Introduction à l'analyse numérique

(S-MATH-020)

Toutes suggestions et corrections peuvent être envoyées à
Christophe.Troestler@umh.ac.be

Table des matières

Chapitre I

Résolution d'équations non-linéaires

Dans ce chapitre, nous allons nous efforcer de présenter quelques méthodes de base pour résoudre des équations du type

$$f(x) = 0 \tag{I.1}$$

où $f : [a,b] \to \mathbb{R}$ est une fonction continue. Malgré son apparente simplicité, (I.1) symbolise en fait une grande diversité de problèmes.

Tout d'abord, on pense bien sûr aux équations algébriques où f est un polynôme. Or on a vu au chapitre II que l'évaluation d'un polynôme, et à fortiori les calcul de ses racines, pouvait être fort sensible à des petites perturbations sur ses coefficients. En d'autres termes, avant même de parler de *méthodes* pour trouver des racines, il faut utiliser les outils du chapitre précedent pour évaluer la *faisabilité* d'une telle démarche. Les fonctions polynomiales offrent déjà un vaste champ d'investigation, par exemple on peut vouloir calculer toutes les racines réelles d'un polynôme ou si on veut inverser une série de puissances jusqu'à un ordre n.

Mais, bien entendu, dans (I.1), f peut être bien autre chose qu'un polynôme. Dès qu'on s'éloigne un peu des exercices des livres scolaires où les questions sont choisies pour que les résultats « tombent bien », on aboutit aisément au problème de trouver une quantité x vérifiant $f(x) = 0$ sans qu'on puisse déterminer x explicitement. Comme exemple élémentaire, considérons simplement le calcul des fonctions telles que $y = \arcsin(x)$, $y = \operatorname{arctg}(x)$,... Les techniques développées ci-après permettent de les évaluer en résolvant les équations $\sin y = x$, $\operatorname{tg} y = x$,... sur des intervalles adéquats.

Une équation du type (I.1) recouvre encore beaucoup d'autres applications. Considérons par exemple le lancement d'un projectile. Selon la loi de Newton, sa trajectoire est décrite dans un repère cartésien par une fonction $t \mapsto x(t) = \big(x_1(t), x_2(t)\big)$ qui doit satisfaire une équation du type

$$\partial_t^2 x(t) = F\big(t, \partial_t x(t), x(t)\big).$$

Chercher par exemple à savoir à quel moment le projectile retombe sur le sol revient à résoudre $x_2(t) = 0$. Ainsi, si on dispose d'une méthode numérique pour estimer $x(t)$, on pourra utiliser les méthodes de ce chapitre pour résoudre $x_2(t) = 0$.

Puisqu'en général la solution d'une équation $f(x) = 0$ ne s'exprime pas par une « formule », on ne peut espérer trouver une solution exacte (ou une représentation symbolique de celle-ci) en un nombre fini d'étapes. Le mieux qu'on puisse faire est d'*approcher* les solutions avec une précision aussi bonne qu'on le souhaite. Mathématiquement, cela signifie qu'on a une suite $(x_n)_{n\in\mathbb{N}}$ de « solutions approchées », c'est-à-dire telle que $x_n \to x^\star$ où $x^\star$ est une racine : $f(x^\star) = 0$.

Avoir des méthodes pour obtenir des solutions de $f(x) = 0$ de manière approchée est intéressant mais, si on veut les appliquer à des problèmes concrets, le temps qu'il faudra attendre pour obtenir la réponse est aussi crucial. Par exemple, en ce qui concerne le projectile heurtant le sol, le coût de calcul de $x(t)$ peut être realtivement élevé et on voudrait donc que la méthode de résolution de $x_2(t) = 0$ converge en aussi peu d'étapes que possible. En effet, le résultat de ce calcul est peut-être utilisé pour prendre des décisions quant à la trajectoire ultérieure du projectile. Cette vitesse de convergence s'exprime ici par le gain de précision qu'on gagne en passant de x_n à x_{n+1}. Nous allons commencer par définir précisément ce concept avant de voir des méthodes de résolution spécifiques.

I.1 Ordre de convergence

Commencons par introduire une notation qui nous sera fort utile.

Définition I.1. Soient (x_n) et (y_n) deux suites. On dit que (x_n) est un *grand O* de (y_n) lorsque $n \to \infty$ si il existe un $n_0 \in \mathbb{N}$ et une constante $C > 0$ tels que

$$\forall n \geqslant n_0, \quad |x_n| \leqslant C y_n$$

Quand c'est le cas, on note $x_n = \mathrm{O}(y_n)$.

Remarque I.2. La notation $x_n = \mathrm{O}(y_n)$ est abusive. En effet, on peut aussi écrire $z_n = \mathrm{O}(y_n)$ pour d'autres suites (z_n). On devrait donc plutôt noter $(x_n) \in \mathrm{O}(y_n)$. Cependant, l'abus $x_n = \mathrm{O}(y_n)$ est fort pratique car il permet d'employer le symbole $\mathrm{O}(y_n)$ dans des expressions. Par exemple, on peut écrire le développement de Taylor de f autour de x_0 comme

$$f(x) = \sum_{i=0}^{n} \tfrac{1}{i!}\partial^i f(x_0)(x-x_0)^i + \mathrm{O}\big(|x-x_0|^{n+1}\big).$$

Le symbole $\mathrm{O}(y_n)$ est alors pensé comme représentant une certaine suite au sujet de laquelle la seule chose qui nous intéresse est qu'elle est bornée par Cy_n pour un certain

C. Lorsqu'on effectue des calculs, la suite représentée par $\mathrm{O}(y_n)$ peut changer à chaque occurrence du symbole $\mathrm{O}(y_n)$. Ainsi on écrira $\mathrm{O}(y_n)+\mathrm{O}(y_n)=\mathrm{O}(y_n)$ et $2\,\mathrm{O}(y_n)=\mathrm{O}(y_n)$!

Cette notation est communément utilisée lorsqu'on parle de la complexité des programmes. À titre d'exemple, on voit immédiatement (faites les détails !) que $n^2-n+17=\mathrm{O}(n^2)$, $\sin(n)=\mathrm{O}(1)$ et $n\sin(n)=\mathrm{O}(n)$.

Intéressons nous maintenant à ce qu'on appelle la convergence linéaire. Prenons une suite $(x_n)_{n\in\mathbb{N}}$ convergeant vers $x^\star$. La question est de caractériser la vitesse avec laquelle x_n se rapproche de $x^\star$. Le cas le plus simple est quand la distance de x_{n+1} à $x^\star$ est au plus une fraction de la distance de x_n à $x^\star$. Cela s'écrit :

$$|x_{n+1}-x^\star|\leqslant c|x_n-x^\star|,\qquad c\in\,]0,1[.$$

En appliquant cette inégalité de manière récurrente, on obtient $|x_n-x^\star|\leqslant c|x_{n-1}-x^\star|\leqslant c^2|x_{n-2}-x^\star|\leqslant\cdots\leqslant c^n|x_0-x^\star|=\mathrm{O}(c^n)$. C'est ce résultat que nous prendrons comme définition de la convergence linéaire.

Définition I.3. On dit qu'une suite (x_n) *converge (au moins) linéairement* ou *à l'ordre* 1 vers $x^\star$ s'il existe un $c\in[0,1[$ tel que

$$x_n=x^\star+\mathrm{O}(c^n)\quad\text{lorsque } n\to\infty. \tag{I.2}$$

Remarquons que, comme $c<1$, la convergence de x_n vers $x^\star$ est une conséquence de (I.2). D'autre part, au vu de ce qu'on vient de dire, si

$$\limsup_{n\to\infty}\frac{|x_{n+1}-x^\star|}{|x_n-x^\star|}<1,$$

alors $x_n\to x^\star$ à l'ordre 1. En particulier, si

$$\lim_{n\to\infty}\frac{x_{n+1}-x^\star}{x_n-x^\star}\in\,]-1,1[$$

(l'écriture suppose implicitement que la limite existe), alors $x_n\to x^\star$ linéairement.

La convergence à des ordres supérieurs à 1 vise à exprimer que x_{n+1} est beaucoup plus proche de $x^\star$ que x_n. C'est le cas si

$$|x_{n+1}-x^\star|\leqslant c|x_n-x^\star|^p$$

où $c > 0$ et $p > 1$. en utilisant cette formule de manière répétée, on obtient

$$\begin{aligned}|x_{n_0+k} - x^\star| &\leqslant c|x_{n_0+k-1} - x^\star|^p \leqslant c\,c^p|x_{n_0+k-2} - x^\star|^{p^2} \\ &\leqslant c\,c^p \cdots c^{p^{k-1}} |x_{n_0} - x^\star|^{p^k} = c^{1+p+\cdots+p^{k-1}} |x_{n_0} - x^\star|^{p^k} \\ &= c^{(p^k-1)/(p-1)}|x_{n_0} - x^\star|^{p^k} = c^{-1/(p-1)} \big(c^{1/(p-1)}\,|x_{n_0} - x^\star|\big)^{p^k}\end{aligned}$$

Dans cette formule, $c^{-1/(p-1)}$ peut être vu comme une constante et, si n_0 est suffisamment grand, on peut avoir que $d := c^{1/(p-1)}|x_{n_0} - x^\star| < 1$. En conclusion, $|x_n - x^\star| \leqslant$ const. d^{p^k} lorsque $n \geqslant n_0$. C'est ce que nous prendrons comme définition.

Définition I.4. On dit qu'une suite (x_n) converge vers $x^\star$ (au moins) *à l'ordre* p s'il existe un $c \in [0,1[$ tel que

$$x_n = x^\star + \mathrm{O}(c^{p^n}) \quad \text{lorsque } n \to \infty. \tag{I.3}$$

Comme pour la convergence linéaire, (I.3) implique que x_n converge vers $x^\star$. On vient de voir que que, si

$$\limsup_{n\to\infty} \frac{|x_{n+1} - x^\star|}{|x_n - x^\star|^p} < \infty \qquad \text{et} \qquad x_n \to x^\star,$$

alors $x_n \to x^\star$ à l'ordre p. La première condition dit que la suite $\big(\frac{|x_{n+1}-x^\star|}{|x_n-x^\star|^p}\big)_n$ est bornée. D'autre part, on peut remarquer que si $x_n \to x^\star$ à l'ordre p, alors $x_n \to x^\star$ à l'ordre q pour tout q entre 1 et p inclus.

I.2 Méthode de la bissection

La méthode de la bissection est un procédé qui permet à la fois de montrer l'existence d'une racine d'une fonction $f : [a,b] \to \mathbb{R}$ et de l'estimer numériquement. L'idée est simple : si f est continue et change de signe sur $[a,b]$, il faut que f s'annule en un certain point de $[a,b]$. Supposons que $f(a) < 0$ et $f(b) > 0$. Choisissons au hasard un point $x_0 \in]a,b[$. Si $f(x_0) = 0$, on a fini. Si $f(x_0) < 0$, alors il doit y avoir une racine dans $]x_0,b[=:]a_1,b_1[$. Sinon, $f(x_0) > 0$ et il doit y avoir une racine dans $]a,x_0[=:]a_1,b_1[$. On recommence la procédure en choisissant x_1 dans $[a_1,b_1]$ et ainsi de suite ce qui donne une suite décroissante d'intervalles $[a_n,b_n]$ contenant chacun une racine. On espère que a_n et b_n sont de bonnes approximations d'une racine $x^\star$: $a_n \leqslant x^\star \leqslant b_n$ et $a_n \to x^\star$, $b_n \to x^\star$. Pour que ce soit le cas, il faut que la longueur de l'intervalle $[a_n,b_n]$ tende vers 0. Il faut donc bien choisir les points x_n (trouvez un exemple où ni a_n ni b_n ne converge vers une racine). Le théorème s'énonce comme suit.

Théorème I.5. *Soit $f : [a,b] \to \mathbb{R}$ une fonction continue. Si $f(a)f(b) < 0$, la fonction f possède au moins une racine dans $]a,b[$. De plus, si on définit par récurrence $[a_0,b_0] = [a,b]$,*

$$x_n = \tfrac{1}{2}(a_n + b_n) \quad et \quad [a_{n+1},b_{n+1}] = \begin{cases} [a_n,x_n] & si\ f(a_n)f(x_n) < 0, \\ [x_n,x_n] = \{x_n\} & si\ f(x_n) = 0, \\ [x_n,b_n] & si\ f(x_n)f(b_n) < 0, \end{cases} \tag{I.4}$$

les trois suites (a_n), (b_n) et (x_n) convergent linéairement *vers la même limite $x^\star$ avec $f(x^\star) = 0$.*

Démonstration. Quitte à remplacer f par $-f$, on peut supposer $f(a) < 0$ et $f(b) > 0$. Commençons par montrer par récurrence que $[a_n,b_n]$ est bien défini et que $f(a_n)f(b_n) < 0$ sauf si $a_n = b_n$ auquel cas $f(a_n) = f(b_n) = 0$. Pour $n = 0$ il n'y a rien à montrer. Supposons que ce soit vrai pour n et montrons le pour $n+1$. De deux choses l'une. Soit $a_n = b_n$ sont racines et alors $a_{n+1} = b_{n+1} = a_n$ sont aussi racines. Soit $a_n \neq b_n$ et $f(a_n)f(b_n) < 0$, ce qui implique que

- si $f(a_n)f(x_n) < 0$ ou $f(x_n)f(b_n) < 0$, $a_{n+1} \neq b_{n+1}$ et $f(a_{n+1})f(b_{n+1}) < 0$;
- sinon, $f(a_n)f(x_n)f(x_n)f(b_n) \geqslant 0$ d'où on déduit que $f(x_n) = 0$ et que $a_{n+1} = b_{n+1} = x_n$ sont des racines de f.

Vu la définition de (I.4), il est aisé de voir que

$$\forall n \in \mathbb{N}, \qquad [a_{n+1},b_{n+1}] \subseteq [a_n,b_n] \quad \text{et} \quad |b_{n+1} - a_{n+1}| \leqslant \tfrac{1}{2}|b_n - a_n|.$$

Cela implique que la suite (x_n) est de Cauchy. Soit en effet $\varepsilon > 0$. Puisque $1/2^n \to 0$, il existe un $n_0 \in \mathbb{N}$ tel que $n \geqslant n_0 \Rightarrow (1/2^n)|b_0 - a_0| \leqslant \varepsilon$. Pour tous les $m \geqslant n \geqslant n_0$, on a que $x_m \in [a_m,b_m] \subseteq [a_n,b_n]$ et dès lors

$$|x_m - x_n| \leqslant |b_n - a_n| \leqslant \tfrac{1}{2}|b_{n-1} - a_{n-1}| \leqslant \cdots \leqslant \frac{1}{2^n}|a_0 - b_0| \leqslant \varepsilon.$$

La suite (x_n) est donc bien de Cauchy. Par conséquent, il existe un $x^\star \in [a,b]$ tel que $x_n \to x^\star$. D'autre part, vu que $|x_n - a_n| \leqslant |b_n - a_n| \xrightarrow[n\to\infty]{} 0$ et $|b_n - x_n| \leqslant |b_n - a_n| \xrightarrow[n\to\infty]{} 0$, il est facile de montrer (faites le !) que (a_n) et (b_n) convergent aussi vers $x^\star$. Puisque $f(a_n)f(b_n) \leqslant 0$ pour tout n, on en déduit en passant à la limite sur n et en utilisant la continuité de f que $f(x^\star)^2 \leqslant 0$, c'est-à-dire $f(x^\star) = 0$.

Nous avons donc montré que f possède une racine — à savoir $x^\star$ — et que les suites (a_n), (b_n) et (x_n) convergent toutes trois vers $x^\star$. Reste à voir que cette convergence est linéaire. Nous allons le voir pour (x_n), l'argument étant le même pour (a_n) et (b_n). Puisque $(x_m)_{m\geqslant n} \subseteq [a_n,b_n]$ et que $x_m \to x^\star$, on a que $x^\star \in [a_n,b_n]$. En conséquence

$$|x_n - x^\star| \leqslant |b_n - a_n| \leqslant \frac{1}{2^n}|b_0 - a_0| = \mathrm{O}(c^n) \tag{I.5}$$

où $c = 1/2 \in]0,1[$. □

Remarque I.6.

- La démonstration ci-dessus montre en fait que les fonctions continues possèdent la *propriété de valeur intermédiaire*, à savoir que pour tout $y \in [f(a), f(b)]$, il existe un $x \in [a,b]$ tel que $f(x) = y$.
- Une estimation du type (I.5) est extrêmement intéressante car elle permet de donner à priori le nombre d'étapes suffisant pour connaître $x^\star$ avec une tolérance ε. En effet, si on veut savoir à partir de quel n on a $|x_n - x^\star| \leqslant \varepsilon$, il suffit de chercher n tel que $(1/2^n)|b-a| \leqslant \varepsilon$. C'est équivalent à $n \geqslant \lceil \log_2(|b-a|/\varepsilon) \rceil$ où $\lceil \xi \rceil$ dénote le plus petit entier $\geqslant \xi$.

I.3 Méthode de la fausse position

La méthode de la bissection est intéressante car elle converge sous des hypothèses très faibles. Malheureusement, la vitesse de convergence n'est pas très élevée : même si le graphe de f est un segment de droite — f est affine —, il faudra de nombreuses itérations avant d'avoir une estimation relativement précise de la racine. D'où l'idée que, au lieu de prendre pour x_n le point milieu de $[a_n, b_n]$, il vaudrait peut-être mieux choisir x_n comme l'intersection du segment de droite joignant $\big(a_n, f(a_n)\big)$ et $\big(b_n, f(b_n)\big)$ avec l'axe « des x » : $\mathbb{R} \times \{0\}$. Cela donne la formule :

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

On peut espérer ainsi que x_n converge plus vite vers $x^\star$. Le désavantage de ce choix est que nous aurons besoin de plus d'hypothèses sur f pour montrer cette convergence.

Théorème I.7. *Soit $f \in \mathcal{C}([a,b];\mathbb{R}) \cap \mathcal{C}^1(]a,b[;\mathbb{R})$ une fonction convexe ou concave* [1] *et $f(a)f(b) < 0$. Définissons a_n, b_n, x_n par la récurrence suivante : $a_0 = a$, $b_0 = b$ et*

$$x_n = a_n - \frac{b_n - a_n}{f(b_n) - f(a_n)} f(a_n), \qquad [a_{n+1}, b_{n+1}] = \begin{cases} [a_n, x_n] & \text{si } f(a_n)f(x_n) < 0, \\ [x_n, b_n] & \text{si } f(x_n)f(b_n) < 0. \end{cases} \tag{I.6}$$

Alors, soit il existe un n tel que $f(x_n) = 0$, soit x_n est bien défini pour tout n et x_n converge à l'ordre 1 vers $x^\star$ où $x^\star$ est l'unique racine de f dans $[a,b]$.

Démonstration. Pour fixer les idées, disons que f est convexe et que $f(a) < 0$, $f(b) > 0$. Les autres cas sont similaires.

1. Voir l'annexe A pour les définitions et quelques propriétés des fonctions convexes et concaves.

(1) *f possède une racine unique $x^\star$ dans $[a,b]$.* En effet, le théorème de la valeur intermédiaire montre l'existence d'une racine $x^\star$ de f. Supposons que $x^{\star\star}$ soit une autre racine de f. Quitte à permuter $x^\star$ et $x^{\star\star}$, on peut supposer $a < x^{\star\star} < x^\star$. Mais dès lors, $x^{\star\star} = \lambda a + (1-\lambda)x^\star$ pour un certain $\lambda \in]0,1[$ et la convexité de f implique que

$$0 = f(x^{\star\star}) \leqslant \lambda f(a) + (1-\lambda) f(x^\star) = \lambda f(a) < 0,$$

ce qui est une contradiction.

(2) *Pour tout n, $a_{n+1} = x_n \nearrow$ et $b_{n+1} = b$.* Pour voir cela, il suffit de montrer que $f(a_n) < 0$ et $f(b_n) > 0$ impliquent $f(x_n) \leqslant 0$. Il suffit ensuite de procéder par récurrence puisque, dès que $f(x_n) < 0$, il découle de (I.6) que $a_{n+1} = x_n > a_n = x_{n-1}$ et $b_{n+1} = b_n = b$ d'où $f(a_{n+1}) < 0$ et $f(b_{n+1}) > 0$ — si $f(x_n) = 0$, on s'arrête là.

Puisque $\big(f(b_n) - f(a_n)\big)/(b_n - a_n) > 0$ et $f(a_n) < 0$, il découle de (I.6) que $x_n > a_n$. D'autre part, $x_n < b_n$ est équivalent à $f(b_n) > 0$. Reste à voir que $f(x_n) \leqslant 0$. Puisque $x_n \in]a_n, b_n[$, il peut s'écrire comme $x_n = \lambda_n a_n + (1-\lambda_n) b_n$ pour un $\lambda_n \in]0,1[$. Par définition de x_n, on a que $\lambda_n f(a_n) + (1-\lambda_n) f(b_n) = 0$. On le voit facilement en y subtituant $\lambda_n = (x_n - b_n)/(a_n - b_n)$ et en utilisant (I.6). Mais alors la convexité de f implique que

$$f(x_n) = f\big(\lambda_n a_n + (1-\lambda_n) b_n\big) \leqslant \lambda_n f(a_n) + (1-\lambda_n) f(b_n) = 0.$$

(3) *$x_n \to x^\star$.* Vu que (x_n) est croissante et majorée par b, $x_n \to x_\infty := \sup_{n \geqslant 0} x_n$. D'autre part, vu le point (2) ci-dessus, (I.6) se réécrit

$$x_{n+1} = x_n - \frac{b - x_n}{f(b) - f(x_n)} f(x_n). \tag{I.7}$$

En passant à la limite dans cette expression multipliée par $f(b) - f(x_n)$, on trouve que $(b - x_\infty) f(x_\infty) = 0$. Par ailleurs $x_n < x^\star$ — en effet, puisque $f(x_n) < 0$, la seule racine de f dans $[a,b]$ doit appartenir à $]x_n, b[$. Dès lors, $x_\infty \leqslant x^\star < b$ et, grâce à $(b - x_\infty) f(x_\infty) = 0$, on conclut que $f(x_\infty) = 0$, d'où il vient $x_\infty = x^\star$.

(4) *$x_n \to x^\star$ linéairement.* En soustrayant $x^\star$ de chaque membre de l'équation (I.7) et en se remémorant que $f(x^\star) = 0$, on trouve

$$\frac{x_{n+1} - x^\star}{x_n - x^\star} = 1 - \frac{b - x_n}{f(b) - f(x_n)} \frac{f(x_n) - f(x^\star)}{x_n - x^\star}.$$

En passant à la limite $n \to \infty$, on a

$$\lim_{n \to \infty} \frac{x_{n+1} - x^\star}{x_n - x^\star} = 1 - \frac{b - x^\star}{f(b) - f(x^\star)} \partial f(x^\star) =: c.$$

Il suffit de montrer que $c \in [0,1[$ pour prouver la convergence linéaire. Vu la convexité de f, la proposition A.5 (2) nous dit que

$$f(b) \geqslant f(x^\star) + \partial f(x^\star)(b-x^\star) \quad \text{et} \quad f(a) \geqslant f(x^\star) + \partial f(x^\star)(a-x^\star).$$

En tenant compte que $a < x^\star < b$, on peut transformer ces deux inégalités en

$$\frac{f(b)-f(x^\star)}{b-x^\star} \geqslant \partial f(x^\star) \qquad \text{et} \qquad \partial f(x^\star) \geqslant \frac{f(x^\star)-f(a)}{x^\star-a} > 0.$$

La première dit que $c \geqslant 0$ et que la seconde que $c < 1$. □

I.4 Méthode de la sécante

La méthode de fausse position conserve la processus de décision de la bissection de manière à être sûr qu'il y a une racine dans $[a_n, b_n]$. La méthode de la sécante est une variante de la méthode de fausse position où on ne cherche plus à préserver cette propriété. On construit x_{n+1} à partir de x_n et x_{n-1} sans se soucier d'encadrer la racine. L'avantage, comme nous allons le voir, est une convergence plus rapide. Le désavantage est que cette convergence ne pourra plus être assurée qu'au voisinage de la racine. C'est ce qu'énoncent précisément les deux théorèmes suivants.

Théorème I.8. *Soit $f \in \mathcal{C}^2(]a,b[;\mathbb{R})$ où $x^\star \in]a,b[$ est un zéro simple [2] de f. Posons $I_\varepsilon :=]x^\star - \varepsilon, x^\star + \varepsilon[$ et*

$$M_\varepsilon := \sup\Big\{ \Big| \frac{\partial^2 f(s)}{2\partial f(t)} \Big| : s,t \in I_\varepsilon \text{ et } \partial f(t) \neq 0 \Big\}.$$

Supposons que $\varepsilon > 0$ soit suffisamment petit pour que $I_\varepsilon \subseteq]a,b[$ et $\varepsilon M_\varepsilon < 1$. Alors, pour tout $x_0 \neq x_1$ dans I_ε, la suite $(x_n)_{n\in\mathbb{N}}$ définie récursivement par

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n) \qquad (n \geqslant 1), \tag{I.8}$$

est bien définie tant que $f(x_n) \neq 0$ et $x_n \xrightarrow[n\to\infty]{} x^\star$.

Remarque I.9. Montrons comme l'affirme l'énoncé qu'il est possible de choisir ε suffisamment petit pour que $I_\varepsilon \subseteq]a,b[$ et $\varepsilon M_\varepsilon < 1$. Tout d'abord, on peut s'arranger pour que $\varepsilon_0 > 0$ soit suffisamment petit pour que $\operatorname{adh} I_{\varepsilon_0} \subseteq]a,b[$. Vu que $\partial f(x^\star) \neq 0$, on peut prendre ε_0 tellement petit que $t \in \operatorname{adh} I_{\varepsilon_0} \Rightarrow \partial f(t) \neq 0$. Dès lors, M_{ε_0} est le supremum de la fonction $I_{\varepsilon_0} \times I_{\varepsilon_0} \to \mathbb{R} : (s,t) \mapsto \big|\partial^2 f(s)/(2f(t))\big|$. Mais $\operatorname{adh} I_{\varepsilon_0} \times$

2. C'est-à-dire que $f(x^\star) = 0$ et $\partial f(x^\star) \neq 0$.

$\operatorname{adh} I_{\varepsilon_0} \to \mathbb{R} : (s,t) \mapsto \left|\partial^2 f(s)/(2f(t))\right|$ est bien définie et continue et son supremum est plus grand que celui de la fonction précédente. Le supremum d'une fonction continue sur un compact est fini. En conséquence $M_{\varepsilon_0} < +\infty$. Pour tout $\varepsilon \leqslant \varepsilon_0$, on aura $\varepsilon M_\varepsilon \leqslant \varepsilon M_{\varepsilon_0}$ et donc, pour avoir $\varepsilon M_\varepsilon < 1$, il suffit de prendre $\varepsilon < 1/M_{\varepsilon_0}$. (Si $M_{\varepsilon_0} = 0$, $\varepsilon \leqslant \varepsilon_0$ suffit.)

Démonstration. (1) $\partial f(t) \neq 0$ *pour tout* $t \in I_\varepsilon$. En effet, nous savons que $M_\varepsilon < 1/\varepsilon < +\infty$. Supposons que $\partial f(t^\star) = 0$ pour un certain $t^\star \in I_\varepsilon$. Soit $\partial^2 f(s) \neq 0$ pour un $s \in I_\varepsilon$ et alors $\left|\partial^2 f(s)/\partial f(t)\right| \xrightarrow[t\to t^\star,\ t\neq t^\star]{} +\infty$, soit $\partial^2 f(s) = 0$ pour tout $s \in I_\varepsilon$ et alors $f(s) = \alpha(x - x^\star)$ avec $\alpha \neq 0$ puisque $\alpha = \partial f(x^\star) \neq 0$, ce qui contredit $\partial f(t^\star) = 0$.

(2) *f est injective sur I_ε et en particulier $x^\star$ est la seule racine de f dans I_ε.* Puisque ∂f est continue et ne s'annule jamais sur I_ε, elle doit être toujours > 0 ou toujours < 0 sur I_ε. Dans le premier cas, f est strictement croissante sur I_ε ; dans le second f est strictement décroissante. C'est dire que f est strictement monotone sur I_ε et donc injective.

(3) *La formule de récurrence (I.8) est bien définie tant que la racine n'est pas atteinte.* Plus précisément, on va montrer par récurrence que, pour tout $n \geqslant 1$,

$$x_n \in I_\varepsilon \quad \text{et} \quad x_n \neq x_{n-1}, \qquad \text{sauf si } f(x_{n-1}) = 0. \tag{I.9}$$

Cela suffit car si $f(x_{n-1}) = 0$ alors, par le point (2), $x_{n-1} = x^\star$ et la racine est atteinte. Si $n = 1$, il n'y a rien à prouver. Supposons que (I.9) soit vrai pour n et montrons que ça le reste pour $n+1$. Puisque $x_n \neq x_{n-1}$ et que f est injective sur I_ε, $f(x_n) \neq f(x_{n-1})$ et x_{n+1} est bien défini. On voit aussi immédiatement que $x_{n+1} = x_n$ si et seulement si $f(x_n) = 0$. Reste à montrer que $x_{n+1} \in I_\varepsilon$. En soustrayant $x^\star$ à chaque membre de (I.8) et en se rappelant que $f(x^\star) = 0$, on trouve

$$\begin{aligned}
x_{n+1} - x^\star &= x_n - x^\star - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n) \\
&= (x_n - x^\star)\Big(1 - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} \frac{f(x_n) - f(x^\star)}{x_n - x^\star}\Big) \\
&= (x_n - x^\star)\Big(1 - \frac{f[x_n, x^\star]}{f[x_{n-1}, x_n]}\Big) \\
&= (x_n - x^\star) \frac{f[x_{n-1}, x_n] - f[x_n, x^\star]}{f[x_{n-1}, x_n]} \\
&= (x_n - x^\star)(x_{n-1} - x^\star) \frac{f[x_{n-1}, x_n, x^\star]}{f[x_{n-1}, x_n]}
\end{aligned}$$

où $f[x_{n-1}, x_n]$, $f[x_n, x^\star]$ et $f[x_{n-1}, x_n, x^\star]$ dénotent les différences divisées de f (voir annexe B). Le théorème B.4 nous dit que

$$f[x_{n-1}, x_n] = \partial f(\eta) \qquad \text{et} \qquad f[x_{n-1}, x_n, x^\star] = \tfrac{1}{2}\partial^2 f(\xi)$$

pour certains $\eta \in [x_{n-1}, x_n]$ et $\xi \in \big[\min\{x_{n-1}, x_n, x^\star\}, \max\{x_{n-1}, x_n, x^\star\}\big]$. Par conséquent

$$|x_{n+1} - x^\star| = |x_n - x^\star|\,|x_{n-1} - x^\star|\left|\frac{\partial^2 f(\xi)}{2\partial f(\eta)}\right| < \varepsilon\varepsilon M_\varepsilon < \varepsilon \tag{I.10}$$

d'où $x_{n+1} \in]x^\star - \varepsilon, x^\star + \varepsilon[= I_\varepsilon$.

(4) *$x_n \xrightarrow[n\to\infty]{} x^\star$ à moins que $x_n = x^\star$ pour un certain n (auquel cas les x_m pour $m > n$ ne sont pas définis).* Au vu de (I.10), on a

$$|x_{n+1} - x^\star| \leqslant |x_n - x^\star|\,(\varepsilon M_\varepsilon)$$

et un argument par récurrence montre qu'alors $|x_{n+1} - x^\star| \leqslant |x_0 - x^\star|(\varepsilon M_\varepsilon)^n \xrightarrow[n\to\infty]{} 0$ puisque $\varepsilon M_\varepsilon < 1$. □

Cette dernière inégalité montre que $|x_n - x^\star| = \mathrm{O}(c^n)$ pour $c := \varepsilon M_\varepsilon \in [0,1[$, c'est-à-dire que $x_n \to x^\star$ à l'ordre 1. En fait, on a mieux comme le montre le théorème suivant.

Théorème I.10. *Sous les hypothèses du théorème précedent, $x_n \to x^\star$ à l'ordre $p = \frac{1}{2}(1+\sqrt{5})$.*

Démonstration. Considérant (I.10) une fois de plus, on voit que

$$|x_{n+1} - x^\star| \leqslant |x_n - x^\star|\,|x_{n-1} - x^\star| M_\varepsilon.$$

Si on pose $E_n := M_\varepsilon\,|x_n - x^\star|$, cela se réécrit

$$E_{n+1} \leqslant E_n E_{n-1}.$$

Nous prétendons que cela implique

$$E_n \leqslant E^{p^n} \qquad \text{où } E := \max\{E_0, E_1^{1/p}\}. \tag{I.11}$$

Pour $n = 0$ et $n = 1$, c'est trivialement vrai par définition de E. Supposons que ce soit vrai pour $n-1$ et n et montrons que c'est vrai pour $n+1$. Ce l'est car

$$E_{n+1} \leqslant E_n E_{n-1} \leqslant E^{p^n} E^{p^{n-1}} = E^{p^{n-1}(1+p)} = E^{p^{n-1}p^2} = E^{p^{n+1}}$$

où on a utilisé le fait que $1 + p = p^2$. Vu la définition de E_n, (I.11) se réécrit

$$|x_n - x^\star| \leqslant \frac{1}{M_\varepsilon} E^{p^n} = \mathrm{O}(E^{p^n}).$$

La preuve sera complète si $E < 1$. On voit facilement que c'est le cas : $E_0 = M_\varepsilon |x_0 - x^\star| < M_\varepsilon \varepsilon < 1$ et de la même manière, $E_1 < 1$. □

Remarque. En relisant la preuve des deux théorèmes ci-dessus, on peut voir qu'en fait on pourrait soit prendre $\varepsilon M_\varepsilon \leqslant 1$ soit définir $I_\varepsilon = [x^\star - \varepsilon, x^\star + \varepsilon]$.

I.5 Méthode de Newton

La méthode de Newton peut être vue comme un cas limite de la méthode de la sécante où les deux points x_{n-1} et x_n sont tellement proches que $(f(x_n)-f(x_{n-1}))/(x_n-x_{n-1})$ est essentiellement la même chose que $\partial f(x_n)$. Ainsi on obtiendra x_{n+1} à partir de x_n en regardant l'intersection de la tangente à f au point x_n avec l'axe « des x ». Comme cette tangente est contituée de l'ensemble des points (x,y) tels que $y = f(x_n)+\partial f(x_n)(x-x_n)$ et que le point recherché est du type $(x_{n+1},0)$, on trouve que

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

Au voisinage de la racine, cette méthode converge plus vite que la méthode de la sécante. C'est ce qu'exprime le théorème suivant.

Théorème I.11. *Soit $f \in \mathcal{C}^2(]a,b[;\mathbb{R})$ et $x^\star \in]a,b[$ une racine simple de f. Pour $\varepsilon > 0$, on définit $I_\varepsilon := [x^\star-\varepsilon, x^\star+\varepsilon]$ et*

$$M_\varepsilon := \sup\Big\{ \frac{\partial^2 f(\xi)}{2\partial f(\eta)} : \xi,\eta \in I_\varepsilon \text{ et } \partial f(\eta) \neq 0 \Big\}.$$

On suppose que $\varepsilon > 0$ est suffisament petit pour que $I_\varepsilon \subseteq]a,b[$ et $\varepsilon M_\varepsilon < 1$. Alors, pour tout $x_0 \in I_\varepsilon$, on peut définir la suite $(x_n)_{n\geqslant 0}$ par récurrence

$$x_{n+1} = x_n - \frac{f(x_n)}{\partial f(x_n)} \tag{I.12}$$

et $x_n \to x^\star$ à l'ordre 2.

Démonstration. (1) De la même manière que pour la méthode de la sécante, on prouve que $\partial f(x) \neq 0$ pour tout $x \in I_\varepsilon$ et que f est injective sur I_ε. En particulier, $x^\star$ est la seule racine de f dans I_ε.

(2) *La suite (x_n) est bien définie.* Puisque $\partial f(x) \neq 0$ pour tout $x \in I_\varepsilon$, la suite est bien définie tant qu'elle reste dans I_ε. Montrons donc par récurrence que $x_n \in I_\varepsilon$ pour tout n. Si $n = 0$, cele découle simplement du choix de x_0. Supposons que $x_n \in I_\varepsilon$ et prouvons que $x_{n+1} \in I_\varepsilon$. En soustrayant $x^\star$ aux deux membres de (I.12) et en utilisant le fait que $f(x^\star) = 0$, on trouve

$$\begin{aligned} x_{n+1}-x^\star &= (x_n-x^\star)\Big(1 - \frac{1}{\partial f(x_n)}\frac{f(x_n)-f(x^\star)}{x_n-x^\star}\Big) \\ &= (x_n-x^\star)\frac{f[x_n,x_n]-f[x_n,x^\star]}{f[x_n,x_n]} = (x_n-x^\star)^2\,\frac{f[x_n,x_n,x^\star]}{f[x_n,x_n]}. \end{aligned}$$

Le théorème B.4 et le développement qui suit impliquent qu'il existe un $\xi_n \in [x_n, x^\star]$ tel que $f[x_n, x_n, x^\star] = \frac{1}{2}\partial^2 f(\xi_n)$. En conséquence,

$$x_{n+1} - x^\star = (x_n - x^\star)^2 \, \frac{\partial^2 f(\xi_n)}{2\partial f(x_n)} \tag{I.13}$$

et donc, puisque $x_n \in I_\varepsilon$, $|x_{n+1} - x^\star| \leqslant |x_n - x^\star|^2 M_\varepsilon \leqslant \varepsilon^2 M_\varepsilon < \varepsilon$. Cela montre que $x_{n+1} \in I_\varepsilon$.

(3) $x_n \to x^\star$ *quadratiquement.* De (I.13), on déduit que $|x_{n+1} - x^\star| \leqslant |x_n - x^\star|(\varepsilon M_\varepsilon)$ et donc que $x_n \to x^\star$. Ceci, combiné avec le fait que que $|x_{n+1} - x^\star| \leqslant M_\varepsilon |x_n - x^\star|^2$, implique la convergence quadratique de x_n vers x^star. □

Nous voudrions maintenant mettre en évidence quelques faiblesses de la méthode de Newton lorsqu'on travaille sur de trop grands voisinages de la racine $x^\star$.

▪ Considérons $f : [-\pi/2, \pi/2] \to \mathbb{R} : x \mapsto \sin x$. Cette fonction possède une racine simple unique, à savoir $x^\star = 0$. La méthode de Newton s'écrit :

$$x_0 \in \,]-\pi/2, \pi/2[, \qquad x_{n+1} = x_n - \operatorname{tg} x_n,\ n \geqslant 0.$$

Choisissons $x_0 = \rho$ où ρ est la racine strictement positive [3] de $\operatorname{tg} x = 2x$. Dans ce cas, $x_1 = x_0 - \operatorname{tg} x_0 = \rho - 2\rho = -\rho$ et $x_2 = x_1 - \operatorname{tg} x_1 = -\rho - \operatorname{tg}(-\rho) = -\rho + \operatorname{tg}\rho = -\rho + 2\rho = \rho$. On ext revenu à x_0 ! Ensuite le processus recommence : $x_3 = -\rho$, $x_4 = \rho$, $x_5 = -\rho$,... La suite $(x_n)_{n\in\mathbb{N}}$ alterne donc entre ρ et $-\rho$. On dit que c'est une *orbite périodique* de période 2 ou un *cycle d'ordre deux*. En conséquence évidemment (x_n) ne converge pas vers 0. Cela met clairement en évidence qu'on doit partir suffisament près de la racine afin d'assurer la convergence de la méthode. Ici on peut montrer [4] que, si $|x_0| < \rho$, alors $x_n \to 0 = x^\star$.

▪ Considérons la fonction $f : [0, +\infty[\to \mathbb{R} : x \mapsto x^{20} - 1$. La seule racine de f est $x^\star = 1$. Comme f est convexe, on peut montrer (c.f. exercice I.3) que la méthode de Newton converge pour tout $x_0 \in \,]0, +\infty[$. C'est à dire que la suite (x_n) définie par

$$x_{n+1} = \frac{19}{20} x_n + \frac{1}{20 x_n^{19}} \qquad (n \geqslant 0),$$

converge vers 1. Prenons par exemple $x_0 = 1/2$. Dès lors $x_1 = 19/40 + 2^{19}/20 \approx 2^{19}/20$ est très grand. Mais alors, dans le calcul de x_2, le terme $1/(20x_1^{19})$ va être très petit si bien que $x_2 \approx (19/20)x_1$. En fait, tant que x_n va rester grand, $x_{n+1} \approx (19/20)x_n$.

3. Prouvez qu'une telle racine existe !

4. Essayez ! Indication : considérons sans perte de généralité $x_0 > 0$. Les sous-suites $(x_{2n})_{n\in\mathbb{N}}$ et $(x_{2n+1})_{n\in\mathbb{N}}$ possèdent les propriétés : $x_{2n} > 0$, $x_{2n} \searrow$, $x_{2n+1} < 0$, $x_{2n+1} \nearrow$. En conséquence, $x_{2n} \to \inf_n x_{2n} =: a \geqslant 0$ et $x_{2n+1} \to \sup_n x_{2n+1} =: b \leqslant 0$. en passant à la limite dans $x_{n+1} = x_n - \operatorname{tg} x_n$, on obtient $a = b - \operatorname{tg} b$ et $b = a - \operatorname{tg} a$. On en déduit $a = -b = 0$.

C'est dire qu'à partir de x_1, la suite va décroître ($19/20 < 1$) mais très lentement ($19/20$ est relativement proche de 1). Ensuite, lorsqu'on arrivera dans un voisinage suffisament petit de la racine, en vertu du théorème I.11, la convergence va s'accélérer : x_n va converger vers 1 à l'ordre 2.

Ce qu'on peut conclude de cet exemple est que la méthode de Newton n'est pas nécessairement plus performante que les autres méthodes si on est trop loin de la racine.

I.6 Méthode du point fixe

Si on regarde la méthode de Newton d'un point de vue abstrait, on voit qu'on obtient x_{n+1} à partir de x_n en évaluant toujours la même expression. Plus précisément, on a $x_{n+1} = \varphi(x_n)$ avec $\varphi(x) := x - f(x)/\partial f(x)$. Si $x_n \to x^\star$, on déduit immédiatement de la continuité de φ que $x^\star = \varphi(x^\star)$. On dit alors que $x^\star$ est un *point fixe* de φ — φ en effet ne « déplace » pas $x^\star$. Or, il se fait que les points fixes de φ correspondent aux zéros simples de f (voyez-vous pourquoi ?).

L'avantage de cette vision plus générale est qu'elle donne un cadre pour rechercher de nouveaux algorithmes. En effet, à toute fonction φ dont les points fixes correspondent aux solutions du problème, on peut associer un schéma récursif $x_{n+1} = \varphi(x_n)$. La question est donc : quels sont les φ intéressants ? Ou plutôt, quelles sont les propriétés que φ doit posséder pour être intéressant. Intéressant ici veut dire deux choses :

- on veut que x_n converge — sa limite $x^\star$ sera alors un point fixe de φ ;
- on veut que $x_n \to x^\star$ aussi vite que possible — l'ordre de convergence doit être aussi élevé que possible.

Nous allons ci-après répondre à ces deux questions. Les développements qui vont suivre vont aussi montrer que la formulation en termes de point fixe est un cadre propice à l'étude de ces questions : les réponses se lisent litéralement sur φ est les dérivées de φ ! Commençons par voir un critère qui implique que le schéma récursif converge.

Théorème I.12. *Soit $\varphi : [a,b] \to [a,b]$ une fonction. On suppose qu'il existe une constante $K \in [0,1[$ telle que*

$$\forall x, y \in [a,b], \qquad |\varphi(x) - \varphi(y)| \leqslant K\,|x-y|. \tag{I.14}$$

Alors, φ possède un unique point fixe $x^\star \in [a,b]$ et, pour tout $x_0 \in [a,b]$, la suite $(x_n)_{n\geqslant 0}$ définie par $x_{n+1} = \varphi(x_n)$ converge vers $x^\star$.

Remarques.

- Une fonction qui satisfait (I.14) pour $K \in [0,+\infty[$ est dite *Lipchitzienne*. Lorsque $K<1$, on dit que φ est une *contraction*.
- Le plus petit K qui satisfait (I.14) (le K optimal) est appelé la constante de Lipschitz de la fonction φ et se note $\mathrm{Lip}(\varphi)$. Ainsi
$$\mathrm{Lip}(\varphi) = \mathrm{Lip}_{[a,b]}(\varphi) := \sup_{\substack{x,y\in[a,b]\\ x\neq y}} \frac{|\varphi(x)-\varphi(y)|}{|x-y|}.$$
- Les fonctions qui satisfont (I.14) sont continues (montrez le !). L'inverse n'est pas vrai (trouvez un exemple).
- Si $\varphi \in \mathcal{C}^1(]a,b[;\mathbb{R})$, on peut montrer (faites le !) grâce au théorème de la moyenne que
$$\mathrm{Lip}(\varphi) = \sup_{x\in]a,b[} |\partial\varphi(x)|.$$
En conséquence, φ sera une contraction si et seulement si $\sup_{x\in]a,b[}|\partial\varphi(x)|<1$. Si de plus φ est dérivable en a et b, il découle de la compacité de $[a,b]$ que φ est une contraction si et seulement si, pour tout $x\in[a,b]$, $|\partial\varphi(x)|<1$.
- Du point de vue de l'existence, l'intérêt de ce théorème est qu'il est valable en dimension supérieure à 1 (voir exercice I.5). En effet, en dimension 1, la continuité suffit (cf. exercice I.4). Notons cependant que, dans ce cas, la convergence des suites (x_n) n'est pas assurée et en fait n'a pas nécessairement lieu. Leur comportement peut d'ailleurs être fort complexe (voir exercice I.19).

Démonstration. Soit $(x_n)_{n\geqslant 0}$ une suite définie par $x_0\in[a,b]$ et $x_{n+1}=\varphi(x_n)$.

(1) *La suite (x_n) est de Cauchy.* En d'autres termes, elle vérifie
$$\forall \varepsilon>0,\ \exists n_0\in\mathbb{N},\ \forall m,n,\quad m\geqslant n\geqslant n_0 \Rightarrow |x_m-x_n|\leqslant\varepsilon.$$
Soit $n,k\in\mathbb{N}$. On a
$$|x_{n+k}-x_n| = \Big|\sum_{0\leqslant i<k} x_{n+i+1}-x_{n+i}\Big| \leqslant \sum_{0\leqslant i<k}|x_{n+i+1}-x_{n+i}|.$$
En utilisant l'inégalité (I.14) de manière répétée, on déduit que $|x_{n+i+1}-x_{n+i}| = |\varphi(x_{n+i})-\varphi(x_{n+i-1})| \leqslant K|x_{n+i}-x_{n+i-1}| = K|\varphi(x_{n+i-1})-\varphi(x_{n+i-2})| \leqslant K^2|x_{n+i-1}-x_{n+i-2}| \leqslant \cdots \leqslant k^{n+i}|x_1-x_0|$. Dès lors,
$$\begin{aligned}|x_{n+k}-x_n| &\leqslant \sum_{0\leqslant i<k} K^{n+i}|x_1-x_0| = K^n\frac{1-K^k}{1-K}\,|x_1-x_0|\\ &\leqslant \frac{K^n}{1-K}\,|x_1-x_0| \xrightarrow[n\to\infty]{} 0.\end{aligned}$$
Ainsi, dès que n est grand et quel que soit $k\geqslant 0$, $|x_{n+k}-x_n|$ est petit. La suite (x_n) est donc de Cauchy.

(2) *La suite (x_n) converge vers un point fixe.* Comme (x_n) est de Cauchy, il découle du fait qur $\mathbb{R}$ est complet que $x_n \to x^\star$ pour un certain $x^\star \in \mathbb{R}$. Vu que $(x_n) \subseteq [a,b]$ et que $[a,b]$ est un ensemble fermé, $x^\star \in [a,b]$.

Par ailleurs, par définition de la suite (x_n), $x_{n+1} = \varphi(x_n)$. En passant à la limite $n \to \infty$ dans cette expression et en utilisant la continuité de φ, on trouve $x^\star = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} \varphi(x_n) = \varphi\big(\lim_{n\to\infty} x_n\big) = \varphi(x^\star)$.

(3) *Le point fixe $x^\star$ est unique.* Soit $x^{\star\star}$ un point fixe de φ. On va prouver que $x^{\star\star}$ ne peut être différent de $x^\star$, c'est-à-dire $x^{\star\star} = x^\star$. En effet, en utilisant (I.14), on a

$$|x^{\star\star} - x^\star| = |\varphi(x^{\star\star}) - \varphi(x^\star)| \leqslant K|x^{\star\star} - x^\star|,$$

et donc $(1-K)|x^{\star\star} - x^\star| \leqslant 0$, ce qui implique $x^{\star\star} - x^\star = 0$ puisque $1 - K > 0$. □

Le théorème I.12 donne un critère pour la convergence des suites sur un intervalle $[a,b]$ Peut-on le particulariser au voisinage d'un point fixe ? Soit donc une fonction $f : \mathbb{R} \circlearrowright \mathbb{R}$ et $x^\star \in \operatorname{int}\operatorname{Dom}\varphi$ un point fixe de φ. Notons $I_\varepsilon := [x^\star - \varepsilon, x^\star + \varepsilon]$ un petit voisinage de $x^\star$. Si on veut montrer que φ est une contraction sur I_ε, il va falloir examiner $\operatorname{Lip}_{I_\varepsilon}(\varphi)$, c'est-à-dire, en vertu des remarques ci-dessus, il va falloir regarder $\partial\varphi$ sur I_ε. Mais, lorsque $\varepsilon > 0$ est petit, les points x de I_ε sont proches de $x^\star$ et donc les dérivées $\partial\varphi(x)$ sont proches de $\partial\varphi(x^\star)$. Ceci suggère qu'il suffit de regarder $\partial\varphi(x^\star)$ pour en déduire que φ est une contraction sur I_ε.

C'est le cas. Supposons que φ soit de classe $\mathcal{C}^1$ et $|\partial\varphi(x^\star)| < 1$. Choisissons c tel que $|\partial\varphi(x^\star)| < c < 1$. Dès lors $c - |\partial\varphi(x^\star)| > 0$. Par continuité de la dérivée, il existe un $\varepsilon > 0$ tel que $I_\varepsilon \subseteq \operatorname{Dom}\varphi$ et, pour tout $x \in I_\varepsilon$, $|\partial\varphi(x) - \partial\varphi(x^\star)| \leqslant c - |\partial\varphi(x^\star)|$. Grâce à l'inégalité $|\partial\varphi(x)| - |\partial\varphi(x^\star)| \leqslant |\partial\varphi(x) - \partial\varphi(x^\star)|$, on déduit que $|\partial\varphi(x)| \leqslant c$ et donc que $\operatorname{Lip}_{I_\varepsilon}(\varphi) \leqslant c < 1$. Pour pouvoir appliquer le théorème I.12, il faut encore que φ soir un endomorphisme de I_ε, c'est-à-dire que $\varphi(I_\varepsilon) \subseteq I_\varepsilon$ — auquel cas on pourra écrire $\varphi : I_\varepsilon \to I_\varepsilon$. Soit donc $x \in I_\varepsilon$. Il faut montrer que $\varphi(x) \in I_\varepsilon$. Cela résulte d'une simple application du théorème de la moyenne :

$$|\varphi(x) - x^\star| = |\varphi(x) - \varphi(x^\star)| = |\partial\varphi(\xi)|\,|x - x^\star| < |x - x^\star| \leqslant \varepsilon$$

où $\xi \in]x, x^\star[$. En conclusion, le théorème I.12 s'applique et on en clonclut que pour tout $x_0 \in I_\varepsilon$, la suite $(x_n)_{n\in IN}$ définie par $x_{n+1} = \varphi(x_n)$ converge bien vers $x^\star$.

Que se passe-t-il si maintenant $|\partial\varphi(x^\star)| > 1$? De nouveau, en utilisant la continuité de la dérivée, on montre (faites le !) qu'il existe un $c > 1$ et un $\varepsilon > 0$ tels que $x \in I_\varepsilon \Rightarrow |\partial\varphi(x)| \geqslant c$. Supposons qu'une suite (x_n) définie par itérations successives de φ entre dans I_ε. Plus précisément, supposons que $x_{n_0} \in I_\varepsilon$. Que peut-on dire de x_n, $n > n_0$? De deux choses l'une : soit x_n quitte I_ε après un certain nombre d'itérations,

soit $\forall n \geqslant n_0,\ x_n \in I_\varepsilon$. Examinons le second cas. Le théorème de la moyenne implique que, pour tout $n > n_0$,

$$|x_n - x^\star| = |\varphi(x_{n-1}) - \varphi(x^\star)| = |\partial\varphi(\xi)|\,|x_{n-1} - x^\star| \geqslant c\,|x_{n-1} - x^\star|$$

où $\xi \in]x_{n-1}, x^\star[\subseteq I_\varepsilon$. En utilisant de manière répétée cette inégalité, on trouve

$$|x_n - x^\star| \geqslant c^{n-n_0}|x_{n_0} - x^\star| \xrightarrow[n\to\infty]{} +\infty$$

où la divergence résulte du fait que $c > 1$. Ceci contredit l'hypothèse $x_n \in I_\varepsilon$ — qui s'écrit $|x_n - x^\star| \leqslant \varepsilon$. Ainsi, en vérité, il est impossible que les suites qui rentrent dans I_ε restent dans I_ε ; elles doivent nécessairement quitter I_ε après un certain temps — en particulier, si on part de $x_0 \in I_\varepsilon$ même fort proche de $x^\star$, la suite ne converge pas vers $x^\star$!

On peut résumer les arguments précédents comme suit.

- Si $|\partial\varphi(x^\star)| < 1$, les suites qui entrent dans un petit voisinage de $x^\star$ convergent vers $x^\star$. On dit que $x^\star$ est un point fixe *attractif*.
- Si $|\partial\varphi(x^\star)| > 1$, même si une suite entre dans un petit voisinage de $x^\star$, elle est forcée d'en ressortir. On dit de $x^\star$ que c'est un point fixe *répulsif*.
- Si $|\partial\varphi(x^\star)| = 1$, on ne peut rien dire.[5] Les deux situations ci-dessus peuvent se produire. Ou aucune d'elles. Cependant on peut penser $|\partial\varphi(x^\star)| = 1$ comme une transition entre $|\partial\varphi(x^\star)| < 1$ et $|\partial\varphi(x^\star)| > 1$, c'est-à-dire entre un point fixe qui était attractif et devient répulsif. De telles situations sont communes et, typiquement, lorsque $|\partial\varphi(x^\star)| = 1$, une *bifurcation* a lieu (voir exercice I.19).

Jusqu'à présent nous avons seulement examiné la convergence — ou non — des suites vers un point fixe. Comme d'habitude, nous voudrions aussi connaître la vitesse de convergence de x_n vers $x^\star$. Globalement, le théorème I.12 ne nous offre qu'une convergence linéaire. En effet, (I.14) implique

$$|x_{n+1} - x^\star| = |\varphi(x_n) - \varphi(x^\star)| \leqslant K\,|x_n - x^\star|.$$

Comment s'exprime le fait que la méthode de Newton est quadratique ? On ne peut espérer avoir cela sur tout $[a,b]$, seulement lorsqu'on est suffisament proche du point fixe $x^\star$. Mais on l'a vu ci-dessus, au voisinage de $x^\star$, c'est la dérivée de φ en $x^\star$ qui est importante. On va donc faire un développement de Taylor avec reste de φ au point $x^\star$. Cela s'écrit :

$$\varphi(x) = \varphi(x^\star) + \partial\varphi(x^\star)(x - x^\star) + \cdots + \frac{\partial^{k-1}\varphi(x^\star)}{(k-1)!}\,(x - x^\star)^{k-1} + \frac{\partial^k\varphi(\xi)}{k!}\,(x - x^\star)^k$$

5. C'est analogue à la situation où $\partial f(x^\star) = 0$ et $\partial^2 f(x^\star) = 0$. On peut trouver des exemples où le point $x^\star$ est un minimum de f, d'autres où il est un maximum de f, et d'autres encore où il n'est ni minimum ni maximum de f. On ne peut donc rien conclure.

où $\xi \in]x, x^\star[$ dépend de x et on a supposé que $\varphi \in \mathcal{C}^k$. Ainsi, si $\partial^i \varphi(x^\star) = 0$ pour $i = 1, \dots, k-1$ et $\partial^k \varphi(x^\star) \neq 0$, on a

$$x_{n+1} - x^\star = \varphi(x_n) - \varphi(x^\star) = \frac{\partial^k \varphi(\xi_n)}{k!} (x_n - x^\star)^k$$

où $\xi_n \in]x_n, x^\star[$. Puisque $\partial \varphi(x^\star) = 0$, on sait qu'il existe un $\varepsilon > 0$ tel que, dès que x_n entre dans I_ε, il y reste et converge vers $x^\star$. si on pose $c := \sup_{x \in I_\varepsilon} |\partial^k \varphi(x)|/k!$, on peut écrire

$$|x_{n+1} - x^\star| = \frac{|\partial^k \varphi(\xi_n)|}{k!} |x_n - x^\star|^k \leqslant c\, |x_n - x^\star|^k.$$

Cela implique que (x_n) converge vers $x^\star$ à l'ordre k. Le fait que $\partial^k \varphi(x^\star) \neq 0$ ne nous permet pas de refaire la même chose avec $k+1$ au lieu de k. Cela indique qu'on n'a pas que $x_n \to x^\star$ à l'ordre $k+1$. Nous venons de prouver le théorème suivant.

Théorème I.13. *Sous les hypothèses du théorème I.12, si on a de plus que* $\varphi \in \mathcal{C}^k(]a,b[;\mathbb{R})$ *et* $\partial \varphi(x^\star) = 0, \dots, \partial^{k-1} \varphi(x^\star) = 0$*, alors* (x_n) *converge vers* $x^\star$ *à l'ordre* k*. Plus précisément, on a* $|x_{n+1} - x^\star| \leqslant c\, |x_n - x^\star|^k$ *où* $c > |\partial^k \varphi(x^\star)|/k!$ *peut être choisi arbitrairement proche de* $|\partial^k \varphi(x^\star)|/k!$*.*

I.7 Exercices

Exercice I.1 Prouver les affirmations de la remarque page 14.

Exercice I.2 Soit $f \in \mathcal{C}^2(\mathbb{R};\mathbb{R})$ et $x^\star$ une racine simple de f. On pose $A(x^\star) := \{x_0 \in \mathbb{R} :$ la suite (x_n) construite à partir de x_0 par le procédé de Newton converge vers $x^\star\}$. Montrez que $A(x^\star)$ est ouvert.

Exercice I.3 Soit $f \in \mathcal{C}^1([a,b];\mathbb{R})$ une fonction convexe (ou concave) telle que $f(a)f(b) < 0$. On suppose que les tangentes à f aux points a et b intersectent l'axe « des x » dans l'intervalle $[a,b]$.

- Quelles sont les inégalités correspondant à cette dernière condition.
- Montrez qu'à un des deux points a ou b, cette condition découle de la convexité (ou concavité) de f et de $f(a)f(b) < 0$.
- Prouvez qu'il existe une racine $x^\star \in]a,b[$ de f et que celle-ci est unique.
- Prouvez que $x^\star$ est un zéro simple de f (i.e., $\partial f(x^\star) \neq 0$). Plus généralement, prouvez que $\partial f(x) \neq 0$ pour tout $x \in [a,b]$ et donc que f est strictement monotone.
- Démontrez que, pour tout $x_0 \in]a,b[$, la suite (x_n) construite par la méthode de Newton converge vers $x^\star$. (Indication : examinez la croissance ou la décroissance de la suite.)

Exercice I.4 Montrez en utilisant la propriété de valeur intermédiaire que toute fonction continue $f : [a,b] \to [a,b]$ possède au moins un point fixe.

Exercice I.5 Énoncez et démontrez un théorème analogue au théorème I.12 pour $\varphi : X \to X$ où X est un sous-ensemble fermé de $\mathbb{R}^N$.

Exercice I.6 Considérons un modèle simplifié d'évolution de la population. Appelons $P(t)$ la population à l'instant t. Après un petit intervalle de temps Δt, la nouvelle population, $P(t+\Delta t)$, se compose de l'ancienne, $P(t)$, à laquelle il faut ajouter le nombre de naissances et soustraire le nombre de décès. Les naissances étant dues à ce qu'une partie de la population se soit reproduite durant la période Δt, elles sont proportionnelles à la population $P(t)$. Par exemple, si x % de la population a un enfant et y % a deux enfants, l'augmentation de la population due aux naissances est de $xP(t) + 2yP(t) = (x+2y)P(t)$. De même pour les décès : le nombre de personnes mortes durant la période Δt doit être proportionnelle à $P(t)$. On peut donc écrire :

$$P(t+\Delta t) = P(t) + (r\Delta t)P(t) \tag{I.15}$$

où $r\Delta t$ est le pourcentage de la population qui est né moins le pourcentage qui est mort durant la période de temps Δt. La constante r marque l'accroissement (positif ou négatif) de la population *par unité de temps*. On l'appelle le *taux de croissance* de la population. Les facteurs économiques et sociaux restant similaires, il semble raisonnable de supposer que r est constant au cours du temps. On peut réécrire (I.15) comme $\big(P(t+\Delta t) - P(t)\big)/\Delta t = rP(t)$. En passant à la limite $\Delta t \to 0$, on trouve que la population P doit satisfaire à l'équation différentielle

$$\partial_t P(t) = rP(t). \tag{I.16}$$

La solution de cette équation est $P(t) = P_0 \mathrm{e}^{rt}$ où P_0 est la population à l'instant $t = 0$.

- Sachant que la population des États-Unis était en 1950 de 151,3 millions et en 1960 de 179,3 millions, calculer le vitesse de croissance r.
- Prédire, grâce à ce modèle, quelle serait la population des États-Unis en l'an 2000.
- Expliquez pourquoi ce modèle n'est pas valable pour des périodes de temps trop longues.

Nous voudrions maintenant modifier ce modèle pour y inclure l'accroissement de la population dû à l'immigration. On sait que le taux d'immigration est resté relativement constant, à savoir de l'ordre de $m := 250000$ personnes par an — ceci est dû à un contrôle visant à éviter un afflux trop important d'immigrants. L'équation (I.15) s'en trouve modifiée comme suit : $P(t+\Delta t) = P(t) + rP(t)\Delta t + m\Delta t$. Comme précédemment, en passant à la limite $\Delta t \to 0$, on trouve une équation différentielle :

$$\partial_t P(t) = rP(t) + m.$$

La solution de cette équation est $P(t) = (P_0 + m/r)\mathrm{e}^{rt} - M/r$.

- Avec les données ci-dessus, déterminez le taux de croissance de la population r.
- On peut écrire la solution $P(t)$ comme $P_0 e^{rt} + (m/r)(e^{rt} - 1)$. Le premier terme indique la croissance de la population en l'absence d'immigration ; le second l'apport de l'immigration. Quel pourcentage de la population en 1960 est dû à l'immigration depuis 1950 ?

Exercice I.7 Soit $f : I \to \mathbb{R}$ où I est un intervalle de $\mathbb{R}$. Montrez que

$$\sup\left\{ \frac{\partial^2 f(\xi)}{2\partial f(\eta)} : \xi, \eta \in I \text{ et } \partial f(\eta) \neq 0 \right\} = \frac{\sup_I |\partial^2 f|}{2 \inf_I |\partial f|}$$

si $\inf_I |\partial f| > 0$.

Exercice I.8 Soit $f :]0, +\infty[\to \mathbb{R} : x \mapsto 4 \ln x - x$.

- Tracez le graphe de f.
- À partir de ce graphe ainsi qu'analytiquement, montrez que la méthode de Newton n'est pas bien définie sur $[e, 4]$ et que, sur $]0, e[$, toutes les suites convergent vers l'unique racine de f dans $]0, 4]$.
- Montrez que, pour tout $x_0 \in]4, +\infty[$, la suite de Newton est bien définie, reste dans $]4, +\infty[$ et converge vers la seule racine de f dans $]4, +\infty[$. (Indication : utilisez l'exercice I.3 sur les intervalles $[a, b]$ où $a > 4$ et proche de 4 et b est grand.)

Exercice I.9 On considère la fonction $f(x) = 4 + 8x^2 - x^4$.

- Combien f possède-t-il de racines ?
- Si on décide d'utiliser la méthode de bissection, quelles sont les paires de points initiaux qu'on peut choisir pour obtenir chacune des racines ?
- Si on opte pour la méthode de Newton, donnez des intervalles autour de chacune des racines sur lesquels la méthode de Newton converge.

Exercice I.10 Soit $f(x) := x^3 - 3x - 3$.

- Prouvez que f possède une seule racine $x^\star$ qui est strictement positive.
- Démontrez que, pour tout $x_0 \in]1, +\infty[$, la suite (x_n) définie par la méthode de Newton converge.
- Trouvez un intervalle autour de $x^\star$ sur lequel la convergence est quadratique.
- Montrez qu'il y a une infinité de points $(\alpha_k)_{k \geqslant 1} \subseteq]-\infty, -1]$ (resp. $(\beta_k)_{k \geqslant 1} \subseteq]-\infty, -1[$) tels que, si $x_0 \in \{\alpha_k : k \geqslant 1\}$, la suite de Newtown partant de x_0 est telle que $x_n = -1$ (resp. $x_n = 1$) pour un certain n. La suite (x_n) n'est donc pas bien définie après ce n (pourquoi ?).

- Soit $A(x^\star)$ tel que défini à l'exercice I.2. Écrivez un algorithme $\mathcal{A}$ qui prenne en entrée un nombre x et ait la propriété suivante : si $\mathcal{A}$ s'arrête sur l'entrée x, alors $x \in A(x^\star)$.
- Utilisez cet algorithme pour écrire un programme qui esquisse l'ensemble $A(x^\star)$. Comparez avec l'exercice I.8

Exercice I.11 Soit $A > 0$. On considère le schéma récursif suivant :

$$x_{n+1} = \frac{x_n(x_n^2 + 3a)}{3x_n^2 + a} =: \varphi(x_n), \quad n \geqslant 0.$$

- Montrez que les seuls points fixes de φ sont 0, $\sqrt{a}$ et $-\sqrt{a}$.
- Calculez la dérivée de φ en ses points fixes. Qu'en concluez vous sur l'attractivité ou la répulsivité locale de ces points ?
- Quel est l'ordre de convergence aux points fixes attractifs ? Comparez avec la méthode de Newton appliquée à la fonction $f(x) = x^2 - a$.
- Pour quelles valeurs de x_0 la suite (x_n) converge-t-elle vers $\sqrt{a}$? (Indication : examinez le signe de $x - \varphi(x)$ et la croissance/décroissance de φ.)

Exercice I.12 Comme nous l'avons dit au début de la section I.6, la méthode de Newton est un cas particulier de la méthode du point fixe avec $\varphi(x) = x - f(x)/\partial f(x)$.

- En appliquant le théorème I.12 à ce φ, quel critère trouve-t-on pour que φ possède un point fixe sur un intervalle I ?
- En utilisant ce critère pour la fonction $f(x) = x^2 - a$, quel voisinage de $\sqrt{a}$ trouve-t-on sur lequel la méthode de Newton converge ? Comparez avec celui donné par le théorème I.11. Quel intervalle est le plus grand ? Comment les conclusions des deux théorèmes se comparent-elles ?
- Quel est l'intervalle maximal autour de $\sqrt{a}$ sur lequel la méthode de Newton converge ? (Indication : voir l'exercice I.3.)

Exercice I.13 L'équation $x^3 + 4x^2 - 10 = 0$ peut se réécrire sous la forme d'un point fixe des trois façons suivantes :

$$x = \varphi_1(x) := \sqrt{\frac{10 - x^3}{4}}$$
$$x = \varphi_2(x) := \frac{10}{x^2 + 4x}$$
$$x = \varphi_3(x) := \sqrt{\frac{10}{x + 4}}$$

- Montrez que l'équation ci-dessus possède une unique racine (qui est positive) et donc que φ_i, $i = 1,2,3$, possèdent un seul point fixe.
- Calculez les dix premières itérées des suites (x_n) définies par $x_{n+1} = \varphi_i(x_n)$ et $x_0 = 1$ pour $i = 1,2,3$. Qu'en déduisez vous ?
- Tracez les graphes des fonctions φ_i. Comment les comportements observés ci-dessus se voient-ils sur ces graphiques ? Observez également la vitesse de convergence.

Exercice I.14 Soit l'équation $x^4 + 2x^2 - x - 3 = 0$.

- Vérifiez que cette équation peur se réécrire sous les quatres formes de type point fixe suivantes :

$$x = \varphi_1(x) := (3 + x - 2x^2)^{1/4}$$
$$x = \varphi_2(x) := \sqrt{\frac{x+3-x^4}{2}}$$
$$x = \varphi_3(x) := \sqrt{\frac{x+3}{x^2+2}}$$
$$x = \varphi_4(x) := \frac{3x^4+2x^2+3}{4x^3+4x-1}$$

- On définit les suites $(x_n^i)_{n\in\mathbb{N}}$ par $x_{n+1}^i = \varphi_i(x_n^i)$. Calculez les dix premières itérations pour chacun des φ_i, $i = 1,2,3,4$, avec $x_0^i = 1 + \varepsilon$ où $\varepsilon > 0$ est un petit nombre aléatoire.
- Que se passe-t-il (et pourquoi) dans les cas suivants : en partant de $x_0^2 = 2 + \varepsilon$ pour φ_2 ; en commencant à $x_0^3 = -3 - \varepsilon$ pour φ_3 ; en partant de $x_0^4 = 0,2367$ pour φ_4 ?

Exercice I.15 En mécanique céleste, le calcul des positions planétaires donne lieu à l'équation de Képler :

$$m = x - E\sin(x)$$

où nous allons considérer les valeurs $m = 0,8$ et $E = 0,2$. Utilisez la méthode du point fixe pour résoudre cette équation en partant des valeurs initiales $x_0 = 1$, 0 et -1 respectivement.

Exercice I.16 Considérons le système

$$\begin{cases} x^2 + y^2 - 1 = 0 \\ y - \mathrm{e}^{-x} = 0 \end{cases}$$

- Interprétez géométriquement ce système.
- Développez un algorithme pour trouver les deux solutions de ce système et prouvez sa convergence.

Exercice I.17 La loi des gazs parfaits s'écrit $pV = nRT$ où p est la pression, V le volume, n le nombre de moles, R la constante universelle des gazs et T la température absolue. Cependant cette loi a le double défaut de n'être valable que pour certains gaz et seulement dans une certaine plage de température et de pression. Une autre équation, connue sous le nom de VAN DER WALL, a donc été proposée :

$$\left(p + \frac{a}{v^2}\right)(v - b) = RT$$

où $v = V/n$ est le volume molaire et a, b des constantes empiriques qui dépendent du gaz. Dans le cadre d'un projet d'ingénierie chimique, on a besoin de connaître le volume molaire v du dioxide de carbonne sur une échelle de températures et de pressions afin de pouvoir construire un conteneur adéquat. On dispose des données suivantes :

$$R = 0,082054\,; \qquad a = 3,592\,; \qquad b = 0,04267.$$

Les pressions et températures auxquelles on va s'intéresser sont respectivement 1, 10 et 100 atm et 300, 500, 700 K.

- Calculez le volume molaire v selon la loi des gazs parfaits.
- Calculez le volume molaire v grâce à l'équation de Van der Wall. Quelle méthode allez-vous utiliser et de quelle approximation de la solution allez-vous partir ?

Exercice I.18 Des études montrent que, t heures après l'administration d'une quantité de A mg d'un médicament dans le sang, la concentration de ce dernier est de $c(t) := Ate^{-t/3}$ mg/ml. De plus, la concentration maximale non-dangereuse est de 1 mg/ml.

- Quelle quantité doit être injectée pour atteindre (sans dépasser) cette concentration maximale ? Quand celle-ci est-elle atteinte ?
- Une seconde injection de médicament sera faite au patient lorsque la concentration de celle ci-dessus sera tombée à 0,25 mg/ml. Déterminez, à la minute prêt, combien de temps après la première cette seconde injection doit être faite.

Exercice I.19 Soit $\varphi_\mu : \mathbb{R} \to \mathbb{R} : x \mapsto 1 - \mu x^2$.

- Montrez que φ_μ est un endomorphisme de $[-1,1]$ (i.e., $\varphi_\mu([-1,1]) \subseteq [-1,1]$) pour $\mu \in [0,2]$.
- Pourvez que $\varphi_\mu : [-1,1] \to [-1,1]$ possède un unique point fixe x_μ pour $\mu \in [0,2[$. Ce point fixe $x_\mu > 0$. Pour $\mu = 2$, φ_μ possède comme points fixes $x_\mu > 0$ et -1.
- *Programmation.* On voudrait connaître le comportement des suites (x_n) définies par $x_{n+1} = \varphi_\mu(x_n)$ lorsque n est grand : convergent-elles toutes vers le point fixe x_μ ? oscillent-elles ?... Essayons de le découvrir expérimentalement par un calcul sur ordinateur. Pour μ fixé, choisissons (disons) 20 points $x_0^{(1)}, \ldots, x_0^{(20)}$ dans $[-1,1]$ et

calculons $(x_n^{(i)})_{n\geqslant 0}$ défini par $x_{n+1}^{(i)} = \varphi_\mu(x_n^{(i)})$. Ignorons (disons) les 300 premières itérations le temps que la suite $(x_n^{(i)})_{n\geqslant 0}$ se « stabilise » — si elle doit le faire — et traçons les 4 suivantes : $x_{301}^{(i)}, \ldots, x_{304}^{(i)}$. Faisons cela pour un nombre suffisant[6] de $\mu \in [0,2]$ et reportons les résultats sur un graphique en (μ, x). On doit obtenir un graphe ressenblant à la figure I.1. Interprétez le graphique ainsi obtenu.

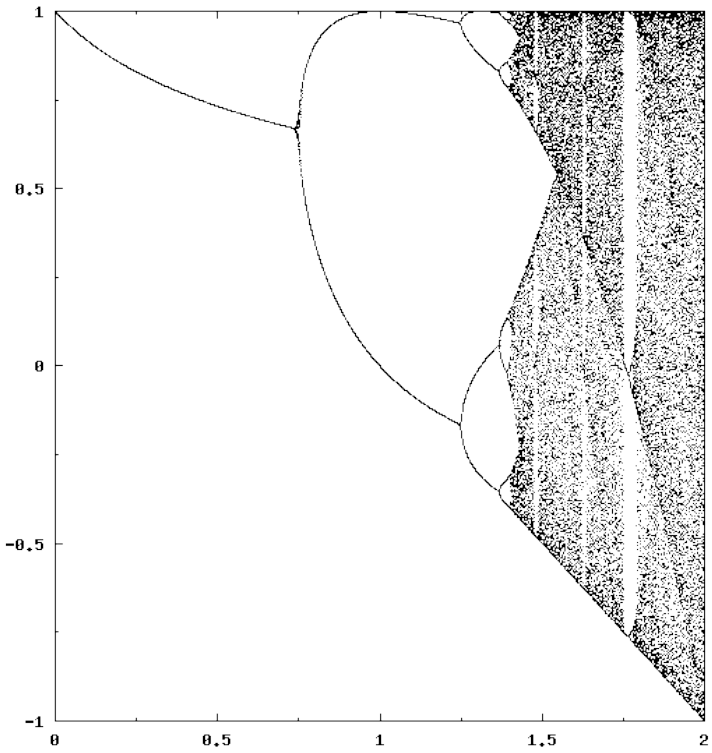

FIGURE I.1 – Diagramme de Feigenbaum

- Grâce aux argument développés dans la section I.6, montrez que la première bifurcation a lieu en $\mu = 3/4$.

Exercice I.20 Étudiez la présentation concernant la méthode de Newton à l'adresse « `http://www.math.uakron.edu/~dpstory/pdf_demos.html` »

6. Pour avoir la continuité des branches malgré un tracé constitué de points, nous avons pris $\mu \in 0,001\,\mathbb{N}$. Notons qu'il faut faire attention à ne pas prendre des paramètres trop grands ou trop petits (selon) car ceux choisis ici produisent déjà $20 \cdot 4 \cdot 2000 = 160000$ points.

Chapitre II

Arithmétique machine

Les ordinateurs, à cause de leur capacité de stockage et de leur vitesse limitées, ne sont capables de manipuler qu'un ensemble fini de nombres réels. Cela implique que les calculs qu'ils exécutent ne sont qu'une approximation des calculs « idéaux » faits sur $\mathbb{R}$. À priori, cela ne semble pas dramatique. Pour des problèmes pratiques, une approximation suffisamment bonne de la solution suffit. Et puis, si les données proviennent d'expériences, elles sont de toute façon déjà entachées d'une erreur.

La situation n'est malheureusement pas aussi idyllique. Des calculs exécutés naïvement peuvent donner des résultats grossièrement erronnés. Comment être sûr que les petites erreurs faites à chaque étape du calcul ne s'ajoutent les unes aux autres pour avoir finalement une influence non négligeable sur le résultat ? On le voit, l'analyse de la *propagation des erreurs* tout au long de l'exécution d'un algorithme est primordiale pour pouvoir avoir confiance dans la validité des réponses produites. D'autre part, vu que, comme on l'a dit ci-dessus, on calcule dans la plupart des cas sur des données approchées, on voudrait s'assurer qu'il n'en découle pas que les résultats en soient modifiés drastiquement. Il s'agit de la *sensibilité du problème* aux perturbations.

Le but de ce chapitre est d'introduire aux concepts et techniques de base de l'arithmétique machine. Nous commencerons par présenter une version quelque peu simplifiée de la représentation des nombres par un ordinateur connue sous le nom de « représentation en virgule flottante ». Grâce à ce modèle, nous introduirons la notion de « précision machine » et nous analyserons comment les erreurs se propagent à travers les opérations algébriques élémentaires. Nous généraliserons ensuite cette analyse au calcul d'une fonction quelconque, ce qui nous amènera à définir le « conditionnement » d'une fonction et d'un algorithme.

II.1 Exemples de problèmes

Avant d'expliquer les parades à la perte de précision — voire aux résultats « loufoques » — que les calculs sur ordinateur peuvent engendrer, nous voudrions aiguiser la perspicacité et l'intérêt du lecteur par divers exemples.

Commencons par le problème élémentaire qui consiste à calculer les racines de l'équation $ax^2+bx+c=0$. La solution est bien connue ; les deux racines sont :

$$x_1 = \frac{-b-\sqrt{b^2-4ac}}{2a}, \quad \text{et} \quad x_2 = \frac{-b+\sqrt{b^2-4ac}}{2a}.$$

Cela donne lieu immédiatement à l'implémentation :

```
void QuadraticSolve(double a, double b, double c,
                    double *x1, double *x2)
{
  double delta;

  delta = sqrt(b*b - 4 * a * c);
  *x1 = (-b - delta)/(2 * a);
  *x2 = (-b + delta)/(2 * a);
}
```

Cette routine est cependant loin d'être satisfaisante. Outre le fait qu'elle ne considère pas les cas $b^2-4ac<0$ et $a=0$, elle souffre d'une possible perte de précision concernant `x1`. Par exemple, si on exécute `QuadraticSolve(100000, -107374182400000.095367431640625, 102400000, &x1, &x2)`, on trouve les valeurs `x1` $= 0.0000009375$ et `x2` $= 1073741824$ alors que les racines exactes sont [1] $x_1 = 2^{-20} = 0.00000095367431640625$ et $x_2 = 2^{30} = 1073741824$. L'erreur commise sur x_1 est $|\texttt{x1} - x_1| = 1.617431640625 \cdot 10^{-8}$, soit de l'ordre de 1.7 % de x_1. On pourrait penser que c'est encore relativement raisonnable, mais cette erreur implique que la deuxième décimale significative de `x1` — et à fortiori les suivantes — est fausse. De plus, comme nous le verrons par la suite, on peut facilement éviter cette erreur en modifiant légèrement l'algorithme. □

Le second exemple provient de test faits pour déterminer la plausibilité d'une conjecture. La voici. On se donne trois matrices carrées $X,Y,Z \in \mathbb{R}^{N\times N}$ et on définit :

$$\begin{aligned} H_1(a) &:= \det\big((a\,\mathrm{id}-Y)+X(a\,\mathrm{id}-Z)\big) \\ H_2(b,c) &:= (b-c)\det\big((b\,\mathrm{id}-Y)(c\,\mathrm{id}-Y)+X(b\,\mathrm{id}-Z)(c\,\mathrm{id}-Z)\big) \\ H_3(a,b,c) &:= H_1(a)H_2(b,c)+H_1(b)H_2(c,a)+H_1(c)H_2(a,b) \end{aligned}$$

1. Tout nombre écrit de manière finie en base 2 s'écrit également de manière finie en base 10. Par exemple $2^{-20} = 5^{20}5^{-20}2^{-20} = 5^{20}10^{-20} = 95367431640625 \cdot 10^{-20}$.

La conjecture dit que la fonction H_3 est identiquement nulle lorsque $\operatorname{rang}(XZ-YX) \leqslant 1$. Comme c'est facile à prouver lorsque $\operatorname{rang}(XZ-YX) = 0$, essayons avec des matrices simples qui donnent un rang égal à 1 :

$$X = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & 1 & 2 \end{pmatrix}, \quad Y = 0, \quad Z = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Pour ces matrices, le résultat du calcul de $H_3(2,30,400)$ naïvement exécuté [2] donne -80. Cependant, si on fait ce calcul à la main, on trouve que $H_1(a) = -86a^2 + 40a^3$ et $H_2(b,c) = (-258c^3 + 86c^4)b^2 + (258c^2 - 40c^4)b^3 + (-86c^2 + 40c^3)b^4$, si bien que $H_3(a,b,c) = 0$ pour tout $a,b,c \in \mathbb{R}$. Ce qui est intéressant avec cet exemple est qu'aucune des données du problème n'est exceptionnelle et on pourrait dès lors être tenté de négliger la possibilité d'une erreur aussi grossière. □

Comme troisième exemple, cherchons à calculer la fonction $f :]-\infty, 1[\to \mathbb{R}$ définie par

$$f(x) := \begin{cases} \ln(1-x)/x & \text{si } x \neq 0, \\ -1 & \text{si } x = 0. \end{cases}$$

L'implémentation qui vient tout de suite à l'esprit consiste à définir la routine :

```
double f1(double x)
{
  if (x == 0)
    return -1.0;
  else
    return log(1-x)/x;
}
```

À première vue, celle-ci se comporte bien. Pour preuve, nous avons tracé à la figure II.1 le graphe de f sur $[-1,1[$ calculé grâce à cette procédure. Un zoom autour de $x = 0$ montre que ce n'est pas le cas. Comme on le voit sur la figure II.2, les valeurs données par `f1` peuvent être fort éloignées de $f(0) = -1$ et, alors que f est continue, le graphe fait des sauts de -2 à 0 ! La raison de ce phénomène est que, lorque y est proche de 1, $\ln(y)$ est très proche de $y-1$ et donc, lorsque x est proche de 0, $f(x)$ est proche de $((1-x)-1)/x$. Algébriquement, cette expression vaut -1. Ceci n'est pas vrai en arithmétique machine car l'addition machine n'est *pas associative*. En fait, si on trace le graphe de `((1-x)-1)/x` sur la figure II.2, il se superpose exactement à celui de f calculé par `f1`.

Est-il possible de calculer f plus précisément ? En fait oui. En l'occurrence, il suffit d'imiter la perte de précision du numérateur au dénominateur. Plus précisément, l'algorithme est :

2. On a simplement recopié les définitions de H_1, H_2, H_3 dans un logiciel du type Octave (`http://www.che.wisc.edu/octave/`) ou SciLab (`http://www-rocq.inria.fr/scilab/`) et effectué le calcul sur un Pentium.

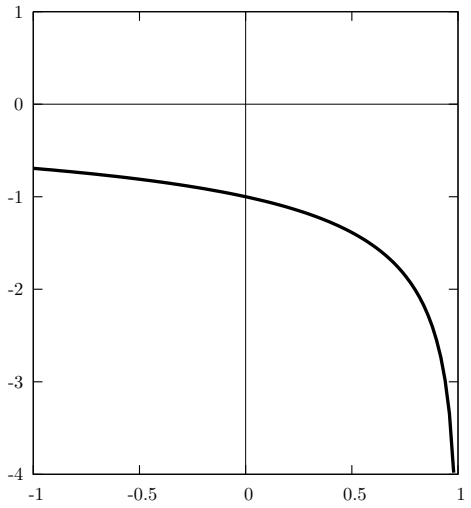

FIGURE II.1 – Graphe de f calculé avec `f1`.

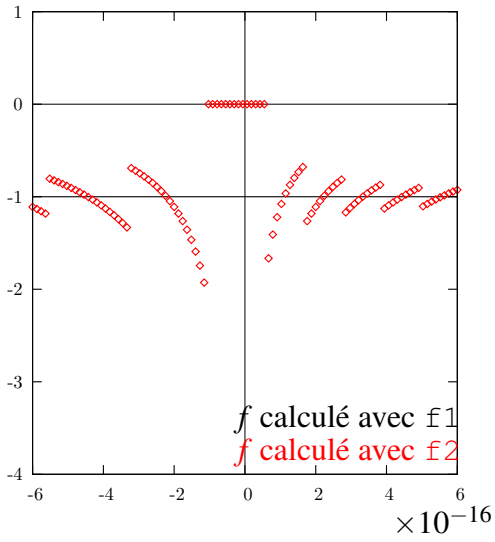

FIGURE II.2 – Graphe de différents calculs de $\ln(1-x)/x$.

```
double f2(double x)
{
  double y;

  y = 1 - x;
  if (y == 1)
    return -1.0;
  else
    return log(y)/(1-y);
}
```

Ainsi qu'on le voit sur la figure II.2, l'évaluation de f au voisinage de 0 par `f2` donne non seulement lieu à une courbe continue (du moins à cette échelle) mais, plus intéressant, plus proche de $f(0) = -1$ comme il se doit. Les techniques que nous verrons dans ce chapitre nous permettront de montrer que `f2` calcule f avec une meilleure précision. □

Clôturons cette série d'exemples par un problème fréquent : l'évaluation d'un polynôme. On se donne donc de coefficients $a_0, a_1, \ldots, a_n$ ainsi qu'une valeur du paramètre x et on désire calculer

$$p(x) := \sum_{i=0}^{n} a_i x^i.$$

Un algorithme efficace et couramment utilisé pour cette tâche est connu sous le nom de « méthode de Horner ». Il consiste à factoriser le polynôme sous la forme $((a_n x + a_{n-1})x + a_{n-2})x + \cdots$ et donne lieu à la procédure :

```
double evalp(double x, double a[], int degree)
{
  double y = a[degree--];

  for(; degree >= 0;  degree--)
    y = y * x + a[degree];
  return y;
}
```

Nous nous proposons d'utiliser cette méthode pour évaluer le polynôme unitaire p_n dont les racines sont $1, 2, \ldots, n$:

$$p_n(x) := \prod_{\nu=1}^{n} (x-\nu) = \sum_{i=0}^{n} a_i x^i.$$

Il est facile de calculer par récurrence [3] les coefficients $(a_i)_{i=0}^n$. Grâce à ceux-ci, nous pouvons évaluer p_n par la méthode d'Horner. Le résultat montré à la figure II.3 pour $n = 23$ (l'écriture `2e+15` est la notation « scientifique » pour le nombre $2\,10^{15}$,...). La

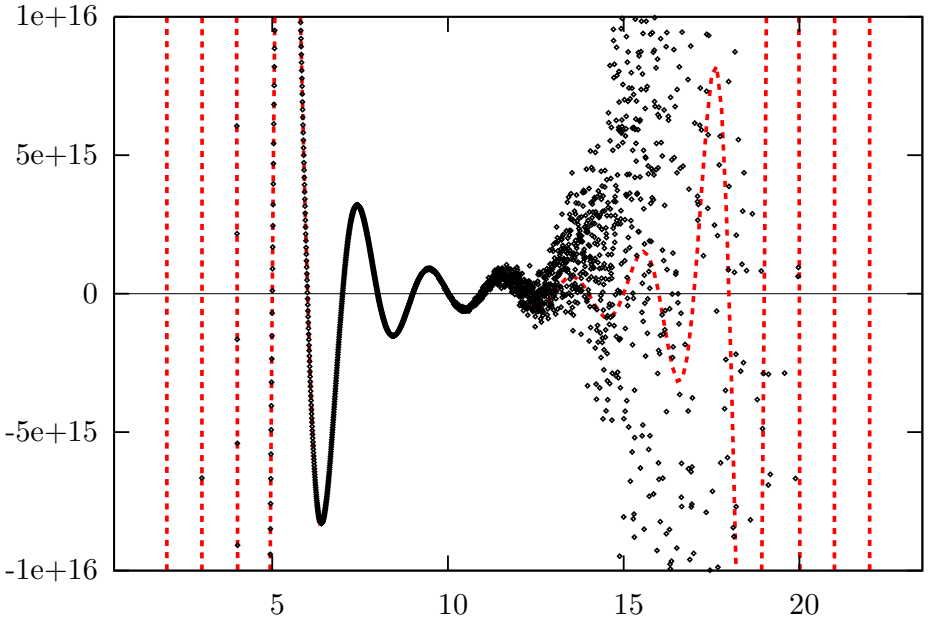

FIGURE II.3 – Différents algorithmes pour évaluer p_{23}.

courbe en rouge représente la valeur exacte [4] de p_n. On y a superposé le graphe de p_n calculé par Horner. On voit qu'à partir de $x = 10$, ce graphe s'étale en un nuage de points qui fini par s'éloigner sensiblement de la valeur exacte. Ainsi, même pour des tâches aussi simples que l'évaluation d'un polynôme, il convient de faire attention.

3. Écrivez et implémentez un algorithme qui rempli cette tâche.
4. Cela se calcule aisément en évaluant directement le produit $\prod_{\nu=1}^{n}(x-\nu)$.

II.2 Nombres en virgule flottante

De nombreuses manières ont été proposées pour représenter les nombres par un ordinateur mais la plus utilisée aujourd'hui est le représentation dite en « virgule flottante ». et en base 2. Cette représentation a été définie très précisément par la norme [5] IEEE-745 qui a été adoptée par la plupart des constructeurs de microprocesseurs. Nous ne ferons ici que décrire les éléments de base de cette norme avec l'espoir que ceci puisse servir de tremplin au lecteur désireux d'approfondir le sujet.

Au cœur de la plupart des micorprocesseurs actuels, les nombres sont représentés en base deux. Rappelons brièvement ce que cela veut dire. Comme vous le savez, la manière courante d'écrire les nombres aujourd'hui est la notation de position en base dix. On se donne donc dix symboles $0, 1, 2 := 1+1, 3 := 2+1, \ldots, 9 := 8+1$. La représentation d'un nombre entier est simplement une suite finie de tels symboles. Par exemple 27821 n'est que le symbole « 2 » suivit du symbole « 7 »,... Pour mettre en évidence l'aspect « suite de symboles », nous écrirons :

2	7	8	2	1

Pour savoir à quel nombre correspond cette suite de symboles, on les interprète selon leur position. Ici la base est $b = 10 := 9+1$. Cela signifie que chaque fois qu'on se déplace d'un digit vers la gauche, la puissance de dix augmente de 1 :

$2[10^4]$	$7[10^3]$	$8[10^2]$	$2[10^1]$	$1[10^0]$

Ainsi, le nombre représenté est $2b^4 + 7b^3 + 8b^2 + 2b^1 + 1b^0 = 2\,10^4 + 7\,10^3 + 8\,10^2 + 2\,10^1 + 1\,10^0$. Ces considérations peuvent paraître comme une lapalissade mais il faut réaliser que c'est grâce au génie de cette représentation que nous sommes capables de calculer aussi facilement aujourd'hui.

Ceci c'était pour les nombres entier. Qu'en est-il des écritures « avec virgule » ? Par exemple, que représente | 0 | , | 2 | ? Simplement, c'est $2/10 = 2\,10^{-1}$. Et | 1 | , | 7 | 4 | représente $1\,10^0 + 7\,10^{-1} + 4\,10^{-2}$. De manière générale,

$\pm$	a_n	$\ldots$	a_1	a_0	,	a_{-1}	a_{-2}	$\ldots$

représente le nombre $\pm(a_n\,10^n + \cdots + a_1\,10^1 + a_0\,10^0 + a_{-1}\,10^{-1} + a_{-2}\,10^{-2} + \cdots)$, c'est-à-dire

$$\pm \sum_{i=-\infty}^{n} a_i\,10^i.$$

5. Voir `http://babbage.cs.qc.edu/courses/cs341/IEEE-754references.html`.

Notons que la suite des décimales a_{-1}, a_{-2},...peut être finie ou infinie et que dans ce dernier cas la série converge. [6]

Représenter les nombres en base deux se fait exactement de la même manière qu'en base dix excepté qu'on remplace « dix » par « deux » ! Ainsi, on se donne deux symboles 0 et 1. À une suite de tels symboles

$$\boxed{\pm}\boxed{a_n}\boxed{\dots}\boxed{a_1}\boxed{a_0}\boxed{,a_{-1}}\boxed{a_{-2}}\boxed{\dots} \qquad a_i \in \{0,1\},$$

on fait correspondre le nombre [7]

$$\pm \sum_{i=-\infty}^{n} a_i 2^i = \pm\left(a_n 2^n + \cdots + a_1 2 + a_0 + a_{-1} 2^{-1} + \cdots\right)$$

Nous noterons ce nombre $(\pm a_n \dots a_1 a_0, a_{-1} a_{-2} \dots)_2$. On a donc $(110)_2 = 2^2 + 2^1 = 5$ et $(1,11)_2 = 2^0 + 2^{-1} + 2^{-2} = 1,75$.

II.3 Erreur absolue et relative

II.4 Arithmétique machine

II.5 Propagation des erreurs — nombre de conditionnement

6. Prouvez le ! Notez que $s_k := \sum_{i=-k}^{n} a_i 10^i$ est croissante et donc converge ssi elle est bornée.
7. Montrez à nouveau que même si la suite des décimales ne devient jamais nulle, le série converge.

II.6 Exercices

Exercice II.1 Donner la représentation binaire des nombres suivants.

- $19,25$;
- 121 ;
- 4100 ;
- $2^n + 2^k - 2$, $(n,k \in \mathbb{N})$.

Exercice II.2 Écrivez en décimal les nombres suivants.

- $(0{,}010101\ldots)_2$. En déduire la représentation de $2^n/3$ en binaire.
- $(0{,}001100110011\ldots)_2$.
- $s_n = 1 + 11 + 111 + \cdots + \underbrace{11\ldots1}_{n \text{ fois}}$ (les nombres doivent êtres lu en binaire).

Exercice II.3 On considère l'ensemble $\mathbb{R}(t,s)$ avec $t = 5$ et $s = 2$ ainsi que les nombres : $a = (111,010110001)_2$, $b = (0,01110011)_2$, $c = (0,000011)_2$, $d = (10011,011101)_2$.

- Calculer $\mathrm{chop}(x)$ et $\mathrm{fl}(x)$ pour $x \in \{a,b,c,d\}$
- Pour quelle valeur de x a-t-on le phénomène d'*underflow* (resp. d'*overflow*).

Exercice II.4 On considère la somme s_n définie pour $n \geqslant 1$ par :

$$s_n = 1 + \sum_{k=1}^{n} \frac{1}{k^2+k}.$$

(1)
- Déterminer a et b tels que $\dfrac{1}{k^2+k} = \dfrac{a}{k} + \dfrac{b}{k+1}$ $(k \geqslant 1)$.
- Montrer que $s_n = 2 - 1/(n+1)$.
- En déduire que : $s_{\underbrace{99\ldots99}_{k \text{ fois}}} = 1,\underbrace{99\ldots99}_{k \text{ fois}}$.

(2)
- Implémentez un programme en C qui calcule s_n selon l'algorithme suivant :

$$s_0 = 1, \qquad s_k = s_{k-1} + \frac{1}{k(k+1)}.$$

- Implémentez un programme en C qui calcule $s_n = s'_0$ selon l'algorithme suivant :

$$s'_n = \frac{1}{n(n+1)}, \qquad s'_{k-1} = s_k + \frac{1}{(k-1)k} \quad (1 \leqslant k < n), \qquad s'_0 = s'_1 + 1.$$

- Comparez les résultats obtenus par les deux programmes pour $n = 9$, $n = 99$, $n = 999$, $n = 9999$, $n = 99999$, $n = 999999$, $n = 9999999$, $n = 99999999$. Qu'en concluez vous ?

Exercice II.5 Écrivez un algorithme qui transforme la représentation décimale d'un nombre réel x en sa représentation binaire. Cet algorithme recevra comme données $x \in \mathbb{R}$ et le nombre $N \in \mathbb{N}$ de digits binaires désirés après la virgule. En sortie, il fournira un tableau de caractères $ba_n \dots a_0,a_{-1} \dots a_{-N}$ où $b \in \{+,-\}$ et $a_k \in \{0,1\}$ tel qu'il existe une suite $(a_k)_{k<-N} \subseteq \{0,1\}$ satisfaisant

$$x = b \sum_{k=-\infty}^{n} a_k$$

Exercice II.6 Soit $x \in \mathbb{R}$. On considère la fonction $\mathrm{fl} : \mathbb{R} \to \mathbb{R}(t,s) : x \mapsto \mathrm{fl}(x)$. Montrer que

$$\frac{|x - \mathrm{fl}(x)|}{|x|} \leqslant 2^{-t} = \mathrm{eps}$$

Exercice II.7 On considère l'ensemble $\mathbb{R}(t,s)$ et le nombre $x_n = 1 = 2^{-n} + 2^{-(n+1)}$. Écrivez explicitement en binaire (en indiquant le nombre de digits) les nombres suivants : x_n, $\mathrm{chop}(x_n)$ et $\mathrm{fl}(x_n)$ pour $n = t-1$, $n = t$ et $n = t+1$.

Exercice II.8 On considère l'algorithme suivant sur une machine qui utilise les nombres en virgule flottante appartenant à $\mathbb{R}(t,s)$:

$$\begin{cases} x := 1 \\ \text{Tant que } 1 + x > 1 \\ \qquad \text{faire } x := x/2 \end{cases}$$

- Ce programme s'arrête-t-il ?
- Si oui, après combien d'itérations et avec quelle valeur de x ?

Discutez deux cas : (1) l'ordinateur utilise $\mathrm{chop}()$ pour arrondir le résultat des opérations arithmétiques ; (2) il utilise $\mathrm{fl}()$.

Exercice II.9 On considère l'addition machine $\oplus$ définie comme suit : $\oplus : \mathbb{R}(t,s) \times \mathbb{R}(t,s) \to \mathbb{R}(t,s) : (x,y) \mapsto x \oplus y := \mathrm{fl}(x+y)$.

- Calculez $s_n = \bigoplus_{i=1}^{n} (i \oplus 2^{-t-1})$ et $s'_n = \sum_{i=1}^{n} (i + 2^{-t-1})$.
- Déterminez $n \in \mathbb{N}$ tel que $|s_n - s'_n| = 1$.

Exercice II.10 Soient $x, y \in \mathbb{R}$. On considère $\tilde{x} = (1+\varepsilon_x)x$, $\tilde{y} = (1+\varepsilon_y)y$ et on effectue l'opération arithmétique exacte sur $\tilde{x}$ et $\tilde{y}$ pour évaluer l'erreur sur le résultat en fonction de l'erreur sur les opérandes. Calculer $\varepsilon_{x\cdot y}$, ε_{x+y} et $\varepsilon_{x/y}$.

Exercice II.11 Au moyen d'un ordinateur qui utilise dix digits décimaux, on se propose de calculer $A = (a+b)+c$ et $B = a+(b+c)$ où $a = 1$, $b = 3\cdot 10^{-10}$ et $c = 3\cdot 10^{-10}$.

- Calculez A et B ;
- A-t-on $A = B$?

Exercice II.12 Au moyen d'un ordinateur qui utilise deux digits décimaux, on veut calculer $X = (a-b)^2$ et $Y = a^2 - 2ab + b^2$ avec $a = 1,8$ et $b = 1,7$.

- Calculer X et Y ?
- A-t-on $X = Y$?
- Quel est le signe de Y ? Commentez.

Exercice II.13 On considère l'équation $x^2 - 56x + 1 = 0$. On donne $\sqrt{783} = 27,982137\ldots$

- Calculez les deux racines x_1 et x_2 de cette équation arrondies (au plus proche) à 5 digits décimaux.
- Montrez que si x_1 et x_2 sont deux racines de l'équation $x^2 - px + q = 0$, alors $x_1 x_2 = q$.
- En déduire une autre façon de calculer x_1 et x_2.
- Comparez les résultats avec ceux obtenus au premier point.

Exercice II.14 On considère un ordinateur qui utilise trois digits décimaux. Calculer la précision machine eps. Soit $x_1 = 1,982$ et $x_2 = 1,984$

- Calculez la moyenne $m = (x_1 + x_2)/2$ en utilisant les opérations machine.
- Que constatez vous sur la façon dont les nombres x_1, x_2 et m sont ordonnés ?

Exercice II.15 La fonction $f(x,y) = x+y$ amplifie les erreurs relatives sur x et y quand $x \approx -y$. Dans plusieurs cas, on peut remédier à cette instabilité en remplaçant $x+y$ par une expression équivalente. Les calculs suivants risquent-ils de donner lieu à une perte de précision et, si oui, comment pouvez-vous y remédier ?

- $y = \sqrt{x+\delta} - \sqrt{x}$ pour $x > 0$ et $|\delta|$ petit.
- $y = \cos(x+\delta) - \cos(x)$, pour $|\delta|$ petit.
- $y = f(x+\delta) - f(x)$, où $|\delta|$ est petit et f est suffisamment dérivable.

Exercice II.16 Calculer le nombre de conditionnement des fonctions suivantes et déterminer les points éventuels où ces fonctions sont numériquement instables.

- $f(x) = x^{1/n}$ où $n > 0, n > 0$;
- $f(x) = x - \sqrt{x^2 - 1}$ où $x > 1$;
- $f(x) = \cos(x)$ où $|x| < \pi/2$;
- $f(x) = \arcsin(x)$ où $|x| < 1$;
- $f(x) = \arcsin\big(\frac{x}{\sqrt{1+x^2}}\big)$;
- $f(x) = x + 1$.

Exercice II.17 Soient f et g deux fonctions. On considère $h(t) := g \circ f(t) = g(f(t))$.

- Exprimer $\operatorname{cond} h$ en fonction de $\operatorname{cond} g$ et $\operatorname{cond} f$.
- Utilisez le point précédent pour calculer $\operatorname{cond} h$ où $h :]-\pi/2, \pi/2[\to \mathbb{R} : t \mapsto h(t) = (1+\sin t)/(1-\sin t)$.

Exercice II.18 Soient f et g deux fonctions définies sur un intervalle de $\mathbb{R}$.

- Calculez $(\operatorname{cond} f \cdot g)(x)$, $(\operatorname{cond} 1/g)(x)$ et $(\operatorname{cond} f/g)(x)$ en fonction de $\operatorname{cond} f$ et $\operatorname{cond} g$.
- Utilisez ces résultats pour déterminer $(\operatorname{cond}\operatorname{tg})(x)$, $(\operatorname{cond}\varphi)(x)$ et $(\operatorname{cond}\psi)(x)$ où $\varphi(x) = 1/(x^2 e^x)$ et $\psi(x) = (1-\cos x)/x$ $(x \neq 0)$.

Exercice II.19 On considère la fonction $f : x \mapsto x + 1$. $x > 0$. On se propose de calculer f en utilisant deux algorithmes :

$$A_1 : f_{A_1}(x) = x + 1 \,; \qquad A_2 : \begin{cases} \text{si } x \neq 1 & f_{A_2}(x) = (x^2-1)/(x-1) \,; \\ \text{sinon} & f_{A_2}(1) = 2. \end{cases}$$

Calculer $\operatorname{cond} A_1$ et $\operatorname{cond} A_2$ et comparer les deux algorithmes.

Exercice II.20 On considère l'équation $f_a(x) := x^n - x^{n-1} - a = 0$ où $n \geqslant 2$.

- Montrez, pour tout $a > O$, que cette équation possède exactement une racine (strictement) positive qu'on note $\xi(a)$.
- On a donc la fonction $\xi :]0, +\infty[\to]0, +\infty[: a \mapsto \xi(a)$ tel que $f_a\big(\xi(a)\big) = 0$. Calculez $(\operatorname{cond}\xi)(a)$. La fonction ξ est-elle bien conditionnée ? (Majorer $(\operatorname{cond}\xi)(a)$.)

Exercice II.21 Mêmes questions pour l'équation $f_a(x) := x^n + ax - 1 = 0$ où $a > 0$ et $n \geqslant 2$?

Exercice II.22 Mêmes questions pour l'équation $f_a(x) := xe^x - a = 0$ où $a \geqslant 0$.

Chapitre III

Interpolation polynomiale

Le but principal de l'interpolation est d'inférer une approximation de la valeur d'une fonction en un point où elle n'est pas connue à partir d'autres informations sur cette fonction (typiquement ses valeurs en d'autres points, des valeurs de la dérivée,...). Il y a deux composantes à cet objectif.

- Puis je, à partir des données dont je dispose, créer une fonction qui me permettrait d'inférer des valeurs en dehors du strict ensemble de données original ? L'efficacité de la méthode est importante et doit si possible s'accomoder facilement de la connaissance de données supplémentaires.
- Si je me donne une fonction, les fonctions calculées par interpolation sont-elles une bonne approximation de cette fonction de départ (et en quel sens) ?

On peut choisir de mettre plus l'accent sur l'une ou l'autre chose selon le problème à traiter.

Par exemple, pour un expérimentateur qui fait des mesures sur la dépendance d'un facteur y en fonction d'un paramètre x, la loi qui donne y en fonction f de x n'est pas nécessairement connue. tout ce dont il dispose est le résultat de mesures : quand je fixe le paramètre x à x_i, j'obtiens le résultat y_i. À partir de cela, il voudrait pouvoir parler de f en des valeurs de x autres que les x_i pour lesquels on a effectué une mesure.

On peut aussi adopter un autre point de vue. Si on veut développer une méthode numérique qui doit s'appliquer à une fonction « quelconque » f — par exemple une méthode d'intégration — on peut penser fixer les paramètres de la méthode sur des fonctions plus simples que f mais qui approximent bien f. Une manière de contruire de telles fonctions plus simples g est précisément d'extraire certaines données de f et de construire g par une méthode d'interpolation. Cela revient à « projeter » f sur l'ensemble des g plus simples. Clairement, ici, la précision avec laquelle g approxime f est importante ainsi que la manière dont la distance entre f et g est mesurée.

Dans ce chapitre, qui ne se veut qu'une brève introduction aux méthodes d'interpolation, nous nous limiterons pour les fonctions « plus simples » aux polynômes et nous ne verrons qu'une manière de construire la projection : celle qui à certains points $(x_i, f(x_i))$, $i = 1, \dots, k$, associe le polynôme qui passe par ces points.

C'est évidemment fort restrictif. On aurait pu considérer d'autres fonctions que les polynômes comme par exemple les fonctions linéaires par morceaux ou les splines. On aurait également pu regarder ce qui se passe si on dispose de plus d'informations comme par exemple les valeurs de la dérivée : $\partial f(x_i)$. D'autre part, on aurait pu partir de la propriété d'approximation : étant donné f, comment trouver la fonction g plus simple qui soit « la plus proche » de f. Cela conduit naturellement à l'approximation des moindres carrés, aux polynômes orthogonaux et aux séries de Fourier.

III.1 Existence et unicité

Comme nous l'avons déjà esquissé dans l'introduction, le problème consiste à trouver un polynôme qui passe par certains points donnés. Plus précisément, soient $x_0, \dots, x_k$ des points tous distincts de $\mathbb{R}$ et $y_0, \dots, y_k$ des nombres réels. On cherche un polynôme p tel que

$$\forall i = 0, \dots, k, \qquad p(x_i) = y_i \,; \tag{III.1}$$

Nous sommes non seulement intéressés par l'existence d'un tel polunôme mais aussi par le nombre de polynômes qui satisfont (III.1). Le polynôme p n'a en effet aucune raison d'être unique. Si on prend $y_0 = \cdots = y_k = 0$, (III.1) dit que $x_0, \dots, x_k$ sont des racines de p. Clairement, $p_0(x) := (x - x_0) \cdots (x - x_k)$ satisfait (III.1). Mais il y en a beaucoup d'autres ! en fait un polynôme p satisfait (III.1) avec les $y_i = 0$ si et seulement si $p(x) = q(x)(x - x_0) \cdots (x - x_k)$ pour un certain polynôme q (montrez cela !). De manière générale, si p est un polynôme qui satisfait (III.1), on peut lui ajouter n'importe quel polynôme qui s'annule en $x_0, \dots, x_k$ et la somme satisfait encore (III.1). On peut alors se demander si l'ensemble des solutions de (III.1) a une structure spéciale. Cet ensemble serait-il « engendré » par certains polynômes ? On l'a vu, on peut ajouter à p n'importe quel polynôme q qui s'annule en $x_0, \dots, x_k$ sans changer la validité de (III.1). Or un tel polynôme q a un degré plus grand ou égal à $k + 1$. Un polynôme de degré k est donné par ses $k + 1$ coefficients. Comme (III.1) est un système (linéaire) de $k + 1$ équations, on peut espérer qu'il existe un unique polynôme de degré k qui le satisfait. C'est ce que montre le théorème suivant.

Théorème III.1. *Soient $(x_0, y_0), \dots, (x_k, y_k)$ des points de $\mathbb{R}^2$ où les x_i sont tous distincts. Alors il existe un et un seul polynôme de degré k qui satisfait*

$$\forall i = 0, \dots, k, \qquad p(x_i) = y_i. \tag{III.2}$$

Ce polynôme p est donné par la formule de Lagrange :

$$p(x) = \sum_{i=0}^{k} \ell_i(x)\, y_i \qquad \text{où } \ell_i(x) := \prod_{\substack{0 \leqslant j \leqslant k \\ j \neq i}} \frac{x - x_j}{x_i - x_j}. \tag{III.3}$$

De plus, tout polynôme q qui satisfait (III.2) s'écrit comme $q(x) = p(x) + r(x)(x - x_0)\cdots(x - x_k)$ où $r(x)$ est un polynôme.

Démonstration. (1) *Existence de p.* Sur la formule (III.3), on voit que $\deg \ell_i = k$ et donc que $\deg p \leqslant k$. D'autre part, on vérifie facilement que

$$\ell_i(x_j) = \begin{cases} 0 & \text{si } i \neq j, \\ 1 & \text{si } i = j. \end{cases}$$

Un simple calcul montre alors que p satisfait (III.2).

(2) *Unicité de p.* Supposons que p' soit une autre solution de (III.2). On a alors que

$$\forall i = 0, \ldots, k, \qquad (p - p')(x_i) = 0.$$

Donc $p - p'$ est un polynôme de degré k qui possède $k+1$ racines. La seule possibilité est qu'il soit nul $p - p' = 0$.

(3) *Caractérisation des solutions de (III.2).* Soit q un polynôme qui vérifie (III.2). Alors, $q - p$ possède (au moins) les racines $x_0, \ldots, x_k$. Cela implique qu'on peut factoriser $q - p$ en $(x - x_0)\cdots(x - x_k)$ fois un polynôme quotient $r(x)$. □

III.2 Interpolation et approximation

Intéressons nous maintenant au cas où les points (x_i, y_i) proviennent d'une fonction. Soit $f \in \mathcal{C}([a,b];\mathbb{R})$ et $x_0, \ldots, x_k$ des points tous distincts de $[a,b]$. Le théorème précédent dit qu'il existe un unique polynôme de degré au plus k dont le graphe passe par les points $\big(x_i, f(x_i)\big)$, $i = 0, \ldots, k$. Nous noterons ce polynôme $p(f, x_0, \ldots, x_k; \bullet)$ ou, lorsque les points x_i sont clairement donnés par le contexte, $P_k f$. Ainsi $P_k f$ est l'unique polynôme de degré au plus k qui satisfait

$$\forall i = 0, \ldots, k, \qquad (P_k f)(x_i) = f(x_i).$$

On peut voir $f \mapsto P_k f$ comme une projection des fonctions sur les polynômes. Notons $\mathbb{P}_k$ l'ensemble des polynômes réels de degré au plus k.

Proposition III.2. *Soit $[a,b]$ in intervalle de $\mathbb{R}$ et $x_0, \ldots, x_k$ des points distincts de $[a,b]$. La fonction $P_k : \mathcal{C}([a,b];\mathbb{R}) \to \mathbb{P}_k : f \mapsto P_k f$ est une projection linéaire, c'est-à-dire*

(1) pour tout $f,g \in \mathcal{C}([a,b];\mathbb{R})$, $P_k(f+g) = P_k(f) + P_k(g)$;
(2) pour tout $f \in \mathcal{C}([a,b];\mathbb{R})$ *et tout* $\alpha \in \mathbb{R}$, $P_k(\alpha f) = \alpha P_k(f)$;
(3) pour tout $f \in \mathcal{C}([a,b];\mathbb{R})$, $P_k(P_k f) = P_k f$.

Pour la condition (3), on suppose implicitement qu'on voit les polynômes comme des cas particuliers de fonctions continues : $\mathbb{P}_k \subseteq \mathcal{C}([a,b];\mathbb{R})$. La condition (3) peut se réécrire comme suit : l'image de P_k dans $\mathcal{C}([a,b];\mathbb{R})$ est invariante sous P_k. En effet, une fonction $g \in \mathcal{C}([a,b];\mathbb{R})$ est dans l'image de P_k si et seulement si elle s'écrit $g = P_k f$ pour un certain f. Mais alors le fait que g soit invariant sous P_k, $P_k g = g$, revient à la condition (3).

Démonstration. (1) Soient $f,g \in \mathcal{C}([a,b];\mathbb{R})$. Il faut montrer que $P_k f + P_k g$ n'est rien d'autre que $P_k(f+g)$. Ce sera le cas si $P_k f + P_k g$ est un polynôme de degré au plus k qui satisfait la condition qui définit $P_k(f+g)$, à savoir

$$\forall i = 0,\dots,k, \qquad (P_k f + P_k g)(x_i) = (f+g)(x_i). \tag{III.4}$$

Or, par définition de $P_k f$ et $P_k g$, on a que $(P_k f)(x_i) = f(x_i)$ et $(P_k g)(x_i) = g(x_i)$. Il suffit d'additionner ces deux égalités pour trouver (III.4). De plus, $P_k f$ et $P_k g$ étant des polynômes de degré au plus k, il en va de même de $P_k f + P_k g$.

(2) La démonstration de la seconde propriété est tout à fait similaire à celle du point (1).

(3) Soit $f \in \mathcal{C}([a,b];\mathbb{R})$. Posons $g := P_k f \in \mathbb{P}_k$. Il faut prouver $P_k g = g$. Or prouver que g vaut $P_k g$ revient à dire que g est un polynôme de degré au plus k qui satisfait la condition qui définit $P_k g$, à savoir

$$\forall i = 0,\dots,k, \qquad g(x_i) = g(x_i).$$

C'est bien évidemment le cas ! □

Maintenant qu'à chaque fonction f on peut associer un polynôme $P_k f$, on pourrait se demander si ce polynôme est une bonne approximation de f. Après tout, $P_k f$ est f sont égaux aux points $x_0,\dots,x_k$ et tout le but de l'interpolation est d'utiliser cette information pour pouvoir parler de ce qui se passe en dehors de ces points ! Plus précisément, si on se donne une infinité de points $(x_i)_{i\in\mathbb{N}}$ tous distincts et $x \notin \{x_i : i \in \mathbb{N}\}$, a-t-on

$$(P_k f)(x) \xrightarrow[k\to\infty]{} f(x) \,?$$

Malheureusement ce n'est pas vrai pour tout f. Alors que l'ensemble des polunômes $\mathbb{P}$ est dense dans l'ensemble des fonctions continues, c'est-à-dire qu'il est toujours possible d'approximer [1] une fonction continue par un polynôme, les polynômes $P_k f$ ne sont pas toujours de bon candidats. [2] Ceci est à rapprocher du même fait concernant le

1. L'approximation dont on parle ici est plus forte que l'approximation en chaque point : il s'agit de l'approximation en norme uniforme. La densité de $\mathbb{P}$ dans $\mathcal{C}([a,b];\mathbb{R})$ s'exprime par $\forall f \in \mathcal{C}([a,b];\mathbb{R})$, $\forall \varepsilon > 0$, $\exists g \in \mathbb{P}$, $|f-g|_\infty \leqslant \varepsilon$ où $|f-g| := \max_{x\in[a,b]} |f(x)-g(x)|$.
2. On connaît des conditions sur f qui impliquent que $|f - P_k f|_\infty \to 0$ lorsque $k \to \infty$.

développement de Taylor. En vérité, on peut penser $P_k f$ comme un développement de Taylor « distribué » sur $k+1$ points différents. Comme nous le verrons par la suite (cf. exercice III.1), $p(f,x_0,\dots,x_k;x) \to \sum_{i=0}^{k} \frac{1}{i!}\partial^i f(x_0)(x-x_0)^i$ lorsque $x_1 \to x_0$, $x_2 \to x_0$,..., et $x_k \to x_0$. Par analogie avec le développement de Taylor, on peut se demander si, quand bien même on n'aurait pas $P_k f(x) \to f(x)$, on pourrait estimer la distance entre $P_k f(x)$ et $f(x)$. Nous recherchons donc un analogue au développement de Taylor avec reste. Le théorème suivant nous donne une telle formule.

Théorème III.3. *Soit $f \in \mathcal{C}^{k+1}(]a,b[;\mathbb{R})$ et $x_0,\dots,x_k$ des points distincts de $]a,b[$. Pour tout $x \in]a,b[$, il existe un $\xi \in]\min\{x_0,\dots,x_k,x\}, \max\{x_0,\dots,x_k,x\}[$ tel que*

$$f(x) = p(f,x_0,\dots,x_k;x) + \frac{\partial^{k+1} f(\xi)}{(k+1)!} \prod_{i=0}^{k} (x-x_i).$$

La preuve suivante est due à Cauchy.

Démonstration. On peut supposer que $x \notin \{x_0,\dots,x_k\}$ car sinon l'égalité est trivialement vraie par définition de $p(f,x_0,\dots,x_k;\bullet) = P_k f$. Définissons la fonction $F :]a,b[\to \mathbb{R}$ par la formule

$$F(t) = f(t) - P_k f(t) - \frac{f(x) - P_k f(x)}{\prod_{i=0}^{k}(x-x_i)} \prod_{i=0}^{k} (t-x_i).$$

Il est clair que $F \in \mathcal{C}^{k+1}(]a,b[;\mathbb{R})$. De plus F possède $k+2$ racines :

$$\forall i = 0,\dots,k, \quad F(x_i) = 0 \qquad \text{et} \qquad F(x) = 0.$$

Entre deux racines consécutives de F, F possède un maximum ou un minimum et donc ∂F s'annule. On a donc que

$$\partial F \text{ s'annule en au moins } k+1 \text{ points.}$$

En répétant cet argument, on déduit successivement que

$$\begin{aligned}
&\partial^2 F \text{ s'annule en au moins } k \text{ points ;}\\
&\partial^3 F \text{ s'annule en au moins } k-1 \text{ points ;}\\
&\qquad\qquad \vdots\\
&\partial^{k+1} F \text{ s'annule en au moins 1 point.}
\end{aligned}$$

Appelons ξ un zéro de $\partial^{k+1} F$. En tenant compte du fait que $P_k f$ est un polynôme de degré au plus k et que donc $\partial^{k+1} P_k f = 0$ et que $\prod_{i=0}^{k}(t-x_i)$ s'écrit comme t^{k+1} plus des termes de degré $\leqslant k$, on trouve

$$0 = \partial^{k+1} F(\xi) = \partial^{k+1} f(\xi) - \frac{f(x) - P_k(x)}{\prod_{i=0}^{k}(x-x_i)} (k+1)!.$$

Cette équation n'est rien d'autre que la thèse. □

III.3 Méthode de calcul

Écrire un programme qui calcule les polunômes interpolants semble fort aisé. Il suffit d'utiliser les formules (III.3). Celles-ci ont cepandant un incovénient majeur. Quel est-il ? Le plus souvent on voudra calculer les polynômes sur différents ensembles de points x_i et en particulier une situation fréquente est de rajouter un point x_{k+1} à un ensemble de points $x_0,\dots,x_k$. Par exemple, si les couples (x_i,y_i) proviennent d'expériences, on voudrait pouvoir facilement inclure le résultat d'une expérience supplémentaire. Or le problème des formules (III.3) est précisément qu'on ne peut déduire du polynôme interpolant sur les points $x_0,\dots,x_k$ celui sur les points $x_0,\dots,x_k,x_{k+1}$. Il faut recommencer tous les calculs !

Soit $f:[a,b]\to\mathbb{R}$ et $(x_i)_{i=0}^{k+1}$ des points de $\mathbb{R}$ tous distincts. Nous sommes intéressés à calculer $p(f,x_0,\dots,x_{k+1};\bullet)$ en utilisant $p(f,x_0,\dots,x_k;\bullet)$. On a certainement la formule suivante :

$$p(f,x_0,\dots,x_k,x_{k+1};x)=p(f,x_0,\dots,x_k;x)+a_{k+1}\prod_{i=0}^{k}(x-x_i) \tag{III.5}$$

pour un certain $a_{k+1}\in\mathbb{R}$. En effet, pour $x=x_i$, $0\leqslant i\leqslant k$, on a bien sûr l'égalité et pour $x=x_{k+1}$, il suffit de prendre (vérifiez le !) $a_{k+1}=\big(y_{k+1}-p(f,x_0,\dots,x_k;x_{k+1})\big)/\prod_{i=0}^{k}(x_{k+1}-x_i)$.

Afin de préciser les éléments dont dépend a_{k+1}, nous allons le noter [3] $f[x_0,\dots,x_{k+1}]$. En utilisant répétitivement (III.5), on a

$$\begin{aligned}p(f,x_0,\dots,x_{k+1};x)=f[x_0]+f[x_0,x_1]\,(x-x_0)+f[x_0,x_1,x_2]\,(x-x_0)(x-x_1)+\cdots\\ \cdots+f[x_0,\dots,x_{k+1}]\,(x-x_0)\cdots(x-x_k).\end{aligned} \tag{III.6}$$

La question est donc maintenant de trouver une procédure de calcul rapide des nombres $f[x_0,\dots,x_\ell]$. Cela va découler des relation ci-dessous.

Théorème III.4. *Soit $f:[a,b]\to\mathbb{R}$ et $(x_i)_{i=0}^{\infty}$ des points de $[a,b]$ tous distincts. On a*

$$\begin{aligned}f[x_0]&=f(x_0)\\ f[x_0,\dots,x_k,x_{k+1}]&=\frac{f[x_0,\dots,x_k]-f[x_1,\dots,x_{k+1}]}{x_0-x_{k+1}}.\end{aligned} \tag{III.7}$$

3. Pour le moment, même si les notations sont identiques, on ne peut dire qu'il s'agit des différences divisées. Cela sera établi au théorème III.4.

Démonstration. On va procéder en deux étapes. Tout d'abord on va montrer que dans $f[x_0,\dots,x_k]$, l'ordre des points $x_0,\dots,x_k$ n'est pas important. Ensuite, nous montrerons (III.7).

▪ *$f[x_0,\dots,x_k]$ est invariant sous permutation des points $x_0,\dots,x_k$.* Soit $x_{\pi(0)},\dots,x_{\pi(k)}$ une permutation des points $x_0,\dots,x_k$. Puisque $p(f,x_0,\dots,x_k;\bullet)$ et $p(f,x_{\pi(0)},\dots,x_{\pi(k)};\bullet)$ sont deux polynômes de degré au plus k qui prennent les valeurs $f(x_i)$ aux points x_i, $0\leqslant i\leqslant k$, ils sont forcément égaux :

$$p(f,x_0,\dots,x_k;x)=p(f,x_{\pi(0)},\dots,x_{\pi(k)};x).$$

En conséquence les coefficients du terme de degré le plus élévé x^k de ces deux polynômes sont égaux. Mais au vu de la formule III.6), ces coefficients sont $f[x_0,\dots,x_k]$ et $f[x_{\pi(0)},\dots,x_{\pi(k)}]$.

▪ $f[x_0]=f(x_0)$. En effet, par définition, $p(f,x_0;\bullet)=f[x_0]$ est le polynôme de degré 0 — donc constant — qui vaut $f(x_0)$ en x_0. Forcément $f[x_0]=p(f,x_0;x_0)=f(x_0)$.

▪ *Le seconde équation de (III.7) est vraie.* Pour faire apparaître $f[x_0,\dots,x_k]$ et $f[x_1,\dots,x_{k+1}]$, nous allons développer $p(f,x_0,\dots,x_{k+1};\bullet)$ de deux manières un peu différentes. Tout d'abord on a

$$\begin{aligned}
&p(f,x_0,\dots,x_{k+1};x)\\
&\quad=p(f,x_0,\dots,x_k;x)+f[x_0,\dots,x_{k+1}]\prod_{i=0}^{k}(x-x_i)\\
&\quad=p(f,x_1,\dots,x_k,x_0;x)+f[x_0,\dots,x_{k+1}]\prod_{i=0}^{k}(x-x_i)\\
&\quad=p(f,x_1,\dots,x_k;x)+f[x_1,\dots,x_k,x_0]\prod_{i=1}^{k}(x-x_i)+f[x_0,\dots,x_{k+1}]\prod_{i=0}^{k}(x-x_i)\\
&\quad=p(f,x_1,\dots,x_k;x)+f[x_0,\dots,x_k]\prod_{i=1}^{k}(x-x_i)+f[x_0,\dots,x_{k+1}]\prod_{i=0}^{k}(x-x_i)
\end{aligned}$$

D'autre part, la formule (III.5) implique aussi que

$$\begin{aligned}
&p(f,x_0,\dots,x_{k+1};x)=p(f,x_1,\dots,x_{k+1},x_0;x)\\
&\quad=p(f,x_1,\dots,x_{k+1};x)+f[x_1,\dots,x_{k+1},x_0]\prod_{i=1}^{k+1}(x-x_i)\\
&\quad=p(f,x_1,\dots,x_k;x)+f[x_1,\dots,x_{k+1}]\prod_{i=1}^{k}(x-x_i)+f[x_0,x_1,\dots,x_{k+1}]\prod_{i=1}^{k+1}(x-x_i)
\end{aligned}$$

En comparant ces deux développements, on déduit que

$$f[x_0,\ldots,x_k]\prod_{i=1}^{k}(x-x_i)+f[x_0,\ldots,x_{k+1}]\prod_{i=0}^{k}(x-x_i)$$
$$=f[x_1,\ldots,x_{k+1}]\prod_{i=1}^{k}(x-x_i)+f[x_0,x_1,\ldots,x_{k+1}]\prod_{i=1}^{k+1}(x-x_i)$$

Comme $\prod_{i=1}^{k}(x-x_i)$ est un facteur commun à tous ces termes, on trouve, après simplification,

$$f[x_0,\ldots,x_k]+f[x_0,\ldots,x_{k+1}]\,(x-x_0)$$
$$=f[x_1,\ldots,x_{k+1}]+f[x_0,x_1,\ldots,x_{k+1}]\,(x-x_{k+1})$$

En regroupant les termes comprenant $f[x_0,x_1,\ldots,x_{k+1}]$, on trouve la formule désirée.

□

Les formules (III.7) s'exploitent comme suit. Calculer le polynôme interpolant sur un seul point est très facile : $p(f,x_0;x)=f[x_0]=f(x_0)$. Si maintenant on ajoute le point x_1, on a $p(f,x_0,x_1;x)=f[x_0]+f[x_0,x_1]\,(x-x_0)$ et il est possible de calculer $f[x_0,x_1]$ en fonction de $f[x_0]$ et $f[x_1]$ par (III.7). Si on ajoute un troisième point x_2, on a $p(f,x_0,x_1,x_2;x)=f[x_0]+f[x_0,x_1]\,(x-x_0)+f[x_0,x_1,x_2]\,(x-x_0)(x-x_1)$ et il faut calculer $f[x_0,x_1,x_2]$. En vertu de (III.7), cette quantité s'exprime en fonction de $f[x_0,x_1]$ — qu'on a déjà calculé — et de $f[x_1,x_2]$ qui à son tour s'exprime en fonction de $f[x_1]$ — déjà calculé — et $f[x_2]=f(x_2)$. On peut mettre en évidence ces relations par le tableau suivant où les flèches indiquent les dépendances.

$$\begin{array}{c|ccccc}
x_0 & f(x_0)=f[x_0] & \searrow & & & \\
x_1 & f(x_1)=f[x_1] & \longrightarrow & f[x_0,x_1] & \searrow & \\
x_2 & f(x_2)=f[x_2] & \longrightarrow & f[x_1,x_2] & \longrightarrow & f[x_0,x_1,x_2] \\
\vdots & \vdots & & \vdots & & \vdots
\end{array}$$

Comme on le voit, calculer le polynôme interpolant pour un point x_{k+1} supplémantaire ne requiert que le calcul d'une nouvelle ligne de $k+1$ éléments.

III.4 Exercices

Exercice III.1 En utilisant la formulation des polynômes interpolants en termes de différences divisées (équation (III.6) et le théorème B.4, prouvez que $p(f,x_0,\ldots,x_k;x)\to\sum_{i=0}^{k}\frac{1}{i!}\partial^i f(x_0)(x-x_0)^i$ lorsque $x_1\to x_0$, $x_2\to x_0$,..., et $x_k\to x_0$.

Annexe A

Fonctions convexes et concaves

Le but de cet appendice est de présenter les définitions de fonction convexe, strictement convexe, concave et strictement concave telles qu'elles sont utilisées dans ce cours ainsi que quelques propriétés de base de celles-ci.

A.1 Fonctions convexes

Définition A.1. Une fonction $f : [a,b] \to \mathbb{R}$ est dite *convexe* si, pour tout $x_1, x_2 \in [a,b]$ et pour tout $\lambda \in [0,1]$, on a

$$f\big(\lambda x_1 + (1-\lambda)x_2\big) \leqslant \lambda f(x_1) + (1-\lambda) f(x_2). \tag{A.1}$$

Un point de la forme $\lambda x_1 + (1-\lambda)x_2$ s'appelle une *combinaison convexe* de x_1 et x_2. Vu que $\lambda x_1 + (1-\lambda)x_2 = x_1 + (1-\lambda)(x_2 - x_1)$, il est facile de voir que $\{\lambda x_1 + (1-\lambda)x_2 : \lambda \in [0,1]\} = [x_1, x_2]$. D'un point de vue géométrique, (A.1) exprime que, sur l'intervalle $[x_1, x_2]$, le graphe de f se trouve en dessous du segment de droite qui joint les points $(x_1, f(x_1))$ et $(x_2, f(x_2))$. En effet, le point du graphe de f correspondant à l'abscisse $\lambda x_1 + (1-\lambda)x_2 \in [x_1, x_2]$ s'écrit

$$\big(\lambda x_1 + (1-\lambda)x_2, f(\lambda x_1 + (1-\lambda)x_2)\big).$$

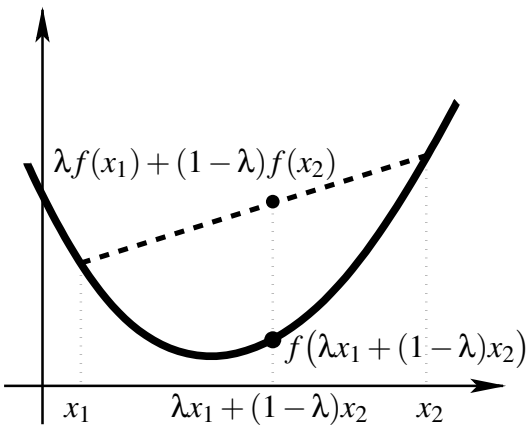

FIGURE A.1 – Graphe d'une fonction convexe

D'autre part, un vecteur directeur du segment de droite joignant les points $(x_1, f(x_1))$ et $(x_2, f(x_2))$ est $(x_2, f(x_2)) - (x_1, f(x_1))$. On peut écrire le segment de droite comme le point $(x_1, f(x_1))$ auquel on ajoute une « partie » du vecteur directeur, disons $t\big(x_2 - x_1, f(x_2) - f(x_1)\big)$ pour $t \in [0,1]$. C'est-à-dire, le segment a pour équation paramétrique

$$\big(x_1, f(x_1)\big) + t\big(x_2 - x_1, f(x_2) - f(x_1)\big)$$

Le point de ce segment au dessus de l'abscisse $\lambda x_1 + (1-\lambda)x_2 = x_1 + (1-\lambda)(x_2 - x_1)$ est celui pour lequel $t = 1-\lambda$. Il a donc pour coordonnées

$$\begin{aligned}&\big(x_1 + (1-\lambda)(x_2 - x_1), f(x_1) + (1-\lambda)(f(x_2) - f(x_1))\big)\\ &\quad = \big(\lambda x_1 + (1-\lambda)x_2, \lambda f(x_1) + (1-\lambda) f(x_2)\big)\end{aligned}$$

On voit donc que, ce que (A.1) dit, c'est que ce point se situe au dessus du point de même abscisse du graphe de f.

Les fonctions convexes d'une variable réelle ont la propriété intéressante d'être continues. Plus précisément, on a :

Proposition A.2. *Si $f : [a,b] \to \mathbb{R}$ est une fonction convexe, alors f est continue sur $]a,b[$.*

Démonstration. Fixons $x \in]a,b[$. Il existe un $\varepsilon_0 > 0$ tel que $[x - \varepsilon_0, x + \varepsilon_0] \subseteq]a,b[$. Notons $\underline{\ell}_+ \leqslant \overline{\ell}_+$ les limites inférieures et supérieures à droite de x, à savoir [1] :

$$\underline{\ell}_+ := \liminf_{h \xrightarrow{>} 0} f(x+h), \qquad \overline{\ell}_+ := \limsup_{h \xrightarrow{>} 0} f(x+h).$$

De même, pour les limites à gauche de x :

$$\underline{\ell}_- := \liminf_{h \xrightarrow{<} 0} f(x+h), \qquad \overline{\ell}_- := \limsup_{h \xrightarrow{<} 0} f(x+h).$$

La fonction f sera continue au point x si nous montrons que [2]

$$\underline{\ell}_- = \overline{\ell}_- = f(x) = \underline{\ell}_+ = \overline{\ell}_+.$$

1. Pour rappel, les définition de limite inférieure et supérieure sont respectivement : $\liminf_{h \xrightarrow{>} 0} f(x+h) = \sup_{\varepsilon > 0} \inf_{0 < h \leqslant \varepsilon} f(x+h)$ et $\limsup_{h \xrightarrow{>} 0} f(x+h) = \inf_{\varepsilon > 0} \sup_{0 < h \leqslant \varepsilon} f(x+h)$.

2. Prouvez que, si $\liminf_{h \xrightarrow{>} 0} f(x+h) = \limsup_{h \xrightarrow{>} 0} f(x+h) = \ell$, alors $\lim_{h \xrightarrow{>} 0} f(x+h)$ existe et vaut ℓ. On a la même chose lorsque $h \xrightarrow{<} 0$.

Nous allons tout d'abord montrer que $\overline{\ell}_+ \leqslant f(x)$. Soit $\varepsilon \in]0,\varepsilon_0[$. Puisque $x+h = x+(h/\varepsilon_0)\varepsilon_0$ est une combinaison convexe de x et $x+\varepsilon_0$, (A.1) nous permet d'écrire

$$\begin{aligned} f(x+h) = f\big(x+\frac{h}{\varepsilon_0}\varepsilon_0\big) &= f\big((1-\frac{h}{\varepsilon_0})x+\frac{h}{\varepsilon_0}(x+\varepsilon_0)\big) \\ &\leqslant (1-\frac{h}{\varepsilon_0})f(x)+\frac{h}{\varepsilon_0}f(x+\varepsilon_0) \\ &\leqslant f(x)+\frac{h}{\varepsilon_0}\big(f(x+\varepsilon_0)-f(x)\big) \end{aligned}$$

Prenant le suprémum sur $h \in]0,\varepsilon]$, on a

$$\sup_{0<h\leqslant\varepsilon} f(x+h) \leqslant f(x)+\frac{\varepsilon}{\varepsilon_0}\big|f(x+\varepsilon_0)-f(x)\big|.$$

En passant à la limite $\varepsilon \to 0$, on obtient comme voulu que $\overline{\ell}_+ \leqslant f(x)$. De la même manière, on montre que $\overline{\ell}_- \leqslant f(x)$.

Nous allons maintenant montrer que $f(x) \leqslant \underline{\ell}_+$ — et similairement que $f(x) \leqslant \underline{\ell}_-$ — ce qui concluera la démontration. Soit de nouveau $\varepsilon \in]0,\varepsilon_0[$ et $h \in]0,\varepsilon]$. Vu que $x = \frac{1}{2}(x+h)+\frac{1}{2}(x-h)$, (A.1) implique que

$$f(x) \leqslant \tfrac{1}{2}f(x+h)+\tfrac{1}{2}f(x-h).$$

En prenant d'abord le suprémum du terme de droite de l'addition puis l'infimum sur h du membre de droite de l'inégalité, on trouve [3] :

$$f(x) \leqslant \tfrac{1}{2}\inf_{0<h\leqslant\varepsilon} f(x+h)+\tfrac{1}{2}\sup_{0<h\leqslant\varepsilon} f(x-h)$$

Puisque c'est vrai pour tout ε petit, on peut passer à la limite $\varepsilon \to 0$, ce qui donne $f(x) \leqslant \frac{1}{2}\underline{\ell}_+ + \frac{1}{2}\overline{\ell}_-$. Or, comme on a prouvé que $\overline{\ell}_- \leqslant f(x)$, cette inégalité implique que $f(x) \leqslant \underline{\ell}_+$. □

Remarque A.3. On ne peut conclure que f est continue sur $[a,b]$. En fait on peut juste dire que les limites de f lorsque $x \xrightarrow{>} a$ et $x \xrightarrow{<} b$ existent et

$$\lim_{x\xrightarrow{>}a} f(x) \leqslant f(a) \quad \text{et} \quad \lim_{x\xrightarrow{<}b} f(x) \leqslant f(b)$$

Esquissons la preuve pour a. Soient $\underline{\ell}_+ \leqslant \overline{\ell}_+$ définis comme ci-dessus avec $x = a$. Par définition des limites sup et inf, il existe deux suites de nombres strictement positifs $(\underline{h}_n)$ et $(\overline{h}_n)$ convergeant vers 0 telles que $f(a+\underline{h}_n) \to \underline{\ell}_+$ et $f(a+\overline{h}_n) \to \overline{\ell}_+$. En choisissant bien des sous-suites, qu'on notera encore $(\underline{h}_n)$ et $(\overline{h}_n)$, on peut supposer que

3. Faites les détails !

$\underline{h}_{n+1} \leqslant \overline{h}_n \leqslant \underline{h}_n$ pour tout n. En conséquence $\overline{h}_n = \lambda_n \underline{h}_{n+1} + (1-\lambda_n)\underline{h}_n$ pour un certain $\lambda_n \in [0,1]$. Puisque $[0,1]$ est compact, on peut supposer qu'à une sous-suite près, encore notée λ_n, on a $\lambda_n \to \lambda$. La convexité de f implique que

$$f(x+\overline{h}_n) \leqslant \lambda_n f(x+\underline{h}_{n+1}) + (1-\lambda_n) f(x+\underline{h}_n).$$

En passant à la limite $n \to \infty$, on trouve

$$\overline{\ell}_+ \leqslant \lambda \underline{\ell}_+ + (1-\lambda)\underline{\ell}_+ = \underline{\ell}_+.$$

Donc $\overline{\ell}_+ = \underline{\ell}_+$ et la limite de $f(x)$ quand $x \xrightarrow{>} a$ existe. L'inégalité $\lim_{x \xrightarrow{>} a} f(x) \leqslant f(a)$ se montre grâce au même type d'argument que celui de la première partie de la démonstration de la proposition (A.2). □

On ne peut montrer plus. En effet ; la fonction $f : [0,1] \to \mathbb{R}$ définie par

$$f(x) = \begin{cases} 1 & \text{si } x = 0 \text{ ou } x = 1 \\ 0 & \text{si } x \in \,]0,1[\end{cases}$$

est bien convexe mais n'est continue ni en 0 ni en 1. □

Nous allons maintenant voir d'autres définitions équivalentes de la notion de convexité lorsque f jouit de plus de régularité.

Proposition A.4. *Soit $f : [a,b] \to \mathbb{R}$ une fonction continue. f est convexe sur $[a,b]$ si et seulement si*

$$f\big(\tfrac{1}{2}(x+y)\big) \leqslant \tfrac{1}{2}\big(f(x)+f(y)\big). \tag{A.2}$$

Démonstration. La nécessité de (A.2) est évidente. Reste à prouver que c'est suffisant pour que f soit convexe. Procédons par contradiction. Si f n'est pas convexe, il existe deux points $x, y \in [a,b]$ et $\lambda \in \,]0,1[$ tels que

$$f\big(\lambda x + (1-\lambda)y\big) > \lambda f(x) + (1-\lambda) f(y). \tag{A.3}$$

On peut supposer sans perte de généralité que $x = 0$, $y = 1$ et $f(x) = 0 = f(1)$. Plus précisément, considérons $\bar{f} : [0,1] \to \mathbb{R}$ définie par

$$\bar{f}(t) := f\big(tx + (1-t)y\big) - t f(x) - (1-t) f(y).$$

Puisque f satisfait (A.2), c'est aussi le cas de $\bar{f}$. En effet, évaluer $\bar{f}$ en $\frac{1}{2}(t+s)$ revient à évaluer f en la moyenne arithmétique de $tx+(1-t)y$ et $sx+(1-s)y$. En termes de $\bar{f}$, (A.3) devient $\bar{f}(\lambda) > 0$. Puisque f est continue, $\bar{f}$ l'est aussi et le maximum de $\bar{f}$ sur le compact $[0,1]$ existe : $0 < \sup_{[0,1]} \bar{f} = \bar{f}(t_0)$ pour un certain $t_0 \in [0,1]$. En tenant compte que $\bar{f}(0) = 0 = \bar{f}(1)$, on a que $t_0 \in \,]0,1[$. On peut donc choisir $\delta \in \big]0, \min\{t_0, 1-t_0\}\big[$ si bien que $[t_0-\delta, t_0+\delta] \subseteq [0,1]$. En utilisant (A.2) aux points $t_0-\delta$ et $t_0+\delta$, on trouve $\bar{f}(t_0) \leqslant \frac{1}{2}\bar{f}(t_0-\delta) + \frac{1}{2}\bar{f}(t_0+\delta)$. Comme $t_0-\delta = 0$ ou $t_0+\delta = 1$, il découle que $\bar{f}(t_0-\delta) = 0$ ou $\bar{f}(t_0+\delta) = 0$. Étant donné cela et le fait que $\bar{f}(t_0)$ est le maximum, on en conclut que $0 < \bar{f}(t_0) \leqslant \frac{1}{2} f(t_0 \pm \delta) \leqslant \frac{1}{2}\bar{f}(t_0)$ ce qui est une contradiction. □

Proposition A.5. *Soit*[4] $f \in \mathcal{C}^1(]a,b[;\mathbb{R}) \cap \mathcal{C}([a,b];\mathbb{R})$. *Les propositions suivantes sont équivalentes :*

(1) f est convexe ;

(2) pour tout $x_0 \in]a,b[$ et tout $x \in [a,b]$, on a

$$f(x) \geqslant f(x_0) + \partial f(x_0)(x - x_0) ;$$

(3) la dérivée de f, ∂f, est croissante sur $]a,b[$.

Démonstration. (1) $\Rightarrow$ (2) Soit $x_0 \in]a,b[$ et $x \in [a,b]$. On peut écrire (A.1) comme

$$f\big(x_0 + (1-\lambda)(x-x_0)\big) \leqslant f(x_0) + (1-\lambda)\big(f(x) - f(x_0)\big)$$

ou encore

$$\frac{f\big(x_0 + (1-\lambda)(x-x_0)\big) - f(x_0)}{(1-\lambda)(x-x_0)}\,(x-x_0) \leqslant f(x) - f(x_0)$$

Comme c'est vrai pour tout $\lambda \in [0,1]$, on peut passer à la limite $\lambda \xrightarrow{<} 1$, ce qui donne

$$\partial f(x_0)(x-x_0) \leqslant f(x) - f(x_0).$$

L'implication est montrée.

(2) $\Rightarrow$ (3) Soient $x_1 < x_2$ dans $]a,b[$. On veut montrer que $\partial f(x_1) \leqslant \partial f(x_2)$. En utilisant (2) avec $x_0 := x_1$ et $x := x_2$ et inversément, on trouve

$$\begin{aligned} f(x_2) &\geqslant f(x_1) + \partial f(x_1)(x_2 - x_1) \\ f(x_1) &\geqslant f(x_2) + \partial f(x_2)(x_1 - x_2) \end{aligned}$$

En tenant compte du fait que $x_2 - x_1 > 0$, on en déduit

$$\partial f(x_1) \leqslant \frac{f(x_2) - f(x_1)}{x_2 - x_1} \leqslant \partial f(x_2).$$

C'est ce qu'il fallait.

(3) $\Rightarrow$ (1) Soient $x_1 < x_2$ dans $[a,b]$. Comme f est continue sur $[x_1,x_2]$ et $\mathcal{C}^1$ sur $]x_1,x_2[$, on peut écrire

$$f\big(x_1 + (1-\lambda)x_2\big) = f(x_1) + \int_{x_1}^{x_1 + (1-\lambda)(x_2 - x_1)} \partial f(s)\,\mathrm{d}s.$$

4. Cela signifie que $f : [a,b] \to \mathbb{R}$ est continue, qu'elle est dérivable en chaque point $x \in]a,b[$ et que $\partial f :]a,b[\to \mathbb{R} : x \mapsto \partial f(x)$ est continue.

En faisant le changement de variable $s = x_1 + (1-\lambda)(t-x_1)$, on obtient

$$f\big(x_1+(1-\lambda)x_2\big) = f(x_1) + \int_{x_1}^{x_2} \partial f\big(\lambda x_1 + (1-\lambda)t\big)(1-\lambda)\,\mathrm{d}t.$$

Maintenant, puisque $\lambda x_1 + (1-\lambda)t \leqslant t$ — car $x_1 \leqslant t$ — et que ∂f est croissante, on déduit

$$\begin{aligned} f\big(\lambda x_1 + (1-\lambda)x_2\big) &\leqslant f(x_1) + (1-\lambda)\int_{x_1}^{x_2} \partial f(t)\,\mathrm{d}t \\ &= f(x_1) + (1-\lambda)\big(f(x_2) - f(x_1)\big) \end{aligned}$$

Ceci conclut la démonstration. □

Si f possède encore plus de régularité, on a l'équivalence suivante.

Proposition A.6. *Si $f \in \mathcal{C}^2(]a,b[;\mathbb{R}) \cap \mathcal{C}([a,b];\mathbb{R})$. Les propositions suivantes sont équivalentes :*

(1) f est convexe ;

(2) $\partial^2 f(x) \geqslant 0$ pour tout $x \in]a,b[$.

Démonstration. Au vu de la proposition A.5, il suffit de montrer que ∂f est croissante si est seulement si $\partial^2 f \geqslant 0$. C'est bien connu mais nous allons le redémontrer dans le but d'être complet.

Tout d'abord, supposons que ∂f est croisssante. Soit $x \in]a,b[$. Si $h > 0$ est assez petit pour que $x+h \in]a,b[$, on a $\partial f(x+h) \geqslant \partial f(x)$ et donc

$$\frac{\partial f(x+h) - \partial f(x)}{h} \geqslant 0.$$

En passant à la limite $h \xrightarrow{>} 0$, on trouve bien $\partial^2 f(x) \geqslant 0$.

Inversément, supposons que $\partial^2 f \geqslant 0$. Soient $x_1 < x_2$ deux points de $]a,b[$. Nous voulons montrer que $\partial f(x_1) \leqslant \partial f(x_2)$. Le théorème de la moyenne nous dit que

$$\partial f(x_2) - \partial f(x_1) = \partial^2 f(\xi)(x_2 - x_1) \tag{A.4}$$

pour un $\xi \in]x_1,x_2[$. Comme $\partial^2 f(\xi) \geqslant 0$ et $x_2 - x_1 > 0$, on a bien $\partial f(x_2) - \partial f(x_1) \geqslant 0$, ce qui conclut la démonstration. □

A.2 Fonctions strictement convexes

Nous allons maintenant passer à la notion de stricte convexité. Comme ici les différentes possibilités ne sont plus équivalentes, il nous faut choisir l'une d'entre elles comme définition. Pour ce cours, nous avons pris la plus naturelle d'autant plus qu'elle n'entraine pas de détails techniques supplémentaires.

Définition A.7. Une fonction $f : [a,b] \to \mathbb{R}$ est dite *strictement convexe* si, pour tout $x_1 \neq x_2 \in [a,b]$ et pour tout $\lambda \in]0,1[$, on a

$$f\big(\lambda x_1 + (1-\lambda)x_2\big) < \lambda f(x_1) + (1-\lambda) f(x_2). \tag{A.5}$$

Proposition A.8. *Soit $f \in \mathcal{C}^1(]a,b[;\mathbb{R}) \cap \mathcal{C}([a,b];\mathbb{R})$. Parmi les propositions suivantes,*

(1) f est strictement convexe sur $[a,b]$;

(2) ∂f est strictement croissante sur $]a,b[$;

(3) pour tout $x_0 \in]a,b[$ et tout $x \in [a,b] \setminus \{x_0\}$, on a

$$f(x) > f(x_0) + \partial f(x_0)(x - x_0)\,;$$

et, si $f \in \mathcal{C}^2(]a,b[;\mathbb{R}) \cap \mathcal{C}([a,b];\mathbb{R})$,

(4) pour tout $x \in]a,b[$, $\partial^2 f(x) > 0$;

on a les implications (1) $\Leftrightarrow$ (2) $\Leftrightarrow$ (3) $\Leftarrow$ (4).

Démonstration. (1) $\Rightarrow$ (3) Supposons que (1) soit vrai mais pas (3). Dès lors, il existe un $x_0 \in]a,b[$ et un $x \in [a,b] \setminus \{x_0\}$ tels que

$$f(x) \leqslant f(x_0) + \partial f(x_0)(x - x_0).$$

D'autre part, puisque (1) implique que f est convexe, on est sûr d'avoir au moins l'inégalité « $\geqslant$ ». Donc

$$f(x) = f(x_0) + \partial f(x_0)(x - x_0). \tag{A.6}$$

On a même mieux. De la convexité de f, on déduit que, pour tout $\lambda \in [0,1]$,

$$\begin{aligned} f\big(\lambda x_0 + (1-\lambda)x\big) &\geqslant f(x_0) + (1-\lambda)\partial f(x_0)(x - x_0) \\ &= f(x_0) + (1-\lambda)\big(f(x) - f(x_0)\big) \\ &= \lambda f(x_0) + (1-\lambda) f(x) \end{aligned}$$

où la première égalité découle de (A.6). Cela contredit le fait que (1) doit être vrai pour $x_1 := x_0$ et $x_2 := x$.

(3) $\Rightarrow$ (2) Il suffit[5] de calquer le « (2) $\Rightarrow$ (3) » de la proposition A.5 en remplacant les inégalités par des inégalités strictes.

(2) $\Rightarrow$ (1) On imite le « (3) $\Rightarrow$ (1) » de la proposition A.5. Pour avoir l'inégalité stricte, il suffit de remarquer que, lorsque $t \in]x_1, x_2[$, on a $\lambda x_1 + (1-\lambda)t < t$ et donc $\partial f\big(\lambda x_1 + (1-\lambda)t\big) < \partial f(t)$. On en déduit

$$f\big(\lambda x_1 + (1-\lambda)x_2\big) < f(x_1) + (1-\lambda)\int_{]x_1,x_2[} \partial f(t)\,\mathrm{d}t$$

et dès lors l'assertion (1).

(4) $\Rightarrow$ (2) Il suffit d'imiter la deuxième partie de la preuve de la proposition A.6. Puisque maintenant $\partial^2 f(\xi) > 0$ dans l'équation (A.4), on a bien que $x_1 < x_2 \Rightarrow \partial f(x_1) < \partial f(x_2)$. □

Remarque A.9. (2) $\not\Rightarrow$ (4) En effet, $f : [-1,1] \to \mathbb{R} : x \mapsto x^4$ est tel que ∂f est strictement croissante sur $]-1,1[$ mais $\partial^2 f(0) = 0$.

A.3 Fonctions concaves et strictement concaves

Définition A.10. Une fonction $f : [a,b] \to \mathbb{R}$ est dite *concave* si, pour tout $x_1, x_2 \in [a,b]$ et pour tout $\lambda \in [0,1]$, on a

$$f\big(\lambda x_1 + (1-\lambda)x_2\big) \geqslant \lambda f(x_1) + (1-\lambda) f(x_2)\,; \tag{A.7}$$

f est dite *strictement concave* si, pour tout $x_1 \neq x_2 \in [a,b]$ et pour tout $\lambda \in]0,1[$, on a

$$f\big(\lambda x_1 + (1-\lambda)x_2\big) > \lambda f(x_1) + (1-\lambda) f(x_2). \tag{A.8}$$

On remarque que f est concave (resp. strictement concave) ssi $-f$ est convexe (resp. strictement convexe). Grâce à cette « dualité », on déduit aisément les propositions suivantes.

Proposition A.11. *Si $f : [a,b] \to \mathbb{R}$ est une fonction concave, alors f est continue sur $]a,b[$.*

Proposition A.12. *Soit $f : [a,b] \to \mathbb{R}$ une fonction continue. f est concave sur $[a,b]$ si et seulement si*

$$f\big(\tfrac{1}{2}(x+y)\big) \geqslant \tfrac{1}{2}\big(f(x) + f(y)\big). \tag{A.9}$$

5. Faites le !

Proposition A.13. *Soit $f \in \mathcal{C}^1(]a,b[;\mathbb{R}) \cap \mathcal{C}([a,b];\mathbb{R})$. Les propositions suivantes sont équivalentes :*

(1) f est concave ;

(2) pour tout $x_0 \in]a,b[$ et tout $x \in [a,b]$, on a

$$f(x) \leqslant f(x_0) + \partial f(x_0)(x - x_0) ;$$

(3) la dérivée de f, ∂f, est décroissante sur $]a,b[$.

Si de plus $f \in \mathcal{C}^2(]a,b[;\mathbb{R}) \cap \mathcal{C}([a,b];\mathbb{R})$. on a aussi l'équivalence avec

(4) $\partial^2 f(x) \leqslant 0$ pour tout $x \in]a,b[$.

Proposition A.14. *Soit $f \in \mathcal{C}^1(]a,b[;\mathbb{R}) \cap \mathcal{C}([a,b];\mathbb{R})$. Parmi les propositions suivantes,*

(1) f est strictement concave sur $[a,b]$;

(2) ∂f est strictement décroissante sur $]a,b[$;

(3) pour tout $x_0 \in]a,b[$ et tout $x \in [a,b] \setminus \{x_0\}$, on a

$$f(x) < f(x_0) + \partial f(x_0)(x - x_0) ;$$

et, si $f \in \mathcal{C}^2(]a,b[;\mathbb{R}) \cap \mathcal{C}([a,b];\mathbb{R})$,

(4) pour tout $x \in]a,b[$, $\partial^2 f(x) < 0$;

on a les implications (1) $\Leftrightarrow$ (2) $\Leftrightarrow$ (3) $\Leftarrow$ (4).

A.4 Exercices

Exercice A.1 Soient $a, b \in \mathbb{R}$. Montrez que

- $x \in [a,b]$ si et seulement si $\exists \lambda \in [0,1],\ x = \lambda a + (1-\lambda)b$.
- $x \in]a,b[$ si et seulement si $\exists \lambda \in]0,1[,\ x = \lambda a + (1-\lambda)b$.

Annexe B

Différences divisées

Les différences divisées sont particulièrement intéressantes car elles sont en quelque sorte des contreparties discrètes de la notion de dérivée. Cela est particulièrement clair au vu du chapitre III où les polynômes interpolants s'expriment sous une forme qui rappelle le développement de Taylor. Le but de cette annexe est de définir les différences divisées pour elles-mêmes et de donner une preuve directe de leur représentation en terme de dérivée.

Définition B.1. Soit $(x_n)_{n\in\mathbb{N}} \subseteq [a,b]$ une suite de nombres réels tous distincts et $f : [a,b] \to \mathbb{R}$. Les *différences divisées* $f[x_0,\dots,x_k]$ ($k \geqslant 0$) sont définies récursivement par

$$f[x_0] := f(x_0), \qquad f[x_0,\dots,x_{k+1}] := \frac{f[x_0,\dots,x_k] - f[x_1,\dots,x_{k+1}]}{x_0 - x_{k+1}}.$$

La première chose que nous allons montrer est que $f[x_0,\dots,x_k]$ ne dépend pas de l'ordre de $x_0,\dots,x_k$. Plus formellement, cela veut dire que si $(\pi(0),\dots,\pi(k))$ est une permutation de $(0,\dots,k)$ (c'est-à-dire que $\pi : \{0,\dots,k\} \to \{0,\dots,k\}$ est une bijection), on a $f[x_{\pi(0)},\dots,x_{\pi(k)}] = f[x_0,\dots,x_k]$. Montrons cela par récurrence sur k. Pour $k = 0$, il n'y a qu'un élément est donc c'est évident. Supposons maintenant qu'on puisse permuter à sa guise les différences divisées sur $k+1$ points et montrons que c'est encore le cas pour $k+2$ points. Par définition des différences divisées, on a

$$f[x_{k+1},x_1,\dots,x_k,x_0] = f[x_0,x_1,\dots,x_k,x_{k+1}]. \tag{B.1}$$

Il suffit de montrer que

$$f[x_0,\dots,x_{k-1},x_{k+1},x_k] = f[x_0,\dots,x_{k-1},x_k,x_{k+1}] \tag{B.2}$$

En effet, par hypothèse de récurrence et vu la définition de $f[x_0,\dots,x_{k+1}]$, on peut permuter $(x_1,\dots,x_k)$ comme bon nous semble. Si $(x_{\pi(0)},\dots,x_{\pi(k+1)})$ est une permutation

de $(x_0,\dots,x_{k+1})$, $x_0 = x_{\pi(i)}$ peut être remis en première position : s'il est en dernière position, il suffit d'utiliser (B.1) ; si x_0 est parmi $x_{\pi(1)},\dots,x_{\pi(k)}$, on peut le mettre en place k par l'hypothèse de récurrence et ensuite en place 0 par (B.2) suivit de (B.1). Par un procédé similaire, on peut mettre x_{k+1} à la place $k+1$ (sans toucher à x_0). Il suffit alors de réordonner les places $1,\dots,k$ ce qui est possible par l'hypohtèse de récurrence.

Prouvons maintenant (B.2). En utilisant la définition des différences divisées et l'hypothèse de récurrence, on obtient :

$$\begin{aligned}
f[x_0,\dots,x_k,x_{k+1}] &= \frac{f[x_0,\dots,x_k]-f[x_1,\dots,x_{k+1}]}{x_0-x_{k+1}} \\
&= \frac{f[x_0,\dots,x_k]-f[x_k,x_1,\dots,x_{k-1},x_{k+1}]}{x_0-x_{k+1}} \\
&= \frac{\dfrac{f[x_0,\dots,x_{k-1}]-f[x_1,\dots,x_k]}{x_0-x_k} - \dfrac{f[x_1,\dots,x_k]-f[x_1,\dots,x_{k-1},x_{k+1}]}{x_k-x_{k+1}}}{x_0-x_{k+1}} \\
&= \frac{(x_k-x_{k+1})f[x_0,\dots,x_{k-1}]-(x_0-x_{k+1})f[x_1,\dots,x_k]+(x_0-x_k)f[x_1,\dots x_{k-1},x_{k+1}]}{(x_0-x_k)(x_0-x_{k+1})(x_k-x_{k+1})}
\end{aligned}$$

Sur la dernière formule, on voit que si on permutte x_k et x_{k+1}, l'expression ne change pas. C'est donc que (B.2) est vrai.

Passons au lien entre les différences divisées et les dérivées. Nous allons commencer par proposer une formule intégrale pour les différences divisées. Pour la comprendre, commençons par examiner les cas de deux points et de trois points. Pour deux points, on a par le théorème fondamental de l'analyse que

$$f[x_0,x_1] = \frac{f(x_0)-f(x_1)}{x_0-x_1} = \int_{x_0}^{x_1} \partial f(x)\frac{\mathrm{d}x}{x_0-x_1}.$$

Si on change la variable d'intégration en t définit par $x_0+t(x_0-x_1)=x$, on trouve

$$f[x_0,x_1] = \int_0^1 \partial f\big(x_0+t(x_0-x_1)\big) = \int_0^1 \partial f\big((1-t)x_0+tx_1)\big).$$

Par symmétrie (en permutant x_0 et x_1), on a évidemment aussi

$$f[x_0,x_1] = \int_0^1 \partial f\big(tx_0+(1-t)x_1)\big).$$

En utilisant cette formule, on peut passer à trois points :

$$\begin{aligned}
f[x_0,x_1,x_2] &= \frac{f[x_0,x_1]-f[x_1,x_2]}{x_0-x_2} \\
&= \int_0^1 \frac{\partial f\big((1-t_1)x_0+t_1x_1\big)-\partial f\big(t_1x_1+(1-t_1)x_2\big)}{x_0-x_2}\,\mathrm{d}t_1 \\
&= \int_0^1 \mathrm{d}t_1 \int_{t_1x_1+(1-t_1)x_2}^{(1-t_1)x_0+t_1x_1} \partial^2 f(x)\frac{\mathrm{d}x}{x_0-x_2}.
\end{aligned}$$

En faisant le changement de variable $t_2 \mapsto x = t_1x_1 + (1-t_1)x_0 + t_2(x_2 - x_0)$, on trouve

$$f[x_0,x_1,x_2] = \int_0^1 \mathrm{d}t_1 \int_0^{1-t_1} \mathrm{d}t_2\, \partial^2 f\big((1-t_1-t_2)x_0 + t_1x_1 + t_2x_2\big).$$

À partir de cela, on peut avoir une intuition de la formule générale.

Lemme B.2. *Si $f \in \mathcal{C}^k(]a,b[;\mathbb{R})$ et $x_0,\dots,x_k$ sont $k+1$ points distincts de $]a,b[$, on a*

$$f[x_0,x_1,\dots,x_k] = \int_0^1 \mathrm{d}t_1 \int_0^{s_1} \mathrm{d}t_2 \cdots \int_0^{s_{k-1}} \mathrm{d}t_k\, \partial^k f\Big(s_kx_0 + \sum_{i=1}^k t_ix_i\Big) \tag{B.3}$$

où $s_i := 1 - \sum_{j=1}^i t_i$ pour $i \geqslant 0$.

Démonstration. Pour $k=1$ et $k=2$, cette formule est vraie par les calculs ci-dessus. Supposons qu'elle soit vraie poir $k+1$ points quelconques et montrons que c'est encore le cas pour $k+2$ points. Par définition des différences divisées et par invariance sous permutation des points, on a

$$\begin{aligned} f[x_0,\dots,x_{k+1}] &= \frac{f[x_0,\dots,x_k] - f[x_1,\dots,x_{k+1}]}{x_0 - x_{k+1}} = \frac{f[x_0,\dots,x_k] - f[x_{k+1},x_1,\dots,x_k]}{x_0 - x_{k+1}} \\ &= \int_0^1 \mathrm{d}t_1 \int_0^{s_1} \mathrm{d}t_2 \cdots \int_0^{s_{k-1}} \mathrm{d}t_k\, \frac{\partial^k f\big(s_kx_0 + \sum_{i=1}^k t_ix_i\big) - \partial^k f\big(\sum_{i=1}^k t_ix_i + s_kx_{k+1}\big)}{x_0 - x_{k+1}}. \end{aligned}$$

Intéressons nous seulement à la fraction. On peut la réécrire comme suit :

$$\int_{\sum_{i=1}^k t_ix_i + s_kx_{k+1}}^{s_kx_0 + \sum_{i=1}^k t_ix_i} \partial^{k+1} f(x)\, \frac{\mathrm{d}x}{x_{k+1} - x_0}.$$

En faisant le changement de variable $t_{k+1} \mapsto x = s_kx_0 + \sum_{i=1}^k t_ix_i + t_{k+1}(x_{k+1} - x_0)$, cette intégrale prend la forme

$$\int_0^{s_k} \partial^{k+1} f\Big((s_k - t_{k+1})x_0 + \sum_{i=0}^{k+1} t_ix_i\Big)\, \mathrm{d}t_{k+1}.$$

Mais, par définition de s_k, on a $s_k - t_{k+1} = s_{k+1}$ et cela prouve la validité de (B.3). □

Nous allons également avoir besoin du théorème de la moyenne suivant.

Théorème B.3. (théorème de la moyenne intégral) *Soit $A \subseteq \mathbb{R}^N$ un ensemble fermé borné de $\mathbb{R}^N$ qui est connexe par arcs et $f : A \to \mathbb{R}$ une fonction continue. Alors il existe un $\xi \in A$ tel que*

$$\int_A f(x)\,\mathrm{d}x = f(\xi)\,\mathrm{mes}(A).$$

Remarques.

- La mesure de A, $\mathrm{mes}(A)$, est définie comme $\int_A 1\,\mathrm{d}x$.
- Un ensemble A est dit *connexe par arcs* si, pour tous $x_0, x_1 \in A$, il existe une fonction continue $\gamma : [0,1] \to A$ tel que $\gamma(0) = x_0$ et $\gamma(1) = x_1$. On dit que γ est un chemin joignant x_0 à x_1.

Démonstration. Comme A est un ensemble compact, la fonction f atteint ses bornes, c'est-à-dire qu'il existe $x_0, x_1 \in A$ tels que $f(x_0) = \min_A f$ et $f(x_1) = \max_A f$. De $f(x_0) \leqslant f(x) \leqslant f(x_1)$, on déduit en intégrant sur A que

$$\mathrm{mes}(A) f(x_0) \leqslant \int_A f(x)\,\mathrm{d}x \leqslant \mathrm{mes}(A) f(x_1).$$

On peut réécrire cela comme

$$\frac{1}{\mathrm{mes}(A)} \int_A f(x)\,\mathrm{d}x \in [f(x_0), f(x_1)].$$

Comme l'ensemble A est connexe par arcs, la propriété de valeur intermédiaire implique qu'il existe un $x \in A$ tel que $(1/\mathrm{mes}(A)) \int_A f = f(\xi)$. □

Nous sommes maintenant en mesure de prouver le théorème principal de cette annexe.

Théorème B.4. *SI $f \in \mathcal{C}^k(]a,b[;\mathbb{R})$ et $x_0, \ldots, x_k$ sont des points tous distincts de $]a,b[$, il existe un $\xi \in \left[\min_{1\leqslant i\leqslant k} x_i, \max_{1\leqslant i\leqslant k} x_i\right]$ tel que*

$$f[x_0, \ldots, x_k] = \frac{1}{k!}\,\partial^k f(\xi). \tag{B.4}$$

Démonstration. La formule intégrale (B.3) peut se mettre sous la forme

$$f[x_0, \ldots, x_k] = \int_A F(t_1, \ldots, t_k)\,\mathrm{d}(t_1, \ldots, t_k)$$

où $A := \left\{(t_1, \ldots t_k) \in \mathbb{R}^k : 0 \leqslant t_1 \leqslant 1 \text{ et } 0 \leqslant t_i \leqslant s_{i-1} \text{ pour } i = 2, \ldots, k\right\}$ et $F(t_1, \ldots, t_k) := \partial^k f(s_k x_0 + \sum_{i=1}^k t_i x_i)$. Par le théorème de la moyenne intégral, il existe $(t_1, \ldots, t_k) \in A$ tel que

$$\int_A F = \mathrm{mes}(A)\, F(t_1, \ldots, t_k).$$

Posons $\xi := s_k x_0 + \sum_{i=1}^k t_i x_i$. Puisque $s_k + \sum_{i=1}^k t_i = 1$, il est facile de voir que $\min_i x_i \leqslant \xi \leqslant \max_i x_i$. Pour compléter la preuve, il suffit de montrer que $\mathrm{mes}(A) = 1/k!$. Par définition,

$$\mathrm{mes}(A) = \int_0^1 \mathrm{d}t_1 \int_0^{s_1} \mathrm{d}t_2 \cdots \int_0^{s_{k-1}} \mathrm{d}t_k 1.$$

Nous allons montrer par récurrence que

$$V_i := \int_0^{s_{k-i}} \mathrm{d}t_{k-i+1} \cdots \int_0^{s_{k-1}} \mathrm{d}t_k 1 = \frac{1}{i!}\, s_{k-i}^i. \tag{B.5}$$

Pour $i = 1$, on a que $V_1 = \int_0^{s_{k-1}} \mathrm{d}t_k = s_{k-1}$. Supposons que la formule (B.5) soit vraie pour i et montrons la pour $i+1$. Par définition,

$$V_{i+1} = \int_0^{s_{k-i-1}} V_i \, \mathrm{d}t_{k-i}.$$

L'hypothèse de récurrence implique que $V_i = (1/i!)s_{k-i}^i = (1/i!)(s_{k-i-1} - t_{k-i})^i$. Donc

$$\begin{aligned} V_{i+1} &= \int_0^{s_{k-i-1}} \frac{1}{i!}(s_{k-i-1} - \tau)^i \, \mathrm{d}\tau \\ &= \frac{1}{i!} \, \frac{-(s_{k-i-1} - \tau)^{i+1}}{i+1} \Bigg]_{\tau=0}^{s_{k-i-1}} = \frac{1}{(i+1)!} s_{k-(i+1)}^{i+1}. \end{aligned}$$

Cela prouve (B.5). En particulier, quand $i = k$, on a

$$\mathrm{mes}(A) = V_k = \frac{1}{k!}\, s_0^k = \frac{1}{k!}.$$

Ceci conclut la preuve. □

Grâce au théorème ci-dessus, on peut définir les différences divisées lorsque plusieurs points sont égaux. Dans le cas extrême où tous les points sont les mêmes, on a envie de poser en vertu de la formule (B.4) que :

$$f[\underbrace{x_0, \ldots, x_0}_{k+1}] = \frac{\partial^k f(x_0)}{k!}.$$

De manière générale, soit $(x_0, x_1, \ldots, x_k) \in \mathbb{R}^{k+1}$ où certains points x_i peuvent être égaux. Il existe une [1] suite $(x_0^{(n)}, \ldots, x_k^{(n)})$ telle que tous les points $x_0^{(n)}, \ldots, x_k^{(n)}$ soient différents et telle que $x_i^{(n)} \xrightarrow[n\to\infty]{} x_i$ pour tout $i = 0, 1, \ldots, k$. On définit

$$f[x_0, x_1, \ldots, x_k] = \lim_{n\to\infty} f[x_0^{(n)}, \ldots, x_k^{(n)}].$$

Pour que cette définition ait un sens, il faut prouver que la limite existe et qu'elle est indépendante de la suite $\big((x_0^{(n)}, \ldots, x_k^{(n)})\big)_{n\in\mathbb{N}}$. Pour cela, la formule intégrale (B.3) nous est de nouveau d'une grande aide. On peut la réécrire de manière condensée comme : lorsque $\xi_0, \ldots, \xi_k$ sont tous différents, on a

$$f[\xi_0, \ldots, \xi_k] = \int_A F(t, \xi_0, \ldots, \xi_k) \, \mathrm{d}t$$

1. En fait il existe une infinité de telles suites !

où $A := \{t = (t_1, \dots, t_k) \in [0,1]^k : t_{i+1} \leqslant 1 - \sum_{j=1}^{i} t_j \text{ pour } 1 \leqslant i < k\}$ et $F(t, \xi_0, \dots, \xi_k) := \partial^k f\big((1 - \sum_{i=1}^{k} t_i)\xi_0 + \sum_{i=1}^{k} t_i \xi_i\big)$. Maintenant, si $(x_0^{(n)}, \dots, x_k^{(n)}) \xrightarrow[n]{} (x_0, \dots, x_k)$, on en déduit que, pour tout t, $F(t, x_0^{(n)}, \dots, x_k^{(n)}) \xrightarrow[n]{} F(t, x_0, \dots, x_k)$. Choisissons ε suffisament petit pour que $K_\varepsilon := \big[\min_{1 \leqslant j \leqslant k} x_j - \varepsilon, \max_{1 \leqslant j \leqslant k} x_j + \varepsilon\big] \subseteq]a, b[$. Pour n assez grand, on peut supposer que $x_i^{(n)} \in K_\varepsilon$ pour tout i. Dès lors, $(1 - \sum_{i=1}^{k} t_i) x_0^{(n)} + \sum_{i=1}^{k} t_i x_i^{(n)} \in K_\varepsilon$. Comme K_ε est compact, $C := \sup_{x \in K_\varepsilon} |\partial^k f(x)| < +\infty$. Donc, $\big|F(t, x_0^{(n)}, \dots, x_k^{(n)})\big| \leqslant C$. Par le théorème de convergence dominée de Lebesgue, on peut conclure que

$$\begin{aligned} \lim_{n \to \infty} \int_A F(t, x_0^{(n)}, \dots, x_k^{(n)}) \, \mathrm{d}t &= \int_A \lim_{n \to \infty} F(t, x_0^{(n)}, \dots, x_k^{(n)}) \, \mathrm{d}t \\ &= \int_A F(t, x_0, \dots, x_k) \, \mathrm{d}t \end{aligned}$$

Cela montre non seulement que la limite existe et est indépendante de la suite $\big((x_0^{(n)}, \dots, x_k^{(n)})\big)_{n \in \mathbb{N}}$ mais de plus la formule intégrale (B.3) est encore valable pour $(x_0, \dots, x_k)$ quand bien même certains points sont égaux. Cela implique également que le théorème B.4 est aussi vrai pour les différences divisées correspondantes $f[x_0, \dots, x_k]$.

Bibliographie

[1] Walter Gautschi. *Numerical analysis : an introduction*. Birkhäuser, 1997.