inférieur à

$$\eta(f)\delta + o(\delta).$$

Exemples 1.12.7. (a) Si $f(x) = \sqrt{x}$, on a

$$\eta(f) = \left| \frac{1}{2\sqrt{x}} \frac{x}{\sqrt{x}} \right| = \frac{1}{2}.$$

(b) Si $f(x) = \sin x$, on a

$$\eta(f) = \left| \frac{\cos x}{\sin x} x \right| = \left| \frac{x}{\operatorname{tg} x} \right|.$$

Considérons à présent le problème du contrôle de l'écart entre $\mathrm{fl}(f_j)(x_1, \ldots, x_n)$ et $f_j(x_1, \ldots, x_n)$ lorsque $x_1, \ldots, x_n$ sont n nombres machines. Nous nous limiterons au cas d'un produit de n facteurs et de la somme de n termes.

Dans le premier cas, il s'agit d'estimer l'écart entre

$$\tilde{y} = (\ldots((x_1\, \mathrm{fl}(\cdot) x_2)\, \mathrm{fl}(\cdot) x_3) \ldots \mathrm{fl}(\cdot) x_n)$$

et $y = x_1 \ldots x_n$. Rappelons que

$$x_1\, \mathrm{fl}(\cdot) x_2 = (x_1 \cdot x_2)(1 + \delta_2) \qquad \text{avec } |\delta_2| < u,$$

où u est l'unité machine. Ainsi,

$$\tilde{y} = y(1 + \delta_2)(1 + \delta_3) \ldots (1 + \delta_n) \qquad \text{avec } |\delta_2| < u, \ldots, |\delta_n| < u.$$

On a donc

$$\delta y = (1 + \delta_2) \ldots (1 + \delta_n) - 1$$

d'où on tire que

$$|\delta y| \leq (1+u)^{n-1} - 1.$$

Remarquons qu'au premier ordre en u, la majorante vaut $(n-1)u$. En fait, on peut montrer (voir Lemme ci-dessous) que si l'on fait l'hypothèse réaliste $(n-1)u < 0.1$, alors $(1+u)^{n-1} - 1 \leq 1.06(n-1)u$. Il en résulte que dans ce cas

$$|\delta y| \leq 1.06(n-1)u$$

et l'erreur relative admet une borne linéaire en le nombre de termes.

Lemme 1.12.8. *Si $\alpha > 0$, on a*

$$\sup\left\{\frac{(1+x)^k - 1}{kx} : x > 0, k > 0, xk < \alpha\right\} = \frac{e^\alpha - 1}{\alpha}.$$

Démonstration. Fixons k et considérons la fonction f définie par

$$f(x) = \frac{(1+x)^k - 1}{kx} = \sum_{l=1}^{k} \frac{(k-1)!}{l!(k-l)!} x^{l-1}$$

pour $x > 0$. Comme cette fonction est croissante pour $x > 0$, on a

$$\sup_{0<x<\frac{\alpha}{k}} \frac{(1+x)^k - 1}{kx} = \frac{(1+\frac{\alpha}{k})^k - 1}{\alpha}.$$

Considérons à présent la fonction g

$$g(x) = x \ln\left(1 + \frac{\alpha}{x}\right)$$

pour $x > 0$. On a

$$\begin{aligned} g'(x) &= \ln\left(1+\frac{\alpha}{x}\right) + x\frac{-\alpha/x^2}{1+\alpha/x} \\ &= \ln\left(1+\frac{\alpha}{x}\right) - \frac{\alpha/x}{1+\alpha/x}. \end{aligned}$$

Cette dernière expression est strictement positive car pour $h > 0$, il existe $\theta \in]0,1[$ avec

$$\ln(1+h) = \ln(1) + \frac{1}{1+\theta h} h$$

ce qui entraîne que

$$\ln(1+h) > \frac{h}{1+h}.$$

Comme $g'(x) > 0$ pour $x > 0$, g est croissant et

$$\sup_{x>0} g(x) = \lim_{x \to \infty} g(x) = \alpha.$$

Il s'ensuit que

$$\sup_{x>0} \left(1 + \frac{\alpha}{x}\right)^x = e^\alpha$$

et la conclusion en découle aisément. □

Dans le cas d'une somme de n termes, posons $y = x_1 + \cdots + x_n$ et

$$\tilde{y} = (\ldots((x_1 \,\mathrm{fl}(+) x_2)\,\mathrm{fl}(+) x_3)\ldots \mathrm{fl}(+) x_n).$$

En procédant comme pour la multiplication, on trouve des nombres $\delta_2, \ldots, \delta_n$ tels que $|\delta_2| \leq u, \ldots, |\delta_n| \leq u$ pour lesquels

$$\tilde{y} = (\ldots((x_1 + x_2)(1 + \delta_2) + x_3)(1 + \delta_3) \cdots + x_n)(1 + \delta_n).$$

En convenant de poser $\delta_1 = 0$, on peut encore écrire

$$\tilde{y} = \sum_{j=1}^{n} x_j (1 + \delta_j) \ldots (1 + \delta_n).$$

Ainsi,

$$\Delta y = \sum_{j=1}^{n} x_j \left[(1 + \delta_j) \ldots (1 + \delta_n) - 1\right]$$

et

$$|\Delta y| \leq \sum_{j=1}^{n} |x_j| \left[(1 + u)^{n-j+1} - 1\right].$$

Si on fait l'hypothèse $nu \leq 0.1$, il vient

$$|\Delta y| \leq \left[\sum_{j=1}^{n} |x_j|(n - j + 1)\right] 1.06u.$$

Cette formule laisse entrevoir que tous les termes ne contribuent pas de manière égale à l'erreur absolue. En particulier, si la suite $|x_1|, \ldots, |x_n|$ est décroissante, il est préférable de sommer les x_j en suivant l'ordre des j décroissants.

Exemple 1.12.9. On a

$$S = \sum_{n=1}^{10000} \frac{1}{n^2} = 1.64483 \pm \frac{1}{2} 10^{-5}.$$

Cependant, si on calcule cette somme sur une machine utilisant une arithmétique flottante correctement arrondie à 6 chiffres significatifs en base 10, on obtient

$$\tilde{S} = 1.64307 \qquad |\Delta S| > 1.7\ 10^{-3}$$

en procédant par ordre des n croissants et

$$\tilde{S} = 1.64483 \qquad |\Delta S| < 0.5\ 10^{-5}$$

en procédant par ordre des n décroissants.

Pour terminer notre étude de la propagation des erreurs, considérons un dernier exemple qui montre comment celle-ci peut rendre totalement inutile du point de vue numérique un algorithme parfaitement correct du point de vue mathématique. Partons du problème consistant à calculer

$$I_m = \int_0^1 \frac{x^m}{5+x} dx$$

pour un naturel m. Clairement

$$I_0 = \int_0^1 \frac{dx}{5+x} = \ln(5+x)|_0^1 = \ln(6/5)$$

et

$$I_m + 5I_{m-1} = \int_0^1 x^{m-1} dx = \left.\frac{x^m}{m}\right|_0^1 = \frac{1}{m}.$$

Pour calculer I_m, on peut donc calculer I_0 puis calculer de proche en proche I_m par

$$I_m = \frac{1}{m} - 5I_{m-1}.$$

Si les calculs sont effectués à 3 décimales, on obtient successivement

$$\begin{aligned} \tilde{I}_0 &= 0.182 \\ \tilde{I}_1 &= 0.090 \\ \tilde{I}_2 &= 0.050 \\ \tilde{I}_3 &= 0.083 \\ \tilde{I}_4 &= -0.165 \end{aligned}$$

Les derniers résultats sont incohérents. En effet, la suite I_m est une suite décroissante de nombres positifs alors que $\tilde{I}_3 > \tilde{I}_2$ et que $\tilde{I}_4 < 0$. En fait, il est clair qu'à chaque étape l'erreur absolue sur I_{m-1} est grosso-modo multipliée par 5, ce qui explique que

les résultats obtenus s'écartent très rapidement des valeurs théoriques. Remarquons cependant que I_m décroît vers 0. Si on pose $\tilde{I}_M = 0$, on a donc

$$|\Delta I_M| \leq I_M \leq I_0 \leq 0.183.$$

Or, la relation

$$I_{m-1} = \frac{1}{5}\left(\frac{1}{m} - I_m\right)$$

permet de calculer de proche en proche $I_{M-1}, I_{M-2}, \ldots, I_1, I_0$ si l'on connaît I_M. De plus, si on remplace I_M par une valeur approchée $\tilde{I}_M$, on a

$$\Delta I_{M-k} = \frac{1}{5^k}\Delta I_M$$

pour $k \in \{1, \ldots, M\}$. Il s'ensuit qu'en partant de l'approximation $\tilde{I}_M = 0$ pour M suffisamment grand, on peut obtenir I_m avec une bonne précision. Par exemple, si on prend $\tilde{I}_4 = 0$, on a $\Delta I_4 \leq 2\ 10^{-1}$ et on tire que

$$\begin{aligned} I_3 &= 0.0500 \pm 4\ 10^{-2} \\ I_2 &= 0.0566 \pm 8\ 10^{-3} \\ I_1 &= 0.0887 \pm 2\ 10^{-3} \\ I_0 &= 0.1822 \pm 3\ 10^{-4} \end{aligned}$$

ce qui donne la valeur de I_0 à 3 décimales exactes.

2 Équations non-linéaires

2.1 Zéros des fonctions réelles d'une variable réelle

Considérons le problème du calcul approché des solutions d'une équation du type

$$f(x) = 0$$

où f est une fonction réelle définie sur un intervalle I de $\mathbb{R}$. Commençons par une méthode élémentaire exigeant très peu de régularité sur f.

2.1.1 Méthode de la bissection

Considérons une fonction réelle et continue f dans un intervalle $[a, b]$ de $\mathbb{R}$. Si f prend des valeurs de signes opposés en a et en b, alors le théorème des valeurs intermédiaires montre que f possède au moins un zéro dans $]a, b[$. La méthode de la bissection est basée sur une utilisation systématique de cette remarque.

Soit ε la précision à laquelle on souhaite déterminer un zéro de f dans $[a, b]$. L'algorithme utilisé est le suivant. On pose

$$a_0 = a, b_0 = b \text{ et } d_0 = b - a.$$

Étant donnés $a_m, b_m \in [a, b]$ tels que $f(a_m)f(b_m) < 0$ et $d_m = b_m - a_m$, on pose $c_m = (a_m + b_m)/2$ et on calcule $f(c_m)$.

— Si $f(a_m)f(c_m) < 0$, on pose $a_{m+1} = a_m$ et $b_{m+1} = c_m$.

— Si $f(a_m)f(c_m) > 0$, on pose $a_{m+1} = c_m$ et $b_{m+1} = b_m$.

— Si $f(c_m) = 0$, c_m est un zéro de f et on peut s'arrêter.

On pose enfin $d_{m+1} = d_m/2$. Si $d_{m+1} > \varepsilon$ on recommence ce qui a été fait ci-dessus en remplaçant m par $m + 1$. Sinon on s'arrête. En cas d'arrêt, il est clair que c_m est une valeur approchée d'un zéro de f à ε près. De plus, comme $d_{m+1} = d_m/2$, l'erreur absolue à l'étape m est bornée par $(b - a)2^{-m-1}$.

Considérons par exemple la fonction $f(x) = x - e^{-x}$. Cette fonction est strictement croissante et $f(0)f(1) < 0$. Elle possède donc un et un seul zéro dans l'intervalle $]0, 1[$. La bissection donne les approximations suivantes :

m	a_m	b_m
0	0	1
1	0.5	1
2	0.5	0.75
5	0.5625	0.59375
10	0.5664062	0.5673828
20	0.56714248	0.56714344

alors qu'une valeur approchée à 10 décimales exactes de la solution est 0.5671432904. A chaque étape, on gagne environ un chiffre binaire. Puisque $10^{-1} \sim 2^{-3.3}$, environ 4 itérations sont nécessaires pour gagner un chiffre décimal.

La méthode de la bissection est donc malheureusement assez lente. Cependant cette lenteur est compensée par le fait que l'on est assuré d'obtenir toujours une suite qui convergent vers un zéro de f dans $[a, b]$. On dit que c'est une méthode *globalement convergente*.

2.1.2 Méthodes itératives

Une autre idée pour résoudre numériquement une équation du type

$$f(x) = 0$$

est de chercher à approcher une de ses solutions λ par une suite x_n définie par une formule de récurrence du type

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

et une valeur initiale x_0. Les méthodes de résolution basées sur cette idée sont dites itératives. Avant de donner les résultats généraux sur ces méthodes, nous allons d'abord traiter un exemple simple.

Processus de Héron

Il s'agit d'une méthode itérative pour le calcul de la racine carrée de $a > 0$. La suite utilisée est définie par

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right) \qquad (*)$$

et une valeur initiale de x_0. Pour étudier la convergence de cette suite, remarquons tout d'abord que

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

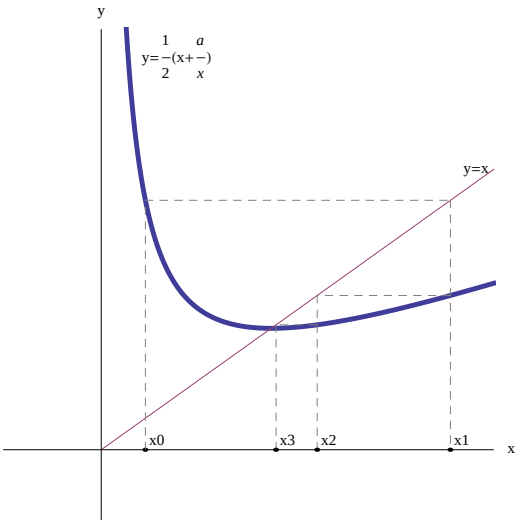
Ainsi, quelque soit $x_n > 0$, on a $x_{n+1} \geq \sqrt{a}$. De plus,

$$x_{n+1} - x_n = \frac{x_n}{2} + \frac{a}{2x_n} - x_n = \frac{a}{2x_n} - \frac{x_n}{2} = \frac{a - x_n^2}{2x_n}$$

et cette dernière expression est négative si $x_n \geq \sqrt{a}$. Il s'ensuit que la suite x_n est décroissante pour $n \geq 1$. Elle converge donc vers une limite $\lambda \geq \sqrt{a}$. De la relation (*), on tire que

$$\lambda = \frac{1}{2}\left(\lambda + \frac{a}{\lambda}\right) \Leftrightarrow \lambda^2 = a.$$

Par conséquent, $\lambda = \sqrt{a}$. Illustrons graphiquement cette convergence.



Remarquons aussi que si Δ_n, δ_n désignent les erreurs absolue et relative à l'étape n, on a

$$|\Delta_{n+1}| \leq \frac{1}{2\sqrt{a}}|\Delta_n|^2$$

et

$$|\delta_{n+1}| \leq \frac{1}{2}|\delta_n|^2.$$

Il s'ensuit que l'on double plus ou moins le nombre de chiffres significatifs exacts à chaque itération. C'est un exemple d'une convergence qui est au moins quadratique. Calculons par exemple $\sqrt{2}$ par cet algorithme. Prenons $x_0 = 1$. On a

$$\begin{aligned} x_0 &= 1 \\ x_1 &= 1.5 \\ x_2 &= 1.416666667 \\ x_3 &= 1.414215686 \\ x_4 &= 1.414213562 \end{aligned}$$

où la dernière valeur ne comporte que des chiffres exacts. Notons que de

$$x_{n+1} - x_n = \frac{a - x_n^2}{2x_n}$$

on tire que

$$x_{n+1} - x_n = \frac{a}{2x_n} - \frac{x_n}{2} \leq \frac{\sqrt{a}}{2} - \frac{x_n}{2}$$

et que par conséquent, on a

$$x_n - \sqrt{a} \leq 2(x_n - x_{n+1}).$$

En y ajoutant

$$x_{n+1} - x_n$$

on trouve que

$$\Delta_{n+1} \leq x_n - x_{n+1}$$

ce qui fournit une condition d'arrêt très simple pour le processus considéré.

Exercice 2.1.1. Étudier de manière similaire la méthode itérative

$$x_{n+1} = x_n(2 - ax_n)$$

fournissant l'inverse de $a > 0$ sans division.

Théorème du point fixe
La convergence d'une méthode itérative générale du type

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

est gouvernée par le résultat suivant :

Théorème 2.1.2. *Soit F un fermé de $\mathbb{R}^n$ et soit*

$$\varphi : F \to F$$

une application continue telle que

$$|\varphi(x) - \varphi(y)| \leq \mu|x - y| \qquad (\forall\, x, y \in F)$$

pour un certain $\mu \in [0, 1[$. Alors, la suite x_n définie par la relation de récurrence

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

et la condition initiale $x_0 \in F$ converge vers le seul point fixe λ de φ dans F. De plus, la convergence de x_n est au moins linéaire de taux μ. En particulier, on a

$$|x_m - \lambda| \leq \mu^m |x_0 - \lambda|.$$

Démonstration. Pour $n \geq 1$, on a

$$|x_{n+1} - x_n| = |\varphi(x_n) - \varphi(x_{n-1})| \leq \mu|x_n - x_{n-1}|$$

d'où l'on tire par récurrence que

$$|x_{n+1} - x_n| \leq \mu^n |x_1 - x_0|$$

pour $n \geq 1$. Ainsi, si $q \geq p \geq 1$, il vient

$$\begin{aligned} |x_{q+1} - x_p| &\leq |x_{q+1} - x_q| + |x_q - x_{q-1}| + \cdots + |x_{p+1} - x_p| \\ &\leq (\mu^q + \mu^{q-1} + \cdots + \mu^p)|x_1 - x_0| \\ &\leq \frac{\mu^{q+1} - \mu^p}{\mu - 1}|x_1 - x_0|. \end{aligned}$$

Comme $0 \leq \mu < 1$, on en tire que la suite x_n est de Cauchy. Cette suite converge donc vers une limite $\lambda \in F$. Or, si on passe à la limite dans la relation

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

en tenant compte de la continuité de φ, on voit que

$$\lambda = \varphi(\lambda).$$

Il s'ensuit que λ est un point fixe de φ dans F. Si λ' est un autre point fixe de φ dans F, on a

$$|\lambda' - \lambda| = |\varphi(\lambda') - \varphi(\lambda)| \leq \mu|\lambda' - \lambda|.$$

Comme $\mu < 1$, cette relation montre que

$$|\lambda' - \lambda| = 0,$$

c'est-à-dire que $\lambda' = \lambda$. Enfin, de la relation

$$|x_{n+1} - \lambda| = |\varphi(x_n) - \varphi(\lambda)| \leq \mu|x_n - \lambda|,$$

on tire que la convergence de la suite x_n est au moins linéaire de taux μ. □

Remarque 2.1.3. (a) Soit φ une fonction réelle de classe C_1 sur un intervalle I de $\mathbb{R}$ et soit $\lambda \in I$ tel que $\varphi(\lambda) = \lambda$. Si $|\varphi'(\lambda)| < 1$, alors le théorème du point fixe s'applique dans un intervalle du type $[\lambda - \delta, \lambda + \delta]$ pour δ suffisamment petit. En effet, par le théorème des accroissements finis, on a

$$\varphi(x) - \varphi(\lambda) = \varphi'(\xi)(x - \lambda)$$

pour un $\xi \in]\lambda, x[$. Il s'ensuit que si δ est choisi de telle sorte qu'il existe $\mu < 1$ tel que

$$|\varphi'(\xi)| \leq \mu$$

si $\xi \in [\lambda - \delta, \lambda + \delta]$, alors

$$|\varphi(x) - \varphi(\lambda)| \leq \mu|x - \lambda| \leq |x - \lambda|$$

pour tout $x \in [\lambda - \delta, \lambda + \delta]$. Comme cette relation entraîne que

$$\varphi([\lambda - \delta, \lambda + \delta]) \subset [\lambda - \delta, \lambda + \delta]\,,$$

le théorème du point fixe est applicable. En particulier, pour toute condition initiale

$$x_0 \in [\lambda - \delta, \lambda + \delta]\,,$$

la suite définie par $x_{m+1} = \varphi(x_m)$ converge vers λ.

(b) La situation considérée ci-dessus se produit en particulier si $\varphi'(\lambda) = 0$. Notons que si $\varphi \in C_p(I)$ et si de plus $\varphi''(\lambda) = 0, \ldots, \varphi^{(p-1)}(\lambda) = 0$ pour un certain $p \geq 2$, alors

$$\lim_{m \to \infty} \frac{x_{m+1} - \lambda}{(x_m - \lambda)^p} = \frac{\varphi^{(p)}(\lambda)}{p!} \qquad (*)$$

lorsque $x_m \neq \lambda$ pour $m \gg 0$. En effet, la formule de Taylor avec reste de Lagrange montre que pour tout $m \in \mathbb{N}$, il existe $\xi_m \in \,]x_m, \lambda[$ tel que

$$x_{m+1} - \lambda = \varphi(x_m) - \varphi(\lambda) = \frac{1}{p!}\varphi^{(p)}(\xi_m)(x_m - \lambda)^p$$

et la conclusion résulte de ce que $\varphi^{(p)}$ est continu. De (*), il résulte que pour tout

$$\mu > \left|\frac{\varphi^{(p)}(\lambda)}{p!}\right|,$$

il existe $M > 0$ tel que

$$|x_{m+1} - \lambda| \leq \mu |x_m - \lambda|^p$$

si $m \geq M$. En particulier, la convergence de x_m est asymptotiquement au moins d'ordre p.

(c) Si la condition $|\varphi'(\lambda)| < 1$ n'est pas satisfaite, on peut souvent remplacer φ par une fonction ψ de classe C_1 telle que $\psi(\lambda) = \lambda$ et pour laquelle $|\psi'(\lambda)| < 1$. Par exemple, si $\varphi'(\lambda) \neq 1$, on peut prendre

$$\psi(x) = \frac{rx + \varphi(x)}{r + 1} \qquad (r \neq -1)$$

pour r convenable. En effet,

$$\psi(\lambda) = \lambda \Leftrightarrow \varphi(\lambda) = \lambda$$

et

$$|\psi'(\lambda)| = \left|\frac{r + \varphi'(\lambda)}{r + 1}\right|$$

peut être rendu inférieur à 1 pour r bien choisi. On peut même l'annuler en prenant $r = -\varphi'(\lambda)$.

Exemple 2.1.4. Comme application de ce qui a été dit en (c) ci-dessus, généralisons la méthode de Héron au cas des racines $n^{\text{èmes}}$ de $a > 0$. Pour résoudre l'équation

$$x^n = a \qquad (a > 0, n \in \mathbb{N}_0),$$

on peut d'abord la réécrire

$$x = ax^{1-n}$$

puis voir sa solution positive $\sqrt[n]{a}$ comme le point fixe positif de

$$\varphi(x) = ax^{1-n}.$$

Cependant, comme

$$\varphi'(x) = a(1-n)x^{-n},$$

on a

$$|\varphi'(\sqrt[n]{a})| = |1-n| = n-1 \geq 1$$

si $n \geq 2$. Si l'on remplace φ par

$$\psi(x) = \frac{(n-1)x + ax^{1-n}}{(n-1)+1} = \frac{1}{n}\left((n-1)x + \frac{a}{x^{n-1}}\right),$$

on a $\psi(\sqrt[n]{a}) = \sqrt[n]{a}$ et $\psi'(\sqrt[n]{a}) = 0$. La méthode itérative

$$x_{m+1} = \psi(x_m)$$

converge donc au moins quadratiquement vers $\sqrt[n]{a}$ si x_0 est assez proche de $\sqrt[n]{a}$. En fait, cette méthode converge pour tout $x_0 > 0$. En effet,

$$\psi'(x) = \frac{1}{n}\left((n-1) + (1-n)ax^{-n}\right) = \frac{n-1}{n}(1 - ax^{-n}),$$

d'où le tableau de variation

	0		$\sqrt[n]{a}$	
ψ'		$-$	0	$+$
ψ		$\searrow$	$\sqrt[n]{a}$	$\nearrow$

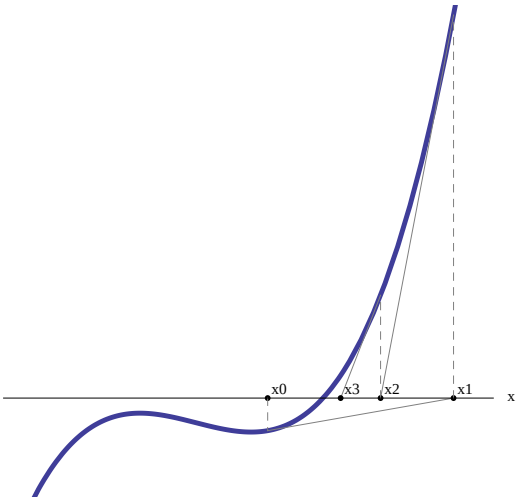
De ce tableau, on tire que $x_m \geq \sqrt[n]{a}$ si $m \geq 1$. Il s'ensuit que

$$x_{m+1} - x_m = ax_m^{1-n} - x_m = x_m(a/x_m^n - 1) \leq 0$$

si $m \geq 1$. La suite $(x_m)_{m\geq 1}$ est donc décroissante et minorée d'où la conclusion.

2.1.3 Méthode de Newton

Soit f une fonction réelle de classe C_1 sur un intervalle I de $\mathbb{R}$ dont la dérivée ne s'annule pas sur cet intervalle et soit $\lambda \in I$ un zéro de f. Supposons que x_m soit une valeur approchée de λ. Considérons la droite τ tangente au graphe de f au point d'abscisse x_m et notons x_{m+1} l'abscisse du point d'intersection de τ avec l'axe des x. Comme $(x, f(x))$ est proche de τ si x est proche de x_m, il est naturel d'espérer que x_{m+1} soit plus proche de λ que ne l'est x_m. La méthode de Newton consiste à partir d'une approximation initiale x_0 de λ et à construire de proche en proche les approximations x_m en suivant le procédé ci-dessus.



L'équation de la tangente au graphe de f en $(x_m, f(x_m))$ est

$$y = f(x_m) + (x - x_m)f'(x_m).$$

Il s'ensuit que

$$x_{m+1} = x_m - \frac{f(x_m)}{f'(x_m)}$$

si $f'(x_m) \neq 0$. La méthode de Newton est donc une méthode itérative associée à la fonction

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

définie pour les x tels que $f'(x) \neq 0$. Remarquons que si $f'(x) \neq 0$, on a

$$\varphi(x) = x \Leftrightarrow f(x) = 0$$

et que si f est de classe C_2, alors

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

s'annule si $f(x) = 0$. On peut donc énoncer le résultat suivant :

Proposition 2.1.5. *Si f est une fonction de classe C_2 au voisinage de $\lambda \in \mathbb{R}$ et si $f(\lambda) = 0$, $f'(\lambda) \neq 0$, alors il existe $\delta > 0$ tel que la suite définie par*

$$x_{m+1} = x_m - \frac{f(x_m)}{f'(x_m)}, \qquad x_0 \in [\lambda - \delta, \lambda + \delta]$$

converge vers λ. De plus, cette convergence est au moins quadratique de taux μ quelque soit

$$\mu > \left| \frac{f''(\lambda)}{2f'(\lambda)} \right|.$$

Si $f''(\lambda) \neq 0$, on a même

$$\lim_{m \to \infty} \frac{x_{m+1} - \lambda}{(x_m - \lambda)^2} = \frac{f''(\lambda)}{2f'(\lambda)}.$$

Démonstration. La première partie provient de la Remarque 2.1.3 (a). La seconde provient de la Remarque 2.1.3 (b) si f est de classe C_3. Pour f de classe C_2, on peut remarquer que

$$0 = f(\lambda) = f(x_m) + (\lambda - x_m) f'(x_m) + \frac{(\lambda - x_m)^2}{2} f''(\xi_m)$$

avec $\xi_m \in]\lambda, x_m[$. Par conséquent,

$$x_{m+1} - \lambda = x_m - \lambda - \frac{f(x_m)}{f'(x_m)} = \frac{(\lambda - x_m)^2}{2} \frac{f''(\xi_m)}{f'(x_m)},$$

d'où la conclusion puisque f' et f'' sont continus et $\xi_m \to \lambda$. □

Le résultat ci-dessus montre que la méthode de Newton n'est en général que localement convergente. On a cependant le résultat suivant :

Proposition 2.1.6. *Soit f une fonction de classe C_2 sur $[a, b]$ telle que*

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

(b) f' ne s'annule pas sur $]a, b[$,

(c) $f'' \geq 0$ (resp. $f'' \leq 0$) sur $]a, b[$.

Alors, la méthode de Newton avec comme condition initiale $x_0 \in]a, b[$ tel que $f(x_0) \geq 0$ (resp. $f(x_0) \leq 0$) converge vers l'unique zéro de f sur $[a, b]$.

Démonstration. Les hypothèses assurent que f', f'' ont un signe constant sur $[a,b]$. Traitons le cas $f' \geq 0$, $f'' \geq 0$, les autres cas étant similaires. Comme la fonction f est strictement croissante sur $]a,b[$, il est clair que $f(a) < 0$ et $f(b) > 0$ et que f s'annule en un seul point $\lambda \in]a,b[$. L'hypothèse $f(x_0) \geq 0$ entraîne que $x_0 \geq \lambda$. Supposons que $x_n \in [\lambda, b[$. Comme

$$0 = f(\lambda) = f(x_n) + (\lambda - x_n)f'(x_n) + \frac{(\lambda - x_n)^2}{2} f''(\xi_n)$$

avec $\xi_n \in [\lambda, x_n]$. Il s'ensuit que

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = \lambda + \frac{(\lambda - x_n)^2}{2} \frac{f''(\xi_n)}{f'(x_n)}.$$

On a donc $x_{n+1} \geq \lambda$ et puisque $f(x_n) \geq 0$ on a aussi

$$x_{n+1} \leq x_n.$$

Par récurrence, on en tire que $x_n \in [\lambda, b[$ et que

$$x_{n+1} \leq x_n$$

pour tout $n \in \mathbb{N}$. Il s'ensuit que la suite x_n est décroissante et minorée par λ. Cette suite possède donc une limite et celle-ci est forcément égale à λ puisque λ est le seul zéro de f sur $[a,b]$. □

2.1.4 Méthode de la sécante

Un des inconvénients de la méthode de Newton est la nécessité de calculer $f'(x_n)$ pour chaque approximation x_n. Si $f'(x)$ n'est pas connu analytiquement à priori, cela peut en effet poser pas mal de difficultés. On peut donc penser à remplacer cette dérivée par son approximation

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

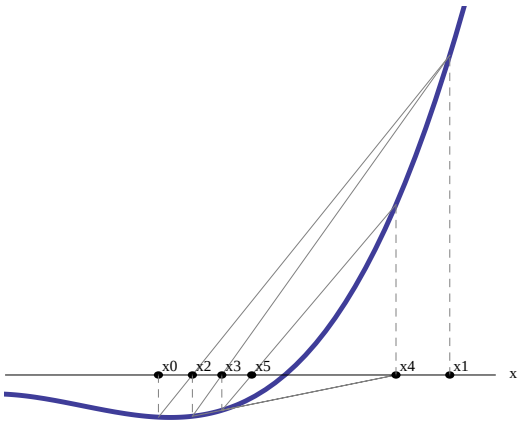
A l'étape $n+1$, l'approximation du zéro λ cherché est alors

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

Cela correspond géométriquement à estimer λ en approchant le graphe de f par la sécante joignant le point $(x_{n-1}, f(x_{n-1}))$ au point $(x_n, f(x_n))$. En effet, celle-ci a pour équation

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

et elle rencontre donc l'axe des x au point d'abscisse x_{n+1}.



La méthode de la sécante consiste à partir de deux approximations x_0 et x_1 distinctes d'un zéro λ de f et à construire les approximations $(x_{n+1})_{n\geq 1}$ de proche en proche au moyen de la formule (*), en s'arrêtant éventuellement si $f(x_n) = f(x_{n-1})$.

L'étude de la convergence de cette méthode est un peu plus délicate que celle de la méthode de Newton.

Commençons par estimer l'écart entre le graphe d'une fonction et la sécante passant par deux des points de ce graphe.

Lemme 2.1.7. *Soit f une fonction de classe C_2 sur $]a, b[$ et soient x_0 et x_1 des points de $]a, b[$ tels que $f(x_0) \neq f(x_1)$. Posons*

$$f[x_0, x_1] = \frac{f(x_0) - f(x_1)}{x_0 - x_1}$$

et

$$g(x) = f(x_0) + (x - x_0)f[x_0, x_1].$$

Alors le graphe de $g(x)$ est la sécante au graphe de $f(x)$ passant par les points $(x_0, f(x_0))$ et $(x_1, f(x_1)$ et pour tout $x \in]a, b[$ il existe ξ dans le plus petit intervalle fermé J contenant x, x_0 et x_1 tel que

$$f(x) - g(x) = \frac{1}{2}(x - x_0)(x - x_1)f''(\xi).$$

Démonstration. Le résultat est clair si $x = x_0$ ou si $x = x_1$. Supposons donc que $x \neq x_0$ et que $x \neq x_1$. Posons

$$h(t) = f(t) - g(t) - (t - x_0)(t - x_1)\rho$$

avec ρ choisi pour que $h(x) = 0$. La fonction h s'annule alors en x, x_0 et x_1. Il s'ensuit que h' s'annule en deux points distincts de J. La fonction h'' s'annule donc en un point ξ de J. Comme

$$h''(t) = f''(t) - 2\rho$$

la conclusion est immédiate. □

Proposition 2.1.8. *Soit f une fonction de classe C_2 sur l'intervalle ouvert I de $\mathbb{R}$ et soit $\lambda \in I$ un zéro de f tel que $f'(\lambda) \neq 0$. Alors il existe $\delta > 0$ tel que la méthode de la sécante associée à des conditions initiales distinctes x_0, $x_1 \in]\lambda - \delta, \lambda + \delta[$ s'arrête avec $x_n = \lambda$ ou fournit une suite x_n qui converge au moins linéairement vers λ. Si de plus $f''(\lambda) \neq 0$ et si $x_0 \neq \lambda$ et $x_1 \neq \lambda$ sont suffisamment proche de λ alors $x_n \neq \lambda$ pour tout $n \geq 0$ et on a*

$$\frac{|x_{n+1} - \lambda|}{|x_n - \lambda|^p} \to \left| \frac{f''(\lambda)}{2f'(\lambda)} \right|^{1/p}$$

pour $p = (1 + \sqrt{5})/2$.

Démonstration. Choisissons d'abord $\delta > 0$ de sorte que f' ne s'annule pas sur $]\lambda - \delta, \lambda + \delta[$ et supposons avoir obtenu par la méthode de la sécante une suite finie $x_0, \ldots, x_n$ d'approximations de λ appartenant à $]\lambda - \delta, \lambda + \delta[$ et essayons de déterminer l'approximation x_{n+1}.

Si $x_n = x_{n-1}$, c'est impossible et la méthode s'arrête. Cependant, dans ce cas la relation

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f[x_{n-1}, x_{n-2}]}$$

montre que $f(x_{n-1}) = 0$ et par conséquent que $x_n = x_{n-1} = \lambda$.

Si $x_n \neq x_{n-1}$, on a $f(x_n) \neq f(x_{n-1})$ et

$$x_{n+1} = x_n - \frac{f(x_n)}{f[x_n, x_{n-1}]}$$

est bien défini. Vu le lemme précédent, il existe ξ dans le plus petit intervalle fermé contenant λ, x_n et x_{n-1} tel que

$$f(\lambda) - f(x_n) - (\lambda - x_n) f[x_n, x_{n-1}] = (\lambda - x_n)(\lambda - x_{n-1}) \frac{f''(\xi)}{2}.$$

Comme $f(\lambda) = 0$, on en tire que

$$\left(x_n - \frac{f(x_n)}{f[x_n, x_{n-1}]} - \lambda \right) f[x_n, x_{n-1}] = (\lambda - x_n)(\lambda - x_{n-1}) \frac{f''(\xi)}{2}.$$

Par conséquent,

$$(x_{n+1} - \lambda) f[x_n, x_{n-1}] = (\lambda - x_n)(\lambda - x_{n-1}) \frac{f''(\xi)}{2}.$$

Ainsi, il existe $\xi' \in]x_n, x_{n-1}[$ tel que

$$\Delta_{n+1} = \Delta_n \Delta_{n-1} \frac{f''(\xi)}{2f'(\xi')}. \qquad (*)$$

Choisissons

$$C > \left| \frac{f''(\lambda)}{2f'(\lambda)} \right|$$

et réduisons éventuellement δ pour que $\delta C < 1/2$ et que

$$\left| \frac{f''(\xi)}{2f'(\xi')} \right| < C$$

pour tous ξ, $\xi' \in]\lambda - \delta, \lambda + \delta[$. Dans ces conditions, on a

$$|\Delta_{n+1}| < C|\Delta_n||\Delta_{n-1}| < C\delta|\Delta_n| < \frac{1}{2}|\Delta_n|;$$

ce qui montre en particulier que $x_{n+1} \in]\lambda - \delta, \lambda + \delta[$.

En itérant le raisonnement ci-dessus, on voit que si l'on part de deux conditions initiales distinctes x_0, $x_1 \in]\lambda - \delta, \lambda + \delta[$ alors la méthode de la sécante s'arrête avec $x_n = \lambda$ ou fournit une suite x_n de $]\lambda - \delta, \lambda + \delta[$ qui converge au moins linéairement vers λ.

Si $f''(\lambda) \neq 0$, il est possible en réduisant éventuellement $\delta > 0$ de faire en sorte que f'' ne s'annule pas sur $]\lambda - \delta, \lambda + \delta[$. Dans ce cas la relation (*) montre que $x_{n+1} \neq \lambda$ si $x_n \neq \lambda$ et si $x_{n-1} \neq \lambda$. On en tire par récurrence que $x_n \neq \lambda$ quelque soit $n \geq 0$. Posons

$$a_n = \left| \frac{\Delta_{n+1}}{\Delta_n \Delta_{n-1}} \right|.$$

D'après ce qui précède,

$$a_n \to \left| \frac{f''(\lambda)}{2f'(\lambda)} \right|.$$

Soit $p > 0$ tel que $p - 1/p = 1$ (i.e. $p = (1 + \sqrt{5})/2$). On a

$$\frac{|\Delta_{n+1}|}{|\Delta_n|^p} = a_n \left(\frac{|\Delta_n|}{|\Delta_{n-1}|^p} \right)^{-1/p}.$$

Posons

$$b_{n+1} = \frac{|\Delta_{n+1}|}{|\Delta_n|^p}.$$

Comme

$$b_{n+1} = a_n b_n^{-1/p}$$

on voit que

$$b_{n+1} = a_n a_{n-1}^{-1/p} a_{n-2}^{1/p^2} \dots a_1^{(-1)^{n-1}/p^{n-1}} b_1^{(-1)^n/p^n}.$$

Ainsi,

$$\begin{aligned}\ln(b_{n+1}) &= \frac{(-1)^n}{p^n}\ln(b_1) + \sum_{k=0}^{n-1}\frac{(-1)^k}{p^k}\ln(a_{n-k}) \\ &= \frac{(-1)^n}{p^n}\ln(b_1) + \int_0^\infty \sum_{k=0}^{n-1}\frac{(-1)^k}{p^k}\ln(a_{n-k})\chi_{]k,k+1]}(t)\,dt\end{aligned}$$

et une application du théorème de Lebesgue montre que

$$\ln(b_{n+1}) \to \ln\left|\frac{f''(\lambda)}{2f'(\lambda)}\right| \sum_{k=0}^{\infty}\frac{(-1)^k}{p^k} = \frac{1}{1+\frac{1}{p}}\ln\left|\frac{f''(\lambda)}{2f'(\lambda)}\right|.$$

On en tire que

$$b_{n+1} \to \left|\frac{f''(\lambda)}{2f'(\lambda)}\right|^{1/p},$$

d'où la conclusion. □

Remarque 2.1.9. (a) Le nombre p est le nombre d'or. Ce nombre apparaît classiquement en cherchant à diviser une longueur l en deux parties $l_1 < l_2$ de sorte que $l/l_2 = l_2/l_1$. Puisque $l = l_1 + l_2$, cette condition peut en effet se réécrire $1 + l_1/l_2 = l_2/l_1$ ce qui entraîne que $p = l_2/l_1$.

(b) Si $f''(\lambda) \neq 0$, la relation $p \approx 1.618$ montre que la méthode de la sécante a une vitesse de convergence intermédiaire entre celle d'une méthode linéaire et celle d'une méthode quadratique.

(c) Si $f''(\lambda) = 0$, on peut montrer que la convergence de la méthode de la sécante est en fait plus rapide que dans le cas où $f''(\lambda) \neq 0$.

2.1.5 Méthode de Steffensen

Soit f une fonction de classe C_1 sur l'intervalle ouvert I de $\mathbb{R}$ et soit $\lambda \in I$ un zéro de f tel que $f'(\lambda) \neq 0$. Si la suite x_n converge vers λ, on a $f(x_n) \to f(\lambda) = 0$; il est donc naturel d'essayer d'approcher la dérivée $f'(x_n)$ utilisée dans la méthode de Newton par le quotient

$$\frac{f(x_n + f(x_n)) - f(x_n)}{(x_n + f(x_n)) - x_n} = \frac{f(x_n + f(x_n)) - f(x_n)}{f(x_n)}.$$

On arrive alors à la méthode itérative dite de *Steffensen* définie par

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

et une condition initiale x_0. Remarquons que puisque

$$\frac{f(x_n + f(x_n)) - f(x_n)}{f(x_n)} = f'(x_n + \theta f(x_n))$$

pour un certain $\theta \in]0,1[$, cette expression est non nulle pour x voisin de λ et distinct de λ. La fonction

$$\varphi(x) = \begin{cases} x - \dfrac{f(x)^2}{f(x + f(x)) - f(x)} & \text{si } x \neq \lambda \\ \lambda & \text{si } x = \lambda \end{cases}$$

est donc bien définie pour x voisin de λ et la méthode de Steffensen est la méthode itérative associée à cette fonction. Pour calculer $\varphi'(x)$, remarquons que

$$\varphi(x) = x - \frac{f(x)}{g(x)} \qquad (*)$$

où $g(x)$ est la fonction définie par

$$g(x) = \int_0^1 f'(x + \theta f(x))\, d\theta. \qquad (**)$$

En effet, on a

$$\frac{\partial}{\partial \theta} f(x + \theta f(x)) = f'(x + \theta f(x)) f(x)$$

et il découle de cette formule que

$$f(x + f(x)) - f(x) = g(x) f(x).$$

De (**), on tire que $g(\lambda) = f'(\lambda)$ et que

$$g'(x) = \int_0^1 f''(x + \theta f(x))(1 + \theta f'(x))\, d\theta$$

si f est de classe C_2 au voisinage de λ. Dans ce cas, φ est de classe C_1 au voisinage de λ et on a

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

Il en résulte en particulier que $\varphi'(\lambda) = 0$ et par conséquent que la méthode de Steffensen converge si x_0 est suffisamment proche de λ. Si f est de classe C_3, alors g et φ sont de classe C_2 et on a

$$\varphi'' = -\frac{g(gf'' - fg'') - (gf' - fg')2g'}{g^3}.$$

Il s'ensuit que

$$\varphi''(\lambda) = -\frac{f''(\lambda) - 2g'(\lambda)}{f'(\lambda)}.$$

Or,

$$g'(\lambda) = f''(\lambda)\left(1 + \frac{f'(\lambda)}{2}\right).$$

Donc

$$\varphi''(\lambda) = \frac{f''(\lambda)(1 + f'(\lambda))}{f'(\lambda)}.$$

Vu la Remarque 2.1.3 (b), on en tire que

$$\lim_{n \to +\infty} \frac{(x_{n+1} - \lambda)}{(x_n - \lambda)^2} = \frac{f''(\lambda)}{2f'(\lambda)}(1 + f'(\lambda)).$$

La convergence est donc exactement d'ordre 2 si $f''(\lambda) \neq 0$.

La méthode de Steffensen est donc une méthode quadratique comme la méthode de Newton, mais elle remplace à chaque itération l'évaluation de f' par une double évaluation de f. Si le calcul de f est un peu lourd, cette méthode peut être moins efficace que la sécante car le travail pour effectuer une étape de Steffensen correspond à peu près à celui effectué lors de deux étapes de la sécante. Ces deux itérations de la sécante correspondent à une méthode d'ordre $p^2 = 1 + p \simeq 2.618$.

2.2 Zéros réels des polynômes réels

Intéressons-nous à présent au calcul approché des zéros réels d'un polynôme à coefficients réels

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

En utilisant la structure particulière de P, nous allons déterminer un intervalle contenant toutes les racines réelles positives (resp. négatives) de l'équation $P(x) = 0$ (en d'autres termes, nous allons localiser les zéros positifs (resp. négatifs) de P). Nous allons également déterminer le nombre de zéros de P dans un semi-intervalle fini de $\mathbb{R}$. Partant de ces informations, une variante de la méthode de la bissection permet de trouver des intervalles $I_1, \ldots, I_r$ contenant chacun une seule racine de P (on dit alors qu'on a séparé les zéros de P). On peut ensuite poursuivre la bissection de chacun

de ces intervalles pour trouver avec la précision voulue le zéro de P qu'il contient. Lorsque la précision atteinte est suffisante, on peut aussi utiliser une méthode comme la méthode de Newton pour affiner rapidement les valeurs approchées obtenues.

2.2.1 Localisation des racines

Parmi les nombreux résultats connus, nous ne présenterons que le suivant :

Théorème 2.2.1 (Lagrange)**.** *Supposons que $a_n > 0$ et que les a_k ne sont pas tous positifs ou nuls. Soit $K < n$ le plus grand naturel tel que $a_K < 0$. Posons*

$$A = \max\{|a_k| : a_k < 0\}.$$

Alors, tout zéro $x > 0$ de P est tel que

$$x < 1 + \sqrt[n-K]{A/a_n}.$$

Démonstration. Supposons que

$$x \geq 1 + \sqrt[n-K]{A/a_n}.$$

Si dans l'expression

$$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$$

on remplace $a_{n-1}, \ldots, a_{K+1}$ par 0 et chaque $a_K, \ldots, a_0$ par $-A$, on obtient une expression plus petite. Ainsi,

$$P(x) \geq a_n x^n - Ax^K - Ax^{K-1} - \cdots - Ax - A \geq a_n x^n - A\frac{x^{K+1}-1}{x-1}.$$

Il en résulte que

$$\begin{aligned} P(x) &> a_n x^n - A\frac{x^{K+1}}{x-1} \\ &> \frac{a_n x^{K+1}}{x-1}\left(x^{n-K-1}(x-1) - \frac{A}{a_n}\right) \\ &> \frac{a_n x^{K+1}}{x-1}\left((x-1)^{n-K} - \frac{A}{a_n}\right). \end{aligned}$$

Or, $(x-1)^{n-K} \geq A/a_n$. Donc, $P(x) > 0$ et la conclusion en découle. □

Remarque 2.2.2. (a) Si tous les a_k $(k < n)$ sont positifs ou nuls, alors

$$P(x) \geq a_n x^n$$

pour $x \geq 0$ et P ne peut avoir de zéro > 0.

(b) Si on suppose simplement $a_n \neq 0$, alors la borne supérieure devient

$$1 + \sqrt[n-K]{A/|a_n|}$$

avec

$$A = \max\{|a_k| : a_k a_n < 0\}$$

s'il existe $k < n$ tel que $a_k a_n < 0$.

Corollaire 2.2.3. *Supposons que $a_0 \neq 0$ et qu'il existe k tel que $a_k a_0 < 0$. Alors, tout zéro $x > 0$ de P est tel que*

$$x > \frac{1}{1 + \sqrt[L]{B/|a_0|}}$$

où

$$L = \min\{k : a_k a_0 < 0\}, \qquad B = \max\{|a_k| : a_k a_0 < 0\}.$$

Démonstration. Il suffit d'appliquer le théorème de Lagrange au polynôme

$$Q(x) = x^n P(1/x).$$

□

Remarque 2.2.4. En utilisant les deux résultats précédents pour

$$R(x) = P(-x)$$

on obtient un encadrement des racines négatives de P.

Exemple 2.2.5. Considérons le polynôme

$$P(x) = x^3 + x^2 - 2x - 3.$$

On a

$$K = 1, \qquad A = 3.$$

Donc les racines strictement positives de P sont inférieures à

$$1 + \sqrt{3} \approx 2.73.$$

On a aussi

$$L = 2, \qquad B = 1.$$

Donc les racines strictement positives de P sont supérieures à

$$1/(1 + \sqrt{1/3}) \approx 0.63.$$

En fait, ce polynôme possède un seul zéro réel $x \approx 1.55$.

2.2.2 Nombre de racines réelles

Dans la suite, nous supposerons que P n'a que des zéros simples (si ce n'est pas le cas, on peut s'y ramener en divisant P par le PGCD de P et de P'). Cette hypothèse signifie aussi que P et P' sont premiers entre eux. Considérons la suite de polynômes de degrés strictement décroissants $(P_k)_{k\geq 0}$ définie par la relation de récurrence

$$P_{k-1} = Q_k P_k - P_{k+1} \qquad (k \geq 1)$$

et les conditions initiales $P_0 = P$, $P_1 = -P'$. Soit P_K le dernier P_k non nul. L'algorithme d'Euclide montre que P_K est à un multiple non nul près le PGCD de P et P' ; c'est donc un polynôme constant. Par construction, la suite $P_0, \ldots, P_K$ est une *suite de Sturm* c'est-à-dire qu'elle possède les propriétés suivantes :

(a) les zéros réels de P_0 sont simples ;

(b) $\operatorname{sgn} P_1(\xi) = -\operatorname{sgn} P_0'(\xi)$ si ξ est un zéro réel de P_0 ;

(c) $P_{k+1}(\xi)P_{k-1}(\xi) < 0$ si ξ est un zéro réel de $P_k(\xi)$;

(d) P_K n'a pas de zéro réel.

Définition 2.2.6. Le *nombre de changements de signes* d'une suite réelle

$$x_1, \ldots, x_K$$

s'obtient en éliminant les zéros et en comptant le nombre de fois qu'un nombre est suivi par un nombre de signe opposé.

Exemple 2.2.7. La suite

$$1, -1, -2, 0, -2, 1, 0, -1, 1$$

présente 4 changements de signes. La suite

$$1, -1, 2, -2, 0, 0, 2, 1$$

en présente aussi 4.

Proposition 2.2.8. *Soit*

$$P_0, \ldots, P_K$$

une suite de Sturm. Pour tout $x \in \mathbb{R}$, notons $N(x)$ le nombre de changements de signes dans la suite

$$P_0(x), \ldots, P_K(x).$$

Alors, le nombre de zéros réels de P_0 dans le semi-intervalle fini $[a, b[\subset \mathbb{R}$ est

$$N(b) - N(a).$$

Démonstration. Voyons comment $N(x)$ varie avec x.

Si aucun P_k ne change de signe en x, ces polynômes ont un signe constant au voisinage de x et $N(x)$ est donc constant au voisinage de x. Si P_0 change de signe en x, alors $P_0(x) = 0$. Comme les zéros réels de P_0 sont simples, $P_0'(x) \neq 0$. Or, $\operatorname{sgn} P_1(x) = -\operatorname{sgn} P_0'(x)$. Il s'ensuit que $P_1(x) \neq 0$ et que pour $h > 0$ petit, on est dans un des cas suivants :

	$x-h$	x	$x+h$
P_0	$-$	0	$+$
P_1	$-$	$-$	$-$

	$x-h$	x	$x+h$
P_0	$+$	0	$-$
P_1	$+$	$+$	$+$

Le nombre de changements de signes de P_0, P_1 reste donc constant en passant de $x-h$ à x et augmente de 1 en passant de x à $x+h$.

Si P_k change de signe en x pour $k > 0$, alors $k < K$ et, pour $h > 0$ petit, les cas de figure suivants peuvent se présenter :

	$x-h$	x	$x+h$
P_{k-1}	$-$	$-$	$-$
P_k	$-$	0	$+$
P_{k+1}	$+$	$+$	$+$

	$x-h$	x	$x+h$
P_{k-1}	$+$	$+$	$+$
P_k	$-$	0	$+$
P_{k+1}	$-$	$-$	$-$

	$x-h$	x	$x+h$
P_{k-1}	$-$	$-$	$-$
P_k	$+$	0	$-$
P_{k+1}	$+$	$+$	$+$

	$x-h$	x	$x+h$
P_{k-1}	$+$	$+$	$+$
P_k	$+$	0	$-$
P_{k+1}	$-$	$-$	$-$

Ainsi, le nombre de changements de signes de la suite P_{k-1}, P_k, P_{k+1} reste constant en passant de $x-h$ à x et à $x+h$.

Il résulte de ces remarques que pour $h > 0$ suffisamment petit, on a

$$N(x-h) = N(x) \qquad \text{et} \qquad N(x+h) = N(x)$$

si $P_0(x) \neq 0$ et

$$N(x-h) = N(x) \qquad \text{et} \qquad N(x+h) = N(x) + 1$$

si $P_0(x) = 0$. La conclusion en résulte. □

Exemple 2.2.9. Soit $P(x) = x^5 - x^3 - x - 1$. La suite de Sturm canonique de P

est

$$\begin{aligned}
P_0(x) &= x^5 - x^3 - x - 1 \\
P_1(x) &= -5x^4 + 3x^2 + 1 \\
P_2(x) &= \frac{2}{5}x^3 + \frac{4}{5}x + 1 \\
P_3(x) &= -13x^2 - \frac{25}{2}x - 1 \\
P_4(x) &= -\frac{385}{338}x - \frac{174}{169} \\
P_5(x) &= \frac{47827}{148225}.
\end{aligned}$$

Ainsi, $K = 5$ et on a le tableau de signes suivant :

x	$-\infty$	0	$+\infty$
$P_0(x)$	$-$	$-$	$+$
$P_1(x)$	$-$	$+$	$-$
$P_2(x)$	$-$	$+$	$+$
$P_3(x)$	$-$	$-$	$-$
$P_4(x)$	$+$	$-$	$-$
$P_5(x)$	$+$	$+$	$+$
$N(x)$	1	3	4

Le polynôme P possède donc deux zéros dans $]-\infty, 0[$ et un zéro dans $[0, +\infty[$. En utilisant le théorème de Lagrange, on voit que ces zéros sont en fait dans $]-2, -1/2[\cup]1/2, 2[$. Un calcul explicite donne les valeurs

$$x_0 = -1, \quad x_1 \approx -0.819173, \quad x_2 \approx 1.38028$$

pour les trois zéros réels de P.

2.2.3 Utilisation d'une suite de Sturm pour le calcul des racines

Bien que le calcul d'une suite de Sturm pour un polynôme P à zéros simples donné soit assez lourd et relativement instable numériquement, on peut si on dispose avec une bonne précision d'une telle suite, donner un algorithme de bissection fournissant tous les zéros réels de P à la précision voulue. Cet algorithme est basé sur le fait que si pour tout $\varepsilon > 0$ fixé, on peut associer à tout intervalle $]a, b]$ de longueur L une liste $I_1, \ldots, I_p$ éventuellement vide de semi-intervalles disjoints de longueur inférieure à ε telle que

(a) $I_1 \cup \cdots \cup I_p \subset [a, b[$ contient tous les zéros de P dans $[a, b[$,

(b) chaque I_k contient au moins un zéro de P,

alors on peut faire de même pour tout intervalle de longueur $2L$. En effet, soit $[a, b[$ de longueur $2L$ et soit $c = \frac{a+b}{2}$ le milieu de $[a, b[$. Comme $[a, c[$ est de longueur L, il existe une liste éventuellement vide $I_1, \ldots, I_r$ de semi-intervalles vérifiant les conditions (a), (b) pour $[a, c[$. De même, il existe une liste $I_{r+1}, \ldots, I_{r+s}$ éventuellement vide vérifiant les conditions (a),(b) pour $[c, b[$. La liste $I_1, \ldots, I_{r+s}$ vérifie alors les conditions (a), (b) pour $[a, b[$. Pour être sûr que cette procédure itérative de construction de la liste $I_1, \ldots, I_p$ s'arrête, il suffit de remarquer que lorsque $[a, b[$ est de longueur inférieure à ε, on peut lui associer la liste vide si $N(b) = N(a)$ et la liste $[a, b[$ si $N(b) > N(a)$. En fait, en général, on peut noter que la liste associée à $[a, b[$ est toujours vide si $N(b) = N(a)$ et utiliser cette remarque pour accélérer le processus. On remarquera que la méthode ci-dessus combinée avec le théorème de Lagrange permet de déterminer à la précision voulue tous les zéros réels de P. La vitesse de convergence est en général seulement linéaire. Aussi a-t-on intérêt à ne l'utiliser que pour obtenir des premières approximations des zéros. On raffinera ensuite ces approximations avec des méthodes plus rapides mais non globalement convergentes telles que la méthode de Newton ou la méthode de la sécante.

2.2.4 Évaluation d'un polynôme et de ses dérivées

Dans ce qui précède, on a eu besoin de calculer très souvent les valeurs d'un polynôme et de ses dérivées en $x_0 \in \mathbb{R}$. Passons en revue les méthodes possibles de calcul de $P(x_0)$ et évaluons dans chaque cas le nombre d'opérations nécessaires.

(a) Pour calculer $P(x_0) = a_0 + a_1 x_0 + \cdots + a_n x_0^n$, on peut être tenté d'additionner les termes dans l'ordre naturel en calculant chaque fois x_0^n au moyen de la fonction puissance n connue de la plupart des machines. Si on procède de la sorte, on effectue n additions et

$$1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}$$

multiplications puisque la fonction puissance n calcule généralement x_0^n sous la forme

$$\underbrace{x_0 \cdot x_0 \cdots x_0}_{n \text{ fois}}.$$

On a donc en tout $\dfrac{n(n+3)}{2}$ opérations.

(b) On peut reprendre le procédé précédent mais en calculant x_0^n progressivement par la récurrence

$$x_0^{n+1} = x_0 \cdot x_0^n.$$

De la sorte, on a en tout n additions et $1 + 2(n-1) = 2n - 1$ multiplications, soit $3n - 1$ opérations.

(c) Enfin, on peut utiliser la méthode de Horner. Rappelons que pour diviser $P(x)$ par $x - x_0$, on forme le tableau

$$\begin{array}{c|cccc} & a_n & a_{n-1} & \cdots & a_0 \\ x_0 & & & & \\ \hline & b_n & b_{n-1} & \cdots & b_0 \end{array}$$

où $b_n = a_n$ et $b_k = a_k + b_{k+1}x_0$. Dans ces conditions, on a

$$P(x) = (b_n x^{n-1} + \cdots + b_2 x + b_1)(x - x_0) + b_0.$$

En particulier, $b_0 = P(x_0)$. Cette méthode de calcul de $P(x_0)$ utilise n multiplications et n additions, soit $2n$ opérations. Elle est donc plus efficace que les deux méthodes précédentes. La méthode de Horner permet aussi de calculer aisément $P^{(k)}(x_0)$. En effet, si on pose

$$P_1(x) = b_n x^{n-1} + \cdots + b_1,$$

alors on peut calculer $P_1(x_0)$ par Horner. De la relation

$$\frac{P(x) - P(x_0)}{x - x_0} = P_1(x) \qquad (x \neq x_0),$$

on tire que

$$P'(x_0) = P_1(x_0).$$

Il s'ensuit que la méthode de Horner est particulièrement efficace en cas d'utilisation de la méthode de Newton pour approcher un zéro d'un polynôme.

Plus généralement, en appliquant itérativement la méthode de Horner, on obtient une suite $P_1, \ldots, P_K$ de polynômes telle que

$$P_k(x) = P_{k+1}(x)(x - x_0) + P_k(x_0)$$

et on vérifie par récurrence descendante sur k que

$$P_k(x) = \sum_{l=k}^{n} P_l(x_0)(x - x_0)^{l-k}.$$

Il en résulte que

$$P(x) = \sum_{l=0}^{n} P_l(x_0)(x - x_0)^l$$

et la formule de Taylor montre que

$$P^{(k)}(x_0) = k!P_k(x_0).$$

Par exemple, si $P(x) = x^3 - x^2 + x + 1$, on obtient

	1	−1	1	1
1		1	0	1
	1	0	1	2
1		1	1	
	1	1	2	
1		1		
	1	2		
1				
	1			

On a donc

$$P(1) = 2, \quad P'(1) = 2, \quad P''(1) = 2 \cdot 2 = 4, \quad P'''(1) = 6 \cdot 1 = 6.$$

2.3 Précision et stabilité des zéros réels

Pour estimer la précision d'une approximation d'un zéro simple d'une fonction d'une variable réelle, on dispose de la majoration suivante :

Proposition 2.3.1. *Soit f une fonction de classe C_1 sur l'intervalle $[a, b]$ de $\mathbb{R}$ dont la dérivée ne s'annule pas sur cet intervalle. Soit $x_0 \in [a, b]$ un zéro de f et soit $\tilde{x_0} \in [a, b]$ une valeur approchée de ce zéro. Alors,*

$$|\Delta x_0| \leq \frac{1}{\inf_{x\in[a,b]} |f'(x)|} |f(\tilde{x_0})|.$$

Démonstration. La formule des accroissements finis montre que

$$f(\tilde{x_0}) - f(x_0) = (\tilde{x_0} - x_0)f'(\xi)$$

où ξ est un point de $[a, b]$ situé entre x_0 et $\tilde{x_0}$. Comme $f(x_0) = 0$, la conclusion en résulte aisément. □

Le nombre de conditionnement d'un zéro simple d'une fonction d'une variable réelle par rapport à un paramètre dont dépend cette fonction peut s'obtenir facilement grâce au théorème des fonctions implicites :

Proposition 2.3.2. *Soit $f(x,\lambda)$ une fonction de classe C_1 au voisinage de $(x_0,\lambda_0) \in \mathbb{R}^2$. Supposons que x_0 soit un zéro simple de $f(x,\lambda_0)$. Alors, il existe un intervalle ouvert $I \ni x_0$ et un intervalle ouvert $J \ni \lambda_0$ tels que pour tout $\lambda \in J$,*

(a) la fonction $f(x,\lambda)$ possède un seul zéro $\varphi(\lambda)$ dans I ;

(b) ce zéro $\varphi(\lambda)$ est simple ;

(c) l'application

$$\varphi : J \to I$$

est de classe C_1 et on a

$$\varphi(\lambda_0) = x_0, \qquad \varphi'(\lambda_0) = -\frac{\frac{\partial f}{\partial \lambda}(x_0,\lambda_0)}{\frac{\partial f}{\partial x}(x_0,\lambda_0)}.$$

En particulier, pour $x_0 \neq 0$, le nombre de conditionnement de φ en $\lambda_0 \neq 0$ est

$$\eta(\varphi) = \left| \frac{\frac{\partial f}{\partial \lambda}(x_0,\lambda_0)}{\frac{\partial f}{\partial x}(x_0,\lambda_0)} \, \frac{\lambda_0}{x_0} \right|.$$

Démonstration. Comme

$$\frac{\partial f}{\partial x}(x_0,\lambda_0) \neq 0$$

le théorème des fonctions implicites montre qu'il existe des intervalles ouverts I, J contenant x_0 et λ_0 et une application $\varphi : J \to I$ de classe C_1 tels que

$$\{(x,\lambda) \in I \times J : f(x,\lambda) = 0\} = \{(\varphi(\lambda),\lambda) : \lambda \in J\}.$$

Quitte à restreindre I et J, on peut même supposer que

$$\frac{\partial f}{\partial x}(x,\lambda) \neq 0$$

si $(x,\lambda) \in I \times J$. Cela étant, pour tout $\lambda \in J$, $\varphi(\lambda)$ est le seul zéro de $f(x,\lambda)$ dans I et que ce zéro est simple. Il s'ensuit que $\varphi(\lambda_0) = x_0$ et en dérivant l'égalité $f(\varphi(\lambda),\lambda) = 0$ par rapport à λ en λ_0, on voit que

$$\frac{\partial f}{\partial x}(x_0,\lambda_0)\varphi'(\lambda_0) + \frac{\partial f}{\partial \lambda}(x_0,\lambda_0) = 0.$$

La conclusion en découle. □

Corollaire 2.3.3. *Le nombre de conditionnement d'un zéro simple non nul x_0 du polynôme réel*

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

par rapport au coefficient a_j est

$$\kappa(j, x_0) = \left| \frac{a_j x_0^{j-1}}{P'(x_0)} \right|.$$

Démonstration. On a

$$\frac{\partial P}{\partial a_j} = x^j$$

et

$$\frac{\partial P}{\partial x} = P'(x).$$

La conclusion en résulte. □

Exemple 2.3.4. Considérons le polynôme de Wilkinson

$$P(x) = (x-1)(x-2)\dots(x-20).$$

Le coefficient de x^{19} dans ce polynôme est $-1-2\cdots-20 = -\frac{21\cdot 20}{2} = -210$. On a aussi $P'(20) = 19!$, donc

$$\kappa(19, 20) = \frac{210 \cdot 20^{18}}{19!} \approx 4.52548\ 10^8.$$

Pour ε petit, une modification de $\varepsilon\%$ de a_{19} engendre donc une modification d'environ

$$4.52548\ 10^8\ \varepsilon\%$$

du zéro $x_0 = 20$ de P. Ce zéro est donc très mal conditionné. A titre d'exemple, donnons à 12 décimales ce qu'il devient lorsque a_{19} est remplacé par $a_{19} + 10^{-t}$ pour quelques valeurs de t :

t	x_0
12	19.999956895369
11	19.999568518530
10	19.995640872330
9	19.950949654306

Lorsque $t = 7$, x_0 "passe" dans le domaine complexe et vaut approximativement

$$19.8694 + 0.4832\ i.$$

Cela ne contredit pas la Proposition 2.3.2 car dans le cas considéré ici, λ est "sorti" de J.

Remarque 2.3.5. Dans le cas d'un zéro multiple, les choses sont plus compliquées. En général, on ne peut assurer que si $f(x_0, \lambda_0) = 0$, alors $f(x, \lambda)$ possède un zéro proche de x_0 si λ est proche de λ_0. On le voit facilement en considérant

$$f(x, \lambda) = x^2 - \lambda$$

avec $\lambda_0 = 0$. Cependant, si f est de classe C_p sur l'intervalle ouvert I et si $x_0 \in I$ est un zéro de multiplicité p de f, alors pour tout x voisin de x_0 on a

$$f(x) = \frac{(x - x_0)^p}{p!} f^{(p)}(\xi)$$

pour un certain $\xi \in]x, x_0[$. En particulier, si $|f(x)| < \varepsilon$, on a

$$|x - x_0| \leq \sqrt[p]{\frac{p!\varepsilon}{|f^{(p)}(\xi)]}},$$

ce qui laisse entrevoir que la sensibilité du zéro x_0 à une perturbation de f est encore plus grande que dans le cas des zéros simples.

3 Systèmes linéaires déterminés

3.1 Méthodes directes de résolution

Soit A une matrice carrée complexe non singulière de dimension n et B un vecteur à n composantes. Dans ce chapitre on se propose d'étudier la résolution numérique du système linéaire

$$AX = B \qquad (*)$$

par des méthodes qui donneraient la solution exacte X après un nombre fini d'étapes s'il était possible d'utiliser une arithmétique de précision infinie. Ces méthodes sont dites *directes* par opposition aux méthodes itératives que nous étudierons plus tard.

Parmi les méthodes directes figure la méthode de Cramer. Cette méthode est basée sur le fait que la solution X du système (*) est donnée par les relations

$$x_1 = \frac{\det\begin{pmatrix} B & C_2 & \cdots & C_n \end{pmatrix}}{\det A}, \ldots, x_n = \frac{\det\begin{pmatrix} C_1 & \cdots & C_{n-1} & B \end{pmatrix}}{\det A}$$

où $C_1, C_2, \ldots, C_n$ désignent les colonnes de A. Malgré leur intérêt théorique et leur simplicité apparente, ces formules sont cependant numériquement assez inefficaces. En effet, on vérifie aisément que le calcul du déterminant d'une matrice carrée de dimension n au moyen de sa définition requiert $(n-1)n!$ multiplications et $n!-1$ additions, soit en tout $n(n!)-1$ opérations élémentaires. Ainsi la méthode précédente requiert en tout $(n+1)[n(n!)-1]+n = n(n+1)!-1 \sim n^2 n!$ opérations élémentaires. Pour $n = 10$, ce nombre est déjà d'environ 400 millions et la méthode considérée est donc sans intérêt même pour des n relativement petits. Heureusement, on peut résoudre le système (*) bien plus efficacement en utilisant, par exemple, la méthode de Gauss.

3.1.1 Méthode de Gauss

L'idée de cette méthode est d'éliminer progressivement les inconnues du système (*) pour se ramener à un système triangulaire de la forme

$$\left\{\begin{array}{rcl} \alpha_{11}x_{\mu_1} + \alpha_{12}x_{\mu_2} + \cdots + \alpha_{1n}x_{\mu_n} &=& \beta_1 \\ \alpha_{22}x_{\mu_2} + \cdots + \alpha_{2n}x_{\mu_n} &=& \beta_2 \\ \vdots & & \\ \alpha_{nn}x_{\mu_n} &=& \beta_n \end{array}\right. \qquad (**)$$

où μ est une permutation de $\{1, \ldots, n\}$. Puis de résoudre celui-ci en calculant d'abord x_{μ_n} puis $x_{\mu_{n-1}}$ et ainsi de suite jusqu'à x_{μ_1}.

Le procédé suivi pour ramener le système

$$\left\{\begin{array}{rcl} a_{11}x_1 + \cdots + a_{1n}x_n &=& b_1 \\ & \vdots & \\ a_{n1}x_1 + \cdots + a_{nn}x_n &=& b_n \end{array}\right.$$

à la forme (**) est le suivant. Comme A est inversible par hypothèse, A possède des éléments non nuls. La première étape de la méthode consiste à choisir l'un de ces éléments. Cet élément non-nul est appelé le *premier pivot*. On permute alors les lignes et les inconnues du système de façon à ce que ce pivot devienne le premier coefficient de la première ligne du système transformé. Celui-ci a alors la forme

$$\left\{\begin{array}{rcl} a'_{11}x'_1 + \cdots + a'_{1n}x'_n &=& b'_1 \\ & \vdots & \\ a'_{n1}x'_1 + \cdots + a'_{nn}x'_n &=& b'_n \end{array}\right.$$

avec $a'_{11} \neq 0$; les inconnues $x'_1, \ldots, x'_n$ étant les inconnues $x_1, \ldots, x_n$ de départ permutées de manière adéquate. On remplace ensuite ce système par le système équivalent obtenu en soustrayant à la $j^{\text{ème}}$ ligne la première ligne multipliée par $\dfrac{a'_{j1}}{a'_{11}}$. Le système obtenu a alors la forme

$$\left\{\begin{array}{rcl} a'_{11}x'_1 + a'_{12}x'_2 + \cdots + a'_{1n}x'_n &=& b'_1 \\ a''_{22}x'_2 + \cdots + a''_{2n}x'_n &=& b''_2 \\ & \vdots & \\ a''_{n2}x'_2 + \cdots + a''_{nn}x'_n &=& b''_n \end{array}\right.$$

et l'inconnue x'_1 a été éliminée des lignes 2 à n. Il suffit alors de procéder de même pour éliminer de proche en proche les autres inconnues.

Pour transformer la méthode dont il a été question ci-dessus en un algorithme implémentable en machine, il faut encore préciser la manière dont on va choisir les pivots successifs. Notons $A^{(m)}$ la matrice des coefficients du système avant l'étape m. Les trois possibilités les plus courantes sont de choisir comme pivot

(i) $a^{(m)}_{mm}$;

(ii) $a^{(m)}_{jm}$ tel que

$$|a^{(m)}_{jm}| = \sup_{m \leq j \leq n} |a^{(m)}_{jm}|;$$

(iii) $a^{(m)}_{jk}$ tel que

$$|a^{(m)}_{jk}| = \sup_{\substack{m \leq j \leq n \\ m \leq k \leq n}} |a^{(m)}_{jk}|.$$

Les algorithmes obtenus s'appellent respectivement *algorithme de Gauss sans pivotage*, *algorithme de Gauss avec pivotage partiel* et *algorithme de Gauss avec pivotage total*.

Étudions d'abord le premier de ces algorithmes.

3.1.2 Algorithme de Gauss sans pivotage

Proposition 3.1.1. *Pour que l'algorithme de Gauss sans pivotage fonctionne pour le système $AX = B$ il est nécessaire et suffisant que tous les mineurs principaux dominants de A soient non nuls.*

Démonstration. Supposons d'abord que l'algorithme de Gauss sans pivotage fonctionne pour le système $AX = B$. Alors, $a_{mm}^{(m)} \neq 0$ pour tout $m \in \{1, \ldots, n\}$ et comme les différentes étapes s'effectuent sans permutation de lignes ou d'inconnues, les propriétés des déterminants montrent que

$$\begin{vmatrix} a_{11}^{(m)} & \cdots & \cdots & a_{1m}^{(m)} \\ & a_{22}^{(m)} & \cdots & a_{2m}^{(m)} \\ & & \ddots & \vdots \\ & & & a_{mm}^{(m)} \end{vmatrix} = \begin{vmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mm} \end{vmatrix}.$$

Il s'ensuit que les mineurs principaux de A sont non nuls. Réciproquement, si les mineurs principaux de A sont non nuls on déduit de proche en proche des égalités précédentes que $a_{mm}^{(m)} \neq 0$ pour $m = 1, 2, \ldots, n$. □

Remarque 3.1.2. La condition trouvée dans la proposition précédente peut paraître très restrictive. Elle est cependant satisfaite génériquement car l'ensemble des matrices carrées qui ont tous leurs mineurs principaux non nuls est un ouvert dense de $\mathbb{C}_n^n$. En effet, soit $A \in \mathbb{C}_n^n$ et soient $M_1, \ldots, M_n$ les sous-matrices principales de A. Comme les sous-matrices principales de $A - \lambda I$ sont $M_1 - \lambda I_1, M_2 - \lambda I_2, \ldots, M_n - \lambda I_n$, on sait que la matrice $A - \lambda I$ a tous ses mineurs principaux non nuls si et seulement si λ n'est valeur propre d'aucune des matrices $M_1, M_2, \ldots, M_n$. L'ensemble des valeurs propres de ces matrices étant fini, pour tout $\varepsilon > 0$, il existe $\lambda \in \mathbb{C}$ tel que $|\lambda| < \varepsilon$ et pour lequel $A - \lambda I$ a tous ses mineurs principaux non nuls.

Proposition 3.1.3. *Le nombre d'opérations élémentaires effectuées pour résoudre le système $AX = B$ par l'algorithme de Gauss sans pivotage est*

$$\frac{4n^3 + 9n^2 - 7n}{6} \sim \frac{2}{3}n^3.$$

Démonstration. Pour construire $(A^{(m+1)}, B^{(m+1)})$ à partir de $(A^{(m)}, B^{(m)})$, il faut pour chaque $j \in \{m+1, \ldots, n\}$, calculer le quotient

$$l_{jm}^{(m)} = \frac{a_{jm}^{(m)}}{a_{mm}^{(m)}}$$

puis soustraire à la j^{e} ligne de $(A^{(m)}, B^{(m)})$ le produit de la m^{e} ligne de $(A^{(m)}, B^{(m)})$ par $l_{jm}^{(m)}$. Parmi les éléments $a_{jk}^{(m+1)}$ obtenus, seuls ceux correspondant à $k \geq m+1$ sont éventuellement non nuls. Le passage de $(A^{(m)}, B^{(m)})$ à $(A^{(m+1)}, B^{(m+1)})$ requiert donc $(n-m)$ divisions , $(n-m)(n-m+1)$ multiplications, $(n-m)(n-m+1)$ soustractions. La phase d'élimination de l'algorithme de Gauss sans pivotage requiert donc en tout

$$\sum_{m=1}^{n-1}(n-m) = \sum_{k=1}^{n-1} k = \frac{n(n-1)}{2}$$

divisions,

$$\begin{aligned}\sum_{m=1}^{n-1}(n-m)(n-m+1) = \sum_{k=1}^{n-1} k(k+1) &= 2\left(1 + \sum_{k=2}^{n-1} \mathrm{C}_{k+1}^2\right) \\ &= 2\left(1 + \sum_{k=2}^{n-1} \mathrm{C}_{k+2}^3 - \mathrm{C}_{k+1}^3\right) = 2\,\mathrm{C}_{n+1}^3 \\ &= \frac{(n+1)n(n-1)}{3}\end{aligned}$$

multiplications et autant de soustractions. Pour résoudre le système triangulaire

$$A^{(n)}X = B^{(n)}$$

il faut encore

$$\begin{aligned}&\sum_{m=1}^{n}[1 \text{ divisions } + (n-m) \text{ multiplications } + (n-m) \text{ soustractions}] \\ &\qquad = n \text{ divisions } + \frac{n(n-1)}{2} \text{ multiplications } + \frac{n(n-1)}{2} \text{ soustractions.}\end{aligned}$$

La résolution complète de $AX = B$ demande donc

$$\frac{n(n+1)}{2} \text{ divisions }, \quad \frac{(2n+5)n(n-1)}{6} \text{ multiplications}$$

et autant de soustractions, soit en tout

$$\frac{4n^3 + 9n^2 - 7n}{6}$$

opérations élémentaires. □

Pour $n = 10$, le nombre d'opérations obtenu dans la proposition précédente est égal à 805. Ceci permet d'apprécier l'énorme différence d'efficacité entre la méthode de Gauss et la méthode de Cramer.

Remarque 3.1.4. (a) Comme on a

$$\det A = a_{11}^{(1)} a_{22}^{(2)} \cdots a_{nn}^{(n)},$$

on peut utiliser l'algorithme de Gauss sans pivotage pour le calcul de $\det A$. Un calcul rapide montre que ce procédé requiert

$$(n-1) + \sum_{m=1}^{n-1} (n-m)(2(n-m)+1) = \frac{4n^3 - 3n^2 + 5n - 6}{6} \sim \frac{2}{3} n^3$$

opérations élémentaires.

(b) On peut aussi utiliser l'algorithme de Gauss sans pivotage pour résoudre simultanément les systèmes $AX_1 = B_1, \ldots, AX_p = B_p$. Le nombre d'opérations élémentaires requises est alors

$$\sum_{m=1}^{n-1} (n-m) + 2 \sum_{m=1}^{n-1} (n-m)(n-m+p) + \sum_{m=1}^{n} p(1 + 2(n-m))$$

soit

$$\frac{4n^3 - 3n^2 - n + p(12n^2 - 6n)}{6} \sim \frac{2}{3} n^3 \qquad (p \text{ fixé}).$$

En particulier, l'inversion de A par cette méthode nécessite

$$\frac{16n^3 - 9n^2 - n}{6} \sim \frac{8}{3} n^3$$

opérations élémentaires.

3.1.3 Décomposition *LU*

Une conséquence intéressante de l'algorithme de Gauss sans pivotage est donnée dans la proposition suivante.

Proposition 3.1.5. *Soit $A \in \mathbb{C}_n^n$ une matrice carrée dont tous les mineurs principaux sont non nuls. Alors, il existe un unique couple (L, U) de matrices de $\mathbb{C}_n^n$ tel que*

(a) L est triangulaire inférieure et a ses éléments diagonaux égaux à 1 ;

(b) U est triangulaire supérieure ;

(c) $LU = A$.

Démonstration. Le passage de $A^{(m)}$ à $A^{(m+1)}$ dans l'algorithme de Gauss sans pivotage peut se traduire par les formules

$$a_{jk}^{(m+1)} = \begin{cases} a_{jk}^{(m)} & \text{si } j \leq m \\ a_{jk}^{(m)} - l_{jm} a_{mk}^{(m)} & \text{si } j > m \end{cases} \qquad (*)$$

où

$$l_{jm} = \frac{a_{jm}^{(m)}}{a_{mm}^{(m)}}$$

si $j > m$. Posons $l_{mm} = 1$ et $l_{jm} = 0$ si $j < m$. Alors, la matrice

$$L = (l_{jk})_{\substack{1 \leq j \leq n \\ 1 \leq k \leq n}}$$

est triangulaire inférieure et a des éléments diagonaux égaux à 1. Posons $U = A^{(n)}$. Par construction, U est triangulaire supérieure. La relation (*) montre que

$$a_{jk}^{(m)} = a_{jk}^{(n)}$$

si $j \leq m$ et que

$$a_{jk}^{(m)} - a_{jk}^{(m+1)} = l_{jm} a_{mk}^{(n)}$$

si $j > m$. En sommant, il vient

$$\sum_{m=1}^{j-1} a_{jk}^{(m)} - a_{jk}^{(m+1)} = \sum_{m=1}^{j-1} l_{jm} a_{mk}^{(n)}$$

d'où la relation

$$a_{jk}^{(1)} - a_{jk}^{(j)} = \sum_{m=1}^{j-1} l_{jm} a_{mk}^{(m)}$$

que l'on peut encore écrire

$$a_{jk} = a_{jk}^{(n)} + \sum_{n=1}^{j-1} l_{jm} a_{mk}^{(n)}$$

ou

$$A = LU.$$

Si le couple L', U' vérifie les conditions de l'énoncé, on a

$$L'U' = LU$$

et $L^{-1}L' = U(U')^{-1}$. La matrice $L^{-1}L'$ est donc à la fois triangulaire inférieure et triangulaire supérieure. Elle est donc diagonale et comme les éléments diagonaux de $L^{-1}L'$ sont égaux à 1, on a en fait

$$L^{-1}L' = I.$$

On en tire que $L = L'$ puis que $U = U'$, ce qui achève la démonstration. □

Remarque 3.1.6. (a) Comme les éléments $a_{jm}^{(m+1)}$ sont nuls pour $j > m$, on peut utiliser leurs places pour stocker les l_{jm} pour $j > m$ et réduire ainsi la place mémoire nécessaire.

(b) La connaissance d'une décomposition LU de la matrice A réduit la résolution d'un système de la forme $AX = B$ à la résolution des deux systèmes triangulaires

$$\begin{aligned} LY &= B \\ UX &= Y \end{aligned}$$

Le premier demande

$$n(n-1)$$

opérations (on évite les divisions car $l_{jj} = 1$), le second demande

$$n^2$$

opérations. La résolution complète de $AX = B$ se fait donc en

$$2n^2 - n \sim 2n^2$$

opérations. Le calcul de la décomposition LU elle-même requiert quant à lui

$$\frac{4n^3 - 3n^2 - n}{6} \sim \frac{2}{3}n^3$$

opérations mais peut être fait une fois pour toutes.

(c) On peut aussi exploiter la décomposition LU de A pour calculer A^{-1}. On a en effet

$$A^{-1} = U^{-1}L^{-1}.$$

Or les calculs de U^{-1} et de L^{-1} sont très simples. Posons $V = U^{-1}$. L'équation $UV = I$ correspond aux égalités

$$\sum_{s=j}^{n} u_{js}v_{sk} = \delta_{jk} \qquad (j, k \in \{1, \ldots, n\}).$$

On en tire que

$$v_{jk} = \left(\delta_{jk} - \sum_{s=j+1}^{n} u_{js} v_{sk} \right) / u_{jj}$$

ce qui permet de calculer les lignes de V par récurrence descendante sur leur indice. Comme V est aussi triangulaire supérieure, le calcul complet requiert

$$\sum_{j=1}^{n} (n-j+1)(1+2(n-j)) = \sum_{k=0}^{n-1} (k+1)(2k+1) = \frac{4n^3+3n^2-n}{6} \sim \frac{2}{3}n^3$$

opérations. Dans le calcul de L^{-1} par une méthode similaire, on gagne les divisions par les éléments diagonaux et on utilise donc

$$\sum_{j=1}^{n} 2(n-j)(n-j+1) = \sum_{k=0}^{n-1} 2k(k+1) = \frac{2n^3-2n}{3} \sim \frac{2}{3}n^3$$

opérations. Enfin, le produit U^{-1} par L^{-1} requiert

$$\sum_{j=1}^{n} \left(\sum_{k \leq j} (2(n-j)) + \sum_{k>j} (2(n-k)+1) \right) = \frac{4n^3-3n^2-n}{6} \sim \frac{2}{3}n^3$$

opérations. Ainsi, le calcul de A^{-1} par la méthode considérée demande en tout

$$2n^3 - n \sim 2n^3$$

opérations, ce qui est encore un peu mieux que pour la méthode considérée en Remarque 3.1.4 (b).

3.1.4 Algorithme de Gauss avec pivotage

Parmi les défauts de l'algorithme de Gauss sans pivotage, on trouve bien sûr celui de ne pas être applicable à toutes les matrices inversibles, mais surtout celui d'être parfois assez instable du point de vue numérique. Considérons par exemple le système

$$\left\{ \begin{array}{rcl} x_1 + x_2 + x_3 & = & 1 \\ x_1 + x_2 + 2x_3 & = & 2 \\ x_1 + 2x_2 + 2x_3 & = & 1 \end{array} \right.$$

qui est non singulier et a $x_1 = 1$, $x_2 = -1$, $x_3 = 1$ pour solution. Après la première phase de l'algorithme de Gauss sans pivotage, il devient

$$\left\{ \begin{array}{rcl} x_1 + x_2 + x_3 & = & 1 \\ x_3 & = & 1 \\ x_2 + x_3 & = & 0 \end{array} \right.$$

et on ne peut poursuivre avec $a_{22}^{(2)}$ comme pivot. De plus, si on perturbe légèrement le système de départ en remplaçant a_{22} par 1.0001 alors le système obtenu après la première étape de l'algorithme de Gauss sans pivotage devient

$$\left\{\begin{array}{rcl} x_1 + x_2 + x_3 & = & 1 \\ 0.0001x_2 + x_3 & = & 1 \\ x_2 + x_3 & = & 0 \end{array}\right.$$

Après la deuxième étape, on obtient donc le système

$$\left\{\begin{array}{rcl} x_1 + x_2 + x_3 & = & 1 \\ 0.0001x_2 + x_3 & = & 1 \\ -9999x_3 & = & -10000 \end{array}\right.$$

Si les solutions de ce système sont calculées en arithmétique flottante à trois chiffres décimaux significatifs, on obtient successivement

$$\tilde{x}_3 = 1, \quad \tilde{x}_2 = 0, \quad \tilde{x}_1 = 0$$

tandis que pour la vraie solution, on a

$$\begin{aligned} x_3 &= 1.0001 \\ x_2 &= -1.0001 \\ x_1 &= 1.0000 \end{aligned}$$

avec quatre décimales exactes. On a donc

$$|\delta x_3| \approx 1.10^{-4}, \quad |\delta x_2| \approx 1, \quad |\delta x_1| \approx 1.$$

Si l'on applique l'algorithme de Gauss avec pivotage partiel, on obtient à la seconde étape le système

$$\left\{\begin{array}{rcl} x_1 + x_2 + x_3 & = & 1 \\ x_2 + x_3 & = & 0 \\ 0.9999x_3 & = & 1 \end{array}\right.$$

et dans les mêmes conditions que ci-dessus, on trouve

$$\tilde{x}_3 = 1, \quad \tilde{x}_2 = -1, \quad \tilde{x}_1 = 1$$

d'où

$$|\delta x_3| \approx 1.10^{-4}, \quad |\delta x_2| \approx 1.10^{-4}, \quad |\delta x_1| < 1.10^{-5}$$

ce qui est beaucoup plus satisfaisant.

On peut montrer en général que l'utilisation du pivotage partiel permet de garder les erreurs relatives sur la solution dans des limites raisonnables. Le prix à payer est que le choix du pivot pour passer de $A^{(m)}$ à $A^{(m+1)}$ requiert $n-m$ comparaisons, ce qui augmente le nombre total d'opérations de

$$\sum_{m=1}^{n}(n-m) = \frac{n(n-1)}{2}$$

mais ne change pas le fait que ce nombre est asymptotiquement $\frac{2}{3}n^3$.

Le pivotage total améliore encore la stabilité de l'algorithme de Gauss. Cependant, dans ce cas, le prix à payer pour le choix du pivot utilisé pour passer de $A^{(m)}$ à $A^{(m+1)}$ est de

$$\sum_{m=1}^{n}((n-m+1)^2-1) \sim \frac{n^3}{3}$$

comparaisons. L'estimation asymptotique du nombre d'opérations passe donc de $\frac{2}{3}n^3$ à n^3. Cela peut engendrer un ralentissement significatif pour les grands systèmes et justifie le fait que l'algorithme de Gauss avec pivotage partiel est le plus souvent utilisé.

3.1.5 Décomposition LU avec pivotage

Tout ce qui a été dit plus haut concernant la décomposition LU peut s'adapter dans le cas des algorithmes de Gauss avec pivotage. Par exemple, dans le cas du pivotage partiel, on aboutit à des facteurs LU d'une matrice de la forme PA où P est une matrice de permutation qui tient compte des échanges de lignes effectués. Nous laissons au lecteur intéressé le soin d'adapter en conséquence l'algorithme donné dans la preuve de la Proposition 3.1.5 et renvoyons à [**?**] pour plus de détails.

3.1.6 Méthode de Choleski

Considérons à présent le cas où la matrice A des coefficients du système étudié est hermitienne définie positive. Dans ce cas, les mineurs principaux de A sont tous strictement positifs et l'algorithme de Gauss sans pivotage fonctionne pour A et conduit à une décomposition de la forme $A = LU$ avec L triangulaire inférieure à éléments diagonaux égaux à 1 et U triangulaire supérieure. Comme $A^* = A$, on a

$$U^*L^* = LU$$

et

$$L^{-1}U^* = U(L^*)^{-1}.$$

Il en résulte que $U(L^*)^{-1}$ est une matrice diagonale $D = \mathrm{diag}(d_1, \ldots, d_n)$. Comme les mineurs principaux de U sont égaux aux mineurs principaux de A et que ceux de L sont égaux à 1, la relation

$$U = DL^*$$

montre que $d_1 > 0, \ldots, d_n > 0$. Posons $\sqrt{D} = \mathrm{diag}(\sqrt{d_1}, \ldots, \sqrt{d_n})$. Il vient

$$A = LU = LDL^* = (L\sqrt{D})(L\sqrt{D})^*.$$

On voit donc qu'il existe une matrice R triangulaire supérieure à éléments diagonaux positifs telle que

$$A = R^*R. \qquad (*)$$

Une décomposition de ce type est appelée *décomposition de Choleski* de A. Elle a les mêmes avantages que la décomposition LU mais elle reflète en plus le caractère hermitien défini positif de A. De plus, ses éléments sont faciles à calculer de proche en proche. En effet, la relation (*) entraîne que

$$a_{jk} = \sum_{s=1}^{j} \overline{r}_{sj} r_{sk}$$

pour $j \leq k$. Ainsi, on a

(a) $r_{jj} = \sqrt{a_{jj} - \sum_{s=1}^{j-1} |r_{sj}|^2}$,

(b) $r_{jk} = \dfrac{a_{jk} - \sum_{s=1}^{j-1} \overline{r}_{sj} r_{sk}}{r_{jj}}$ si $k > j$

et ces relations permettent de calculer de proche en proche les lignes de R. Le nombre d'opérations requises pour ce calcul est

$$\begin{aligned}
&\sum_{j=1}^{n} \left[(j-1) + (j-2) + 1 + 1 + \sum_{k=j+1}^{n} (j-1) + (j-2) + 1 + 1 \right] \\
&\quad = \sum_{j=1}^{n} \left[(2j-1) + (n-j)(2j-1) \right] \\
&\quad = \sum_{j=1}^{n} (n-j+1)(2j-1) \\
&\quad = \frac{2n^3 + 3n^2 + n}{6} \sim \frac{n^3}{3}.
\end{aligned}$$

Si on utilise cette décomposition pour résoudre le système $AX = B$, on doit encore ajouter $2n^2$ opérations soit un total de

$$\frac{2n^3 + 15n^2 + n}{6} \sim \frac{n^3}{3}$$

opérations. Asymptotiquement, on a donc gagné un facteur 2 par rapport au nombre d'opérations requises par l'algorithme de Gauss sans pivotage. Pour n petit, le gain est moins sensible. Par exemple, pour $n = 10$, le nombre d'opérations passe de 805 à 585.

3.2 Stabilité des solutions

3.2.1 Normes matricielles et nombre de conditionnement

Définition 3.2.1. Si $X \in \mathbb{C}^n$, on pose

$$|X|_1 = \sum_{j=1}^{n} |x_j|, \qquad |X|_2 = \sqrt{\sum_{j=1}^{n} |x_j|^2}, \qquad |X|_\infty = \max_{j\in\{1,\dots,n\}} |x_j|.$$

Proposition 3.2.2. *Pour $p \in \{1, 2, \infty\}$, l'application*

$$|\cdot|_p : \mathbb{C}^n \to [0, +\infty[$$

est une norme sur $\mathbb{C}^n$ (i.e.

(a) $|\alpha X|_p = |\alpha||X|_p$;

(b) $|X + Y|_p \leq |X|_p + |Y|_p$;

(c) $|X|_p = 0 \Leftrightarrow X = 0$*).*

De plus, si p, $q \in \{1, 2, \infty\}$, il existe $C_{pq} > 0$ tel que

$$|X|_p \leq C_{pq}|X|_q$$

ce qui montre que les normes $|\cdot|_p$ et $|\cdot|_q$ sont équivalentes.

Démonstration. Laissée à titre d'exercice. □

Définition 3.2.3. Soit $A \in \mathbb{C}_n^m$ et soient p, $q \in \{1, 2, \infty\}$. Posons

$$\|A\|_{pq} = \sup_{|X|_q<1} |AX|_p.$$

On vérifie aisément que l'application

$$\|\cdot\|_{pq} : \mathbb{C}_n^m \to [0, +\infty[$$

est une norme sur $\mathbb{C}_n^m$. On dit que c'est la *norme matricielle subordonnée aux normes vectorielles* $|\cdot|_p$ et $|\cdot|_q$. Dans la suite, nous considérerons uniquement le cas $p = q$ et poserons pour simplifier

$$\|A\|_p = \|A\|_{pp}.$$

Proposition 3.2.4.

(a) Soit $A \in \mathbb{C}_n^m$ et soit $X \in \mathbb{C}^n$. Alors,

$$|AX|_p \leq \|A\|_p |X|_p.$$

(b) Soit $A \in \mathbb{C}_n^n$ une matrice inversible et soient X, $B \in \mathbb{C}^n$ tels que

$$AX = B.$$

Si $\tilde{X}$, $\tilde{B}$ sont des valeurs approchées de X, B telles que

$$A\tilde{X} = \tilde{B},$$

alors

$$|\Delta X|_p \leq \left\|A^{-1}\right\|_p |\Delta B|_p.$$

De plus, si $B \neq 0$, on a $X \neq 0$ et

$$\frac{|\Delta X|_p}{|X|_p} \leq \left\|A^{-1}\right\|_p \|A\|_p \frac{|\Delta B|_p}{|B|_p}.$$

Démonstration.

(a) Cela résulte directement des définitions.

(b) Comme

$$A(\tilde{X} - X) = \tilde{B} - B,$$

on a

$$\Delta X = A^{-1}\Delta B$$

et la conclusion résulte de (a). □

Définition 3.2.5. Le nombre

$$\eta_p(A) = \left\|A^{-1}\right\|_p \|A\|_p$$

s'appelle le *nombre de conditionnement de la matrice A*. Il donne une indication de la sensibilité des solutions du système linéaire

$$AX = B$$

à une perturbation sur B.

3.2.2 Calcul de $\|A\|_1$, $\|A\|_2$ et $\|A\|_\infty$

Pour calculer ce nombre, il faut pouvoir calculer $\|A\|_p$. C'est facile dans le cas $p = 1$ ou $p = \infty$ car on a :

Proposition 3.2.6. *Si $A \in \mathbb{C}_n^m$, alors*

(a) $\|A\|_1 = \sup\limits_{1\le k\le n} \sum\limits_{j=1}^{m} |a_{jk}|$;

(b) $\|A\|_\infty = \sup\limits_{1\le j\le m} \sum\limits_{k=1}^{n} |a_{jk}|$.

Démonstration. (a) Pour tout $X \in \mathbb{C}^n$ on a

$$\begin{aligned} |AX|_1 = \sum_{j=1}^{m} \left| \sum_{k=1}^{n} a_{jk}x_k \right| &\le \sum_{j=1}^{m}\sum_{k=1}^{n} |a_{jk}||x_k| \\ &\le \sum_{k=1}^{n} \left(\sum_{j=1}^{m} |a_{jk}| \right) |x_k| \\ &\le \sup_{1\le k\le n} \sum_{j=1}^{m} |a_{jk}||X|_1. \end{aligned}$$

Ainsi,

$$\|A\|_1 \le \sup_{1\le k\le n} \sum_{j=1}^{m} |a_{jk}|.$$

De plus, pour

$$X = E_k$$

on a

$$\sum_{j=1}^{m} |a_{jk}| = |AX|_1 \le \|A\|_1 |X|_1 \le \|A\|_1 .$$

Il s'ensuit que

$$\sup_{1\le k\le n} \sum_{j=1}^{m} |a_{jk}| \le \|A\|_1 .$$

(b) Pour tout $X \in \mathbb{C}^n$, on a

$$\left| \sum_{k=1}^{n} a_{jk}x_k \right| \le \left(\sum_{k=1}^{n} |a_{jk}| \right) |X|_\infty$$

d'où l'on tire que

$$|AX|_\infty = \sup_{1\le j\le n}\left|\sum_{k=1}^{n} a_{jk}x_k\right| \le \sup_{1\le j\le n}\sum_{k=1}^{n}|a_{jk}||X|_\infty.$$

Ainsi,

$$\|A\|_\infty \le \sup_{1\le j\le n}\sum_{k=1}^{n}|a_{jk}|.$$

Fixons $j\in\{1,\dots,n\}$ et posons

$$c_k = \begin{cases}\overline{a_{jk}}/|a_{jk}| & \text{si } a_{jk}\neq 0\\ 1 & \text{sinon.}\end{cases}$$

Pour $X=\sum_{l=1}^{n} c_k E_k$, on a

$$(AX)_j = \sum_{k=1}^{n} c_k a_{jk} = \sum_{k=1}^{n}|a_{jk}|.$$

Il s'ensuit que

$$\sum_{k=1}^{n}|a_{jk}| \le |AX|_\infty \le \|A\|_\infty |X|_\infty.$$

Comme $|X|_\infty = 1$, on en tire que

$$\sum_{k=1}^{n}|a_{jk}| \le \|A\|_\infty .$$

Cette relation ayant lieu pour tout j dans $\{1,\dots,n\}$, on voit que

$$\sup_{1\le j\le n}\sum_{k=1}^{n}|a_{jk}| \le \|A\|_\infty$$

et la conclusion en découle aisément. □

Dans le cas où $p=2$, le calcul de $\|A\|_p$ est un peu plus délicat et il faut faire appel à la notion de valeur singulière.

Définition 3.2.7. Soit $A\in\mathbb{C}_n^m$. La matrice

$$A^*A \in \mathbb{C}_n^n$$

est une matrice hermitienne semi-définie positive ; ses valeurs propres sont donc positives ou nulles. Ordonnons-les par valeurs décroissantes et notons-les $\rho_1,\dots,\rho_n$. La k^{e} *valeur singulière* de A est alors

$$\sigma_k = \sqrt{\rho_k}.$$

Dans la suite, nous noterons r le nombre de valeurs singulières non nulles.

Proposition 3.2.8. *Il existe des matrices $U \in \mathbb{C}_n^n$ et $V \in \mathbb{C}_m^m$ unitaires telles que*

$$V^*AU = \begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}.$$

En particulier, $\operatorname{rg} A = r$.

Démonstration. On sait qu'il existe une base orthonormée $U_1, \ldots, U_n$ formée de vecteurs propres de A^*A. Ordonnons-la de sorte que

$$A^*AU_k = \sigma_k^2 U_k.$$

On a alors

$$\langle AU_j, AU_k \rangle = \langle U_j, A^*AU_k \rangle = \sigma_k^2 \delta_{jk}.$$

Il s'ensuit que les vecteurs $AU_1, \ldots, AU_r$ sont non nuls et deux à deux orthogonaux et que $AU_{r+1} = 0, \ldots, AU_n = 0$. Posons

$$V_1 = \frac{AU_1}{|AU_1|}, \ldots, V_r = \frac{AU_r}{|AU_r|}$$

et complétons $V_1, \ldots, V_r$ par $V_{r+1}, \ldots, V_m$ de manière à obtenir une base orthonormée de $\mathbb{C}^m$. Comme

$$AU_k = \begin{cases} \sigma_k V_k & \text{si } k \in \{1, \ldots, r\} \\ 0 & \text{si } k \in \{r+1, \ldots, n\} \end{cases}$$

la matrice de A par rapport aux bases $U_1, \ldots, U_n$ et $V_1, \ldots, V_m$ est

$$\begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}.$$

La conclusion en découle. □

Remarque 3.2.9. (a) Si on a

$$V^*AU = \begin{pmatrix} \sigma'_1 & & & & & \\ & \ddots & & & & \\ & & \sigma'_k & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}$$

avec V, U unitaires et $\sigma'_1 \geq \sigma'_2 \cdots \geq \sigma'_k > 0$, alors

$$U^*A^*AU = \begin{pmatrix} \sigma'^2_1 & & & & & \\ & \ddots & & & & \\ & & \sigma'^2_k & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}$$

et les $\sigma'^2_1, \ldots, \sigma'^2_k$ sont les valeurs propres de A^*A. Il s'ensuit que

$$\sigma'_1 = \sigma_1, \ldots, \sigma'_k = \sigma_k.$$

(b) Avec les notations utilisées dans la preuve de la proposition précédente, on a bien sûr

$$\operatorname{Ker} A = \rangle U_{r+1}, \ldots, U_n \langle = \rangle U_1, \ldots, U_r \langle^{\perp}$$
$$\operatorname{Im} A = \rangle V_1, \ldots, V_r \langle = \rangle V_{r+1}, \ldots, V_n \langle^{\perp}.$$

Corollaire 3.2.10. *Si $A \in \mathbb{C}^n_n$, on a*

$$\|A\|_2 = \sigma_1.$$

Démonstration. Il est clair que

$$\|A\|_2 = \|V^*AU\|_2$$

quelque soient les matrices unitaires U, $V \in \mathbb{C}^n_n$. D'après ce qui précède, tout revient donc à traiter le cas où

$$A = \begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}.$$

Dans ce cas, on a

$$|AX|_2^2 = \sigma_1^2|x_1|^2 + \cdots + \sigma_r^2|x_r|^2 \leq \sigma_1^2|x|_2^2$$

et

$$\|A\|_2 \leq \sigma_1.$$

Comme on a aussi

$$\sigma_1^2 = |AE_1|_2^2 \leq \|A\|_2^2$$

on voit que $\sigma_1 = \|A\|_2$; d'où la conclusion. □

3.2.3 Influence d'une perturbation d'un sytème sur sa solution

Soit $|\cdot|$ une norme vectorielle sur $\mathbb{C}^n$ et soit $\|\cdot\|$ la norme matricielle qui lui est subordonnée.

Lemme 3.2.11. *Si A, $B \in \mathbb{C}_n^n$ alors*

$$\|AB\| \leq \|A\| \|B\|, \qquad \|I\| = 1.$$

Démonstration. Cela résulte directement des définitions et de 3.2.4. □

Lemme 3.2.12. *Soit E une matrice telle que $\|E\| < 1$. Alors la matrice $I - E$ est inversible et on a*

$$(I - E)^{-1} = \sum_{k=0}^{\infty} E^k.$$

En particulier,

$$\left\|(I - E)^{-1}\right\| \leq \sum_{k=0}^{\infty} \|E\|^k = \frac{1}{1 - \|E\|}.$$

Démonstration. Vu le lemme précédent, on a clairement

$$\left\| \sum_{k=p}^{q} E^k \right\| \leq \sum_{k=p}^{q} \|E\|^k$$

pour tous $p, q \in \mathbb{N}$. Comme la série

$$\sum_{k=0}^{+\infty} \|E\|^k$$

est convergente puisque $\|E\| < 1$, elle est de Cauchy. Vu la majoration précédente, il en est donc de même de la série matricielle

$$\sum_{k=0}^{+\infty} E^k.$$

Cette série est donc convergente. De plus, comme

$$(I-E)\sum_{k=0}^{K} E^k = \sum_{k=0}^{K} E^k - \sum_{k=1}^{K+1} E^k = I - E^{K+1}$$

on a aussi

$$(I-E)\sum_{k=0}^{+\infty} E^k = I,$$

ce qui démontre la première assertion de l'énoncé. La seconde en est une conséquence directe. □

Proposition 3.2.13. *Soit $A \in \mathbb{C}_n^n$ une matrice inversible et soient X, $B \in \mathbb{C}^n$ tels que*

$$AX = B.$$

Si $\tilde{A}$, $\tilde{X}$, $\tilde{B}$ sont des valeurs approchées de A, X, B telles que

$$\tilde{A}\tilde{X} = \tilde{B}$$

alors

$$\frac{|\Delta X|}{|X|} \leq \frac{\eta(A)}{1-\eta(A)\frac{\|\Delta A\|}{\|A\|}}\left(\frac{\|\Delta A\|}{\|A\|} + \frac{|\Delta B|}{|B|}\right)$$

si $\eta(A)\dfrac{\|\Delta A\|}{\|A\|} < 1$.

Démonstration. De

$$(A+\Delta A)(X+\Delta X) = (B+\Delta B)$$

on tire que

$$(A+\Delta A)\Delta X = \Delta B - (\Delta A)X$$

puis que

$$(I + A^{-1}\Delta A)\Delta X = A^{-1}(\Delta B - (\Delta A)X).$$

Comme

$$\left\|A^{-1}\Delta A\right\| \leq \left\|A^{-1}\right\| \|\Delta A\| \leq \eta(A)\frac{\|\Delta A\|}{\|A\|}$$

nos hypothèses et le lemme précédent montrent que

$$(I + A^{-1}\Delta A)$$

est inversible et on a

$$\Delta X = (I + A^{-1}\Delta A)^{-1}A^{-1}(\Delta B - (\Delta A)X).$$

On en tire que

$$|\Delta X| \leq \frac{1}{1 - \|A^{-1}\Delta A\|} \left\|A^{-1}\right\| (|\Delta B| + \|\Delta A\| \, |X|)$$

puis que

$$\frac{|\Delta X|}{|X|} \leq \frac{\eta(A)}{1 - \eta(A)\frac{\|\Delta A\|}{\|A\|}} \left(\frac{|\Delta B|}{\|A\| \, |X|} + \frac{\|\Delta A\|}{\|A\|} \right)$$

et comme $|B| \leq \|A\| \, |X|$, la conclusion en résulte. □

3.3 Méthodes itératives de résolution

Plutôt que de résoudre le système linéaire

$$AX = B$$

par une méthode directe comme la méthode de Gauss, on peut aussi chercher à construire de manière itérative une suite X_k d'approximations de la solution X qui converge vers celle-ci. Parmi les méthodes classiques de ce type, nous étudierons uniquement les méthodes de Jacobi et de Gauss-Seidel. Ces deux méthodes sont voisines et présupposent que les éléments diagonaux de A sont non nuls.

3.3.1 Méthode de Jacobi

L'idée de la méthode de Jacobi est de réécrire le système

$$\left\{ \begin{array}{rcl} a_{11}x_1 + \cdots + a_{1n}x_n & = & b_1 \\ & \vdots & \\ a_{n1}x_1 + \cdots + a_{nn}x_n & = & b_n \end{array} \right.$$

sous la forme

$$\left\{ \begin{array}{rcl} x_1 & = & \dfrac{b_1}{a_{11}} - \dfrac{a_{12}}{a_{11}}x_2 - \cdots - \dfrac{a_{1n}}{a_{11}}x_n \\ & \vdots & \\ x_n & = & \dfrac{b_n}{a_{nn}} - \dfrac{a_{n1}}{a_{nn}}x_1 - \cdots - \dfrac{a_{nn-1}}{a_{nn}}x_{n-1} \end{array} \right.$$

De cette manière, on remplace le problème initial par un problème de recherche de point fixe pour la transformation

$$\varphi_J : \mathbb{C}^n \to \mathbb{C}^n \qquad (x_j)_{1\leq j\leq n} \mapsto \left(\frac{b_j}{a_{jj}} - \sum_{k\neq j} \frac{a_{jk}}{a_{jj}} x_k \right)_{1\leq j\leq n}.$$

Il est alors naturel de chercher à approcher la solution X du système $AX = B$ par une suite X_m définie par la relation de récurrence

$$X_{m+1} = \varphi_J(X_m) \qquad (m \geq 0)$$

et un choix de condition initiale X_0. Pour bien comprendre la nature de la formule de récurrence ci-dessus, le plus simple est de donner une formule matricielle pour $\varphi_J(X)$. Pour cela, posons

$$D = \mathrm{diag}(a_{11}, \ldots, a_{nn})$$

et écrivons $D^{-1}A$ sous la forme $L+I+U$ où L (resp. U) est une matrice triangulaire inférieure (resp. supérieure) dont les éléments diagonaux sont nuls. On a alors

$$\varphi_J(X) = D^{-1}B - (L+U)X$$

ce qui met bien en relief le fait que φ_J est une affinité.

3.3.2 Méthode de Gauss-Seidel

Lorsque l'on calcule X_{m+1} à partir de X_m dans la méthode de Jacobi, on calcule progressivement les composantes $x_{m+1,j}$ pour $j = 1, \ldots, n$ au moyen de la formule

$$x_{m+1,j} = \frac{b_j}{a_{jj}} - \sum_{k \neq j} \frac{a_{jk}}{a_{jj}} x_{m,k}. \qquad (*)$$

Lors du calcul de $x_{m+1,j}$ on dispose donc déjà des valeurs pour $x_{m+1,1}, \ldots, x_{m+1,j-1}$. L'idée de la méthode de Gauss-Seidel est de remplacer

$$x_{m,1}, \ldots, x_{m,j-1}$$

par

$$x_{m+1,1}, \ldots, x_{m+1,j-1}$$

dans la formule (*) en espérant ainsi améliorer l'approximation de $x_{m+1,j}$ obtenue. La formule de calcul de $x_{m+1,j}$ est donc

$$x_{m+1,j} = \frac{b_j}{a_{jj}} - \sum_{k<j} \frac{a_{jk}}{a_{jj}} x_{m+1,k} - \sum_{k>j} \frac{a_{jk}}{a_{jj}} x_{m,k}.$$

Matriciellement cela correspond à la formule

$$X_{m+1} = D^{-1}B - LX_{m+1} - UX_m$$

que l'on peut encore réécrire

$$(I + L)X_{m+1} = D^{-1}B - UX_m$$

où

$$X_{m+1} = (I + L)^{-1}D^{-1}B - (I + L)^{-1}UX_m.$$

La suite X_m fournie par la méthode de Gauss-Seidel est donc définie par la relation de récurrence

$$X_{m+1} = \varphi_{\mathrm{GS}}(X_m)$$

où $\varphi_{\mathrm{GS}} : \mathbb{C}^n \to \mathbb{C}^n$ est l'affinité

$$X \mapsto (I + L)^{-1}D^{-1}B - (I + L)^{-1}UX.$$

L'étude de la convergence des méthodes considérées ci-dessus est donc subordonnée à celle de la convergence des méthodes itératives affines et c'est à celle-ci que nous allons à présent nous attacher.

3.3.3 Convergence des méthodes itératives affines

Soit φ une transformation affine de $\mathbb{C}^n$ possédant un et un seul point fixe Λ. Pour un tel φ, il existe une unique matrice $A \in \mathbb{C}_n^n$ et un unique vecteur $B \in \mathbb{C}^n$ tels que

$$\varphi(X) = AX + B \qquad (*)$$

pour tout $X \in \mathbb{C}^n$. Les points fixes de φ étant les solutions du système

$$X = AX + B$$

l'hypothèse sur φ revient à exiger que $A - I$ soit inversible ou ce qui revient au même que 1 ne soit pas une valeur propre de A. Essayons de dégager une condition nécessaire et suffisante pour que toute suite X_m vérifiant

$$X_{m+1} = AX_m + B \qquad (m \geq 0) \qquad (**)$$

converge vers Λ (on dit alors que la méthode itérative *converge globalement* vers Λ). Si ce phénomène a lieu, la suite X_m définie par la relation de récurrence (**) et la condition initiale $X_0 = \Lambda + Z$ converge vers Λ quel que soit $Z \in \mathbb{C}^n$. Comme pour cette suite on a

$$X_{m+1} - \Lambda = A(X_m - \Lambda) \qquad (m \geq 0)$$

on voit que

$$A^m Z = X_m - \Lambda \to 0$$

pour tout $Z \in \mathbb{C}^n$. Il s'ensuit que $A^m \to 0$ dans $\mathbb{C}_n^n$. Réciproquement, si $A^m \to 0$ et si X_m est une suite vérifiant (**), on a

$$(X_m - \Lambda) = A^m(X_0 - \Lambda) \to 0$$

et $X_m \to \Lambda$. Nous voyons donc que la méthode itérative (**) converge globalement vers Λ si et seulement si $A^m \to 0$. Pour transformer cette condition en une condition plus explicite, nous aurons besoin de la notion de rayon spectral.

Définition 3.3.1. Soit $A \in \mathbb{C}_n^n$. Le *rayon spectral* de la matrice A est le maximum des modules des valeurs propres de A. On le note $\rho(A)$. On a donc par définition

$$\rho(A) = \sup\{|\lambda| : \lambda \text{ valeur propre de } A\}.$$

Proposition 3.3.2. *Soit $A \in \mathbb{C}_n^n$. Alors une condition nécessaire et suffisante pour que $A^m \to 0$ est que $\rho(A) < 1$.*

Démonstration. Supposons que $A^m \to 0$. Soit λ une valeur propre de A telle que $|\lambda| = \rho(A)$ et soit V un vecteur propre non nul associé à λ. De la relation

$$AV = \lambda V$$

on tire que

$$A^m V = \lambda^m V.$$

Puisque $A^m \to 0$ et que $V \neq 0$, on en tire que $\lambda^m \to 0$. Ainsi $\rho(A) = |\lambda| < 1$. Réciproquement, si $\rho(A) < 1$, alors toutes les valeurs propres de A sont de module strictement inférieur à 1. Notons P une matrice qui réduit A à la forme canonique de Jordan. On a

$$P^{-1}AP = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_p \end{pmatrix}$$

où chaque J_j est une matrice carrée de dimension n_j de la forme

$$\begin{pmatrix} \lambda_j & 1 & & \\ & \ddots & \ddots & \\ & & \lambda_j & 1 \\ & & & \lambda_j \end{pmatrix} = \lambda_j I + N.$$

Comme $N^{n_j} = 0$, on a

$$(\lambda_j I + N)^m = \sum_{k=0}^{n_j} \mathrm{C}_m^k \, \lambda_j^{m-k} N^k$$

et la relation $|\lambda_j| < 1$ entraîne que

$$J_j^m \to 0.$$

Il s'ensuit que

$$P^{-1}A^mP = (P^{-1}AP)^m \to 0$$

puis que $A^m \to 0$. □

Corollaire 3.3.3. *La méthode itérative (**) converge globalement vers Λ si et seulement si $\rho(A) < 1$.*

Le rayon spectral $\rho(A)$ est en général assez délicat à calculer. On a cependant le résultat suivant :

Proposition 3.3.4. *Si $A \in \mathbb{C}_n^n$ alors $\rho(A)$ est la borne inférieure de $\|A\|$ lorsque $\|\cdot\|$ parcourt l'ensemble des normes matricielles subordonnées à des normes vectorielles de $\mathbb{C}^n$.*

Démonstration. Si V est un vecteur propre non nul de A et si $\|\cdot\|$ est une norme matricielle subordonnée à la norme vectorielle $|\cdot|$, la relation

$$AV = \lambda V$$

entraîne que

$$|\lambda||V| = \|AV\| \leq \|A\|\,|V|.$$

Il s'ensuit que $|\lambda| \leq \|A\|$, ce qui montre que $\|A\|$ majore le rayon spectral de A. Pour conclure, il suffit de montrer que pour tout $\varepsilon > 0$ il existe une norme matricielle $\|\cdot\|$ subordonnée à une norme vectorielle $|\cdot|$ telle que $\|A\| \leq \rho(A) + \varepsilon$. Fixons $\varepsilon > 0$ et notons P une matrice qui réduit A à la forme canonique de Jordan. On a

$$P^{-1}AP = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_p \end{pmatrix}$$

avec J_j comme dans la preuve de 3.3.2. Posons $P_\varepsilon = \mathrm{diag}(1, \varepsilon, \ldots, \varepsilon^{n-1})$. Un calcul direct montre que

$$P_\varepsilon^{-1}J_jP_\varepsilon = \begin{pmatrix} \lambda_j & \varepsilon & & \\ & \ddots & \ddots & \\ & & \lambda_j & \varepsilon \\ & & & \lambda_j \end{pmatrix}.$$

Il s'ensuit que

$$\left\|P_\varepsilon^{-1}P^{-1}APP_\varepsilon\right\|_1 \leq \sup_{1\leq j\leq p}|\lambda_j| + \varepsilon \leq \rho(A) + \varepsilon.$$

Comme $A \mapsto \|P_\varepsilon^{-1}P^{-1}APP_\varepsilon\|_1$ est une norme matricielle subordonnée à la norme vectorielle $X \mapsto \|P_\varepsilon^{-1}P^{-1}X\|_1$ on obtient la conclusion souhaitée. □

Remarque 3.3.5. Le résultat précédent montre en particulier que si $\|\cdot\|$ est une norme matricielle subordonnée à une norme vectorielle $|\cdot|$, alors la condition $\|A\| < 1$ est suffisante pour que toute suite vérifiant (**) converge vers Λ. De plus, dans ce cas comme

$$|X_{m+1} - \Lambda| \leq \|A\|\,|X_m - \Lambda|$$

la convergence est au moins linéaire de taux $\|A\|$. En choisissant bien $|\cdot|$ et $\|\cdot\|$ on peut d'ailleurs rendre ce taux aussi proche de $\rho(A)$ que l'on veut. Notons également que de la relation

$$(X_{m+1} - \Lambda) = A(X_m - \Lambda)$$

on tire que

$$(X_{m+1} - \Lambda) = A(X_m - X_{m+1}) + A(X_{m+1} - \Lambda).$$

Ainsi,

$$|X_{m+1} - \Lambda| \leq \|A\|\,|X_{m+1} - X_m| + \|A\|\,|X_{m+1} - A|$$

et

$$|X_{m+1} - \Lambda| \leq \frac{\|A\|}{1 - \|A\|}|X_{m+1} - X_m|.$$

Comme le second membre de cette relation ne contient que des quantités calculables on pourra s'en servir pour estimer la précision obtenue en approchant Λ par X_{m+1}. Bien sûr, pour être complet, une étude des erreurs d'arrondis s'imposerait mais nous ne la ferons pas ici.

3.3.4 Convergence des méthodes de Jacobi et de Gauss-Seidel

Définition 3.3.6. Une matrice $A \in \mathbb{C}_n^n$ est *strictement diagonalement dominante* si

$$|a_{jj}| > \sum_{k\neq j}|a_{jk}|.$$

Proposition 3.3.7. *Soit $B \in \mathbb{C}^n$ et soit $A \in \mathbb{C}_n^n$ une matrice strictement diagonalement dominante. Alors A est inversible et la méthode de Jacobi (resp. Gauss-Seidel) fournit une suite X_m qui converge vers la solution de*

$$AX = B$$

quelque soit la condition initiale X_0.

Démonstration. Méthode de Jacobi. On sait que

$$\varphi_J(X) = D^{-1}B - (L+U)X.$$

Tout revient donc à montrer que

$$\rho(L+U) < 1$$

car alors $A = D(I + (L+U))$ est clairement inversible. Puisque

$$\|L+U\|_\infty = \sup_{1\leq j\leq n} \sum_{k\neq j} \frac{|a_{jk}|}{|a_{jj}|}$$

nos hypothèses entraînent que $\|L+U\|_\infty < 1$. La conclusion découle donc de 3.3.2.

Méthode de Gauss-Seidel. On sait que

$$\varphi_{\mathrm{GS}}(X) = (I+L)^{-1}D^{-1}B - (I+L)^{-1}UX.$$

Il suffit donc de montrer que

$$\rho((I+L)^{-1}U) < 1.$$

Fixons $X \in \mathbb{C}^n$ et posons

$$Y = (I+L)^{-1}UX.$$

Comme $(I+L)Y = UX$, on a

$$y_j = \sum_{k>j} u_{jk}x_k - \sum_{k<j} l_{jk}y_k$$

pour tout $j \in \{1,\ldots,n\}$. Pour j tel que $|y_j| = |Y|_\infty$, on a donc

$$|Y|_\infty \leq \left(\sum_{k>j} |u_{jk}|\right) |X|_\infty + \left(\sum_{k<j} |l_{jk}|\right) |Y|_\infty.$$

Il s'ensuit que

$$|Y|_\infty \leq \frac{\sum_{k>j} |u_{jk}|}{1 - \sum_{k<j} |l_{jk}|} |X|_\infty.$$

Par conséquent,

$$\left\|(I+L)^{-1}U\right\|_\infty \leq \sup_{1\leq j\leq n} \frac{\sum_{k>j} |a_{jk}|}{|a_{jj}| - \sum_{k<j} |a_{jk}|}$$

et nos hypothèses entraînent que

$$\left\|(I+L)^{-1}U\right\|_\infty < 1.$$

□

Remarque 3.3.8. Il résulte aussi de la majoration obtenue à la fin de la preuve précédente que

$$\left\|(I+L)^{-1}U\right\|_\infty \leq \|L+U\|_\infty .$$

On pourrait don être tenté de croire que la méthode de Gauss-Seidel est toujours plus rapide que la méthode de Jacobi. En fait, ce n'est pas le cas en général car la majoration précédente n'entraîne pas que

$$\rho((I+L)^{-1}U) \leq \rho(L+U).$$

Cette dernière condition est cependant réalisée dans beaucoup de cas rencontrés en pratique (voir e.g. [**?**][Theorem 8.2.14]).

Terminons par un autre cas intéressant de convergence de la méthode de Gauss-Seidel.

Proposition 3.3.9. *Soit $A \in \mathbb{C}_n^n$ une matrice hermitienne définie positive. Alors la méthode de Gauss-Seidel fournit une suite X_n qui converge vers la solution de*

$$AX = B$$

quelque soit la condition initiale X_0.

Démonstration. Puisque $A = D(L+I+U)$ est hermitienne définie positive, D est à éléments diagonaux strictement positifs et $DU = (DL)^*$. Soit V un vecteur propre non nul de valeur propre λ de $(I+L)^{-1}U$. De la relation

$$(I+L)^{-1}UV = \lambda V$$

on tire que

$$DUV = \lambda(DV + DLV)$$

puis que

$$\overline{\langle DLV, V\rangle} = \langle V, DLV\rangle = \langle (DL)^*V, V\rangle = \langle DUV, V\rangle = \lambda(\langle DV, V\rangle + \langle DLV, V\rangle).$$

Posons $\alpha = \Re\langle DLV, V\rangle$, $\beta = \Im\langle DLV, V\rangle$, $\gamma = \langle DV, V\rangle$. Il vient

$$\alpha - i\beta = \lambda(\gamma + \alpha + i\beta).$$

Comme la matrice A et hermitienne définie positive,

$$\langle D(L+I+U)V, V\rangle = \langle DLV, V\rangle + \langle DV, V\rangle + \langle DUV, V\rangle = 2\alpha + \gamma$$

est strictement positif. Comme on a aussi $\gamma > 0$, on voit que

$$|\gamma + \alpha + i\beta|^2 = \alpha^2 + \beta^2 + 2\gamma\alpha + \gamma^2 > \alpha^2 + \beta^2.$$

Il s'ensuit que $\gamma + \alpha + i\beta \neq 0$ et que

$$|\lambda|^2 = \frac{\alpha^2 + \beta^2}{(\gamma + \alpha)^2 + \beta^2} = \frac{\alpha^2 + \beta^2}{\alpha^2 + \beta^2 + 2\gamma\alpha + \gamma^2} < 1.$$

Ainsi, $\rho((I + L)^{-1}U) < 1$ et la conclusion découle de 3.3.2. □

Remarque 3.3.10. L'énoncé précédent peut paraître plus restrictif qu'il n'est réellement. En effet, si $A \in \mathbb{C}_n^n$ est inversible, le système

$$AX = B$$

est équivalent au système

$$A^*AX = A^*B$$

et comme la matrice A^*A est hermitienne définie positive, la méthode de Gauss-Seidel permet d'en approcher sans problème la solution.

4 Interpolation et approximation polynomiale

4.1 Interpolation à pas variable

Le problème de l'interpolation d'une fonction f définie sur $[a, b]$ par un polynôme est le suivant : étant donnés $n+1$ points distincts $x_0, x_1, \ldots, x_n \in [a, b]$, trouver un polynôme P de degré le plus bas possible tel que

$$\begin{aligned} P(x_0) &= f(x_0) \\ &\vdots \\ P(x_n) &= f(x_n) \end{aligned}$$

L'idée étant que si un tel polynôme est connu, alors on devrait pouvoir utiliser $P(x)$ pour approcher $f(x)$ sur $[a, b]$. La résolution théorique de ce problème est contenue dans le résultat suivant :

Proposition 4.1.1. *Si $x_0, \ldots, x_n$ sont $n+1$ points distincts de $\mathbb{R}$ et si $y_0, \ldots, y_n$ sont $n+1$ réels, alors il existe un et un seul polynôme réel P de degré inférieur ou égal à n tel que*

$$P(x_0) = y_0, \ldots, P(x_n) = y_n.$$

Démonstration. Tout polynôme réel de degré $\leq n$ s'écrit de manière unique sous la forme

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

L'ensemble $\mathbb{R}[X]_n$ de ces polynômes est donc un espace vectoriel réel de dimension $n+1$. L'application

$$\mathbb{R}[X]_n \to \mathbb{R}^{n+1} \qquad P \mapsto \begin{pmatrix} P(x_0) \\ \vdots \\ P(x_n) \end{pmatrix}$$

est linéaire et injective car un polynôme de degré $\leq n$ qui a $n+1$ zéros distincts est nul. Il s'ensuit que cette application est aussi surjective, d'où la conclusion. □

Remarque 4.1.2. On peut baser le calcul de P sur la proposition précédente, car pour trouver $a_0, \ldots, a_n$ il suffit de résoudre le système de Cramer

$$\begin{pmatrix} 1 & x_0 & \cdots & x_0^n \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ \vdots \\ y_n. \end{pmatrix}$$

On notera que la matrice de ce système est une matrice de Vandermonde. Cette méthode est en fait peu efficace. On lui préfère généralement des méthodes basées

sur des formules explicites pour $P(x)$ telles que les formules de Lagrange et de Newton établies ci-dessous.

Proposition 4.1.3 (Formule de Lagrange)**.** *Dans les conditions de la Proposition 4.1.1, le polynôme P est donné par*

$$P(x) = \sum_{j=0}^{n} y_j L_j(x)$$

où L_j est le polynôme de Lagrange de degré n

$$\frac{(x-x_0)\dots\widehat{(x-x_j)}\dots(x-x_n)}{(x_j-x_0)\dots\widehat{(x_j-x_j)}\dots(x_j-x_n)}$$

caractérisé par

$$L_j(x_k) = \delta_{jk}.$$

Démonstration. Il suffit de remarquer par calcul direct que l'on a bien

$$L_j(x_k) = \delta_{jk}$$

puis de constater que

$$\sum_{j=0}^{n} y_j L_j(x_k) = \sum_{j=0}^{n} y_j \delta_{jk} = y_k.$$

□

Remarque 4.1.4. Cette formule est commode si on doit calculer P pour beaucoup de valeurs différentes des y_j, les x_j restant fixes. Cependant, elle requiert le calcul des polynômes de Lagrange $L_j(x)$ qui peut être assez long.

Exemple 4.1.5. Pour $n = 2$, considérons la table

x_j	0	1	3
y_j	1	3	2

On a

$$L_0(x) = \frac{(x-1)(x-3)}{(0-1)(0-3)}, \quad L_1(x) = \frac{(x-0)(x-3)}{(1-0)(1-3)}, \quad L_2(x) = \frac{(x-0)(x-1)}{(3-0)(3-1)}$$

et

$$P(x) = y_0 L_0(x) + y_1 L_1(x) + y_2 L_2(x).$$

Ainsi, par exemple,

$$P(2) = 1\frac{-1}{3} + 3\frac{-2}{-2} + 2\frac{2}{6} = \frac{-1}{3} + 3 + \frac{2}{3} = \frac{1}{3} + 3 = \frac{10}{3}.$$

Définition 4.1.6. Dans les conditions de la Proposition 4.1.1, on définit les $k^{\text{èmes}}$ différences divisées $y_{j_0\dots j_k}$ par récurrence sur k au moyen des formules

$$y_{j_0\dots j_k} = \frac{y_{j_1\dots j_k} - y_{j_0\dots j_{k-1}}}{x_{j_k} - x_{j_0}}.$$

Remarque 4.1.7. Pour calculer à la main les différences divisées, on emploie souvent le schéma suivant :

	$k=0$	$k=1$	$k=2$	$\cdots$	$\cdots$	$\cdots$	$k=n$
x_0	y_0						
		y_{01}					
x_1	y_1		y_{012}				
		y_{12}		$\cdot$			
x_2	y_2		$\cdot$		$\cdot$		
		$\cdot$		$\cdot$		$\cdot$	
$\cdot$	$\cdot$		$\cdot$		$\cdot$		$y_{01\dots n}$
		$\cdot$		$\cdot$		$\cdot$	
x_{n-2}	$y_{(n-2)}$		$\cdot$		$\cdot$		
		$y_{(n-2)(n-1)}$		$\cdot$			
x_{n-1}	$y_{(n-1)}$		$y_{(n-2)(n-1)n}$				
		$y_{(n-1)n}$					
x_n	y_n						

Exemple 4.1.8. En reprenant les données de l'Exemple 4.1.5, on a le schéma

	$k=0$	$k=1$	$k=2$
0	1		
		2	
1	3		$-5/6$
		$-1/2$	
3	2		

Proposition 4.1.9 (Formule de Newton)**.** *Avec les notations de la Proposition 4.1.1, on a*

$$P(x) = \sum_{j=0}^{n} y_{0\dots j}(x - x_0)\dots(x - x_{j-1}).$$

Démonstration. Procédons par récurrence sur n. Pour $n = 0$, il n'y a rien à établir. Pour $k \leq n$, notons $P_{j_0\dots j_k}$ le polynôme associé aux suites $x_{j_0}, \dots, x_{j_k}$, $y_{j_0}, \dots, y_{j_k}$. On a

$$P_{j_0\dots j_k}(x) = \frac{(x - x_{j_0})P_{j_1\dots j_k}(x) - (x - x_{j_k})P_{j_0\dots j_{k-1}}(x)}{x_{j_k} - x_{j_0}} \qquad (*)$$

car le second membre vérifie les conditions caractérisant $P_{j_0\dots j_k}$. Nous savons déjà que

$$P_{j_0\dots j_{k-1}}(x) = \sum_{l=0}^{k-1} y_{j_0\dots j_l}(x - x_{j_0})\dots(x - x_{j_{l-1}})$$

et que

$$P_{j_0\dots j_k}(x) - P_{j_0\dots j_{k-1}}(x) = \mu(x - x_{j_0})\dots(x - x_{j_{k-1}}) \qquad (\mu \in \mathbb{R}).$$

Le nombre μ est donc le coefficient du terme dominant de

$$P_{j_0\dots j_k}(x).$$

D'après (*), ce coefficient est

$$\frac{y_{j_1\dots j_k} - y_{j_0\dots j_{k-1}}}{x_{j_k} - x_{j_0}} = y_{j_0\dots j_k}.$$

Il s'ensuit que

$$P_{j_0\dots j_k}(x) = \sum_{l=0}^{k-1} y_{j_0\dots j_l}(x - x_{j_0})\dots(x - x_{j_{l-1}}) + y_{j_0\dots j_k}(x - x_{j_0})\dots(x - x_{j_{k-1}}),$$

d'où la conclusion. □

Exemple 4.1.10. En reprenant les données de l'Exemple 4.1.8, on trouve

$$P(x) = 1 + 2(x - 0) - \tfrac{5}{6}(x - 0)(x - 1).$$

Corollaire 4.1.11. *Les différences divisées*

$$y_{j_0\dots j_k}$$

sont invariantes par permutation des indices.

Démonstration. Cela résulte de ce que le polynôme P associé aux suites

$$x_0, \dots, x_k \qquad \text{et} \qquad y_0, \dots, y_k$$

est invariant lorsque l'on permute conjointement ces deux suites. □

Remarque 4.1.12. On peut calculer la valeur en x du polynôme

$$P(x) = \sum_{j=0}^{n} y_{0\dots j}(x - x_0)\dots(x - x_{j-1})$$

de manière efficace en s'inspirant de la méthode d'Horner. Cela revient à réécrire $P(x)$ sous la forme

$$P(x) = (\dots(y_{0\dots n}(x - x_{n-1}) + y_{0\dots n-1})(x - x_{n-2}) + \dots y_{01})(x - x_0) + y_0.$$

Proposition 4.1.13. *Soit f une fonction de classe C_{n+1} sur l'intervalle I de $\mathbb{R}$. Supposons que $x_0, \ldots, x_n$ soient des points distincts de I et soient*

$$y_0 = f(x_0), \ldots, y_n = f(x_n).$$

Alors, pour tout $x \in I$, il existe $\xi \in I$ tel que

$$f(x) - P(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)\ldots(x - x_n).$$

Démonstration. Si $x \in I \setminus \{x_0, \ldots, x_n\}$ il est possible de trouver $\mu \in \mathbb{R}$ tel que

$$f(x) - P(x) = \mu(x - x_0)\ldots(x - x_n).$$

Pour un tel μ, la fonction

$$g(t) = f(t) - P(t) - \mu(t - x_0)\ldots(t - x_n)$$

s'annule en $x, x_0, \ldots, x_n$. Comme ces $n + 2$ points sont distincts, des applications répétées de Rolle montrent qu'il existe $\xi \in I$ tel que

$$g^{(n+1)}(\xi) = 0.$$

Puisque

$$g^{(n+1)}(t) = f^{(n+1)}(t) - \mu(n+1)!,$$

la conclusion en découle. □

Définition 4.1.14. Soit f une fonction définie sur l'intervalle $[a, b]$ de $\mathbb{R}$ et soient $x_0, \ldots, x_n$ des points distincts de $[a, b]$. Si

$$y_0 = f(x_0), \ldots, y_n = f(x_n),$$

on pose

$$f[x_0, \ldots, x_n] = y_{01\ldots n}.$$

Proposition 4.1.15. *Dans les conditions de le définition précédente, le polynôme d'interpolation de f en $x_0, \ldots, x_n$ est donné par*

$$P(x) = \sum_{k=0}^{n} f[x_0, \ldots, x_k](x - x_0)\ldots(x - x_{k-1})$$

et on a

$$f(x) - P(x) = f[x_0, \ldots, x_n, x](x - x_0)\ldots(x - x_n)$$

si $x \notin \{x_0, \ldots, x_n\}$.

Démonstration. La première partie découle directement de la Proposition 4.1.9. Pour tout $x_{n+1} \notin \{x_0, \ldots, x_n\}$, le polynôme d'interpolation de f en $x_0, \ldots, x_n, x_{n+1}$ est donc

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

En x_{n+1}, on a donc

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

La conclusion s'obtient en prenant $x_{n+1} = x$. □

Remarque 4.1.16. (a) On pourrait croire que l'écart entre la valeur de f en x et celle du polynôme P_n interpolant f aux points $a = x_0, \ldots, x_n = b$ peut être rendu aussi petit que l'on veut si on fait en sorte que

$$\max_{k \in \{0,\ldots,n-1\}} |x_{k+1} - x_k|$$

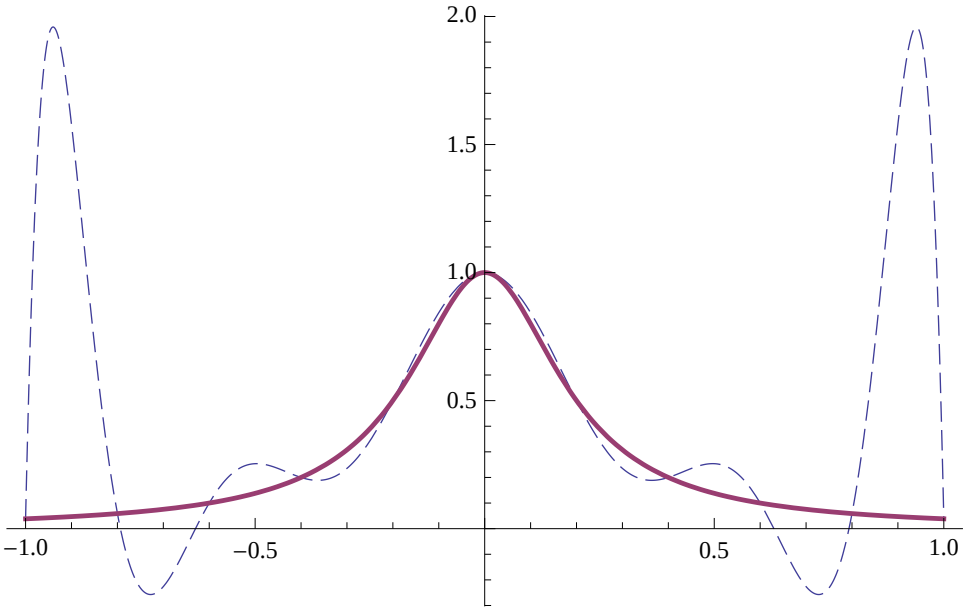
soit suffisamment petit. En fait, ce n'est en général pas le cas même si f est très régulier sur $[a, b]$. Dans le cas où $x_{k+1} - x_k$ ne dépend pas de k, c'est là une des facettes du phénomène de Runge. Par exemple, si

$$f(x) = \frac{1}{1 + 25x^2}$$

et si $[a, b] = [-1, 1]$, on peut montrer que le polynôme P_n interpolant f aux points

$$x_k = -1 + \frac{2k}{n} \qquad (k = 0, \ldots, n)$$

ne tend pas vers f lorsque $n \to \infty$. Sur la figure ci-dessous, on a représenté f en gras et P_n en pointillés dans le cas $n = 10$:



(b) Si l'on fixe n et que l'on pose

$$h = \max_{k\in\{0,\dots,n-1\}} |x_{k+1} - x_k|$$

alors pour $x \in [x_0, x_1]$, on a

$$|(x - x_0)\dots(x - x_n)| \leq h(h)(2h)\dots(nh) \leq h^{n+1}n!.$$

Si f est de classe C_{n+1} sur $[a, b]$ et si on a $x_0 = a$ et $h < \frac{b-a}{n}$, alors

$$\sup_{x\in[x_0,x_1]} |f(x) - P(x)| \leq \frac{\sup_{[a,b]} |f^{(n+1)}|}{n+1} h^{n+1}.$$

Comme le second membre tend vers 0 avec h, on peut donc rendre

$$\sup_{x\in[x_0,x_1]} |f(x) - P(x)|$$

aussi petit que l'on veut à condition de rendre h suffisamment petit. Ceci montre que malgré (a), l'interpolation polynomiale (même à pas constant) peut quand même être utilisée pour approcher les valeurs de $f(x)$.

4.2 Interpolation à pas constant

Nous allons montrer que dans le cas où

$$x_k = x_0 + kh \qquad (k \in \{0, \dots, n\})$$

pour $h > 0$ fixé, la formule de Newton peut se réécrire d'une manière très simple en termes de différences finies.

Définition 4.2.1. Si $y = (y_n)_{n\geq 0}$ est une suite de réels, alors on note Δy la suite $(y_{n+1} - y_n)_{n\geq 0}$. En d'autres termes, on pose

$$\Delta y_n = y_{n+1} - y_n.$$

Proposition 4.2.2. *L'opérateur $\Delta : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ est linéaire et on a*

$$\Delta = \tau - \mathrm{id}$$

où

$$\tau : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$$

est défini par

$$\tau(y)_n = y_{n+1}.$$

Démonstration. C'est immédiat. □

Corollaire 4.2.3. *On a*

$$\Delta^p y_n = \sum_{k=0}^{p} \mathrm{C}_p^k (-1)^{p-k} y_{n+k}.$$

Démonstration. Comme τ et id commutent, il vient

$$\Delta^p = \sum_{k=0}^{p} \mathrm{C}_p^k \, \tau^k (-1)^{p-k} \, \mathrm{id}$$

d'où la conclusion puisque

$$\tau^k y_n = y_{n+k}.$$

□

Proposition 4.2.4. *Soient $x_0, \ldots, x_n$ des réels distincts et $y_0, \ldots, y_n$ des réels. Supposons que $x_k = x_0 + kh$ ($k \in \{0, \ldots, n\}$). Alors,*

$$y_{01\ldots n} = \frac{\Delta^n y_0}{n! h^n}.$$

Démonstration. Procédons par récurrence sur n. On a

$$y_{01} = \frac{y_1 - y_0}{x_1 - x_0} = \frac{\Delta y_0}{h}$$

et le résultat est donc vrai pour $n = 1$. Supposons-le vrai pour n et démontrons-le pour $n + 1$. On a

$$y_{0\ldots n+1} = \frac{y_{1\ldots n+1} - y_{0\ldots n}}{x_{n+1} - x_0} = \frac{\Delta^n y_1 - \Delta^n y_0}{(n+1)hn!h^n} = \frac{\Delta^{n+1} y_0}{(n+1)!h^{n+1}},$$

d'où la conclusion. □

Corollaire 4.2.5. *Dans les conditions de la proposition précédente, le polynôme P de degré $\leq n$ tel que $P(x_0) = y_0, \ldots, P(x_n) = y_n$ est donné par*

$$P(x) = \sum_{k=0}^{n} \frac{\Delta^k y_0}{k!} q(q-1)\ldots(q-k+1)$$

où $q = (x - x_0)/h$. Si $y_0 = f(x_0), \ldots, y_n = f(x_n)$ pour f de classe C_{n+1} sur un intervalle $[a, b]$ contenant $\{x_0, \ldots, x_n, x\}$, alors il existe $\xi \in [a, b]$ tel que

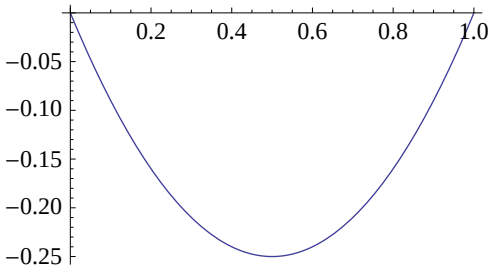
$$f(x) - P(x) = \frac{f^{(n+1)}(\xi) h^{n+1}}{(n+1)!} q(q-1)\ldots(q-n).$$

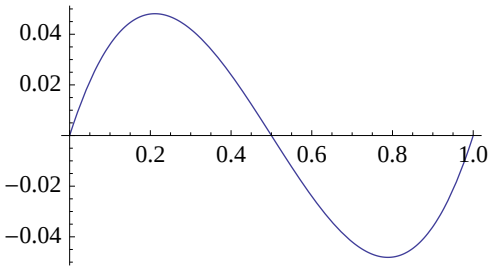
Démonstration. Compte tenu de la proposition précédente, cela résulte directement de la Proposition 4.1.9 et de la Proposition 4.1.13. □

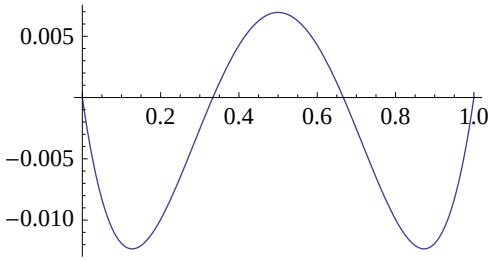
Remarque 4.2.6. Le terme d'erreur fait intervenir le polynôme

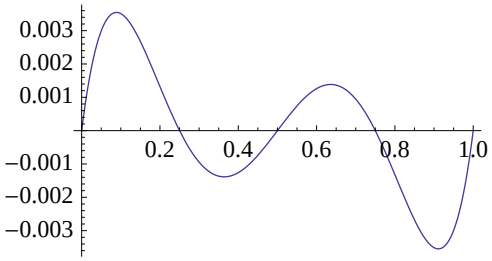
$$E(x) = (x - x_0) \dots (x - x_n) = h^{n+1} q(q-1) \dots (q-n)$$

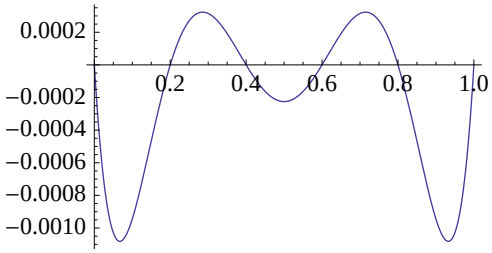
de manière essentielle. Il est donc intéressant de comprendre le comportement de ce polynôme pour $x \in [x_0, x_n]$. On trouvera page suivante le graphe de $E(x)$ pour $x_0 = 0, x_n = 1$ et $n \in \{1, \dots, 10\}$. Sur ces graphes on constate que le polynôme $E(x)$ est bien plus petit près du centre de l'intervalle que près du bord. Ceci révèle une propriété assez générale de l'interpolation à pas constant qui est une autre facette du phénomène de Runge.

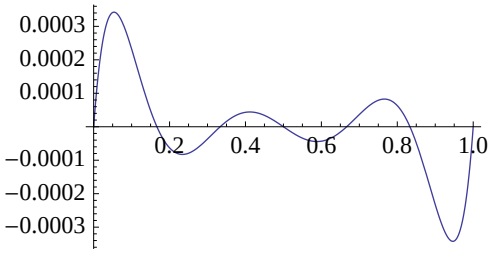

0.2 0.4 0.6 0.8 1.0
−0.05 −0.10 −0.15 −0.20 −0.25

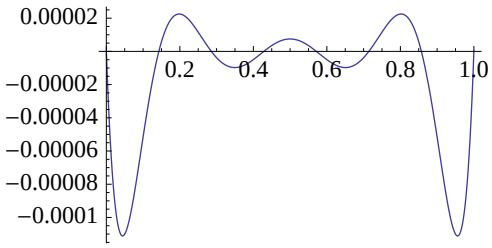

0.04 0.02
0.2 0.4 0.6 0.8 1.0
−0.02 −0.04

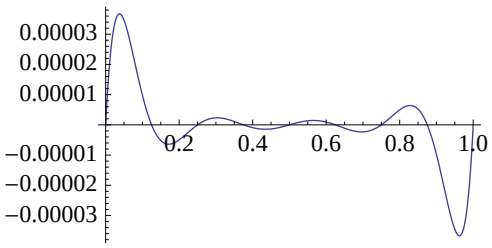

0.005
0.2 0.4 0.6 0.8 1.0
−0.005 −0.010

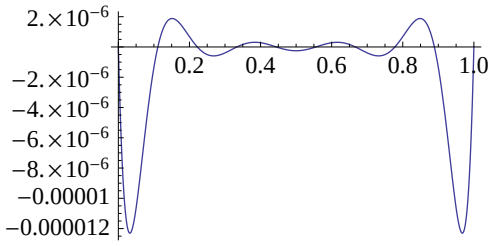

0.003 0.002 0.001
0.2 0.4 0.6 0.8 1.0
−0.001 −0.002 −0.003

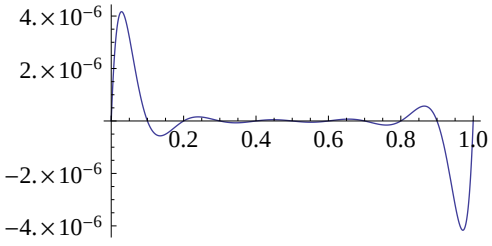

0.0002
0.2 0.4 0.6 0.8 1.0
−0.0002 −0.0004 −0.0006 −0.0008 −0.0010


0.0003 0.0002 0.0001
0.2 0.4 0.6 0.8 1.0
−0.0001 −0.0002 −0.0003


0.00002
0.2 0.4 0.6 0.8 1.0
−0.00002 −0.00004 −0.00006 −0.00008 −0.0001


0.00003 0.00002 0.00001
0.2 0.4 0.6 0.8 1.0
−0.00001 −0.00002 −0.00003


2.×10−6
0.2 0.4 0.6 0.8 1.0
−2.×10−6 −4.×10−6 −6.×10−6 −8.×10−6 −0.00001 −0.000012


4.×10−6 2.×10−6
0.2 0.4 0.6 0.8 1.0
−2.×10−6 −4.×10−6

L'interpolation linéaire et l'interpolation quadratique étant souvent utilisées, calculons le maximum de $E(x)$ dans ces deux cas simples.

Proposition 4.2.7. *(a) On a*

$$\max_{q\in[0,1]} q(q-1) = \frac{1}{4}.$$

(b) On a

$$\max_{q\in[0,2]} q(q-1)(q-2) = \frac{2}{3\sqrt{3}}.$$

Démonstration. (a) Posons $g(q) = q(q-1) = q^2 - q$. Il vient

$$g'(q) = 2q - 1.$$

Ainsi, $g'(q) = 0$ si et seulement si $q = 1/2$ et

$$\max_{q\in[0,1]} |g(q)| = \max(|g(0)|, |g(1/2)|, |g(1)|) = 1/4.$$

(b) Posons $g(q) = q(q-1)(q-2) = q^3 - 3q^2 + 2q$. Il vient

$$g'(q) = 3q^2 - 6q + 2.$$

Ainsi,

$$g'(q) = 0 \Leftrightarrow q = \frac{3 \pm \sqrt{9-6}}{3} = 1 \pm \frac{1}{\sqrt{3}}.$$

Or,

$$g(1 \pm \frac{1}{\sqrt{3}}) = \left(1 \pm \frac{1}{\sqrt{3}}\right)\left(\pm\frac{1}{\sqrt{3}}\right)\left(-1 \pm \frac{1}{\sqrt{3}}\right) = \left(\frac{1}{3} - 1\right)\left(\pm\frac{1}{\sqrt{3}}\right) = \mp\frac{2}{3\sqrt{3}}$$

et $g(0) = g(1) = g(2) = 0$. Donc,

$$\max_{q\in[0,2]} |g(q)| = \frac{2}{3\sqrt{3}}.$$

□

Exemple 4.2.8. Utilisons les résultats ci-dessus pour construire une table de $\sin x$ pour x entre $0°$ et $90°$ permettant d'obtenir $\sin x$ à 2 décimales par interpolation quadratique. Si h est le pas en radiant de cette table, il suffit que

$$\frac{h^3}{3!}\frac{2}{3\sqrt{3}} < \frac{1}{2}10^{-2} \Leftrightarrow h < 0.427.$$

Le pas en degré de la table doit donc être inférieur à 24°. Prenons ce pas égal à 22.5°. On a à trois décimales

$$\sin(22.5°) = \sqrt{\frac{1-\cos(45°)}{2}} = \frac{\sqrt{2-\sqrt{2}}}{2} = 0.383$$
$$\sin(45°) = 0.707$$
$$\sin(67.5°) = 0.924$$

d'où la table

x_k	$y_k = \sin x_k$	Δy_k	$\Delta^2 y_k$
0°	0.000		
		0.383	
22.5°	0.383		−0.059
		0.324	
45°	0.707		−0.107
		0.217	
67.5°	0.924		−0.141
		0.076	
90°	1.000		

Pour $x_0 = 0°$, $x_1 = 22.5°$, $x_2 = 45°$, on a

$$P(x) = 0.383q - 0.059\frac{q(q-1)}{2} \qquad (q = \frac{x}{22.5}).$$

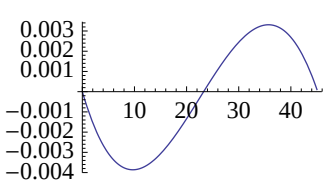
Pour $x_0 = 22.5°$, $x_1 = 45°$, $x_2 = 67.5°$, on a

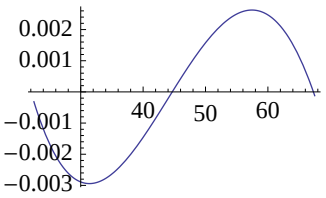
$$P(x) = 0.383 + 0.324q - 0.107\frac{q(q-1)}{2} \qquad (q = \frac{x-22.5}{22.5}).$$

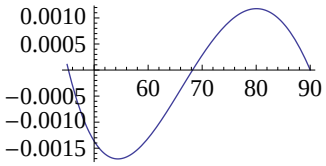
Pour $x_0 = 45°$, $x_1 = 67.5°$, $x_2 = 90°$, on a

$$P(x) = 0.707 + 0.217q - 0.141\frac{q(q-1)}{2} \qquad (q = \frac{x-45}{22.5}).$$

On trouvera, ci-dessous, les graphes des erreurs associées.







4.3 Interpolation de Tchebycheff

Définition 4.3.1. Pour tout $n \in \mathbb{N}$, on pose

$$T_n(x) = \cos(n \arccos x)$$

pour tout $x \in [-1, 1]$.

Proposition 4.3.2. *On a*

(a) $T_0(x) = 1$;

(b) $T_1(x) = x$;

(c) $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$ $(n \geq 1)$.

En particulier, $T_n(x)$ *est un polynôme de degré* n *en* x *dont le coefficient de* x^n *est* 2^{n-1} *si* $n \geq 1$. *De plus,* $T_n(x)$ *est pair (resp. impair) si* n *est pair (resp. impair).*

Démonstration. (a) et (b) sont évidents. Pour obtenir (c), il suffit de remarquer que pour $\theta = \arccos x$ $(x \in [-1, 1])$, on a

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

et par conséquent,

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

La conclusion concernant le caractère polynomial de $T_n(x)$, sa parité et la valeur du coefficient de x^n s'obtient ensuite par une récurrence directe. □

Exemples 4.3.3. On a

$$T_2(x) = 2x^2 - 1, \qquad T_3(x) = 4x^3 - 3x, \qquad T_4(x) = 8x^4 - 8x^2 + 1.$$

Proposition 4.3.4. *(a) Le polynôme* $T_n(x)$ *a* n *zéros distincts. Ces zéros sont situés dans* $[-1, 1]$ *et sont donnés par*

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

(b) Le polynôme $T_n(x)$ *a* $(n+1)$ *extrema locaux dans* $[-1, 1]$. *Ces extrema sont aussi globaux dans* $[-1, 1]$ *et sont donnés par*

$$x'_k = \cos\left(\frac{k\pi}{n}\right) \qquad (k = 0, \ldots, n).$$

De plus, on a

$$T_n(x'_k) = (-1)^k.$$

Démonstration. (a) Pour $x \in [-1,1]$, on a

$$\cos(n \arccos x) = 0$$

si et seulement si

$$n \arccos x = \frac{\pi}{2} + k\pi \qquad (k \in \mathbb{Z}).$$

Comme $\arccos x \in [0,\pi]$, les seuls $x \in [-1,1]$ qui annulent $T_n(x)$ sont ceux pour lesquels

$$\arccos x = \frac{\pi}{2n} + k\frac{\pi}{n} \qquad (k = 0,\ldots,n-1).$$

Il s'ensuit que les zéros de $T_n(x)$ situés dans $[-1,1]$ sont les

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

Comme $T_n(x)$ est un polynôme de degré n, ces x_k sont les seuls zéros de $T_n(x)$.

(b) On sait que

$$\cos n\theta \in [-1,1]$$

et que

$$\cos n\theta = 1 \Leftrightarrow \theta = \frac{2k\pi}{n} \quad (k \in \mathbb{Z})$$
$$\cos n\theta = -1 \Leftrightarrow \theta = \frac{(2k+1)\pi}{n} \quad (k \in \mathbb{Z}).$$

Il s'ensuit que sur $[-1,1]$, $T_n(x)$ atteint son maximum global 1 en

$$x'_{2k} = \cos\left(\frac{2k\pi}{n}\right) \qquad k = 0,\ldots,\left[\frac{n}{2}\right]$$

et son minimum global -1 en

$$x'_{2k+1} = \cos\left(\frac{2k+1}{n}\pi\right) \qquad k = 0,\ldots,\left[\frac{n-1}{2}\right].$$

La conclusion en résulte. □

Proposition 4.3.5 (Propriété de minimax)**.** *Pour tout polynôme $P(x)$ de degré $n \geq 1$ et de coefficient dominant* 1*, on a*

$$\max_{x\in[-1,1]} |P(x)| \geq \max_{x\in[-1,1]} |2^{1-n}T_n(x)| = 2^{1-n}.$$

Démonstration. Procédons par l'absurde. Si ce n'est pas vrai, il existe un polynôme $P(x)$ de degré $n \geq 1$ et de coefficient dominant 1 tel que

$$|P(x)| < 2^{1-n}$$

pour tout $x \in [-1, 1]$. Pour un tel polynôme, on a alors

$$\begin{aligned}
&P(x'_0) < 2^{1-n} T_n(x'_0) \\
&P(x'_1) > 2^{1-n} T_n(x'_1) \\
&P(x'_2) < 2^{1-n} T_n(x'_2) \\
&\qquad \vdots \\
&P(x'_n) \begin{cases} < 2^{1-n} T_n(x'_n) & \text{si } n \text{ est pair} \\ > 2^{1-n} T_n(x'_n) & \text{si } n \text{ est impair} \end{cases}
\end{aligned}$$

Il s'ensuit que le polynôme

$$Q(x) = P(x) - 2^{1-n} T_n(x)$$

change au moins n fois de signe sur $[-1, 1]$. Ce polynôme possède donc au moins n zéros distincts dans $]-1, 1[$. Comme le coefficient de x^n dans $P(x)$ et dans $2^{1-n} T_n(x)$ est 1, $Q(x)$ est de degré inférieur ou égal à $n - 1$. Il s'ensuit que $Q = 0$; d'où une contradiction puisqu'alors

$$P(x) = 2^{1-n} T_n(x).$$

□

Corollaire 4.3.6. *Si $x_0, \ldots, x_n$ sont $n + 1$ points distincts de $[a, b]$,*

$$M = \max_{x \in [a,b]} |(x - x_0) \ldots (x - x_n)|$$

est minimum lorsque

$$x_k = \frac{b + a}{2} + \frac{b - a}{2} \cos\left(\frac{2k + 1}{2n + 2}\pi\right).$$

Dans ce cas,

$$M = 2\left(\frac{b - a}{4}\right)^{n+1}.$$

Démonstration. Considérons d'abord le cas où $a = -1$, $b = 1$. Comme $P(x) = (x - x_0) \ldots (x - x_n)$ est un polynôme de degré $n + 1$ et de coefficient dominant 1, la proposition précédente montre que

$$\max_{x \in [-1,1]} |P(x)| \geq \max_{x \in [-1,1]} |2^{-n} T_{n+1}(x)| = 2^{-n}.$$

Or,

$$2^{-n}T_{n+1}(x) = (x - x_0)\dots(x - x_n)$$

pour

$$x_k = \cos\left(\frac{2k+1}{2n+2}\pi\right) \qquad (k = 0, \dots, n),$$

d'où la conclusion. Dans le cas général, il suffit de remarquer que

$$\begin{aligned}
&\max_{x\in[a,b]} |(x - x_0)\dots(x - x_n)| \\
&\quad = \max_{t\in[-1,1]} \left|\left(\frac{b+a}{2} + \frac{b-a}{2}t - x_0\right)\dots\left(\frac{b+a}{2} + \frac{b-a}{2}t - x_n\right)\right| \\
&\quad = \left(\frac{b-a}{2}\right)^{n+1} \max_{t\in[-1,1]} |(t - t_0)\dots(t - t_n)|
\end{aligned}$$

avec

$$t_k = \left(x_k - \frac{b+a}{2}\right) \Big/ \left(\frac{b-a}{2}\right) \qquad (k = 0, \dots, n)$$

et d'utiliser le résultat obtenu lorsque $a = -1$, $b = 1$. □

Remarque 4.3.7. On a vu que si on remplace f de classe C_{n+1} sur $[a, b]$ par son polynôme d'interpolation aux points $x_0, \dots, x_n$, l'erreur absolue commise est

$$\left|(x - x_0)\dots(x - x_n)\frac{f^{(n+1)}(\xi)}{(n+1)!}\right|$$

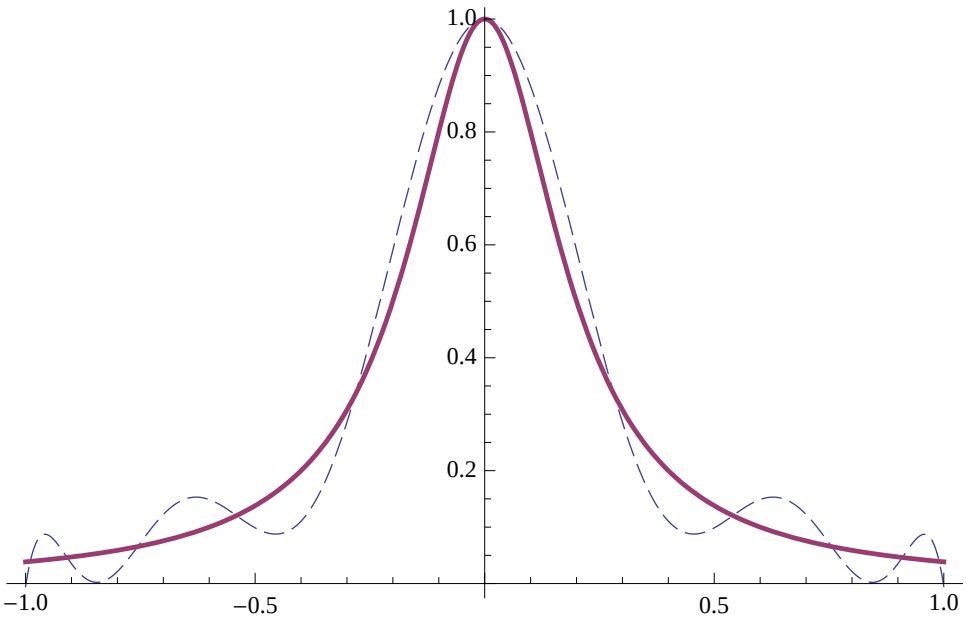
pour un certain $\xi \in [a, b]$. Le corollaire précédent montre que le choix de

$$x_k = \frac{b+a}{2} + \frac{b-a}{2}\cos\left(\frac{2k+1}{2n+2}\pi\right)$$

fait en sorte que le facteur qui ne dépend pas de f soit globalement le plus petit possible. Le polynôme d'interpolation correspondant est appelé *polynôme d'interpolation de Tchebycheff* et les x_k sont les *abscisses de Tchebycheff* associées à l'intervalle $[a, b]$. Pour ce type d'interpolation, l'erreur absolue commise est majorée par

$$2\left(\frac{b-a}{4}\right)^{n+1} \sup_{\xi\in[a,b]} \frac{|f^{(n+1)}(\xi)|}{(n+1)!}.$$

Contrairement à ce qui se passe dans le cas de l'interpolation à pas constant, on peut montrer que le polynôme d'interpolation de Tchebycheff de $f \in C_1([a, b])$ tend uniformément vers f sur $[a, b]$ si $n \to \infty$. Par exemple dans le cas considéré dans la Remarque 4.1.16, on constate sur la figure ci-dessous que l'interpolation de Tchebycheff est bien meilleure que l'interpolation à pas constant.



Exemple 4.3.8. Pour $f(x) = e^x$ et $[a, b] = [0, 1]$, l'erreur absolue associée à l'interpolation de Tchebycheff de degré $n = 5$ est majorée par

$$2\frac{1}{(4)^6}\frac{e}{6!} \approx 2\ 10^{-6}.$$

4.4 Stabilité de l'interpolation polynomiale

Définition 4.4.1. Soient $x_0, \ldots, x_n$ des points distincts de $[a, b]$. La *fonction de Lebesgue* associée à ces points est la fonction

$$\lambda_n(x) = \sum_{j=0}^{n} |L_j(x)|$$

où $L_j(x)$ est le $j^{\text{ème}}$ polynôme de Lagrange associé aux points $x_0, \ldots, x_n$. Le *nombre de Lebesgue* associé à ces points est

$$\Lambda_n = \sup_{x\in[a,b]} |\lambda_n(x)|.$$

Remarque 4.4.2. Comme

$$\sum_{j=0}^{n} L_j(x)$$

interpole la fonction 1 en $x_0, \ldots, x_n$ qui est un polynôme de degré $\leq n$, on a

$$1 = \sum_{j=0}^{n} L_j(x).$$

Il s'ensuit que

$$1 = \left| \sum_{j=0}^{n} L_j(x) \right| \leq \sum_{j=0}^{n} |L_j(x)| = \lambda_n(x).$$

En particulier, on a $\Lambda_n \geq 1$.

Proposition 4.4.3. *Soient $x_0, \ldots, x_n$ des points distincts de $[a, b]$ et, pour toute suite de réels $y_0, \ldots, y_n$, soit*

$$P(y; x)$$

le polynôme d'interpolation associé à $x_0, \ldots, x_n$ et $y_0, \ldots, y_n$. Alors,

$$\max_{x \in [a,b]} |P(y; x) - P(y + \Delta y; x)| \leq \Lambda_n \max_{j \in \{0,\ldots,n\}} |\Delta y_j|$$

pour toute suite de réels $\Delta y_0, \ldots, \Delta y_n$. De plus, Λ_n est la plus petite constante à jouir de cette propriété.

Démonstration. On a

$$P(y; x) = \sum_{j=0}^{n} y_j L_j(x)$$

et

$$P(y + \Delta y; x) = \sum_{j=0}^{n} (y_j + \Delta y_j) L_j(x).$$

Il s'ensuit que

$$P(y + \Delta y; x) - P(y; x) = \sum_{j=0}^{n} \Delta y_j L_j(x)$$

et que

$$|P(y + \Delta y; x) - P(y; x)| \leq \left(\sum_{j=0}^{n} |L_j(x)| \right) \max_{j \in \{0,\ldots,n\}} |\Delta y_j|$$

et on obtient la majoration annoncée. Si on a

$$\max_{x \in [a,b]} |P(y; x) - P(y + \Delta y; x)| \leq C \max_{j \in \{0,\ldots,n\}} |\Delta y_j|$$

pour tout $\Delta y_0, \ldots, \Delta y_n$, on a en particulier

$$\sum_{j=0}^{n} \Delta y_j L_j(x) \leq C$$

si $|\Delta y_0| = \cdots = |\Delta y_n| = 1$. Posons $\Delta y_j = \operatorname{sgn} L_j(x)$ si $L_j(x) \neq 0$ et $\Delta y_j = 1$ sinon. Il vient alors

$$\sum_{j=0}^{n} |L_j(x)| \leq C.$$

Par conséquent, $\lambda_n(x) \leq C$ pour tout $x \in [a, b]$ et on a

$$\Lambda_n \leq C.$$

□

Remarque 4.4.4. La proposition précédente montre que Λ_n est en quelque sorte le nombre de conditionnement du problème de l'interpolation polynomiale.

On peut montrer que

$$\Lambda_n \sim \frac{2^{n+1}}{en \ln n} \qquad (n \to \infty)$$

si les points $x_0, \ldots, x_n$ sont équidistants. En particulier, dans ce cas, Λ_n tend exponentiellement vers ∞ si n tend vers ∞. Le problème de l'interpolation à pas constant est donc numériquement très instable pour n grand. C'est là encore une des facettes du phénomène de Runge.

Si les points $x_0, \ldots, x_n$ sont les abscisses de Tchebycheff, alors on peut montrer que

$$\Lambda_n \sim \frac{2}{\pi} \ln n \qquad (n \to \infty).$$

Dans ce cas, Λ_n tend vers ∞ de manière beaucoup moins rapide que dans le cas de l'interpolation à pas constant. Le problème de l'interpolation de Tchebycheff reste donc relativement stable même pour n assez grand.

Si on désigne par Λ_n^o la borne inférieure des constantes de Lebesgue associée aux $(n+1)$-uples de points de $[a, b]$, on peut montrer que

$$\Lambda_n^o \sim \frac{2}{\pi} \ln n \qquad (n \to \infty).$$

C'est pourquoi on considère souvent que l'interpolation de Tchebycheff est quasi-optimale.

4.5 Relations avec l'approximation polynomiale

Définition 4.5.1. Soit f une fonction continue sur $[a, b]$. Posons

$$E_n(f) = \inf_{P \in \mathbb{R}[x]_n} \max_{x \in [a,b]} |f(x) - P(x)|$$

où $\mathbb{R}[x]_n$ désigne l'ensemble des polynômes réels de degré $\leq n$.

Remarque 4.5.2. Le nombre $E_n(f)$ est en fait la distance de f à $\mathbb{R}[x]_n$ dans l'espace des fonctions continues sur $[a,b]$ muni de la métrique

$$d(f,g) = \max_{x\in[a,b]} |f(x) - g(x)|.$$

Proposition 4.5.3. *Si f est une fonction continue sur $[a,b]$, il existe $P \in \mathbb{R}[x]_n$ tel que*

$$d(f,P) = \max_{x\in[a,b]} |f(x) - P(x)| = E_n(f).$$

Démonstration. L'ensemble $\mathbb{R}[x]_n$ est un espace vectoriel E de dimension $n+1$ et l'inégalité triangulaire montre que

$$P \mapsto d(f,P)$$

est continue sur E. Si $\rho > E_n(f)$, il existe $P_0 \in E$ tel que

$$d(f,P_0) < \rho.$$

Si $P \in E$ est tel que $d(P,P_0) > 2\rho$, on a alors

$$d(f,P) \geq d(P,P_0) - d(f,P_0) > 2\rho - \rho = \rho.$$

Il s'ensuit que

$$E_n(f) = \inf_{\{P:d(P,P_0)\leq 2\rho\}} d(P,f).$$

Comme $\{P : d(P,P_0) \leq 2\rho\}$ est borné et fermé dans $\mathbb{R}[x]_n$, la fonction continue

$$P \mapsto d(P,f)$$

y atteint son minimum ; d'où la conclusion. □

Remarque 4.5.4. On peut réénoncer le théorème d'approximation de Weierstrass en disant que

$$\lim_{n\to\infty} E_n(f) = 0.$$

Exemple 4.5.5. Soit $n \geq 1$. Pour $f(x) = x^{n+1}$ et $[a,b] = [-1,1]$, on a

$$E_n(f) = d(f,P_n) = 2^{-n}$$

avec $P_n = x^{n+1} - 2^{-n}T_{n+1}(x)$. En effet, tout polynôme Q de degré inférieur ou égal à n peut s'écrire sous la forme

$$Q = x^{n+1} - R$$

où R est un polynôme de degré $n+1$ et de coefficient dominant 1. Dans ce cas, on a

$$d(f,Q) = \max_{x\in[-1,1]} |R(x)|.$$

La conclusion résulte donc de la propriété de minimax des polynômes de Tchebycheff. Comme P_n est en fait de degré $n-1$, on a aussi

$$E_{n-1}(f) = E_n(f) = 2^{-n}.$$

Remarque 4.5.6. L'exemple précédent est à la base d'une technique permettant de réduire le degré d'une approximation polynomiale. Par exemple, supposons que

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

pour $x \in [-1,1]$ et considérons l'approximation polynomiale de f donnée par la somme partielle

$$\sum_{n=0}^{N} a_n x^n.$$

Dans cette somme, remplaçons x^N par le polynôme P_{N-1} considéré dans l'exemple précédent. On obtient alors une nouvelle approximation polynomiale de f qui ne contient plus de terme en x^N. En itérant ce procédé, on arrive aisément à des approximations de f du type

$$\sum_{n=0}^{M} b_n x^n$$

avec $M < N$ qui sont souvent meilleures que les sommes partielles

$$\sum_{n=0}^{M} a_n x^n.$$

Exemple 4.5.7. Travaillons sur $[-1,1]$. On a

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots.$$

Ainsi l'erreur commise en approchant $\cos x$ par

$$1 - \frac{x^2}{2!} \qquad (*)$$

est inférieure à $1/24 \approx 0.042$. L'erreur commise en approchant $\cos x$ par

$$1 - \frac{x^2}{2} + \frac{x^4}{4!}$$

est quant à elle inférieure à $1/(6!) \approx 0.0014$. On a

$$P_3(x) = x^4 - 2^{-3}T_4(x) = x^4 - x^4 + x^2 - \frac{1}{8} = x^2 - \frac{1}{8}$$

et

$$|x^4 - P_3(x)| \leq \frac{1}{8}.$$

Ainsi,

$$1 - \frac{x^2}{2} + \frac{1}{24}\left(x^2 - \frac{1}{8}\right) = \frac{191}{192} - \frac{11}{24}x^2$$

approche $\cos x$ à moins de $1/(6!) + 1/(8 \cdot 24) \approx 0.007$; ce qui est nettement mieux que la précision obtenue avec (*). Remarquons cependant que par cette formule, $\cos 0$ est approché par $\dfrac{191}{192}$ alors que l'on sait très bien que $\cos 0 = 1$. Si on souhaite préserver cette propriété, on peut utiliser plutôt le développement

$$\frac{\cos x - 1}{x} = \frac{-x}{2!} + \frac{x^3}{4!} - \frac{x^5}{6!} + \dots$$

Comme $P_2(x) = x^3 - 2^{-2}T_3(x) = x^3 - x^3 + \dfrac{3}{4}x = \dfrac{3}{4}x$, on arrive à l'approximation

$$1 + x\left(\frac{-x}{2} + \frac{1}{32}x\right) = 1 - \frac{15}{32}x^2$$

dont l'erreur est inférieure à $1/(6!) + 1/(4 \cdot 4!) \approx 0.012$.

Proposition 4.5.8. *Soient $x_0, \dots, x_n$ des points de $[a, b]$ et soit $f \in C_0([a, b])$. Alors, on a*

$$\max_{x \in [a,b]} |f(x) - P(y; x)| \leq (1 + \Lambda_n) E_n(f)$$

si $y_0 = f(x_0), \dots, y_n = f(x_n)$.

Démonstration. Soit $Q(x)$ un polynôme de degré $\leq n$ et soient $z_0 = Q(x_0), \dots, z_n = Q(x_n)$. On a alors

$$Q(x) = P(z; x)$$

et par conséquent,

$$\begin{aligned} \max_{x \in [a,b]} |Q(x) - P(y; x)| &\leq \Lambda_n \max_{j \in \{0,\dots,n\}} |z_j - y_j| \\ &\leq \Lambda_n \max_{x \in [a,b]} |Q(x) - f(x)|. \end{aligned}$$

Ainsi,

$$\begin{aligned}\max_{x\in[a,b]}|f(x)-P(y;x)| &\leq \max_{x\in[a,b]}|f(x)-Q(x)|+\max_{x\in[a,b]}|Q(x)-P(y;x)|\\ &\leq (1+\Lambda_n)\max_{x\in[a,b]}|Q(x)-f(x)|.\end{aligned}$$

Comme cette majoration est vraie pour tout polynôme Q de degré $\leq n$, la conclusion en découle. □

Pour conclure cette section, donnons sans démonstration le résultat suivant :

Proposition 4.5.9 (Jackson)**.** *Si f est de classe C_k sur $[a,b]$, alors*

$$E_n(f) \leq \frac{\pi^k}{(n+1)n\dots(n-k+2)}\left(\frac{b-a}{4}\right)^k \sup_{\xi\in[a,b]}|f^{(k)}(\xi)|$$

si $n \geq k$. En particulier,

$$E_n(f) = O(n^{-k}) \qquad (n\to\infty).$$

Remarque 4.5.10. Soit $f \in C_1([a,b])$. En combinant les deux propositions précédentes avec la Remarque 4.4.4, on comprend pourquoi les polynômes d'interpolation de Tchebycheff tendent vers f si $n \to \infty$.

4.6 Régression polynomiale

Soient $x_0,\dots,x_n$ des réels deux à deux distincts et soient $y_0,\dots,y_n$ des réels. Plutôt que de chercher un polynôme $P(x)$ pour lequel on a exactement

$$P(x_0)=y_0,\dots,P(x_n)=y_n$$

comme dans le problème de l'interpolation polynomiale, on peut chercher un polynôme $P(x)$ de degré $\leq m$ fixé pour lequel la somme

$$\sum_{j=0}^{n}|y_j-P(x_j)|^2$$

soit la plus petite possible. Le problème auquel on aboutit est alors appelé *problème de régression polynomiale de degré $\leq m$* ou encore *problème des moindres carrés pour les polynômes de degré $\leq m$.*

Comme tout polynôme $P(x)$ de degré $\leq m$ peut s'écrire

$$P(x) = a_m x^m + \dots + a_1 x + a_0$$

le problème de la régression polynomiale de degré $\leq m$ revient à déterminer les réels $a_0, \ldots, a_m$ qui rendent

$$\sum_{j=0}^{n} \left| y_j - \sum_{k=0}^{m} a_k x_j^k \right|^2$$

minimum. Si on pose

$$Y = \begin{pmatrix} y_0 \\ \vdots \\ y_n \end{pmatrix} \qquad \text{et} \qquad V = \begin{pmatrix} 1 & x_0 & \cdots & x_0^m \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^m \end{pmatrix}$$

et si on note $\|\cdot\|$ la norme euclidienne usuelle, le problème revient donc à trouver les vecteurs

$$A = \begin{pmatrix} a_0 \\ \vdots \\ a_m \end{pmatrix}$$

pour lesquels

$$\|Y - VA\|$$

est minimum. Comme

$$L = \{VA : A \in \mathbb{R}^{m+1}\}$$

est un sous-espace vectoriel de $\mathbb{R}^{m+1}$, on sait par la géométrie que $\|Y - VA\|$ est minimum lorsque la droite joignant Y à VA est orthogonale à L. En effet, dans ce cas on a

$$\begin{aligned} \|Y - VB\|^2 &= \|(Y - VA) + V(A - B)\|^2 \\ &= \|Y - VA\|^2 + \|V(A - B)\|^2 \\ &\geq \|Y - VA\|^2 \end{aligned}$$

pour tout $B \in \mathbb{R}^{m+1}$. Comme l'orthogonalité entre L et la droite joignant Y à VA se traduit par la condition

$$\langle Y - VA, VB \rangle = 0 \qquad \forall B \in \mathbb{R}^{m+1},$$

elle est équivalente à la condition

$$\tilde{V}(Y - VA) = 0.$$

On arrive donc au résultat suivant :

Proposition 4.6.1. *Soient $x_0, \ldots, x_n$ des réels deux à deux distincts et soient $y_0, \ldots, y_n$ des réels quelconques. Posons*

$$Y = \begin{pmatrix} y_0 \\ \vdots \\ y_n \end{pmatrix} \qquad et \qquad V = \begin{pmatrix} 1 & x_0 & \cdots & x_0^m \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^m \end{pmatrix}$$

et supposons que $m \leq n$. Alors, le polynôme

$$a_m x^m + \cdots + a_0$$

est solution du problème de la régression polynomiale de degré $\leq m$ si et seulement si le vecteur

$$A = \begin{pmatrix} a_0 \\ \vdots \\ a_m \end{pmatrix}$$

est l'unique solution du système linéaire

$$(\tilde{V}V)A = (\tilde{V}Y).$$

Démonstration. Vu ce qui précède, la seule chose à justifier est le caractère déterminé du système

$$(\tilde{V}V)A = (\tilde{V}Y).$$

Comme ce système a $(m+1)$ lignes et $(m+1)$ inconnues, il suffit de montrer qu'une relation du type

$$(\tilde{V}V)B = 0$$

avec $B \in \mathbb{R}^{m+1}$ entraîne que $B = 0$. Or, une relation de ce type entraîne que

$$\|VB\|^2 = \tilde{B}\tilde{V}VB = 0$$

et donc que $VB = 0$. Vu la forme de V, cette égalité signifie que le polynôme

$$b_m x^m + \cdots + b_1 x + b_0$$

s'annule en les $n+1$ points $x_0, \ldots, x_n$. Comme $n \geq m$, cela ne peut arriver que si

$$b_m = 0, \ldots, b_1 = 0, b_0 = 0.$$

□

Remarque 4.6.2. Dans les conditions du résultat précédent, la matrice $\tilde{V}V$ est en fait hermitienne définie positive car on a

$$\left\langle B, \tilde{V}VB \right\rangle = \|VB\|^2 > 0$$

si $B \neq 0$. On peut donc résoudre aisément le système

$$(\tilde{V}V)A = \tilde{V}Y$$

par la méthode de Choleski. Remarquons également que puisque $V_{jk} = x_{j-1}^{k-1}$, on a

$$(\tilde{V}V)_{jk} = \sum_{l=1}^{n+1} V_{lj}V_{lk} = \sum_{l=1}^{n+1} x_{l-1}^{j-1}x_{l-1}^{k-1} = \sum_{l=0}^{n} x_l^{j+k-2}.$$

Si on convient de poser

$$\mu_p = \sum_{l=0}^{n} x_l^p$$

pour $p = 0, \ldots, 2m$, on a donc

$$(\tilde{V}V)_{jk} = \mu_{j+k-2};$$

ce qui montre qu'il est suffisant de calculer les $2m+1$ réels $\mu_0, \ldots, \mu_{2m}$ pour connaître explicitement les $(m+1)^2$ éléments de $\tilde{V}V$.

Remarque 4.6.3. Une solution du problème de la régression polynomiale de degré $\leq m$ fournit un polynôme P de degré $\leq m$ pour lequel on n'a pas toujours

$$y_j = P(x_j)$$

pour tout $j \in \{0, \ldots, n\}$. Cependant la moyenne des écarts

$$y_j - P(x_j)$$

est nulle. En effet, pour $B = E_1$, on a

$$VB = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

et la relation

$$\langle Y - VA, VB \rangle = 0$$

entraîne que

$$\sum_{j=0}^{n}(y_j - P(x_j)) = 0.$$

Cela étant, l'expression

$$\sum_{j=0}^{n}(y_j - P(x_j))^2$$

représente la variance de la distribution des écarts. Cette variance est donc inférieure à celle que l'on obtiendrait si on remplaçait P par un autre polynôme de degré $\leq m$ donnant une distribution des écarts de moyenne nulle.

5 Intégration

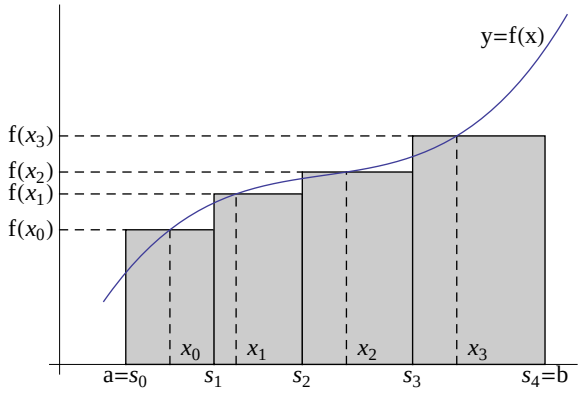
5.1 Méthode des rectangles

Soit f une fonction continue sur $[a,b] \subset \mathbb{R}$. On sait par l'interprétation de Cauchy-Riemann de l'intégrale que

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \sum_{k=0}^{n-1} (s_{k+1}^{(n)} - s_k^{(n)}) f(x_k^{(n)})$$

si $a = s_0^{(n)} < s_1^{(n)} < \cdots < s_n^{(n)} = b$, si $x_k^{(n)} \in \left[s_k^{(n)}, s_{k+1}^{(n)}\right]$ et si

$$\lim_{n\to\infty} \max_{k\in\{0,\ldots,n-1\}} |s_{k+1}^{(n)} - s_k^{(n)}| = 0.$$



En d'autres termes, on peut approcher l'intégrale considérée à la précision voulue par la somme des aires (algébriques) des rectangles de base $s_{k+1}^{(n)} - s_k^{(n)}$ et de hauteur $f(x_k^{(n)})$ si on fait en sorte que l'écart maximum entre $s_{k+1}^{(n)}$ et $s_k^{(n)}$ soit suffisamment petit.

5.1.1 Méthode des rectangles de rapport ρ

Pour pouvoir tirer une méthode pratique d'intégration du résultat précédent, nous devons décider comment choisir les $s_k^{(n)}$ et les $x_k^{(n)}$. Une façon simple de procéder est de choisir une subdivision de $[a,b]$ en intervalles de même longueur et de faire en sorte que la position relative de $x_k^{(n)}$ dans $\left[s_k^{(n)}, s_{k+1}^{(n)}\right]$ soit constante. On est alors amener à choisir $\rho \in [0,1]$ et à poser

$$s_k^{(n)} = a + kh^{(n)} \qquad (k \in \{0,\ldots,n\})$$

$$x_k^{(n)} = s_k^{(n)} + \rho h^{(n)} \qquad (k \in \{0,\ldots,n-1\})$$

pour $h^{(n)} = (b-a)/n$. L'approximation de

$$\int_a^b f(x)\,dx$$

associée à ces choix est alors

$$R_\rho(f, h^{(n)}) = \sum_{k=0}^{n-1} h^{(n)} f(x_k^{(n)}).$$

Cette méthode d'intégration approchée est connue sous le nom de *méthode des rectangles de rapport* ρ.

Proposition 5.1.1 (Majoration de l'erreur locale)**.** *Soient $x \in \mathbb{R}$, $\varepsilon > 0$, $\rho \in [0,1]$ et f une fonction de classe C_1 sur*

$$I = [x - \rho\varepsilon, x + (1-\rho)\varepsilon]\,.$$

Posons

$$e(h) = f(x)h - \int_{x-\rho h}^{x+(1-\rho)h} f(t)\,dt$$

pour tout $h \in [0, \varepsilon]$. Alors, pour un tel h, on a

$$|e(h)| \leq \frac{\rho^2 + (1-\rho)^2}{2} \sup_{\xi \in I} |f'(\xi)| h^2.$$

En particulier, $e(h) = O(h^2)$ pour $h \to 0^+$.

Démonstration. On a

$$e(h) = f(x)h - \int_{x-\rho h}^{x+(1-\rho)h} f(t)\,dt$$

donc $e(h)$ est de classe C_2 sur $[0, \varepsilon]$ et on a

$$e'(h) = f(x) - f(x + (1-\rho)h)(1-\rho) + f(x - \rho h)(-\rho)$$

et

$$e''(h) = \rho^2 f'(x - \rho h) - (1-\rho)^2 f'(x + (1-\rho)h).$$

On en tire que $e(0) = e'(0) = 0$ et la formule de Taylor montre qu'il existe $\theta \in [0,1]$ tel que

$$e(h) = \frac{e''(\theta h)}{2} h^2.$$

En particulier,

$$|e(h)| \leq \frac{\rho^2 + (1-\rho)^2}{2} \sup_{\xi \in I} |f'(\xi)| h^2$$

d'où la conclusion. □

Proposition 5.1.2 (Majoration de l'erreur globale)**.** *Soit f une fonction de classe C_1 sur*

$$I = [a, b] \subset \mathbb{R},$$

et soit $h = (b-a)/n$ avec $n \in \mathbb{N}_0$. Posons

$$E(h) = R_\rho(f, h) - \int_a^b f(x)\, dx.$$

Alors

$$|E(h)| \leq \frac{\rho^2 + (1-\rho)^2}{2}(b-a) \sup_{\xi \in I} |f'(\xi)| h.$$

En particulier, $E(h) = O(h)$ pour $h \to 0^+$.

Démonstration. En laissant tomber l'exposant (n) dans les notations, il vient

$$\begin{aligned} E(h) &= \sum_{k=0}^{n-1} f(x_k)h - \int_a^b f(x)\, dx \\ &= \sum_{k=0}^{n-1} f(x_k)h - \sum_{k=0}^{n-1} \int_{s_k}^{s_{k+1}} f(x)\, dx \\ &= \sum_{k=0}^{n-1} \left[f(x_k)h - \int_{x_k-\rho h}^{x_k+(1-\rho)h} f(x)\, dx \right]. \end{aligned}$$

Vu le résultat précédent, on a donc

$$|E(h)| \leq \sum_{k=0}^{n-1} \frac{\rho^2 + (1-\rho)^2}{2} \sup_{\xi \in I} |f'(\xi)| h^2 \leq \frac{\rho^2 + (1-\rho)^2}{2} \sup_{\xi \in I} |f'(\xi)| n h^2$$

d'où la conclusion. □

Proposition 5.1.3 (Terme d'ordre 2 de l'erreur locale)**.** *Plaçons nous dans les conditions de la proposition 5.1.1 et supposons que f soit de classe C_2 sur I. Alors,*

$$e(h) = \left(\rho - \frac{1}{2} \right) f'(x) h^2 + O(h^3)$$

où

$$O(h^3) = -\frac{\rho^3 + (1-\rho)^3}{6} f''(\xi) h^3.$$

pour un $\xi \in I$.

Démonstration. En gardant les notations introduites dans la preuve de la proposition 5.1.1, on voit que

$$e''(0) = (\rho^2 - (1-\rho)^2)f'(x) = (2\rho - 1)f'(x)$$

et que

$$e'''(h) = -\rho^3 f''(x - \rho h) - (1-\rho)^3 f''(x + (1-\rho)h).$$

et la conclusion résulte de la combinaison de la formule de Taylor et du théorème de la moyenne. □

Proposition 5.1.4 (Terme d'ordre 1 de l'erreur globale)**.** *Plaçons nous dans les conditions de la proposition 5.1.2 et supposons que f soit de classe C_2 sur I. Alors,*

$$E(h) = \left(\rho - \frac{1}{2}\right)(f(b) - f(a))h + O(h^2)$$

pour $h \to 0^+$.

Démonstration. On a

$$E(h) = \sum_{k=0}^{n-1} \left[f(x_k)h - \int_{x_k - \rho h}^{x_k + (1-\rho)h} f(t)\, dt \right].$$

Vu le résultat précédent, il vient

$$E(h) = \sum_{k=0}^{n-1} \left(\rho - \frac{1}{2}\right) f'(x_k)h^2 - \frac{\rho^3 + (1-\rho)^3}{6} f''(\xi_k)h^3.$$

On en tire que

$$\begin{aligned} E(h) &= \left(\rho - \frac{1}{2}\right) R_\rho(f', h)h - \sum_{k=0}^{n-1} \frac{\rho^3 + (1-\rho)^3}{6} f''(\xi_k)h^3 \\ &= \left(\rho - \frac{1}{2}\right) R_\rho(f', h)h - \frac{\rho^3 + (1-\rho)^3}{6}(b-a)f''(\xi)h^2 \end{aligned}$$

pour un $\xi \in I$. Or, vu la proposition 5.1.2

$$R_\rho(f', h) = \int_a^b f'(x)\, dx + O(h).$$

On a donc

$$E(h) = \left(\rho - \frac{1}{2}\right) \int_a^b f'(x)\, dx + O(h^2)$$

et la conclusion en découle. □

Les résultats précédents montre que la méthode des rectangles de rapport ρ est au moins d'ordre 1 et qu'elle est même en général exactement d'ordre 1 si $\rho \neq 1/2$ alors qu'elle est au moins d'ordre 2 lorsque $\rho = 1/2$. La méthode des rectangles de rapport $1/2$ est donc en général plus précise que les méthode des rectangles de rapport $\rho \neq 1/2$; c'est pourquoi nous allons étudier cette méthode plus en détails.

5.1.2 Méthode des rectangles de rapport $1/2$

Proposition 5.1.5 (Erreur locale)**.** *Soient $x \in \mathbb{R}$, $\varepsilon > 0$ et f une fonction de classe C_2 sur*

$$I = \left[x - \frac{\varepsilon}{2}, x + \frac{\varepsilon}{2}\right].$$

Posons

$$e(h) = f(x)h - \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f(t)\, dt$$

pour tout $h \in [0, \varepsilon]$. Alors, pour un tel h, il existe $\xi \in I$ tel que

$$e(h) = -\frac{1}{24} f''(\xi) h^3.$$

Démonstration. En reprenant les notations introduites dans la preuve de la proposition 5.1.1 et en posant $\rho = 1/2$, on voit que $e(h) \in C_3([0, \varepsilon])$ et que

$$e'''(h) = -\frac{1}{8}\left(f''\left(x - \frac{h}{2}\right) + f''\left(x + \frac{h}{2}\right)\right).$$

Comme $e''(0) = 0$, la formule de Taylor montre que

$$e(h) = -\frac{1}{24} \frac{f''\left(x - \frac{\theta h}{2}\right) + f''\left(x + \frac{\theta h}{2}\right)}{2} h^3$$

pour un $\theta \in [0, 1]$ bien choisi. Le théorème de la moyenne permet alors de conclure. □

Proposition 5.1.6 (Erreur globale)**.** *Soit f de classe C_2 sur $I = [a, b]$. Posons*

$$E(h) = R_{1/2}(f, h) - \int_a^b f(x)\, dx.$$

Alors il existe $\xi \in [a, b]$ tel que

$$E(h) = -\frac{b-a}{24} f''(\xi) h^2.$$

Démonstration. On sait que

$$E(h) = \sum_{k=0}^{n-1} \left[f(x_k)h - \int_{x_k-h/2}^{x_k+h/2} f(t)\ dt \right]$$

avec $x_k = a + (k + \frac{1}{2})h$. Vu la proposition précédente, il existe donc des $\xi_k \in [a, b]$ tels que

$$E(h) = \sum_{k=0}^{n-1} -\frac{1}{24} f''(\xi_k)h^3 = -\frac{b-a}{24} \left[\frac{1}{n} \sum_{k=0}^{n-1} f''(\xi_k) \right] h^2$$

et la conclusion résulte du théorème de la moyenne. □

Proposition 5.1.7 (Terme d'ordre 2 de l'erreur globale)**.** *Plaçons nous dans les conditions de la proposition précédente et supposons de plus que f soit de classe C_3 sur I. Alors,*

$$E(h) = -\frac{f'(b) - f'(a)}{24} h^2 + O(h^3)$$

si $h \to 0^+$.

Démonstration. Vu la forme de l'erreur locale, il est clair que

$$E(h) = \left[\sum_{k=0}^{n-1} -\frac{1}{24} f''(x_k)h^3 \right] + O(h^3).$$

Ainsi

$$E(h) = -\frac{1}{24} R_{1/2}(f'', h)h^2 + O(h^3).$$

Comme

$$R_{1/2}(f'', h) = \int_a^b f''(x)\ dx + O(h),$$

on voit que

$$E(h) = -\frac{1}{24} \left(\int_a^b f''(x)\ dx \right) h^2 + O(h^3)$$

d'où la conclusion. □

Exemple 5.1.8. Considérons l'intégrale

$$I = \int_0^1 e^{-x^2}\ dx.$$

Comme

$$e^{-x^2} = \sum_{n=0}^{\infty} \frac{(-x^2)^n}{n!} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} x^{2n},$$

on a

$$I = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \left[\frac{x^{2n+1}}{2n+1} \right]_0^1 = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \frac{1}{2n+1}.$$

Si on prend pour valeur approchée de I la somme

$$\sum_{n=0}^{N-1} \frac{(-1)^n}{n!(2n+1)},$$

des N premiers termes de la série précédente, l'erreur absolue est majorée par

$$\frac{1}{N!(2N+1)}.$$

Pour $N = 8$, cette erreur est inférieure à $1.5\ 10^{-6}$. On a donc

$$I \approx \sum_{n=0}^{7} \frac{(-1)^n}{n!(2n+1)} \approx 0.74682$$

à 5 décimales exactes. Si on approche I par la méthode des rectangles de rapport $1/2$ et de pas $h = 1/n$, l'erreur absolue est inférieure à

$$\varepsilon = \frac{h^2}{24} \sup_{\xi \in [0,1]} |f''(\xi)|$$

où $f(x) = e^{-x^2}$. Comme

$$f'(x) = -2xe^{-x^2}, \qquad f''(x) = (-2x)^2 e^{-x^2} - 2e^{-x^2} = 2(2x^2 - 1)e^{-x^2}$$

on a donc

$$\varepsilon = \frac{1}{12n^2}.$$

Si on veut obtenir 3 décimales exactes, il suffit par conséquent que

$$\frac{1}{12n^2} \leq 0.5\ 10^{-3}$$

ce qui a lieu si $n > \sqrt{500/3} \approx 12.9$, c'est-à-dire si $n \geq 13$. En effectuant les calculs pour $n = 1, \ldots, 14$, on trouve les tables suivantes :

n	$R_{1/2}(1/n)$	Erreur
1	0.778801	$\approx 3\ 10^{-2}$
2	0.754598	$\approx 8\ 10^{-3}$
3	0.750252	$\approx 3\ 10^{-3}$
4	0.748747	$\approx 2\ 10^{-3}$
5	0.748053	$\approx 1\ 10^{-3}$
6	0.747677	$\approx 9\ 10^{-4}$
7	0.747451	$\approx 6\ 10^{-4}$

n	$R_{1/2}(1/n)$	Erreur
8	0.747304	$\approx 5\ 10^{-4}$
9	0.747203	$\approx 4\ 10^{-4}$
10	0.747131	$\approx 3\ 10^{-4}$
11	0.747078	$\approx 3\ 10^{-4}$
12	0.747037	$\approx 2\ 10^{-4}$
13	0.747006	$\approx 2\ 10^{-4}$
14	0.746981	$\approx 2\ 10^{-4}$

La précision n'est donc pas essentiellement meilleure que celle donnée par l'estimation théorique.

Si l'on veut obtenir 5 décimales exactes, un calcul analogue au précédent montre qu'il suffit de prendre $n = 130$. On a alors effectivement

$$R_{1/2}(1/n) \approx 0.74682.$$

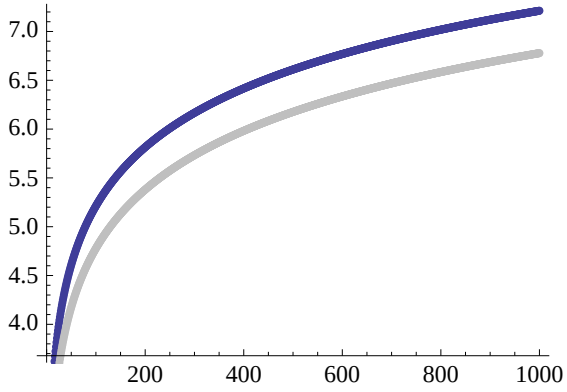
L'estimation théorique du "nombre" de décimales exactes pour un n donné est

$$\log_{10}(6) + 2\log_{10}(n)$$

puisque

$$\frac{1}{12n^2} \leq 0.5\ 10^{-d}$$

si et seulement si $6n^2 \geq 10^d$. Cette estimation est représentée en grisé sur la figue suivante. La valeur réelle correspondante est représentée en noir.



5.2 Accélération de la convergence de la méthode des rectangles

Comme on vient de le voir sur l'exemple précédent, la méthode des rectangles de rapport ρ reste une méthode assez lente même lorsque $\rho = 1/2$. Un moyen simple pour accélérer sa convergence et obtenir une méthode d'ordre au moins $p + 1$ est de soustraire à l'expression $R_\rho(f, h)$ les termes d'ordre inférieur ou égal à p de son développement asymptotique pour $h \to 0^+$. La clef pour obtenir ce développement asymptotique est fournie par une suite de polynômes particuliers dont l'origine remonte aux travaux de J. Bernoulli sur les sommes du type

$$1^p + 2^p + \cdots + n^p$$

où p et n sont des entiers naturels non-nuls.

5.2.1 Polynômes de Bernoulli

Définition 5.2.1. Les *polynômes de Bernoulli* sont les polynômes réels $B_k(x)$ ($k \in \mathbb{N}$) caractérisés par les relations :

$$\begin{aligned} &B_0(x) = 1\,; \\ &B_k'(x) = kB_{k-1}(x) \quad (k \geq 1)\,; \\ &\textstyle\int_0^1 B_k(x)\,dx = 0 \qquad (k \geq 1). \end{aligned}$$

Le *nombres de Bernoulli* sont quant à eux les réels

$$B_k = B_k(0) \qquad (k \in \mathbb{N})$$

Remarque 5.2.2. Il résulte de la définition précédente que

$$B_k(1) - B_k(0) = \int_0^1 B_k'(x)\,dx = k\int_0^1 B_{k-1}(x)\,dx = 0$$

si $k \geq 2$. On a donc aussi

$$B_k = B_k(1)$$

si $k \geq 2$.

Les polynômes de Bernoulli sont déterminés par les nombres de Bernoulli. En effet :

Proposition 5.2.3. *On a*

$$B_k(x) = \sum_{l=0}^{k} C_k^l B_{k-l}(x_0)(x - x_0)^l$$

pour tout $x_0 \in \mathbb{R}$. En particulier,

$$B_k(x) = \sum_{l=0}^{k} C_k^l B_{k-l} x^l$$

Démonstration. Il résulte de la définition précédente que

$$B_k^{(l)}(x) = \frac{k!}{(k-l)!} B_{k-l}(x)$$

et la conclusion découle donc de la formule de Taylor en x_0. □

Les nombres de Bernoulli peuvent se calculer facilement par récurrence grâce la formule suivante :

Proposition 5.2.4. *On a*

$$\sum_{l=0}^{k} C_{k+1}^{l} B_l = \delta_{k0}$$

pour tout $k \geq 0$. En particulier, les nombres de Bernoulli sont rationnels.

Démonstration. Le résultat est clair si $k = 0$ et on peut donc supposer que $k \geq 1$. Dans ce cas, il résulte de la définition des polynômes de Bernoulli que

$$\int_0^1 B_k(x)\, dx = 0.$$

En tenant compte de la proposition précédente, on voit alors que

$$\sum_{l=0}^{k} C_k^l B_{k-l} \int_0^1 x^l\, dx = 0$$

et par conséquent que

$$\sum_{l=0}^{k} \frac{k!}{(l+1)!(k-l)!} B_{k-l} = 0.$$

Il suffit alors de multiplier cette égalité par $(k+1)$ et d'utiliser $(k-l)$ comme indice de sommation pour conclure. □

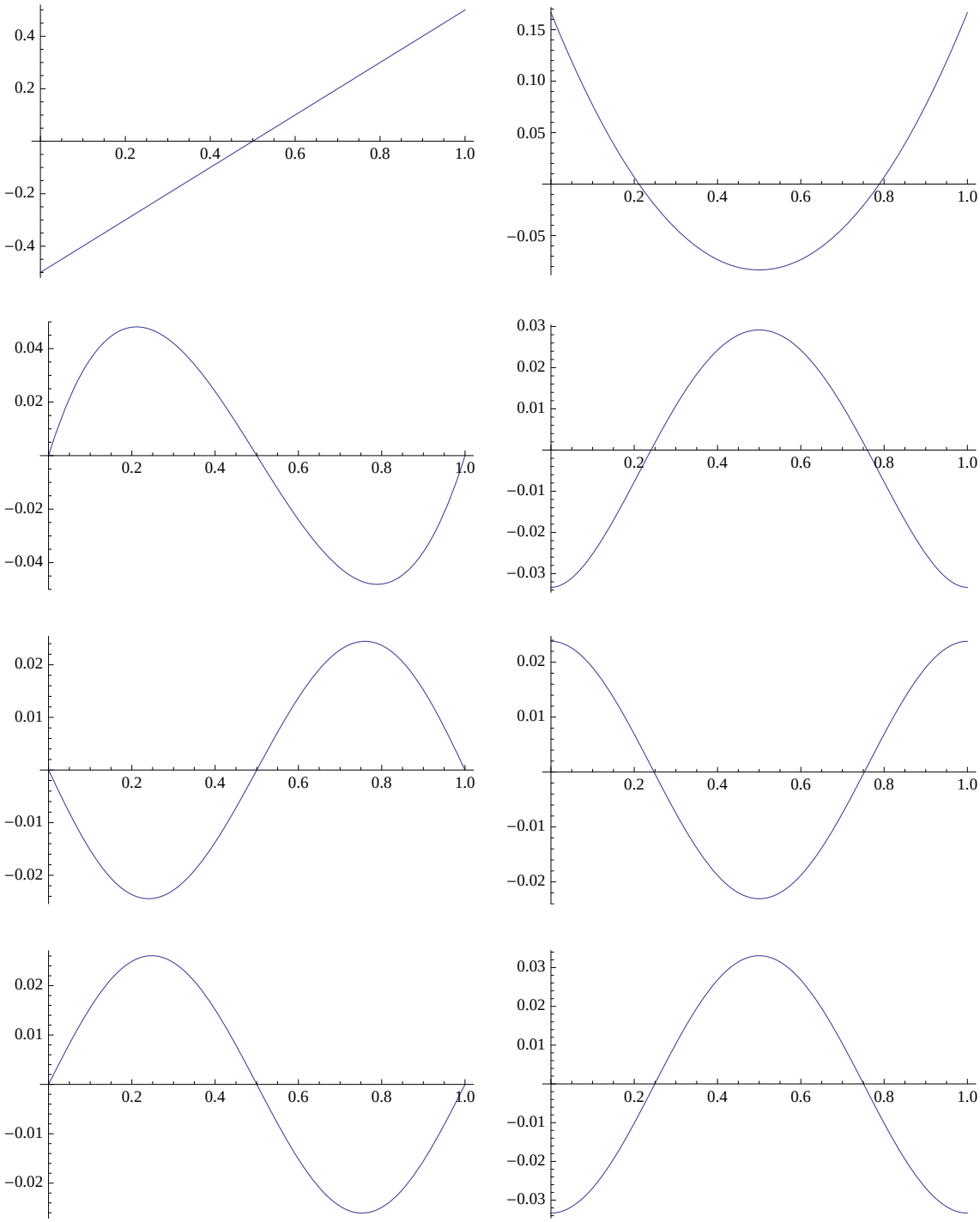
En combinant les deux résultats précédent, il est aisé d'obtenir explicitement les différents polynômes de Bernoulli.

Exemples 5.2.5. Les premiers nombres et polynômes de Bernoulli sont donnés par

la table suivante :

k	B_k	$B_k(x)$
1	$-\frac{1}{2}$	$-\frac{1}{2}+x$
2	$\frac{1}{6}$	$\frac{1}{6}-x+x^2$
3	0	$\frac{1}{2}x-\frac{3}{2}x^2+x^3$
4	$-\frac{1}{30}$	$-\frac{1}{30}+x^2-2x^3+x^4$
5	0	$-\frac{1}{6}x+\frac{5}{3}x^3-\frac{5}{2}x^4+x^5$
6	$\frac{1}{42}$	$\frac{1}{42}-\frac{1}{2}x^2+\frac{5}{2}x^4-3x^5+x^6$
7	0	$\frac{1}{6}x-\frac{7}{6}x^3+\frac{7}{2}x^5-\frac{7}{2}x^6+x^7$
8	$-\frac{1}{30}$	$-\frac{1}{30}+\frac{2}{3}x^2-\frac{7}{3}x^4+\frac{14}{3}x^6-4x^7+x^8$

et leurs graphes sur $[0,1]$ sont donnés ci-dessous :



Ces exemples semblent indiquer que le nombre de Bernoulli d'indice k est nul si k est un naturel impair différent de 1. Cette propriété est confirmée par le résultat suivant :

Proposition 5.2.6. *Pour tout $k \in \mathbb{N}$, on a*

$$B_k(x) = (-1)^k B_k(1-x).$$

En particulier, on a

$$B_k(1/2) = 0$$

si k est un naturel impair et

$$B_k(0) = 0 = B_k(1)$$

et si k est un naturel impair différent de 1.

Démonstration. Posons

$$P_k(x) = (-1)^k B_k(1-x).$$

On a

$$P_k'(x) = (-1)^{k+1} B_k'(1-x) = (-1)^{k-1} k B_{k-1}(1-x) = k P_{k-1}(x)$$

et

$$\int_0^1 P_k(x)\, dx = (-1)^k \int_0^1 B_k(1-x)\, dx = (-1)^k \int_0^1 B_k(x)\, dx = 0.$$

Comme on a aussi $P_0(x) = 1$, il est clair que

$$P_k(x) = B_k(x)$$

pour tout $k \in \mathbb{N}$; d'où la conclusion. □

5.2.2 Développement asymptotique de $R_\rho(f,h)$

Grâce aux propriétés des polynômes de Bernoulli que nous venons détablir, nous pouvons maintenant obtenir aisément un développement limité de $R_\rho(f,h)$ selon les puissances de h :

Proposition 5.2.7. *Si $p \in \mathbb{N}$ et si $f \in C_{p+1}([a,b])$, alors*

$$R_\rho(f,h) = \int_a^b f(x)\, dx + \sum_{k=1}^{p} B_k(\rho) \left(f^{(k-1)}(b) - f^{(k-1)}(a)\right) \frac{h^k}{k!} + \int_a^b \left[B_{p+1}(\rho) - \widehat{B}_{p+1}\left(\rho - \frac{x-a}{h}\right)\right] f^{(p+1)}(x)\, dx\, \frac{h^{p+1}}{(p+1)!}$$

où $\widehat{B}_{p+1}(x)$ désigne la fonction 1-périodique sur $\mathbb{R}$ qui coïncident avec $B_{p+1}(x)$ sur $[0,1[$.

Démonstration. En posant $x = a + th$, on peut se réduire au cas où $a = 0$, $b = n$ et $h = 1$ et en décomposant l'intégrale on peut même se limiter à traiter le cas où $n = 1$. Tout revient alors à montrer que

$$\int_0^1 \frac{\widehat{B}_{p+1}(\rho - t) - B_{p+1}(\rho)}{(p+1)!} f^{(p+1)}(t)\, dt \\ = -f(\rho) + \int_0^1 f(t)\, dt + \sum_{k=1}^{p} \frac{B_k(\rho)}{k!} \left(f^{(k-1)}(1) - f^{(k-1)}(0)\right).$$

Or des intégrations par parties montrent que

$$\int_0^\rho \frac{\widehat{B}_{p+1}(\rho - t) - B_{p+1}(\rho)}{(p+1)!} f^{(p+1)}(t)\, dt \\ = \int_0^\rho \frac{B_{p+1}(\rho - t) - B_{p+1}(\rho)}{(p+1)!} f^{(p+1)}(t)\, dt \\ = \frac{B_{p+1}(0) - B_{p+1}(\rho)}{(p+1)!} f^{(p)}(\rho) + \int_0^\rho \frac{B_p(\rho - t)}{p!} f^{(p)}(t)\, dt$$

et

$$\int_\rho^1 \frac{\widehat{B}_{p+1}(\rho - t) - B_{p+1}(\rho)}{(p+1)!} f^{(p+1)}(t)\, dt \\ = \int_\rho^1 \frac{B_{p+1}(\rho + 1 - t) - B_{p+1}(\rho)}{(p+1)!} f^{(p+1)}(t)\, dt \\ = -\frac{B_{p+1}(1) - B_{p+1}(\rho)}{(p+1)!} f^{(p)}(\rho) + \int_\rho^1 \frac{B_p(\rho + 1 - t)}{p!} f^{(p)}(t)\, dt.$$

Ainsi

$$\int_0^1 \frac{\widehat{B}_{p+1}(\rho - t) - B_{p+1}(\rho)}{(p+1)!} f^{(p+1)}(t)\, dt \\ = \frac{B_{p+1}(0) - B_{p+1}(1)}{(p+1)!} f^{(p)}(\rho) + \int_0^1 \frac{\widehat{B}_p(\rho - t)}{p!} f^{(p)}(t)\, dt.$$

Si $p = 0$, cette formule devient

$$\int_0^1 \widehat{B}_1(\rho - t) - B_1(\rho) f'(t)\, dt = -f(\rho) + \int_0^1 f(t)\, dt$$

comme attendu puisque $B_1(0) = -1/2$, $B_1(1) = 1/2$ et $\widehat{B}_0(\rho - t) = 1$. Si $p \geq 1$, on

a $B_{p+1}(0) = B_{p+1}(1)$ et on a

$$\begin{aligned}\int_0^1 &\frac{\widehat{B}_{p+1}(\rho - t) - B_{p+1}(\rho)}{(p+1)!} f^{(p+1)}(t)\, dt \\ &= \int_0^1 \frac{\widehat{B}_p(\rho - t)}{p!} f^{(p)}(t)\, dt. \\ &= \int_0^1 \frac{\widehat{B}_p(\rho - t) - B_p(\rho)}{p!} f^{(p)}(t)\, dt + \frac{B_p(\rho)}{p!}\left(f^{(p-1)}(1) - f^{(p-1)}(0)\right)\end{aligned}$$

d'où la conclusion par récurrence sur p. □

Corollaire 5.2.8. *Si $p \in \mathbb{N}$ et si $f \in C_{p+1}([a,b])$, alors*

$$R_\rho(f,h) = \int_a^b f(x)\, dx + \sum_{k=1}^{p} B_k(\rho)\left(f^{(k-1)}(b) - f^{(k-1)}(a)\right)\frac{h^k}{k!} + O(h^{p+1})$$

pour $h \to 0^+$ avec

$$|O(h^{p+1})| \leq (b-a)\int_0^1 |B_{p+1}(x) - B_{p+1}(\rho)|\, dx \sup_{\xi\in[a,b]} |f^{(p+1)}(\xi)|\, \frac{h^{p+1}}{(p+1)!}.$$

De plus, si ρ est un extremum global de $B_{p+1}(x)$ sur $[0,1]$, alors il existe $\xi \in [a,b]$ tel que

$$O(h^{p+1}) = (b-a)B_{p+1}(\rho)f^{(p+1)}(\xi)\frac{h^{p+1}}{(p+1)!}.$$

Démonstration. Il est clair que

$$\begin{aligned}&\left|\int_a^b \left[B_{p+1}(\rho) - \widehat{B}_{p+1}\left(\rho - \frac{x-a}{h}\right)\right] f^{(p+1)}(x)\, dx\right| \\ &\qquad \leq \int_a^b \left|B_{p+1}(\rho) - \widehat{B}_{p+1}\left(\rho - \frac{x-a}{h}\right)\right| dx \sup_{\xi\in[a,b]} |f^{(p+1)}(\xi)|.\end{aligned}$$

En utilisant le changement de variable $x = a + th$ et la 1-périodicité de la fonction $\widehat{B}_{p+1}$ on voit que

$$\begin{aligned}\int_a^b \left|B_{p+1}(\rho) - \widehat{B}_{p+1}\left(\rho - \frac{x-a}{h}\right)\right| dx &= h\int_0^n |B_{p+1}(\rho) - \widehat{B}_{p+1}(\rho - t)|\, dt \\ &= nh\int_0^1 |B_{p+1}(\rho) - \widehat{B}_{p+1}(\rho - t)|\, dt \\ &= (b-a)\int_{\rho-1}^{\rho} |B_{p+1}(\rho) - \widehat{B}_{p+1}(t)|\, dt \\ &= (b-a)\int_0^1 |B_{p+1}(\rho) - B_{p+1}(t)|\, dt\end{aligned}$$

et on en tire la première partie de l'énoncé. Supposons à présent que ρ soit un extremum global de $B_{p+1}(x)$ sur $[0,1]$. La fonction

$$\left[B_{p+1}(\rho) - \widehat{B}_{p+1}\left(\rho - \frac{x-a}{h}\right)\right]$$

est alors de signe constant sur $[a,b]$ et le théorème de la moyenne montre qu'il existe un $\xi \in [a,b]$ pour lequel

$$\begin{aligned}&\int_a^b \left[B_{p+1}(\rho) - \widehat{B}_{p+1}\left(\rho - \frac{x-a}{h}\right)\right] f^{(p+1)}(x)\,dx \\ &= \int_a^b \left[B_{p+1}(\rho) - \widehat{B}_{p+1}\left(\rho - \frac{x-a}{h}\right)\right]\,dx f^{(p+1)}(\xi).\end{aligned}$$

En procédant comme ci-dessus, on voit alors que

$$\begin{aligned}\int_a^b \left[B_{p+1}(\rho) - \widehat{B}_{p+1}\left(\rho - \frac{x-a}{h}\right)\right]\,dx &= (b-a)\int_0^1 \left[B_{p+1}(\rho) - B_{p+1}(t)\right]\,dt \\ &= (b-a)B_{p+1}(\rho)\end{aligned}$$

puisque

$$\int_0^1 B_{p+1}(t)\,dt = 0.$$

Il s'ensuit que

$$\int_a^b \left[B_{p+1}(\rho) - \widehat{B}_{p+1}\left(\rho - \frac{x-a}{h}\right)\right] f^{(p+1)}(x)\,dx = (b-a)B_{p+1}(\rho)f^{(p+1)}(\xi)$$

comme attendu. □

Corollaire 5.2.9. *On a*

$$R_\rho(P,h) = \int_a^b P(x)\,dx + \sum_{k=1}^{p} \frac{B_k(\rho)}{k!}\left(P^{(k-1)}(b) - P^{(k-1)}(a)\right)h^k$$

pour tout polynôme P de degré $\leq p$.

Démonstration. Cela découle directement du résultat précédent puisque $P^{(p+1)}(x) = 0$ pour tout polynôme P de degré $\leq p$. □

Remarque 5.2.10. Le corollaire précédente permet de calculer facilement une somme de Bernoulli du type

$$S_p(n) = \sum_{k=0}^{n-1} k^p.$$

En effet, si nous renons $P = x^p$, $h = 1$, $a = 0$, $b = n$ et $\rho = 0$, il vient

$$\sum_{k=0}^{n-1} k^p = \sum_{k=0}^{p} B_k \ \frac{p!}{(p-k+1)!} n^{p-k+1} \ \frac{1}{k!} = \sum_{k=0}^{p} C_p^k B_k \frac{n^{p+1-k}}{p+1-k}.$$

Ainsi

$$S_p(n) = \int_0^n B_p(x)\ dx = \frac{B_{p+1}(n) - B_{p+1}(0)}{p+1}.$$

Il en résulte en particulier que $S_p(n)$ est un polynôme de degré $p+1$ en n dont le terme constant est nul et dont le terme linéaire a B_p pour coefficient. Bien sûr, on a aussi

$$B_p(x) = S_p'(x).$$

Pour pouvoir utiliser la forme simplifiée du reste du développement limité de $R_\rho(f,h)$ selon les puissances de h, il est nécessaire de connaître les extrema des $B_p(x)$ sur $[0,1]$. C'est pourquoi nous allons nous intéresser maintenant aux comportement des polynômes de Bernoulli sur cet intervalle.

5.2.3 Etude des $B_k(x)$ sur $[0,1]$

Proposition 5.2.11. *Le tableau de signes du polynôme $B_1(x)$ sur $[0,1]$ est le suivant :*

	0		1/2		1
$B_1(\rho)$	−	−	0	+	+

Démonstration. C'est évident car

$$B_1(x) = x - \frac{1}{2}.$$

□

Proposition 5.2.12. *Si $l \geq 1$, le tableau de signes de $(-1)^l B_{2l+1}(x)$ sur $[0,1]$ est le suivant :*

	0		1/2		1
$(-1)^l B_{2l+1}(x)$	0	−	0	+	0

Démonstration. La proposition 5.2.6 montre que

$$B_{2l+1}(1) = B_{2l+1}(1/2) = B_{2l+1}(0) = 0$$

et il suffit donc de montrer que

$$(-1)^l B_{2l+1}(x) \begin{cases} < 0 & \text{si } x \in]0, 1/2[\\ > 0 & \text{si } x \in]1/2, 1[\end{cases}$$

pour $l \geq 0$. Procédons par récurrence. D'une part, vu la forme du tableau de signe de $B_1(x)$, le cas $l = 0$ est clair. D'autre part, si $l > 0$ et si

$$(-1)^{l-1} B_{2l-1}(x) \begin{cases} < 0 & \text{si } x \in]0, 1/2[\\ > 0 & \text{si } x \in]1/2, 1[\end{cases}$$

alors la relation

$$(-1)^l B''_{2l+1}(x) = -(2l+1)(2l)(-1)^{l-1} B_{2l-1}(x)$$

montre que la fonction $(-1)^l B_{2l+1}(x)$ est strictement convexe sur $]0, 1/2[$ et strictement concave sur $]1/2, 1[$. Comme cette fonction s'annule en 0, $1/2$ et 1, on en tire que dans ce cas on a aussi

$$(-1)^l B_{2l+1}(x) \begin{cases} < 0 & \text{si } x \in]0, 1/2[\\ > 0 & \text{si } x \in]1/2, 1[\end{cases}$$

ce qui permet de conclure. □

Proposition 5.2.13. *Si $l \geq 1$, alors le polynôme $B_{2l}(x)$ a exactement deux zéros sur $[0, 1]$. Le plus petit d'entre eux ρ_l^* se trouve dans l'intervalle $]0, 1/2[$ et le plus grand est égal à $1 - \rho_l^*$. De plus, le tableau de signes de $(-1)^{l-1} B_{2l}(x)$ est le suivant :*

	0		ρ_l^*		$1/2$		$1-\rho_l^*$		1
$(-1)^{l-1} B_{2l}(x)$	$+$	$+$	0	$-$	$-$	$-$	0	$+$	$+$

Démonstration. Puisque

$$B'_{2l}(x) = 2l B_{2l-1}(x),$$

la proposition 5.2.12 montre que le tableau de variation de $(-1)^{l-1} B_{2l}(x)$ est le suivant :

	0		$1/2$		1
$(-1)^{l-1} B_{2l}(\rho)$	max	$\searrow$	min	$\nearrow$	max

la croissance et la décroissance étant stricte.

Comme

$$B_{2l}(0) = B_{2l} = B_{2l}(1),$$

on en tire que

$$(-1)^{l-1} B_{2l}\left(\frac{1}{2}\right) < (-1)^{l-1} B_{2l}(x) < (-1)^{l-1} B_{2l}$$

si $x \in]0, 1/2[\cup]1/2, 1[$. Par intégration, on en déduit que

$$(-1)^{l-1} B_{2l}(1/2) < 0 < (-1)^{l-1} B_{2l}$$

et par conséquent que

$$(-1)^{l-1} B_{2l}$$

est un rationnel strictement positif.

Il en résulte aussi que $(-1)^{l-1} B_{2l}(x)$ ne s'annule pas en 0, 1/2, 1 et s'annule une seule fois sur $]0, 1/2[$ et sur $]1/2, 1[$. Si ρ_l^* désigne le zéro de $B_{2l}(x)$ sur $]0, 1/2[$, alors la proposition 5.2.6 montre que $1 - \rho_l^*$ est le zéro de $B_{2l}(x)$ sur $]1/2, 1[$. On est donc conduit au tableau de signe figurant dans l'énoncé. □

Corollaire 5.2.14. *Si $l \geq 1$, le tableau de variation de $(-1)^l B_{2l+1}(x)$ sur $[0, 1]$ est le suivant :*

	0		ρ_l^*		$1/2$		$1 - \rho_l^*$		1
$(-1)^l B_{2l+1}(x)$	0	$\searrow$	*min*	$\nearrow$	0	$\nearrow$	*max*	$\searrow$	0

Démonstration. Cela résulte directement de la relation

$$(-1)^l B'_{2l+1}(x) = -(2l+1)(-1)^{l-1} B_{2l}(x)$$

et de la proposition précédente. □

Remarque 5.2.15. Si $l \geq 1$, les résultats précédents montrent en particulier que les nombres de Bernoulli de la forme B_{2l+1} sont nuls et que ceux de la forme B_{2l} sont non nuls et du même signe que $(-1)^{l-1}$. C'est pourquoi on définissait classiquement le l-ème nombre de Bernoulli comme étant le rationnel strictement positif

$$(-1)^{l-1} B_{2l}.$$

Suivant l'usage moderne (voir *e.g.* [**?**]), nous noterons ce nombre B_l^*.

5.2.4 Développement asymptotique de $R_0(f, h)$, $R_1(f, h)$ et $R_{1/2}(f, h)$

Les résultats ci-dessus montrent que le développement asymptotique à l'ordre p de $R_\rho(f, h)$ est particulièrement simple lorsque $\rho \in \{0, 1/2, 1\}$ et que p est impair. En effet :

Proposition 5.2.16. *Soit p un naturel impair et soit f une fonction de classe C_{p+1} sur $[a,b]$. Alors*

$$R_0(f,h) = \int_a^b f(x)\,dx - \frac{1}{2}(f(b)-f(a))h + \sum_{l=1}^{\lfloor p/2\rfloor} \frac{B_{2l}}{(2l)!}\left(f^{(2l-1)}(b) - f^{(2l-1)}(a)\right)h^{2l} + O(h^{p+1})$$

avec

$$\begin{aligned} O(h^{p+1}) &= \int_a^b \left[B_{p+1} - \widehat{B}_{p+1}\left(\frac{x-a}{h}\right)\right] f^{(p+1)}(x)\,dx \frac{h^{p+1}}{(p+1)!} \\ &= (b-a)B_{p+1} f^{(p+1)}(\xi)\frac{h^{p+1}}{(p+1)!} \end{aligned}$$

pour un $\xi \in [a,b]$.

Démonstration. Compte tenu de la Proposition 5.2.7 et du Corollaire 5.2.8, la première formule découle directement de ce que

$$B_k = B_k(0)$$

est nul si k est impair et différent de 1 et que

$$B_1 = -\frac{1}{2}.$$

De ces mêmes résultats, il découle aussi que

$$O(h^{p+1}) = \int_a^b \left[B_{p+1} - \widehat{B}_{p+1}\left(-\frac{x-a}{h}\right)\right] f^{(p+1)}(x)\,dx.$$

Comme la Proposition 5.2.6 montre que

$$\widehat{B}_{p+1}\left(-\frac{x-a}{h}\right) = \widehat{B}_{p+1}\left(1-\frac{x-a}{h}\right) = (-1)^{p+1}\widehat{B}_{p+1}\left(\frac{x-a}{h}\right),$$

la seconde formule est également claire. Pour conclure, il suffit alors d'utiliser à nouveau le Corollaire 5.2.8 en se rappelant que $x = 0$ est un extremum global de $B_{p+1}(x)$ sur $[0,1]$. □

Proposition 5.2.17. *Soit p un naturel impair et soit f une fonction de classe C_{p+1} sur $[a,b]$. Alors*

$$R_1(f,h) = \int_a^b f(x)\,dx + \frac{1}{2}(f(b)-f(a))h + \sum_{l=1}^{\lfloor p/2\rfloor} \frac{B_{2l}}{(2l)!}\left(f^{(2l-1)}(b) - f^{(2l-1)}(a)\right)h^{2l} + O(h^{p+1})$$

avec

$$O(h^{p+1}) = \int_a^b \left[B_{p+1} - \widehat{B}_{p+1}\left(\frac{x-a}{h}\right)\right] f^{(p+1)}(x)\, dx \frac{h^{p+1}}{(p+1)!}$$
$$= (b-a) B_{p+1} f^{(p+1)}(\xi) \frac{h^{p+1}}{(p+1)!}$$

pour un $\xi \in [a,b]$.

Démonstration. Il suffit de procéder comme dans la preuve de la proposition précédente ou de remarquer que

$$R_1(f,h) = R_0(f,h) + f(b)h - f(a)h.$$

□

Proposition 5.2.18. *Soit* p *un naturel impair et soit* f *une fonction de classe* C_{p+1} *sur* $[a,b]$. *Alors*

$$R_{1/2}(f,h) = \int_a^b f(x)\, dx$$
$$+ \sum_{l=1}^{\lfloor p/2 \rfloor} \frac{B_{2l}(1/2)}{(2l)!} \left(f^{(2l-1)}(b) - f^{(2l-1)}(a)\right) h^{2l} + O(h^{p+1})$$

avec

$$O(h^{p+1}) = \int_a^b \left[B_{p+1}(1/2) - \widehat{B}_{p+1}\left(\frac{1}{2} - \frac{x-a}{h}\right)\right] f^{(p+1)}(x)\, dx \frac{h^{p+1}}{(p+1)!}$$
$$= (b-a) B_{p+1}(1/2) f^{(p+1)}(\xi) \frac{h^{p+1}}{(p+1)!}$$

pour un $\xi \in [a,b]$. *De plus, on a*

$$B_k(1/2) = \left(\frac{1}{2^{k-1}} - 1\right) B_k$$

pour $k \geq 0$.

Démonstration. La première partie découle directement de la Proposition 5.2.7 et du Corollaire 5.2.8, compte tenu du fait que

$$B_k(1/2) = 0$$

si k est un naturel impair. De plus, il est clair que

$$\begin{aligned} R_{1/2}(f,h) + R_0(f,h) &= \sum_{k=0}^{n-1} f\left(a + kl + \frac{h}{2}\right) h + \sum_{k=0}^{n-1} f(a+kh)h \\ &= 2\left(\sum_{k=0}^{n-1} f\left(a + (2k+1)\frac{h}{2}\right)\frac{h}{2} + \sum_{k=0}^{n-1} f\left(a + 2k\frac{h}{2}\right)\frac{h}{2}\right) \\ &= 2R_0(f, \frac{h}{2}). \end{aligned}$$

Ainsi, pour toute fonction f de classe C_{p+1} sur $[a,b]$, on a

$$\begin{aligned} R_{1/2}(f,h) &= 2R_0(f,h/2) - R_0(f,h) \\ &= \sum_{k=2}^{p} \frac{B_k}{k!}\left(\frac{1}{2^{k-1}} - 1\right)\left(f^{(k-1)}(b) - f^{(k-1)}(a)\right) h^k + O(h^{p+1}). \end{aligned}$$

La conclusion s'obtient alors en comparant ce développement avec celui obtenu dans la première partie. □

5.2.5 Formule sommatoire d'Euler-Maclaurin

Un sous-produit intéressant de la Proposition 5.2.16 est la célèbre formule sommatoire d'Euler-Maclaurin :

Proposition 5.2.19. *Soient $M < N$ des entiers relatifs et soit q un naturel non nul. Supposons que f soit une fonction de classe C_{2q} sur $[M,N]$. Alors*

$$\begin{aligned} \sum_{m=M}^{N} f(m) = \int_M^N f(x)\,dx + \frac{f(M)+f(N)}{2} \\ + \sum_{l=1}^{q-1} \frac{B_{2l}}{(2l)!}\left(f^{(2l-1)}(N) - f^{(2l-1)}(M)\right) \\ + \int_M^N \frac{B_{2q} - \widehat{B}_{2q}(x)}{(2q)!} f^{(2q)}(x)\,dx. \end{aligned}$$

De plus

$$\int_M^N \frac{B_{2q} - \widehat{B}_{2q}(x)}{(2q)!} f^{(2q)}(x)\,dx = (N-M)\frac{B_{2q}}{(2q)!} f^{(2q)}(\xi)$$

pour un $\xi \in [M,N]$.

Démonstration. Il suffit de prendre $a = M$, $b = N$, $h = 1$ et $p = 2q - 1$ dans la Proposition 5.2.16 et de remarquer que, dans ce cas, on a $n = N - M$ et

$$R_0(f, h) = \sum_{k=0}^{n-1} f(M + k) = \sum_{m=M}^{N-1} f(m).$$

□

La formule sommatoire d'Euler-Maclaurin a de nombreuses applications tant en analyse qu'en analyse numérique. Elle montre par exemple que :

Exemple 5.2.20. Pour tout $q \in \mathbb{N}_0$, on a

$$\begin{aligned}\sum_{m=1}^{N} \frac{1}{m} &= \int_1^N \frac{1}{x}\, dx + \frac{1 + \frac{1}{N}}{2} \\ &\quad + \sum_{l=1}^{q-1} \frac{B_{2l}}{(2l)!}(-1)^{2l-1}(2l-1)! \left(\frac{1}{N^{2l}} - 1\right) \\ &\quad + \int_1^N \frac{B_{2q} - \widehat{B}_{2q}(x)}{(2q)!}\, \frac{(-1)^{2q}(2q)!}{x^{2q+1}}\, dx.\end{aligned}$$

Ainsi,

$$\sum_{m=1}^{N} \frac{1}{m} = \ln N + \gamma + \frac{1}{2N} - \sum_{l=1}^{q-1} \frac{B_{2l}}{2l}\, \frac{1}{N^{2l}} - \int_N^\infty \frac{B_{2q} - \widehat{B}_{2q}(x)}{x^{2q+1}}\, dx$$

où

$$\gamma = \frac{1}{2} + \sum_{l=1}^{q-1} \frac{B_{2l}}{2l} + \int_1^\infty \frac{B_{2q} - \widehat{B}_{2q}(x)}{x^{2q+1}}\, dx$$

est la constante d'Euler. Il s'ensuit que

$$\sum_{m=1}^{N} \frac{1}{m} = \ln N + \gamma + \frac{1}{2N} - \sum_{l=1}^{q-1} \frac{B_{2l}}{2l}\, \frac{1}{N^{2l}} + O\left(\frac{1}{N^{2q}}\right).$$

Pour trouver la valeur de γ, on peut utiliser les approximations

$$s_q(N) = \sum_{m=1}^{N} \frac{1}{m} - \ln N - \frac{1}{2N} + \sum_{l=1}^{q-1} \frac{B_{2l}}{2l}\, \frac{1}{N^{2l}}$$

pour différentes valeurs de q. Comme l'erreur associée à cette approximation change de signe avec q, on peut facilement l'estimer. On trouvera ci-dessous la table des

valeurs de $s_q(N)$ pour $q = 1, 2, 3, 4$ et $N = 1, \ldots, 10$.

N	$s_1(N)$	$s_2(N)$	$s_3(N)$	$s_4(N)$
1	0.5000000000	0.5833333333	0.5750000000	0.5789682540
2	0.5568528194	0.5776861528	0.5771653194	0.5772273234
3	0.5680543780	0.5773136373	0.5772107566	0.5772162000
4	0.5720389722	0.5772473055	0.5772147535	0.5772157223
5	0.5738954209	0.5772287542	0.5772154209	0.5772156749
6	0.5749071974	0.5772220123	0.5772155822	0.5772156673
7	0.5755184224	0.5772191026	0.5772156319	0.5772156656
8	0.5759156012	0.5772176845	0.5772156500	0.5772156651
9	0.5761881211	0.5772169277	0.5772156575	0.5772156650
10	0.5763831610	0.5772164943	0.5772156610	0.5772156649

La valeur de γ à 8 décimales est donc 0.57721566.

5.3 Méthode des trapèzes

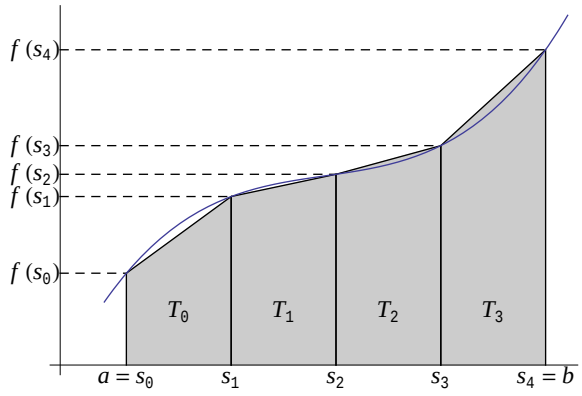
Dans la méthode des trapèzes, on approche l'intégrale

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

par la somme

$$T(f, h) = \sum_{k=0}^{n-1} \frac{f(s_k) + f(s_{k+1})}{2} h$$

où $s_k = a + kh$ avec $h = (b-a)/n$. En d'autres termes, on approche I par la somme des aires (algébriques) des trapèzes T_k de sommets $(s_k, 0)$, $(s_{k+1}, 0)$, $(s_{k+1}, f(s_{k+1}))$, $(s_k, f(s_k))$.



Comme

$$T(f,h) = \sum_{k=0}^{n-1} \frac{f(s_k)}{2} h + \sum_{k=1}^{n} \frac{f(s_k)}{2} h,$$

il est clair que

$$T(f,h) = R_0(f,h) + \frac{1}{2}(f(b) - f(a))h = R_1(f,h) - \frac{1}{2}(f(b) - f(a))h.$$

Il s'ensuit que $T(f,h)$ coïncide avec la méthode d'ordre 2 obtenue en corrigeant la méthode des rectangles à points initiaux (resp. finaux) par le terme d'ordre 1 de son développement asymptotique selon les puissances de h. Vu ce qui précède, on a donc le résultat suivant :

Proposition 5.3.1. *Soit p un naturel impair et soit f une fonction de classe C_{p+1} sur $[a,b]$. Alors*

$$\begin{aligned} T(f,h) = &\int_a^b f(x)\,dx \\ &+ \sum_{l=1}^{\lfloor p/2 \rfloor} \frac{B_{2l}}{(2l)!} \left(f^{(2l-1)}(b) - f^{(2l-1)}(a)\right) h^{2l} + O(h^{p+1}) \end{aligned}$$

avec

$$\begin{aligned} O(h^{p+1}) &= \int_a^b \left[B_{p+1} - \widehat{B}_{p+1}\left(\frac{x-a}{h}\right)\right] f^{(p+1)}(x)\,dx \frac{h^{p+1}}{(p+1)!} \\ &= (b-a) B_{p+1} f^{(p+1)}(\xi) \frac{h^{p+1}}{(p+1)!} \end{aligned}$$

pour un $\xi \in [a,b]$.

Corollaire 5.3.2. *Si f est de classe C_2 sur $[a,b]$, on a*

$$T(f,h) - \int_a^b f(x)\,dx = \frac{(b-a)}{12} f''(\xi) h^2$$

pour un $\xi \in [a,b]$.

Démonstration. Cela découle directement de ce qui précède compte tenu du fait que $B_2 = 1/6$. □

Remarque 5.3.3. La formule d'erreur donnée ci-dessus montre que la méthode des trapèzes est en général moins précise que celle des rectangles de rapport 1/2. Elle a cependant un gros avantage sur celle-ci. En effet, on a

$$T(f,h) = h\left(\frac{f(a)}{2} + f(a+h) + \cdots + f(a+(n-1)h) + \frac{f(b)}{2}\right)$$

et

$$T\left(f, \frac{h}{2}\right) = \frac{h}{2}\left(\frac{f(a)}{2} + f\left(a + \frac{h}{2}\right) + f(a+h) + \cdots \right.$$
$$\left. + f(a + (n-1)h) + f\left(a + (n-1)h + \frac{h}{2}\right) + \frac{f(b)}{2}\right).$$

Donc

$$T\left(f, \frac{h}{2}\right) = \frac{1}{2}T(f,h) + \frac{h}{2}\sum_{k=0}^{n-1} f\left(a + kh + \frac{h}{2}\right)$$

ce qui permet de calculer $T(f, h/2)$ à partir de $T(f,h)$ grâce à une multiplication, deux divisions, n additions et n évaluations de f. Dans le cas du calcul de $R_{1/2}(f, h/2)$, on ne peut exploiter les calculs effectués pour obtenir $R_{1/2}(f,h)$ et on a donc besoin d'une multiplication, de $2n$ additions et $2n$ évaluations de f. Cette particularité de la méthode des trapèzes fait, comme nous allons le voir, qu'elle se prête très bien à l'extrapolation de Richardson.

5.4 Méthode de Romberg

5.4.1 Extrapolation de Richardson

Proposition 5.4.1. *Soit f défini sur $]0, H[$ et tel que*

$$f(h) = a_0 + a_1 h^{p_1} + O(h^{p_2})$$

pour $h \to 0^+$ avec a_0, $a_1 \in \mathbb{R}$, $p_2 > p_1 > 0$. Fixons $\theta > 1$ et posons

$$f_1(h) = f(h) + \frac{f(h) - f(\theta h)}{\theta^{p_1} - 1}$$

pour tout $h \in]0, H/\theta[$. Alors,

$$f_1(h) = a_0 + O(h^{p_2})$$

pour $h \to 0^+$.

Démonstration. Par hypothèse,

$$f(h) = a_0 + a_1 h^{p_1} + O(h^{p_2})$$

et

$$f(\theta h) = a_0 + a_1 \theta^{p_1} h^{p_1} + O(h^{p_2})$$

si $h \to 0^+$. Il s'ensuit que

$$f(h) - f(\theta h) = a_1(1 - \theta^{p_1})h^{p_1} + O(h^{p_2})$$

si $h \to 0^+$. On en tire que

$$\frac{f(h) - f(\theta h)}{\theta^{p_1} - 1} = -a_1 h^{p_1} + O(h^{p_2})$$

puis que

$$f_1(h) = a_0 + O(h^{p_2})$$

si $h \to 0^+$. □

Corollaire 5.4.2. *Soit f défini sur $]0, H[$ et tel que*

$$f(h) = a_0 + a_1 h^{p_1} + a_2 h^{p_2} + \cdots + a_q h^{p_q} + O(h^{p_{q+1}})$$

pour $h \to 0^+$ avec $a_0, \ldots, a_q \in \mathbb{R}$ et $0 < p_1 < p_2 < \cdots < p_q < p_{q+1}$. Fixons $\theta > 1$ et définissons $f_0, f_1, \ldots, f_q$ par récurrence en posant $f_0 = f$ et

$$f_l(h) = f_{l-1}(h) + \frac{f_{l-1}(h) - f_{l-1}(\theta h)}{\theta^{p_l} - 1} \qquad (h \in \left]0, H\theta^{-l}\right[)$$

pour $l = 1, \ldots, q$. Alors,

$$f_l(h) = a_0 + O(h^{p_{l+1}})$$

pour $l = 1, \ldots, q$.

Corollaire 5.4.3. *Dans les conditions du corollaire précédent, fixons $h_0 > 0$ et posons*

$$F_{l,m} = f_l(h_0\theta^{-m})$$

pour $l \in \{0, \ldots, q\}$. Alors,

$$F_{l,m} = F_{l-1,m} + \frac{F_{l-1,m} - F_{l-1,m-1}}{\theta^{p_l} - 1}$$

pour $l = 1, \ldots, q$ et

$$F_{l,m} = a_0 + O(\theta^{-mp_{l+1}})$$

si $m \to \infty$.

Remarque 5.4.4. Pour utiliser le résultat précédent en pratique, on forme le tableau

$$\begin{array}{ccccc} F_{0,0} & & & & \\ F_{0,1} & F_{1,1} & & & \\ F_{0,2} & F_{1,2} & F_{2,2} & & \\ \vdots & \vdots & \vdots & \ddots & \\ F_{0,q} & F_{1,q} & F_{2,n} & \cdots & F_{q,q} \\ \vdots & \vdots & \vdots & & \vdots \\ F_{0,m} & F_{1,m} & F_{2,m} & \cdots & F_{q,m} \end{array}$$

La colonne l de ce tableau fournit une approximation de a_0 avec une erreur en $O(\theta^{-mp_{l+1}})$. En fait, on ne calcule pas toutes les colonnes mais seulement quelques-unes d'entre elles et on admet que $F_{l,m}$ approche a_0 avec une erreur inférieure à ε si $|F_{l-1,m} - F_{l-1,m-1}| < \varepsilon$. Ce critère d'arrêt heuristique n'est malheureusement pas toujours justifié.

5.4.2 Application à la méthode des trapèzes

Soit q un naturel et soit $f \in C_{2q+2}([a,b])$. Alors il résulte de la Proposition 5.3.1 que

$$T(f,h) = \int_a^b f(x)\,dx + \sum_{l=1}^{q} \frac{B_{2l}}{(2l)!}\left(f^{(2l-1)}(b) - f^{(2l-1)}(a)\right) h^{2l} + O(h^{2q+2}).$$

On peut donc appliquer l'extrapolation de Richardson à la fonction $h \mapsto T(f,h)$. Si le coefficient θ est choisi égal à 2, cela donne naissance à une méthode très efficace pour approcher

$$\int_a^b f(x)\,dx$$

puisque $T(f,h/2)$ peut se calculer rapidement à partir de $T(f,h)$. C'est cette méthode qui est connue sous le nom de méthode de Romberg.

Exemple 5.4.5. Pour

$$I = \int_0^1 e^{-x^2}\,dx,$$

la méthode de Romberg donne le tableau suivant :

m	$F_{0,m}$	$F_{1,m}$	$F_{2,m}$	$F_{3,m}$	$F_{4,m}$	$F_{5,m}$
0	0.6839397206					
1	0.7313702518	0.7471804289				
2	0.7429840978	0.7468553798	0.7468337098			
3	0.7458656148	0.7468261205	0.7468241699	0.7468240185		
4	0.7465845968	0.7468242574	0.7468241332	0.7468241326	0.7468241331	
5	0.7467642547	0.7468241406	0.7468241328	0.7468241328	0.7468241328	0.7468241328

alors que la valeur de I à 12 décimales est 0.746824132812.

5.5 Méthodes de Newton-Cotes

Soit f une fonction continue sur $[a, b] \subset \mathbb{R}$. Pour approcher

$$\int_a^b f(x)\, dx$$

on peut aussi utiliser

$$\int_a^b P(x)\, dx$$

où $P(x)$ est le polynôme interpolant f en $a = x_0 < x_1 \cdots < x_n = b$. Les méthodes de Newton-Cotes sont basées sur ce procédé dans le cas où les points $x_0, \ldots, x_n$ sont équidistants. Les cas les plus connus sont ceux pour lesquels $n = 1$ (méthode des trapèzes localisée) et $n = 2$ (méthode de Simpson localisée). En général, on sait par la formule de Lagrange que

$$P(x) = \sum_{j=0}^{n} f(x_j) L_j(x).$$

Il s'ensuit que

$$\int_a^b P(x)\, dx = \sum_{j=0}^{n} f(x_j) \int_a^b L_j(x)\, dx.$$

Puisque

$$L_j(x) = \frac{(x - x_0) \cdots \widehat{(x - x_j)} \cdots (x - x_n)}{(x_j - x_0) \cdots \widehat{(x_j - x_j)} \cdots (x_j - x_n)}$$

et que $x_j = a + jh$, avec $h = (b - a)/n$, on a $L_j(x) = \varphi_j\left(\dfrac{x - a}{h}\right)$, avec

$$\varphi_j(t) = \frac{(t - 0) \cdots \widehat{(t - j)} \cdots (t - n)}{(j - 0) \cdots \widehat{(j - j)} \cdots (j - n)}.$$

Il s'ensuit que

$$\int_a^b L_j(x)\,dx = h\int_0^n \varphi_j(t)\,dt.$$

Si nous posons

$$c_j = \frac{1}{n}\int_0^n \varphi_j(t)\,dt,$$

c_j ne dépend que de n et il vient

$$\int_a^b P(x)\,dx = (b-a)\sum_{j=0}^{n} c_j f(x_j).$$

La méthode de Newton-Cotes de degré n consiste donc à approcher

$$\int_a^b f(x)\,dx$$

par

$$(b-a)\sum_{j=0}^{n} c_j f(x_j).$$

Les nombres c_j sont les *coefficients de Cotes* de la méthode. Désignons par s le dénominateur commun de c_j et par σ_j l'entier sc_j. On a alors la table suivante :

n	s	sc_0	sc_1	sc_2	sc_3	sc_4	sc_5	sc_6	sc_7	sc_8	sc_9	sc_{10}
1	2	1	1									
2	6	1	4	1								
3	8	1	3	3	1							
4	90	7	32	12	32	7						
5	288	19	75	50	50	75	19					
6	840	41	216	27	272	27	216	41				
7	17280	751	3577	1323	2989	2989	1323	3577	751			
8	28350	989	5888	−928	10496	−4540	10496	−928	5888	989		
9	89600	2857	15741	1080	19344	5778	5778	19344	1080	15741	2857	
10	598752	16067	106300	−48525	272400	−260550	427368	−260550	272400	−48525	106300	16067

Vérifions par exemple les cas $n = 1$ et $n = 2$.
Cas n=1. On a

$$\varphi_0(t) = \frac{t-1}{0-1} = 1-t, \qquad \varphi_1(t) = \frac{t-0}{1-0} = t.$$

Ainsi,

$$c_0 = \int_0^1 \varphi_0(t)\,dt = t - \frac{t^2}{2}\Big|_0^1 = \frac{1}{2};$$
$$c_1 = \int_0^1 \varphi_1(t)\,dt = \frac{t^2}{2}\Big|_0^1 = \frac{1}{2}.$$

Cas $n = 2$. On a

$$\varphi_0(t) = \frac{(t-1)(t-2)}{(0-1)(0-2)} = \frac{1}{2}(t^2 - 3t + 2);$$
$$\varphi_1(t) = \frac{(t-0)(t-2)}{(1-0)(1-2)} = -t^2 + 2t;$$
$$\varphi_2(t) = \frac{(t-0)(t-1)}{(2-0)(2-1)} = \frac{1}{2}(t^2 - t).$$

Donc,

$$2c_0 = \frac{1}{2}\left(\frac{t^3}{3} - 3\frac{t^2}{2} + 2t\right)\Big|_0^2 = \frac{1}{2}\left(\frac{8}{3} - 6 + 4\right) = \frac{1}{3}$$
$$2c_1 = \frac{-t^3}{3} + t^2\Big|_0^2 = -\frac{8}{3} + 4 = \frac{4}{3}$$
$$2c_2 = \frac{1}{2}\left(\frac{t^3}{3} - \frac{t^2}{2}\right)\Big|_0^2 = \frac{1}{2}\left(\frac{8}{3} - \frac{4}{2}\right) = \frac{1}{3}.$$

Proposition 5.5.1. *La méthode de Newton-Cotes de degré n est exacte pour tout polynôme de degré $\leq n$. Si n est pair, elle reste exacte pour tout polynôme de degré $\leq n + 1$.*

Démonstration. Si f est un polynôme de degré $\leq n$, alors f est égal au polynôme P qui l'interpole aux points $x_0, \ldots, x_n$. On a donc

$$\int_a^b f(x)\,dx = \int_a^b P(x)\,dx = (b-a)\sum_{j=0}^{n} c_j f(x_j)$$

et la méthode de Newton-Cotes de degré n est exacte pour f.

Supposons que $n = 2k$ ($k \in \mathbb{N}$). Pour montrer que la méthode de Newton-Cotes de degré n est exacte pour tout polynôme de degré $n + 1$, il suffit, d'après ce qui précède, de montrer qu'elle est exacte pour $f(x) = x^{2k+1}$ et $[a, b] = [-1, 1]$. Dans ce cas,

$$x_j = -1 + \frac{j}{k} \qquad (j = 0, \ldots, 2k)$$

et les abscisses d'interpolation sont les rationnels

$$-\frac{k}{k}, -\frac{k-1}{k}, \ldots, -\frac{1}{k}, 0, \frac{1}{k}, \ldots, \frac{k}{k}.$$

Comme le polynôme

$$P(x) = x^{2k+1} - x\prod_{j=1}^{k}\left(x^2 - \frac{j^2}{k^2}\right)$$

est de degré $\leq 2k$ et prend les mêmes valeurs que x^{2k+1} en les points $x_0, \ldots, x_n$, on a

$$\int_a^b P(x)\,dx = (b-a)\sum_{j=0}^{n} c_j f(x_j).$$

Or,

$$\int_{-1}^{1} x^{2k+1}\,dx - \int_{-1}^{1} P(x)\,dx = \int_{-1}^{1} x\prod_{j=1}^{k}\left(x^2 - \frac{j^2}{k^2}\right)\,dx$$

et cette dernière intégrale est nulle puisque l'intégrand est impair. La méthode de Newton-Cotes de degré n est donc exacte pour f. □

Remarque 5.5.2. Si une formule d'intégration approchée du type

$$\int_a^b f(x)\,dx \approx \sum_{j=0}^{J} \mu_j f(x_j)$$

avec $J = n$, $x_j = a + jh$ et $h = (b-a)/n$ est exacte pour les polynômes de degré $\leq n$, alors elle coïncide avec la méthode de Newton-Cotes de degré n et on a

$$\mu_j = (b-a)c_j.$$

On le voit directement en remplaçant f par le polynôme qui l'interpole en $x_0, \ldots, x_n$.

Proposition 5.5.3 (Peano)**.** *Soit $n \geq 0$ et soient $x_0, \ldots, x_J$ des points de $[a,b]$. Supposons que la formule d'intégration approchée*

$$\int_a^b f(x)\,dx \approx \sum_{j=0}^{J} \mu_j f(x_j)$$

soit exacte pour les polynômes de degré $\leq n$. Alors, sur $C_{n+1}([a,b])$ son reste

$$R(f) = \int_a^b f(x)\,dx - \sum_{j=0}^{J} \mu_j f(x_j)$$

est une fonctionnelle linéaire et on a

$$R(f) = \int_a^b K(t) f^{(n+1)}(t)\,dt$$

où

$$K(t) = \frac{1}{n!} R\left[(x-t)_+^n\right] \qquad \text{et} \qquad (x-t)_+^n = \begin{cases} (x-t)^n & \text{si } x \geq t \\ 0 & \text{si } x < t \end{cases}$$

Démonstration. La linéarité de R sur $C_{n+1}([a,b])$ est immédiate. Par la formule de Taylor avec reste intégral, on a

$$f(x) = f(a) + f'(a)(x-a) + \cdots + f^{(n)}(a)\frac{(x-a)^n}{n!} + \frac{1}{n!}\int_a^x f^{(n+1)}(t)(x-t)^n\,dt.$$

Il s'ensuit que

$$R(f) = R\left(\frac{1}{n!}\int_a^x f^{(n+1)}(t)(x-t)^n\,dt\right) = R\left(\frac{1}{n!}\int_a^b f^{(n+1)}(t)(x-t)_+^n\,dt\right)$$

puisque $R(P) = 0$ si P est un polynôme de degré $\leq n$. Comme

$$f^{(n+1)}(t)(x-t)_+^n$$

est intégrable en (x,t) sur $[a,b]\times[a,b]$, on a

$$R\left(\frac{1}{n!}\int_a^b f^{(n+1)}(t)(x-t)_+^n\,dt\right) = \frac{1}{n!}\int_a^b f^{(n+1)}(t)R\left((x-t)_+^n\right)\,dt$$

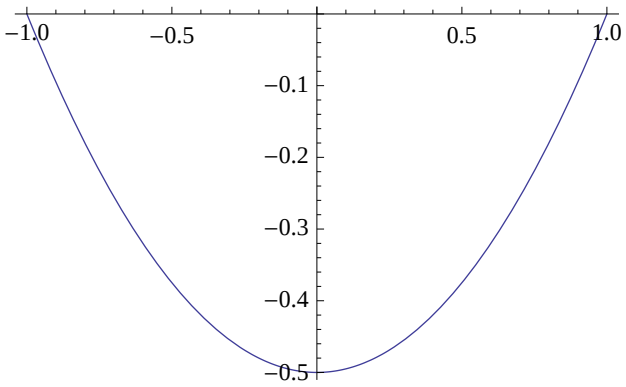
et la conclusion en découle. □

Définition 5.5.4. La fonction $K(t)$ de la proposition précédente est le *noyau de Peano d'ordre n* de la formule d'intégration approchée considérée.

Exemples 5.5.5.
(a) Le noyau de Peano d'ordre 1 de la méthode des trapèzes localisée sur $[-1,1]$ est

$$\begin{aligned}
K(t) &= \frac{1}{1}R[(x-t)_+] \\
&= \int_{-1}^1 (x-t)_+ dx - 2\frac{(1-t)_+ + (-1-t)_+}{2} \\
&= \int_t^1 (x-t)\,dx - (1-t) \\
&= \frac{(x-t)^2}{2}\Big|_t^1 - (1-t) \\
&= \frac{(1-t)^2}{2} - (1-t) = \frac{t^2-1}{2}.
\end{aligned}$$

Son graphe est



(b) Le noyau de Peano d'ordre 3 de la méthode de Simpson localisée sur $[-1,1]$ est

$$K(t) = \frac{1}{3!} R[(x-t)_+^3].$$

Donc,

$$6K(t) = \int_{-1}^{1} (x-t)_+^3 \, dx - \frac{2}{6}[(1-t)_+^3 + 4(0-t)_+^3 + (-1-t)_+^3].$$

Si $0 < t \leq 1$, on trouve donc

$$6K(t) = \frac{(1-t)^4}{4} - \frac{(1-t)^3}{3}$$

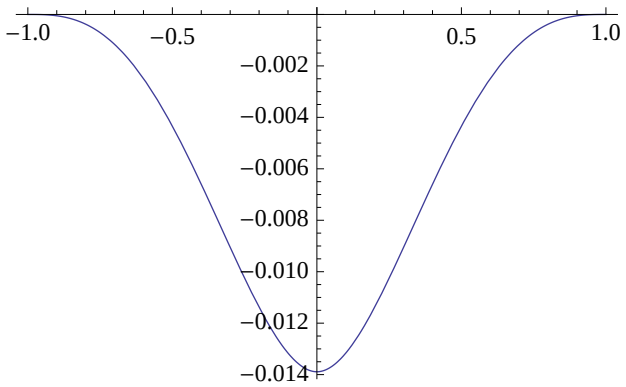
et après simplification

$$K(t) = -\frac{(1-t)^3(1+3t)}{72}.$$

Comme $R(f)$ ne change pas si on remplace $f(x)$ par $f(-x)$, $K(t)$ est pair et on a en fait

$$K(t) = -\frac{(1-|t|)^3(1+3|t|)}{72}$$

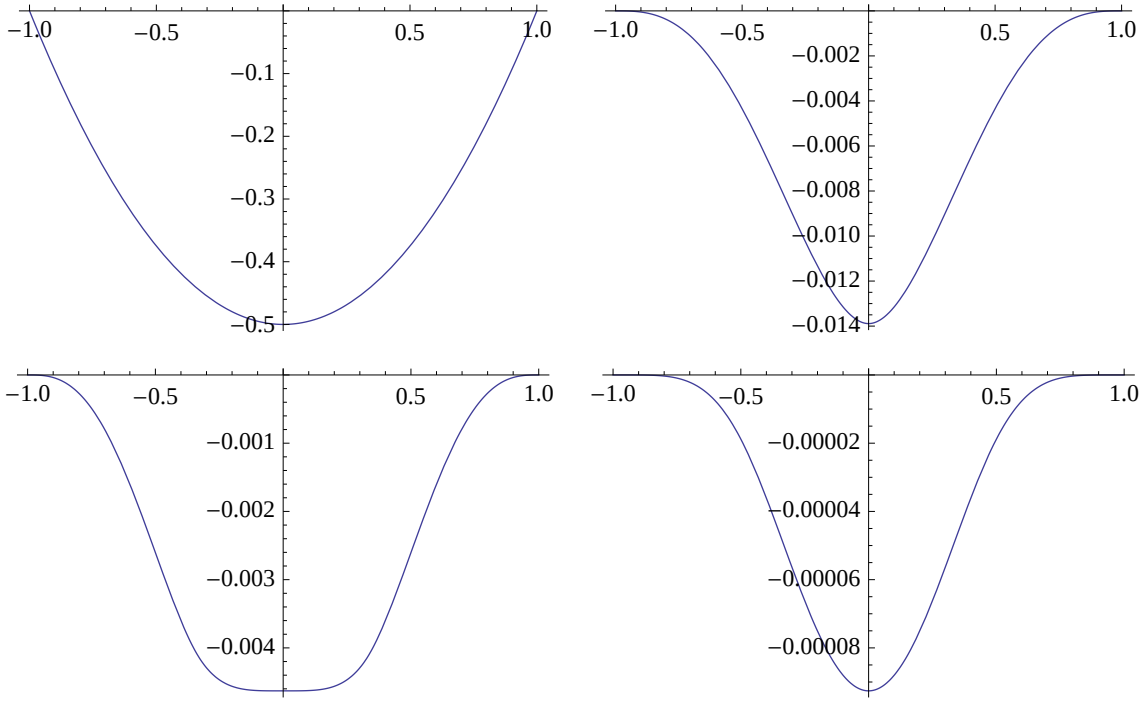
sur $[-1,1]$. Son graphe est



Remarque 5.5.6. Les calculs ci-dessus montrent que les noyaux de Peano des méthodes de Newton-Cotes de degré 1 et 2 sont négatifs sur l'intervalle d'intégration. En fait, on peut montrer plus généralement que le noyau de Peano

$$K_n(t) = \begin{cases} \frac{1}{n!} R(x-t)_+^n & \text{si } n \text{ est impair} \\ \frac{1}{(n+1)!} R(x-t)_+^{n+1} & \text{si } n \text{ est pair} \end{cases}$$

associé à la méthode de Newton-Cotes de degré n, reste de signe constant sur l'intervalle d'intégration. On trouvera ci-dessous les graphes de ces noyaux pour $n = 3, \ldots, 6$.



Proposition 5.5.7. *Soit*

$$\int_a^b f(x)\, dx \approx \sum_{j=0}^{J} \mu_j f(x_j)$$

une formule d'intégration approchée exacte pour les polynômes de degré $\leq n$ et soit $K(t)$ son noyau de Peano d'ordre n. Si $K(t)$ est de signe constant sur $[a,b]$, alors pour tout $f \in C_{n+1}([a,b])$, il existe $\xi \in [a,b]$ tel que

$$R(f) = \left[\int_a^b K(t)\, dt\right] f^{(n+1)}(\xi).$$

De plus,

$$\int_a^b K(t)\,dt = \frac{1}{(n+1)!}R(x^{n+1}).$$

Démonstration. Par la Proposition 5.5.3, on a

$$R(f) = \int_a^b K(t) f^{(n+1)}(t)\,dt.$$

Comme $K(t)$ est de signe constant, le théorème de la moyenne montre que

$$R(f) = \left(\int_a^b K(t)\,dt\right) f^{(n+1)}(\xi)$$

pour un $\xi \in [a,b]$. Dans le cas où $f = x^{n+1}$, on a donc

$$R(x^{n+1}) = (n+1)! \int_a^b K(t)\,dt$$

et la conclusion en résulte. □

Corollaire 5.5.8. *Soit $n \in \mathbb{N}$ et soit*

$$p_n = \begin{cases} n+1 & \text{si } n \text{ est impair;} \\ n+2 & \text{si } n \text{ est pair.} \end{cases}$$

Alors, pour tout $f \in C_{p_n}([a,b])$, il existe $\xi \in [a,b]$ tel que

$$\int_a^b f(x)\,dx = (nh)\sum_{j=0}^{n} c_j f(a+jh) + K_n h^{p_n+1} f^{(p_n)}(\xi)$$

où $h = (b-a)/n$ et où K_n est une constante qui ne dépend que de n.

Démonstration. La proposition précédente s'applique à la méthode de Newton-Cotes de degré n. Ainsi, pour tout $f \in C_{p_n}([0,n])$, il existe $\theta \in [0,n]$ tel que

$$\int_0^n f(t)\,dt = n\sum_{j=0}^{n} c_j f(j) + K_n f^{(p_n)}(\theta)$$

avec

$$K_n = R\left[\frac{t^{p_n}}{p_n!}\right];$$

R étant le reste de la formule de Newton-Cotes de degré n sur $[0,n]$. Si f est de classe C_{p_n} sur $[a,b]$, alors

$$f(a+th)$$

est de classe C_{p_n} sur $[0, n]$. Vu ce qui précède, il existe donc $\theta \in [0, n]$ tel que

$$\int_0^n f(a+th)\,dt = n\sum_{j=0}^{n} c_j f(a+jh) + K_n h^{p_n} f^{(p_n)}(a+\theta h).$$

La conclusion résulte alors de ce que

$$\int_a^b f(x)\,dx = h\int_0^n f(a+th)\,dt.$$

□

Exemples 5.5.9.
(a) Cas $n = 1$. On a $p_n = 2$ et

$$K_n = \int_0^1 \frac{t^2}{2}\,dt - \frac{1}{2}(0+\frac{1}{2}) = \frac{1}{6} - \frac{1}{4} = -\frac{1}{12}.$$

(b) Cas $n = 2$. On a $p_n = 4$ et

$$K_n = \int_0^2 \frac{t^4}{24}\,dt - \frac{2}{6}(0+\frac{4}{24}+\frac{16}{24}) = \frac{1}{3}\,\frac{4}{5} - \frac{1}{3}\,\frac{5}{6} = -\frac{1}{90}.$$

(c) En procédant de même pour $n = 3, 4, 5, 6$, on trouve la table suivante :

n	p_n	$-K_n$
3	4	3/80
4	6	8/945
5	6	275/12096
6	8	9/1400

Définition 5.5.10. La *formule de Newton-Cotes composée de degré n et de pas* $h = (b-a)/(nN)$ est la formule d'intégration approchée

$$\int_a^b f(x)\,dx = \sum_{k=0}^{N-1} nh \sum_{j=0}^{n} c_j f(a+(kn+j)h)$$

obtenue en approchant les intégrales

$$\int_{a+knh}^{a+(k+1)nh} f(x)\,dx \qquad (k = 0, \ldots, N-1)$$

par la formule de Newton-Cotes de degré n

$$nh\sum_{j=0}^{n} c_j f(a+(kn+j)h).$$

Pour $n = 1$ et $n = 2$, on obtient la méthode des trapèzes et la méthode de Simpson.

Proposition 5.5.11. *Si $f \in C_{p_n}([a,b])$, alors le reste de la formule de Newton-Cotes composée de degré n considérée ci-dessus est égal à*

$$K_n \frac{(b-a)}{n} h^{p_n} f^{(p_n)}(\xi)$$

pour un $\xi \in [a,b]$. En particulier, son module est majoré par

$$|K_n| \frac{b-a}{n} h^{p_n} \sup_{\xi \in [a,b]} |f^{(p_n)}(\xi)|$$

et se comporte comme

$$O(h^{p_n})$$

si $h \to 0^+$.

Démonstration. On a

$$\int_{a+knh}^{a+(k+1)nh} f(x)\,dx = nh \sum_{j=0}^{n} c_j f(a+knh+jh) + K_n h^{p_n+1} f^{(p_n)}(\xi_k)$$

avec $\xi_k \in [a+knh, a+(k+1)nh]$. Il s'ensuit que

$$\int_a^b f(x)\,dx = \sum_{k=0}^{N-1} nh \sum_{j=0}^{n} c_j f(a+knh+jh) + K_n h^{p_n+1} \sum_{k=0}^{N-1} f^{(p_n)}(\xi_k).$$

Par la formule de la moyenne,

$$\frac{1}{N} \sum_{k=0}^{N-1} f^{(p_n)}(\xi_k) = f^{(p_n)}(\xi)$$

pour un $\xi \in [a,b]$. Ainsi, le reste cherché est

$$K_n h^{p_n+1} N f^{(p_n)}(\xi) = K_n h^{p_n} \frac{(b-a)}{n} f^{(p_n)}(\xi),$$

d'où la conclusion. □

Exemples 5.5.12. Pour la méthode des trapèzes, on retrouve bien sûr un reste égal à

$$-\frac{(b-a)}{12} h^2 f''(\xi)$$

et pour la méthode de Simpson, on trouve un reste égal à

$$-\frac{(b-a)}{180} h^4 f^{(4)}(\xi).$$

Table des matières