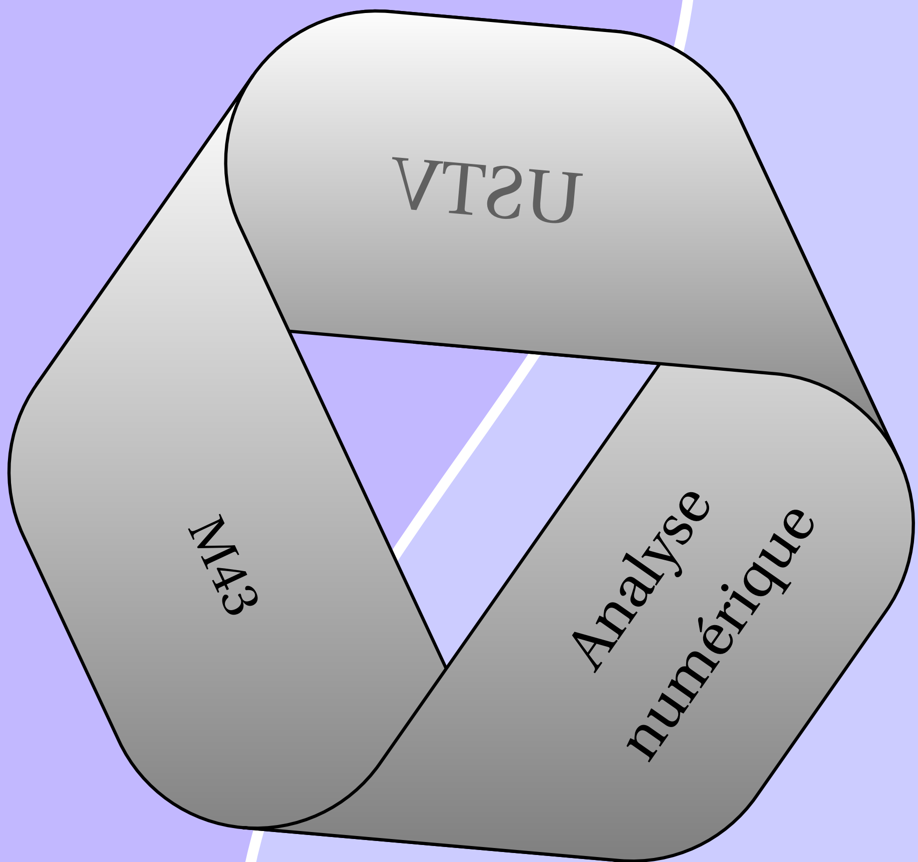
L2

2011/2012

Recueil d'exercices corrigés
et aide-mémoire



G. Faccanoni

Dernière mise-à-jour
Jeudi 10 mai 2012

Avertissement : ces notes sont régulièrement mises à jour et corrigées, ne vous étonnez pas si vous découvrez des erreurs. Merci de me les communiquer. Toutes les remarques ou questions permettant d'en améliorer la rédaction peuvent être envoyées à l'adresse `gloria.faccanoni@univ-tln.fr`

Gloria FACCANONI

IMATH Bâtiment U-318
Université du Sud Toulon-Var
Avenue de l'université
83957 LA GARDE - FRANCE

☎ 0033 (0)4 94 14 23 81

✉ `gloria.faccanoni@univ-tln.fr`
ℹ `http://faccanoni.univ-tln.fr`

Table des matières

1. Résolution d'équations non linéaires

Recherche de la solution de l'équation non linéaire $f(x)=0$ où f est une fonction donnée

Théorème des zéros d'une fonction continue

Soit une fonction continue $f\colon [a,b]\to\mathbb{R}$, si $f(a)f(b)<0$, alors il existe $\alpha\in]a,b[$ tel que $f(\alpha)=0$.

Méthode de dichotomie et méthode de Lagrange

Soit deux points a_0 et b_0 (avec $a_0<b_0$) d'images par f de signe contraire (*i.e.* $f(a_0)\cdot f(b_0)<0$). En partant de $I_0=[a_0,b_0]$, les méthodes *de dichotomie* et *de* Lagrange (appelée aussi *Regula falsi*) produisent une suite de sous-intervalles $I_k=[a_k,b_k]$, $k\ge 0$, avec $I_k\subset I_{k-1}$ pour $k\ge 1$ et tels que $f(a_k)\cdot f(b_k)<0$.

- ▷ Dans la *méthode de dichotomie*, on découpe l'intervalle $[a_k;b_k]$ en deux intervalles de même longueur, *i.e.* on divise $[a_k;b_k]$ en $[a_k;c_k]$ et $[c_k;b_k]$ où c_k est
$$c_k=\frac{a_k+b_k}{2}.$$
- ▷ Dans la *méthode de Lagrange*, plutôt que de diviser l'intervalle $[a_k;b_k]$ en deux intervalles de même longueur, on découpe $[a_k;b_k]$ en $[a_k;c_k]$ et $[c_k;b_k]$ où c_k est l'abscisse du point d'intersection de la droite passant par $(a_k,f(a_k))$ et $(b_k,f(b_k))$ et l'axe des abscisses, i.e. le zéro de la fonction
$$g(c)=\frac{f(b_k)-f(a_k)}{b_k-a_k}(c-a_k)+f(a_k)$$
qui est
$$c_k=a_k-\frac{b_k-a_k}{f(b_k)-f(a_k)}f(a_k)=\frac{a_kf(b_k)-b_kf(a_k)}{f(b_k)-f(a_k)}.$$

Dans les deux cas, pour l'itération suivante, on pose soit $[a_{k+1};b_{k+1}]=[a_k;c_k]$ soit $[a_{k+1};b_{k+1}]=[c_k;b_k]$ de sorte à ce que $f(a_{k+1})\cdot f(b_{k+1})<0$. Les algorithmes s'écrivent alors comme suit :

Dichotomie :

Require: a, b, ε, $f\colon[a,b]\to\mathbb{R}$
- $k\leftarrow 0$
- $a_k\leftarrow a$
- $b_k\leftarrow b$
- $x_k\leftarrow \frac{a_k+b_k}{2}$
- **while** $|b_k-a_k|>\varepsilon$ **do**
  - **if** $f(a_k)f(x_k)<0$ **then**
    - $a_{k+1}\leftarrow a_k$
    - $b_{k+1}\leftarrow x_k$
  - **else**
    - $a_{k+1}\leftarrow x_k$
    - $b_{k+1}\leftarrow b_k$
  - **end if**
  - $x_{k+1}\leftarrow\frac{a_{k+1}+b_{k+1}}{2}$
  - $k\leftarrow k+1$
- **end while**

Lagrange :

Require: a, b, ε, $f\colon[a,b]\to\mathbb{R}$
- $k\leftarrow 0$
- $a_k\leftarrow a$
- $b_k\leftarrow b$
- $x_k\leftarrow a_k-\frac{b_k-a_k}{f(b_k)-f(a_k)}f(a_k)$
- **while** $|b_k-a_k|>\varepsilon$ **do**
  - **if** $f(a_k)f(x_k)<0$ **then**
    - $a_{k+1}\leftarrow a_k$
    - $b_{k+1}\leftarrow x_k$
  - **else**
    - $a_{k+1}\leftarrow x_k$
    - $b_{k+1}\leftarrow b_k$
  - **end if**
  - $x_{k+1}\leftarrow a_{k+1}-\frac{b_{k+1}-a_{k+1}}{f(b_{k+1})-f(a_{k+1})}f(a_{k+1})$
  - $k\leftarrow k+1$
- **end while**

Remarque

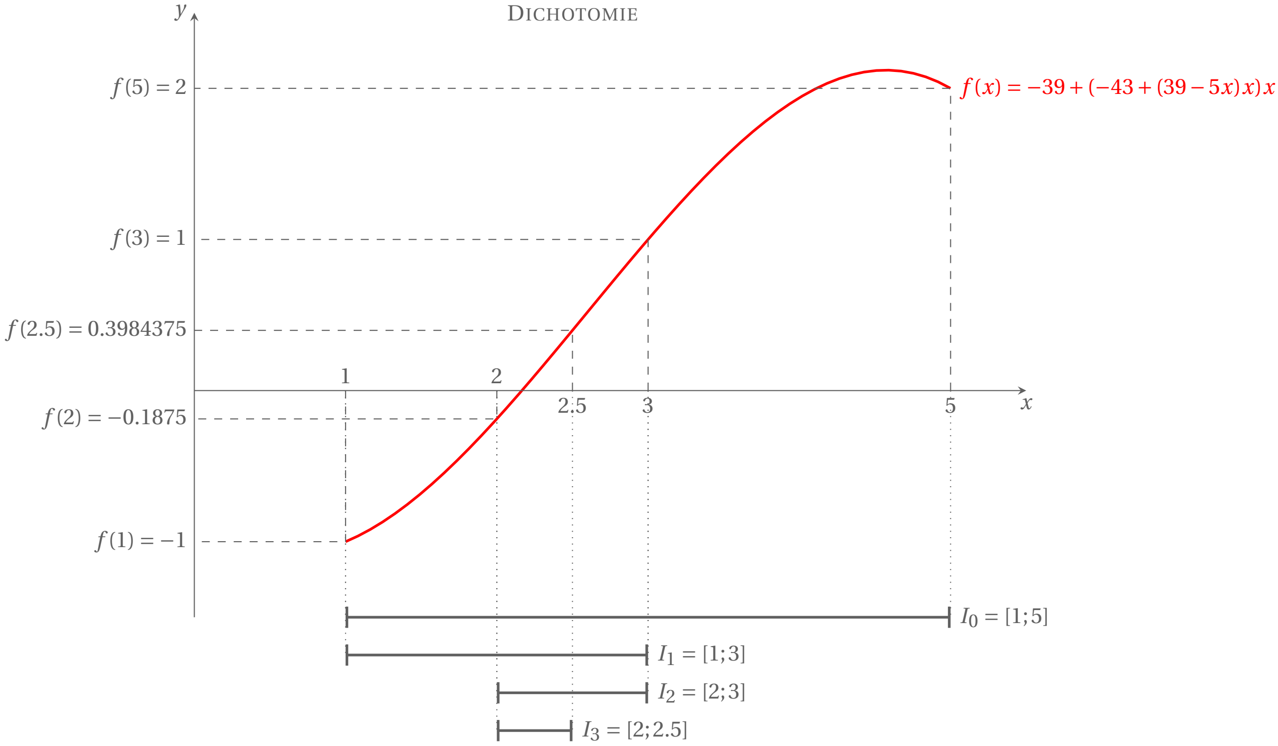
Avec la méthode de la dichotomie, les itération s'achèvent à la m-ème étape quand $|x_m-\alpha|\le|I_m|<\varepsilon$, où ε est une tolérance fixée et $|I_m|$ désigne la longueur de l'intervalle I_m. Clairement $I_k=\frac{b-a}{2^k}$, donc pour avoir une erreur $|x_m-\alpha|<\varepsilon$

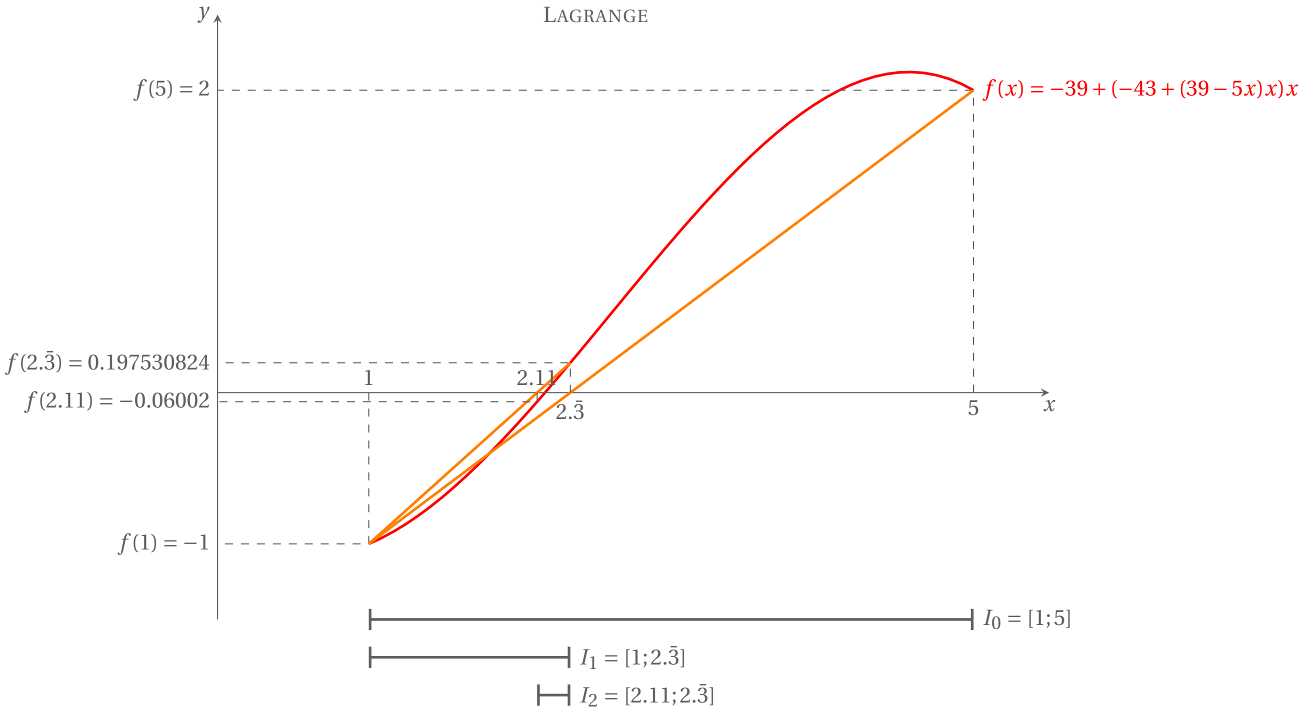
on doit prendre

$$m \geq \log_2 \frac{b-a}{\varepsilon}.$$

Exemple

Soit $f(x) = -39 - 43x + 39x^2 - 5x^3$. On cherche a estimer $x \in [1;5]$ tel que $f(x) = 0$.





Méthode de point fixe

Soit $f : [a,b] \to \mathbb{R}$. Il est toujours possible de transformer le problème $f(x) = 0$ en un problème équivalent $x - \varphi(x) = 0$, où la fonction auxiliaire $\varphi : [a,b] \to \mathbb{R}$ a été choisie de manière à ce que $\varphi(\alpha) = \alpha$ quand $f(\alpha) = 0$. Approcher les zéros de f se ramène donc au problème de la détermination des points fixes de φ, ce qui se fait en construisant la suite récurrente

$$\begin{cases} x_{k+1} = \varphi(x_k), \\ x_0 \text{ donné.} \end{cases}$$

On utilise alors l'algorithme itératif suivant :

Require: x_0, ε, $\varphi : [a,b] \to \mathbb{R}$
  $k \leftarrow 0$
  while $|x_{k+1} - x_k| > \varepsilon$ **do**
    $x_{k+1} \leftarrow \varphi(x_k)$
    $k \leftarrow k+1$
  end while

Convergence des itérations de point fixe

On se donne x_0 et on considère la suite $x_{k+1} = \varphi(x_k)$ pour $k \geq 0$. Si

Stabilité : $\varphi(x) \in [a,b]$ pour tout $x \in [a,b]$

Régularité : $\varphi \in \mathscr{C}^1([a,b])$

Contraction : il existe $K < 1$ tel que $|\varphi'(x)| \leq K$ pour tout $x \in [a,b]$

alors φ a un unique point fixe α dans $[a,b]$ et la suite $x_{k+1} = \varphi(x_k)$ converge vers α pour tout choix de x_0 dans $[a,b]$. De plus, on a

$$\lim_{k\to\infty} \frac{x_{k+1} - \alpha}{x_k - \alpha} = \varphi'(\alpha).$$

Ce théorème assure la convergence, avec un ordre 1, de la suite $(x_k)_{k\in\mathbb{N}}$ vers le point fixe α pour tout choix d'une valeur initiale $x_0 \in [a;b]$. Il constitue donc un exemple de résultat de convergence globale. Mais en pratique, il est souvent difficile de déterminer a priori l'intervalle $[a;b]$; dans ce cas, le résultat de convergence suivant peut être utile.

Théorème d'Ostrowski

Soit α un point fixe d'une fonction φ continue et différentiable dans un intervalle $[c;d]$ contenant α. Si $|\varphi'(\alpha)| < 1$, alors il existe un intervalle $[a;b] \subset [c;d]$ tel que la suite $(x_k)_k$ converge vers α pour tout $x_0 \in [a;b]$.

Remarque

- si $0 < \varphi'(\alpha) < 1$ la suite converge de façon monotone, c'est-à-dire, l'erreur $x_k - \alpha$ garde un signe constant quand k varie ;
- si $-1 < \varphi'(\alpha) < 0$ la suite converge de façon oscillante, c'est-à-dire, l'erreur $x_k - \alpha$ change de signe quand k varie ;
- si $|\varphi'(\alpha)| > 1$ la suite diverge. Plus précisément, si $\varphi'(\alpha) > 1$ la suite diverge de façon monotone, tandis que pour $\varphi'(\alpha) < -1$ elle diverge en oscillant ;
- si $|\varphi'(\alpha)| = 1$, on ne peut en général tirer aucune conclusion : selon le problème considéré, il peut y avoir convergence ou divergence.
  Par exemple,
  - *soit $\phi(x) = x - x^3$ qui admet $\alpha = 0$ comme point fixe. On a $\phi'(\alpha) = 1$ et $x_k \to \alpha$ pour tout $x_0 \in [-1;1]$ car*
    - *si $x_0 = \pm 1$ alors $x_k = \alpha$ pour tout $k \geq 1$,*
    - *si $x_0 \in [-1,1]$ alors $x_k \in]-1,1[$ pour tout $k \geq 1$ et la suite est monotone ;*
  - *considérons maintenant $\phi(x) = x + x^3$ qui admet aussi $\alpha = 0$ comme point fixe. À nouveau $\phi'(\alpha) = 1$ mais dans ce cas la suite diverge pour tout choix de $x_0 \neq 0$.*

Ordre de convergence

Soit α un point fixe d'une fonction $\varphi \in \mathscr{C}^{p+1}$ pour un entier $p \geq 1$ dans un intervalle $[a;b]$ contenant α. Si $\varphi^{(i)}(\alpha) = 0$ pour $1 \leq i \leq p$ et $\varphi^{(p+1)}(\alpha) \neq 0$, alors la méthode de point fixe associée à la fonction d'itération φ est d'ordre $p+1$.

Méthodes de point fixe particulièrement connues

Méthode de la Corde : $\varphi(x_k) = x_k - \dfrac{b-a}{f(b)-f(a)} f(x_k)$ ordre : 1

Méthode de Newton : $\varphi(x_k) = x_k - \dfrac{f(x_k)}{f'(x_k)}$ ordre : $\begin{cases} 2 & \text{si } \alpha \text{ est une racine simple} \\ 1 & \text{sinon} \end{cases}$

Ordre de convergence de la méthode de Newton

Soit la méthode de Newton pour le calcul de ℓ zéro de f. Cette méthode peut être mise sous la forme d'une itération de point fixe $u_{n+1} = \phi(u_n)$ en posant

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

Si $f(\ell) \neq 0$ (*i.e.* si ℓ est racine simple), on trouve

$$\phi'(x) = 1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2}, \qquad \phi'(\ell) = 0,$$

$$\phi''(x) = \frac{f''(x)}{f'(x)} + \frac{f(x)f'''(x)}{(f'(x))^2} - 2\frac{f(x)(f''(x))^2}{(f'(x))^3}, \qquad \phi''(\ell) = \frac{f''(\ell)}{f'(\ell)}.$$

La méthode de Newton est donc d'ordre 2. Si la racine ℓ est de multiplicité $m > 1$, alors la méthode n'est plus du second ordre. En effet, $f(x) = (x-\ell)^m h(x)$ où h est une fonction telle que $h(\ell) \neq 0$. On a alors

$$\phi(x) = 1 - \frac{f(x)}{f'(x)} = 1 - \frac{(x-\ell)h(x)}{mh(x) + (x-\ell)h'(x)},$$

$$\phi'(x) = \frac{h(x)\big(m(m-1)h(x) + 2(x-\ell)h'(x) + (x-\ell)^2 h''(x)\big)}{\big(mh(x) + (x-\ell)h'(x)\big)^2}, \qquad \phi'(\ell) = 1 - \frac{1}{m}.$$

Si la valeur de m est connue *a priori*, on peut retrouver la convergence quadratique en modifiant la méthode de Newton comme suit :

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

Critères d'arrêt

Supposons que $(x_n)_n$ soit une suite qui converge vers ℓ zéro de la fonction f. Nous avons le choix entre deux types de critères d'arrêt pour interrompre le processus itératif d'approximation de ℓ : ceux basés sur le résidu et ceux basés sur l'incrément. Nous désignerons par ε une tolérance fixée pour le calcul approché de ℓ et par $e_n = \ell - x_n$ l'erreur absolue. Nous supposerons de plus f continûment différentiable dans un voisinage de la racine.

Contrôle du résidu : les itérations s'achèvent dès que $|f(x_n)| < \varepsilon$. Il y a des situations pour lesquelles ce test s'avère trop restrictif ou, au contraire, trop optimiste.

- ▷ si $|f'(\ell)| \simeq 1$ alors $|e_n| \simeq \varepsilon$: le test donne donc une indication satisfaisante de l'erreur ;
- ▷ si $|f'(\ell)| \ll 1$, le test n'est pas bien adapté car $|e_n|$ peut être assez grand par rapport à ε ;
- ▷ si enfin $|f'(\ell)| \gg 1$ alors $|e_n| \ll \varepsilon$ et le test est trop restrictif.

Dans l'exercice précédent $f'(\ell) = 2\sqrt{2} \gg 1$: le test est trop restrictif (comparer la colonne $|f(x_n)|$ à la colonne $|\ell - x_n|$).

Contrôle de l'incrément : les itérations s'achèvent dès que $|x_{n+1} - x_n| < \varepsilon$. Soit $(x_n)_n$ la suite produite par la méthode de point fixe $x_{n+1} = \phi(x_n)$. Comme $\ell = \phi(\ell)$ et $x_{n+1} = \phi(x_n)$, on obtient par un développement au premier ordre

$$e_{n+1} = \ell - x_{n+1} = \phi(\ell) - \phi(x_n) = \phi'(\xi_n)(\ell - x_n) = \phi'(\xi_n)e_n, \qquad \xi_n \in I_{\ell,x_n}$$

I_{ℓ,x_n} étant l'intervalle d'extrémités ℓ et x_k. En utilisant l'identité

$$e_n = (\ell - x_{n+1}) + (x_{n+1} - x_n) = e_{n+1} + (x_{n+1} - x_n) = \phi'(\xi_n)e_n + (x_{n+1} - x_n),$$

on en déduit que

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

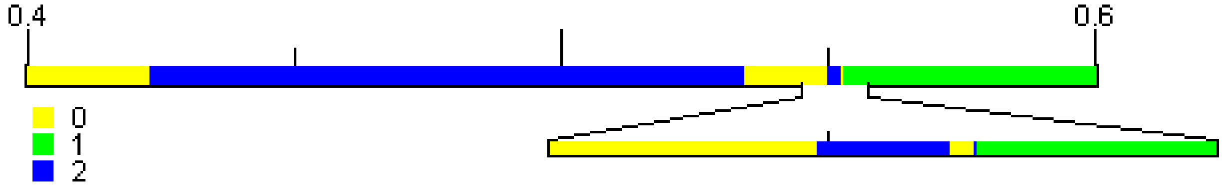
Par conséquent, ce critère fournit un estimateur d'erreur satisfaisant si $\phi'(x) \simeq 0$ dans un voisinage de ℓ. C'est le cas notamment des méthodes d'ordre 2, dont la méthode de Newton. Cette estimation devient d'autant moins bonne que ϕ' s'approche de 1.

Les fractales de Newton

Prenons la fonction $f: x \mapsto x^3 - 3x^2 + 2x$. Ce polynôme admet 3 racines : 0, 1 et 2. Que se passe-t-il lorsque l'on applique la méthode de Newton en partant de $x_0 = 0.4$? En partant de $x_0 = 0.5$? En partant de $x_0 = 0.6$? De $x_0 = 0.5527$? De $x_0 = 0.55275$? De $x_0 = 0.5528$?

▷ si $x_0 = 0.4$ alors $x_n \to 0$
▷ si $x_0 = 0.4$ alors $x_n \to 2$
▷ si $x_0 = 0.4$ alors $x_n \to 1$
▷ si $x_0 = 0.4$ alors $x_n \to 0$
▷ si $x_0 = 0.4$ alors $x_n \to 2$
▷ si $x_0 = 0.4$ alors $x_n \to 1$

En choisissant une couleur par limite possible (0 :jaune, 1 :vert et 2 :bleu), on retrouve quelque chose de fractal !



Source : `http://eljjdx.canalblog.com/archives/2008/08/30/10303555.html`

Exercices

Exercice 1.1

Décrire les méthodes de la dichotomie et de LAGRANGE et les utiliser pour calculer le zéro de la fonction

$$f(x) = x^3 - 4x - 8.95$$

dans l'intervalle $[2;3]$ avec une précision de 10^{-2}.

SOLUTION. En partant de $I_0 = [a, b]$, les méthodes de la dichotomie et de LAGRANGE produisent une suite de sous-intervalles $I_k = [a_k, b_k]$, $k \geq 0$, avec $I_k \subset I_{k-1}$, $k \geq 1$, et tels que $f(a_k)f(b_k) < 0$. Dans notre cas on a

$k \leftarrow 0$
$a_k \leftarrow 2$
$b_k \leftarrow 3$
while $|b_k - a_k| > 0.01$ **do**
  $x_k \leftarrow g(a_k, b_k)$
  $k \leftarrow k+1$
  if $(a_k^3 - 4a_k - 8.95)(x_k^3 - 4x_k - 8.95) < 0$ **then**
    $a_{k+1} \leftarrow a_k$
    $b_{k+1} \leftarrow x_k$
  else
    $a_{k+1} \leftarrow x_k$
    $b_{k+1} \leftarrow b_k$
  end if
end while

avec

$$g(a_k, b_k) = \begin{cases} \frac{a_k+b_k}{2} & \text{pour la méthode de la dichotomie,} \\ \frac{a_k f(b_k) - b_k f(a_k)}{f(b_k) - f(a_k)} & \text{pour la méthode de la LAGRANGE.} \end{cases}$$

Dichotomie

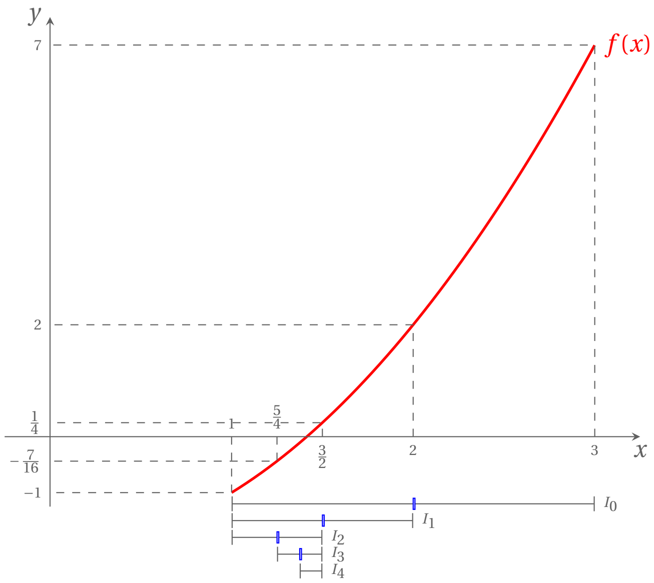
k	a_k	x_k	b_k	signe de $f(a_k)$	signe de $f(x_k)$	signe de $f(b_k)$
0	2.000000	2.5000000	3.00000	−	−	+
1	2.500000	2.7500000	3.00000	−	+	+
2	2.500000	2.6250000	2.75000	−	−	+
3	2.625000	2.6875000	2.75000	−	−	+
4	2.687500	2.7187500	2.75000	−	+	+
5	2.687500	2.7031250	2.71875	−	−	+
6	2.703125	2.7109375	2.71875	−	+	+

LAGRANGE

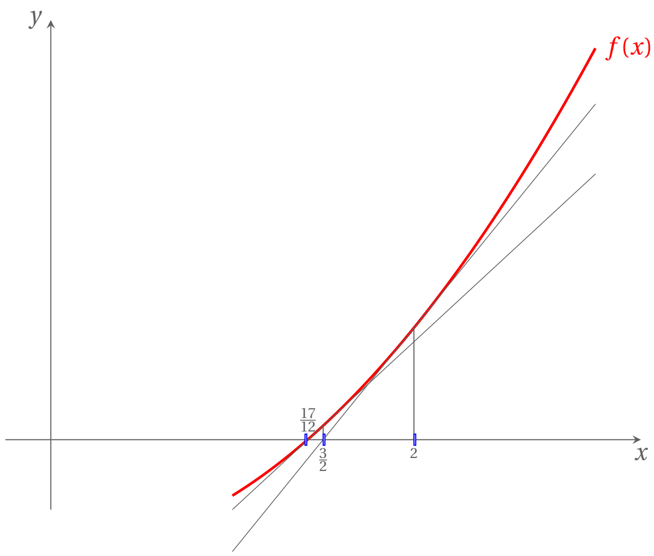
k	a_k	x_k	b_k	signe de $f(a_k)$	signe de $f(x_k)$	signe de $f(b_k)$
0	2.000000	2.596666667	3.00000	−	−	+
1	2.596666667	2.690262642	3.00000	−	−	+
2	2.690262642	2.702092263	3.00000	−	−	+
3	2.702092263	2.703541518	3.00000	−	−	+
4	2.703541518	2.703718378	3.00000	−	−	+
5	2.703718378	2.703739951	3.00000	−	−	+
6	2.703739951	2.703742582	3.00000	−	−	+

Exercice 1.2

Déterminer la suite des premiers 3 itérés des méthodes de dichotomie dans l'intervalle $[1,3]$ et de Newton avec $x_0 = 2$ pour l'approximation du zéro de la fonction $f(x) = x^2 - 2$. Combien de pas de dichotomie doit-on effectuer pour améliorer d'un ordre de grandeur la précision de l'approximation de la racine ?



(a) Méthode de la dichotomie.



(b) Méthode de Newton.

FIGURE 1.1.: Approximation du zéro de la fonction $f(x) = x^2 - 2$.

SOLUTION. On cherche les zéros de la fonction $f(x) = x^2 - 2$:

▷ Méthode de la dichotomie : en partant de $I_0 = [a, b]$, la méthode de la dichotomie produit une suite de sous-intervalles $I_k = [a_k, b_k]$ avec $I_{k+1} \subset I_k$ et tels que $f(a_k)f(b_k) < 0$. Plus précisément
  ▷ on pose $a_0 = a$, $b_0 = b$, $x_0 = \frac{a_0+b_0}{2}$,
  ▷ pour $k \geq 0$
    ▷ si $f(a_k)f(x_k) < 0$ on pose $a_{k+1} = a_k$, $b_{k+1} = x_k$ sinon on pose $a_{k+1} = x_k$, $b_{k+1} = b_k$
    ▷ et on pose $x_{k+1} = \frac{a_k+b_k}{2}$.

  Voir la figure 1.1a.

▷ Méthode de Newton :

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{x_k^2 - 2}{2x_k} = \frac{1}{2}x_k + \frac{1}{x_k}.$$

  Voir la figure 1.1b.

Donc on a le tableau suivant

	x_0	x_1	x_2	x_3
Dichotomie	2	$\frac{3}{2} = 1{,}5$	$\frac{5}{4} = 1{,}25$	$\frac{11}{8} = 1{,}375$
Newton	2	$\frac{3}{2} = 1{,}5$	$\frac{17}{12} = 1{,}41\bar{6}$	$\frac{17}{24} + \frac{12}{17} \simeq 1{,}4142156$

On rappelle qu'avec la méthode de la dichotomie, les itération s'achèvent à la m-ème étape quand $|x_m - \alpha| \leq |I_m| < \varepsilon$, où ε est une tolérance fixée et $|I_m|$ désigne la longueur de l'intervalle I_m. Clairement $I_k = \frac{b-a}{2^k}$, donc pour avoir $|x_m - \alpha| < \varepsilon$ on doit prendre

$$m \geq \log_2 \frac{b-a}{\varepsilon}.$$

Améliorer d'un ordre de grandeur la précision de l'approximation de la racine signifie avoir

$$|x_k - \alpha| = \frac{|x_j - \alpha|}{10}$$

donc on doit effectuer $k - j = \log_2(10) \simeq 3{,}3$ itérations de dichotomie.

Exercice 1.3

1. Donner la suite définissant la méthode de Newton pour la recherche d'un zéro de fonction. Justifier l'expression de la suite.

2. Écrire l'algorithme pour une convergence à 10^{-6} près.
3. Déterminer l'ordre de convergence minimale de cette suite.

SOLUTION.

1. Supposons $f \in \mathscr{C}^1$ et $f'(\alpha) \neq 0$ (c'est-à-dire α est une racine simple de f). La méthode de Newton revient à calculer le zéro de f en remplaçant localement f par sa tangente : en partant de l'équation de la tangente à la courbe $(x, f(x))$ au point $x^{(k)}$
$$y(x) = f(x^{(k)}) + f'(x^{(k)})(x - x^{(k)})$$
et en faisant comme si $x^{(k+1)}$ vérifiait $y(x^{(k+1)}) = 0$, on obtient
$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}.$$
Étant donné une valeur initiale $x^{(0)}$, cette formule permet de construire une suite $x^{(k)}$.
2. Algorithmes pour une convergence à $\varepsilon = 10^{-6}$:
Require: $x^{(0)}$
  while $|x^{(k+1)} - x^{(k)}| > 10^{-6}$ **do**
    $x^{(k+1)} \leftarrow x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}$
  end while
3. La relation précédent peut être mise sous la forme d'une itération de point fixe $x^{(k+1)} = g(x^{(k)})$ avec
$$g(x) = x - \frac{f(x)}{f'(x)}.$$
Si α est racine simple, c'est-à-dire si $f'(\alpha) \neq 0$, on trouve $g'(\alpha) = 0$ et $g''(\alpha) = \frac{f''(\alpha)}{f'(\alpha)}$: la méthode de Newton est donc d'ordre 2. Si la racine α est de multiplicité $m > 1$, alors $g'(\alpha) = 1 - \frac{1}{m}$ et la méthode n'est que d'ordre 1. Si la valeur de m est connue à priori, on peut retrouver la convergence quadratique de la méthode de Newton en modifiant la méthode comme suit :
$$x^{(k+1)} = x^{(k)} - m\frac{f(x^{(k)})}{f'(x^{(k)})}.$$

Exercice 1.4

On veut calculer le zéro de la fonction
$$f(x) = x^2 - 2$$
dans l'intervalle $[0;2]$.

1. On applique la méthode de LAGRANGE : écrire l'algorithme et l'utiliser pour remplir le tableau (on s'arrêtera au plus petit k qui vérifie $|f(x_k)| < 10^{-4}$).

k	a_k	x_k	b_k	signe de $f(a_k)$	$f(x_k)$	signe de $f(b_k)$	$\lvert x_k - \sqrt{2}\rvert$
0	0.00000	1.00000	2.00000	–	-1.00000	+	0.41421
1							
⋮							

2. On applique la méthode de NEWTON : écrire l'algorithme et l'utiliser pour remplir le tableau (on s'arrêtera au plus petit k qui vérifie $|f(x_k)| < 10^{-4}$). Le point de départ x_0 est donné.

k	x_k	$f(x_k)$	$\lvert x_k - \sqrt{2}\rvert$
0	1.00000		
1			
⋮			

SOLUTION.

1. En partant de $I_0 = [a, b]$, la méthode de LAGRANGE produit une suite de sous-intervalles $I_k = [a_k, b_k]$, $k \geq 0$, avec $I_k \subset I_{k-1}$, $k \geq 1$, et tels que $f(a_k)f(b_k) < 0$. Dans notre cas on a
  $k \leftarrow 0$
  $a_k \leftarrow 0$
  $b_k \leftarrow 2$
  $x_k \leftarrow a_k$

while $|x_k^2-2|>0.0001$ **do**
  $x_k \leftarrow \frac{a_k b_k+2}{a_k+b_k}$
  if $(a_k^2-2)(x_k^2-2)<0$ **then**
    $a_{k+1} \leftarrow a_k$
    $b_{k+1} \leftarrow x_k$
  else
    $a_{k+1} \leftarrow x_k$
    $b_{k+1} \leftarrow b_k$
  end if
  $k \leftarrow k+1$
end while

k	a_k	x_k	b_k	signe de $f(a_k)$	$\lvert f(x_k)\rvert$	signe de $f(b_k)$	$\lvert x_k-\sqrt{2}\rvert$
0	0.00000	1.00000	2.00000	–	|-1.00000|>0.0001	+	0.41421
1	1.00000	1.33333	2.00000	–	|-0.22222|>0.0001	+	0.08088
2	1.33333	1.40000	2.00000	–	|-0.04000|>0.0001	+	0.01421
3	1.40000	1.41176	2.00000	–	|-0.00692|>0.0001	+	0.00245
4	1.41176	1.41379	2.00000	–	|-0.00119|>0.0001	+	0.00042
5	1.41379	1.41414	2.00000	–	|-0.00020|>0.0001	+	0.00007
6	1.41414	1.41420	2.00000	–	|-0.00004|<0.0001	+	0.00001

2. La méthode de Newton est une méthode de point fixe avec fonction d'itération $\phi(x)=x-\frac{f(x)}{f'(x)}$ ce qui donne l'algorithme suivant :

$k \leftarrow 0$
$x_k \leftarrow 1.00000$
while $|x_k^2-2|>10^{-4}$ **do**
  $x_{k+1} \leftarrow \frac{x_k}{2}+\frac{1}{x_k}$
  $k \leftarrow k+1$
end while

k	x_k	$\lvert f(x_k)\rvert$	$\lvert x_k-\sqrt{2}\rvert$
0	1.00000	|-1.00000|>0.0001	0.41421
1	1.50000	|0.25000|>0.0001	0.08579
2	1.41667	|0.00695|>0.0001	0.00246
3	1.41422	|0.00002|<0.0001	0.00001

Exercice 1.5

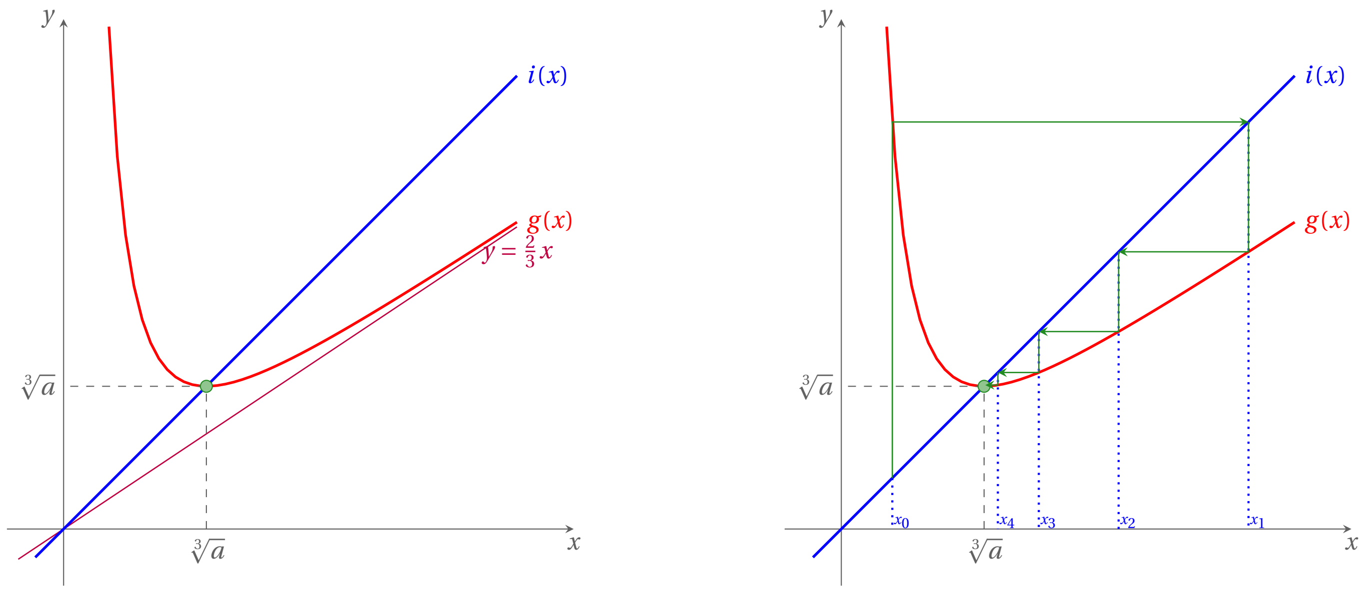
Le but de cet exercice est de calculer la racine cubique d'un nombre positif a. Soit g la fonction définie sur $\mathbb{R}_+^*$ par

$$g(x)=\frac{2}{3}x+\frac{1}{3}\frac{a}{x^2} \qquad (a>0 \text{ fixé}).$$

1. Faire l'étude complète de la fonction g.
2. Comparer g à l'identité.
3. Soit la suite $(x_n)_{n\in\mathbb{N}}$ définie par
$$x_{n+1}=g(x_n), \qquad x_0>0.$$
À l'aide des graphe de g et de l'identité sur $\mathbb{R}_+^*$, dessiner la suite $(x_n)_{n\in\mathbb{N}}$ sur l'axe des abscisses. Observer graphiquement la convergence.
4. Justifier mathématiquement la convergence observée graphiquement. En particulier, montrer que cette suite est décroissante à partir du rang 1.
5. Calculer l'ordre de convergence de la suite.
6. Écrire l'algorithme défini par la suite $(x_n)_{n\in\mathbb{N}}$ qui permet de déterminer $\sqrt[3]{a}$ à une précision de 10^{-6}.
7. Expliciter la méthode de Newton pour la recherche du zéro de la fonction f définie par $f(x)=x^3-a$. Que remarque-t-on ?

Solution.

1. Étude de la fonction $g\colon \mathbb{R}_+^* \to \mathbb{R}$ définie par $g(x)=\frac{2}{3}x+\frac{1}{3}\frac{a}{x^2}$:
   ⋆ $g(x)>0$ pour tout $x\in\mathbb{R}_+^*$;



(a) Graphe de g comparé au graphe de $i(x) = x$.

(b) Étude graphique de la convergence de la méthode de point fixe.

FIGURE 1.2.: Exercice 1.5

- ⋆ $\lim_{x\to 0^+} g(x) = \lim_{x\to+\infty} g(x) = +\infty$;
  $\lim_{x\to+\infty} \frac{g(x)}{x} = \frac{2}{3}$ et $\lim_{x\to+\infty} g(x) - \frac{2}{3}x = 0$ donc $y = \frac{2}{3}x$ est un asymptote ;
- ⋆ $g'(x) = \frac{2}{3x^3}(x^3 - a)$;
- ⋆ g est croissante sur $[\sqrt[3]{a}, +\infty[$, décroissante sur $[0, \sqrt[3]{a}]$;
- ⋆ $x = \sqrt[3]{a}$ est un minimum absolu et $g(\sqrt[3]{a}) = \sqrt[3]{a}$.

x	0		$\sqrt[3]{a}$		$+\infty$
$g'(x)$		$-$		$+$	
$g(x)$	$+\infty$	$\searrow$	$\sqrt[3]{a}$	$\nearrow$	$+\infty$

2. Graphe de g comparé au graphe de $i(x) = x$: voir la figure 1.2a. On vérifie analytiquement qu'il existe une et une seule intersection entre la courbe d'équation $y = g(x)$ et la droite d'équation $y = x$:
$$g(x) = x \iff \frac{2}{3}x + \frac{1}{3}\frac{a}{x^2} = x \iff x^3 = a.$$
3. Étude graphique de la convergence de la méthode de point fixe : voir la figure 1.2a.
4. On en déduit que pour tout $x > 0$ on a $g(x) \ge \sqrt[3]{a}$. Donc, pour tout $k > 0$, $x_k = g(x_{k-1}) \ge \sqrt[3]{a}$. Vérifions les hypothèses du théorème de point fixe qui fournit une condition suffisante de convergence de la suite :
   4.1. pour tout x dans $[\sqrt[3]{a}, +\infty[$ on a $g(x) > \sqrt[3]{a}$ donc $g([\sqrt[3]{a}, +\infty[) \subset [\sqrt[3]{a}, +\infty[$ (*i.e.* l'intervalle $\sqrt[3]{a}, +\infty[$ est stable) ;
   4.2. $g \in \mathscr{C}^1([\sqrt[3]{a}, +\infty[)$;
   4.3. pour tout x dans $[\sqrt[3]{a}, +\infty[$ on a
   $$|g'(x)| = \left|\frac{2}{3}\left(1 - \frac{a}{x^3}\right)\right| < 1$$
   donc g est contractante.

   Alors la méthode converge vers α point fixe de g. De plus, pour tout $\alpha \in [\sqrt[3]{a}, +\infty[$ on a $\alpha = g(\alpha) \iff \alpha = \sqrt[3]{a}$: la méthode permet donc de calculer de façon itérative la racine cubique de a.
5. Étant donné que
$$g'(\alpha) = 0, \qquad g''(\alpha) = \frac{2a}{\alpha^4} \neq 0$$
la méthode de point fixe converge à l'ordre 2.

Algorithm 1 Calcul de $x = g(x)$

Require: $x_0 > 0$
while $|x_{k+1} - x_k| > 10^{-6}$ **do**
  $x_{k+1} \leftarrow g(x_k)$
end while

6. Algorithme de point fixe : Quelques remarques à propos du critère d'arrêt basé sur le contrôle de l'incrément. Les itérations s'achèvent dès que $|x_{k+1} - x_k| < \varepsilon$; on se demande si cela garantît-t-il que l'erreur absolue e_{k+1} est elle aussi inférieur à ε. L'erreur absolue à l'itération $(k+1)$ peut être évaluée par un développement de Taylor au premier ordre
$$\mathrm{e}_{k+1} = |g(\alpha) - g(x_k)| = |g'(z_k)\mathrm{e}_k|$$
avec z_k compris entre α et x_k. Donc
$$|x_{k+1} - x_k| = |\mathrm{e}_{k+1} - \mathrm{e}_k| = |g'(z_k) - 1|\mathrm{e}_k \simeq |g'(\alpha) - 1|\mathrm{e}_k.$$
Puisque $g'(\alpha) = 0$, on a bien $|x_{k+1} - x_k| \simeq \mathrm{e}_k$.
7. La méthode de Newton est une méthode de point fixe avec $g(x) = x - \frac{f(x)}{f'(x)}$. Ici elle s'écrit
$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{x_k^3 - a}{3x_k^2} = x_k - \frac{1}{3}x_k + \frac{a}{3x_k^2} = \frac{2}{3}x_k + \frac{a}{3x_k^2}$$
autrement dit la méthode de point fixe assignée est la méthode de Newton (qu'on sait être d'ordre de convergence égale à 2 lorsque la racine est simple).

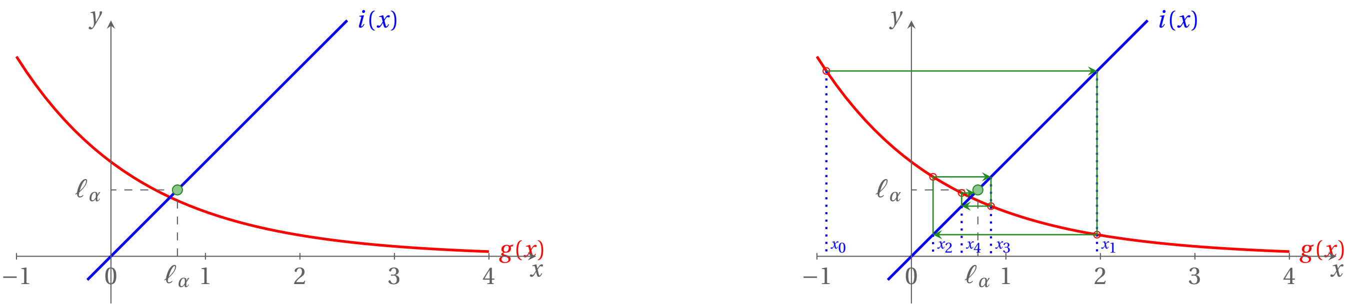
Exercice 1.6

On veut résoudre l'équation $e^{-\alpha x} = x$ avec $0 < \alpha < 1$.

1. Vérifier que cette équation admet une unique solution, notée ℓ_α, dans $\mathbb{R}$.
2. Soit $g\colon \mathbb{R} \to \mathbb{R}$ la fonction définie par $g(x) = e^{-\alpha x}$. On définit la suite récurrente
$$\begin{cases} u_0 \in \mathbb{R} \\ u_{n+1} = g(u_n). \end{cases} \tag{1.1}$$
On veut montrer que u_n converge vers ℓ_α. Pour cela, comparer d'abord le graphe de g à l'identité et observer graphiquement la convergence, ensuite justifier mathématiquement la convergence observée graphiquement.
3. Écrire la méthode de Newton pour résoudre l'équation $e^{-\alpha x} = x$ avec $0 < \alpha < 1$. Parmi la méthode de Newton et la méthode de point fixe (1.1), laquelle faut-il préférer vis-à-vis de la vitesse de convergence ?

Solution.

1. Deux méthodes (équivalentes) possibles :
   Méthode 1 : La fonction $g\colon x \mapsto e^{-\alpha x}$ est monotone décroissante, $\lim_{x\to-\infty} e^{-\alpha x} = +\infty$ et $\lim_{x\to+\infty} e^{-\alpha x} = 0$; par conséquente elle intersecte la droite d'équation $y = x$ une et une seule fois. Notons ce point ℓ_α. Comme la fonction $x \mapsto e^{-\alpha x}$ est positive pour tout $x \in \mathbb{R}$ tandis que la fonction $x \mapsto x$ est positive si et seulement si $x > 0$, on en déduit que $\ell_\alpha > 0$.
   Méthode 2 : La fonction $f\colon x \mapsto e^{-\alpha x} - x$ est monotone décroissante, $\lim_{x\to-\infty} e^{-\alpha x} - x = +\infty$ et $\lim_{x\to+\infty} e^{-\alpha x} - x = -\infty$; par le théorème des valeurs intermédiaires on conclut qu'il existe un et un seul $\ell_\alpha \in \mathbb{R}$ tel que $f(\ell_\alpha) = 0$. Comme $f(0) > 0$, on peut appliquer à nouveau le théorème des valeurs intermédiaires à l'intervalle $[0;\infty[$ et en déduire que $\ell_\alpha > 0$. De plus, comme $f(1) < e^{-1} - 1 < 0$, on peut conclure que $\ell_\alpha \in]0;1[$.
2. Le graphe de la fonction g est celui en figure 1.6. On en déduit que
   - ▷ la suite $(u_n)_n$ converge pour tout $u_0 \in \mathbb{R}$;
   - ▷ $g(\mathbb{R}) =]0;+\infty[$ et $g(]0;+\infty[) =]0;1[$ ainsi $u_1 \in]0;+\infty[$ et $u_n \in]0;1[$ pour tout $n > 1$;
   - ▷ la convergence n'est pas monotone : la sous-suite des termes d'indice pair est monotone croissante tandis que la sous-suite des termes d'indice impair est monotone décroissante (ce qui veut dire d'une part qu'on ne pourra pas utiliser les théorèmes du type «monotone+bornée=convergente» pour prouver la convergence, d'autre part on voit aussi que ni l'intervalle $[\ell_\alpha;+\infty[$ ni l'intervalle $[0;\ell_\alpha]$ sont stables) ;
   - ▷ $|g'(x)|$ n'est pas bornée pour tout $x \in \mathbb{R}$ (croissance exponentielle à $-\infty$). Plus particulièrement, $|g'(x)| < 1$ ssi $e^{\alpha x} > \alpha$ ssi $x > \ln(\alpha)/\alpha$. Comme $0 < \alpha < 1$, on conclut que $|g'(x)| < 1$ pour tout $x \geq 0$.



(a) Graphe de g comparé au graphe de $i(x) = x$.

(b) Étude graphique de la convergence de la méthode de point fixe.

FIGURE 1.3.: Exercice 1.6

Cette étude préliminaire suggère d'utiliser le théorème de point fixe dans l'intervalle $]0;+\infty[$. On a
- ▷ $g \in \mathscr{C}^\infty(]0;+\infty[)$,
- ▷ $g(]0;+\infty[) \subset]0;+\infty[$,
- ▷ $|g'(x)| < 1$ pour tout $x \in]0;+\infty[$,

on peut alors utiliser le théorème de point fixe pour conclure que la suite $(u_n)_{n\in\mathbb{N}}$ converge vers ℓ_α pour tout $u_0 \in]0;+\infty[$. Comme $g(x) \in]0;+\infty[$ pour tout $x \in \mathbb{R}$, alors $u_n \in]0;+\infty[$ pour tout $n \in \mathbb{N}^*$, on peut donc conclure que la suite $(u_n)_{n\in\mathbb{N}}$ converge vers ℓ_α pour tout $u_0 \in \mathbb{R}$.

3. Soit $f(x) = e^{-\alpha x} - x$. La méthode de Newton (qui s'applique à f et non à g) définit la suite récurrente

$$\begin{cases} u_0 \in \mathbb{R} \\ u_{n+1} = u_n - \frac{e^{-\alpha u_n} - u_n}{-\alpha e^{-\alpha u_n} - 1}. \end{cases} \tag{1.2}$$

La méthode de point fixe (1.1) n'est que d'ordre 1 car $g'(\ell_\alpha) \neq 0$ tandis que la méthode de Newton, qui est encore une méthode de point fixe, est d'ordre 2.

Exercice 1.7

Soit f une application de $\mathbb{R}$ dans $\mathbb{R}$ définie par $f(x) = \exp(x^2) - 4x^2$. On se propose de trouver les racines réelles de f.

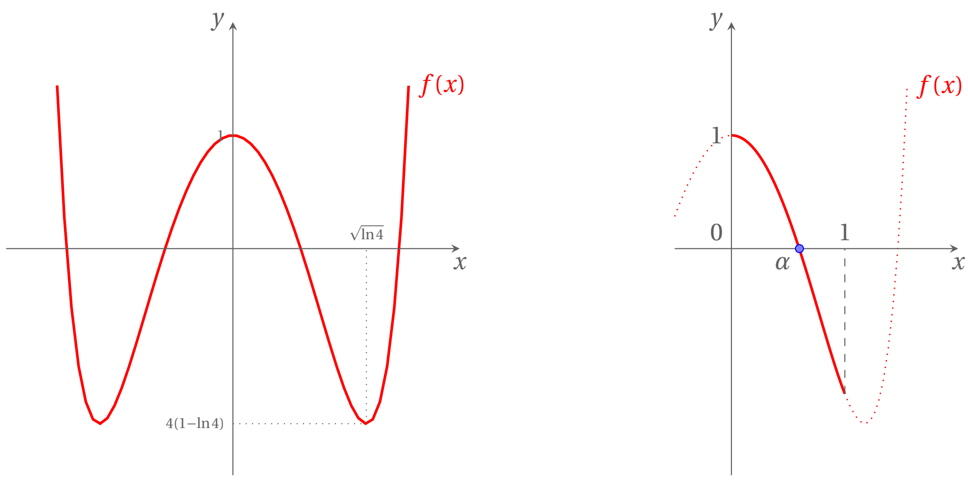
1. Situer les 4 racines de f (*i.e.* indiquer 4 intervalles disjoints qui contiennent chacun une et une seule racine).
2. Montrer qu'il y a une racine α comprise entre 0 et 1.
3. Soit la méthode de point fixe
$$\begin{cases} x_{k+1} = \phi(x_k), \\ x_0 \in]0,1[, \end{cases} \tag{1.3}$$
avec ϕ l'application de $\mathbb{R}$ dans $\mathbb{R}$ définie par $\phi(x) = \frac{\sqrt{\exp(x^2)}}{2}$. Examiner la convergence de cette méthode et en préciser l'ordre de convergence.
4. Écrire la méthode de Newton pour la recherche des zéros de la fonction f.
5. Entre la méthode de Newton et la méthode de point fixe (1.3), quelle est la plus efficace ? Justifier la réponse.

SOLUTION. On cherche les zéros de la fonction $f(x) = \exp(x^2) - 4x^2$.

1. On remarque que $f(-x) = f(x)$: la fonction est paire. On fait donc une brève étude sur $[0,+\infty[$:
   - ▷ $f(0) = 1$ et $\lim_{x\to+\infty} f(x) = +\infty$,
   - ▷ $f'(x) = 0$ pour $x = 0$ et $x = \sqrt{\ln 4}$ et on a $f(0) = 1$ et $f(\sqrt{\ln 4}) = 4(1 - \ln 4) < 0$; f est croissante pour $x > \sqrt{\ln 4}$ et décroissante pour $0 < x < \sqrt{\ln 4}$.

   On a
   - ▷ une racine dans l'intervalle $]-\infty, -\sqrt{\ln 4}[$,
   - ▷ une racine dans l'intervalle $]-\sqrt{\ln 4}, 0[$,
   - ▷ une racine dans l'intervalle $]0, \sqrt{\ln 4}[$,
   - ▷ une racine dans l'intervalle $]\sqrt{\ln 4}, \infty[$.

   Voir la figure 1.4a pour le graphe de f sur $\mathbb{R}$.
2. Puisque $f(0) = 1 > 0$ et $f(1) = e - 4 < 0$, pour le théorème des valeurs intermédiaires il existe au moins un $\alpha \in]0,1[$ tel que $f(\alpha) = 0$. Puisque $f'(x) = 2x\exp(x^2) - 8x = 2x(\exp(x^2) - 2^2) < 2x(e-4) < 0$ pour tout $x \in]0,1[$, ce α est unique. Voir la figure 1.4b.



(a) Graphe de $f(x) = \exp(x^2) - 4x^2$. (b) Zoom.

FIGURE 1.4.: Exercice 1.7

3. Étude de la convergence de la méthode (1.3) :
   3.1. pour tout x dans $]0,1[$ on a
   $$0 < \sqrt{\frac{\exp(x^2)}{4}} < \sqrt{\frac{e}{4}} < 1$$
   donc $\phi\colon]0,1[\to]0,1[$;
   3.2. $\phi \in \mathscr{C}^1(]0,1[)$;
   3.3. pour tout x dans $]0,1[$ on a
   $$|\phi'(x)| = \left|\frac{x\sqrt{\exp(x^2)}}{2}\right| = \left|x\phi(x)\right| < |x| < 1$$
   donc ϕ est contractante.

   Alors la méthode (1.3) converge vers α point fixe de ϕ. De plus, pour tout $\alpha \in]0,1[$,
   $$\alpha = \phi(\alpha) \iff 2\alpha = \sqrt{\exp(\alpha^2)} \iff 4\alpha^2 = \exp(\alpha^2) \iff f(\alpha) = 0;$$
   donc α, point fixe de ϕ, est un zéro de f.
   Étant donné que
   $$\phi'(\alpha) = \alpha\phi(\alpha) = \alpha^2 \neq 0,$$
   la méthode de point fixe (1.3) converge seulement à l'ordre 1.
4. La méthode de Newton est une méthode de point fixe avec $\phi(x) = x - \frac{f(x)}{f'(x)}$. Ici donc elle s'écrit
   $$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{\exp(x_k^2) - 4x_k^2}{2x_k\exp(x_k^2) - 8x_k} = x_k - \frac{\exp(x_k^2) - 4x_k^2}{2x_k(\exp(x_k^2) - 4)}.$$
5. Puisque α est une racine simple de f, la méthode de Newton converge à l'ordre 2 tandis que la méthode de point fixe (1.3) converge seulement à l'ordre 1 : la méthode de Newton est donc plus efficace.

Exercice 1.8

On cherche à évaluer $\sqrt{5}$ à l'aide d'un algorithme n'autorisant que les opérations élémentaires. Soit $(x_n)_{n\in\mathbb{N}}$ la suite définie par récurrence
$$\begin{cases} x_0 = 1, \\ x_{n+1} = \dfrac{10x_n}{x_n^2+5} & \forall n \in \mathbb{N}. \end{cases}$$

1. Montrer que si la suite converge, alors elle converge vers 0 ou $\sqrt{5}$.
2. Soit la fonction g définie sur $[1;\sqrt{5}]$ par $g(x) = \frac{10x}{x^2+5}$. Étudier g et la comparer à l'identité.
3. Montrer que la suite $(x_n)_{n\in\mathbb{N}}$ est croissante et majorée par $\sqrt{5}$. Conclure.
4. Déterminer l'ordre de convergence de cette suite.

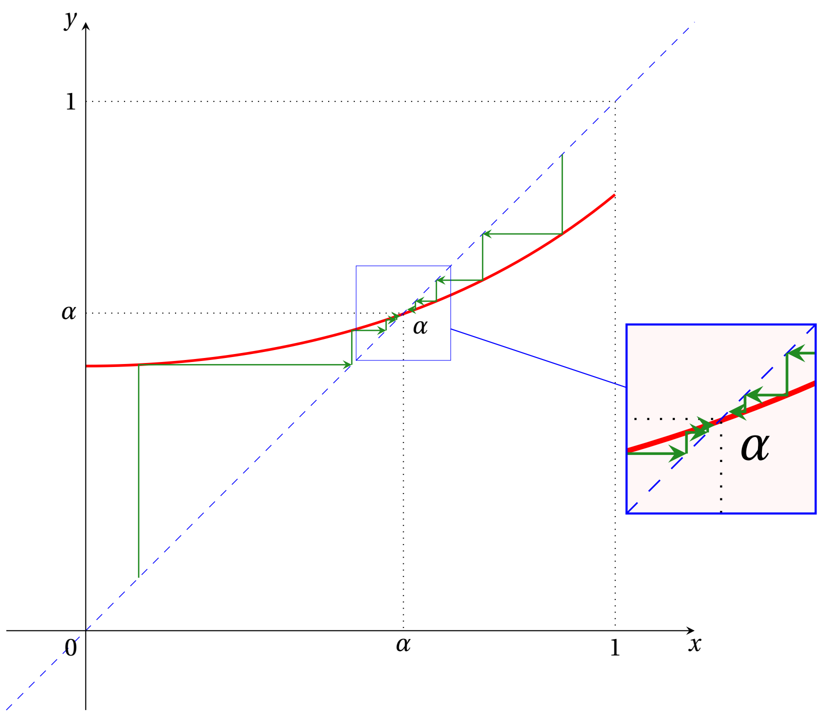


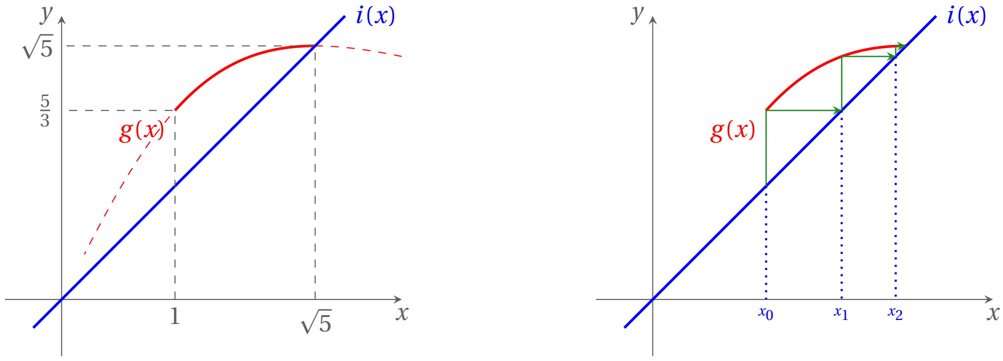
FIGURE 1.5.: Exercice 1.7 : convergence de la méthode de point fixe.

SOLUTION.

1. Supposons qu'il existe $\ell \in \mathbb{R}$ tel que $x_n \xrightarrow[n\to+\infty]{} \ell$.
   - ▷ Par définition de convergence on a $\ell = \frac{10\ell}{\ell^2+5}$ et par conséquent $\ell \in \left\{-\sqrt{5}, 0, \sqrt{5}\right\}$.
   - ▷ On prouve par récurrence que
     - ▷ si $x_0 = 0$ alors $x_n = 0$ pour tout $n \in \mathbb{N}$ donc $\ell = 0$,
     - ▷ si $x_0 > 0$ alors $x_n > 0$ pour tout $n \in \mathbb{N}$ donc $\ell \geq 0$,
     - ▷ si $x_0 < 0$ alors $x_n < 0$ pour tout $n \in \mathbb{N}$ donc $\ell \leq 0$.

     Comme $x_0 = 1 > 0$, alors $x_n > 0$ pour tout $n \in \mathbb{N}$ et $\ell \in \left\{0, \sqrt{5}\right\}$.
2. Soit la fonction g définie sur $[1;\sqrt{5}]$ par $g(x) = \frac{10x}{x^2+5}$. On étudie la fonction g :
   - ⋆ $g(x) > 0$ pour tout $x \in [1;\sqrt{5}]$;
   - ⋆ $g(1) = \frac{5}{3}$, $g(\sqrt{5}) = \sqrt{5}$;
   - ⋆ $g'(x) = -10\frac{x^2-5}{(x^2+5)^2}$;
   - ⋆ g est croissante sur $[1;\sqrt{5}[$ et $g'(\sqrt{5}) = 0$.

   Graphe de g comparé au graphe de $i(x) = x$: voir la figure 1.6a. On vérifie analytiquement qu'il existe une et une seule intersection entre la courbe d'équation $y = g(x)$ et la droite d'équation $y = x$ dans $[1;\sqrt{5}]$:
   $$g(x) = x \iff \frac{10x}{x^2+5} = x \iff x^2 = 5.$$
3. On a $g(x) \in [5/3;\sqrt{5}]$ pour tout $x \in [1;\sqrt{5}]$ et on a vu au point précédent que g est croissante et $g(\sqrt{5}) = \sqrt{5}$.
   De plus, $g(x) \geq x$ car
   $$g(x) = \frac{10x}{x^2+5} \geq \frac{10x}{(\sqrt{5})^2+5} = x,$$
   par conséquent la suite $x_{k+1} = g(x_k) \geq x_k$ est croissante.
   Comme $g(x) \leq (\sqrt{5}) = \sqrt{5}$ alors la suite $x_{k+1} = g(x_k) \leq \sqrt{5}$ est bornée. On a ainsi une suite croissante et borné, ce qui implique qu'elle converge. Comme au premier point on a montré que si elle converge vers ℓ alors $\ell \in \left\{0, \sqrt{5}\right\}$, on conclut que $x_n \xrightarrow[n\to+\infty]{} \sqrt{5}$. Pour l'étude graphique de la convergence de la méthode de point fixe voir la figure 1.6b.
   Dans ce cas, on ne peut pas utiliser le théorème de point fixe pour prouver la convergence de la suite sur l'intervalle $[1;\sqrt{5}]$. En effet
   - ▷ g est au moins de classe $\mathscr{C}^1([1;\sqrt{5}])$



(a) Graphe de g comparé au graphe de $i(x) = x$.

(b) Étude graphique de la convergence de la méthode de point fixe $x_{k+1} = g(x_k)$.

FIGURE 1.6.: Exercice 1.8

- ▷ $g([1;\sqrt{5}]) = [5/3;\sqrt{5}] \subset [1;\sqrt{5}]$
- ▷ mais $0 \le g'(x) < 1$ ssi $x \in [\sqrt{-10+5\sqrt{5}};\sqrt{5}]$ (et on a $\sqrt{-10+5\sqrt{5}} > 1$).

En revanche, on peut utiliser le théorème de point fixe pour prouver la convergence de la suite sur l'intervalle $[5/3;\sqrt{5}]$ car

- ▷ g est au moins de classe $\mathscr{C}^1([5/3;\sqrt{5}])$
- ▷ $g([5/3;\sqrt{5}]) \subset [5/3;\sqrt{5}]$
- ▷ $0 \le g'(x) < 1$ pour tout $x \in [5/3;\sqrt{5}]$.

4. Comme $g'(\sqrt{5}) = 0$ et $g''(\sqrt{5}) \neq 0$, la méthode de point fixe associée à la fonction d'itération g est d'ordre 2.

Exercice 1.9

L'objectif de cet exercice est de déterminer le zéro d'une fonction $\mathscr{C}^2(\mathbb{R},\mathbb{R})$ vérifiant $-2 < f'(x) < -1$ sur $\mathbb{R}$. On définit la suite $\{x_n\}_{n\in\mathbb{N}}$ de $\mathbb{R}$ par la récurrence suivante

$$x_{n+1} = g(x_n) = x_n + \alpha f(x_n),$$

où $\alpha > 0$ et $x_0 \in \mathbb{R}$ sont donnés.

1. Montrer que $\lim\limits_{x\to-\infty} f(x) = +\infty$ et $\lim\limits_{x\to+\infty} f(x) = -\infty$.
2. En déduire qu'il existe un unique ℓ élément de $\mathbb{R}$ tel que $f(\ell) = 0$.
3. Montrer que si $0 < \alpha < 1$, la fonction g définie par $g(x) = x + \alpha f(x)$ vérifie
$$-1 < 1 - 2\alpha < g'(x) < 1 - \alpha \qquad \text{sur } \mathbb{R}.$$
4. En déduire la convergence de la suite $\{x_n\}_{n\in\mathbb{N}}$ si $0 < \alpha < 1$.
5. La suite converge-t-elle pour $\alpha = -\frac{1}{f'(\ell)}$?
6. Donner l'ordre de convergence de la suite $\{x_n\}_{n\in\mathbb{N}}$ pour $0 < \alpha < 1$ en distinguant le cas $\alpha = \frac{1}{f'(\ell)}$.
7. Peut-on choisir $\alpha = -\frac{1}{f'(\ell)}$ d'un point de vue pratique ?
8. On choisit alors d'approcher $\alpha = -\frac{1}{f'(\ell)}$ par $\alpha_n = -\frac{1}{f'(x_n)}$ et la suite $\{x_n\}_{n\in\mathbb{N}}$ est définie par
$$x_{n+1} = g(x_n) = x_n + \alpha_n f(x_n).$$
Quel est le nom de cette méthode itérative ? Montrer que la suite $\{x_n\}_{n\in\mathbb{N}}$ converge quel que soit $x_0 \in R$.

SOLUTION.

1. Puisque f est de classe $\mathscr{C}^2(\mathbb{R},\mathbb{R})$ et $f'(x) < 0$ sur $\mathbb{R}$ alors f est monotone décroissante. De plus, $f'(x) < -1$ sur $\mathbb{R}$ donc
$$\lim_{x\to-\infty} f(x) = +\infty \qquad \lim_{x\to+\infty} f(x) = -\infty.$$
NB : seul la condition $f'(x) < -1$ permet de conclure car une fonction peut être monotone décroissante mais avoir une limite finie ! En effet, la condition $f'(x) < -1$ garantie que la fonction décroit plus vite qu'une droite comme on peut

facilement vérifier :

$$\lim_{x\to\pm\infty}\frac{f(x)}{x}=\lim_{x\to\pm\infty}\frac{f'(x)}{1}\le -1.$$

2. Puisque $\lim\limits_{x\to-\infty} f(x)=+\infty>0$ et $\lim\limits_{x\to+\infty} f(x)=-\infty<0$, pour le théorème des valeurs intermédiaires il existe au moins un $\ell\in\mathbb{R}$ tel que $f(\ell)=0$. Puisque $f'(x)<0$ pour tout $x\in\mathbb{R}$, ce ℓ est unique.
3. Considérons la fonction g définie par $g(x)=x+\alpha f(x)$ alors g est de classe $\mathscr{C}^2(\mathbb{R},\mathbb{R})$ et
$$g'(x)=1+\alpha f'(x) \qquad \text{sur } \mathbb{R}.$$
Puisque $f'(x)<-1$ et $0<\alpha<1$ on a
$$g'(x)<1-\alpha<1 \qquad \text{sur } \mathbb{R}$$
et puisque $f'(x)>-2$ et $0<\alpha<1$ alors
$$g'(x)>1-2\alpha>-1 \qquad \text{sur } \mathbb{R}.$$
Autrement dit
$$|g'(x)|<1 \qquad \text{sur } \mathbb{R}.$$
4. Soit $0<\alpha<1$. On étudie la suite
$$x_{n+1}=g(x_n)$$
et on va vérifier qu'il s'agit d'une méthode de point fixe pour le calcul du zéro ℓ de f.
   4.1. On vérifie d'abord que, si la suite converge vers un point fixe de g, ce point est bien un zéro de f (ici le réciproque est vrai aussi) : soit $\ell\in\mathbb{R}$, alors
   $$\ell=g(\ell)\iff \ell=\ell+\alpha f(\ell)\iff 0=\alpha f(\ell)\iff f(\ell)=0;$$
   4.2. vérifions maintenant que la suite converge vers un point fixe de g (et donc, grâce à ce qu'on a vu au point précédant, elle converge vers l'unique zéro de f) :
      4.2.1. on a évidemment que $g\colon\mathbb{R}\to\mathbb{R}$;
      4.2.2. on a déjà remarqué que $g\in\mathscr{C}^1(\mathbb{R},\mathbb{R})$;
      4.2.3. pour tout x dans $\mathbb{R}$ on a prouvé que $|g'(x)|<1$, *i.e.* que g est contractante.

   Alors la suite $x_{n+1}=g(x_n)$ converge vers ℓ point fixe de g et zéro de f.
5. Si $\alpha=-\frac{1}{f'(\ell)}$ alors
$$x_{n+1}=g(x_n)=x_n-\frac{f(x_n)}{f'(\ell)},$$
qui converge car $-2<f'(\ell)<-1$ ssi $\frac12<\alpha<1$ et donc on rentre dans le cas de $0<\alpha<1$.
6. Étant donné que
$$g'(\ell)=1+\alpha f'(\ell)$$
   - ▷ la méthode de point fixe converge à l'ordre 2 si $\alpha f'(\ell)=-1$,
   - ▷ la méthode de point fixe converge à l'ordre 1 si $-2<\alpha f'(\ell)<0$ mais $\alpha f'(\ell)\neq-1$,
   - ▷ la méthode de point fixe ne converge pas si $\alpha f'(\ell)<-2$ ou $\alpha f'(\ell)>0$.

   Étant donné que $-2<f'(\ell)<-1$ et que $0<\alpha<1$ on peut conclure que
   - ▷ la méthode de point fixe converge à l'ordre 2 si $\alpha=-\frac{1}{f'(\ell)}$,
   - ▷ la méthode de point fixe converge à l'ordre 1 si $\alpha\neq-\frac{1}{f'(\ell)}$.
7. D'un point de vue pratique on ne peut pas choisir $\alpha=-\frac{1}{f'(\ell)}$ car on ne connaît pas ℓ.
8. Si on choisit d'approcher $\alpha=-\frac{1}{f'(\ell)}$ par $\alpha_n=-\frac{1}{f'(x_n)}$ et on considère la suite $\{x_n\}_{n\in\mathbb{N}}$ définie par
$$x_{n+1}=g(x_n)=x_n+\alpha_n f(x_n),$$
on obtient la méthode de Newton (qui est d'ordre 2).
De plus, comme $-2<f'(x)<-1$ on rentre dans le cas $0<\alpha<1$ donc la suite $\{x_n\}_{n\in\mathbb{N}}$ converge quel que soit $x_0\in R$.

Exercice 1.10

Soit g la fonction définie sur $\mathbb{R}_+^*$ par
$$g(x)=\frac{2x^3+4x^2+10}{3x^2+8x}.$$

1. Faire l'étude complète de la fonction g. (On admettra que $x^3+4x^2-10=0$ admet comme unique solution $m \approx 1{,}36$ et que $g(m)=m$.)
2. Comparer g à l'identité.
3. Soit la suite $(x_n)_{n\in\mathbb{N}}$ définie par
$$x_{n+1}=g(x_n), \qquad x_0>0.$$
À l'aide des graphe de g et de l'identité sur $\mathbb{R}_+^*$, dessiner la suite $(x_n)_{n\in\mathbb{N}}$ sur l'axe des abscisses. Observer graphiquement la convergence. En particulier, montrer que cette suite est décroissante à partir du rang 1.
4. Expliciter (sans la vérifier) la condition nécessaire pour la convergence observée graphiquement.
5. Écrire l'algorithme défini par la suite $(x_n)_{n\in\mathbb{N}}$ qui permet de déterminer le point fixe à une précision de ε.
6. Expliciter la méthode de Newton pour la recherche du zéro de la fonction f définie par $f(x)=x^3+4x^2-10$. Que remarque-t-on ?
7. Donner l'ordre de convergence de la suite.

Solution.

1. Étude de la fonction $g\colon \mathbb{R}_+^* \to \mathbb{R}$ définie par $g(x)=\frac{2x^3+4x^2+10}{3x^2+8x}$:
   - ⋆ $g(x)>0$ pour tout $x\in\mathbb{R}_+^*$;
   - ⋆ $\lim\limits_{x\to 0^+} g(x) = \lim\limits_{x\to+\infty} g(x) = +\infty$;
     $\lim\limits_{x\to+\infty} \frac{g(x)}{x} = \frac{2}{3}$ et $\lim\limits_{x\to+\infty} g(x)-\frac{2}{3}x = -\frac{4}{9}$ donc $y=\frac{2}{3}x-\frac{4}{9}$ est un asymptote ;
   - ⋆ $g'(x)=\frac{2(3x+4)(x^3+4x^2-10)}{x^2(3x+8)^2}$;
   - ⋆ g est croissante sur $[m,+\infty[$, décroissante sur $[0,m]$ où $m\approx 1{,}36$;
   - ⋆ $x=m$ est un minimum absolu et $g(m)=m$.

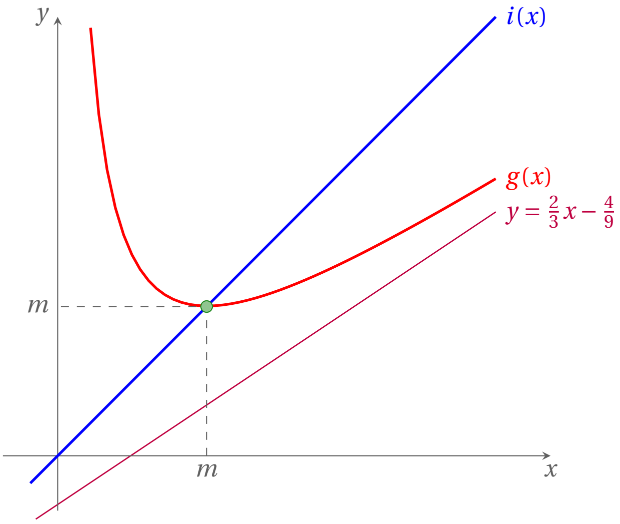
x	0		m		$+\infty$
$g'(x)$		$-$		$+$	
$g(x)$	$+\infty$	$\searrow$	m	$\nearrow$	$+\infty$

2. Graphe de g comparé au graphe de $i(x)=x$: voir la figure 1.7a. On vérifie analytiquement qu'il existe une et une seule intersection entre la courbe d'équation $y=g(x)$ et la droite d'équation $y=x$:
$$g(x)=x \iff \frac{2x^3+4x^2+10}{3x^2+8x}=x \iff x^3+4x^2-10=0 \iff x=m \iff f(x)=0.$$
3. Pour l'étude graphique de la convergence de la méthode de point fixe voir la figure 1.7b.
4. On en déduit que pour tout $x>0$ on a $g(x)\ge m$. Donc, pour tout $k>0$, $x_k=g(x_{k-1})\ge m$. Pour étudier la convergence de la méthode vérifions si on peut appliquer le théorème de point fixe :
   - 4.1. pour tout x dans $[m,+\infty[$ on a $g(x)>m$ donc $g([m,+\infty[)\subset[m,+\infty[$;
   - 4.2. $g\in\mathscr{C}^1([m,+\infty[)$;
   - 4.3. pour tout x dans $[m,+\infty[$, on a $|g'(x)|=\left|\frac{(6x^2+8x)-g(x)(6x+8)}{3x^2+8x}\right|<1$ alors g est contractante.

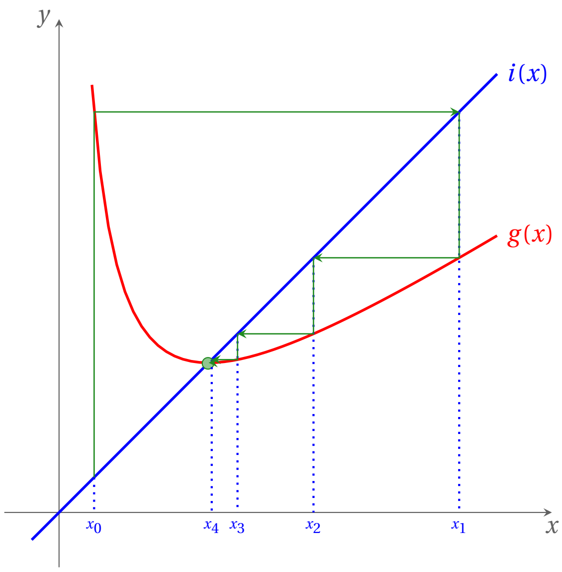
   Si les conditions précédentes sont vérifiées alors la méthode converge vers m point fixe de g. De plus, pour tout $\alpha\in[m,+\infty[$: $\alpha=g(\alpha) \iff \alpha=m$ donc le point fixe de g est racine de f.
5. Algorithme de point fixe :

Algorithm 2 Calcul de $x=g(x)$

Require: $x_0>0$
Require: $g\colon x\mapsto g(x)$
  while $|x_{k+1}-x_k|>\varepsilon$ **do**
    $x_{k+1}\leftarrow g(x_k)$
  end while



(a) Graphe de g comparé au graphe de i.



(b) Étude graphique de la convergence de la méthode de point fixe.

FIGURE 1.7.

6. La méthode de Newton est une méthode de point fixe avec $g(x) = x - \frac{f(x)}{f'(x)}$. Ici donc elle s'écrit

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{x_k^3 + 4x_k^2 - 10}{3x_k^2 + 8x_k} = g(x_k)$$

autrement dit la méthode de point fixe assignée est la méthode de Newton.

7. Étant donné que la méthode de point fixe donnée est la méthode de Newton et que la racine m de f est simple, elle converge à l'ordre 2.

Quelques remarques à propos du critère d'arrêt basé sur le contrôle de l'incrément. Les itérations s'achèvent dès que $|x_{k+1} - x_k| < \varepsilon$; on se demande si cela garantît-t-il que l'erreur absolue e_{k+1} est elle aussi inférieur à ε. L'erreur absolue à l'itération $(k+1)$ peut être évaluée par un développement de Taylor au premier ordre

$$e_{k+1} = |g(\alpha) - g(x_k)| = |g'(z_k)e_k|$$

avec z_k compris entre m et x_k. Donc

$$|x_{k+1} - x_k| = |e_{k+1} - e_k| = |g'(z_k) - 1|e_k \simeq |g'(m) - 1|e_k.$$

Puisque $g'(x) = 2\frac{3x+4}{x^2(3x+8)^2} f(x)$, alors $g'(m) = 0$ donc on a bien $|x_{k+1} - x_k| \simeq e_k$.

Exercice 1.11

On se propose de calculer $\sqrt[4]{\frac{1}{3}}$ en trouvant les racines réelles de l'application f de $\mathbb{R}$ dans $\mathbb{R}$ définie par $f(x) = x^4 - \frac{1}{3}$.

1. Situer les 2 racines de f (i.e. indiquer 2 intervalles disjoints qui contiennent chacun une et une seule racine). En particulier, montrer qu'il y a une racine α comprise entre 0 et 1.
2. Soit g la fonction définie sur $[0;1]$ par

$$g(x) = \frac{x(9x^4 + 5)}{3(5x^4 + 1)}.$$

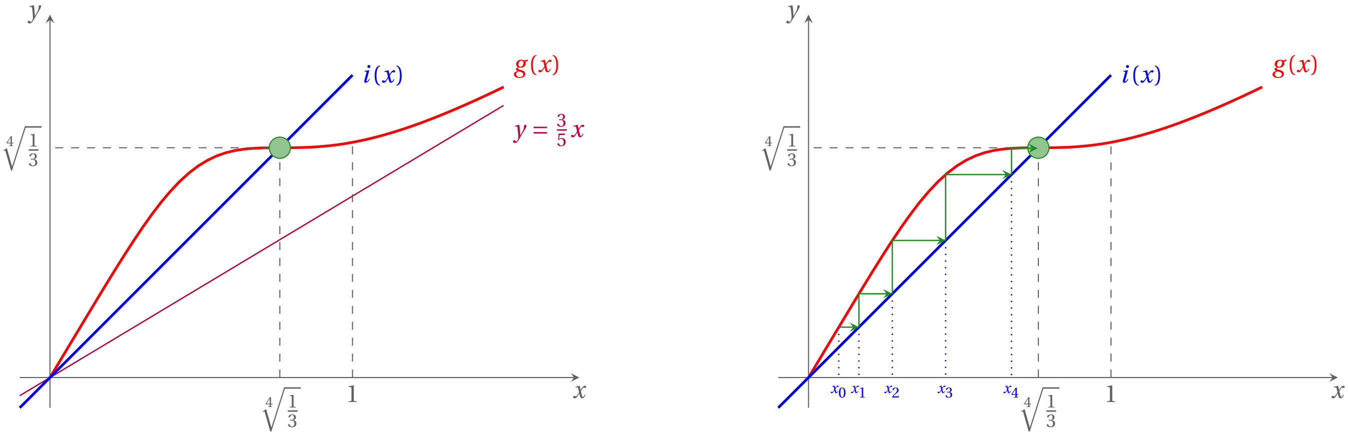
2.1. Faire l'étude complète de la fonction g et la comparer à l'identité.

2.2. Soit la suite $(x_n)_{n\in\mathbb{N}}$ définie par

$$x_{n+1} = g(x_n), \qquad x_0 \in]0;1[.$$

À l'aide des graphe de g et de l'identité sur $[0;1]$, dessiner la suite $(x_n)_{n\in\mathbb{N}}$ sur l'axe des abscisses. Observer graphiquement la convergence.

2.3. Justifier mathématiquement la convergence observée graphiquement.



(a) Graphe de g comparé au graphe de i. (b) Étude graphique de la convergence de la méthode de point fixe.

FIGURE 1.8.

2.4. Calculer l'ordre de convergence de la suite.

2.5. Écrire l'algorithme défini par la suite $(x_n)_{n\in\mathbb{N}}$ qui permet de déterminer $\sqrt[4]{\frac{1}{3}}$ à une précision de ε.

3. Expliciter la méthode de Newton pour la recherche du zéro de la fonction f.

4. Entre la méthode de Newton et la méthode de point fixe $x_{k+1} = g(x_k)$, quelle est la plus efficace ? Justifier la réponse.

SOLUTION.

1. f est paire ; comme $f'(x) = 4x^3$, f est croissante pour $x > 0$ et décroissante pour $x < 0$; puisque $f(0) < 0$ et $f(-1) = f(1) > 0$, on conclut que il n'y a que deux racines réelles distinctes : $\alpha \in]0;1[$ et $-\alpha \in]-1;0[$.

2. On étudie la fonction $g(x) = \frac{x(9x^4+5)}{3(5x^4+1)}$ pour $x \ge 0$.

2.1. ▷ $g(x) \ge 0$ pour tout $x \ge 0$ et $g(x) = 0$ ssi $x = 0$;

▷ $g'(x) = \frac{5(9x^8-6x^4+1)}{3(5x^4+1)^2} = \frac{5}{3}\left(\frac{3x^4-1}{5x^4+1}\right)^2$ donc $g'(x) \ge 0$ pour tout $x \in]0;1[$ et $g'(x) = 0$ ssi $x = \sqrt[4]{\frac{1}{3}}$. De plus, $g\left(\sqrt[4]{\frac{1}{3}}\right) = \sqrt[4]{\frac{1}{3}}$.

▷ Enfin, $g''(x) = \frac{10}{3}\frac{3x^4-1}{5x^4+1}\frac{32x^3}{(5x^4+1)^2} = \sqrt{\frac{20}{3}g'(x)}\frac{32x^3}{(5x^4+1)^2} = \frac{320x^3(3x^4-1)}{(5x^4+1)^3}$ donc $g''(x) = 0$ ssi $x = 0$ ou $x = \sqrt[4]{\frac{1}{3}}$, g est concave pour $x \in \left]0;\sqrt[4]{\frac{1}{3}}\right[$, convexe pour $x > \sqrt[4]{\frac{1}{3}}$.

▷ Pour le graphe de g comparé au graphe de $i(x) = x$ pour $x \in [0;1]$ voir la figure 1.8a.

▷ On vérifie analytiquement qu'il existe une et une seule intersection entre la courbe d'équation $y = g(x)$ et la droite d'équation $y = x$:

$$g(x) = x \iff \frac{x(9x^4+5)}{3(5x^4+1)} = x \iff 9x^4+5 = 3(5x^4+1) \iff x^4 = \frac{1}{3} \iff f(x) = 0.$$

2.2. Pour l'étude graphique de la convergence de la méthode de point fixe voir la figure 1.8b.

2.3. Étudions la convergence de la méthode. On remarque que

$$\frac{x_{k+1}}{x_k} = \frac{9x_k^4+5}{3(5x_k^4+1)} > 1 \iff x_k < \sqrt[4]{\frac{1}{3}}$$

donc la suite récurrente

$$\begin{cases} x_0 \in \left]0;\sqrt[4]{\frac{1}{3}}\right[\\ x_{k+1} = g(x_k) \end{cases}$$

est monotone croissante et majorée par $\sqrt[4]{\frac{1}{3}}$: elle est donc convergente vers $\ell \le \sqrt[4]{\frac{1}{3}}$. Comme $\ell = g(\ell)$ ssi $\ell = \sqrt[4]{\frac{1}{3}}$, on conclut qu'elle converge vers $\sqrt[4]{\frac{1}{3}}$. De même, la suite récurrente

$$\begin{cases} x_0 \in \left]\sqrt[4]{\frac{1}{3}};0\right[\\ x_{k+1} = g(x_k) \end{cases}$$

est monotone décroissante et minoré par $\sqrt[4]{\frac{1}{3}}$: elle est donc convergente vers $\ell \le \sqrt[4]{\frac{1}{3}}$. Comme $\ell = g(\ell)$ ssi $\ell = \sqrt[4]{\frac{1}{3}}$, on conclut qu'elle converge vers $\sqrt[4]{\frac{1}{3}}$.

Par conséquent, quelque soit le point initiale, la méthode de point fixe donnée converge vers $\sqrt[4]{\frac{1}{3}}$ point fixe de g (et racine de f).

Soulignons qu'on ne peut pas utiliser le théorème de point fixe pour prouver la convergence de la méthode car g n'est pas contractante sur $[0;1]$. En effet, dans $[0;1]$ on a

$$|g'(x)| < 1 \iff g'(x) < 1 \iff 5(3x^4-1)^2 < 3(5x^4+1)^2 \iff 15x^8+30x^4-1>0 \iff x^4 > -1+\sqrt{\frac{16}{15}} \in]0;1[.$$

2.4. Si on pose $\alpha = \sqrt[4]{\frac{1}{3}}$ alors $g(\alpha)=\alpha$, $g'(\alpha)=0$, $g''(\alpha)=0$ et $g'''(\alpha) = -320\alpha^2\frac{25\alpha^8-22\alpha^4+1}{(5\alpha^4+1)^4} = \frac{15\sqrt{3}}{2}$: on conclut que la suite converge à l'ordre 3.

2.5. Algorithme de point fixe :

Algorithm 3 Calcul de $x = g(x)$

Require: $x_0 > 0$
Require: $g\colon x \mapsto g(x)$
  while $|x_{k+1} - x_k| > \varepsilon$ **do**
    $x_{k+1} \leftarrow g(x_k)$
  end while

3. Entre la méthode de Newton et la méthode de point fixe $x_{k+1} = g(x_k)$, la plus efficace est la méthode de point fixe $x_{k+1} = g(x_k)$ car elle est d'ordre 3 tandis que celle de Newton n'est que d'ordre 2.

2. Interpolation

Étant donné $n+1$ points $\{(x_i, y_i)\}_{i=0}^n$, estimer $y(x)$

Étant donné $n+1$ couples (x_i, y_i), le problème consiste à trouver une fonction $\varphi = \varphi(x)$ telle que $\varphi(x_i) = y_i$; on dit alors que φ interpole $\{y_i\}$ aux nœuds $\{x_i\}$. On parle d'*interpolation polynomiale* quand φ est un polynôme, d'*approximation trigonométrique* quand φ est un polynôme trigonométrique et d'interpolation polynomiale par morceaux (ou d'interpolation par fonctions *splines*) si φ est polynomiale par morceaux. Les quantités y_i peuvent, par exemple, représenter les valeurs aux nœuds x_i d'une fonction f connue analytiquement ou des données expérimentales. Dans le premier cas, l'approximation a pour but de remplacer f par une fonction plus simple en vue d'un calcul numérique d'intégrale ou de dérivée. Dans l'autre cas, le but est d'avoir une représentation synthétique de données expérimentales dont le nombre peut être très élevé.

Polynôme de Lagrange

Considérons $n+1$ couples (x_i, y_i), le problème est de trouver un polynôme $\Pi_m(x) = a_0 + a_1 x + \dots a_m x^m \in \mathbb{P}_m$, appelé *polynôme d'interpolation* ou *polynôme interpolant*, tel quel

$$\Pi_m(x_i) = y_i, \qquad i = 0, \dots n.$$

Les points x_i sont appelés nœuds d'interpolation. Il s'agit d'un système linéaire de $n+1$ équations et $m+1$ inconnues. Si $m = n$ on a le résultat suivant :

Théorème

Étant donné $n+1$ points distincts $x_0, \dots, x_n$ et $n+1$ valeurs correspondantes $y_0, \dots, y_n$, il existe un unique polynôme $\Pi_n \in \mathbb{P}_n$ tel que $\Pi_n(x_i) = y_i$, pour $i = 0, \dots n$ qu'on peut écrire sous la forme

$$\Pi_n(x) = \sum_{i=0}^{n} y_i L_i(x) \in \mathbb{P}_n \qquad \text{où} \quad L_i(x) = \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j}.$$

Cette relation est appelée formule d'interpolation de Lagrange et les polynômes L_i sont les polynômes caractéristiques (de Lagrange).

Remarque

Si n est petit on peut calculer directement les coefficients a_0, a_1, …, a_n en résolvant le système linéaire de $n+1$ équations

$$\begin{cases} a_0 + a_1 x_0 + \dots a_n x_0^n = y_0 \\ a_0 + a_1 x_1 + \dots a_n x_1^n = y_1 \\ \dots \\ a_n + a_1 x_n + \dots a_n x_n^n = y_n \end{cases} \qquad \text{i.e.} \qquad \begin{pmatrix} 1 & x_0 & \dots & x_0^n \\ 1 & x_1 & \dots & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \dots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Erreur

Si $y_i = f(x_i)$ pour $i = 0, 1, \dots, n$, $f\colon I \to \mathbb{R}$ étant une fonction donnée de classe $\mathscr{C}^{n+1}(I)$ où I est le plus petit intervalle contenant les nœuds $\{x_i\}$, l'erreur d'interpolation au point $x \in I$ est donné par

$$E_n(x) \equiv f(x) - \Pi_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \omega_{n+1}(x)$$

où $\xi \in I$ et $\omega_{n+1}(x) \prod_{i=0}^{n}(x-x_j)$.

Exemple

Pour $n=2$ le polynôme de Lagrange s'écrit

$$\begin{aligned} P(x) = y_0 & \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} \\ + y_1 & \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} \\ + y_2 & \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \end{aligned}$$

Polynôme d'Hermite ou polynôme osculateur

On peut généraliser l'interpolation de Lagrange pour prendre en compte, en plus des valeurs nodales, les valeurs de la dérivée du polynôme interpolateur dans ces nœuds.

Considérons $n+1$ triplets (x_i, y_i, y'_i), le problème est de trouver un polynôme $\Pi_m(x) = a_0+a_1x+\dots a_mx^m \in \mathbb{P}_m$ tel quel

$$\begin{cases} \Pi_m(x_i) = y_i, \\ \Pi'_m(x_i) = y'_i, \end{cases} \qquad i = 0, \dots n.$$

Il s'agit d'un système linéaire de $2(n+1)$ équations et $m+1$ inconnues. Si $m = 2n+1$ on a le résultat suivant :

Théorème

Étant donné $n+1$ points distincts $x_0, \dots, x_n$ et $n+1$ couples correspondantes $(y_0, y'_0), \dots, (y_n, y'_n)$, il existe un unique polynôme $\Pi_N \in \mathbb{P}_N$ tel que $\Pi_N(x_i) = y_i$ et $\Pi'_N(x_i) = y'_i$, pour $i = 0, \dots n$ qu'on peut écrire sous la forme

$$Q(x) = \sum_{i=0}^{n} y_i A_i(x) + y'_i B_i(x) \in \mathbb{P}_N \qquad \text{où} \qquad \begin{cases} L_i(x) & = \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j}, \\ c_i & = \sum_{\substack{j=0\\ j\neq i}}^{n} \frac{1}{x_i-x_j}, \\ A_i(x) & = (1-2(x-x_i)c_i)(L_i(x))^2, \\ B_i(x) & = (x-x_i)(L_i(x))^2, \\ N & = 2n+1. \end{cases}$$

qu'on peut réécrire comme

$$Q(x) = \sum_{i=0}^{n} (y_i D_i(x) + y'_i(x-x_i))(L_i(x))^2 \qquad \text{où} \qquad \begin{cases} L_i(x) & = \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j}, \\ c_i & = \sum_{\substack{j=0\\ j\neq i}}^{n} \frac{1}{x_i-x_j}, \\ D_i(x) & = 1-2(x-x_i)c_i, \\ N & = 2n+1. \end{cases}$$

Cette relation est appelée formule d'interpolation de Hermite.

Remarque

Si n est petit on peut calculer directement les coefficients $a_0, a_1, \dots, a_N$ en résolvant le système linéaire de $N+1 = 2n+2$

équations

$$\begin{cases} a_0 + a_1 x_0 + \dots a_N x_0^N = y_0 \\ a_0 + a_1 x_1 + \dots a_N x_1^N = y_1 \\ \dots \\ a_n + a_1 x_n + \dots a_N x_n^N = y_n \\ a_1 + a_2 x_0 + \dots N a_N x_0^{N-1} = y_0' \\ a_1 + a_2 x_0 + \dots N a_N x_1^{N-1} = y_1' \\ \dots \\ a_n + a_1 x_n + \dots N a_N x_n^{N-1} = y_n' \end{cases} \qquad \text{i.e.} \qquad \begin{pmatrix} 1 & x_0 & \dots & x_0^N \\ 1 & x_1 & \dots & x_1^N \\ \vdots & & \vdots & \vdots \\ 1 & x_n & \dots & x_n^N \\ 0 & x_0 & \dots & N x_0^{N-1} \\ 0 & x_1 & \dots & N x_1^{N-1} \\ \vdots & & \vdots & \vdots \\ 0 & x_n & \dots & N x_n^{N-1} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ \vdots \\ \vdots \\ a_N \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \\ y_0' \\ y_1' \\ \vdots \\ y_n' \end{pmatrix}$$

Exemple

Pour $n = 2$ le polynôme de Hermite s'écrit

$$\begin{aligned} Q(x) = {} & y_0 \left(1 - 2(x - x_0)\left(\frac{1}{x_0 - x_1} + \frac{1}{x_0 - x_2}\right)\right)\left(\frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}\right)^2 + y_0'(x - x_0)\left(\frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}\right)^2 \\ & + y_1 \left(1 - 2(x - x_1)\left(\frac{1}{x_1 - x_0} + \frac{1}{x_1 - x_2}\right)\right)\left(\frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}\right)^2 + y_1'(x - x_1)\left(\frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}\right)^2 \\ & + y_2 \left(1 - 2(x - x_2)\left(\frac{1}{x_2 - x_0} + \frac{1}{x_2 - x_1}\right)\right)\left(\frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}\right)^2 + y_2'(x - x_2)\left(\frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}\right)^2, \end{aligned}$$

qu'on peut réécrire comme

$$\begin{aligned} Q(x) = {} & \left(y_0 \left(1 - 2(x - x_0)\left(\frac{1}{x_0 - x_1} + \frac{1}{x_0 - x_2}\right)\right) + y_0'(x - x_0)\right)\left(\frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}\right)^2 \\ & + \left(y_1 \left(1 - 2(x - x_1)\left(\frac{1}{x_1 - x_0} + \frac{1}{x_1 - x_2}\right)\right) + y_1'(x - x_1)\right)\left(\frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}\right)^2 \\ & + \left(y_2 \left(1 - 2(x - x_2)\left(\frac{1}{x_2 - x_0} + \frac{1}{x_2 - x_1}\right)\right) + y_2'(x - x_2)\right)\left(\frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}\right)^2. \end{aligned}$$

Algorithmes

LAGRANGE :

Require: $t, n, \{(x_i, y_i)\}_{i=0}^n$
  $p \leftarrow 0$
  for $i = 0$ to n **do**
    $L_i \leftarrow 1$
    for $j = 0$ to n **do**
      if $j \neq i$ **then**
        $L_i \leftarrow \dfrac{t - x_j}{x_i - x_j} \times L_i$
      end if
    end for
    $p \leftarrow p + y_i \times L_i$
  end for
  return p

HERMITE :

Require: $t, n, \{(x_i, y_i, y_i')\}_{i=0}^n$
  $p \leftarrow 0$
  for $i = 0$ to n **do**
    $L_i \leftarrow 1$
    for $j = 0$ to n **do**
      if $j \neq i$ **then**
        $L_i \leftarrow \dfrac{t - x_j}{x_i - x_j} \times L_i$
        $c_i \leftarrow \dfrac{1}{x_i - x_j} + c_i$
      end if
    end for
    $p \leftarrow p + \big(y_i \times (1 - 2(t - x_i) \times c_i) + y_i' \times (t - x_i)\big) \times L_i^2$
  end for
  return p

Splines

C'est une méthode d'interpolation par morceaux possédant des propriétés de régularité globale.

Définition

Étant donné $n+1$ points distincts $x_0, \dots, x_n$ de $[a;b]$ avec $a = x_0 < x_1 < \dots < x_n = b$, la fonction $s_k(x)\colon [a;b] \to \mathbb{R}$ est une spline de degré k relative aux nœuds $\{x_i\}$ si

$$\begin{cases} s_k(x)|_{[x_i;x_i+1]} \in \mathbb{P}_k, & i = 0, 1, \dots, n-1, \\ s_k \in \mathscr{C}^{k-1}([a;b]). \end{cases}$$

Évidemment tout polynôme de degré k est une spline, mais en pratique une spline est constituée de polynômes différents sur chaque sous-intervalle. Il peut donc y avoir des discontinuité de la dérivée k-ième aux nœuds internes $x_1, \dots, x_{n-1}$.

Splines linéaires

Étant donné $n+1$ points distincts $x_0, \dots, x_n$ de $[a;b]$ avec $a = x_0 < x_1 < \dots < x_n = b$, la fonction $\ell(x)\colon [a;b] \to \mathbb{R}$ est une spline linéaire relative aux nœuds $\{x_i\}$ si

$$\begin{cases} \ell(x)|_{[x_i;x_i+1]} \in \mathbb{P}_k, & i = 0, 1, \dots, n-1, \\ \ell \in \mathscr{C}^{0}([a;b]). \end{cases}$$

Autrement dit, dans chaque sous-intervalle $[x_i; x_i+1]$, la fonction ℓ est le segment qui connecte le point (x_i, y_i) au point (x_{i+1}, y_{i+1}) ; elle s'écrit donc

$$\ell(x)|_{[x_i;x_i+1]} = y_i + \frac{y_{i+1} - y_i}{x_{i+1} - x_i}(x - x_i)$$

Erreur

Si $y_i = f(x_i)$ pour $i = 0, 1, \dots, n$, $f\colon [a;b] \to \mathbb{R}$ étant une fonction donnée de classe $\mathscr{C}^2([a;b])$, l'erreur d'interpolation au point $x \in [a;b]$ est donné par

$$\max_{x \in I} |f(x) - \ell(x)| \le \frac{(b-a)^2}{8} \max_{x \in I} |f''(x)|.$$

Exercices

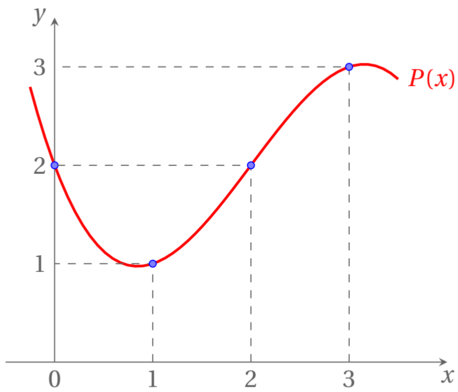
Exercice 2.1

Construire le polynôme de Lagrange P qui interpole les points $(0,2)$, $(1,1)$, $(2,2)$ et $(3,3)$.

Solution. Le polynôme d'interpolation de Lagrange de degré n sur l'ensemble des $n+1$ points $\{(x_i, y_i)\}_{i=0}^{n}$ s'écrit

$$p_n(x) = \sum_{i=0}^{n}\left(y_i \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j}\right).$$

Ici $n=3$ donc on a



$$\begin{aligned}
P(x) &= y_0\frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + y_1\frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}\\
&\quad + y_2\frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + y_3\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} =\\
&= 2\frac{(x-1)(x-2)(x-3)}{(0-1)(0-2)(0-3)} + \frac{(x-0)(x-2)(x-3)}{(1-0)(1-2)(1-3)}\\
&\quad + 2\frac{(x-0)(x-1)(x-3)}{(2-0)(2-1)(2-3)} + 3\frac{(x-0)(x-1)(x-2)}{(3-0)(3-1)(3-2)} =\\
&= \frac{(x-1)(x-2)(x-3)}{-3} + \frac{x(x-2)(x-3)}{2}\\
&\quad - x(x-1)(x-3) + \frac{x(x-1)(x-2)}{2} = -\frac{1}{3}x^3 + 2x^2 - \frac{8}{3}x + 2.
\end{aligned}$$

Sinon, comme on cherche un polynôme de degré 3, il s'agit de trouver les 4 coefficients a_0, a_1, a_2 et a_3 solution du système linéaire

$$\begin{cases} a_0 + a_1\cdot 0 + a_2\cdot 0^2 + a_3\cdot 0^3 = 2\\ a_0 + a_1\cdot 1 + a_2\cdot 1^2 + a_3\cdot 1^3 = 1\\ a_0 + a_1\cdot 2 + a_2\cdot 2^2 + a_3\cdot 2^3 = 2\\ a_0 + a_1\cdot 3 + a_2\cdot 3^2 + a_3\cdot 3^3 = 3\end{cases} \qquad \text{i.e.} \qquad \begin{pmatrix}1&0&0&0\\1&1&1&1\\1&2&4&8\\1&3&9&27\end{pmatrix}\begin{pmatrix}a_0\\a_1\\a_2\\a_3\end{pmatrix} = \begin{pmatrix}2\\1\\2\\3\end{pmatrix}$$

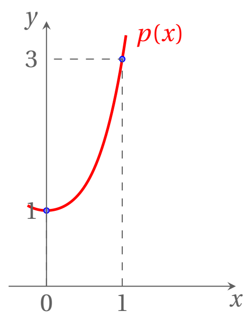
Exercice 2.2

Trouver le polynôme de l'espace vectoriel $\text{Vec}\{1+x^2, x^4\}$ qui interpole les points $(0,1)$ et $(1,3)$.

Solution. Il s'agit de trouver un polynôme $p(x)$ qui soit combinaison linéaire des deux polynômes assignés (*i.e.* $p(x) = \alpha(1+x^2) + \beta(x^4)$) et qui interpole les deux points $(0,1)$ et $(1,3)$:

$$\begin{cases} p(0) = 1,\\ p(1) = 3,\end{cases} \qquad \Leftrightarrow \qquad \begin{cases} \alpha(1+0^2) + \beta(0^4) = 1,\\ \alpha(1+1^2) + \beta(1^4) = 3,\end{cases}$$

d'où $\alpha = 1$ et $\beta = 1$. Le polynôme cherché est donc le polynôme $p(x) = 1 + x^2 + x^4$.



Exercice 2.3

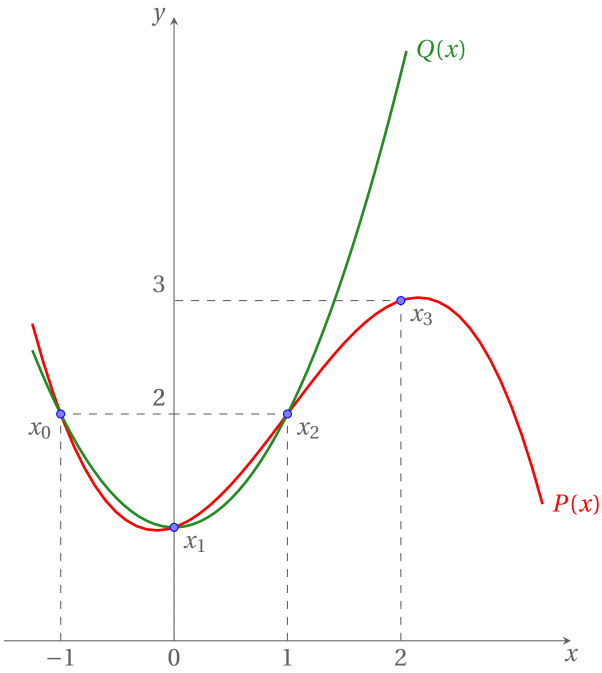
1. Construire le polynôme de Lagrange P qui interpole les points $(-1,2)$, $(0,1)$, $(1,2)$ et $(2,3)$.

2. Soit Q le polynôme de Lagrange qui interpole les points $(-1,2)$, $(0,1)$, $(1,2)$. Montrer qu'il existe un réel λ tel que :
$$Q(x)-P(x)=\lambda(x+1)x(x-1).$$

SOLUTION. Le polynôme d'interpolation de Lagrange de degré n sur l'ensemble des $n+1$ points $\{(x_i,y_i)\}_{i=0}^n$ s'écrit
$$p_n(x)=\sum_{i=0}^n\left(y_i\prod_{\substack{j=0\\j\neq i}}^n\frac{x-x_j}{x_i-x_j}\right).$$

1. Ici $n=3$ donc on a
$$\begin{aligned}P(x)&=y_0\frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)}+y_1\frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}\\&+y_2\frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)}+y_3\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}\\&=\frac{x(x-1)(x-2)}{-3}+\frac{(x+1)(x-1)(x-2)}{2}-(x+1)x(x-2)+\frac{(x+1)x(x-1)}{2}=\\&=-\frac{1}{3}x^3+x^2+\frac{1}{3}x+1.\end{aligned}$$

2. Par construction
$$\begin{aligned}Q(-1)&=P(-1),\\Q(0)&=P(0),\\Q(1)&=P(1),\end{aligned}$$
donc le polynôme $Q(x)-P(x)$ s'annule en -1, en 0 et en 1, ceci signifie qu'il existe un polynôme $R(x)$ tel que
$$Q(x)-P(x)=R(x)(x+1)x(x-1).$$
Puisque $P(x)$ a degré 3 et $Q(x)$ a degré 2, le polynôme $Q(x)-P(x)$ a degré 3, donc le polynôme $R(x)$ qu'on a mis en facteur a degré 0 (*i.e.* $R(x)$ est une constante).
Si on n'a pas remarqué ça, on peut tout de même faire tous les calculs : dans ce cas $n=2$ donc on a
$$\begin{aligned}Q(x)&=y_0\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}+y_1\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}+y_2\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}\\&=x(x-1)-(x+1)(x-1)+(x+1)x\\&=x^2+1.\end{aligned}$$
Ainsi
$$\begin{aligned}Q(x)-P(x)&=y_0\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}\left[1-\frac{x-x_3}{x_0-x_3}\right]+y_1\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}\left[1-\frac{x-x_3}{x_1-x_3}\right]\\&+y_2\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}\left[1-\frac{x-x_3}{x_2-x_3}\right]-y_3\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}\\&=-y_0\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)}-y_1\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}\\&-y_2\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)}-y_3\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}\\&=-\left[\frac{y_0}{(x_0-x_1)(x_0-x_2)(x_0-x_3)}+\frac{y_1}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}\right.\\&\left.+\frac{y_2}{(x_2-x_0)(x_2-x_1)(x_2-x_3)}+\frac{y_3}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}\right](x-x_0)(x-x_1)(x-x_2)\\&=\frac{(x+1)x(x-1)}{3}\end{aligned}$$
et $\lambda=\frac{1}{3}$. Sinon directement
$$Q(x)-P(x)=x^2+1+\frac{1}{3}x^3-x^2+\frac{1}{3}x-1=\frac{1}{3}x^3+\frac{1}{3}x=\frac{(x+1)x(x-1)}{3}=\lambda x(x+1)(x-1)$$
avec $\lambda=\frac{1}{3}$.



Exercice 2.4

1. Construire le polynôme de Lagrange P qui interpole les trois points $(-1, e)$, $(0,1)$ et $(1, e)$.
2. Sans faire de calculs, donner l'expression du polynôme de Lagrange Q qui interpole les trois points $(-1,-1)$, $(0,0)$ et $(1,-1)$.
3. Trouver le polynôme de l'espace vectoriel $\text{Vec}\{1, x, x^2\}$ qui interpole les trois points $(-1,-1)$, $(0,0)$ et $(1,-1)$.

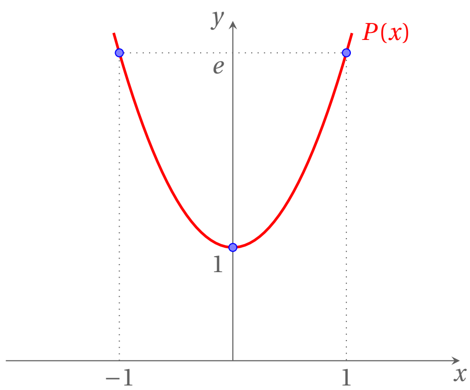
Solution.

1. Le polynôme d'interpolation de Lagrange de degré n sur l'ensemble des $n+1$ points $\{(x_i, y_i)\}_{i=0}^n$ s'écrit

$$p_n(x) = \sum_{i=0}^{n}\left(y_i \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j}\right).$$

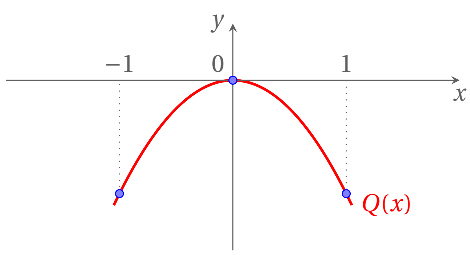
Ici $n=2$ donc on a

$$\begin{aligned}P(x) &= y_0\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + y_1\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + y_2\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} =\\ &= e\frac{x(x-1)}{2} - (x+1)(x-1) + e\frac{(x+1)x}{2} =\\ &= (e-1)x^2+1.\end{aligned}$$



2. Il suffit de changer les coefficients y_i dans l'expression précédente :

$$Q(x) = -\frac{x(x-1)}{2} - \frac{(x+1)x}{2} = -x^2.$$



3. Il s'agit de trouver un polynôme $p(x)$ qui soit combinaison linéaire des deux polynômes assignés (*i.e.* $p(x) = \alpha + \beta x + \gamma x^2$) et qui interpole les trois points $(-1,-1)$, $(0,0)$ et $(1,-1)$:

$$\begin{cases} p(-1) = 1, \\ p(0) = 0, \\ p(1) = -1, \end{cases} \qquad \Leftrightarrow \qquad \begin{cases} \alpha - \beta + \gamma = -1, \\ \alpha = 0, \\ \alpha + \beta + \gamma = -1, \end{cases}$$

d'où $\alpha = 0$, $\beta = 0$ et $\gamma = -1$. Le polynôme cherché est donc le polynôme $p(x) = -x^2$.

Exercice 2.5

1. Construire le polynôme de Lagrange P qui interpole les points $(-1,1)$, $(0,1)$, $(1,2)$ et $(2,3)$.
2. Soit Q le polynôme de Lagrange qui interpole les points $(-1,1)$, $(0,1)$, $(1,2)$. Montrer qu'il existe un réel λ tel que :

$$Q(x) - P(x) = \lambda(x+1)x(x-1).$$

Solution. Le polynôme d'interpolation de Lagrange de degré n sur l'ensemble des $n+1$ points $\{(x_i, y_i)\}_{i=0}^n$ s'écrit

$$p_n(x) = \sum_{i=0}^n \left(y_i \prod_{\substack{j=0\\ j\neq i}}^n \frac{x - x_j}{x_i - x_j} \right).$$

1. Ici $n = 3$ donc on a

$$\begin{aligned} P(x) &= y_0 \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \\ &\quad + y_2 \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \\ &= \frac{x(x-1)(x-2)}{-6} + \frac{(x+1)(x-1)(x-2)}{2} - (x+1)x(x-2) + \frac{(x+1)x(x-1)}{2} = \\ &= -\frac{1}{6}x^3 + \frac{1}{2}x^2 + \frac{2}{3}x + 1. \end{aligned}$$

2. Par construction

$$\begin{aligned} Q(-1) &= P(-1), \\ Q(0) &= P(0), \\ Q(1) &= P(1), \end{aligned}$$

donc le polynôme $Q(x) - P(x)$ s'annule en -1, en 0 et en 1, ceci signifie qu'il existe un polynôme $R(x)$ tel que

$$Q(x) - P(x) = R(x)(x+1)x(x-1).$$

Puisque $P(x)$ a degré 3 et $Q(x)$ a degré 2, le polynôme $Q(x) - P(x)$ a degré 3, donc le polynôme $R(x)$ qu'on a mis en facteur a degré 0 (*i.e.* $R(x)$ est une constante).
Si on n'a pas remarqué ça, on peut tout de même faire tous les calculs : dans ce cas $n = 2$ donc on a

$$\begin{aligned} Q(x) &= y_0 \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + y_1 \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + y_2 \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \\ &= \frac{x(x-1)}{2} - (x+1)(x-1) + (x+1)x \\ &= \frac{1}{2}x^2 + \frac{1}{2}x + 1. \end{aligned}$$

Ainsi

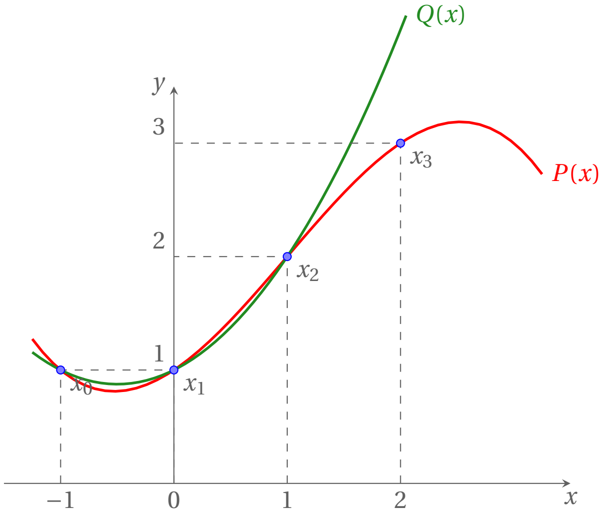
$$\begin{aligned} Q(x) - P(x) &= y_0 \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} \left[1 - \frac{x-x_3}{x_0-x_3} \right] + y_1 \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} \left[1 - \frac{x-x_3}{x_1-x_3} \right] \\ &\quad + y_2 \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \left[1 - \frac{x-x_3}{x_2-x_3} \right] - y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \end{aligned}$$

$$
\begin{aligned}
&= -y_0\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} - y_1\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}\\
&\quad - y_2\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} - y_3\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}\\
&= -\left[\frac{y_0}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + \frac{y_1}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}\right.\\
&\quad \left.+ \frac{y_2}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + \frac{y_3}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}\right](x-x_0)(x-x_1)(x-x_2)\\
&= \frac{(x+1)x(x-1)}{6}
\end{aligned}
$$

et $\lambda = \frac{1}{6}$. Sinon directement

$$
Q(x)-P(x) = \frac{1}{2}x^2+\frac{1}{2}x+1+\frac{1}{6}x^3-\frac{1}{2}x^2-\frac{2}{3}x-1 = \frac{1}{6}x^3-\frac{1}{6}x = \frac{1}{6}x(x^2-1) = \lambda x(x+1)(x-1)
$$

avec $\lambda = \frac{1}{6}$.



Exercice 2.6

1. Construire le polynôme de Lagrange P qui interpole les trois points $(-1,\alpha)$, $(0,\beta)$ et $(1,\alpha)$ où α et β sont des réels.
2. Si $\alpha = \beta$, donner le degré de P.
3. Montrer que P est pair. Peut-on avoir P de degré 1 ?

Solution.

1. Construire le polynôme de Lagrange P qui interpole les trois points $(-1,\alpha)$, $(0,\beta)$ et $(1,\alpha)$ où α et β sont des réels.
   Le polynôme d'interpolation de Lagrange de degré n sur l'ensemble des $n+1$ points $\{(x_i,y_i)\}_{i=0}^n$ s'écrit

$$
p_n(x) = \sum_{i=0}^{n}\left(y_i\prod_{\substack{j=0\\ j\neq i}}^{n}\frac{x-x_j}{x_i-x_j}\right).
$$

Ici $n=2$ donc on a

$$
\begin{aligned}
P(x) &= y_0\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + y_1\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + y_2\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}+\\
&= \alpha\frac{x(x-1)}{2} + \beta\frac{(x+1)(x-1)}{-1} + \alpha\frac{(x+1)x}{2} =\\
&= \frac{\alpha}{2}x(x-1) - \beta(x+1)(x-1) + \frac{\alpha}{2}x(x+1)\\
&= (\alpha-\beta)x^2+\beta.
\end{aligned}
$$

2. Si $\alpha=\beta$, $P=\alpha$ qui est un polynôme de degré 0.
3. $P(-x)=P(x)$ donc P est pair. Donc P ne peut pas être de degré 1 car un polynôme de degré 1 est de la forme a_0+a_1x qui ne peut pas être pair.

Exercice 2.7

Soit f une fonction de classe $\mathscr{C}^1([-1,1])$ et p le polynôme interpolateur d'Hermite (de degré ≤ 3) de f vérifiant

$$p(-1)=f(-1),\qquad p'(-1)=f'(-1),\qquad p(1)=f(1),\qquad p'(1)=f'(1).$$

Écrire le polynôme p.

Solution. On a deux méthodes pour calculer le polynôme interpolateur d'Hermite :

Première méthode : le polynôme interpolateur d'Hermite s'écrit

$$p(x)=\sum_{i=0}^{n}\left\{\left[y_i(1-2(x-x_i)c_i)+y_i'(x-x_i)\right]\prod_{\substack{j=0\\j\neq i}}^{n}\frac{(x-x_j)^2}{(x_i-x_j)^2}\right\}\qquad\text{où}\qquad c_i=\sum_{\substack{j=0\\j\neq i}}^{n}\frac{1}{x_i-x_j}.$$

Pour $n=1$ on a alors

$$\begin{aligned}p(x)&=y_0\left(1-2(x-x_0)\left(\frac{1}{x_0-x_1}\right)\right)\left(\frac{(x-x_1)}{(x_0-x_1)}\right)^2+y_0'(x-x_0)\left(\frac{(x-x_1)}{(x_0-x_1)}\right)^2\\&+y_1\left(1-2(x-x_1)\left(\frac{1}{x_1-x_0}\right)\right)\left(\frac{(x-x_0)}{(x_1-x_0)}\right)^2+y_1'(x-x_1)\left(\frac{(x-x_0)}{(x_1-x_0)}\right)^2.\end{aligned}$$

Dans notre cas $x_0=-1$, $x_1=1$, $y_0=f(-1)$, $y_1=f(1)$, $y_0'=f'(-1)$, $y_1'=f'(1)$ donc

$$\begin{aligned}p(x)&=\frac{1}{4}\left[f(-1)(x+2)(x-1)^2+f'(-1)(x+1)(x-1)^2+f(1)(2-x)(x+1)^2+f'(1)(x-1)(x+1)^2\right]\\&=\frac{1}{4}\left[f(-1)(x^3-3x+2)+f'(-1)(x^3-x^2-x+1)+f(1)(-x^3+3x+2)+f'(1)(x^3+x^2-x-1)\right]\\&=\frac{2f(-1)+f'(-1)+2f(1)-f'(1)}{4}+\frac{3f(1)-3f(-1)-f'(-1)-f'(1)}{4}x\\&+\frac{f'(1)-f'(-1)}{4}x^2+\frac{f(-1)+f'(-1)-f(1)+f'(1)}{4}x^3.\end{aligned}$$

Le polynôme interpolateur d'Hermite est donc le polynôme

$$p(x)=\alpha+\beta x+\gamma x^2+\delta x^3$$

où

$$\begin{aligned}\alpha&=\frac{2f(-1)+2f(1)+f'(-1)-f'(1)}{4}, & \beta&=\frac{-3f(-1)+3f(1)-f'(-1)-f'(1)}{4},\\ \gamma&=\frac{-f'(-1)+f'(1)}{4}, & \delta&=\frac{f(-1)-f(1)+f'(-1)+f'(1)}{4}.\end{aligned}$$

Deuxième méthode : le polynôme interpolateur d'Hermite est un polynôme de degré $2n+1$. On cherche donc un polynôme

$$p(x)=\alpha+\beta x+\gamma x^2+\delta x^3$$

tel que

$$p(-1)=f(-1),\qquad p'(-1)=f'(-1),\qquad p(1)=f(1),\qquad p'(1)=f'(1),$$

c'est-à-dire tel que

$$\begin{cases}\alpha-\beta+\gamma-\delta=f(-1),\\ \alpha+\beta+\gamma+\delta=f(1),\\ \beta-2\gamma+3\delta=f'(-1),\\ \beta+2\gamma+3\delta=f'(1).\end{cases}$$

En utilisant la méthode d'élimination de Gauss on obtient :

$$
[\mathbb{A}|\mathbf{b}] = \left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & f(-1) \\ 1 & 1 & 1 & 1 & f(1) \\ 0 & 1 & -2 & 3 & f'(-1) \\ 0 & 1 & 2 & 3 & f'(1) \end{array}\right) \xrightarrow{L_2 \leftarrow L_2 - L_1} \left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & f(-1) \\ 0 & 2 & 0 & 2 & f(1)-f(-1) \\ 0 & 1 & -2 & 3 & f'(-1) \\ 0 & 1 & 2 & 3 & f'(1) \end{array}\right)
$$

$$
\xrightarrow[L_4 \leftarrow L_4 - \frac{1}{2} L_3]{L_3 \leftarrow L_3 - \frac{1}{2} L_2} \left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & f(-1) \\ 0 & 2 & 0 & 2 & f(1)-f(-1) \\ 0 & 0 & -2 & 2 & f'(-1) - \frac{f(1)-f(-1)}{2} \\ 0 & 0 & 2 & 2 & f'(1) - \frac{f(1)-f(-1)}{2} \end{array}\right) \xrightarrow{L_4 \leftarrow L_4 + L_3} \left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & f(-1) \\ 0 & 2 & 0 & 2 & f(1)-f(-1) \\ 0 & 0 & -2 & 2 & f'(-1) - \frac{f(1)-f(-1)}{2} \\ 0 & 0 & 0 & 4 & f'(1)+f'(-1)-f(1)+f(-1) \end{array}\right)
$$

ainsi

$$
\alpha = \frac{2f(-1)+2f(1)+f'(-1)-f'(1)}{4}, \qquad \beta = \frac{-3f(-1)+3f(1)-f'(-1)-f'(1)}{4},
$$
$$
\gamma = \frac{-f'(-1)+f'(1)}{4}, \qquad \delta = \frac{f(-1)-f(1)+f'(-1)+f'(1)}{4}.
$$

Exercice 2.8

1. Construire le polynôme de Lagrange p qui interpole les points $(-1,0)$, $(0,0)$, $(1,0)$ et $(2,0)$.
2. Construire l'ensemble des polynômes de degré 4 qui interpolent les points $(-1,0)$, $(0,0)$, $(1,0)$ et $(2,0)$.
3. Construire le polynôme d'Hermite Q qui interpole les points $(-1,0,1)$ et $(2,0,-1)$.

Solution.

1. Le polynôme d'interpolation de Lagrange de degré n sur l'ensemble des $n+1$ points $\{(x_i, y_i)\}_{i=0}^n$ s'écrit

$$
p_n(x) = \sum_{i=0}^n \left(y_i \prod_{\substack{j=0\\ j\neq i}}^n \frac{x-x_j}{x_i-x_j} \right).
$$

   Ici $n=3$ et $y_i=0$ pour $i=0,1,2,3$ donc $p_3(x)=0$.

2. Comme les points donnés appartiennent tous à la droite d'équation $y=0$, il s'agit de construire les polynômes de degré 4 qui ont 4 racines réelles distinctes $\{x_1, x_2, x_3, x_4\}$. Ils sont tous de la forme $r_a(x) = a(x-x_1)(x-x_2)(x-x_3)(x-x_4)$; ici donc $r_a(x) = a(x+1)x(x-1)(x-2) = a(x^4-2x^3-x^2+2x)$.

3. Étant donné $n+1$ points distincts $x_0, \dots, x_n$ et $n+1$ couples correspondants $(y_0, y'_0), \dots, (y_n, y'_n)$, le polynôme d'Hermite Q de degré $N=2n+1$ tel que $Q(x_i)=y_i$ et $Q'(x_i)=y'_i$, pour $i=0,\dots n$ s'écrit

$$
Q(x) = \sum_{i=0}^n y_i A_i(x) + y'_i B_i(x) \in \mathbb{P}_N \qquad \text{où} \qquad \begin{cases} L_i(x) = \prod\limits_{\substack{j=0\\ j\neq i}}^n \frac{x-x_j}{x_i-x_j}, \\ c_i = \sum\limits_{\substack{j=0\\ j\neq i}}^n \frac{1}{x_i-x_j}, \\ A_i(x) = (1-2(x-x_i)c_i)(L_i(x))^2, \\ B_i(x) = (x-x_i)(L_i(x))^2. \end{cases}
$$

   Ici $n=1$ et le polynôme de Hermite s'écrit

$$
\begin{aligned}
Q(x) &= y_0 A_0 + y'_0 B_0 + y_1 A_1 + y'_0 B_1 = B_0 - B_1 \\
&= (x-x_0)\left(\frac{x-x_1}{x_0-x_1}\right)^2 - (x-x_1)\left(\frac{x-x_0}{x_1-x_0}\right)^2 = (x+1)\left(\frac{x-2}{-3}\right)^2 - (x-2)\left(\frac{x+1}{3}\right)^2 \\
&= \frac{(x-2)^2(x+1)-(x-2)(x+1)^2}{9} = \frac{-3(x-2)(x+1)}{9} = \frac{-x^2+x+2}{3}.
\end{aligned}
$$

   Si on a oublié la formule, il suffit de remarquer qu'on cherche un polynôme de degré 3 qui a comme racines -1 et 2 et donc qui s'écrit $Q(x) = (x+1)(x-2)(ax+b) = ax^3 + (-a+b)x^2 + (-b-2a)x - 2b$; de plus on sait que $Q'(-1)=1$ et $Q'(2)=-1$, on trouve alors a et b en résolvant le système linéaire

$$
\begin{cases} 3a(-1)^2 + 2(-a+b)(-1) + (-b-2a) = 1, \\ 3a(2)^2 + 2(-a+b)(2) + (-b-2a) = -1, \end{cases} \iff \begin{cases} 3a+2a-2b-b-2a = 1, \\ 12a-4a+4b-b-2a = -1, \end{cases} \iff \begin{cases} a = 0, \\ b = -1/3. \end{cases}
$$

On obtient le polynôme $Q(x) = \frac{-(x+1)(x-2)}{3}$.
Une autre idée pour calculer le polynôme Q sans utiliser la formule ni la remarque précédente est de calculer directement le polynôme selon la définition : on cherche un polynôme de degré 3, donc de la forme $Q(x) = a_0 + a_1x + a_2x^2 + a_3x^3$, qui vérifie $Q(-1) = 0$, $Q(2) = 0$, $Q'(-1) = 1$ et $Q'(2) = -1$. On doit alors résoudre le système linéaire

$$\begin{cases} a_0 - a_1 + a_2 - a_3x^3 = 0 \\ a_0 + 2a_1 + 4a_2 + 8a_3x^3 = 0 \\ a_1 - 2a_2 + 3a_3x^3 = 1 \\ a_1 + 4a_2 + 12a_3x^3 = -1 \end{cases}$$

qu'on peut réécrire sous la forme $\mathbb{A}\mathbf{a} = \mathbf{b}$ avec

$$\mathbb{A} = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 2 & 4 & 8 \\ 0 & 1 & -2 & 3 \\ 0 & 1 & 4 & 12 \end{pmatrix}, \qquad \mathbf{a} = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} \qquad \text{et} \qquad \mathbf{b} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}$$

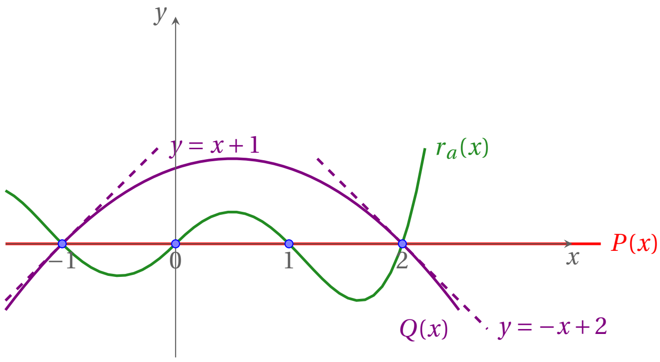
On utilise la méthode d'élimination de Gauss :

$$(\mathbb{A}|\mathbf{b}) = \left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & 0 \\ 1 & 2 & 4 & 8 & 0 \\ 0 & 1 & -2 & 3 & 1 \\ 0 & 1 & 4 & 12 & -1 \end{array}\right) \xrightarrow{L_2 \leftarrow L_2 - L_1} \left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & 0 \\ 0 & 3 & 2 & 9 & 0 \\ 0 & 1 & -2 & 3 & 1 \\ 0 & 1 & 4 & 12 & -1 \end{array}\right)$$
$$\xrightarrow[L_4 \leftarrow L_4 - L_3/3]{L_3 \leftarrow L_3 - L_2/3} \left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & 0 \\ 0 & 3 & 2 & 9 & 0 \\ 0 & 0 & -3 & 0 & 1 \\ 0 & 0 & 3 & 9 & -1 \end{array}\right) \xrightarrow{L_4 \leftarrow L_4 + L_3} \left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & 0 \\ 0 & 3 & 2 & 9 & 0 \\ 0 & 0 & -3 & 0 & 1 \\ 0 & 0 & 0 & 9 & 0 \end{array}\right)$$

et finalement on obtient

$$a_3 = 0, \qquad a_2 = -\frac{1}{3}, \qquad a_1 = \frac{1}{3}, \qquad a_0 = \frac{2}{3},$$

d'où $Q(x) = \dfrac{-x^2 + x + 2}{3}$.



Exercice 2.9

L'espérance de vie dans un pays a évoluée dans le temps selon le tableau suivant :

Année	1975	1980	1985	1990
Espérance	72,8	74,2	75,2	76,4

Utiliser l'interpolation de Lagrange pour estimer l'espérance de vie en 1977, 1983 et 1988. La comparer avec une interpolation linéaire par morceaux.

Solution. Le polynôme d'interpolation de Lagrange de degré n sur l'ensemble des $n+1$ points $\{(x_i, y_i)\}_{i=0}^n$ s'écrit

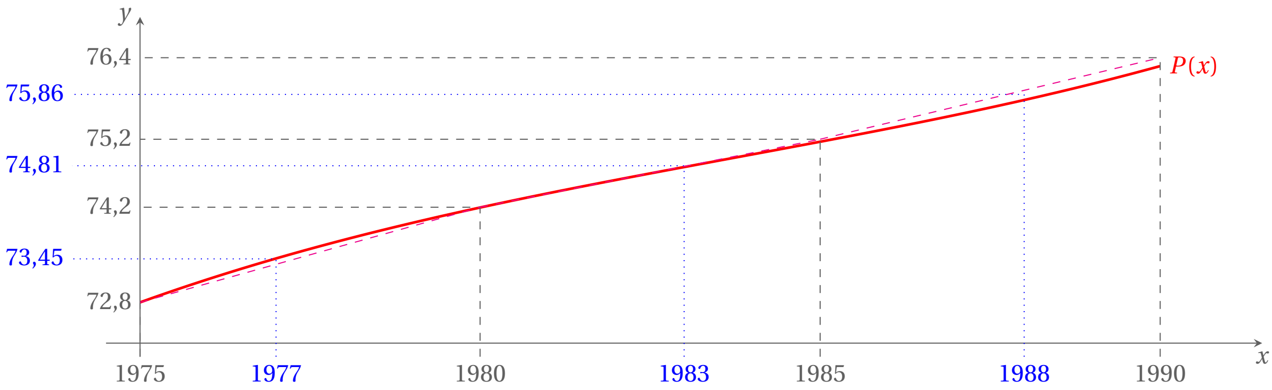
$$p_n(x) = \sum_{i=0}^{n} \left(y_i \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} \right).$$

Ici $n = 3$ et si on choisit de poser $x_0 = 0$ pour l'année 1975, $x_1 = 5$ pour l'année 1980 etc., on a

$$\begin{aligned}
P(x) = y_0 &\frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \\
&+ y_2 \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} = \\
&= 72{,}8 \frac{(x-5)(x-10)(x-15)}{(0-5)(0-10)(0-15)} + 74{,}2 \frac{(x-0)(x-10)(x-15)}{(5-0)(5-10)(5-15)} \\
&+ 75{,}2 \frac{(x-0)(x-5)(x-15)}{(10-0)(10-5)(10-15)} + 76{,}4 \frac{(x-0)(x-5)(x-10)}{(15-0)(15-5)(15-10)} = \\
&= \frac{-72{,}8(x-5)(x-10)(x-15) + 3\times 74{,}2x(x-10)(x-15) - 3\times 75{,}2x(x-5)(x-15) + 76{,}4x(x-5)(x-10)}{750}
\end{aligned}$$

On a alors que

- ▷ l'espérance de vie en 1977 correspond à $P(2) = 73{,}45$,
- ▷ l'espérance de vie en 1983 correspond à $P(8) = 74{,}81$,
- ▷ l'espérance de vie en 1988 correspond à $P(13) = 75{,}86$.



Remarque : il est intéressant de considérer une interpolation linéaire par morceaux (splines de degré 1) ; on note que l'espérance de vie est sous-estimé en 1977 et sur-estimé en 1988 par rapport à l'interpolation précédente car

- ▷ l'espérance de vie en 1977 correspond à $\frac{74{,}2-72{,}8}{5-0}2 + 72{,}8 = 73{,}36 < P(2)$,
- ▷ l'espérance de vie en 1983 correspond à $\frac{75{,}2-74{,}2}{10-5}8 + 73{,}2 = 74{,}8 \sim P(8)$,
- ▷ l'espérance de vie en 1988 correspond à $\frac{76{,}4-74{,}2}{15-10}13 + 72{,}8 = 75{,}92 > P(13)$.

Exercice 2.10

Pour calculer le zéro d'une fonction $y = f(x)$ inversible sur un intervalle $[a; b]$ on peut utiliser l'interpolation : après avoir évalué f sur une discrétisation x_i de $[a; b]$, on interpole l'ensemble $\left\{(y_i, x_i)\right\}_{i=0}^{n}$ et on obtient un polynôme $x = p(y)$ tel que

$$f(x) = 0 \qquad \Longleftrightarrow \qquad x = p(0).$$

Utiliser cette méthode pour évaluer l'unique racine α de la fonction $f(x) = e^x - 2$ dans l'intervalle $[0; 1]$ avec trois points d'interpolation.

Comparer ensuite le résultat obtenu avec l'approximation du zéro de f obtenue par la méthode de Newton en 3 itérations à partir de $x_0 = 0$.

SOLUTION. Calculons d'abord les valeurs à interpoler

i	x_i	y_i
0	0	-1
1	$\frac{1}{2}$	$\sqrt{e}-2$
2	1	$e-2$

Le polynôme d'interpolation de Lagrange de degré n sur l'ensemble des $n+1$ points $\{(y_i, x_i)\}_{i=0}^{n}$ s'écrit

$$p_n(y) = \sum_{i=0}^{n} \left(x_i \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{y-y_j}{y_i-y_j} \right).$$

Ici $n=2$ donc on a

$$\begin{aligned}
p(y) &= x_0 \frac{(y-y_1)(y-y_2)}{(y_0-y_1)(y_0-y_2)} + x_1 \frac{(y-y_0)(y-y_2)}{(y_1-y_0)(y_1-y_2)} + x_2 \frac{(y-y_0)(y-y_1)}{(y_2-y_0)(y_2-y_1)} \\
&= \frac{1}{2} \frac{(y+1)(y-e+2)}{(\sqrt{e}-2+1)(\sqrt{e}-2-e+2)} + \frac{(y+1)(y-\sqrt{e}+2)}{(e-2+1)(e-2-\sqrt{e}+2)}.
\end{aligned}$$

Par conséquent une approximation de la racine de f est $p(0) = \frac{1}{2}\frac{-e+2}{(\sqrt{e}-2+1)(\sqrt{e}-2-e+2)} + \frac{-\sqrt{e}+2}{(e-2+1)(e-2-\sqrt{e}+2)} \approx 0.7087486785$.

La méthode de Newton s'écrit

$$\begin{cases} x_0 = 0, \\ x_{k+1} = x_k - \frac{e^{x_k}-2}{e^{x_k}} = x_k - 1 + \frac{2}{e^{x_k}}, \end{cases}$$

on obtient ainsi la suite

k	x_k
0	0
1	1
2	$\frac{2}{e} \approx 0.7357588825$
3	$\frac{2-e}{e} - \frac{2}{e^{\frac{2}{e}}} \approx 0.6940422999$

Remarque : comme il n'y a que trois points d'interpolation, on pourrait calculer directement le polynôme interpolateur de f plutôt que de sa réciproque et chercher les zéros de ce polynôme directement car il s'agit d'un polynôme de degré 2. Cependant cette idée ne peu pas être généralisée au cas de plus de trois points d'interpolation car on ne connait pas de formule générale pour le calcul des zéros d'un polynôme de degré $n \geq 3$.

Exercice 2.11

Soit f une fonction continue dont on connait les valeurs uniquement pour t entier, c'est-à-dire on suppose connues les valeurs $f(\kappa)$ pour tout $\kappa \in \mathbb{Z}$. Si $t \in \mathbb{R} \setminus \mathbb{Z}$, on définit une approximation $p(t)$ de $f(t)$ en interpolant la fonction f par un polynôme de degré 3 aux quatre points entiers les plus proches de t. Calculer $p(t)$ et écrire un algorithme qui fournit $p(t)$.

SOLUTION. Soit $\ell = E[t]$ la partie entière[1] de t. Alors $t \in [\ell; \ell+1]$ et il s'agit de définir le polynôme p interpolant les points

$$(\kappa-1, f(\kappa-1)), \qquad (\kappa, f(\kappa)), \qquad (\kappa+1, f(\kappa+1)), \qquad (\kappa+2, f(\kappa+2)),$$

ce qui donne

$$\begin{aligned}
P(t) &= \sum_{i=0}^{3} \left(f(\kappa-1+i) \prod_{\substack{j=0\\ j\neq i}}^{3} \frac{t-(\kappa-1+j)}{(\kappa-1+i)-(\kappa-1+j)} \right) = \sum_{i=0}^{3} \left(f(\kappa-1+i) \prod_{\substack{j=0\\ j\neq i}}^{3} \frac{t-\kappa+1-j}{i-j} \right) \\
&= -\frac{f(\kappa-1)}{6}(t-\kappa)(t-\kappa-1)(t-\kappa-2) + \frac{f(\kappa)}{2}(t-\kappa+1)(t-\kappa-1)(t-\kappa-2) \\
&\quad - \frac{f(\kappa+1)}{2}(t-\kappa+1)(t-\kappa)(t-\kappa-2) + \frac{f(\kappa+2)}{6}(t-\kappa+1)(t-\kappa)(t-\kappa-1)
\end{aligned}$$

Require: $f\colon \mathbb{Z} \to \mathbb{R}$, t
  $\kappa \leftarrow E[t]$
  $x_0 \leftarrow \kappa - 1$

1. Pour tout nombre réel x, la partie entière notée $E(x)$ est le plus grand entier relatif inférieur ou égal à x. Par exemple, $E(2.3) = 2$, $E(-2) = -2$ et $E(-2.3) = -3$. La fonction partie entière est aussi notée $[x]$ (ou $\lfloor x \rfloor$ par les anglo-saxons). On a toujours $E(x) \leq x < E(x)+1$ avec égalité si et seulement si x est un entier relatif. Pour tout entier relatif k et et pour tout nombre réel x, on a $E(x+k) = E(x)+k$. L'arrondi à l'entier le plus proche d'un réel x peut être exprimé par $E(x+0.5)$.

$x_1 \leftarrow \kappa$
$x_2 \leftarrow \kappa + 1$
$x_3 \leftarrow \kappa + 2$
$y \leftarrow 0$
for $i = 0$ to 3 **do**
  $L \leftarrow 1$
  for $j = 0$ to 3 **do**
    if $j \neq i$ **then**
      $L \leftarrow \dfrac{t - x_j}{x_i - x_j} \times L$
    end if
  end for
  $y \leftarrow y + f(x_i) \times L$
end for
return y

3. Quadrature

Calculer $\int_a^b f(x)\,dx$ où f est une fonction donnée

Soit f une fonction réelle intégrable sur l'intervalle $[a;b]$. Le calcul explicite de l'intégrale définie $I(f) = \int_a^b f(x)dx$ peut être difficile, voire impossible. On appelle *formule de quadrature* ou *formule d'intégration numérique* toute formule permettant de calculer une approximation de $I(f)$. Une possibilité consiste à remplacer f par une approximation f_n, où n est un entier positif, et calculer $I(f_n)$ au lieu de $I(f)$. En posant $I_n(f) = I(f_n)$ (la dépendance par rapports aux extrémités a et b sous-entendue), on a

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

Si f est de classe $\mathscr{C}^0$ sur $[a;b]$, l'erreur de quadrature $E_n(f) = |I_n(f) - I(f)|$ satisfait

$$E_n(f) \leq \int_a^b |f(x) - f_n(x)|dx \leq (b-a)\|f - f_n\|_\infty.$$

L'approximation f_n doit être facilement intégrable, ce qui est le cas si, par exemple, $f_n = \sum_{i=0}^n \xi_i x^i \in \mathbb{R}_n[x]$ car

$$I(f) \approx I_n(f) = \int_a^b f_n(x)dx = \int_a^b \left(\sum_{i=0}^n \xi_i x^i\right) dx = \sum_{i=0}^n \xi_i \left(\int_a^b x^i dx\right) = \sum_{i=0}^n \frac{\xi_i}{i+1}\left[x^{i+1}\right]_a^b = \sum_{i=0}^n \frac{b^{i+1} - a^{i+1}}{i+1}\xi_i.$$

Une approche naturelle consiste à prendre $f_n = \Pi_n f = \sum_{i=0}^n f(x_i)L_i(x)$, le polynôme d'interpolation de LAGRANGE de f sur un ensemble de $n+1$ nœuds distincts $\{x_i\}_{i=0}^{i=n}$. Ainsi on déduit

$$I_n(f) = \sum_{i=0}^n \left(f(x_i)\int_a^b L_i(x)dx\right) \qquad \text{où} \qquad L_i(x) = \prod_{\substack{j=0\\ j\neq i}}^n \frac{x - x_j}{x_i - x_j}.$$

Il s'agit d'un cas particulier de la formule de quadrature suivante

$$I_n(f) = \sum_{i=0}^n \alpha_i f(x_i)$$

qui est une somme pondérée des valeurs de f aux points x_i : on dit que ces points sont les nœuds de la formule de quadrature et que les nombres $\alpha_i \in \mathbb{R}$ sont les coefficients ou encore les poids. La formule de quadrature de LAGRANGE peut être généralisée au cas où on connaît les valeurs de la dérivée de f : ceci conduit à la formule de quadrature d'HERMITE. Les formules de Lagrange et d'HERMITE sont toutes les deux des *formules de quadrature interpolatoires*, car la fonction f est remplacée par son polynôme d'interpolation.

Degré d'exactitude

On définit le degré d'exactitude d'une formule de quadrature comme le plus grand entier $r \geq 0$ pour lequel $I_n(q) = I(q)$ pour tout polynôme $q \in \mathbb{R}_r[x]$.

Astuce

Si q est un polynôme de $\mathbb{R}_r[x]$, il existe $\alpha_0, \alpha_1, \dots, \alpha_r$ tels que $q(x) = \sum_{k=0}^r \alpha_k x^k$. Alors $I(q) = \int_a^b q(x)dx = \sum_{k=0}^r \alpha_k (\int_a^b x^k) = \sum_{k=0}^r \alpha_k I(x^k)$. Pour vérifier qu'une formule de quadrature I_n a degré d'exactitude r il suffit alors de vérifier que $I_n(x^k) = I(x^k)$ pour tout $k = 0 \dots r$.

Théorème

Toute formule de quadrature interpolatoire utilisant $n+1$ nœuds distincts a un degré d'exactitude au moins égale à n.
En effet, si $f \in \mathbb{R}_n[x]$, alors $\Pi_n f \equiv f$.

La réciproque aussi est vraie : une formule de quadrature utilisant $n+1$ nœuds distincts et ayant un degré d'exactitude au moins égale à n est nécessairement de type interpolatoire.

Le degré d'exactitude peut même atteindre $2n+1$ dans le cas des formules de quadrature de Gauss.

Stabilité

Une formule de quadrature $I_n(f) = \sum_{i=0}^{n} \alpha_i f(x_i)$ est dite stable s'il existe $M \in \mathbb{R}_+^*$ tel que $\sum_{i=0}^{n} |\alpha_i| \le M$.

Théorème

Une méthode de quadrature de type interpolation est convergente sur $\mathscr{C}[a;b]$ ssi les formules sont stables.

Formule de quadrature composite

On décompose l'intervalle d'intégration $[a;b]$ en m sous-intervalles $T_j = [y_j; y_{j+1}]$ tels que $y_j = a + jH$ où $H = \frac{b-a}{m}$ pour $j = 0, 1, \dots, m$. On utilise alors sur chaque sous-intervalle une formule interpolatoire de nœuds $\left\{x_k^{(j)}\right\}_{k=0}^{n}$ et de poids $\left\{\alpha_k^{(j)}\right\}_{k=0}^{n}$. Puisque

$$I(f) = \int_a^b f(x)\mathrm{d}x = \sum_{j=0}^{m-1} \int_{y_j}^{y_{j+1}} f(x)\mathrm{d}x,$$

une formule de quadrature interpolatoire composite est obtenue en remplaçant $I(f)$ par

$$I_{n,m}(f) = \sum_{j=0}^{m-1} \sum_{k=0}^{n} \alpha_k^{(j)} f(x_k^{(j)}).$$

Changement de variable affine

Souvent on définit d'abord une formule de quadrature sur l'intervalle $[0;1]$ ou sur l'intervalle $[-1;1]$ et puis on la généralise à l'intervalle $[x_i; x_{i+1}]$ par un changement de variable affine.
Soit $x \in [a;b]$ et soit $y \in [c;d]$, on cherche une transformation $y = g(x)$ qui envoie l'intervalle $[a;b]$ dans l'intervalle $[c;d]$ ainsi

$$\int_c^d g(y)\mathrm{d}y = \int_a^b f(g(x))g'(x)\mathrm{d}x.$$

Si $g'(x)$ est une constante, *i.e.* si g est une transformation affine $g(x) = mx + q$, alors

$$\int_c^d g(y)\mathrm{d}y = m\int_a^b f(mx+q)\mathrm{d}x.$$

Pour déterminer cette transformation affine, on doit résoudre le système linéaire

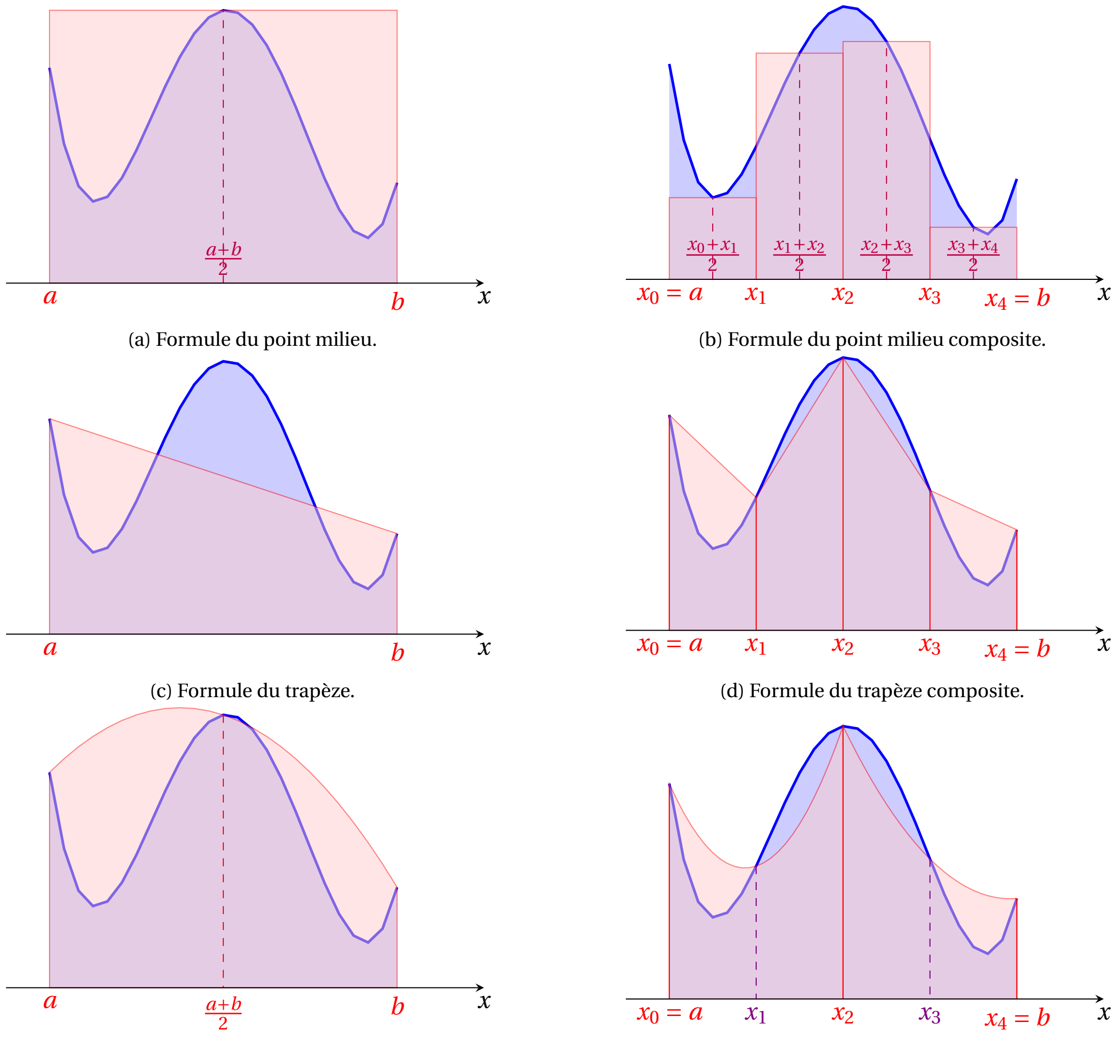
$$\begin{cases} c = ma + q, \\ d = mb + q. \end{cases}$$

On obtient

$$m = \frac{d-c}{b-a}, \qquad q = \frac{cb-ad}{b-a}.$$

Par conséquent $y = \dfrac{d-c}{b-a}x + \dfrac{cb-ad}{b-a}$ d'où

$$\int_c^d f(y)\mathrm{d}y = \frac{d-c}{b-a}\int_a^b f\left(\frac{d-c}{b-a}x + \frac{cb-ad}{b-a}\right)\mathrm{d}x.$$



(a) Formule du point milieu.

(b) Formule du point milieu composite.

(c) Formule du trapèze.

(d) Formule du trapèze composite.

(e) Formule de Cavalieri-Simpson.

(f) Formule de Cavalieri-Simpson composite.

FIGURE 3.1.: Formules de quadrature pour $n = 0, 1, 2$.

Exemple

Transformer l'intervalle $[0;1]$ dans l'intervalle $[x_i; x_{i+1}]$ par un changement de variable affine.
On a $y = (x_{i+1} - x_i)x + x_i$ et

$$\int_{x_i}^{x_{i+1}} f(y)\,\mathrm{d}y = (x_{i+1} - x_i) \int_0^1 f((x_{i+1} - x_i)x + x_i)\,\mathrm{d}x.$$

On voit que lorsque $x = 0$ alors $y = x_i$, lorsque $x = 1$ alors $y = x_{i+1}$, ou encore lorsque $x = 1/2$ alors $y = \frac{x_i + x_{i+1}}{2}$ etc.

Exemple

Transformer l'intervalle $[-1;1]$ dans l'intervalle $[x_i; x_{i+1}]$ par un changement de variable affine.
On a $y = \frac{x_{i+1}-x_i}{2} x + \frac{-x_{i+1}-x_i}{2}$, qu'on peut réécrire $y = x_i + (1+x)\frac{x_{i+1}-x_i}{2}$ et

$$\int_{x_i}^{x_{i+1}} f(y)\,\mathrm{d}y = \frac{x_{i+1} - x_i}{2} \int_{-1}^{1} f\left(x_i + (1+x)\frac{x_{i+1} - x_i}{2}\right)\mathrm{d}x.$$

Exemples de formules de quadrature interpolatoires

1. La formule du *rectangle* ou du *point milieu* est obtenue en remplaçant f par une constante égale à la valeur de f au milieu de $[a;b]$ (polynôme de degré 0), ce qui donne

$$I_0(f) = (b-a)f\left(\frac{a+b}{2}\right).$$

Si $f \in \mathscr{C}^2([a;b])$ alors l'erreur de quadrature est

$$E_0(f) = \frac{h^3}{3}|f''(\eta)|, \qquad h = \frac{b-a}{2}, \quad \eta \in]a;b[.$$

Le degré d'exactitude de la formule du point milieu est 1.
On décompose maintenant l'intervalle d'intégration $[a;b]$ en m sous-intervalles de largeur $H = \frac{b-a}{m}$ avec $m \ge 1$ En introduisant les nœuds de quadrature $x_k = a + \frac{2k+1}{2}H$ pour $k = 0, 1, \dots, m-1$ on obtient la *formule composite du point milieu*

$$I_{o,m}(f) = H\sum_{k=0}^{m-1} f\left(a + \frac{2k+1}{2}H\right).$$

Si $f \in \mathscr{C}^2([a;b])$ alors l'erreur de quadrature est

$$E_{0,m}(f) = \frac{b-a}{24}H^2|f''(\eta)|, \qquad \eta \in]a;b[.$$

2. La formule du *trapèze* est obtenue en remplaçant f par le segment qui relie $(a, f(a))$ à $(b, f(b))$ (polynôme de Lagrange de degré 1), ce qui donne

$$I_1(f) = \frac{b-a}{2}\big(f(a) + f(b)\big).$$

Si $f \in \mathscr{C}^2([a;b])$ alors l'erreur de quadrature est

$$E_1(f) = \frac{h^3}{12}|f''(\eta)|, \qquad h = b-a, \quad \eta \in]a;b[.$$

Le degré d'exactitude de la formule du point milieu est 1, comme celle du point milieu.
Pour obtenir la *formule du trapèze composite*, on décompose l'intervalle d'intégration $[a;b]$ en m sous-intervalles de largeur $H = \frac{b-a}{m}$ avec $m \ge 1$. En introduisant les nœuds de quadrature $x_k = a + kH$ pour $k = 0, 1, \dots, m-1$ on obtient

$$I_{1,m}(f) = \frac{H}{2}\sum_{k=0}^{m-1}\big(f(x_k) + f(x_{k+1})\big) = H\left(\frac{1}{2}f(a) + \sum_{k=1}^{m-1} f(a+kH) + \frac{1}{2}f(b)\right).$$

Si $f \in \mathscr{C}^2([a;b])$ alors l'erreur de quadrature est

$$E_{0,m}(f) = \frac{b-a}{12}H^2|f''(\eta)|, \qquad \eta \in]a;b[.$$

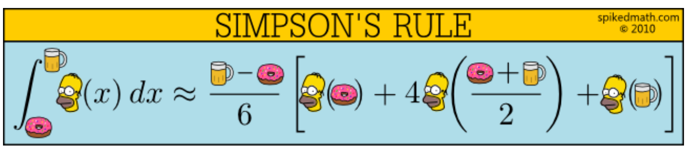
3. La formule de *Cavalieri-Simpson* est obtenue en remplaçant f par la parabole qui interpole $(a, f(a))$, $(b, f(b))$ et $\left(\frac{a+b}{2}, f\left(\frac{a+b}{2}\right)\right)$ (polynôme de Lagrange de degré 2), ce qui donne

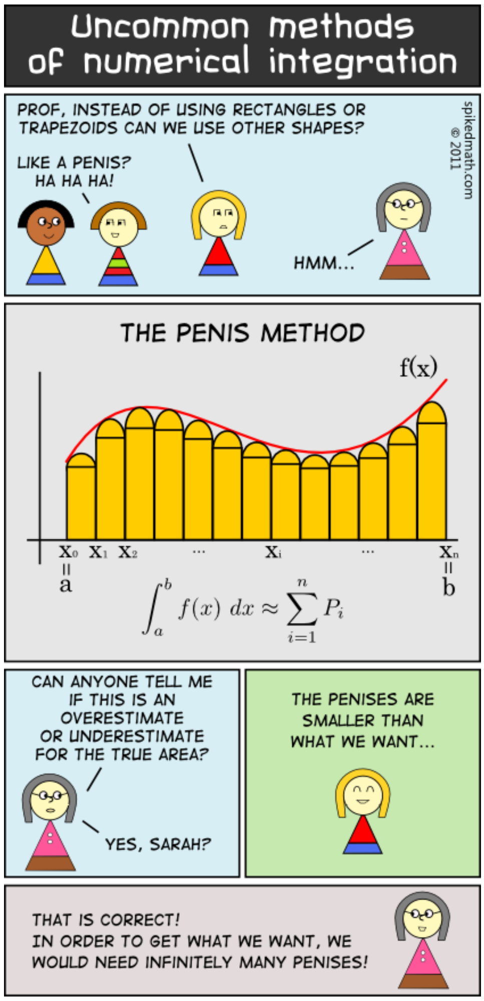
$$I_2(f) = \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right).$$

Si $f \in \mathscr{C}^4([a;b])$ alors l'erreur de quadrature est

$$E_2(f) = \frac{h^5}{90}|f^{(4)}(\eta)|, \qquad h = \frac{b-a}{2}, \quad \eta \in]a;b[.$$

Le degré d'exactitude de la formule du point milieu est 3.





Pour obtenir la *formule composite*, on décompose l'intervalle d'intégration $[a;b]$ en m sous-intervalles de largeur $H=\frac{b-a}{2m}$ avec $m\ge 1$. En introduisant les nœuds de quadrature $x_k=a+kH$ pour $k=0,1,\dots,m$ on obtient

$$I_{2,m}(f)=\frac{H}{3}\left(f(a)+2\sum_{r=1}^{m-1}f(x_{2r})+4\sum_{s=0}^{s-1}f(x_{2s+1})+f(b)\right)=\frac{H}{3}\left(f(a)+2\sum_{r=1}^{m-1}f(a+rH)+4\sum_{s=0}^{m-1}f\left(a+\frac{2s+1}{2}H\right)+f(b)\right).$$

Si $f\in\mathscr{C}^4([a;b])$ alors l'erreur de quadrature est

$$E_{2,m}(f)=\frac{b-a}{180}\left(\frac{H}{4}\right)^4|f^{(4)}(\eta)|,\qquad \eta\in]a;b[.$$

Algorithmes

Méthode du rectangle

Require: f, a, $b>a$, $n>0$
  $h\leftarrow\frac{b-a}{n}$
  $s\leftarrow 0$
  for $i=0$ to $n-1$ **do**
    $s\leftarrow s+f\left(a+\left(i+\frac{1}{2}\right)h\right)$
  end for
  return $I\leftarrow hs$

Méthode des trapèzes

Require: f, a, $b>a$, $n>0$
  $h\leftarrow\frac{b-a}{n}$
  $s\leftarrow\frac{f(a)+f(b)}{2}$
  for $i=1$ to $n-1$ **do**
    $s\leftarrow s+f(a+ih)$
  end for
  return $I\leftarrow hs$

Méthode de Simpson

Require: f, a, $b>a$, $n>0$
  $h\leftarrow\frac{b-a}{2n}$
  $s_1\leftarrow 0$
  $s_2\leftarrow f(a+h)$
  for $i=1$ to $n-1$ **do**
    $s_1\leftarrow s_1+f(a+2ih)$
    $s_2\leftarrow s_2+f(a+(2i+1)h)$
  end for
  return $I\leftarrow\frac{h}{3}\left(f(a)+f(b)+2s_1+4s_2\right)$

Exercices

Exercice 3.1

Estimer $\int_0^{5/2} f(x)\,\mathrm{d}x$ à partir des données

x	0	$^1/_2$	1	$^3/_2$	2	$^5/_2$
$f(x)$	$^3/_2$	2	2	1.6364	1.2500	0.9565

en utilisant la méthode des trapèzes composite.

Solution. La méthode des trapèzes composite à $m+1$ points pour calculer l'intégrale d'une fonction f sur l'intervalle $[a,b]$ s'écrit

$$\int_a^b f(t)dt \simeq h\left(\frac{1}{2}f(a)+\sum_{i=1}^{m-1} f(a+ih)+\frac{1}{2}f(b)\right) \qquad \text{avec } h=\frac{b-a}{m}.$$

Ici on a $a=0$, $b=5/2$, $h=1/2$ donc

$$\int_0^{5/2} f(x)\,\mathrm{d}x \simeq \frac{1}{2}\left(\frac{3}{4}+2+2+1.6364+1.2500+\frac{0.9565}{2}\right)=4.057325$$

Exercice 3.2

Estimer $\int_0^{\pi}\sin(x)\,\mathrm{d}x$ en utilisant la méthode des trapèzes composite avec 8 et puis 16 sous-intervalles en prenant en compte l'erreur.

Solution. La méthode des trapèzes composite à $m+1$ points pour calculer l'intégrale d'une fonction f sur l'intervalle $[a,b]$ s'écrit

$$\int_a^b f(t)dt \simeq h\left(\frac{1}{2}f(a)+\sum_{i=1}^{m-1} f(a+ih)+\frac{1}{2}f(b)\right) \qquad \text{avec } h=\frac{b-a}{m}$$

et l'erreur est donné par

$$E=-\frac{b-a}{12}h^2 f''(\xi)$$

avec $a<\xi<b$.

Ici on a $a=0$, $b=\pi$. Avec 8 sous-intervalles on a $h=\pi/8$ donc

$$\int_0^{\pi}\sin(x)\,\mathrm{d}x \simeq \frac{\pi}{8}\left(\frac{\sin(0)}{2}+\sum_{i=1}^{7}\sin(i\pi/8)+\frac{\sin(\pi)}{2}\right)\approx 1.97423$$

et l'erreur est

$$E=\frac{\pi^3}{768}\sin(\xi)$$

pour $\xi\in]0;\pi[$. Comme on ne connait pas la valeur de ξ, on ne peut pas connaitre E mais on peut en déterminer les bornes :

$$E_{\min}=\frac{\pi^3}{768}\sin(0)=0 \qquad\qquad E_{\max}=\frac{\pi^3}{768}\sin(\pi/2)=\frac{\pi^3}{768}\simeq 0.04037$$

ainsi

$$(1.97423-0)\le\int_0^{\pi}\sin(x)\,\mathrm{d}x\le(1.97423+0.04037)=2.01460$$

La valeur exacte est bien évidemment 2.

Avec 16 sous-intervalles on a $h=\pi/16$ et les nouveaux noeuds se trouvent au milieus des sous-intervalles précédents : $x_j=\pi/16+j\pi/8=(1+2j)\pi/16$ pour $j=0,1,\dots,7$, ainsi

$$\int_0^{\pi}\sin(x)\,\mathrm{d}x \simeq \frac{1.97423}{2}+\frac{\pi}{16}\sum_{j=0}^{7}\sin((1+2j)\pi/16)\approx 1.99358$$

et le limites de l'erreur deviennent (observons que E est divisé par 4 lorsque h est divisé par 2) :

$$E_{\min} = 0 \qquad\qquad E_{\max} \simeq \frac{0.04037}{4} = 0.01009$$

ainsi

$$1.99358 \le \int_0^{\pi} \sin(x)\,\mathrm{d}x \le (1.99358 + 0.01009) = 2.00367.$$

Exercice 3.3

On considère l'intégrale

$$I = \int_1^2 \frac{1}{x}\,\mathrm{d}x.$$

1. Calculer la valeur exacte de I.
2. Évaluer numériquement cette intégrale par la méthode des trapèzes avec $m = 3$ sous-intervalles.
3. Pourquoi la valeur numérique obtenue à la question précédente est-elle supérieure à $\ln(2)$? Est-ce vrai quelque soit m ? Justifier la réponse. (On pourra s'aider par un dessin.)
4. Quel nombre de sous-intervalles m faut-il choisir pour avoir une erreur inférieure à 10^{-4} ? On rappelle que l'erreur de quadrature associée s'écrit, si $f \in \mathscr{C}^2([a;b])$,

$$|E_m| = \left|\frac{(b-a)^4}{12m^2} f''(\xi)\right|, \qquad \xi \in]a;b[.$$

Solution.

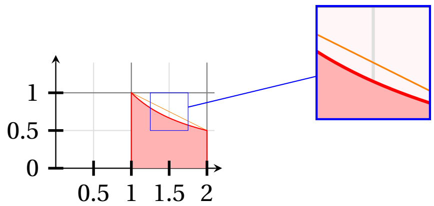
1. Une primitive de $\frac{1}{x}$ est $F(x) = \ln(x)$. La valeur exacte est alors $I = \Big[\ln(x)\Big]_{x=1}^{x=2} = \ln(2)$.
2. La méthode des trapèzes composite à $m+1$ points pour calculer l'intégrale d'une fonction f sur l'intervalle $[a,b]$ s'écrit

$$\int_a^b f(t)dt \simeq h\left(\frac{1}{2}f(a) + \sum_{i=1}^{m-1} f(a+ih) + \frac{1}{2}f(b)\right) \qquad \text{avec } h = \frac{b-a}{m}.$$

   Ici on a $f(x) = \frac{1}{x}$, $a = 1$, $b = 2$, $m = 3$ d'où $h = \frac{1}{3}$ et on obtient

$$I \simeq \frac{1}{3}\left(\frac{1}{2}f(1) + f(1+1/3) + f(1+2/3) + \frac{1}{2}f(2)\right) = \frac{1}{3}\left(\frac{1}{2} + \frac{3}{4} + \frac{3}{5} + \frac{1}{4}\right) = \frac{21}{30} = 0{,}7.$$

3. La valeur numérique obtenue à la question précédente est supérieure à $\ln(2)$ car la fonction $f(x) = \frac{1}{x}$ est convexe. On peut se convaincre à l'aide d'un dessin que les trapèzes sont au-dessus de la courbe $y = 1/x$, l'aire sous les trapèzes sera donc supérieure à l'aire sous la courbe. Pour bien visualiser la construction considérons $m = 1$:



   Cela reste vrai quelque soit le pas h choisi car la fonction est convexe ce qui signifie qu'une corde définie par deux points de la courbe $y = 1/x$ sera toujours au-dessus de la courbe et par le raisonnement précédant l'aire sous les trapèzes sera supérieure à l'aire exacte.
4. L'erreur est majorée par

$$|E| \le \frac{(b-a)^4}{12m^2} \sup_{\xi\in]a;b[} |f''(\xi)|.$$

   Donc ici on a

$$|E| \le \frac{1}{12m^2} \max_{\xi\in]1;2[} \frac{2}{\xi^3} = \frac{1}{6m^2}.$$

   Pour que $|E| < 10^{-4}$ il suffit que $\frac{1}{6m^2} < 10^{-4}$, *i.e.* $m > 10^2/\sqrt{6} \approx 40{,}8$. À partir de 41 sous-intervalles, l'erreur de quadrature est inférieure à 10^{-4}.

Exercice 3.4 *Intégration*

Étant donnée l'égalité

$$\pi = 4\left(\int_0^{+\infty} e^{-x^2}\,\mathrm{d}x\right)^2 = 4\left(\int_0^{10} e^{-x^2}\,\mathrm{d}x + \epsilon\right)^2,$$

avec $0 < \epsilon < 10^{-44}$, utiliser la méthode des trapèzes composite à 10 intervalles pour estimer la valeur de π.

Solution. La méthode des trapèzes composite à m intervalles pour calculer l'intégrale d'une fonction f sur l'intervalle $[a,b]$ s'écrit

$$\int_a^b f(t)\,dt \simeq h\left(\frac{1}{2}f(a) + \sum_{i=1}^{m-1} f(a+ih) + \frac{1}{2}f(b)\right) \qquad \text{avec } h = \frac{b-a}{m}.$$

Ici on a $f(x) = e^{-x^2}$, $a = 0$, $b = 10$, $m = 10$ d'où $h = 1$ et on obtient

$$I \simeq \frac{1}{2} + \sum_{i=1}^{10} e^{-i^2} + \frac{1}{2e^{100}} = \frac{1}{2} + \frac{1}{e} + \frac{1}{e^4} + \frac{1}{e^9} + \frac{1}{e^{16}} + \frac{1}{e^{25}} + \frac{1}{e^{36}} + \frac{1}{e^{49}} + \frac{1}{e^{64}} + \frac{1}{e^{81}} + \frac{1}{2e^{100}} \approx 3.1422.$$

Exercice 3.5

Soit f une fonction $\mathscr{C}^\infty(\mathbb{R},\mathbb{R})$.

1. On considère l'approximation
$$\int_{-1}^1 f(x)\,\mathrm{d}x \approx \frac{2}{3}\Big(2f\big(-\tfrac{1}{2}\big) - f(0) + 2f\big(\tfrac{1}{2}\big)\Big).$$
Quel est le degré d'exactitude de cette formule de quadrature ?
2. On se donne les points $\{x_i\}_{i=0}^{i=n}$ de subdivision de l'intervalle $[a;b]$: $x_i = a + ih$ avec $h = \frac{b-a}{n}$. À l'aide d'un changement de variable affine, en déduire une formule de quadrature pour l'intégrale
$$\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x.$$
En tirer une formule de quadrature composite pour l'intégrale $\int_a^b f(x)\,\mathrm{d}x$.
3. Écrire l'algorithme pour approcher $\int_a^b f(x)\,\mathrm{d}x$.

Solution.

1. On a

k	$p_k(x) = x^k$	$\int_{-1}^1 p_k(x)\,\mathrm{d}x$	$\frac{2}{3}(2p_k(-1/2) - p_k(0) + 2p_k(1/2))$	Degré d'exactitude
0	1	2	2	au moins 0
1	x	0	0	au moins 1
2	x^2	2/3	2/3	au moins 2
3	x^3	0	0	au moins 3
4	x^4	2/5	1/6	3

La formule est donc exacte de degré 3.

2. Soit $x = mt + q$, alors
$$\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x = m\int_{-1}^1 f(mt+q)\,\mathrm{d}t \quad \text{avec} \quad \begin{cases} x_i = -m + q, \\ x_{i+1} = m + q, \end{cases}$$
d'où le changement de variable $x = x_i + (t+1)\frac{x_{i+1}-x_i}{2}$. On déduit la formule de quadrature (exacte sur l'espace des polynôme de degré au plus 3)
$$\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x = \frac{x_{i+1}-x_i}{2}\int_{-1}^1 f\left(x_i + (t+1)\frac{x_{i+1}-x_i}{2}\right)\mathrm{d}t \approx \frac{x_{i+1}-x_i}{3}\left[2f\left(\frac{x_i + \frac{x_{i+1}+x_i}{2}}{2}\right) - f\left(\frac{x_{i+1}+x_i}{2}\right) + 2f\left(\frac{\frac{x_{i+1}+x_i}{2} + x_{i+1}}{2}\right)\right].$$
Soit $h = x_{i+1} - x_i = \frac{b-a}{n}$. La formule précédente se réécrit
$$\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x \approx \frac{h}{3}\left[2f\left(x_i + \frac{h}{4}\right) - f\left(x_i + \frac{h}{2}\right) + 2f\left(x_i + \frac{3h}{4}\right)\right].$$

et la formule de quadrature composite déduite de cette approximation est

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x \approx \frac{h}{3}\sum_{i=0}^{n-1}\left[2f\left(x_i+\frac{h}{4}\right)-f\left(x_i+\frac{h}{2}\right)+2f\left(x_i+\frac{3h}{4}\right)\right].$$

3. Algorithme d'approximation de $\int_a^b f(x)\,\mathrm{d}x$

Require: f, a, $b>a$, $n>0$
$h \leftarrow \frac{b-a}{n}$
$s \leftarrow 0$
for $i=0$ to $n-1$ **do**
$x \leftarrow a+ih$
$s \leftarrow s+2f\left(x+\frac{h}{4}\right)-f\left(x+\frac{h}{2}\right)+2f\left(x+\frac{3h}{4}\right)$
end for
return $I \leftarrow \frac{h}{3}s$

Exercice 3.6

Soit f une fonction $\mathscr{C}^\infty(\mathbb{R},\mathbb{R})$. On se donne les points $\{x_i\}_{i=0}^{i=n}$ de subdivision de l'intervalle $[a;b]$: $x_i = a+ih$ avec $h=\frac{b-a}{n}$. Le but de l'exercice est de trouver une formule de quadrature à $2n$ points pour approcher l'intégrale

$$\int_a^b f(x)\,\mathrm{d}x. \tag{3.1}$$

On propose dans un premier temps (question 1 à 4) de construire la formule de quadrature à deux points suivantes :

$$\int_{-1}^{1} g(x)\,\mathrm{d}x \approx g(-\alpha)+g(\alpha), \tag{3.2}$$

où $0<\alpha<1$ est à déterminer.

1. Choisir α pour rendre la formule de quadrature exacte pour des polynômes de degré le plus élevé possible. Quel est alors le degré d'exactitude de cette formule de quadrature ?
2. À l'aide d'un changement de variable affine, étendre cette formule de quadrature pour l'intégrale suivante :

$$\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x.$$

3. En déduire une formule de quadrature à $2n$ points, notée F, pour le calcul approché de (3.1). Cette formule de quadrature est-elle stable ?
4. Écrire l'algorithme du calcul de F.

Solution.

1. On a

k	$p_k(x)=x^k$	$\int_{-1}^1 p_k(x)\,\mathrm{d}x$	$p_k(-\alpha)+p_k(\alpha)$	Degré d'exactitude
0	1	2	2	au moins 0
1	x	0	0	au moins 1
2	x^2	$2/3$	$2\alpha^2$	1 si $\alpha\neq 1/\sqrt{3}$, au moins 2 si $\alpha=1/\sqrt{3}$
		Soit $\alpha=1/\sqrt{3}$		
3	x^3	0	0	au moins 3
4	x^4	$2/5$	$2/9$	3

Donc la formule de quadrature a degré d'exactitude 1 si $\alpha\neq 1/\sqrt{3}$ et degré d'exactitude 3 si $\alpha=1/\sqrt{3}$.

2. Par le changement de variable $y=x_i+(x+1)\frac{x_{i+1}-x_i}{2}$ on déduit la formule de quadrature

$$\begin{aligned}\int_{x_i}^{x_{i+1}} f(y)\,\mathrm{d}y &= \frac{x_{i+1}-x_i}{2}\int_{-1}^{1} f\left(x_i+(x+1)\frac{x_{i+1}-x_i}{2}\right)\mathrm{d}x\\ &\approx \frac{x_{i+1}-x_i}{2}\left[f\left(x_i+\left(1-\sqrt{\tfrac{1}{3}}\right)\frac{x_{i+1}-x_i}{2}\right)+f\left(x_i+\left(1+\sqrt{\tfrac{1}{3}}\right)\frac{x_{i+1}-x_i}{2}\right)\right].\end{aligned}$$

3. Si $h=x_{i+1}-x_i=\frac{b-a}{n}$ (*i.e.* si on considère une subdivision de l'intervalle $[a;b]$ équirépartie) alors on trouve la formule de quadrature composite (*i.e.* sur n sous-intervalles et à $2n$ points)

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{h}{2}\sum_{i=0}^{n-1}\left[f\left(x_i+h\left(1-\sqrt{\tfrac{1}{3}}\right)\right)+f\left(x_i+h\left(1+\sqrt{\tfrac{1}{3}}\right)\right)\right]=\frac{h}{2}\sum_{i=0}^{n-1}\left[f\left(a+h\left(i+1-\sqrt{\tfrac{1}{3}}\right)\right)+f\left(a+h\left(i+1+\sqrt{\tfrac{1}{3}}\right)\right)\right].$$

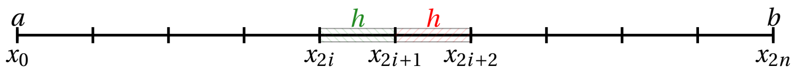
Cette formule de quadrature est stable puisque tous les coefficients sont positifs.

4. Algorithme du calcul de F :

Require: f, a, $b>a$, $n>0$
$h \leftarrow \frac{b-a}{n}$
$\alpha_1 \leftarrow a+\left(1-\sqrt{\frac{1}{3}}\right)h$
$\alpha_2 \leftarrow a+\left(1+\sqrt{\frac{1}{3}}\right)h$
for $i=0$ to $n-1$ **do**
  $s \leftarrow s+f(\alpha_1+ih)+f(\alpha_2+ih)$
end for
return $I \leftarrow \frac{h}{2}s$

Exercice 3.7 *Interpolation et Intégration*

1. Soit f une fonction de classe $\mathscr{C}^1([-1,1])$ et p le polynôme de Lagrange qui interpole f aux points -1, 0 et 1. Écrire le polynôme p.
2. En déduire une méthode de quadrature pour approcher l'intégrale
$$\int_{-1}^{1} f(t)\,\mathrm{d}t.$$
3. Étudier le degré d'exactitude de la formule de quadrature ainsi trouvée.
4. À l'aide d'un changement de variable affine, en déduire une formule de quadrature pour l'intégrale
$$\int_{x_{2i}}^{x_{2i+2}} f(x)\,\mathrm{d}x.$$
5. Soit $h=\frac{b-a}{2n}$ et $x_i=a+ih$ pour $i=0,\dots,2n$. On subdivise l'intervalle $[a;b]$ en n intervalles $[x_{2i};x_{2i+2}]$ de largeur $2h$.



En déduire la formule de quadrature composite pour le calcul approché de
$$\int_{a}^{b} f(x)\,\mathrm{d}x.$$
6. Écrire l'algorithme associé à cette formule de quadrature.

Solution.

1. On a trois points, donc le polynôme interpolateur de Lagrange est un polynôme de $\mathbb{R}_2[x]$. On cherche alors les coefficients α, β et γ du polynôme $p(x)=\alpha+\beta x+\gamma x^2$ tels que
$$\begin{cases} f(-1)=\alpha-\beta+\gamma, & (3.3a)\\ f(0)=\alpha, & (3.3b)\\ f(1)=\alpha+\beta+\gamma. & (3.3c)\end{cases}$$
L'équation (3.3b) donne $\alpha=f(0)$, la somme (3.3c)+(3.3a) donne $\gamma=\frac{f(1)-f(-1)}{2}$ et enfin la soustraction (3.3c)−(3.3a) donne $\beta=\frac{f(1)-f(-1)}{2}$.
2. On en déduit la méthode de quadrature
$$\int_{-1}^{1} f(t)\,\mathrm{d}t \approx \int_{-1}^{1} p(t)\,\mathrm{d}t = 2(\alpha+\gamma/3)=\frac{f(-1)+4f(0)+f(1)}{3}.$$
3. Par construction, cette formule de quadrature a degré d'exactitude au moins 2. De plus

k	$p_k(x)=x^k$	$\int_{-1}^{1} p_k(x)\,\mathrm{d}x$	$\frac{p_k(-1)+4p_k(0)+p_k(1)}{3}$	Degré d'exactitude
3	x^3	0	0	au moins 3
4	x^4	2/5	2/3	3

La formule est exacte pour les polynômes de degré au plus 3.

4. Soit $x = mt + q$, alors
$$\int_{x_{2i}}^{x_{2i+2}} f(x)\,\mathrm{d}x = m\int_{-1}^{1} f(mt+q)\,\mathrm{d}t \quad \text{avec} \quad \begin{cases} x_{2i} = -m+q, \\ x_{2i+2} = m+q, \end{cases}$$
d'où le changement de variable $x = x_{2i} + (t+1)\frac{x_{2i+2}-x_{2i}}{2}$. On déduit la formule de quadrature (exacte sur l'espace des polynôme de degré au plus 3)
$$\int_{x_{2i}}^{x_{2i+2}} f(x)\,\mathrm{d}x = \frac{x_{2i+2}-x_{2i}}{2}\int_{-1}^{1} f\left(x_{2i} + (t+1)\frac{x_{2i+2}-x_{2i}}{2}\right)\mathrm{d}t \approx \frac{x_{2i+2}-x_{2i}}{6}\left[f(x_{2i}) + 4f(x_{2i+1}) + f(x_{2i+2})\right].$$

5. $h = \frac{b-a}{2n} = \frac{x_{i+1}-x_i}{2}$ pour $i = 0,\dots,2n$. On subdivise l'intervalle $[a;b]$ en n intervalles $[x_{2i};x_{2i+2}]$ de largeur $2h$. On trouve ainsi la formule de quadrature composite
$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x = \sum_{i=0}^{n-1}\int_{x_{2i}}^{x_{2i+2}} f(x)\,\mathrm{d}x &\approx \sum_{i=0}^{n-1}\frac{x_{2i+2}-x_{2i}}{6}\left[f(x_{2i}) + 4f(x_{2i+1}) + f(x_{2i+2})\right] \\ &= \frac{h}{3}\sum_{i=0}^{n-1}\left[f(x_{2i}) + 4f(x_{2i+1}) + f(x_{2i+2})\right] \\ &= \frac{h}{3}\left[f(a) + f(b) + 2\sum_{i=1}^{n-1} f(x_{2i}) + 4\sum_{i=0}^{n-1} f(x_{2i}+h)\right] \\ &= \frac{h}{3}\left[f(a) + f(b) + 2\sum_{i=1}^{n-1} f(a+2ih) + 4\sum_{i=0}^{n-1} f(a+(2i+1)h)\right].\end{aligned}$$

6. Algorithme du calcul associé à cette formule de quadrature

Require: f, a, $b > a$, $n > 0$
  $h \leftarrow \frac{b-a}{2n}$
  $s_1 \leftarrow 0$
  $s_2 \leftarrow s_2 + f(a+h)$
  for $i = 1$ to $n-1$ **do**
    $s_1 \leftarrow s_1 + f(a+2ih)$
    $s_2 \leftarrow s_2 + f(a+(2i+1)h)$
  end for
  return $\frac{h}{3}\left[f(a) + f(b) + 2s_1 + 4s_2\right]$

Exercice 3.8

Soit f une fonction $\mathscr{C}^\infty(\mathbb{R},\mathbb{R})$. On se donne les points $\{x_i\}_{i=0}^{i=2n}$ de subdivision de l'intervalle $[a;b]$: $x_i = a + ih$ avec $h = \frac{b-a}{2n}$. Le but de l'exercice est de trouver une formule de quadrature à $2n+1$ points basée sur la formule de SIMPSON pour approcher
$$\int_a^b f(x)\,\mathrm{d}x. \tag{3.4}$$
On propose dans un premier temps (question 1 à 3) de construire la formule de quadrature à 3 points de Simpson :
$$\int_{-1}^{1} g(x)\,\mathrm{d}x \approx \alpha g(-1) + \beta g(0) + \alpha g(1), \tag{3.5}$$
où les réels α et β sont à déterminer.

1. Déterminer α et β pour que la formule de quadrature (3.5) ait degré d'exactitude maximale.
2. À l'aide d'un changement de variable affine, en déduire une formule de quadrature exacte sur l'espace des polynôme de degré au plus 3 pour l'intégrale suivante :
$$\int_{x_{2i}}^{x_{2i+2}} f(x)\,\mathrm{d}x.$$
3. En déduire une formule de quadrature à $2n$ points, notée F, pour le calcul approché de (3.4). Cette formule de quadrature est-elle stable ?
4. Écrire l'algorithme du calcul de F.
5. Soit x un élément de $[x_i;x_{i+1}]$. Écrire une formule de Taylor $f(x) = P_i(x) + R_i(x)$ à l'ordre 3 pour f en x, avec $P_i \in \mathbb{P}_3$. Majorer R_i sur $[x_i;x_{i+1}]$ en fonction de h.
6. En déduire une estimation d'erreur entre (3.4) et F.

SOLUTION.

1. On a

k	$p_k(x)=x^k$	$\int_{-1}^1 p_k(x)\,\mathrm{d}x$	$\alpha p_k(-1)+\beta p_k(0)+\alpha p_k(1)$	Degré d'exactitude
0	1	2	$2\alpha+\beta$	même pas 0 si $\beta\neq 2(1-\alpha)$, au moins 0 si $\beta=2(1-\alpha)$
Soit $\beta=2(1-\alpha)$				
1	x	0	0	au moins 1
2	x^2	$2/3$	2α	1 si $\alpha\neq 1/3$, au moins 2 si $\alpha=1/3$
Soit $\alpha=1/3$				
3	x^3	0	0	au moins 3
4	x^4	$2/5$	$2/3$	3

   Si $\beta\neq 2(1-\alpha)$ la formule de quadrature n'est même pas exacte pour une constante, si $\beta=2(1-\alpha)$ mais $\alpha\neq 1/3$, elle est exacte pour les polynômes de degré au plus 1, si $\alpha=1/3$ et $\beta=4/3$ la formule est exacte pour les polynômes de degré au plus 3.

2. Soit $x=mt+q$, alors
$$\int_{x_{2i}}^{x_{2i+2}} f(x)\,\mathrm{d}x = m\int_{-1}^{1} f(mt+q)\,\mathrm{d}t \quad \text{avec} \quad \begin{cases} x_{2i}=-m+q,\\ x_{2i+2}=m+q,\end{cases}$$
   d'où le changement de variable $x=x_{2i}+(t+1)\frac{x_{2i+2}-x_{2i}}{2}$. On déduit la formule de quadrature (exacte sur l'espace des polynôme de degré au plus 3)
$$\int_{x_{2i}}^{x_{2i+2}} f(x)\,\mathrm{d}x = \frac{x_{2i+2}-x_{2i}}{2}\int_{-1}^{1} f\left(x_{2i}+(t+1)\frac{x_{2i+2}-x_{2i}}{2}\right)\mathrm{d}t \approx \frac{x_{2i+2}-x_{2i}}{6}\left[f(x_{2i})+4f(x_{2i+1})+f(x_{2i+2})\right].$$

3. On trouve ainsi la formule de quadrature composite (*i.e.* sur n sous-intervalles)
$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=0}^{n-1}\int_{x_{2i}}^{x_{2i+2}} f(x)\,\mathrm{d}x \approx \sum_{i=0}^{n-1}\frac{x_{2i+2}-x_{2i}}{6}\left[f(x_{2i})+4f(x_{2i+1})+f(x_{2i+2})\right].$$
   Si $h=\frac{x_{i+1}-x_i}{2}=\frac{b-a}{2n}$ (*i.e.* si on considère une subdivision de l'intervalle $[a;b]$ équirépartie) alors on a
$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x \approx \frac{h}{3}\sum_{i=0}^{n-1}\left[f(x_{2i})+4f(x_{2i+1})+f(x_{2i+2})\right] &= \frac{h}{3}\left[f(a)+f(b)+2\sum_{i=1}^{n-1}f(x_{2i})+4\sum_{i=0}^{n-1}f(x_{2i}+h)\right]\\ &= \frac{h}{3}\left[f(a)+f(b)+2\sum_{i=1}^{n-1}f(a+2ih)+4\sum_{i=0}^{n-1}f(a+(2i+1)h)\right].\end{aligned}$$
   Cette formule de quadrature est stable puisque tous les coefficients sont positifs et on a
$$\frac{h}{3}\left[1+1+2\sum_{i=1}^{n-1}1+4\sum_{i=0}^{n-1}1\right]=\frac{b-a}{6n}\left[2+2(n-1)+4n\right]=\frac{b-a}{6n}6n=(b-a).$$

4. Algorithme du calcul de F :
   Require: $f\colon[a,b]\to\mathbb{R}$, a, $b>a$, $n>0$
   $H\leftarrow\frac{b-a}{n}$
   $s_1\leftarrow 0$
   $s_2\leftarrow s_2+f(a+H/2)$
   for $i=1$ to $n-1$ **do**
   $s_1\leftarrow s_1+f(a+iH)$
   $s_2\leftarrow s_2+f(a+(i+1)H/2)$
   end for
   return $I\leftarrow\frac{H}{6}\left[f(a)+f(b)+2s_1+4s_2\right]$

5. Soit x un élément de $[x_{2i};x_{2i+2}]$. Une formule de TAYLOR à l'ordre 3 pour f en x s'écrit
$$f(x)=P_i(x)+R_i(x),$$
   avec
$$P_i(x)=f(x_{2i})+(x-x_{2i})f'(x_{2i})+(x-x_{2i})^2\frac{f''(x_{2i})}{2}+(x-x_{2i})^3\frac{f'''(x_{2i})}{6}\qquad\in\mathbb{P}_3$$

et le reste de LAGRANGE

$$R_i(x) = (x - x_{2i})^4 \frac{f^{IV}(\xi)}{24} \qquad \text{avec} \quad \xi \in]x_{2i}; x_{2i+2}[.$$

On peut majorer R_i sur $[x_{2i}; x_{2i+2}]$ en fonction de $H = x_{2i+2} - x_{2i}$:

$$|R_i(x)| \le \frac{H^4}{24} \max|f^{IV}(\xi)| = \frac{b-a}{n}\frac{H^3}{24} \max|f^{IV}(\xi)|.$$

6. On en déduit l'estimation d'erreur entre (3.4) et F suivante [1]

$$\begin{aligned}
\left|\int_a^b f(x)\,\mathrm{d}x - F\right| &\le \left|\sum_{i=0}^{n-1}\int_{x_{2i}}^{x_{2i+2}} P_i(x)\,\mathrm{d}x - F\right| + \left|\sum_{i=0}^{n-1}\int_{x_{2i}}^{x_{2i+2}} R_i(x)\,\mathrm{d}x\right| \\
&\le nH|R_i(x_{2i+2})| + \left|\sum_{i=0}^{n-1}\int_{x_{2i}}^{x_{2i+2}} R_i(x)\,\mathrm{d}x\right| \\
&\le nH\frac{b-a}{n}\frac{H^3}{24}\max|f^{IV}(\xi)| + nH\frac{b-a}{n}\frac{H^3}{24}\max|f^{IV}(\xi)| \\
&= (b-a)\frac{H^4}{12}\sup|f^{IV}(\xi)|.
\end{aligned}$$

Exercice 3.9

Soit f une fonction $\mathscr{C}^\infty(\mathbb{R},\mathbb{R})$. On se donne les points $\{x_i\}_{i=0}^{i=n}$ de subdivision de l'intervalle $[a;b]$: $x_i = a + ih$ avec $h = \frac{b-a}{n}$. Le but de l'exercice est de trouver une formule de quadrature à $3n$ points pour approcher l'intégrale

$$\int_a^b f(x)\,\mathrm{d}x. \tag{3.6}$$

On propose dans un premier temps de construire la formule de quadrature à trois points suivantes :

$$\int_{-1}^1 g(x)\,\mathrm{d}x \approx \frac{2}{3}\big(g(-\alpha) + g(0) + g(\alpha)\big), \tag{3.7}$$

où le réel $0 < \alpha < 1$ sera à déterminer par la suite.

1. Déterminer α pour que la formule de quadrature (3.7) ait degré d'exactitude maximale. Quel est alors le degré d'exactitude de cette formule de quadrature ?
2. À l'aide d'un changement de variable affine, étendre cette formule de quadrature pour l'intégrale suivante :

$$\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x.$$

3. En déduire une formule de quadrature à $3n$ points, notée F, pour le calcul approché de (3.6). Cette formule de quadrature est-elle stable ?

SOLUTION.

1. On a

k	$p_k(x) = x^k$	$\int_{-1}^1 p_k(x)\,\mathrm{d}x$	$\frac{2}{3}(p_k(-\alpha) + \beta p_k(0) + p_k(\alpha))$	Degré d'exactitude
0	1	2	2	au moins 0
1	x	0	0	au moins 1
2	x^2	$2/3$	$2\alpha^2/3$	1 si $\alpha \neq 1/\sqrt{2}$, au moins 2 si $\alpha = 1/\sqrt{2}$
			Soit $\alpha = 1/\sqrt{2}$	
3	x^3	0	0	au moins 3
4	x^4	$2/5$	$1/3$	3

Si $\alpha \neq 1/\sqrt{2}$ la formule de quadrature est exacte pour les polynômes de degré au plus 1, si $\alpha = 1/\sqrt{2}$ la formule est exacte pour les polynômes de degré au plus 3.

1. N.B. : le polynôme P_i n'est pas le polynôme d'interpolation en x_{2i}, x_{2i+2} et $(x_{2i} + x_{2i+2})/2$ donc $\int_{x_{2i}}^{x_{2i+2}} P_i(x)\,\mathrm{d}x - F \neq 0$.

2. Par le changement de variable $y = x_i + (x+1)\frac{x_{i+1}-x_i}{2}$ on déduit la formule de quadrature

$$\begin{aligned}\int_{x_i}^{x_{i+1}} f(y)\,\mathrm{d}y &= \frac{x_{i+1}-x_i}{2}\int_{-1}^{1} f\left(x_i+(x+1)\frac{x_{i+1}-x_i}{2}\right)\mathrm{d}x \\ &\approx \frac{x_{i+1}-x_i}{3}\left[f\left(x_i+\left(1-\sqrt{\tfrac{1}{2}}\right)\frac{x_{i+1}-x_i}{2}\right)+f\left(x_i+\frac{x_{i+1}-x_i}{2}\right)+f\left(x_i+\left(1+\sqrt{\tfrac{1}{2}}\right)\frac{x_{i+1}-x_i}{2}\right)\right].\end{aligned}$$

3. Si $H = x_{i+1}-x_i = \frac{b-a}{n}$ (*i.e.* si on considère une subdivision de l'intervalle $[a;b]$ équirépartie) alors on trouve la formule de quadrature composite (*i.e.* sur n sous-intervalles et à $3n$ points)

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &\approx \frac{H}{3}\sum_{i=0}^{n-1}\left[f\left(x_i+H\left(1-\tfrac{1}{\sqrt{2}}\right)\right)+f\left(x_i+\tfrac{H}{2}\right)+f\left(x_i+H\left(1+\tfrac{1}{\sqrt{2}}\right)\right)\right] \\ &= \frac{H}{3}\sum_{i=0}^{n-1}\left[f\left(a+H\left(i+1-\tfrac{1}{\sqrt{2}}\right)\right)+f\left(x_i+\tfrac{H}{2}\right)+f\left(a+H\left(i+1+\tfrac{1}{\sqrt{2}}\right)\right)\right].\end{aligned}$$

Cette formule de quadrature est stable puisque tous les coefficients sont positifs.

Exercice 3.10

Soit f une fonction $\mathscr{C}^\infty(\mathbb{R},\mathbb{R})$. On se donne les points $\{x_i\}_{i=0}^{i=n}$ de subdivision de l'intervalle $[a;b]$: $x_i = a+ih$ avec $h = \frac{b-a}{n}$. Le but de l'exercice est de trouver une formule de quadrature à n points pour approcher l'intégrale définie

$$\int_a^b f(x)\,\mathrm{d}x. \tag{3.8}$$

On propose dans un premier temps (question 1 à 2) de construire la formule de quadrature à deux points :

$$\int_{-1}^{1} g(x)\,\mathrm{d}x \approx \frac{4}{3}g(-w/2)+\frac{2}{3}g(w), \tag{3.9}$$

où $0 < w \le 1$ est à déterminer.

1. Déterminer w pour que la formule de quadrature (3.9) soit exacte pour toute fonction g polynomiale de degré $m > 1$ et donner la plus grande valeur de m.
2. À l'aide d'un changement de variable affine, en déduire une formule de quadrature pour l'intégrale suivante :

$$\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x.$$

3. En déduire une formule de quadrature à $2n$ points, notée F, pour le calcul approché de (3.8). Cette formule de quadrature est-elle stable ?
4. Écrire l'algorithme du calcul de F.

SOLUTION.

1. On a

k	$p_k(x)=x^k$	$\int_{-1}^{1} p_k(x)\,\mathrm{d}x$	$\frac{2}{3}(2p_k(-w/2)+p_k(w))$	Degré d'exactitude
0	1	2	2	au moins 0
1	x	0	0	au moins 1
2	x^2	$2/3$	w^2	1 si $w \ne \sqrt{2/3}$, au moins 2 si $w=\sqrt{2/3}$
			Soit $w=\sqrt{2/3}$	
3	x^3	0	$-w^3/8$??????	2

Si $w \ne \sqrt{2/3}$ la formule de quadrature est exacte pour les polynômes de degré au plus 1, si $w = \sqrt{2/3}$ la formule est exacte pour les polynômes de degré au plus 2.

2. Par le changement de variable $y = x_i + (x+1)\frac{x_{i+1}-x_i}{2}$ on déduit la formule de quadrature

$$\begin{aligned}\int_{x_i}^{x_{i+1}} f(y)\,\mathrm{d}y &= \frac{x_{i+1}-x_i}{2}\int_{-1}^{1} f\left(x_i+(x+1)\frac{x_{i+1}-x_i}{2}\right)\mathrm{d}x \\ &\approx \frac{x_{i+1}-x_i}{3}\left[f\left(x_i+(1+\sqrt{\tfrac{2}{3}})\frac{x_{i+1}-x_i}{2}\right)+2f\left(x_i+(1-\sqrt{\tfrac{1}{6}})\frac{x_{i+1}-x_i}{2}\right)\right].\end{aligned}$$

3. Si $H = x_{i+1} - x_i = \frac{b-a}{n}$ (*i.e.* si on considère une subdivision de l'intervalle $[a;b]$ équirépartie) alors on trouve la formule de quadrature composite (*i.e.* sur n sous-intervalles et à $2n$ points)

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &\approx \frac{H}{3}\sum_{i=0}^{n-1}\left[f\left(x_i + H\left(1+\sqrt{\tfrac{2}{3}}\right)\right) + 2f\left(x_i + H\left(1-\sqrt{\tfrac{1}{6}}\right)\right)\right]\\ &= \frac{H}{3}\sum_{i=0}^{n-1}\left[f\left(a + H\left(i+1+\sqrt{\frac{2}{3}}\right)\right) + 2f\left(a + H\left(i+1-\sqrt{\frac{1}{6}}\right)\right)\right].\end{aligned}$$

   Cette formule de quadrature est stable puisque tous les coefficients sont positifs.

4. Algorithme du calcul de F :

   Require: f, a, $b > a$, $n > 0$
   $H \leftarrow \frac{b-a}{n}$
   $\alpha_1 \leftarrow a + H(1+\sqrt{2/3})$
   $\alpha_2 \leftarrow a + H(1-\sqrt{1/6})$
   for $i = 0$ to $n-1$ **do**
   $s \leftarrow s + f(\alpha_1 + iH) + 2f(\alpha_2 + iH)$
   end for
   return $I \leftarrow \frac{H}{3}s$

Exercice 3.11

1. Soit $0 < \alpha \le 1$ un nombre réel donné et soit $\omega_1, \omega_2, \omega_3$ trois nombres réels. Considérons la formule de quadrature

$$\int_{-1}^{1} g(t)\,\mathrm{d}t \approx \omega_1 g(-\alpha) + \omega_2 g(0) + \omega_3 g(\alpha).$$

   Calculer α, $\omega_1, \omega_2, \omega_3$ pour que le degré d'exactitude de la formule de quadrature soit de 5.

2. À l'aide d'un changement de variable affine, en déduire une formule de quadrature pour l'intégrale

$$\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x.$$

3. Soit $h = \frac{b-a}{n}$ et $x_i = a + ih$ pour $i = 0, \dots, n$. On subdivise l'intervalle $[a;b]$ en n intervalles $[x_i; x_{i+1}]$ de largeur h. En déduire la formule de quadrature composite pour le calcul approché de

$$\int_a^b f(x)\,\mathrm{d}x.$$

4. Écrire l'algorithme associé à cette formule de quadrature.

Solution.

1. On a

<table>
<tr><th>k</th><th>$p_k(x) = x^k$</th><th>$\int_{-1}^{1} p_k(x)\,\mathrm{d}x$</th><th>$\omega_1 p_k(-\alpha) + \omega_2 p_k(0) + \omega_3 p_k(\alpha)$</th></tr>
<tr><td>0</td><td>1</td><td>2</td><td>$\omega_1 + \omega_2 + \omega_3$</td></tr>
<tr><td colspan="4">Soit $\omega_2 = 2 - \omega_1 - \omega_3$</td></tr>
<tr><td>1</td><td>x</td><td>0</td><td>$-\alpha\omega_1 + \alpha\omega_3$</td></tr>
<tr><td colspan="4">Soit $\omega_3 = \omega_1$ et donc $\omega_2 = 2 - 2\omega_1$</td></tr>
<tr><td>2</td><td>x^2</td><td>$\frac{2}{3}$</td><td>$2\alpha^2\omega_1$</td></tr>
<tr><td colspan="4">Soit $\omega_1 = \frac{1}{3\alpha^2}$ et donc $\omega_3 = \frac{1}{3\alpha^2}$ et $\omega_2 = 2 - \frac{2}{3\alpha^2}$</td></tr>
<tr><td>3</td><td>x^3</td><td>0</td><td>0</td></tr>
<tr><td>4</td><td>x^4</td><td>$\frac{2}{5}$</td><td>$\frac{2}{3}\alpha^2$</td></tr>
<tr><td colspan="4">Soit $\alpha = \sqrt{\frac{3}{5}}$ et donc $\omega_1 = \omega_3 = \frac{5}{9}$ et $\omega_2 = \frac{8}{9}$</td></tr>
<tr><td>5</td><td>x^5</td><td>0</td><td>0</td></tr>
<tr><td>6</td><td>x^6</td><td>$\frac{2}{7}$</td><td>$\frac{6}{25}$</td></tr>
</table>

Si $\alpha = \sqrt{\frac{3}{5}}$, $\omega_1 = \omega_3 = \frac{5}{9}$ et $\omega_2 = \frac{8}{9}$ alors la formule est exacte pour les polynômes de degré au plus 5 (il s'agit de la formule de GAUSS-LEGENDRE à 3 points).
Remarquons que si on choisit $\alpha = 1$ on retrouve la formule de SIMPSON.

2. Soit $x = mt + q$, alors

$$\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x = m\int_{-1}^{1} f(mt+q)\,\mathrm{d}t \quad \text{avec} \quad \begin{cases} x_i = -m+q, \\ x_{i+1} = m+q, \end{cases}$$

d'où le changement de variable $x = x_i + (t+1)\frac{x_{i+1}-x_i}{2}$. On déduit la formule de quadrature (exacte sur l'espace des polynôme de degré au plus 5)

$$\begin{aligned}\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x &= \frac{x_{i+1}-x_i}{2}\int_{-1}^{1} f\left(x_i + (t+1)\frac{x_{i+1}-x_i}{2}\right)\mathrm{d}t \\ &\approx \frac{x_{i+1}-x_i}{18}\left[5f\left(x_i + \left(1-\sqrt{\frac{3}{5}}\right)\frac{x_{i+1}-x_i}{2}\right) + 8f\left(\frac{x_{i+1}+x_i}{2}\right) + 5f\left(x_i + \left(1+\sqrt{\frac{3}{5}}\right)\frac{x_{i+1}-x_i}{2}\right)\right].\end{aligned}$$

3. $h = \frac{b-a}{n} = x_{i+1} - x_i$ pour $i = 0,\dots,n$. On subdivise l'intervalle $[a;b]$ en n intervalles $[x_i;x_{i+1}]$ de largeur h. On trouve ainsi la formule de quadrature composite

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x = \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x &\approx \sum_{i=0}^{n-1}\frac{x_{i+1}-x_i}{18}\left[5f\left(x_i + \left(1-\sqrt{\frac{3}{5}}\right)\frac{x_{i+1}-x_i}{2}\right) + 8f\left(\frac{x_{i+1}+x_i}{2}\right) + 5f\left(x_i + \left(1+\sqrt{\frac{3}{5}}\right)\frac{x_{i+1}-x_i}{2}\right)\right] \\ &= \frac{h}{18}\sum_{i=0}^{n-1}\left[5f\left(x_i + \left(1-\sqrt{\frac{3}{5}}\right)h\right) + 8f\left(x_i + \frac{h}{2}\right) + 5f\left(x_i + \left(1+\sqrt{\frac{3}{5}}\right)h\right)\right] \\ &= \frac{h}{18}\sum_{i=0}^{n-1}\left[5f\left(a + \left(i+1-\sqrt{\frac{3}{5}}\right)h\right) + 8f\left(a + \left(i+\frac{1}{2}\right)h\right) + 5f\left(a + \left(i+1+\sqrt{\frac{3}{5}}\right)h\right)\right].\end{aligned}$$

4. Algorithme du calcul associé à cette formule de quadrature

Require: f, a, $b > a$, $n > 0$
 $h \leftarrow \frac{b-a}{n}$
 $c_1 \leftarrow a + \left(1 - \sqrt{3/5}\right)h$
 $c_2 \leftarrow a + \frac{1}{2}h$
 $c_3 \leftarrow a + \left(1 + \sqrt{3/5}\right)h$
 $s \leftarrow 0$
 for $i = 0$ to $n-1$ **do**
  $s \leftarrow s + 5f(c_1 + ih) + 8f(c_2 + ih) + 5f(c_3 + ih)$
 end for
 return $\frac{h}{18}s$

Exercice 3.12

Soit f une fonction de classe $\mathscr{C}^1([-1,1])$ et p le polynôme interpolateur d'Hermite (de degré ≤ 3) de f vérifiant

$$p(-1) = f(-1), \qquad p'(-1) = f'(-1), \qquad p(1) = f(1), \qquad p'(1) = f'(1).$$

1. Écrire le polynôme p.
2. En déduire la méthode d'intégration numérique élémentaire

$$\int_{-1}^{1} f(s)\,\mathrm{d}s \approx f(-1) + f(1) + \frac{1}{3}\left(f'(-1) - f'(1)\right).$$

3. Connaissant la formule sur $[-1;1]$, en déduire la formule de quadrature des trapèzes-Hermite sur l'intervalle $[a;b]$ par exemple grâce au changement de variable $y = a + (x+1)\frac{b-a}{2}$.

SOLUTION. 1. On a deux méthodes pour calculer le polynôme interpolateur d'Hermite :

Première méthode : le polynôme interpolateur d'Hermite s'écrit

$$p(x) = \sum_{i=0}^{n} y_i A_i(x) + y_i' B_i(x)$$

où

$$\begin{aligned}
A_i(x) &= (1-2(x-x_i)c_i)(L_i(x))^2,\\
B_i(x) &= (x-x_i)(L_i(x))^2,\\
L_i(x) &= \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j},\\
c_i &= \sum_{\substack{j=0\\ j\neq i}}^{n} \frac{1}{x_i-x_j}.
\end{aligned}$$

Pour $n=1$ on a alors

$$\begin{aligned}
p(x) = y_0\left(1-2(x-x_0)\left(\frac{1}{x_0-x_1}\right)\right)\left(\frac{(x-x_1)}{(x_0-x_1)}\right)^2 + y_0'(x-x_0)\left(\frac{(x-x_1)}{(x_0-x_1)}\right)^2\\
+ y_1\left(1-2(x-x_1)\left(\frac{1}{x_1-x_0}\right)\right)\left(\frac{(x-x_0)}{(x_1-x_0)}\right)^2 + y_1'(x-x_1)\left(\frac{(x-x_0)}{(x_1-x_0)}\right)^2.
\end{aligned}$$

Dans notre cas $x_0=-1$, $x_1=1$, $y_0=f(-1)$, $y_1=f(1)$, $y_0'=f'(-1)$, $y_1'=f'(1)$ donc

$$\begin{aligned}
p(x) &= \frac{1}{4}\left[f(-1)(x+2)(x-1)^2 + f'(-1)(x+1)(x-1)^2 + f(1)(2-x)(x+1)^2 + f'(1)(x-1)(x+1)^2\right]\\
&= \frac{1}{4}\left[f(-1)(x^3-3x+2) + f'(-1)(x^3-x^2-x+1) + f(1)(-x^3+3x+2) + f'(1)(x^3+x^2-x-1)\right]\\
&= \frac{2f(-1)+f'(-1)+2f(1)-f'(1)}{4} + \frac{3f(1)-3f(-1)-f'(-1)-f'(1)}{4}x\\
&+ \frac{f'(1)-f'(-1)}{4}x^2 + \frac{f(-1)+f'(-1)-f(1)+f'(1)}{4}x^3.
\end{aligned}$$

Le polynôme interpolateur d'HERMITE est donc le polynôme

$$p(x) = \alpha + \beta x + \gamma x^2 + \delta x^3$$

où

$$\alpha = \frac{2f(-1)+2f(1)+f'(-1)-f'(1)}{4}, \qquad \beta = \frac{-3f(-1)+3f(1)-f'(-1)-f'(1)}{4},$$
$$\gamma = \frac{-f'(-1)+f'(1)}{4}, \qquad \delta = \frac{f(-1)-f(1)+f'(-1)+f'(1)}{4}.$$

Deuxième méthode : le polynôme interpolateur d'Hermite est un polynôme de degré $2n+1$. On cherche donc un polynôme

$$p(x) = \alpha + \beta x + \gamma x^2 + \delta x^3$$

tel que

$$p(-1) = f(-1), \qquad p'(-1) = f'(-1), \qquad p(1) = f(1), \qquad p'(1) = f'(1),$$

c'est-à-dire tel que

$$\begin{cases}
\alpha-\beta+\gamma-\delta = f(-1),\\
\alpha+\beta+\gamma+\delta = f(1),\\
\beta-2\gamma+3\delta = f'(-1),\\
\beta+2\gamma+3\delta = f'(1).
\end{cases}$$

On obtient

$$\alpha = \frac{2f(-1)+2f(1)+f'(-1)-f'(1)}{4}, \qquad \beta = \frac{-3f(-1)+3f(1)-f'(-1)-f'(1)}{4},$$
$$\gamma = \frac{-f'(-1)+f'(1)}{4}, \qquad \delta = \frac{f(-1)-f(1)+f'(-1)+f'(1)}{4}.$$

2. En intégrant le polynôme ainsi trouvé on en déduit

$$\int_{-1}^{1} p(x)\,\mathrm{d}x = \left[\alpha x + \frac{\beta}{2}x^2 + \frac{\gamma}{3}x^3 + \frac{\delta}{4}x^4\right]_{-1}^{1}$$

$$
\begin{aligned}
&= 2\alpha + \frac{2}{3}\gamma = \frac{2f(-1)+2f(1)+f'(-1)-f'(1)}{2} + \frac{-f'(-1)+f'(1)}{6} \\
&= \frac{6f(-1)+6f(1)+3f'(-1)-3f'(1)-f'(-1)+f'(1)}{6} \\
&= f(-1)+f(1)+\frac{1}{3}(f'(-1)-f'(1)).
\end{aligned}
$$

Remarque : la formule est au moins exacte de degré 3 par construction. Elle n'est pas exacte de degré supérieure à 3 car si $f(x)=x^4$ alors

$$
\begin{gathered}
\int_{-1}^{1} f(x)\,\mathrm{d}x = \left[\frac{1}{5}x^5\right]_{-1}^{1} = \frac{2}{5} = \frac{6}{15} \\
\neq \\
f(-1)+f(1)+\frac{1}{3}(f'(-1)-f'(1)) = 1+1+\frac{1}{3}(4+4) = \frac{14}{3} = \frac{70}{15}
\end{gathered}
$$

3. Connaissant la formule sur $[-1;1]$, on en déduit la formule sur un intervalle $[a;b]$ quelconque par le changement de variable $y=a+(x+1)\frac{b-a}{2}$ qui donne[2]

$$
\begin{aligned}
\int_a^b f(y)\mathrm{d}y &= \frac{b-a}{2}\int_{-1}^{1} f\left(a+(x+1)\frac{b-a}{2}\right)\mathrm{d}x \\
&= \frac{b-a}{2}\left[f(a)+f(b)+\frac{b-a}{6}(f'(a)-f'(b))\right] \\
&= \frac{b-a}{2}(f(a)+f(b)) + \frac{(b-a)^2}{12}(f'(a)-f'(b)).
\end{aligned}
$$

Exercice 3.13

Soit f une fonction $\mathscr{C}^\infty([0;1],\mathbb{R})$. On se donne les points $\{x_i\}_{i=0}^{i=n}$ de subdivision de l'intervalle $[0;1]$: $x_i = ih$ avec $h=\frac{1}{n}$. Le but de l'exercice est de trouver une formule de quadrature pour approcher

$$
\int_0^1 f(x)\,\mathrm{d}x. \tag{3.10}
$$

1. Soit i un entier fixé ($1\le i\le n-1$). Trouver m_i un point du segment $[x_i;x_{i+1}]$ et a, b et c trois coefficients réels tels que la formule de quadrature suivante, sur l'intervalle $[x_i;x_{i+1}]$, soit exacte pour p un polynôme de degré le plus haut possible :
$$
\int_{x_i}^{x_{i+1}} p(x)\,\mathrm{d}x = ap(x_i)+bp(m_i)+cp(x_{i+1}).
$$
2. En déduire en fonction de a, b et c la formule de quadrature $Q(f)$
$$
Q(f) = \sum_{i=0}^{n}\alpha_i f(x_i) + \sum_{i=0}^{n-1}\beta_i f(m_i)
$$
pour le calcul approché de 3.10 construite sur la formule de quadrature précédente pour chaque intervalle du type $[x_i;x_{i+1}]$. Cette formule de quadrature est-elle stable ?
3. On rappelle que si p interpole f en k points $y_1<y_2<\dots<y_k$, on a l'estimation d'erreur
$$
\forall x\in[y_1;y_k], \qquad |f(x)-p(x)| \le \frac{\sup_{\xi\in[y_1;y_k]}|f^{(k)}(\xi)|}{k!}\prod_{j=1}^{k}(x-y_j).
$$
En déduire une estimation de l'erreur de quadrature entre (3.10) et Q
$$
E(h) = \int_0^1 f(x)\,\mathrm{d}x - Q(f).
$$
La dépendance en h dans cette estimation d'erreur est-elle optimale ?
4. Écrire l'algorithme qui calcule $Q(f)$.

2. Rappel : si $y=a+(x+1)\frac{b-a}{2}$ alors $\mathrm{d}y=\frac{b-a}{2}\mathrm{d}x$ et $f'(y)=\frac{b-a}{2}f'(x)$.

SOLUTION.

1. Pour simplifier le calcul, on se ramène à l'intervalle $[0;1]$. Soit x un élément de l'intervalle $[x_i;x_{i+1}]$ et y un élément de l'intervalle $[0;1]$. On transforme l'intervalle $[x_i;x_{i+1}]$ dans l'intervalle $[0;1]$ par le changement de variable affine $y=\frac{1}{x_{i+1}-x_i}x-\frac{x_i}{x_{i+1}-x_i}$. On note $h=x_{i+1}-x_i$. Alors $y=\frac{x-x_i}{h}$ et on a $\int_0^1 f(y)\,\mathrm{d}y=\frac{1}{h}\int_{x_i}^{x_{i+1}} f(\frac{x-x_i}{h})dx$. Comme $\int_{x_i}^{x_{i+1}} f(t)\,\mathrm{d}t\approx af(x_i)+bf(m_i)+cf(x_{i+1})$, alors $\int_0^1 f(y)\,\mathrm{d}y=\frac{1}{h}\int_{x_i}^{x_{i+1}} f(\frac{x-x_i}{h})\,\mathrm{d}x\approx\frac{1}{h}\left(af(0)+bf(\frac{m_i-x_i}{h})+cf(1)\right)$. On note alors $A=\frac{a}{h}$, $B=\frac{b}{h}$, $C=\frac{c}{h}$, $M=\frac{m_i-x_i}{h}$ d'où $m_i=(1-M)x_i+Mx_{i+1}$. Rechercher a, b, c et m_i revient à chercher A, B, C et M avec
$$\begin{cases}m_i=(1-M)x_i+Mx_{i+1},\\ a=Ah,\\ b=Bh,\\ c=Ch\end{cases}$$
tels que
$$\int_0^1 p(x)\,\mathrm{d}x=Ap(0)+Bp(M)+Cp(1),$$
où $p(x)$ est un polynôme. Si $p\in\mathbb{P}^3$ (*i.e.* si $p(x)=d_0+d_1x+d_2x^2+d_3x^3$) on a
$$\begin{array}{ccc}\int_0^1 p(x)\,\mathrm{d}x & = & Ap(0)+Bp(M)+Cp(1)\\ \| & & \|\\ \left[d_0x+\frac{d_1}{2}x^2+\frac{d_2}{3}x^3+\frac{d_3}{4}x^4\right]_0^1 & & Ad_0+Bd_0+Cd_0\\ \| & & \|\\ d_0+\frac{d_1}{2}+\frac{d_2}{3}+\frac{d_3}{4} & & (A+B+C)d_0+(BM+C)d_1+(BM^2+C)d_2+(BM^3+C)d_3\end{array}$$
Par conséquent, pour que la formule soit exacte de degré au moins 3 il faut que
$$\begin{cases}A+B+C=1\\ BM+C=\frac{1}{2}\\ BM^2+C=\frac{1}{3}\\ BM^3+C=\frac{1}{4}\end{cases}\iff\begin{cases}A+B+C=1\\ BM=\frac{1}{2}-C\\ (\frac{1}{2}-C)M=\frac{1}{3}-C\\ (\frac{1}{3}-C)M=\frac{1}{4}-C\end{cases}\iff\begin{cases}A=\frac{1}{6},\\ B=\frac{2}{3},\\ C=\frac{1}{6},\\ M=\frac{1}{2}.\end{cases}$$
La méthode
$$\int_0^1 f(x)\,\mathrm{d}x=\frac{1}{6}f(0)+\frac{2}{3}f\left(\frac{1}{2}\right)+\frac{1}{6}f(1),$$
est exacte pour tout polynôme de degré au moins 3.
Soit maintenant $f(x)=x^4$. On a
$$\int_0^1 f(x)\,\mathrm{d}x=\left[\frac{x^5}{5}\right]_0^1=\frac{1}{5}$$
mais
$$\frac{1}{6}f(0)+\frac{2}{3}f\left(\frac{1}{2}\right)+\frac{1}{6}f(1)=\frac{1}{6}+\frac{2}{3}\left(\frac{1}{2}\right)^4+\frac{1}{6}=\frac{5}{24},$$
donc la formule de quadrature est exacte de degré 3.
Si on revient aux variables initiales, on trouve
$$\begin{cases}m_i=\frac{1}{2}x_i+\frac{1}{2}x_{i+1},\\ a=\frac{1}{6}h,\\ b=\frac{2}{3}h,\\ c=\frac{1}{6}h\end{cases}$$

2. L'intégrale
$$\int_0^1 f(x)\mathrm{d}x=\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x$$
peut être calculée numériquement en utilisant la formule précédente pour approcher chaque intégrale
$$\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x\approx\frac{h}{6}\left[f(x_i)+4f\left(\frac{x_i+x_{i+1}}{2}\right)+f(x_{i+1})\right].$$

On obtient ainsi

$$\begin{aligned}
\int_0^1 f(x)\mathrm{d}x &= \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x\\
&\approx \sum_{i=0}^{n-1}\frac{h}{6}\left[f(x_i)+4f\left(\frac{x_i+x_{i+1}}{2}\right)+f(x_{i+1})\right]\\
&=\frac{h}{6}\left[\sum_{i=0}^{n-1}f(x_i)+\sum_{i=0}^{n-1}f(x_{i+1})+4\sum_{i=0}^{n-1}f\left(\frac{x_i+x_{i+1}}{2}\right)\right]\\
&=\frac{h}{6}\left[f(a)+f(b)+2\sum_{i=1}^{n-1}f(x_i)+4\sum_{i=0}^{n-1}f\left(\frac{x_i+x_{i+1}}{2}\right)\right]\\
&=\sum_{i=0}^{n}\alpha_i f(x_i)+\sum_{i=0}^{n-1}\beta_i f(m_i)=Q(f)\qquad\text{avec}\quad \beta_i=\frac{2h}{3},\quad \alpha_i=\begin{cases}\frac{h}{3} & \text{si } i=1,\dots,n-1,\\ \frac{h}{6} & \text{sinon.}\end{cases}
\end{aligned}$$

Cette formule de quadrature est stable puisque tous les coefficients α_i et β_i sont positifs et on a

$$\sum_{i=0}^{n}\alpha_i+\sum_{i=0}^{n-1}\beta_i=\frac{h}{6}+\sum_{i=1}^{n-1}\frac{h}{3}+\frac{h}{6}+\sum_{i=0}^{n-1}\frac{2h}{3}=\frac{1}{n}\left(\frac{1}{6}+\frac{1}{3}\sum_{i=1}^{n-1}1+\frac{1}{6}+\frac{2}{3}\sum_{i=0}^{n-1}1\right)=1.$$

3. On reconnait la formule de Cavalieri-Simpson : remarquons alors que $Q(f)=\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}}p(x)\mathrm{d}x$ avec p le polynôme qui interpole $(x_i,f(x_i))$, $(m_i,f(m_i))$ et $(x_{i+1},f(x_{i+1}))$. Par conséquent l'erreur de quadrature entre (3.10) et Q est

$$\begin{aligned}
|E(h)| &= \left|\int_0^1 f(x)\,\mathrm{d}x - Q(f)\right|\\
&=\left|\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x-\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}}p(x)\mathrm{d}x\right|\\
&\le \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}}\left|f(x)-p(x)\right|\mathrm{d}x\\
&\le \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}}\frac{\sup_{\xi\in[x_i;x_{i+1}]}|f'''(\xi)|}{6}(x-x_i)(x-m_i)(x-x_{i+1})\mathrm{d}x\\
&\le Dh^4.
\end{aligned}$$

4. Algorithme

Require: $x\mapsto f$
Require: $n>0$
  $a \leftarrow \frac{1}{6n}$
  $b \leftarrow \frac{2}{3n}$
  $c \leftarrow \frac{1}{6n}$
  $I \leftarrow af(0)$
  for $i=1$ to $n-1$ **do**
    $I \leftarrow I+(a+c)f\left(\frac{i}{n}\right)+bf\left(\frac{i-\frac{1}{2}}{n}\right)$
  end for
  return $I \leftarrow I+cf(1)+bf\left(\frac{n-\frac{1}{2}}{n}\right)$

4. Systèmes linéaires

Résoudre l'ensemble d'équations linéaires $\mathbb{A}\mathbf{x} = \mathbf{b}$

Définition : système linéaire

Soit $n, p, \geq 1$ des entiers. Un SYSTÈME LINÉAIRE $n \times p$ est un ensemble de n équations linéaires à p inconnues de la forme

$$(S) \quad \begin{cases} a_{11}x_1 & + & \dots & + & a_{1p}x_p & = & b_1, \\ \vdots & & & & \vdots & & \vdots \\ a_{n1}x_1 & + & \dots & + & a_{np}x_p & = & b_n. \end{cases}$$

▷ Les COEFFICIENTS a_{ij} et les SECONDES MEMBRES b_i sont des éléments donnés de $\mathbb{R}$. Les INCONNUES $x_1, x_2, \dots, x_p$ sont à chercher dans $\mathbb{R}$.
▷ Le SYSTÈME HOMOGÈNE associé à (S) est le système obtenu en remplaçant les b_i par 0.
▷ Une SOLUTION de (S) est un p-uplet $(x_1, x_2, \dots, x_p)$ qui vérifie simultanément les n équations de (S). Résoudre (S) signifie chercher toutes les solutions.
▷ Un système est IMPOSSIBLE, ou incompatible, s'il n'admet pas de solution. Un système est POSSIBLE, ou compatible, s'il admet une ou plusieurs solutions.
▷ Deux systèmes sont ÉQUIVALENTS s'ils admettent les mêmes solutions.

Écriture matricielle

Si on note

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \qquad \mathbb{A} = \begin{pmatrix} a_{11} & \dots & a_{1p} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{np} \end{pmatrix}$$

le système (S) est équivalent à l'écriture matricielle $\mathbb{A}\mathbf{x} = \mathbf{b}$.

Dans ce chapitre, nous ne traiterons que des systèmes linéaires carrés d'ordre n à coefficients réels, autrement dit $\mathbb{A} = (a_{i,j}) \in \mathbb{R}^{n \times n}$ et $\mathbf{b} = (b_i) \in \mathbb{R}^n$. Dans ce cas, on est assuré de l'existence et de l'unicité de la solution si une des conditions équivalentes suivantes est remplie :

1. $\mathbb{A}$ est inversible (*i.e.* $\det(\mathbb{A}) \neq 0$) ;
2. le système homogène $\mathbb{A}\mathbf{x} = \mathbf{0}$ admet seulement la solution nulle.

La solution du système peut alors être calculée par la formule de CRAMER (voir annexe). Cependant cette formule est d'une utilité pratique limitée à cause du calcul des déterminants qui est très couteux. Pour cette raison, des méthodes numériques alternatives aux formules de CRAMER ont été développées. Elles sont dites directes si elles fournissent la solution du système en un nombre fini d'étapes, et itératives si elles nécessitent (théoriquement) un nombre infini d'étapes. Notons dès à présent que le choix entre une méthode directe et une méthode itérative pour la résolution d'un système dépend non seulement de l'efficacité théorique des algorithmes, mais aussi du type de matrice, des capacités de stockage en mémoire et enfin de l'architecture de l'ordinateur.

Systèmes mal conditionnés

Considérons le système de deux équations à deux inconnues suivant :

$$\begin{cases} 3.218613x_1 + 6.327917x_2 = 10.546530, \\ 3.141592x_1 + 4.712390x_2 = 7.853982. \end{cases}$$

Ce système et non singulier et sa solution est donnée par $x_1 = x_2 = 1$. Considérons maintenant un système d'équations voisin (le carré indique un changement de décimale) :

$$\begin{cases} 3.21861\boxed{1}x_1 + 6.327917x_2 = 10.546530, \\ 3.14159\boxed{4}x_1 + 4.712390x_2 = 7.85398\boxed{0}. \end{cases}$$

Ce système et non singulier et sa solution est donnée par $x_1 = -5$, $x_2 = 5$.

On voit donc que, bien que ces deux systèmes soient voisins, leurs solutions sont très différentes. On parle dans ce cas de systèmes mal conditionnés. Résoudre un système mal conditionné avec un ordinateur peut être une affaire délicate si l'ordinateur calcule avec trop peu de chiffres significatifs. Dans l'exemple précédent nous observons que, si l'ordinateur ne retient que 6 chiffres significatifs, il complètement inespéré d'obtenir une solution raisonnablement proche de la solution.

Conditionnement d'une matrice

Le conditionnement d'une matrice $\mathbb{A} \in \mathbb{R}^{n\times n}$ non singulière est défini par

$$K(\mathbb{A}) = \|\mathbb{A}\|\|\mathbb{A}^{-1}\|(\geq 1),$$

où $\|\cdot\|$ est une norme matricielle subordonnée. En général, $K(\mathbb{A})$ dépend du choix de la norme ; ceci est signalé en introduisant un indice dans la notation. Par exemple, on a les deux normes matricielles suivantes :

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

Cas particulier

Si $\mathbb{A}$ est symétrique et définie positive [a],

$$K_2(\mathbb{A}) = \|\mathbb{A}\|_2\|\mathbb{A}^{-1}\|_2 = \frac{\lambda_{\max}}{\lambda_{\min}}$$

où $\lambda_{\max}$ (resp. $\lambda_{\min}$) est la plus grande (resp. petite) valeur propre de $\mathbb{A}$.

a. $\mathbb{A} \in \mathbb{R}^{n\times n}$ est
- ▷ symétrique si $a_{ij} = a_{ji}$ pour tout $i, j = 1,\dots,n$,
- ▷ définie positive si pour tout vecteurs $\mathbf{x} \in \mathbb{R}^n$ avec $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x}^T\mathbb{A}\mathbf{x} > 0$.

Considérons un système non singulier $\mathbb{A}\mathbf{x} = \mathbf{b}$. Si $\delta\mathbf{b}$ est une perturbation de $\mathbf{b}$ et si on résout $\mathbb{A}\mathbf{y} = \mathbf{b} + \delta\mathbf{b}$, on obtient par linéarité $\mathbf{y} = \mathbf{x} + \delta\mathbf{x}$ avec $\mathbb{A}\delta\mathbf{x} = \delta\mathbf{b}$. La question est de savoir s'il est possible de majorer l'erreur relative $\|\delta\mathbf{x}\|/\|\mathbf{x}\|$ sur la solution du système en fonction de l'erreur relative $\|\delta\mathbf{b}\|/\|\mathbf{b}\|$ commise sur le second membre. Il est possible de démontrer que

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq K(\mathbb{A})\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}$$

où $K(\mathbb{A})$ est le nombre de conditionnement de la matrice $\mathbb{A}$. On voit alors que plus le conditionnement de la matrice est grand, plus la solution du système linéaire est sensible aux perturbations des données. Cependant, le fait qu'un système linéaire soit bien conditionné n'implique pas nécessairement que sa solution soit calculée avec précision. Il faut en plus utiliser des algorithmes stables. Inversement, le fait d'avoir une matrice avec un grand conditionnement n'empêche pas nécessairement le système global d'être bien conditionné pour des choix particuliers du second membre.

Si $\|\delta\mathbf{b}\|/\|\mathbf{b}\|$ est de l'ordre de la précision relative $\eta = 10^{-p}$ du calculateur, alors $\|\delta\mathbf{x}\|/\|\mathbf{x}\|$ pourrait, au pire, être égal à

$$K(\mathbb{A})\eta = 10^{\log_{10}(K(\mathbb{A}))}10^{-p} = 0^{\log_{10}(K(\mathbb{A})-p)}.$$

Si on calcul la solution du système linéaire avec un ordinateur à p chiffres significatifs en valeur décimale, on ne pourra pas garantir à priori plus de

$$E(p - \log_{10}(K(\mathbb{A})))$$

chiffre significatifs sur la solution. Si on applique cette règle au système linéaire de l'exemple, il est facile de vérifier que $K(\mathbb{A}) \simeq 10^7$, par conséquent nous pouvons perdre jusqu'à 7 chiffres significatifs lors de sa résolution. Il faut donc un ordinateur calculant avec 10 chiffres significatifs pour être sûr d'obtenir les 3 premiers chiffres de la solution.

Méthode (directe) d'élimination de Gauss et factorisation $\mathbb{LU}$

Matrices et systèmes triangulaires

On dit qu'une matrice carrée $\mathbb{A} = (a_{ij})_{1\le i,j\le n}$ est TRIANGULAIRE SUPÉRIEURE (respectivement triangulaire INFÉRIEURE) si $i > j \implies a_{ij} = 0$ (resp. si $i < j \implies a_{ij} = 0$).
Si la matrice est triangulaire supérieure (resp. triangulaire inférieure), on dira que le système linéaire est un système triangulaire supérieur (resp. triangulaire inférieur).
Pour résoudre le système triangulaire $\mathbb{A}\mathbf{x} = \mathbf{b}$,

▷ si $\mathbb{A}$ est une matrice triangulaire inférieure, on a $x_1 = \frac{b_1}{a_{11}}$ et on déduit les inconnues $x_2, x_3, \dots x_n$ grâce à la relation $x_i = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j\right)$;

▷ si $\mathbb{A}$ est une matrice triangulaire supérieure on a $x_n = \frac{b_n}{a_{nn}}$ et on déduit les inconnues $x_{n-1}, x_{n-2}, \dots x_1$ grâce à la relation $x_i = \frac{1}{a_{ii}}\left(b_i - \sum_{j=i+1}^{n} a_{ij}x_j\right)$.

Propriété

Le déterminant d'une matrice triangulaire est égal au produit des éléments diagonaux.

La méthode du pivot de Gauss transforme le système $\mathbb{A}\mathbf{x} = \mathbf{b}$ en un système équivalent (c'est-à-dire ayant la même solution) de la forme $\mathbb{U}\mathbf{x} = \mathbf{y}$, où $\mathbb{U}$ est une matrice triangulaire supérieure et $\mathbf{y}$ est un second membre convenablement modifié. Enfin on résout le système triangulaire $\mathbb{U}\mathbf{x} = \mathbf{y}$.

Méthode du pivot de Gauss

Soit $\mathbb{A} = (a_{ij})_{\substack{1\le i\le n\\ 1\le j\le p}}$ la matrice des coefficients du système $\mathbb{A}\mathbf{x} = \mathbf{b}$. En permutant éventuellement deux lignes du système, on peut supposer $a_{ii} \neq 0$ (appelé pivot de l'étape i).
Étape j : pour $i > k$, les transformations

$$L_i \leftarrow L_i - \frac{a_{ik}}{a_{kk}} L_k$$

éliminent l'inconnue x_k dans les lignes L_i.
En réitérant le procédé pour i de 1 à n, on aboutit à un système triangulaire supérieur.

Exemple

Soit le système linéaire

$$\begin{cases} x_1 + 2x_2 + 3x_3 + 4x_4 = 1, \\ 2x_1 + 3x_2 + 4x_3 + x_4 = 2, \\ 3x_1 + 4x_2 + x_3 + 2x_4 = 3, \\ 4x_1 + x_2 + 2x_3 + 3x_4 = 4. \end{cases}$$

1. Résolution par la méthode du pivot de Gauss :

$$\begin{cases} x_1 + 2x_2 + 3x_3 + 4x_4 = 1 \\ 2x_1 + 3x_2 + 4x_3 + x_4 = 2 \\ 3x_1 + 4x_2 + x_3 + 2x_4 = 3 \\ 4x_1 + x_2 + 2x_3 + 3x_4 = 4 \end{cases} \xrightarrow{\substack{L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - 3L_1 \\ L_4 \leftarrow L_4 - 4L_1}} \begin{cases} x_1 + 2x_2 + 3x_3 + 4x_4 = 1 \\ -x_2 - 2x_3 - 7x_4 = 0 \\ -2x_2 - 8x_3 - 10x_4 = 0 \\ -7x_2 - 10x_3 - 13x_4 = 0 \end{cases}$$

$$\xrightarrow{\substack{L_3 \leftarrow L_3 - 2L_2 \\ L_4 \leftarrow L_4 - 7L_2}} \begin{cases} x_1 + 2x_2 + 3x_3 + 4x_4 = 1 \\ -x_2 - 2x_3 - 7x_4 = 0 \\ -4x_3 + 4x_4 = 0 \\ 4x_3 + 36x_4 = 0 \end{cases} \xrightarrow{L_4 \leftarrow L_4 + L_3} \begin{cases} x_1 + 2x_2 + 3x_3 + 4x_4 = 1 \\ -x_2 - 2x_3 - 7x_4 = 0 \\ -4x_3 + 4x_4 = 0 \\ 40x_4 = 0 \end{cases}$$

donc $x_4 = 0$, $x_3 = 0$, $x_2 = 0$ et $x_1 = 1$.

2. Résolution par la méthode du pivot de Gauss en écriture matricielle :

$$[\mathbb{A}|\mathbf{b}] = \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 2 & 3 & 4 & 1 & 2 \\ 3 & 4 & 1 & 2 & 3 \\ 4 & 1 & 2 & 3 & 4 \end{array}\right) \xrightarrow{\substack{L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - 3L_1 \\ L_4 \leftarrow L_4 - 4L_1}} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & -2 & -8 & -10 & 0 \\ 0 & -7 & -10 & -13 & 0 \end{array}\right)$$

$$\xrightarrow[L_4 \leftarrow L_4 - 7L_2]{L_3 \leftarrow L_3 - 2L_2} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 4 & 36 & 0 \end{array}\right) \xrightarrow{L_4 \leftarrow L_4 + L_3} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 0 & 40 & 0 \end{array}\right)$$

donc $x_4 = 0$, $x_3 = 0$, $x_2 = 0$ et $x_1 = 1$.

Si on a plusieurs systèmes dont seul le second membre change, il peut être utile de factoriser une fois pour toute la matrice $\mathbb{A}$ et résoudre ensuite des systèmes triangulaires.

Algorithme de factorisation $\mathbb{L}\mathbb{U}$ sans pivot

Soit le système linéaire $\mathbb{A}\mathbf{x} = \mathbf{b}$.

1. On factorise la matrice $\mathbb{A} \in \mathbb{R}^{n\times n}$ sous la forme d'un produit de deux matrices $\mathbb{A} = \mathbb{L}\mathbb{U}$. Les termes non nuls de $\mathbb{U}$ et les termes non nuls en-dessous de la diagonale principale de $\mathbb{L}$ sont mémorisés encore dans la matrice $\mathbb{A}$ et sont ainsi calculées :

   for $k = 1$ to $n-1$ **do**
     for $i = k+1$ to n **do**
       $a_{ik} \leftarrow \dfrac{a_{ik}}{a_{kk}}$
       for $j = k+1$ to n **do**
         $a_{ij} \leftarrow a_{ij} - a_{ik}a_{kj}$
       end for
     end for
   end for

2. Résoudre le système linéaire consiste simplement à résoudre successivement

   2.1. le système triangulaire inférieur $\mathbb{L}\mathbf{y} = \mathbf{b}$: les éléments non nuls de la matrice triangulaire inférieure $\mathbb{L}$ sont donné par $\ell_{ij} = a_{ij}$ pour $i = 1,\dots,n$ et $j = 1,\dots,i-1$ et $\ell_{ii} = 1$ pour tout $i = 1,\dots,n$, donc l'algorithme s'écrit

   $y_1 \leftarrow b_1$
   for $i = 2$ to n **do**
     $s_i \leftarrow 0$
     for $j = 1$ to $i-1$ **do**
       $s_i \leftarrow s_i + a_{ij}x_j$
     end for
     $y_i \leftarrow b_i - s_i$
   end for

   2.2. le système triangulaire supérieure $\mathbb{U}\mathbf{x} = \mathbf{y}$: les éléments non nuls de la matrice triangulaire supérieure $\mathbb{U}$ sont donné par $u_{ij} = a_{ij}$ pour $i = 1,\dots,n$ et $j = i,\dots,n$, donc l'algorithme s'écrit

   $x_n \leftarrow \dfrac{y_n}{a_{nn}}$
   for $i = n-1$ to 1 by -1 **do**
     $s_i \leftarrow 0$
     for $j = 1$ to $i-1$ **do**
       $s_i \leftarrow s_i + a_{ij}y_j$
     end for
     $x_i \leftarrow \dfrac{y_i - s_i}{a_{ii}}$
   end for

Attention

Pour une matrice quelconque $\mathbb{A} \in \mathbb{R}^{n\times n}$, la factorisation $\mathbb{L}\mathbb{U}$ existe et est unique si et seulement si les sous-matrices principales $\mathbb{A}_i$ de $\mathbb{A}$ d'ordre $i = 1,\dots,n-1$ (celles que l'on obtient en restreignant $\mathbb{A}$ à ses i premières lignes et colonnes) ne sont pas singulières (autrement dit si les mineurs principaux, *i.e.* les déterminants des sous-matrices principales, sont non nuls).

On peut identifier des classes de matrices particulières pour lesquelles les hypothèses de cette proposition sont satisfaites :

Proposition

Si la matrice $\mathbb{A} \in \mathbb{R}^{n\times n}$ est symétrique et définie positive ou si est à diagonale dominante [a] alors la factorisation $\mathbb{L}\mathbb{U}$ existe et est unique.

a. $\mathbb{A} \in \mathbb{R}^{n\times n}$ est

- ▷ symétrique si $a_{ij} = a_{ji}$ pour tout $i, j = 1, \dots, n$,
- ▷ définie positive si pour tout vecteurs $\mathbf{x} \in \mathbb{R}^n$ avec $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x}^T \mathbb{A}\mathbf{x} > 0$,
- ▷ à diagonale dominante par lignes si $|a_{ii}| \geq \sum_{\substack{j=1\\ j\neq i}}^n |a_{ij}|$, pour $i = 1, \dots, n$ (à diagonale dominante stricte par lignes si l'inégalité est stricte),
- ▷ à diagonale dominante par colonnes si $|a_{ii}| \geq \sum_{\substack{j=1\\ j\neq i}}^n |a_{ji}|$, pour $i = 1, \dots, n$ (à diagonale dominante stricte par colonnes si l'inégalité est stricte),

Une technique qui permet d'effectuer la factorisation $\mathbb{LU}$ pour toute matrice $\mathbb{A}$ inversible, même quand les hypothèses de cette proposition ne sont pas vérifiées, est la méthode du pivot par ligne : il suffit d'effectuer une permutation convenable des lignes de la matrice originale $\mathbb{A}$ à chaque étape k où un terme diagonal a_{kk} s'annule.

Algorithme de Gauss avec pivot

Dans la méthode d'élimination de Gauss les pivot $a_{kk}^{(k)}$ doivent être différents de zéro. Si la matrice est inversible mais un pivot est zéro (ou numériquement proche de zéro), on peut permuter deux lignes avant de poursuivre la factorisation. L'algorithme modifié s'écrit alors

for $k = 1$ to $n-1$ **do**
  for $i = k+1$ to n **do**
    Chercher $\bar{r}$ tel que $|a_{\bar{r}k}^{(k)}| = \max_{r=k,\dots,n} |a_{rk}^{(k)}|$ et échanger la ligne k avec la ligne $\bar{r}$
    $\ell_{ik} \leftarrow \dfrac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$
    for $j = k+1$ to n **do**
      $a_{ij}^{(k+1)} \leftarrow a_{ij}^{(k)} - \ell_{ik}^{(k)} a_{kj}^{(k)}$
    end for
  end for
end for

Une fois calculées les matrices $\mathbb{L}$ et $\mathbb{U}$ et la matrice des permutations $\mathbb{P}$ (*i.e.* la matrice telle que $\mathbb{PA} = \mathbb{LU}$), résoudre le système linéaire consiste simplement à résoudre successivement le système triangulaire inférieur $\mathbb{L}\mathbf{y} = \mathbb{P}\mathbf{b}$ puis le système triangulaire supérieure $\mathbb{U}\mathbf{x} = \mathbf{y}$.

Déterminant

La factorisation $\mathbb{LU}$ permet de calculer le déterminant de $\mathbb{A}$ en $O(n^3)$ car $\det(\mathbb{A}) = \det(\mathbb{L})\det(\mathbb{U}) = \prod_{k=1}^n u_{kk}$.

Inverse d'une matrice

Le calcul explicite de l'inverse d'une matrice peut être effectué en utilisant la factorisation $\mathbb{LU}$ comme suit. En notant $\mathbb{X}$ l'inverse d'une matrice régulière $\mathbb{A} \in \mathbb{R}^{n\times n}$, les vecteurs colonnes de $\mathbb{X}$ sont les solutions des systèmes linéaires

$$\mathbb{A}\mathbf{x}_i = \mathbf{e}_i, \qquad \text{pour } i = 1, \dots, n.$$

En supposant que $\mathbb{PA} = \mathbb{LU}$, où $\mathbb{P}$ est la matrice de changement de pivot partiel, on doit résoudre $2n$ systèmes triangulaires de la forme

$$\mathbb{L}\mathbf{y}_i = \mathbb{P}\mathbf{e}_i, \qquad \mathbb{U}\mathbf{x}_i = \mathbf{y}_i, \qquad \text{pour } i = 1, \dots, n.$$

c'est-à-dire une suite de systèmes linéaires ayant la même matrice mais des seconds membres différents.

Exemple

Soit les systèmes linéaires

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \qquad \text{et} \qquad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 10 \\ 10 \\ 10 \\ 10 \end{pmatrix}.$$

1. Résoudre les systèmes linéaires par la méthode du pivot de Gauss.
2. Factoriser la matrice $\mathbb{A}$ (sans utiliser la technique du pivot) et résoudre les systèmes linéaires.
3. Calculer le déterminant de $\mathbb{A}$.
4. Calculer $\mathbb{A}^{-1}$.

1. Résolution par la méthode du pivot de Gauss du premier système

$$[\mathbb{A}|\mathbf{b}] = \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 2 & 3 & 4 & 1 & 2 \\ 3 & 4 & 1 & 2 & 3 \\ 4 & 1 & 2 & 3 & 4 \end{array}\right) \xrightarrow[L_4 \leftarrow L_4 - 4L_1]{\substack{L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - 3L_1}} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & -2 & -8 & -10 & 0 \\ 0 & -7 & -10 & -13 & 0 \end{array}\right) \xrightarrow{\substack{L_3 \leftarrow L_3 - 2L_2 \\ L_4 \leftarrow L_4 - 7L_2}} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 4 & 36 & 0 \end{array}\right)$$

$$\xrightarrow{L_4 \leftarrow L_4 + L_3} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 0 & 40 & 0 \end{array}\right)$$

donc

$$x_4 = 0, \quad x_3 = 0, \quad x_2 = 0, \quad x_1 = 1.$$

Résolution par la méthode du pivot de Gauss du second système

$$(\mathbb{A}|\mathbf{b}) = \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 2 & 3 & 4 & 1 & 10 \\ 3 & 4 & 1 & 2 & 10 \\ 4 & 1 & 2 & 3 & 10 \end{array}\right) \xrightarrow[L_4 \leftarrow L_4 - 4L_1]{\substack{L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - 3L_1}} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 0 & -1 & -2 & -7 & -10 \\ 0 & -2 & -8 & -10 & -20 \\ 0 & -7 & -10 & -13 & -30 \end{array}\right) \xrightarrow[L_4 \leftarrow L_4 - 7L_2]{L_3 \leftarrow L_3 - 2L_2} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 0 & -1 & -2 & -7 & -10 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 4 & 36 & 40 \end{array}\right)$$

$$\xrightarrow{L_4 \leftarrow L_4 + L_3} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 0 & -1 & -2 & -7 & -10 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 0 & 40 & 40 \end{array}\right)$$

donc

$$\begin{cases} x_1 + 2x_2 + 3x_3 + 4x_4 = 10 \\ -x_2 - 2x_3 - 7x_4 = -10 \\ -4x_3 + 4x_4 = 0 \\ 40x_4 = 40 \end{cases} \implies x_4 = 1, \quad x_3 = 1, \quad x_2 = 1, \quad x_1 = 1.$$

2. Factorisation de la matrice $\mathbb{A}$:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{pmatrix} \xrightarrow[L_4 \leftarrow L_4 - 4L_1]{\substack{L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - 3L_1}} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & -1 & -2 & -7 \\ 3 & -2 & -8 & -10 \\ 4 & -7 & -10 & -13 \end{pmatrix} \xrightarrow[L_4 \leftarrow L_4 - 7L_2]{L_3 \leftarrow L_3 - 2L_2} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & -1 & -2 & -7 \\ 3 & 2 & -4 & 4 \\ 4 & 7 & 4 & 36 \end{pmatrix} \xrightarrow{L_4 \leftarrow L_4 + L_3} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & -1 & -2 & -7 \\ 3 & 2 & -4 & 4 \\ 4 & 7 & -1 & 40 \end{pmatrix}$$

donc

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 4 & 7 & -1 & 1 \end{pmatrix} \qquad \mathbb{U} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -1 & -2 & -7 \\ 0 & 0 & -4 & 4 \\ 0 & 0 & 0 & 40 \end{pmatrix}$$

Pour résoudre le premier système linéaire on résout les systèmes triangulaires $\mathbb{L}\mathbf{y} = \mathbf{b}$

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 4 & 7 & -1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \implies y_1 = 1, \quad y_2 = 0, \quad y_3 = 0, \quad y_4 = 0$$

et $\mathbb{U}\mathbf{x} = \mathbf{y}$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -1 & -2 & -7 \\ 0 & 0 & -4 & 4 \\ 0 & 0 & 0 & 40 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \implies x_4 = 0, \quad x_3 = 0, \quad x_2 = 0, \quad x_1 = 1.$$

Pour résoudre le second système linéaire on résout les systèmes triangulaires $\mathbb{L}\mathbf{y} = \mathbf{b}$

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 4 & 7 & -1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 10 \\ 10 \\ 10 \\ 10 \end{pmatrix} \implies y_1 = 10, \quad y_2 = -10, \quad y_3 = 0, \quad y_4 = 40$$

et $\mathbb{U}\mathbf{x} = \mathbf{y}$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -1 & -2 & -7 \\ 0 & 0 & -4 & 4 \\ 0 & 0 & 0 & 40 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 10 \\ -10 \\ 0 \\ 40 \end{pmatrix} \implies x_4 = 1, \quad x_3 = 1, \quad x_2 = 1, \quad x_1 = 1.$$

3. Le déterminant de $\mathbb{A}$ est $u_{11}u_{22}u_{33}u_{44} = 1 \times (-1) \times (-4) \times 40 = 160$.

4. Pour calculer $\mathbb{A}^{-1}$ on résout les quatre systèmes linéaires

$$\begin{pmatrix}1&2&3&4\\2&3&4&1\\3&4&1&2\\4&1&2&3\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}=\begin{pmatrix}1\\0\\0\\0\end{pmatrix}\text{ i.e. }\begin{pmatrix}1&0&0&0\\2&1&0&0\\3&2&1&0\\4&7&-1&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix}=\begin{pmatrix}1\\0\\0\\0\end{pmatrix}\implies\begin{pmatrix}1\\-2\\1\\11\end{pmatrix}\text{ puis }\begin{pmatrix}1&2&3&4\\0&-1&-2&-7\\0&0&-4&4\\0&0&0&40\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}=\begin{pmatrix}1\\-2\\1\\11\end{pmatrix}\implies\begin{pmatrix}-9/40\\1/40\\1/40\\11/40\end{pmatrix}$$

$$\begin{pmatrix}1&2&3&4\\2&3&4&1\\3&4&1&2\\4&1&2&3\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}=\begin{pmatrix}0\\1\\0\\0\end{pmatrix}\text{ i.e. }\begin{pmatrix}1&0&0&0\\2&1&0&0\\3&2&1&0\\4&7&-1&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix}=\begin{pmatrix}0\\1\\0\\0\end{pmatrix}\implies\begin{pmatrix}0\\1\\-2\\-9\end{pmatrix}\text{ puis }\begin{pmatrix}1&2&3&4\\0&-1&-2&-7\\0&0&-4&4\\0&0&0&40\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}=\begin{pmatrix}0\\1\\-2\\-9\end{pmatrix}\implies\begin{pmatrix}1/40\\1/40\\11/40\\-9/40\end{pmatrix}$$

$$\begin{pmatrix}1&2&3&4\\2&3&4&1\\3&4&1&2\\4&1&2&3\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}=\begin{pmatrix}0\\0\\1\\0\end{pmatrix}\text{ i.e. }\begin{pmatrix}1&0&0&0\\2&1&0&0\\3&2&1&0\\4&7&-1&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix}=\begin{pmatrix}0\\0\\1\\0\end{pmatrix}\implies\begin{pmatrix}0\\0\\1\\1\end{pmatrix}\text{ puis }\begin{pmatrix}1&2&3&4\\0&-1&-2&-7\\0&0&-4&4\\0&0&0&40\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}=\begin{pmatrix}0\\0\\1\\1\end{pmatrix}\implies\begin{pmatrix}1/40\\11/40\\-9/40\\1/40\end{pmatrix}$$

$$\begin{pmatrix}1&2&3&4\\2&3&4&1\\3&4&1&2\\4&1&2&3\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}=\begin{pmatrix}0\\0\\0\\1\end{pmatrix}\text{ i.e. }\begin{pmatrix}1&0&0&0\\2&1&0&0\\3&2&1&0\\4&7&-1&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix}=\begin{pmatrix}0\\0\\0\\1\end{pmatrix}\implies\begin{pmatrix}0\\0\\0\\1\end{pmatrix}\text{ puis }\begin{pmatrix}1&2&3&4\\0&-1&-2&-7\\0&0&-4&4\\0&0&0&40\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}=\begin{pmatrix}0\\0\\0\\1\end{pmatrix}\implies\begin{pmatrix}11/40\\-9/40\\1/40\\1/40\end{pmatrix}$$

et finalement

$$\mathbb{A}^{-1}=\begin{pmatrix}-9/40&1/40&1/40&11/40\\1/40&1/40&11/40&-9/40\\1/40&11/40&-9/40&1/40\\11/40&-9/40&1/40&1/40\end{pmatrix}=\frac{1}{40}\begin{pmatrix}-9&1&1&11\\1&1&11&-9\\11&11&-9&1\\11&-9&1&1\end{pmatrix}.$$

Méthodes itératives

Une méthode itérative pour le calcul de la solution d'un système linéaire $\mathbb{A}\mathbf{x}=\mathbf{b}$ avec $\mathbb{A}\in\mathbb{R}^{n\times n}$ est une méthode qui construit une suite de vecteurs $\mathbf{x}^{(k)}=(x_1^{(k)},x_2^{(k)},\dots,x_n^{(k)})^T\in\mathbb{R}^n$ convergent vers le vecteur solution exacte $\mathbf{x}=(x_1,x_2,\dots,x_n)^T$ pour tout vecteur initiale $\mathbf{x}^{(0)}=(x_1^{(0)},x_2^{(0)},\dots,x_n^{(0)})^T\in\mathbb{R}^n$ lorsque k tend vers $+\infty$. Dans ces notes on ne verra que deux méthodes itératives :

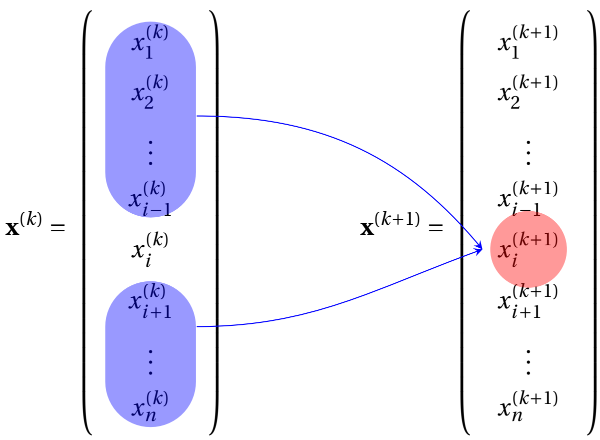
▷ la méthode de Jacobi,

▷ la méthode de Gauss-Seidel.

Méthode de Jacobi

Soit $\mathbf{x}^0=(x_1^0,x_2^0,\dots,x_n^0)$ un vecteur donné. La méthode de Jacobi définit la composante x_i^{k+1} du vecteur $\mathbf{x}^{k+1}$ à partir des composantes x_j^k du vecteur $\mathbf{x}^k$ pour $j\neq i$ de la manière suivante :

$$x_i^{k+1}=\frac{b_i-\sum\limits_{\substack{j=1\\j\neq i}}^{n}a_{ij}x_j^k}{a_{ii}},\qquad i=1,\dots,n$$



Proposition

Si la matrice $\mathbb{A}$ est à diagonale dominante stricte, la méthode de Jacobi converge.

La méthode de GAUSS-SIDEL est une amélioration de la méthode de JACOBI dans laquelle les valeurs calculées sont utilisées au fur et à mesure du calcul et non à l'issue d'une itération comme dans la méthode de JACOBI.

Méthode de GAUSS-SIDEL

Soit $\mathbf{x}^0 = (x_1^0, x_2^0, \dots, x_n^0)$ un vecteur donné. La méthode de GAUSS-SIDEL définit la composante x_i^{k+1} du vecteur $\mathbf{x}^{k+1}$ à partir des composantes x_j^{k+1} du vecteur $\mathbf{x}^{k+1}$ pour $j < i$ et des composantes x_j^k du vecteur $\mathbf{x}^k$ pour $j \ge i$ de la manière suivante :

$$x_i^{k+1} = \frac{b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij} x_j^k}{a_{ii}}, \qquad i = 1, \dots, n$$

$$\mathbf{x}^{(k)} = \begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \\ \vdots \\ x_{i-1}^{(k)} \\ x_i^{(k)} \\ x_{i+1}^{(k)} \\ \vdots \\ x_n^{(k)} \end{pmatrix} \qquad \mathbf{x}^{(k+1)} = \begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ \vdots \\ x_{i-1}^{(k+1)} \\ x_i^{(k+1)} \\ x_{i+1}^{(k+1)} \\ \vdots \\ x_n^{(k+1)} \end{pmatrix}$$

Proposition

Si la matrice $\mathbb{A}$ est à diagonale dominante stricte ou si elle est symétrique et définie positive, la méthode de GAUSS-SEIDEL converge.

Algorithmes

Ces algorithmes tentent de résoudre le système d'équations linéaires $\mathbb{A}\mathbf{x} = \mathbf{b}$ d'inconnue $\mathbf{x}$. La matrice $\mathbb{A}$, de taille $n \times n$, doit être inversible et le second membre $\mathbf{b}$ doit être de longueur n. Les itérations s'arrêtent quand le rapport entre la norme du k-ème residu est inférieure ou égale à TOLL, le nombre d'itérations effectuées est alors renvoyé dans iter. MaxITER est le nombre maximum d'itérations.

JACOBI

Require: $\mathbb{A} = (a_{ij})_{1\le i,j\le n}$, $\mathbf{b} = (b_i)_{1\le i\le n}$, $\mathbf{x}$, MaxITER, TOLL
- iter ← 0
- $r \leftarrow \|\mathbf{b} - \mathbb{A}\mathbf{x}\|$
- **while** (r >TOLL & iter<MaxITER) **do**
  - iter ← iter +1
  - $\mathbf{y} \leftarrow \mathbf{x}$
  - **for** i from 1 to n **do**
    - $s \leftarrow 0$
    - **for** j from 1 to $i-1$ **do**
      - $s \leftarrow s + a_{ij} y_j$
    - **end for**
    - **for** j from $i+1$ to n **do**
      - $s \leftarrow s + a_{ij} y_j$
    - **end for**
    - $x_i \leftarrow (b_i - s)/a_{ii}$
  - **end for**
  - $r \leftarrow \|\mathbf{b} - \mathbb{A}\mathbf{x}\|$
- **end while**

GAUSS-SEIDEL

Require: $\mathbb{A} = (a_{ij})_{1\le i,j\le n}$, $\mathbf{b} = (b_i)_{1\le i\le n}$, $\mathbf{x}$, MaxITER, TOLL
- iter ← 0
- $r \leftarrow \|\mathbf{b} - \mathbb{A}\mathbf{x}\|$
- **while** (r >TOLL & iter<MaxITER) **do**
  - iter ← iter +1
  - $\mathbf{y} \leftarrow \mathbf{x}$
  - **for** i from 1 to n **do**
    - $s \leftarrow 0$
    - **for** j from 1 to $i-1$ **do**
      - $s \leftarrow s + a_{ij} x_j$
    - **end for**
    - **for** j from $i+1$ to n **do**
      - $s \leftarrow s + a_{ij} y_j$
    - **end for**
    - $x_i \leftarrow (b_i - s)/a_{ii}$
  - **end for**
  - $r \leftarrow \|\mathbf{b} - \mathbb{A}\mathbf{x}\|$
- **end while**

Il n'y a pas de résultat général établissant que la méthode de GAUSS-SEIDEL converge toujours plus vite que celle de JACOBI. On peut cependant l'affirmer dans certains cas, comme le montre la proposition suivante

Proposition

Soit $\mathbb{A}$ une matrice tridiagonale de taille $n \times n$ inversible dont les coefficients diagonaux sont tous non nuls. Alors les méthodes de JACOBI et de GAUSS-SEIDEL sont soit toutes les deux convergentes soit toutes les deux divergentes. En cas de convergence, la méthode de GAUSS-SEIDEL est plus rapide que celle de JACOBI.

Exemple

Considérons le système linéaire

$$\begin{pmatrix} 4 & 2 & 1 \\ -1 & 2 & 0 \\ 2 & 1 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 9 \end{pmatrix}$$

mis sous la forme

$$\begin{cases} x = 1 - \frac{y}{2} - \frac{z}{4}, \\ y = 1 + \frac{x}{2}, \\ z = \frac{9}{4} - \frac{x}{2} - \frac{y}{4}. \end{cases}$$

Soit $\mathbf{x}^{(0)} = (0,0,0)$ le vecteur initial.

▷ En calculant les itérées avec la méthode de JACOBI on trouve

$$\mathbf{x}^{(1)} = \begin{pmatrix} 1 - \frac{0}{2} - \frac{0}{4} \\ 1 + \frac{0}{2} \\ \frac{9}{4} - \frac{0}{2} - \frac{0}{4} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 9/4 \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} 1 - \frac{1}{2} - \frac{9/4}{4} \\ 1 + \frac{1}{2} \\ \frac{9}{4} - \frac{1}{2} - \frac{1}{4} \end{pmatrix} = \begin{pmatrix} -1/16 \\ 3/2 \\ 3/2 \end{pmatrix}, \quad \mathbf{x}^{(3)} = \begin{pmatrix} 1 - \frac{3/2}{2} - \frac{3/2}{4} \\ 1 + \frac{-1/16}{2} \\ \frac{9}{4} - \frac{-1/16}{2} - \frac{3/2}{4} \end{pmatrix} = \begin{pmatrix} -1/8 \\ -1/32 \\ 61/32 \end{pmatrix}, \quad \mathbf{x}^{(4)} = \begin{pmatrix} 1 - \frac{-1/32}{2} - \frac{61/32}{4} \\ 1 + \frac{-1/8}{2} \\ \frac{9}{4} - \frac{-1/8}{2} - \frac{-1/32}{4} \end{pmatrix} = \begin{pmatrix} 5/128 \\ 15/16 \\ 265/128 \end{pmatrix}.$$

La suite $\mathbf{x}^{(k)}$ converge vers $(0,1,2)$ la solution du système.

▷ En calculant les itérées avec la méthode de GAUSS-SEIDEL on trouve

$$\mathbf{x}^{(1)} = \begin{pmatrix} 1 - \frac{0}{2} - \frac{0}{4} \\ 1 + \frac{1}{2} \\ \frac{9}{4} - \frac{1}{2} - \frac{3/2}{4} \end{pmatrix} = \begin{pmatrix} 1 \\ 3/2 \\ 11/8 \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} 1 - \frac{3/2}{2} - \frac{11/8}{4} \\ 1 + \frac{-3/32}{2} \\ \frac{9}{4} - \frac{-3/32}{2} - \frac{61/64}{4} \end{pmatrix} = \begin{pmatrix} -3/32 \\ 61/64 \\ 527/256 \end{pmatrix}, \quad \mathbf{x}^{(3)} = \begin{pmatrix} 1 - \frac{-3/32}{2} - \frac{61/64}{4} \\ 1 + \frac{9/1024}{2} \\ \frac{9}{4} - \frac{9/1024}{2} - \frac{2047/2048}{4} \end{pmatrix} = \begin{pmatrix} 9/1024 \\ 2047/2048 \\ 16349/8192 \end{pmatrix},$$

La suite $\mathbf{x}^{(k)}$ converge vers $(0,1,2)$ la solution du système.

Exercices

Exercice 4.1

Soit le système linéaire

$$\begin{pmatrix}6&1&1\\2&4&0\\1&2&6\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}12\\0\\6\end{pmatrix}.$$

1. Approcher la solution avec la méthode de JACOBI avec 3 itérations à partir de $\mathbf{x}^{(0)}=(2,2,2)$.
2. Approcher la solution avec la méthode de GAUSS-SEIDEL avec 3 itérations à partir de $\mathbf{x}^{(0)}=(2,2,2)$.
3. Résoudre les systèmes linéaires par la méthode d'élimination de Gauss.
4. Factoriser la matrice $\mathbb{A}$ (sans utiliser la technique du pivot) et résoudre les systèmes linéaires.

SOLUTION.

1. Méthode de JACOBI :

$$\mathbf{x}^{(0)}=\begin{pmatrix}2\\2\\2\end{pmatrix},\quad \mathbf{x}^{(1)}=\begin{pmatrix}\frac{12-(1\times2+1\times2)}{6}\\ \frac{0-(2\times2+0\times2)}{4}\\ \frac{6-(1\times2+2\times2)}{6}\end{pmatrix}=\begin{pmatrix}{}^{4}/_{3}\\-1\\0\end{pmatrix},\quad \mathbf{x}^{(2)}=\begin{pmatrix}\frac{12-(1\times(-1)+1\times0)}{6}\\ \frac{0-(2\times\frac{4}{3}+0\times0)}{4}\\ \frac{6-(1\times\frac{4}{3}+2\times(-1))}{6}\end{pmatrix}=\begin{pmatrix}{}^{13}/_{6}\\-{}^{2}/_{3}\\{}^{10}/_{9}\end{pmatrix},\quad \mathbf{x}^{(3)}=\begin{pmatrix}\frac{12-(1\times\frac{-2}{3}+1\times\frac{10}{9})}{6}\\ \frac{0-(2\times\frac{13}{6}+0\times\frac{10}{9})}{4}\\ \frac{6-(1\times\frac{13}{6}+2\times\frac{-2}{3})}{6}\end{pmatrix}=\begin{pmatrix}{}^{52}/_{27}\\-{}^{13}/_{12}\\{}^{31}/_{36}\end{pmatrix}$$

ainsi

$$\mathbf{x}\approx\begin{pmatrix}1.926\\-1.083\\0.861\end{pmatrix}.$$

2. Méthode de GAUSS-SEIDEL :

$$\mathbf{x}^{(0)}=\begin{pmatrix}2\\2\\2\end{pmatrix},\quad \mathbf{x}^{(1)}=\begin{pmatrix}\frac{12-(1\times2+1\times2)}{6}\\ \frac{0-(2\times\frac{4}{3}+0\times2)}{4}\\ \frac{6-(1\times\frac{4}{3}+2\times\frac{-2}{3})}{6}\end{pmatrix}=\begin{pmatrix}{}^{4}/_{3}\\-\frac{2}{3}\\1\end{pmatrix},\quad \mathbf{x}^{(2)}=\begin{pmatrix}\frac{12-(1\times\frac{-2}{3}+1\times1)}{6}\\ \frac{0-(2\times\frac{35}{18}+0\times1)}{4}\\ \frac{6-(1\times\frac{35}{18}+2\times\frac{-35}{36})}{6}\end{pmatrix}=\begin{pmatrix}{}^{35}/_{18}\\-{}^{35}/_{36}\\1\end{pmatrix},\quad \mathbf{x}^{(3)}=\begin{pmatrix}\frac{12-(1\times\frac{35}{18}+1\times\frac{-35}{36})}{6}\\ \frac{0-(2\times\frac{431}{216}+0\times1)}{4}\\ \frac{6-(1\times\frac{431}{216}+2\times\frac{-431}{432})}{6}\end{pmatrix}=\begin{pmatrix}{}^{431}/_{216}\\-{}^{431}/_{432}\\1\end{pmatrix}$$

ainsi

$$\mathbf{x}\approx\begin{pmatrix}1.995\\-0.995\\1\end{pmatrix}.$$

3. Méthode d'élimination de Gauss :

$$(\mathbb{A}|\mathbf{b})=\left(\begin{array}{ccc|c}6&1&1&12\\2&4&0&0\\1&2&6&6\end{array}\right)\xrightarrow[L_3\leftarrow L_3-\frac{1}{6}L_1]{L_2\leftarrow L_2-\frac{2}{6}L_1}\left(\begin{array}{ccc|c}6&1&1&12\\0&\frac{11}{3}&-\frac{1}{3}&-4\\0&\frac{11}{6}&\frac{35}{6}&4\end{array}\right)\xrightarrow{L_3\leftarrow L_3-\frac{\frac{11}{6}}{\frac{11}{3}}L_2}\left(\begin{array}{ccc|c}6&1&1&12\\0&\frac{11}{3}&-\frac{1}{3}&-4\\0&0&6&6\end{array}\right)$$

donc

$$\begin{cases}6x_1+x_2+x_3=12,\\ \frac{11}{3}x_2-\frac{1}{3}x_3=-4\\ 6x_3=6\end{cases}\qquad\Longrightarrow\qquad x_3=1,\quad x_2=-1,\quad x_1=2.$$

4. Factorisation de la matrice $\mathbb{A}$:

$$\begin{pmatrix}6&1&1\\2&4&0\\1&2&6\end{pmatrix}\xrightarrow[L_3\leftarrow L_3-\frac{1}{6}L_1]{L_2\leftarrow L_2-\frac{2}{6}L_1}\begin{pmatrix}6&1&1\\\frac{2}{6}&\frac{11}{3}&-\frac{1}{3}\\\frac{1}{6}&\frac{11}{6}&\frac{35}{6}\end{pmatrix}\xrightarrow{L_3\leftarrow L_3-\frac{\frac{11}{6}}{\frac{11}{3}}L_2}\begin{pmatrix}6&1&1\\\frac{2}{6}&\frac{11}{3}&-\frac{1}{3}\\\frac{1}{6}&\frac{\frac{11}{6}}{\frac{11}{3}}&6\end{pmatrix}$$

donc

$$\mathbb{L}=\begin{pmatrix}1&0&0\\\frac{1}{3}&1&0\\\frac{1}{6}&\frac{1}{2}&1\end{pmatrix}\qquad \mathbb{U}=\begin{pmatrix}6&1&1\\0&\frac{11}{3}&-\frac{1}{3}\\0&0&6\end{pmatrix}$$

Pour résoudre le système linéaire on résout les systèmes triangulaires $\mathbb{L}\mathbf{y}=\mathbf{b}$

$$\begin{pmatrix}1&0&0\\\frac{1}{3}&1&0\\\frac{1}{6}&\frac{1}{2}&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix}=\begin{pmatrix}12\\0\\6\end{pmatrix}\qquad\Longrightarrow\qquad y_1=12,\quad y_2=-4,\quad y_3=6$$

et $\mathbb{U}\mathbf{x} = \mathbf{y}$

$$\begin{pmatrix} 6 & 1 & 1 \\ 0 & \frac{11}{3} & -\frac{1}{3} \\ 0 & 0 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ -4 \\ 6 \end{pmatrix} \qquad \Longrightarrow \qquad x_3 = 1, \quad x_2 = -1, \quad x_1 = 2.$$

Exercice 4.2

Soit $\mathbb{A}$ une matrice, $\mathbb{A} \in \mathcal{M}_{n,n}(\mathbb{R})$.

1. Rappeler les conditions nécessaires et suffisantes pour l'existence d'une factorisation $\mathbb{L}\mathbb{U}$ de la matrice $\mathbb{A}$ et préciser les définitions de $\mathbb{L}$ et $\mathbb{U}$.
2. On suppose $\mathbb{L}$ et $\mathbb{U}$ construites (*i.e.* on dispose de tous les coefficients $\ell_{i,j}$ et $u_{i,j}$ de $\mathbb{L}$ et $\mathbb{U}$), écrire l'algorithme de résolution de $\mathbb{A}\mathbf{x} = \mathbf{b}$, avec $\mathbf{b} \in \mathcal{M}_{n,1}(\mathbb{R})$ donné.
3. Soit la matrice $\mathbb{A}$ suivante :
$$\begin{pmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{pmatrix}.$$
Construire à la main les matrices $\mathbb{L}$ et $\mathbb{U}$ de la factorisation $\mathbb{L}\mathbb{U}$.

SOLUTION.

1. Pour une matrice quelconque $\mathbb{A} \in \mathcal{M}_{n,n}(\mathbb{R})$, la factorisation $\mathbb{L}\mathbb{U}$ (sans pivot) existe et est unique ssi les sous-matrices principales $\mathbb{A}_i$ de $\mathbb{A}$ d'ordre $i = 1, \dots, n-1$ (celles que l'on obtient en restreignant $\mathbb{A}$ à ses i premières lignes et colonnes) ne sont pas singulières (autrement dit si les mineurs principaux, *i.e.* les déterminants des sous-matrices principales, sont non nuls). On peut identifier des classes de matrices particulières pour lesquelles les hypothèses de cette proposition sont satisfaites. Mentionnons par exemple : les matrices à diagonale strictement dominante, les matrices réelles symétriques définies positives. Une technique qui permet d'effectuer la factorisation $\mathbb{L}\mathbb{U}$ pour toute matrice $\mathbb{A}$ inversible, même quand les hypothèses de cette proposition ne sont pas vérifiées, est la méthode du pivot par ligne : il suffit d'effectuer une permutation convenable des lignes de la matrice originale $\mathbb{A}$ à chaque étape k où un terme diagonal a_{kk} s'annule.
2. Une fois calculées les matrices $\mathbb{L}$ et $\mathbb{U}$, résoudre le système linéaire $\mathbb{A}\mathbf{x} = \mathbf{b}$, avec $\mathbf{b} \in \mathcal{M}_{n,1}(\mathbb{R})$ donné consiste simplement à résoudre successivement
   2.1. le système triangulaire inférieur $\mathbb{L}\mathbf{y} = \mathbf{b}$ par l'algorithme
$$y_1 = b_1, \qquad y_i = b_i - \sum_{j=1}^{i-1} \ell_{ij} y_j, \qquad i = 2, \dots, n$$
   2.2. le système triangulaire supérieure $\mathbb{U}\mathbf{x} = \mathbf{y}$ par l'algorithme
$$x_n = \frac{y_n}{u_{nn}}, \qquad x_i = \frac{1}{u_{ii}}\left(y_i - \sum_{j=i+1}^{n} u_{ij} x_j\right), \qquad j = n-1, \dots, 1$$
3. Factorisation :
$$\begin{pmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{pmatrix} \xrightarrow[L_3 \leftarrow L_3 - \frac{-1}{3} L_1]{L_2 \leftarrow L_2 - \frac{-1}{3} L_1} \begin{pmatrix} 3 & -1 & -1 \\ 0 & 8/3 & -4/3 \\ 0 & -4/3 & 8/3 \end{pmatrix} \xrightarrow{L_3 \leftarrow L_3 - \frac{-4/3}{8/3} L_2} \begin{pmatrix} 3 & -1 & -1 \\ 0 & 8/3 & -4/3 \\ 0 & 0 & 2 \end{pmatrix}.$$
Par conséquent
$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 \\ -1/3 & 1 & 0 \\ -1/3 & -1/2 & 1 \end{pmatrix} \qquad \text{et} \qquad \mathbb{U} = \begin{pmatrix} 3 & -1 & -1 \\ 0 & 8/3 & -4/3 \\ 0 & 0 & 2 \end{pmatrix}.$$

Exercice 4.3

Calculer, lorsqu'il est possible, la factorisation $\mathbb{L}\mathbb{U}$ des matrices suivantes :

$$\mathbb{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 7 & 8 & 9 \end{pmatrix}, \qquad \mathbb{B} = \begin{pmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \\ 2 & 4 & 5 \end{pmatrix}.$$

Comment peut-on modifier l'algorithme de factorisation pour pouvoir toujours aboutir à une factorisation $\mathbb{L}\mathbb{U}$ lorsque la matrice est inversible ?

Solution. Pour une matrice quelconque $\mathbb{A} \in \mathscr{M}_{n,n}(\mathbb{R})$, la factorisation $\mathbb{L}\mathbb{U}$ (sans pivot) existe et est unique ssi les sous-matrices principales $\mathbb{A}_i$ de $\mathbb{A}$ d'ordre $i = 1, \dots, n-1$ (celles que l'on obtient en restreignant $\mathbb{A}$ à ses i premières lignes et colonnes) ne sont pas singulières (autrement dit si les mineurs principaux, *i.e.* les déterminants des sous-matrices principales, sont non nuls).

Matrice $\mathbb{A}$: comme $\det(\mathbb{A}) \neq 0$, la matrice $\mathbb{A}$ est bien inversible. Puisque $\det(\mathbb{A}_1) = a_{11} = 1 \neq 0$ mais $\det(\mathbb{A}_2) = a_{11}a_{22} - a12a_{21} = 0$, on ne peut pas factoriser $\mathbb{A}$ sans utiliser la technique du pivot. En effet,

$$\mathbb{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 7 & 8 & 9 \end{pmatrix} \xrightarrow[L_3 \leftarrow L_3 - \frac{7}{1}L_1]{L_2 \leftarrow L_2 - \frac{2}{1}L_1} \begin{pmatrix} 1 & 2 & 3 \\ 0 & 0 & -1 \\ 0 & -6 & -12 \end{pmatrix}$$

La factorisation $\mathbb{L}\mathbb{U}$ ne peut pas être calculée car à la prochaine étape il faudrait effectuer le changement $L_3 \leftarrow L_3 - \frac{-6}{0}L_2$.

Matrice $\mathbb{B}$:

$$\mathbb{A}_2 = \begin{pmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \\ 2 & 4 & 5 \end{pmatrix} \xrightarrow[L_3 \leftarrow L_3 - \frac{2}{1}L_1]{L_2 \leftarrow L_2 - \frac{7}{1}L_1} \begin{pmatrix} 1 & 2 & 3 \\ 0 & -6 & -12 \\ 0 & 0 & -1 \end{pmatrix}$$

La factorisation $\mathbb{L}\mathbb{U}$ de la matrice $\mathbb{B}$ est donc

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 \\ 7 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix}, \qquad \mathbb{U} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -6 & -12 \\ 0 & 0 & -1 \end{pmatrix}.$$

Lorsqu'un pivot est nul, la méthode de Gauss pour calculer la factorisation $\mathbb{L}\mathbb{U}$ de la matrice $\mathbb{A}$ n'est plus applicable. De plus, si le pivot n'est pas nul mais très petit, l'algorithme conduit à des erreurs d'arrondi importantes. C'est pourquoi des algorithmes qui échangent les éléments de façon à avoir le pivot le plus grand possible ont été développés. Les programmes optimisés intervertissent les lignes à chaque étape de façon à placer en pivot le terme de coefficient le plus élevé : c'est la méthode du pivot partiel. Pour la matrice $\mathbb{A}$ cela aurait donné

$$\mathbb{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 7 & 8 & 9 \end{pmatrix} \xrightarrow{L_2 \leftrightarrow L_3} \begin{pmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \\ 2 & 4 & 5 \end{pmatrix} \xrightarrow[L_3 \leftarrow L_3 - \frac{2}{1}L_1]{L_2 \leftarrow L_2 - \frac{7}{1}L_1} \begin{pmatrix} 1 & 2 & 3 \\ 0 & -6 & -12 \\ 0 & 0 & -1 \end{pmatrix}.$$

Bien évidemment, il faut garder trace de cet échange de lignes pour qu'il puisse être répercuté sur le terme source et sur l'inconnue lors de la résolution du système linéaire ; ceci est réalisé en introduisant une nouvelle matrice $\mathbb{P}$, dite matrice pivotale, telle que $\mathbb{P}\mathbb{A} = \mathbb{L}\mathbb{U}$: la résolution du système linéaire $\mathbb{A}\mathbf{x} = \mathbf{b}$ est donc ramené à la résolution des deux systèmes triangulaires $\mathbb{L}\mathbf{y} = \mathbb{P}\mathbf{b}$ et $\mathbb{U}\mathbf{x} = \mathbf{y}$. Dans notre exemple cela donne

$$\mathbb{P} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Exercice 4.4

Après avoir utilisé la méthode d'élimination de Gauss, une matrice symétrique $\mathbb{A}$ a été transformée en la matrice triangulaire supérieure

$$\mathbb{U} = \begin{pmatrix} 4 & -2 & 1 & 0 \\ 0 & 3 & -3/2 & 1 \\ 0 & 0 & 3 & -3/2 \\ 0 & 0 & 0 & 35/12 \end{pmatrix}.$$

Calculer la matrice $\mathbb{A}$.

Solution. Comme $\mathbb{A}=\mathbb{L}\mathbb{U}$, il suffit de calculer $\mathbb{L}$. En divisant chaque ligne de $\mathbb{U}$ par son terme diagonal on trouve

$$\mathbb{L}^T=\begin{pmatrix}1&-1/2&1/4&0\\0&1&-1/2&1/3\\0&0&1&-1/2\\0&0&0&1\end{pmatrix}.$$

Alors

$$\mathbb{A}=\mathbb{L}\mathbb{U}=\begin{pmatrix}1&0&0&0\\-1/2&1&0&0\\1/4&-1/2&1&0\\0&1/3&-1/2&1\end{pmatrix}\begin{pmatrix}4&-2&1&0\\0&3&-3/2&1\\0&0&3&-3/2\\0&0&0&35/12\end{pmatrix}=\begin{pmatrix}4&-2&1&0\\-2&4&-2&1\\1&-2&4&-2\\0&1&-2&4\end{pmatrix}.$$

Exercice 4.5

Soit α un paramètre réel et soient les matrices $\mathbb{A}_\alpha$, $\mathbb{P}$ et le vecteur $\mathbf{b}$ définis par

$$\mathbb{A}_\alpha=\begin{pmatrix}2&4&1\\\alpha&-2&-1\\2&3&2\end{pmatrix},\qquad \mathbb{P}=\begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix},\qquad \mathbf{b}=\begin{pmatrix}0\\-3/2\\-1\end{pmatrix}.$$

1. À quelle condition sur α, la matrice $\mathbb{A}_\alpha$ est inversible ?
2. À quelle condition sur α, la matrice $\mathbb{A}_\alpha$ admet-elle une décomposition $\mathbb{L}\mathbb{U}$ (sans pivot) ?
3. Soit $\alpha=-1$. Calculer, si elle existe, la décomposition $\mathbb{L}\mathbb{U}$ de la matrice $\mathbb{M}=\mathbb{P}\mathbb{A}_\alpha$.
4. Soit $\alpha=-1$. Résoudre le système linéaire $\mathbb{A}\mathbf{x}=\mathbf{b}$ en résolvant le système linéaire $\mathbb{M}\mathbf{x}=\mathbb{P}\mathbf{b}$.

Solution.

1. La matrice $\mathbb{A}_\alpha$ est inversible si et seulement si $\det(\mathbb{A})\neq 0$. Comme
$$\begin{aligned}\det(\mathbb{A})=\det\begin{pmatrix}2&4&1\\\alpha&-2&-1\\2&3&2\end{pmatrix}&=(2\times(-2)\times 2)+(4\times(-1)\times 2)+(1\times\alpha\times 3)-(2\times(-1)\times 3)-(4\times\alpha\times 2)-(1\times(-2)\times 2)\\&=(-8)+(-8)+(3\alpha)-(-6)-(8\alpha)-(-4)\\&=-6-5\alpha,\end{aligned}$$
la matrice $\mathbb{A}_\alpha$ est inversible si et seulement si $\alpha\neq-\frac{6}{5}$.
2. Pour une matrice $\mathbb{A}$ carrée d'ordre n quelconque, la factorisation de Gauss existe et est unique si et seulement si les sous-matrices principales $\mathbb{A}_i$ de $\mathbb{A}$ d'ordre $i=1,\dots,n-1$ (celles que l'on obtient en restreignant $\mathbb{A}$ à ses i premières lignes et colonnes) ne sont pas singulières (autrement dit si les mineurs principaux, *i.e.* les déterminants des sous-matrices principales, sont non nuls).
Pour la matrice $\mathbb{A}_\alpha$ on a les sous-matrices principales suivantes :
$$\mathbb{A}_1=\begin{pmatrix}2\end{pmatrix},\qquad \det(\mathbb{A}_1)=2;$$
$$\mathbb{A}_2=\begin{pmatrix}2&4\\\alpha&-2\end{pmatrix},\qquad \det(\mathbb{A}_2)=-4(1+\alpha).$$
Par conséquent, la matrice $\mathbb{A}_\alpha$ admet une décomposition $\mathbb{L}\mathbb{U}$ (sans pivot) si et seulement si $\alpha\neq-1$.
3. Si $\alpha=-1$ la matrice $\mathbb{A}_\alpha$ n'admet pas de décomposition $\mathbb{L}\mathbb{U}$ sans pivot. La matrice $\mathbb{P}$ échange les lignes 2 et 3 de la matrice $\mathbb{A}$ et on obtient la matrice
$$\mathbb{P}\mathbb{A}_{-1}=\begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix}\begin{pmatrix}2&4&1\\-1&-2&-1\\2&3&2\end{pmatrix}=\begin{pmatrix}2&4&1\\2&3&2\\-1&-2&-1\end{pmatrix}.$$
La matrice $\mathbb{M}$ admet une décomposition $\mathbb{L}\mathbb{U}$ (sans pivot) et l'on a
$$\begin{pmatrix}2&4&1\\2&3&2\\-1&-2&-1\end{pmatrix}\xrightarrow[L_3\leftarrow L_3-\frac{-1}{2}L_1]{L_2\leftarrow L_2-L_1}\begin{pmatrix}2&4&1\\0&-1&1\\0&0&-\frac{1}{2}\end{pmatrix}$$

Par conséquent, on obtient la décomposition $\mathbb{L}\mathbb{U}$ suivante de la matrice $\mathbb{M}$:

$$\mathbb{L}=\begin{pmatrix}1&0&0\\1&1&0\\-\frac{1}{2}&0&1\end{pmatrix},\qquad \mathbb{U}=\begin{pmatrix}2&4&1\\0&-1&1\\0&0&-\frac{1}{2}\end{pmatrix}.$$

4. Pour résoudre le système linéaire $\mathbb{M}\mathbf{x}=\mathbb{P}\mathbf{b}$ il suffit de résoudre les deux systèmes triangulaires suivantes :
   ▷ $\mathbb{L}\mathbf{y}=\mathbb{P}\mathbf{b}$:
   $$y_1=0,\qquad y_2=-1-y_1=-1,\qquad y_3=-\frac{3}{2}+\frac{1}{2}y_1=-\frac{3}{2};$$
   ▷ $\mathbb{U}\mathbf{x}=\mathbf{y}$:
   $$x_3=\frac{-3}{2}(-2)=3,\qquad x_2=(-1-x_3)/(-1)=4,\qquad x_1=(0-x_2-4x_3)/2=-\frac{19}{2}.$$

Exercice 4.6

Considérons les deux matrices carrées d'ordre $n>3$:

$$\mathbb{A}=\begin{pmatrix}\alpha&0&0&0&&\dots&\beta\\0&\alpha&0&0&0&\dots&\beta\\0&0&\alpha&0&\ddots&&\vdots\\&0&\ddots&\ddots&&\dots&\beta\\\vdots&\vdots&&\ddots&&0&\beta\\0&0&&&0&\alpha&\beta\\\beta&\beta&\dots&&\beta&\beta&\alpha\end{pmatrix}\qquad \mathbb{B}=\begin{pmatrix}\beta&0&\dots&\dots&0&0&\alpha\\\beta&&0&0&0&\alpha&0\\\vdots&&&0&\ddots&&0\\&&&\ddots&&\dots&\vdots\\\vdots&0&\alpha&0&&0&0\\\beta&\alpha&0&&0&\alpha&0\\\alpha&\beta&\beta&\dots&&\beta&\beta\end{pmatrix}$$

avec α et β réels non nuls.

1. Vérifier que la factorisation $\mathbb{L}\mathbb{U}$ de la matrice $\mathbb{B}$ ne peut pas être calculée sans utiliser la technique du pivot.
2. Calculer analytiquement le nombre d'opérations nécessaires pour calculer la factorisation $\mathbb{L}\mathbb{U}$ de la matrice $\mathbb{A}$.
3. Exprimer le déterminant de la matrice $\mathbb{A}$ sous forme récursive en fonction des coefficients de la matrice et de sa dimension n.
4. Sous quelles conditions sur α et β la matrice $\mathbb{A}$ est définie positive ? Dans ce cas, exprimer le conditionnement de la matrice en fonction des coefficients et de la dimension n.

Solution.

1. La factorisation $\mathbb{L}\mathbb{U}$ de la matrice $\mathbb{B}$ ne peut pas être calculée sans utiliser la technique du pivot car l'élément pivotale au deuxième pas est nul. Par exemple, si $n=4$, on obtient :
   $$\mathbb{B}^{(1)}=\begin{pmatrix}\beta&0&0&\alpha\\\beta&0&\alpha&0\\\beta&\alpha&0&0\\\alpha&\beta&\beta&\beta\end{pmatrix}\xrightarrow[L_4\leftarrow L_4-\frac{\alpha}{\beta}L_1]{\substack{L_2\leftarrow L_2-L_1\\L_3\leftarrow L_3-L_1}}\mathbb{B}^{(2)}=\begin{pmatrix}\beta&0&0&\alpha\\0&\boxed{0}&\alpha&-\alpha\\0&\alpha&0&-\alpha\\0&\beta&\beta&\beta-\frac{\alpha^2}{\beta}\end{pmatrix}.$$
2. La matrice $\mathbb{A}$ est une matrice «en flèche» : pour en calculer la factorisation $\mathbb{L}\mathbb{U}$ il suffit de transformer la dernière ligne, ce qui requiert le calcul de l'unique multiplicateur $m=\beta/\alpha$ et l'exécution de $n-1$ produits et sommes. Le coût globale est donc de l'ordre de n.
3. Le déterminant δ_n de la matrice $\mathbb{A}$ de dimension n coïncide avec le déterminant de la matrice $\mathbb{U}$. Comme $u_{ii}=\alpha$ pour tout $i<n$ et $u_{nn}=\alpha-(n-1)\beta^2/\alpha$, on conclut que
   $$\delta_n=\alpha^n-(n-1)\alpha^{n-2}\beta^2.$$
4. Les valeurs propres de la matrice $\mathbb{A}$ sont les racines du déterminant de la matrice $\mathbb{A}-\lambda\mathbb{I}$. Suivant le même raisonnement du point précédant, ce déterminant s'écrit
   $$(\alpha-\lambda)^n-(n-1)(\alpha-\lambda)^{n-2}\beta^2$$
   dont les racines sont
   $$\lambda_{1,2}=\alpha\pm\sqrt{(n-1)\beta},\qquad \lambda_3=\dots=\lambda_n=\alpha.$$

Par conséquent, pour que la matrice $\mathbb{A}$ soit définie positive il faut que les valeurs propres soient tous positifs, ce qui impose

$$\alpha > 0, \qquad |\beta| < \frac{\alpha}{\sqrt{n-1}}.$$

Dans ce cas, le conditionnement de la matrice en norme 2 est

$$K_2(\mathbb{A}) = \begin{cases} \frac{\alpha+\beta\sqrt{n-1}}{\alpha-\beta\sqrt{n-1}} & \text{si } \beta \geq 0, \\ \frac{\alpha-\beta\sqrt{n-1}}{\alpha+\beta\sqrt{n-1}} & \text{sinon.} \end{cases}$$

Exercice 4.7

Donner une condition suffisante sur le coefficient α pour avoir convergence des méthodes de Jacobi et Gauss-Seidel pour la résolution d'un système linéaire associé à la matrice

$$\mathbb{A} = \begin{pmatrix} \alpha & 0 & 1 \\ 0 & \alpha & 0 \\ 1 & 0 & \alpha \end{pmatrix}$$

Solution. Une condition suffisante pour la convergence des méthodes de Jacobi et de Gauss-Seidel est que $\mathbb{A}$ est à diagonale strictement dominante. $\mathbb{A}$ vérifie cette condition si et seulement si $|\alpha| > 1$.

Exercice 4.8

Considérons le système linéaire $\mathbb{A}\mathbf{x} = \mathbf{b}$ avec

$$\mathbb{A} = \begin{pmatrix} \alpha & 0 & \gamma \\ 0 & \alpha & \beta \\ 0 & \delta & \alpha \end{pmatrix}$$

avec α, β, γ et δ des paramètres réels. Donner des conditions suffisantes sur les coefficients pour avoir

1. convergence de la méthode de Jacobi
2. convergence de la méthode de Gauss-Seidel.

Solution.

1. Une condition suffisante pour que la méthode de Jacobi converge est que la matrice soit à dominance diagonale stricte, ce qui équivaut à imposer

$$\begin{cases} |\alpha| > |\gamma|, \\ |\alpha| > |\beta|, \\ |\alpha| > |\delta|, \end{cases}$$

c'est-à-dire $|\alpha| > \max\{|\beta|, |\gamma|, |\delta|\}$.

2. La condition précédente est aussi suffisante pour la convergence de la méthode de Gauss-Seidel. Une autre condition suffisante pour la convergence de cette méthode est que la matrice soit symétrique définie positive. Pour la symétrie il faut que

$$\begin{cases} \gamma = 0, \\ \beta = \delta, \end{cases}$$

on obtient ainsi la matrice

$$\mathbb{A} = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \alpha & \beta \\ 0 & \beta & \alpha \end{pmatrix}.$$

Elle est définie positive si ses valeurs propres sont positifs. On a

$$\lambda_1 = \alpha, \qquad \lambda_2 = \alpha - \beta, \qquad \lambda_3 = \alpha + \beta,$$

donc il faut que $\alpha > |\beta|$.

Exercice 4.9

Écrire les formules de la méthode d'élimination de Gauss pour une matrice de la forme

$$\mathbb{A} = \begin{pmatrix} a_{1,1} & a_{1,2} & 0 & \dots & & 0 \\ a_{2,1} & a_{2,2} & a_{2,3} & 0 & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ \vdots & & & & a_{n-1,n-1} & a_{n-1,n} \\ a_{n,1} & a_{n,2} & \dots & & a_{n,n-1} & a_{n,n} \end{pmatrix}.$$

Quelle est la forme finale de la matrice $\mathbb{U} = \mathbb{A}^{(n)}$? Étant donné la forme particulière de la matrice $\mathbb{A}$, indiquer le nombre minimal d'opérations nécessaire pour calculer $\mathbb{U}$ ainsi que celui pour la résolution des systèmes triangulaires finaux.

Solution. Comme la matrice a une seule sur-diagonale non nulle, les formules de la méthode d'élimination de Gauss deviennent

$$\begin{aligned} a_{i,j}^{(k+1)} &= a_{i,j}^{(k)} + m_{i,k} a_{k,j}^{(k)}, & i,j &= k+1, \\ m_{i,k} &= \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}, & i &= k+1. \end{aligned}$$

La coût est donc de l'ordre de n et la matrice $\mathbb{U}$ est bidiagonale supérieure.

Exercice 4.10

Soit $\alpha \in \mathbb{R}^*$ et considérons les matrices carrées de dimension n

$$\mathbb{A} = \begin{pmatrix} \alpha & 0 & \dots & -\alpha \\ 0 & \ddots & & \vdots \\ \vdots & & \alpha & -\alpha \\ -\alpha & \dots & -\alpha & -\alpha \end{pmatrix}, \qquad \mathbb{B} = \begin{pmatrix} \frac{\beta}{\alpha} & -\frac{\gamma}{\alpha} & \dots & -\frac{\gamma}{\alpha} \\ -\frac{\gamma}{\alpha} & \ddots & & \vdots \\ \vdots & & \frac{\beta}{\alpha} & -\frac{\gamma}{\alpha} \\ -\frac{\gamma}{\alpha} & \dots & -\frac{\gamma}{\alpha} & \frac{\gamma}{\alpha} \end{pmatrix}.$$

1. Calculer γ et β pour que $\mathbb{B}$ soit l'inverse de $\mathbb{A}$.
2. Calculer le conditionnement $K_\infty(\mathbb{A})$ en fonction de n et en calculer la limite pour n qui tend vers l'infini.

Solution.

1. Par définition, $\mathbb{B}$ est la matrice inverse de $\mathbb{A}$ si $\mathbb{AB} = \mathbb{BA} = \mathbb{I}$. Comme

$$\mathbb{AB} = \begin{pmatrix} \beta+\gamma & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \beta+\gamma & 0 \\ -\beta+(n-3)\gamma & \dots & -\beta+(n-3)\gamma & (n-2)\gamma \end{pmatrix},$$

il faut que

$$\begin{cases} \beta+\gamma = 1 \\ -\beta+(n-3)\gamma = 0 \\ (n-2)\gamma = 1 \end{cases}$$

ce qui donne

$$\beta = \frac{n-3}{n-2}, \qquad \gamma = \frac{1}{n-2}.$$

2. On trouve immédiatement $\|\mathbb{A}\|_\infty = n|\alpha|$ tandis que
$$\|\mathbb{A}^{-1}\|_\infty = \frac{1}{|\alpha|}\max\left\{n, \frac{n}{n-2}\right\} = \frac{2}{|\alpha|}.$$
On conclut que le conditionnement $K_\infty(\mathbb{A})$ en fonction de n est
$$K_\infty(\mathbb{A}) = n|\alpha|\frac{2}{|\alpha|} = 2n.$$
La matrice est donc mal conditionnée pour n grand.

Exercice 4.11

On suppose que le nombre réel $\varepsilon > 0$ est assez petit pour que l'ordinateur arrondisse $1+\varepsilon$ en 1 et $1+(1/\varepsilon)$ en $1/\varepsilon$ (ε est plus petit que l'erreur machine (relative), par exemple, $\varepsilon = 2^{-30}$ en format 32 *bits*). Simuler la résolution par l'ordinateur des deux systèmes suivants :
$$\begin{cases}\varepsilon a + b = 1\\ 2a+b=0\end{cases} \qquad \text{et} \qquad \begin{cases}2a+b=0\\ \varepsilon a + b = 1\end{cases}$$
On appliquera pour cela la méthode du pivot de Gauss et on donnera les décompositions $\mathbb{L}\mathbb{U}$ des deux matrices associées à ces systèmes. On fournira également la solution exacte de ces systèmes. Commenter.

Solution.

Premier système :
$$\begin{pmatrix}\varepsilon & 1\\ 2 & 1\end{pmatrix}\begin{pmatrix}a\\ b\end{pmatrix} = \begin{pmatrix}1\\ 0\end{pmatrix}.$$
Factorisation $\mathbb{L}\mathbb{U}$:
$$\begin{pmatrix}\varepsilon & 1\\ 2 & 1\end{pmatrix} \xrightarrow{L_2 \leftarrow L_2 - \frac{2}{\varepsilon}L_1} \begin{pmatrix}\varepsilon & 1\\ 0 & 1-\frac{2}{\varepsilon}\end{pmatrix}$$
donc
$$\mathbb{L} = \begin{pmatrix}1 & 0\\ \frac{2}{\varepsilon} & 1\end{pmatrix} \qquad \mathbb{U} = \begin{pmatrix}\varepsilon & 1\\ 0 & 1-\frac{2}{\varepsilon}\end{pmatrix}$$
Pour résoudre le système linéaire on résout les systèmes triangulaires $\mathbb{L}\mathbf{y} = \mathbf{b}$ et $\mathbb{U}\mathbf{x} = \mathbf{y}$:
$$\begin{pmatrix}1 & 0\\ \frac{2}{\varepsilon} & 1\end{pmatrix}\begin{pmatrix}y_1\\ y_2\end{pmatrix} = \begin{pmatrix}1\\ 0\end{pmatrix} \implies y_1 = 1, \quad y_2 = -\frac{2}{\varepsilon};$$
$$\begin{pmatrix}\varepsilon & 1\\ 0 & 1-\frac{2}{\varepsilon}\end{pmatrix}\begin{pmatrix}a\\ b\end{pmatrix} = \begin{pmatrix}1\\ -\frac{2}{\varepsilon}\end{pmatrix} \implies b = -\frac{2}{\varepsilon\left(1-\frac{2}{\varepsilon}\right)}, \quad a = \frac{1+\frac{2}{\varepsilon(1-\frac{2}{\varepsilon})}}{\varepsilon}.$$
Mais avec l'ordinateur, comme $1+\varepsilon \approx 1$ et $1+(1/\varepsilon) \approx 1/\varepsilon$, on obtient
$$\widetilde{\mathbb{L}} = \begin{pmatrix}1 & 0\\ \frac{2}{\varepsilon} & 1\end{pmatrix} \qquad \widetilde{\mathbb{U}} = \begin{pmatrix}\varepsilon & 1\\ 0 & -\frac{2}{\varepsilon}\end{pmatrix}$$
Pour résoudre ce système linéaire approché on résout les systèmes triangulaires $\widetilde{\mathbb{L}}\mathbf{y} = \mathbf{b}$ et $\widetilde{\mathbb{U}}\mathbf{x} = \mathbf{y}$:
$$\begin{pmatrix}1 & 0\\ \frac{2}{\varepsilon} & 1\end{pmatrix}\begin{pmatrix}y_1\\ y_2\end{pmatrix} = \begin{pmatrix}1\\ 0\end{pmatrix} \implies y_1 = 1, \quad y_2 = -\frac{2}{\varepsilon};$$
$$\begin{pmatrix}\varepsilon & 1\\ 0 & -\frac{2}{\varepsilon}\end{pmatrix}\begin{pmatrix}a\\ b\end{pmatrix} = \begin{pmatrix}1\\ -\frac{2}{\varepsilon}\end{pmatrix} \implies b = 1, \quad a = 0.$$

Second système :
$$\begin{pmatrix}2 & 1\\ \varepsilon & 1\end{pmatrix}\begin{pmatrix}a\\ b\end{pmatrix} = \begin{pmatrix}0\\ 1\end{pmatrix}.$$
Factorisation $\mathbb{L}\mathbb{U}$:
$$\begin{pmatrix}2 & 1\\ \varepsilon & 1\end{pmatrix} \xrightarrow{L_2 \leftarrow L_2 - \frac{\varepsilon}{2}L_1} \begin{pmatrix}2 & 1\\ 0 & 1-\frac{\varepsilon}{2}\end{pmatrix}$$

donc

$$\mathbb{L}=\begin{pmatrix}1&0\\\frac{\varepsilon}{2}&1\end{pmatrix}\qquad\qquad \mathbb{U}=\begin{pmatrix}2&1\\0&1-\frac{2}{\varepsilon}\end{pmatrix}$$

Pour résoudre le système linéaire on résout les systèmes triangulaires $\mathbb{L}\mathbf{y}=\mathbf{b}$ et $\mathbb{U}\mathbf{x}=\mathbf{y}$:

$$\begin{pmatrix}1&0\\\frac{\varepsilon}{2}&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\end{pmatrix}=\begin{pmatrix}0\\1\end{pmatrix}\qquad\Longrightarrow\qquad y_1=0,\quad y_2=1;$$

$$\begin{pmatrix}2&1\\0&1-\frac{2}{\varepsilon}\end{pmatrix}\begin{pmatrix}a\\b\end{pmatrix}=\begin{pmatrix}0\\1\end{pmatrix}\qquad\Longrightarrow\qquad b=-\frac{2}{\varepsilon\left(1-\frac{2}{\varepsilon}\right)},\quad a=\frac{1+\frac{2}{\varepsilon\left(1-\frac{2}{\varepsilon}\right)}}{\varepsilon}.$$

Mais avec l'ordinateur, comme $1+\varepsilon\approx 1$ et $1+(1/\varepsilon)\approx 1/\varepsilon$, on obtient

$$\widetilde{\mathbb{L}}=\begin{pmatrix}1&0\\\frac{\varepsilon}{2}&1\end{pmatrix}\qquad\qquad \widetilde{\mathbb{U}}=\begin{pmatrix}2&1\\0&-\frac{2}{\varepsilon}\end{pmatrix}$$

Pour résoudre ce système linéaire approché on résout les systèmes triangulaires $\widetilde{\mathbb{L}}\mathbf{y}=\mathbf{b}$ et $\widetilde{\mathbb{U}}\mathbf{x}=\mathbf{y}$:

$$\begin{pmatrix}1&0\\\frac{\varepsilon}{2}&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\end{pmatrix}=\begin{pmatrix}0\\1\end{pmatrix}\qquad\Longrightarrow\qquad y_1=0,\quad y_2=1;$$

$$\begin{pmatrix}2&1\\0&-\frac{2}{\varepsilon}\end{pmatrix}\begin{pmatrix}a\\b\end{pmatrix}=\begin{pmatrix}0\\1\end{pmatrix}\qquad\Longrightarrow\qquad b=-\frac{\varepsilon}{2},\quad a=\frac{\varepsilon}{4}.$$

Exercice 4.12

Rappeler l'algorithme vu en cours pour calculer la décomposition $\mathbb{L}\mathbb{U}$ d'une matrice $\mathbb{A}$ et la solution du système $\mathbb{A}\mathbf{x}=\mathbf{b}$ où le vecteur colonne $\mathbf{b}$ est donné. On appliquera ces algorithmes pour les cas suivants :

$$\begin{pmatrix}1&1&1\\2&1&3\\-3&2&4\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}1\\1\\1\end{pmatrix}\quad\text{et}\quad\begin{pmatrix}1&2&3&4\\2&-5&7&1\\3&1&1&5\\2&2&0&3\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}=\begin{pmatrix}1\\1\\1\\1\end{pmatrix}\quad\text{et}\quad\begin{pmatrix}1&1&1&1\\1&-2&3&4\\1&4&6&8\\1&0&0&0\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}=\begin{pmatrix}1\\1\\1\\1\end{pmatrix}$$

Donner, en fonction de n (nombre de lignes et de colonnes de $\mathbb{A}$), une majoration du nombres d'opérations effectuées par l'ordinateur pour calculer la décomposition $\mathbb{L}\mathbb{U}$ de $\mathbb{A}$ avec l'algorithme donné en cours. Donner aussi une estimation du nombres d'opérations effectuées pour résoudre le système $\mathbb{A}\mathbf{x}=\mathbf{b}$ quand la décomposition $\mathbb{L}\mathbb{U}$ est connue.

SOLUTION. Premier système :

$$\left(\begin{array}{ccc|c}1&1&1&1\\2&1&3&1\\-3&2&4&1\end{array}\right)\xrightarrow[L_3\leftarrow L_3-\frac{-3}{1}L_1]{L_2\leftarrow L_2-\frac{2}{1}L_1}\left(\begin{array}{ccc|c}1&1&1&1\\0&-1&1&-1\\0&5&7&4\end{array}\right)\xrightarrow{L_3\leftarrow L_3-\frac{5}{-1}L_2}\left(\begin{array}{ccc|c}1&1&1&1\\0&-1&1&-1\\0&0&12&-1\end{array}\right)$$

donc

$$\mathbb{L}=\begin{pmatrix}1&0&0\\2&1&0\\-3&-5&1\end{pmatrix}\qquad\mathbb{U}=\begin{pmatrix}1&1&1\\0&-1&1\\0&0&12\end{pmatrix}$$

Il ne reste à résoudre que le système triangulaire

$$\begin{cases}x_1+x_2+x_3=1\\-x_2+x_3=-1\\12x_3=-1\end{cases}\Longrightarrow x_3=-\frac{1}{12},\quad x_2=\frac{11}{12},\quad x_1=\frac{1}{6}.$$

Deuxième système :

$$\left(\begin{array}{cccc|c}1&2&3&4&1\\2&-5&7&1&1\\3&1&1&5&1\\2&2&0&3&1\end{array}\right)\xrightarrow{\substack{L_2\leftarrow L_2-\frac{2}{1}L_1\\L_3\leftarrow L_3-\frac{3}{1}L_1\\L_4\leftarrow L_4-\frac{2}{1}L_1}}\left(\begin{array}{cccc|c}1&2&3&4&1\\0&-9&1&-7&-1\\0&-5&-8&-7&-2\\0&-2&-6&-5&-1\end{array}\right)$$

$$\xrightarrow[L_4 \leftarrow L_4 - \frac{-2}{-9} L_2]{L_3 \leftarrow L_3 - \frac{-5}{-9} L_2} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -9 & 1 & -7 & -1 \\ 0 & 0 & -\frac{77}{9} & -\frac{28}{9} & -\frac{13}{9} \\ 0 & 0 & -\frac{56}{9} & -\frac{31}{9} & -\frac{7}{9} \end{array}\right) \xrightarrow{L_4 \leftarrow L_4 - \frac{56/9}{77/9} L_2} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -9 & 1 & -7 & -1 \\ 0 & 0 & -\frac{77}{9} & -\frac{28}{9} & -\frac{13}{9} \\ 0 & 0 & 0 & -\frac{13}{11} & \frac{3}{11} \end{array}\right)$$

donc

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & \frac{5}{9} & 1 & 0 \\ 2 & \frac{2}{9} & \frac{56}{77} & 1 \end{pmatrix} \qquad \mathbb{U} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -9 & 1 & -7 \\ 0 & 0 & -\frac{77}{9} & -\frac{28}{9} \\ 0 & 0 & 0 & -\frac{13}{11} \end{pmatrix}$$

Il ne reste à résoudre que le système triangulaire

$$\begin{cases} x_1 + 2x_2 + 3x_3 + 4x_4 = 1 \\ -9x_2 + x_3 - 7x_4 = -1 \\ -\frac{77}{9} x_3 - \frac{28}{9} x_4 = -\frac{13}{9} \\ -\frac{13}{11} x_4 = \frac{3}{11} \end{cases} \implies x_4 = -\frac{3}{13}, \quad x_3 = \frac{23}{91}, \quad x_2 = \frac{29}{91}, \quad x_1 = \frac{48}{91}.$$

<u>Troisième système :</u>

$$\left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 1 \\ 1 & -2 & 3 & 4 & 1 \\ 1 & 4 & 6 & 8 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{array}\right) \xrightarrow[\substack{L_3 \leftarrow L_3 - L_1 \\ L_4 \leftarrow L_4 - L_1}]{L_2 \leftarrow L_2 - L_1} \left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 1 \\ 0 & -3 & 2 & 3 & -0 \\ 0 & 3 & 5 & 7 & 0 \\ 0 & -1 & -1 & -1 & 0 \end{array}\right)$$

$$\xrightarrow[L_4 \leftarrow L_4 - \frac{-1}{-3} L_2]{L_3 \leftarrow L_3 - (-1) L_2} \left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 1 \\ 0 & -3 & 2 & 3 & -0 \\ 0 & 0 & 7 & 10 & 0 \\ 0 & 0 & -\frac{5}{3} & -2 & 0 \end{array}\right) \xrightarrow{L_4 \leftarrow L_4 - \frac{-5/3}{7} L_2} \left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 1 \\ 0 & -3 & 2 & 3 & -0 \\ 0 & 0 & 7 & 10 & 0 \\ 0 & 0 & 0 & \frac{8}{21} & 0 \end{array}\right)$$

donc

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & \frac{1}{3} & \frac{-5}{21} & 1 \end{pmatrix} \qquad \mathbb{U} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & -3 & 2 & 3 \\ 0 & 0 & 7 & 10 \\ 0 & 0 & 0 & \frac{8}{21} \end{pmatrix}$$

Il ne reste à résoudre que le système triangulaire

$$\begin{cases} x_1 + x_2 + x_3 + x_4 = 1 \\ -3x_2 + 2x_3 + 3x_4 = 0 \\ 7x_3 + 10x_4 = 0 \\ \frac{8}{21} x_4 = 0 \end{cases} \implies x_4 = 0, \quad x_3 = 0, \quad x_2 = 0, \quad x_1 = 1.$$

Exercice 4.13

Écrire les méthodes itératives de Gauss, Jacobi et Gauss-Seidel pour les systèmes suivants :

$$\begin{cases} 10a + b = 11 \\ 2a + 10b = 12 \end{cases} \quad \text{et} \quad \begin{cases} 2a + 10b = 12 \\ 10a + b = 11. \end{cases}$$

Pour chacun de ces méthodes et systèmes, on calculera le rayon spectral de la matrice associée à la méthode. On illustrera les résultats théoriques de convergence/non-convergence en calculant les 3 premiers itérés en prenant comme point de départ le vecteur $(a, b) = (0, 0)$.

Solution.

Gauss ▷ Premier système :

$$\left(\begin{array}{cc|c} 10 & 1 & 11 \\ 2 & 10 & 12 \end{array}\right) \xrightarrow{L_2 \leftarrow L_2 - \frac{2}{10} L_1} \left(\begin{array}{cc|c} 10 & 1 & 11 \\ 0 & \frac{49}{5} & \frac{49}{5} \end{array}\right) \implies \begin{cases} 10a + b = 11 \\ \frac{49}{5} b = \frac{49}{5} \end{cases} \implies \begin{cases} a = 1 \\ b = 1. \end{cases}$$

▷ Second système :

$$\left(\begin{array}{cc|c} 2 & 10 & 12 \\ 10 & 1 & 11 \end{array}\right) \xrightarrow{L_2\leftarrow L_2-\frac{10}{2}L_1} \left(\begin{array}{cc|c} 2 & 10 & 12 \\ 0 & -49 & -49 \end{array}\right) \implies \begin{cases} 2a+10b=12 \\ -49b=-49 \end{cases} \implies \begin{cases} a=1 \\ b=1. \end{cases}$$

Jacobi ▷ Premier système :

$$\begin{cases} 10a+b=11 \\ 2a+10b=12 \end{cases} \iff \begin{cases} a=\frac{11-b}{10} \\ b=\frac{12-2a}{10} \end{cases}$$

La matrice étant à diagonale dominante stricte, la méthode converge et on a

$$\mathbf{x}^{(0)}=\begin{pmatrix}0\\0\end{pmatrix},\quad \mathbf{x}^{(1)}=\begin{pmatrix}\frac{11-0}{10}\\ \frac{12-0}{10}\end{pmatrix}=\begin{pmatrix}11/10\\ 12/10\end{pmatrix},\quad \mathbf{x}^{(2)}=\begin{pmatrix}\frac{11-\frac{12}{10}}{10}\\ \frac{12-2\frac{11}{10}}{10}\end{pmatrix}=\begin{pmatrix}49/50\\ 49/50\end{pmatrix},\quad \mathbf{x}^{(3)}=\begin{pmatrix}\frac{11-\frac{49}{50}}{10}\\ \frac{12-2\frac{49}{50}}{10}\end{pmatrix}=\begin{pmatrix}501/500\\ 502/500\end{pmatrix}.$$

▷ Second système :

$$\begin{cases} 2a+10b=12 \\ 10a+b=11 \end{cases} \iff \begin{cases} a=\frac{12-10b}{2} \\ b=11-10a \end{cases}$$

La méthode ne converge pas, en effet on a

$$\mathbf{x}^{(0)}=\begin{pmatrix}0\\0\end{pmatrix},\quad \mathbf{x}^{(1)}=\begin{pmatrix}\frac{12-0}{2}\\ 11-0\end{pmatrix}=\begin{pmatrix}6\\11\end{pmatrix},\quad \mathbf{x}^{(2)}=\begin{pmatrix}\frac{12-10\times 11}{2}\\ 11-10\times 6\end{pmatrix}=\begin{pmatrix}-49\\-49\end{pmatrix},\quad \mathbf{x}^{(3)}=\begin{pmatrix}\frac{12-10\times(-49)}{2}\\ 11-10\times(-49)\end{pmatrix}=\begin{pmatrix}251\\501\end{pmatrix}.$$

Gauss-Seidel ▷ Premier système :

$$\begin{cases} 10a+b=11 \\ 2a+10b=12 \end{cases} \iff \begin{cases} a=\frac{11-b}{10} \\ b=\frac{12-2a}{10} \end{cases}$$

La matrice étant à diagonale dominante stricte, la méthode converge et on a

$$\mathbf{x}^{(0)}=\begin{pmatrix}0\\0\end{pmatrix},\quad \mathbf{x}^{(1)}=\begin{pmatrix}\frac{11-0}{10}\\ \frac{12-2\frac{11}{10}}{10}\end{pmatrix}=\begin{pmatrix}11/10\\ 49/50\end{pmatrix},\quad \mathbf{x}^{(2)}=\begin{pmatrix}\frac{11-\frac{49}{50}}{10}\\ \frac{12-2\frac{501}{500}}{10}\end{pmatrix}=\begin{pmatrix}501/500\\ 2499/2500\end{pmatrix},\quad \mathbf{x}^{(3)}=\begin{pmatrix}\frac{11-\frac{2499}{2500}}{10}\\ \frac{12-2\frac{25001}{25000}}{10}\end{pmatrix}=\begin{pmatrix}25001/25000\\ 12499/125000\end{pmatrix}.$$

▷ Second système :

$$\begin{cases} 2a+10b=12 \\ 10a+b=11 \end{cases} \iff \begin{cases} a=\frac{12-10b}{2} \\ b=11-10a \end{cases}$$

La méthode ne converge pas, en effet on a

$$\mathbf{x}^{(0)}=\begin{pmatrix}0\\0\end{pmatrix},\quad \mathbf{x}^{(1)}=\begin{pmatrix}\frac{12-0}{2}\\ 11-10\times 6\end{pmatrix}=\begin{pmatrix}6\\-49\end{pmatrix},\quad \mathbf{x}^{(2)}=\begin{pmatrix}\frac{12-10\times(-49)}{2}\\ 11-10\times 251\end{pmatrix}=\begin{pmatrix}251\\-2499\end{pmatrix},\quad \mathbf{x}^{(3)}=\begin{pmatrix}\frac{12-10\times(-2499)}{2}\\ 11-10\times(12501)\end{pmatrix}=\begin{pmatrix}12501\\-124999\end{pmatrix}.$$

Exercice 4.14

Résoudre les systèmes linéaires suivants :

$$\begin{cases} x-5y-7z=3 \\ 2x-13y-18z=3 \\ 3x-27y-36z=3 \end{cases} \quad\text{et}\quad \begin{cases} x-5y-7z=6 \\ 2x-13y-18z=0 \\ 3x-27y-36z=-3 \end{cases} \quad\text{et}\quad \begin{cases} x-5y-7z=0 \\ 2x-13y-18z=3 \\ 3x-27y-36z=6. \end{cases}$$

Solution. Le trois systèmes s'écrivent sous forme matricielle

$$\begin{pmatrix}1&-5&-7\\2&-13&-18\\3&-27&-36\end{pmatrix}\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}3\\3\\3\end{pmatrix} \quad\text{et}\quad \begin{pmatrix}1&-5&-7\\2&-13&-18\\3&-27&-36\end{pmatrix}\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}6\\0\\-3\end{pmatrix} \quad\text{et}\quad \begin{pmatrix}1&-5&-7\\2&-13&-18\\3&-27&-36\end{pmatrix}\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}0\\3\\6\end{pmatrix}$$

On remarque que seul le terme source change. On calcul d'abord la décomposition $\mathbb{LU}$ de la matrice $\mathbb{A}$:

$$\begin{pmatrix}1&-5&-7\\2&-13&-18\\3&-27&-36\end{pmatrix} \xrightarrow[L_3\leftarrow L_3-3L_1]{L_2\leftarrow L_2-2L_1} \begin{pmatrix}1&-5&-7\\0&-3&-4\\0&-12&-15\end{pmatrix} \xrightarrow{L_3\leftarrow L_3-4L_2} \begin{pmatrix}1&-5&-7\\0&-3&-4\\0&0&1\end{pmatrix}$$

donc

$$\mathbb{L}=\begin{pmatrix}1&0&0\\2&1&0\\3&4&1\end{pmatrix}\qquad\qquad \mathbb{U}=\begin{pmatrix}1&-5&-7\\0&-3&-4\\0&0&1\end{pmatrix}$$

Pour résoudre chaque système linéaire on résout les systèmes triangulaires $\mathbb{L}\mathbf{y}=\mathbf{b}$ et $\mathbb{U}\mathbf{x}=\mathbf{y}$.

1. Pour le premier système on a

$$\begin{pmatrix}1&0&0\\2&1&0\\3&4&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix}=\begin{pmatrix}3\\3\\3\end{pmatrix}\qquad\Longrightarrow\qquad y_1=3,\quad y_2=-3,\quad y_3=6;$$

$$\begin{pmatrix}1&-5&-7\\0&-3&-4\\0&0&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}3\\-3\\6\end{pmatrix}\qquad\Longrightarrow\qquad x_3=6,\quad x_2=-7,\quad x_1=10.$$

2. Pour le seconde système on a

$$\begin{pmatrix}1&0&0\\2&1&0\\3&4&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix}=\begin{pmatrix}6\\0\\-3\end{pmatrix}\qquad\Longrightarrow\qquad y_1=6,\quad y_2=-12,\quad y_3=27;$$

$$\begin{pmatrix}1&-5&-7\\0&-3&-4\\0&0&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}6\\-12\\27\end{pmatrix}\qquad\Longrightarrow\qquad x_3=27,\quad x_2=-32,\quad x_1=35.$$

3. Pour le dernier système on a

$$\begin{pmatrix}1&0&0\\2&1&0\\3&4&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix}=\begin{pmatrix}0\\3\\6\end{pmatrix}\qquad\Longrightarrow\qquad y_1=0,\quad y_2=3,\quad y_3=-6;$$

$$\begin{pmatrix}1&-5&-7\\0&-3&-4\\0&0&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}6\\-12\\27\end{pmatrix}\qquad\Longrightarrow\qquad x_3=-6,\quad x_2=7,\quad x_1=-7.$$

Exercice 4.15

Soit $\mathbb{A}$ une matrice, $\mathbb{A}\in\mathcal{M}_{n,n}(\mathbb{R})$.

1. Rappeler la méthode de JACOBI pour la résolution du système $\mathbb{A}\mathbf{x}=\mathbf{b}$, avec $\mathbf{b}\in\mathcal{M}_{n,1}(\mathbb{R})$ donné.
2. Soit la matrice $\mathbb{A}$ suivante :
$$\begin{pmatrix}4&-1&-1\\-1&3&-1\\-1&-1&4\end{pmatrix}.$$
La méthode de JACOBI est-elle convergente pour cette matrice ?
3. Construire à la main les matrices $\mathbb{L}$ et $\mathbb{U}$ de la factorisation $\mathbb{L}\mathbb{U}$ pour la matrice ci-dessus.

SOLUTION.

1. La méthode de JACOBI est une méthode itérative pour le calcul de la solution d'un système linéaire qui construit une suite de vecteurs $\mathbf{x}^{(k)}\in\mathbb{R}^n$ convergent vers la solution exacte $\mathbf{x}$ pour tout vecteur initiale $\mathbf{x}^{(0)}\in\mathbb{R}^n$:
$$x_i^{k+1}=\frac{b_i-\sum\limits_{\substack{j=1\\j\neq i}}^{n}a_{ij}x_j^k}{a_{ii}},\qquad i=1,\dots,n.$$
2. Comme $|4|>|-1|+|-1|$, $|3|>|-1|+|-1|$ et $|4|>|-1|+|-1|$, la matrice $\mathbb{A}$ est à diagonale dominante stricte donc la méthode de JACOBI converge
3. Factorisation :
$$\begin{pmatrix}4&-1&-1\\-1&3&-1\\-1&-1&4\end{pmatrix}\xrightarrow[L_3\leftarrow L_3-\frac{-1}{4}L_1]{L_2\leftarrow L_2-\frac{-1}{4}L_1}\begin{pmatrix}4&-1&-1\\0&11/4&-5/4\\0&-5/4&15/4\end{pmatrix}\xrightarrow{L_3\leftarrow L_3-\frac{-5/4}{11/4}L_2}\begin{pmatrix}3&-1&-1\\0&11/4&-5/4\\0&0&35/11\end{pmatrix}.$$

Par conséquent

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 \\ -1/4 & 1 & 0 \\ -1/4 & -5/11 & 1 \end{pmatrix} \qquad \text{et} \qquad \mathbb{U} = \begin{pmatrix} 4 & -1 & -1 \\ 0 & 11/4 & -5/4 \\ 0 & 0 & 35/11 \end{pmatrix}.$$

Exercice 4.16

Soit la matrice $\mathbb{A} \in \mathbb{R}^{n\times n}$ dont les éléments vérifient

- ▷ $a_{ij} = 1$ si $i = j$ ou $i = n$,
- ▷ $a_{ij} = -1$ si $i < j$,
- ▷ $a_{ij} = 0$ sinon.

Montrer que $\mathbb{A}$ admet une factorisation $\mathbb{LU}$ avec

- ▷ $\ell_{ii} = 1$ pour $i = 1, \dots, n$,
- ▷ $\ell_{ij} = 0$ si $i < n$ et $i \neq j$,
- ▷ $\ell_{nj} = 2^{j-1}$ si $j < n$;
- ▷ $u_{ij} = a_{ij}$ pour i=1,...,n-1, j=1,...,n,
- ▷ $u_{nj} = 0$ si $j < n$,
- ▷ $u_{nn} = 2^{n-1}$.

Solution. Factorisation $\mathbb{LU}$ de la matrice $\mathbb{A}$:

$$\begin{pmatrix} 1 & -1 & \dots & \dots & \dots & -1 \\ 0 & 1 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & 1 & -1 \\ 1 & 1 & 1 & \dots & 1 & 1 \end{pmatrix} \xrightarrow{L_n \leftarrow L_n - \frac{1}{1}L_1} \begin{pmatrix} 1 & -1 & \dots & \dots & \dots & -1 \\ 0 & 1 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & 1 & -1 \\ 0 & 2 & 2 & \dots & 2 & 2 \end{pmatrix} \xrightarrow{L_n \leftarrow L_n - \frac{2}{1}L_2} \begin{pmatrix} 1 & -1 & \dots & \dots & \dots & -1 \\ 0 & 1 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & 1 & -1 \\ 0 & 0 & 4 & \dots & 4 & 4 \end{pmatrix}$$

[...]

$$\xrightarrow{L_n \leftarrow L_n - \frac{2^{n-2}}{1}L_{n-1}} \begin{pmatrix} 1 & -1 & \dots & \dots & \dots & -1 \\ 0 & 1 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & 1 & -1 \\ 0 & 0 & 0 & \dots & 0 & 2^{n-1} \end{pmatrix}$$

par conséquent on obtient les matrices

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \ddots & & \vdots \\ 0 & 0 & 1 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \dots & 0 & 1 & 0 \\ 1 & 2 & 4 & \dots & 2^{n-2} & 1 \end{pmatrix} \qquad \text{et} \qquad \mathbb{U} = \begin{pmatrix} 1 & -1 & -1 & \dots & \dots & -1 \\ 0 & 1 & -1 & \ddots & & \vdots \\ 0 & 0 & 1 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & -1 & -1 \\ 0 & 0 & \dots & 0 & 1 & -1 \\ 0 & 0 & 0 & \dots & 0 & 2^{n-1} \end{pmatrix}.$$

Exercice 4.17

Soit les systèmes linéaires

$$\begin{cases} 4x_1 + 3x_2 + 3x_3 = 10 \\ 3x_1 + 4x_2 + 3x_3 = 10 \\ 3x_1 + 3x_2 + 4x_3 = 10 \end{cases} \tag{4.1}$$

$$\begin{cases} 4x_1 + x_2 + x_3 = 6 \\ x_1 + 4x_2 + x_3 = 6 \\ x_1 + x_2 + 4x_3 = 6 \end{cases} \tag{4.2}$$

1. Rappeler une condition suffisante de convergence pour les méthodes de Jacobi et de Gauss-Seidel. Rappeler une

autre condition suffisante de convergence pour la méthode de GAUSS-SEIDEL (mais non pour la méthode de Jacobi). Les systèmes (4.1) et (4.2) vérifient-ils ces conditions ?

2. Écrire les méthodes de JACOBI et de GAUSS-SEIDEL pour ces deux systèmes linéaires.
3. On illustrera les résultats théoriques de convergence/non-convergence de ces deux schémas en prenant comme point de départ le vecteur $(x_1, x_2, x_3) = (0, 0, 0)$ et en calculant les 3 premiers itérés **dans l'un des cas suivant** (vous êtes libre de choisir) :
   3.1. avec la méthode de JACOBI pour le système (4.1),
   3.2. avec la méthode de GAUSS-SEIDEL pour le système (4.1),
   3.3. avec la méthode de JACOBI pour le système (4.2),
   3.4. avec la méthode de GAUSS-SEIDEL pour le système (4.2).
4. On comparera le résultat obtenu avec la solution exacte (qu'on calculera à l'aide de la méthode d'élimination de Gauss).

SOLUTION. Écrivons les deux systèmes sous forme matricielle $\mathbb{A}\mathbf{x} = \mathbf{b}$:

$$\underbrace{\begin{pmatrix} 4 & 3 & 3 \\ 3 & 4 & 3 \\ 3 & 3 & 4 \end{pmatrix}}_{\mathbb{A}_1} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 10 \\ 10 \\ 10 \end{pmatrix} \qquad \text{et} \qquad \underbrace{\begin{pmatrix} 4 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 4 \end{pmatrix}}_{\mathbb{A}_2} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 6 \\ 6 \end{pmatrix}$$

1. Rappelons deux propriétés de convergence :
   ▷ Si la matrice $\mathbb{A}$ est à diagonale dominante stricte, les méthodes de JACOBI et de GAUSS-SEIDEL convergent.
   ▷ Si la matrice $\mathbb{A}$ est symétrique et définie positive, la méthode de GAUSS-SEIDEL converge.
   Comme $4 > 1 + 1$, la matrice $\mathbb{A}_2$ est à diagonale dominante stricte : les méthodes de JACOBI et de GAUSS-SEIDEL convergent.
   Comme $4 < 3 + 3$, la matrice $\mathbb{A}_1$ n'est pas à diagonale dominante stricte : les méthodes de JACOBI et de GAUSS-SEIDEL peuvent ne pas converger. Cependant elle est symétrique et définie positive (car les valeurs propres[1] sont $\lambda_1 = \lambda_2 = 1$ et $\lambda_3 = 10$) : la méthode de GAUSS-SEIDEL converge.
2. Pour les systèmes donnés les méthodes de JACOBI et GAUSS-SEIDEL s'écrivent

	$\mathbb{A}_1\mathbf{x} = \mathbf{b}$	$\mathbb{A}_2\mathbf{x} = \mathbf{b}$
Jacobi	$\begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 10 - 3x_2^{(k)} - 3x_3^{(k)} \\ 10 - 3x_1^{(k)} - 3x_3^{(k)} \\ 10 - 3x_1^{(k)} - 3x_2^{(k)} \end{pmatrix}$	$\begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 6 - x_2^{(k)} - x_3^{(k)} \\ 6 - x_1^{(k)} - x_3^{(k)} \\ 6 - x_1^{(k)} - x_2^{(k)} \end{pmatrix}$
Gauss-Seidel	$\begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 10 - 3x_2^{(k)} - 3x_3^{(k)} \\ 10 - 3x_1^{(k+1)} - 3x_3^{(k)} \\ 10 - 3x_1^{(k+1)} - 3x_2^{(k+1)} \end{pmatrix}$	$\begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 6 - x_2^{(k)} - x_3^{(k)} \\ 6 - x_1^{(k+1)} - x_3^{(k)} \\ 6 - x_1^{(k+1)} - x_2^{(k+1)} \end{pmatrix}$

3. On obtient les suites suivantes
   3.1. JACOBI pour le système (4.1) :

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \implies \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(1)} = \frac{1}{4} \begin{pmatrix} 10 - 3 \times 0 - 3 \times 0 \\ 10 - 3 \times 0 - 3 \times 0 \\ 10 - 3 \times 0 - 3 \times 0 \end{pmatrix} = \begin{pmatrix} \frac{5}{2} \\ \frac{5}{2} \\ \frac{5}{2} \end{pmatrix}$$

$$\implies \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(2)} = \frac{1}{4} \begin{pmatrix} 10 - 3 \times \frac{5}{2} - 3 \times \frac{5}{2} \\ 10 - 3 \times \frac{5}{2} - 3 \times \frac{5}{2} \\ 10 - 3 \times \frac{5}{2} - 3 \times \frac{5}{2} \end{pmatrix} = \begin{pmatrix} -\frac{5}{4} \\ -\frac{5}{4} \\ -\frac{5}{4} \end{pmatrix} \implies \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(3)} = \frac{1}{4} \begin{pmatrix} 10 - 3 \times \frac{-5}{4} - 3 \times \frac{-5}{4} \\ 10 - 3 \times \frac{-5}{4} - 3 \times \frac{-5}{4} \\ 10 - 3 \times \frac{-5}{4} - 3 \times \frac{-5}{4} \end{pmatrix} = \begin{pmatrix} \frac{35}{8} \\ \frac{35}{8} \\ \frac{35}{8} \end{pmatrix}$$

1. $\det \mathbb{A}_1(\lambda) = (4-\lambda)^3 + 27 + 27 - 9(4-\lambda) - 9(4-\lambda) - 9(4-\lambda) = 64 - 48\lambda + 12\lambda^2 - \lambda^3 + 54 - 108 + 27\lambda = -\lambda^3 + 12\lambda^2 - 21\lambda + 10$. Une racine évidente est $\lambda = 1$ et on obtient $\det \mathbb{A}_1(\lambda) = (\lambda - 1)(-\lambda^2 + 11\lambda - 10) = (\lambda - 1)^2(\lambda - 10)$.

3.2. GAUSS-SEIDEL pour le système (4.1) :

$$\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}^{(0)}=\begin{pmatrix}0\\0\\0\end{pmatrix}\implies\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}^{(1)}=\frac{1}{4}\begin{pmatrix}10-3\times0-3\times0\\10-3\times\frac{5}{2}-3\times0\\10-3\times\frac{5}{2}-3\times\frac{5}{8}\end{pmatrix}=\begin{pmatrix}\frac{5}{2}\\\frac{5}{8}\\\frac{5}{32}\end{pmatrix}$$

$$\implies\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}^{(2)}=\frac{1}{4}\begin{pmatrix}10-3\times\frac{5}{8}-3\times\frac{5}{32}\\10-3\times\frac{245}{128}-3\times\frac{5}{32}\\10-3\times\frac{245}{128}-3\times\frac{485}{512}\end{pmatrix}=\begin{pmatrix}\frac{245}{128}\\\frac{485}{512}\\\frac{725}{2048}\end{pmatrix}\implies\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}^{(3)}=\begin{pmatrix}\frac{12485}{8192}\\\frac{35765}{32768}\\\frac{70565}{131072}\end{pmatrix}$$

3.3. JACOBI pour le système (4.2) :

$$\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}^{(0)}=\begin{pmatrix}0\\0\\0\end{pmatrix}\implies\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}^{(1)}=\frac{1}{4}\begin{pmatrix}6-1\times0-1\times0\\6-1\times0-1\times0\\6-1\times0-1\times0\end{pmatrix}=\begin{pmatrix}\frac{3}{2}\\\frac{3}{2}\\\frac{3}{2}\end{pmatrix}$$

$$\implies\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}^{(2)}=\frac{1}{4}\begin{pmatrix}6-1\times\frac{3}{2}-1\times\frac{3}{2}\\6-1\times\frac{3}{2}-1\times\frac{3}{2}\\6-1\times\frac{3}{2}-1\times\frac{3}{2}\end{pmatrix}=\begin{pmatrix}\frac{3}{4}\\\frac{3}{4}\\\frac{3}{4}\end{pmatrix}\implies\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}^{(3)}=\frac{1}{4}\begin{pmatrix}6-1\times\frac{3}{4}-1\times\frac{3}{4}\\6-1\times\frac{3}{4}-1\times\frac{3}{4}\\6-1\times\frac{3}{4}-1\times\frac{3}{4}\end{pmatrix}=\begin{pmatrix}\frac{9}{8}\\\frac{9}{8}\\\frac{9}{8}\end{pmatrix}$$

3.4. GAUSS-SEIDEL pour le système (4.2) :

$$\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}^{(0)}=\begin{pmatrix}0\\0\\0\end{pmatrix}\implies\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}^{(1)}=\frac{1}{4}\begin{pmatrix}6-1\times0-1\times0\\6-1\times\frac{3}{2}-1\times0\\6-1\times\frac{3}{2}-1\times\frac{9}{8}\end{pmatrix}\begin{pmatrix}\frac{3}{2}\\\frac{9}{8}\\\frac{27}{32}\end{pmatrix}$$

$$\implies\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}^{(2)}=\frac{1}{4}\begin{pmatrix}6-1\times\frac{9}{8}-1\times\frac{27}{32}\\6-1\times\frac{129}{128}-1\times\frac{27}{32}\\6-1\times\frac{129}{128}-1\times\frac{531}{512}\end{pmatrix}=\begin{pmatrix}\frac{129}{128}\\\frac{531}{512}\\\frac{2025}{2048}\end{pmatrix}\implies\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}^{(3)}=\frac{1}{4}\begin{pmatrix}6-1\times\frac{531}{512}-1\times\frac{2025}{2048}\\6-1\times\frac{8139}{8192}-1\times\frac{2025}{2048}\\6-1\times\frac{8139}{8192}-1\times\frac{32913}{32768}\end{pmatrix}=\begin{pmatrix}\frac{8139}{8192}\\\frac{32913}{32768}\\\frac{131139}{131072}\end{pmatrix}$$

4. Calcul de la solution exacte à l'aide de la méthode d'élimination de Gauss :
   ▷ Système (4.1) :

$$\left(\begin{array}{ccc|c}4&3&3&10\\3&4&3&10\\3&3&4&10\end{array}\right)\xrightarrow[L_3\leftarrow L_3-\frac{3}{4}L_1]{L_2\leftarrow L_2-\frac{3}{4}L_1}\left(\begin{array}{ccc|c}4&3&3&10\\0&7/4&3/4&5/2\\0&3/4&7/4&5/2\end{array}\right)\xrightarrow{L_3\leftarrow L_3-\frac{3/4}{7/4}L_2}\left(\begin{array}{ccc|c}4&3&3&10\\0&7/4&3/4&5/2\\0&0&10/7&10/7\end{array}\right)\implies\mathbf{x}=\begin{pmatrix}1\\1\\1\end{pmatrix}$$

   ▷ Système (4.2) :

$$\left(\begin{array}{ccc|c}4&1&1&6\\1&4&1&6\\1&1&4&6\end{array}\right)\xrightarrow[L_3\leftarrow L_3-\frac{1}{4}L_1]{L_2\leftarrow L_2-\frac{1}{4}L_1}\left(\begin{array}{ccc|c}4&1&1&6\\0&15/4&3/4&9/2\\0&3/4&15/4&9/2\end{array}\right)\xrightarrow{L_3\leftarrow L_3-\frac{3/4}{15/4}L_2}\left(\begin{array}{ccc|c}4&1&1&6\\0&15/4&3/4&9/2\\0&0&18/5&18/5\end{array}\right)\implies\mathbf{x}=\begin{pmatrix}1\\1\\1\end{pmatrix}$$

5. Équations différentielles ordinaires

Calculer la fonction $t \mapsto y(t)$ qui vérifie l'EDO $y'(t) = f(t, y(t))$ et la condition $y(t_0) = y_0$

Les équations différentielles décrivent l'évolution de nombreux phénomènes dans des domaines variés. Une équation différentielle est une équation impliquant une ou plusieurs dérivées d'une fonction inconnue. Si toutes les dérivées sont prises par rapport à une seule variable, on parle d'équation différentielle ordinaire. Une équation mettant en jeu des dérivées partielles est appelée équation aux dérivées partielles. On dit qu'une équation différentielle (ordinaire ou aux dérivées partielles) est d'ordre p si elle implique des dérivées d'ordre au plus p. Dans le présent chapitre, nous considérons des équations différentielles ordinaires d'ordre un.

Équations différentielles

Une équation différentielle (EDO) est une équation, dont l'inconnue est une fonction y, exprimée sous la forme d'une relation $F(y, y', y'', \dots, y^{(n)}) = g(x)$ dans laquelle cohabitent à la fois $y = y(x)$ et ses dérivées y', y'', ... (n est appelé l'ordre de l'équation). Si la fonction g, appelée «second membre» de l'équation, est nulle, on dit que l'équation en question est homogène.

Nous pouvons nous limiter aux équations différentielles du premier ordre, car une équation d'ordre $p > 1$ peut toujours se ramener à un système de p équations d'ordre 1. Une équation différentielle ordinaire admet généralement une infinité de solutions. Pour en sélectionner une, on doit imposer une condition supplémentaire qui correspond à la valeur prise par la solution en un point de l'intervalle d'intégration. On considérera par conséquent des problèmes, dits de CAUCHY ainsi défini :

Problème de CAUCHY

Soit $f\colon I \times \mathbb{R} \to \mathbb{R}$ une fonction donnée et y' la dérivée de y par rapport à t. On appelle *problème de* CAUCHY le problème

trouver $y\colon I \subset \mathbb{R} \to \mathbb{R}$ tel que

$$\begin{cases} y'(t) = f(t, y(t)), & \forall t \in I, \\ y(t_0) = y_0, \end{cases}$$

avec t_0 un point de I et y_0 une valeur appelée donnée initiale.

Exemple

On se donne $f(t, y(t)) = 3t - 3y(t)$ et $y_0 = \alpha$ (un nombre quelconque). On cherche une fonction $y\colon t \in \mathbb{R}^+ \mapsto y(t) \in \mathbb{R}$ qui satisfait

$$\begin{cases} y'(t) = 3t - 3y(t), & \forall t > 0, \\ y(0) = \alpha. \end{cases}$$

Sa solution est donnée par $y(t) = (\alpha - 1/3)e^{3t} + t + 1/3$.

Exemple *Non unicité de la solution d'un problème de* CAUCHY

On se donne $f(t, y(t)) = \sqrt[3]{y(t)}$ et $y_0 = 0$. On cherche une fonction $y\colon t \in \mathbb{R}^+ \mapsto y(t) \in \mathbb{R}$ qui satisfait

$$\begin{cases} y'(t) = \sqrt[3]{y(t)}, & \forall t > 0, \\ y(0) = 0. \end{cases}$$

On vérifie que les fonctions $y_1(t) = 0$ et $y_{2,3}(t) = \pm\sqrt{8t^3/27}$, pour tout $t \ge 0$, sont toutes trois solutions du problème de CAUCHY donné. C'est exemple montre qu'un problème de CAUCHY n'as pas nécessairement de solution unique.

Exemple *Non existence sur $\mathbb{R}$ de la solution d'un problème de* CAUCHY

On se donne $f(t, y(t)) = (y(t))^3$ et $y_0 = 1$. On cherche une fonction $y\colon t \in \mathbb{R}^+ \mapsto y(t) \in \mathbb{R}$ qui satisfait

$$\begin{cases} y'(t) = (y(t))^3, & \forall t > 0, \\ y(0) = 1. \end{cases}$$

On vérifie que la solution y est donnée par $y(t) = 1/\sqrt{1-2t}$ qui n'est définie que pour $t \in [0; 1/2[$. C'est exemple montre qu'un problème de CAUCHY n'as pas toujours une solution pour tout $t \in [0; +\infty[$ puisqu'ici la solution explose lorsque t tend vars la valeur $1/2$ (en effet, nous avons $\lim_{t \to (1/2)^-} y(t) = +\infty$).

Les trois exemples ci-dessus montrent que l'étude mathématique de l'existence et de l'unicité de solutions d'un problème de CAUCHY peut être une affaire délicate. Dans ce chapitre, nous nous contentons de rappeler un résultat d'existence et d'unicité global, au sens où on peut intégrer le problème de CAUCHY jusqu'à $t = \infty$:

Théorème de CAUCHY-Lipschitz

Soit un problème de CAUCHY. Si la fonction $f(t, y)$ est

1. continue par rapport à ses deux variables ;
2. lipschitzienne par rapport à sa deuxième variable, c'est-à-dire qu'il existe une constante positive L (appelée constante de Lipschitz) telle que

$$|f(t, y_1) - f(t, y_2)| \le L|y_1 - y_2|, \qquad \forall t \in I, \quad \forall y_1, y_2 \in \mathbb{R},$$

alors la solution $y = y(t)$ du problème de CAUCHY existe, est unique et appartient à $\mathscr{C}^1(I)$.

Cas particulier

Soient a et g deux fonctions continues d'un intervalle I dans $\mathbb{R}$, si $f(t, y(t)) = a(t)y(t) + g(t)$ et si $(t_0, y_0) \in I \times \mathbb{R}$, alors il existe une unique solution y de l'équation différentielle telle que $y(t_0) = y_0$.

Remarque

Graphiquement, ce théorème signifie que par tout point du plan dont l'abscisse est dans I, il passe une courbe intégrale et une seule, autrement dit deux trajectoires ne peuvent pas se croiser. En particulier, si une équation différentielle admet comme solution la solution nulle, alors toute autre solution est soit toujours positive soit toujours négative.

Exemple

On se donne $f(t, y(t)) = |y(t)| + \sin(y(t)) + e^{-t^2/2}$ et $y_0 = 1$. On cherche une fonction $y\colon t \in \mathbb{R}^+ \mapsto y(t) \in \mathbb{R}$ qui satisfait

$$\begin{cases} y'(t) = (y(t))^3, & \forall t > 0, \\ y(0) = 1. \end{cases}$$

On vérifie facilement que

$$\begin{aligned} \left| f(t, y_1(t)) - f(t, y_2(t)) \right| &= \left| |y_1(t)| + \sin(y_1(t)) - |y_2(t)| - \sin(y_2(t)) \right| \\ &= \left| |y_1(t)| - |y_2(t)| + \int_{y_2}^{y_1} \sin(\vartheta)\, d\vartheta \right| \\ &\le \left| |y_1(t)| - |y_2(t)| + |y_1(t) - y_2(t)| \right| \\ &\le \left| y_1(t) - y_2(t) \right| \end{aligned}$$

Par le théorème de CAUCHY-Lipschitz, le problème de CAUCHY a une solution globale unique $y(t)$. Cependant, il n'est pas possible de donner une expression explicite de la solution.

En pratique, on ne peut expliciter les solutions que pour des équations différentielles ordinaires très particulières. Dans certains cas, on ne peut exprimer la solution que sous forme implicite. Dans d'autres cas, on ne parvient même pas à représenter la solution sous forme implicite. Pour ces raisons, on cherche des méthodes numériques capables d'approcher la solution de toutes les équations différentielles qui admettent une solution.

Problème de CAUCHY numériquement mal posé

Une fois calculée la solution numérique $\{u_n\}_{n=1}^{N_h}$, il est légitime de chercher à savoir dans quelle mesure l'erreur $|y(t_n) - u_n|$ est petite ou non pour $n = 1, 2, \dots$. Nous essayons de répondre à cette question en reprenant le premier exemple du chapitre. On se donne $f(t, y(t)) = 3t - 3y(t)$ et $y_0 = \alpha$ (un nombre quelconque). On cherche une fonction $y \colon t \in \mathbb{R}^+ \mapsto y(t) \in \mathbb{R}$ qui satisfait

$$\begin{cases} y'(t) = 3t - 3y(t), & \forall t > 0, \\ y(0) = \alpha. \end{cases}$$

Nous avons vu que sa solution est donnée par $y(t) = (\alpha - 1/3)e^{3t} + t + 1/3$. Si nous cherchons à résoudre le problème de CAUCHY jusqu'à $t = 10$ avec $\alpha = 1/3$, nous obtenons $y(10) = 10 + 1/3 = 31/3$. Par contre, si nous faisons le calcul avec l'approximation $\alpha = 0.333333$ au lieu de $1/3$, nous avons $y(10) = (0.333333 - 1/3)e^{30} + 10 + 1/3 = -e^{30}/3000000 + 31/3$ ce qui représente une différence avec la précédente valeur de $e^{30}/3000000 \approx 10^7/3$. Cet exemple nous apprend qu'une petite erreur sur la condition initiale (erreur relative d'ordre 10^{-6}) peut provoquer une très grande erreur sur $y(10)$ (erreur relative d'ordre 10^6). Ainsi, si le calculateur mis à notre disposition ne calcul qu'avec 6 chiffres significatifs (en virgule flottante), alors $\alpha = 1/3$ devient $\alpha = 0.333333$ et il est inutile d'essayer d'inventer une méthode numérique pour calculer $y(10)$. En effet, la seule erreur sur la condition initiale provoque déjà une erreur inadmissible sur la solution. Nous sommes en présence ici d'un problème numériquement mal posé, appelé aussi problème mal conditionné.

Schémas numériques

Considérons le problème de CAUCHY et supposons que l'on ait montré l'existence d'une solution y.

Le principe de toutes ces méthodes est de subdiviser l'intervalle $I = [t_0, T]$, avec $T < +\infty$, en N_h intervalles de longueur $h = (T - t_0)/N_h = t_{n+1} - t_n$; h est appelé le pas de discrétisation. Alors, pour chaque nœud $t_n = t_0 + nh$ $(1 \le n \le N_h)$ on cherche la valeur inconnue u_n qui approche $y(t_n)$. L'ensemble des valeurs $\{u_0 = y_0, u_1, \dots, u_{N_h}\}$ représente la solution numérique. Les schémas qu'on va construire permettent de calculer u_{n+1} à partir de u_n et il est donc possible de calculer successivement $u_1, u_2, \dots$, en partant de u_0.

Si nous intégrons l'EDO $y'(t) = f(t, y(t))$ entre t_n et t_{n+1} nous obtenons

$$y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} f(t, y(t))dt.$$

Soit u_n une approximation de $y(t_n)$ et u_{n+1} une approximation de $y(t_{n+1})$. On peut construire différentes schémas selon la formule de quadrature utilisée pour approcher le membre de droite.

▷ Si on utilise la formule de quadrature du rectangle à gauche, *i.e.*

$$\int_{t_n}^{t_{n+1}} f(t, y(t))dt \approx hf(t_n, y(t_n))$$

on obtient le *schéma d'*EULER *progressif*

$$\begin{cases} u_0 = y(y_0) = y_0, \\ u_{n+1} = u_n + hf(t_n, u(t_n)) & n = 0, 1, 2, \dots \end{cases}$$

Il s'agit d'un schéma explicite car il permet d'expliciter u_{n+1} en fonction de u_n.

▷ Si on utilise la formule de quadrature du rectangle à droite, *i.e.*

$$\int_{t_n}^{t_{n+1}} f(t, y(t))dt \approx hf(t_{n+1}, y(t_{n+1}))$$

on obtient le *schéma d'*EULER *rétrograde*

$$\begin{cases} u_0 = y(y_0) = y_0, \\ u_{n+1} - hf(t_{n+1}, u_{n+1}) = u_n & n = 0, 1, 2, \dots \end{cases}$$

Il s'agit d'un schéma implicite car il ne permet pas d'expliciter directement u_{n+1} en fonction de u_n lorsque la fonction f n'est pas triviale.

▷ Si on utilise la formule de quadrature du point du milieu, *i.e.*

$$\int_{t_n}^{t_{n+1}} f(t,y(t))dt \approx hf\left(t_n+\frac{h}{2}, y\left(t_n+\frac{h}{2}\right)\right)$$

on obtient un nouveau schéma :

$$\begin{cases} u_0 = y(y_0) = y_0, \\ u_{n+1} = u_n + hf(t_n+h/2, u_{n+1/2}) \quad n = 0,1,2,\dots \end{cases}$$

où $u_{n+1/2}$ est une approximation de $y(t_n+h/2)$. Nous pouvons utiliser une prédiction d'EULER progressive pour approcher le $u_{n+1/2}$ dans le terme $f(t_n+h/2, u_{n+1/2})$ par $\tilde{u}_{n+1/2} = u_n + (h/2)f(t_n, u_n)$. Nous avons construit ainsi un nouveau schéma appelé *schéma d'*EULER *modifié* qui s'écrit

$$\begin{cases} u_0 = y(y_0) = y_0, \\ \tilde{u}_{n+1/2} = u_n + (h/2)f(t_n, u_n), \\ u_{n+1} = u_n + hf(t_n+h/2, \tilde{u}_{n+1/2}) \quad n = 0,1,2,\dots \end{cases}$$

Il s'agit d'un schéma explicite car il permet d'expliciter u_{n+1} en fonction de u_n.

▷ Si on utilise la formule du trapèze, *i.e.*

$$\int_{t_n}^{t_{n+1}} f(t,y(t))dt \approx \frac{h}{2}\big(f(t_n, y(t_n)) + f(t_{n+1}, y(t_{n+1}))\big)$$

on obtient le *schéma du trapèze ou de* CRANCK-NICHOLSON

$$\begin{cases} u_0 = y(y_0) = y_0, \\ u_{n+1} - hf(t_{n+1}, u_{n+1}) = u_n + hf(t_n, u_n) \quad n = 0,1,2,\dots \end{cases}$$

Il s'agit à nouveau d'un schéma implicite car il ne permet pas d'expliciter directement u_{n+1} en fonction de u_n lorsque la fonction f n'est pas triviale. En fait, ce schéma fait la moyenne des schémas d'EULER progressif et rétrograde.

▷ Pour éviter le calcul implicite de u_{n+1} dans le schéma du trapèze, nous pouvons utiliser une prédiction d'EULER progressive et remplacer le u_{n+1} dans le terme $f(t_{n+1}, u_{n+1})$ par $\tilde{u}_{n+1} = u_n + hf(t_n, u_n)$. Nous avons construit ainsi un nouveau schéma appelé *schéma de* HEUN. Plus précisément, la méthode de HEUN s'écrit

$$\begin{cases} u_0 = y(y_0) = y_0, \\ \tilde{u}_{n+1} = u_n + hf(t_n, u_n), \\ u_{n+1} = u_n + h\big(f(t_n, u_n) + f(t_{n+1}, \tilde{u}_{n+1})\big) \quad n = 0,1,2,\dots \end{cases}$$

Remarque

Considérons le schéma d'EULER rétrograde. Si nous voulons calculer u_{n+1}, nous définissons la fonction

$$g(x) = x - hf(t_{n+1}, x) - u_n$$

et nous cherchons un zéro de $g(x)$ en prenant par exemple la méthode de NEWTON. Ainsi nous pouvons poser $x_0 = u_0$ et $x_{m+1} = x_m - g(x_m)/g'(x_m)$, $m = 0,1,\dots$. Puisque $g'(x) = 1 - h\partial_x f(t_{n+1}, x)$, nous obtenons donc dans ce cas le schéma

$$\begin{cases} x_0 = u_n, \\ x_{m+1} = x_m - \frac{x_m - hf(t_{n+1},x) - u_n}{1 - h\partial_x f(t_{n+1},x)} \quad m = 0,1,2,\dots \end{cases}$$

et $u_{n+1} = \lim_{m\to\infty} x_m$ pour autant que f soit suffisamment régulière et que x_0 soit suffisamment proche de u_{n+1}, ce qui est le cas si le pas h est suffisamment petit.

Stabilité

Dans la section précédente, on a considéré la résolution du problème de CAUCHY sur des intervalles bornés. Dans ce cadre, le nombre N_h de sous-intervalles ne tend vers l'infini que quand h tend vers zéro. Il existe cependant de nombreuses situations dans lesquelles le problème de CAUCHY doit être intégré sur des intervalles en temps très grands ou même infini. Dans ce cas, même pour h fixé, N_h tend vers l'infini. On s'intéresse donc à des méthodes capables d'approcher la solution pour des intervalles en temps arbitrairement grands, même pour des pas de temps h «assez grands».

A première vue, il semble que le schéma d'EULER progressif soit préférable au schéma d'EULER rétrograde puisque ce dernier n'est pas explicite. Cependant, nous verrons sur un exemple que le schéma progressif peut engendrer des difficultés que le schéma rétrograde n'engendre pas.

Considérons le problème de CAUCHY dans le cas particulier où $f(t, y(t)) = -\beta y(t)$ avec β un nombre réel positif donné. Sa solution est trivialement $y(t) = y_0 e^{-\beta t}$. Puisque β est positif, ce problème est numériquement bien posé : $y(t)$ décroît exponentiellement lorsque t tend vers l'infini. Pour discrétiser la demi-droite $t \ge 0$, nous choisissons un nombre réel $h > 0$ et nous posons $t_n = nh$ avec $n = 0, 1, 2, \dots$ Nous allons étudier dans ce cadre les deux schémas d'EULER, à savoir le schéma d'EULER progressif et le schéma d'EULER rétrograde.

- ▷ Le schéma d'EULER progressif devient

$$u_{n+1} = (1 - \beta h) u_n, \qquad n = 0, 1, 2, \dots$$

et par suite

$$u_{n+1} = (1 - \beta h)^n u_0, \qquad n = 0, 1, 2, \dots$$

Bien que la solution $y(t)$ tend vers zéro lorsque t tend vers l'infini, nous voyons que si $u_0 \neq 0$ et $1 - \beta h < -1$ alors u_n tend vers l'infini en alternant de signe lorsque t tend vers l'infini. Nous dirons dans ce cas que le schéma d'EULER progressif est instable. Pour éviter de phénomène, il convient donc d'imposer $-1 \le 1 - \beta h$, ce qui aura pour effet de limiter h à

$$h \le \frac{2}{\beta}.$$

Cette condition est appelée condition de stabilité : elle limite le pas h d'avance en t lorsqu'on utilise le schéma d'EULER progressif.

- ▷ Le schéma d'EULER rétrograde devient dans le cadre de notre exemple

$$(1 + \beta h) u_{n+1} = u_n, \qquad n = 0, 1, 2, \dots$$

et par suite

$$u_{n+1} = \frac{1}{(1 - \beta h)^{n+1}} u_0, \qquad n = 0, 1, 2, \dots$$

Dans ce cas nous voyons que pour tout $h > 0$ nous avons $\lim_{n\to\infty} u_n = 0$, le schéma d'EULER rétrograde est donc toujours stable, sans limitations sur h.

Définition

La propriété

$$\lim_{n\to+\infty} u_n = 0$$

est appelée stabilité absolue (ou A-stabilité).

✎✎✎✎✎✎✎✎✎✎✎✎✎✎✎ Exercices ✎✎✎✎✎✎✎✎✎✎✎✎✎✎

Exercice 5.1 *Stabilité de la méthode d'EULER explicite en fonction du pas*

On considère le problème de CAUCHY

$$\begin{cases} y'(t) = -y(t), \\ y(0) = 1, \end{cases}$$

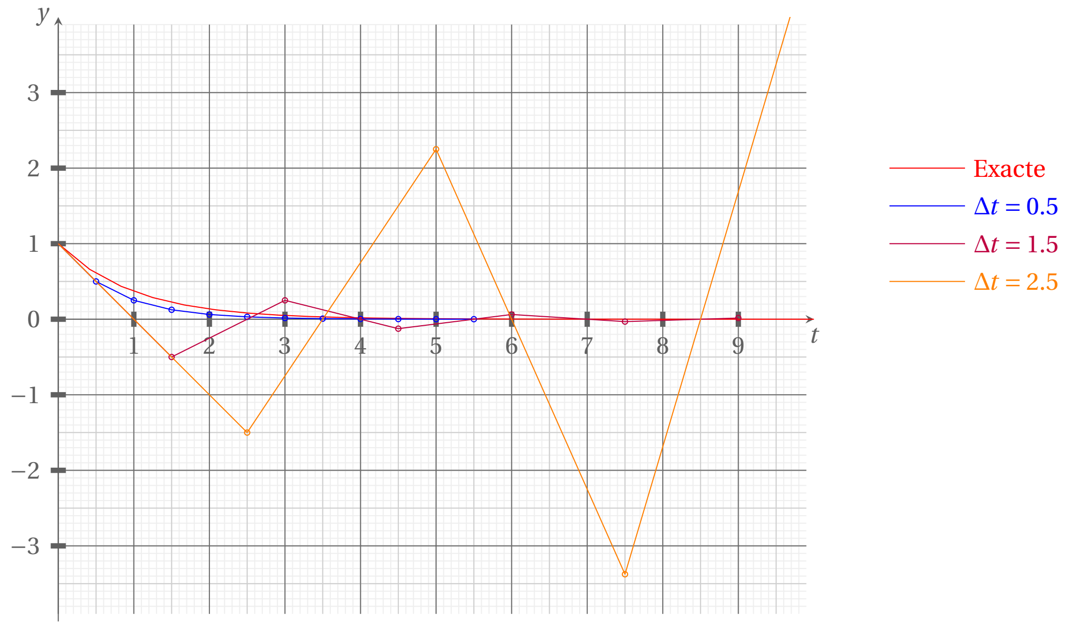
sur l'intervalle $[0;10]$.

1. Calculer la solution exacte du problème de CAUCHY.
2. Soit Δt le pas temporel. Écrire la méthode d'EULER explicite pour cette équation différentielle ordinaire (EDO).
3. En déduire une forme du type
$$y_{k+1} = g(\Delta t, k)$$
avec $g(\Delta t, k)$ à préciser (autrement dit, l'itérée en t_k ne dépend que de Δt et k et ne dépend pas de y_k).
4. Utiliser la formulation ainsi obtenue pour tracer les solutions
   - ▷ exacte,
   - ▷ obtenue avec la méthode d'EULER avec $\Delta t = 2.5$,
   - ▷ obtenue avec la méthode d'EULER avec $\Delta t = 1.5$,
   - ▷ obtenue avec la méthode d'EULER avec $\Delta t = 0.5$.
5. Que peut-on en déduire sur la stabilité de la méthode ?

SOLUTION.

1. Il s'agit d'une EDO à variables séparables. L'unique solution constante de l'EDO est la fonction $y(t) \equiv 0$, toutes les autres solutions sont du type $y(t) = Ce^{-t}$. Donc l'unique solution du problème de CAUCHY est la fonction $y(t) = e^{-t}$ définie pour tout $t \in \mathbb{R}$.
2. La méthode d'EULER est une méthode d'intégration numérique d'EDO du premier ordre de la forme $y'(t) = F(t, y(t))$ C'est une méthode itérative : la valeur y à l'instant $t+\Delta t$ se déduisant de la valeur de y à l'instant t par l'approximation linéaire
$$y(t+\Delta t) \approx y(t) + y'(t)\Delta t = y(t) + F(t, y(t))\Delta t.$$
En choisissant un pas de discrétisation Δt, nous obtenons une suite de valeurs (t_k, y_k) qui peuvent être une excellente approximation de la fonction $y(t)$ avec
$$\begin{cases} t_k = t_0 + k\Delta t, \\ y_k = y_{k-1} + F(t_{k-1}, y_{k-1})\Delta t. \end{cases}$$
La méthode d'EULER explicite pour cette EDO s'écrit donc
$$y_{k+1} = (1-\Delta t) y_k.$$
3. En procédant par récurrence sur k, on obtient
$$y_{k+1} = (1-\Delta t)^{k+1}.$$
4. On a donc
   - ▷ si $\Delta t = 2.5$ alors $y_k = \left(-\frac{3}{2}\right)^k$,
   - ▷ si $\Delta t = 1.5$ alors $y_k = \left(-\frac{1}{2}\right)^k$,
   - ▷ si $\Delta t = 0.5$ alors $y_k = \left(\frac{1}{2}\right)^k$.

   Ci-dessous sont tracées sur l'intervalle $[0;10]$, les courbes représentatives de la solution exacte et de la solution calculée par la méthode d'EULER explicite. En faisant varier le pas Δt nous pouvons constater que si $\Delta t = 2.5$ l'erreur commise entre la solution exacte et la solution calculée est amplifiée d'un pas à l'autre.



NB : les trois premières itérées ont la même pente (se rappeler de la construction géométrique de la méthode d'EULER).

5. De la formule $y_{k+1} = (1-\Delta t)^{k+1}$ on déduit que
   - si $0 < \Delta t < 1$ alors la solution numérique est stable et convergente,
   - si $1 < \Delta t < 2$ alors la solution numérique oscille mais est encore convergente,
   - si $\Delta t > 2$ alors la solution numérique oscille et divergente.

   En effet, on sait que la méthode est absolument stable si et seulement si $|1-\Delta t| < 1$.

   Remarque : la suite obtenue est une suite géométrique de raison $q = 1-\Delta t$. On sait qu'une telle suite
   - diverge si $|q| > 1$ ou $q = -1$,
   - est stationnaire si $q = 1$,
   - converge vers 0 $|q| < 1$.

Exercice 5.2

Soit $\beta > 0$ un nombre réel positif et considérons le problème de CAUCHY

$$\begin{cases} y'(t) = -\beta y(t), & \text{pour } t > 0, \\ y(0) = y_0, \end{cases} \tag{5.1}$$

où y_0 est une valeur donnée. Soit $\Delta t > 0$ un pas de temps donné, soit $t_i = i\Delta t$ pour $i \in \mathbb{N}$ et y_i une approximation de $y(t_i)$.

1. Écrire le schéma de HEUN permettant de calculer y_{i+1} à partir de y_i.
2. Sous quelle condition la relation $\lim_{i\to+\infty} y_i = 0$ a-t-elle lieu ? Autrement dit, sous quelle condition le schéma de HEUN est A-stable ?

SOLUTION.

1. Le problème (5.1) est un problème du type

   trouver $y\colon I \subset \mathbb{R}^+ \to \mathbb{R}$ tel que

$$\begin{cases} y'(t) = f(t, y(t)), & \forall t \in I, \\ y(t_0) = y_0. \end{cases}$$

   Le principe des méthodes d'approximation est de subdiviser l'intervalle I en sous-intervalles de longueur Δt et, pour chaque nœud $t_n = t_0 + i\Delta t$ ($i \geq 0$), on cherche la valeur inconnue y_i qui approche $y(t_i)$. L'ensemble de valeurs $\{y_0, y_1, \dots\}$ représente la solution numérique. Le schéma de HEUN permet de calculer y_{i+1} à partir de y_i et il est donc possible de calculer successivement y_1, y_2,... en partant de y_0.

   Si nous intégrons l'EDO $y'(t) = f(t, y(t))$ entre t_i et t_{i+1} nous obtenons

$$y(t_{i+1}) - y(t_i) = \int_{t_i}^{t_{i+1}} f(t, y(t))\, \mathrm{d}t.$$

   Soit y_i une approximation de $y(t_i)$ et y_{i+1} une approximation de $y(t_{i+1})$. Si on utilise la formule du trapèze, *i.e.*

$$\int_{t_i}^{t_{i+1}} f(t, y(t))\, \mathrm{d}t \approx \frac{\Delta t}{2}\big(f(t_i, y(t_i)) + f(t_{i+1}, y(t_{i+1}))\big)$$

on obtient le *schéma du trapèze ou de* CRANCK-NICHOLSON

$$\begin{cases} y_0 = y(t_0), \\ y_{i+1} - \frac{\Delta t}{2} f(t_{i+1}, y_{i+1}) = y_i + \frac{\Delta t}{2} f(t_i, y_i), \quad \text{pour } i = 0, 1, 2, \dots \end{cases}$$

Il s'agit d'un schéma implicite car il ne permet pas d'écrire directement y_{i+1} en fonction de y_i lorsque la fonction f n'est pas triviale. Pour éviter le calcul implicite de y_{i+1} dans le schéma du trapèze, nous pouvons utiliser une prédiction d'EULER progressive et remplacer le y_{i+1} dans le terme $f(t_{i+1}, y_{i+1})$ par $\tilde{y}_{i+1} = y_i + \Delta t f(t_i, y_i)$. Nous avons construit ainsi le *schéma de* HEUN. Plus précisément, cette méthode s'écrit

$$\begin{cases} y_0 = y(t_0), \\ y_{i+1} = y_i + \frac{\Delta t}{2} \left(f(t_i, y_i) + f\left(t_{i+1}, y_i + \Delta t f(t_i, y_i)\right) \right), \quad \text{pour } i = 0, 1, 2, \dots \end{cases}$$

2. En appliquant le schéma de HEUN au problème (5.1) on obtient la suite définie par récurrence suivante

$$\begin{cases} y_0 = y(t_0), \\ y_{i+1} = \left(1 - \beta\Delta t + \frac{(\beta\Delta t)^2}{2}\right) y_i. \end{cases}$$

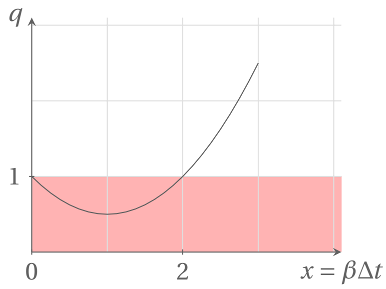
Par induction on obtient

$$y_i = \left(1 - \beta\Delta t + \frac{(\beta\Delta t)^2}{2}\right)^i y_0.$$

Par conséquent, $\lim_{i\to+\infty} y_i = 0$ si et seulement si

$$\left|1 - \beta\Delta t + \beta^2 \frac{(\Delta t)^2}{2}\right| < 1.$$

Notons x le produit $\beta\Delta t$ et q le polynôme $q(x) = \frac{1}{2}x^2 - x + 1$ dont le graphe est représenté en figure.



Nous avons $|q(x)| < 1$ si et seulement si $0 < x < 2$. La relation $\lim_{i\to+\infty} y_i = 0$ est donc satisfaite si et seulement si

$$\Delta t < \frac{2}{\beta}.$$

Exercice 5.3

On considère le problème de CAUCHY

$$\begin{cases} y'(t) = -(y(t))^m + \cos(t), \quad \text{pour } t > 0, \\ y(0) = 0, \end{cases} \tag{5.2}$$

où m est un entier impair.

1. Montrer que le problème (5.2) possède une solution unique globale.
2. Soit $\Delta t > 0$ un pas de temps donné, soit $t_i = i\Delta t$ pour $i \in \mathbb{N}$ et y_i une approximation de $y(t_i)$. Écrire le schéma d'EULER rétrograde permettant de calculer y_{i+1} à partir de y_i.
3. À partir du schéma obtenu au point précédent, écrire un seul pas de la méthode de NEWTON pour calculer une nouvelle approximation de y_{i+1}. En déduire ainsi un nouveau schéma explicite.

SOLUTION.

1. Le problème (5.2) est un problème du type
   trouver $y\colon I\subset\mathbb{R}^+\to\mathbb{R}$ tel que
$$\begin{cases} y'(t)=f(t,y(t)), & \forall t\in I,\\ y(t_0)=y_0, \end{cases}$$
   Si la fonction $f(t,y)$ est
   1.1. continue par rapport à ses deux variables ;
   1.2. lipschitzienne par rapport à sa deuxième variable, c'est-à-dire qu'il existe une constante positive L (appelée constante de Lipschitz) telle que
$$|f(t,y_1)-f(t,y_2)|\le L|y_1-y_2|,\qquad \forall t\in I,\quad \forall y_1,y_2\in\mathbb{R},$$
   alors la solution $y=y(t)$ du problème de CAUCHY existe, est unique et appartient à $\mathscr{C}^1(I)$.
   Dans notre cas, $f(t,y(t))=-(y(t))^m+\cos(t)$ et $y_0=0$. Puisque m est impair et $\partial_t f(y,y(t))=-m(y(t))^{m-1}\le 0$, la fonction f est décroissante et satisfait donc
$$\big(f(t_1,y(t_1))-f(t_2,y(t_2))\big)(t_1-t_2)\le 0\qquad \forall t_1,t_2\in\mathbb{R}.$$
   Par conséquent, le problème (5.2) possède une solution unique globale.
2. Le problème (5.1) est un problème du type
   trouver $y\colon I\subset\mathbb{R}^+\to\mathbb{R}$ tel que
$$\begin{cases} y'(t)=f(t,y(t)), & \forall t\in I,\\ y(t_0)=y_0. \end{cases}$$
   Le principe des méthodes d'approximation est de subdiviser l'intervalle I en sous-intervalles de longueur Δt et, pour chaque nœud $t_n=t_0+i\Delta t$ ($i\ge 0$), on cherche la valeur inconnue y_i qui approche $y(t_i)$. L'ensemble de valeurs $\{y_0,y_1,\dots\}$ représente la solution numérique. Le schéma de EULER rétrograde établit une relation entre y_i et y_{i+1} et il est donc possible de calculer successivement y_1, y_2,..., en partant de y_0.
   Si nous intégrons l'EDO $y'(t)=f(t,y(t))$ entre t_i et t_{i+1} nous obtenons
$$y(t_{i+1})-y(t_i)=\int_{t_i}^{t_{i+1}} f(t,y(t))\,\mathrm{d}t.$$
   Soit y_i une approximation de $y(t_i)$ et y_{i+1} une approximation de $y(t_{i+1})$. Si on utilise la formule du rectangle à droite, *i.e.*
$$\int_{t_i}^{t_{i+1}} f(t,y(t))\,\mathrm{d}t\approx\Delta t f(t_{i+1},y(t_{i+1}))$$
   on obtient le *schéma d'*EULER *rétrograde* :
$$\begin{cases} y_0=0,\\ y_{i+1}=y_i+\Delta t\big(-(y_{i+1})^m+\cos(t_{i+1})\big). \end{cases}$$
   Ce schéma est implicite car il ne permet pas de calculer y_{i+1} directement à partir de y_i.
3. Il s'agit de trouver y_{i+1} tel que
$$y_{i+1}=y_i+\Delta t\big(-y_{i+1}^m+\cos(t_{i+1})\big).$$
   Pour déterminer y_{i+1} nous devons chercher le zéro de la fonction g définie par
$$g(x)=x+\Delta t x^m-\big(\Delta t\cos(t_{i+1})+y_i\big).$$
   La méthode de NEWTON pour approcher le zéro de g s'écrit
$$x_{k+1}=x_k-\frac{g(x_k)}{g'(x_k)}=x_k-\frac{x_k+\Delta t x_k^m-\big(\Delta t\cos(t_{i+1})+y_i\big)}{1+m\Delta t x_k^{m-1}}.$$
   Le premier pas de la méthode de NEWTON s'écrit donc
$$x_1=x_0-\frac{x_0+\Delta t x_0^m-\big(\Delta t\cos(t_{i+1})+y_i\big)}{1+m\Delta t x_0^{m-1}}.$$
   Choisissons $x_0=y_i$ comme valeur de départ. Nous pouvons utiliser x_1 comme approximation de y_{i+1} et on obtient le schéma
$$\begin{cases} y_0=0,\\ y_{i+1}=-\dfrac{y_i+\Delta t y_i^m-\big(\Delta t\cos(t_{i+1})+y_i\big)}{1+m\Delta t y_i^{m-1}}. \end{cases}$$

Exercice 5.4

Soit le problème de CAUCHY :

$$\begin{cases} u'(t)+10u(t)=0, & \forall t\in\mathbb{R},\\ u(0)=u_0>0. \end{cases} \tag{5.3}$$

1. Montrer qu'il existe une unique solution globale $u\in\mathscr{C}^\infty(\mathbb{R},\mathbb{R})$ que vous préciserez explicitement.
2. Soit le schéma numérique de Cranck-Nicholson défini par la suite $(u_n)_{n\in\mathbb{N}}$ vérifiant

$$\frac{u_{n+1}-u_n}{\Delta t}+5(u_{n+1}+u_n)=0, \qquad \forall n\in\mathbb{N},$$

   pour $\Delta t>0$ fixé.
   Montrer que la suite $(u_n)_{n\in\mathbb{N}}$ est une suite géométrique dont vous préciserez la raison.
3. Montrer que la raison r de la suite vérifie $|r|<1$ pour tout $\Delta t>0$. Ce schéma est-il inconditionnellement A-stable ?
4. Sous quelle condition sur $\Delta t>0$ le schéma génère-t-il une suite positive ?
5. Donner l'expression de u_n en fonction de n.
6. Soit $T>0$ fixé, soit $n^*=n^*(\Delta t)$ tel que $T-\Delta t<n^*\Delta t\le T$. Montrer que

$$\lim_{\Delta t\to 0} u_{n^*}=u_0e^{-10T}.$$

7. Soit $(v_n)_{n\in\mathbb{N}}$ la suite définissant le schéma d'EULER explicite pour l'équation différentielle (5.3). Montrer que

$$\lim_{\Delta t\to 0} v_{n^*}=u_0e^{-10T}.$$

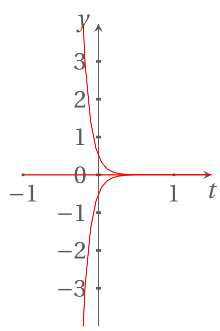
   Montrer que la suite (u_{n^*}) converge plus vite que (v_{n^*}) lorsque $\Delta t\to 0$.

SOLUTION. C'est un problème de CAUCHY du type

$$\begin{cases} u'(t)=f(t,u(t)), & \forall t\in\mathbb{R},\\ u(0)=u_0>0, \end{cases} \tag{5.4}$$

avec $f(t,u(t))=g(u(t))=-10u(t)$.

1. Comme $g\in\mathscr{C}^1(\mathbb{R},\mathbb{R})$, d'après CAUCHY-Lipschitz, il existe $T>0$ et une unique solution $u\in\mathscr{C}^1([-T,T],\mathbb{R})$. Par récurrence, en exploitant l'EDO et la régularité de g, on grimpe en régularité sur u ainsi $u\in\mathscr{C}^\infty([-T,T],\mathbb{R})$. La fonctionne nulle est solution de l'équation différentielle $f(t,0)=0$, par l'unicité de la solution du problème de CAUCHY, pour tout $t\in[-T,T]$, $u(t)>0$ si $u_0>0$ et $u(t)<0$ si $u_0>0$ (autrement dit, deux trajectoires ne peuvent pas se croiser). De plus, u est décroissante si $u_0>0$ et croissante si $u_0<0$. On en déduit par le théorème des extrémités que la solution u admet un prolongement sur $\mathbb{R}$ solution de l'EDO.
   Pour en calculer la solution, on remarque qu'il s'agit d'une EDO à variables séparables. L'unique solution constante est $u(t)\equiv 0$, toutes les autres solutions sont du type $u(t)=Ce^{-10t}$. En prenant en compte la condition initiale on conclut que l'unique solution du problème de CAUCHY est $u(t)=u_0e^{-10t}$ définie pour tout $t\in\mathbb{R}$.



2. Soit le schéma numérique de Cranck-Nicholson défini par la suite $(u_n)_{n\in\mathbb{N}}$ vérifiant

$$\frac{u_{n+1}-u_n}{\Delta t}+5(u_{n+1}+u_n)=0, \qquad \forall n\in\mathbb{N},$$

   pour $\Delta t>0$ fixé. On obtient une formule de récurrence rendue explicite par un calcul élémentaire :

$$u_{n+1}=-5\Delta t\,u_{n+1}-5\Delta t\,u_n+u_n$$

d'où

$$u_{n+1} = \frac{1-5\Delta t}{1+5\Delta t} u_n.$$

Il s'agit d'une suite géométrique de raison

$$r = \frac{1-5\Delta t}{1+5\Delta t}.$$

3. Pour tout $\Delta t > 0$ on a

$$r = \frac{1-5\Delta t}{1+5\Delta t} = 1 - \frac{10\Delta t}{1+5\Delta t}$$

et

$$-1 < 1 - \frac{10\Delta t}{1+5\Delta t} < 1.$$

Ce schéma est donc inconditionnellement A-stable car $|u_{n+1}| = |r^{n+1} u_0| \le |u_0|$.

4. Le schéma génère une suite positive ssi

$$1 - \frac{10\Delta t}{1+5\Delta t} > 0$$

i.e. ssi

$$\Delta t < \frac{1}{5}.$$

5. Par récurrence on obtient

$$u_n = \left(\frac{1-5\Delta t}{1+5\Delta t}\right)^n u_0.$$

6. Soit $T > 0$ fixé et considérons $n^* = n^*(\Delta t)$ tel que $T - \Delta t < n^* \Delta t \le T$. En se rappelant que

$$\lim_{x\to 0} \frac{\ln(1+\alpha x)}{\alpha x} = 1$$

et en observant que

$$\begin{array}{ccccc}
\left(\frac{1-5\Delta t}{1+5\Delta t}\right)^{\frac{T}{\Delta t}-1} & \le & \left(\frac{1-5\Delta t}{1+5\Delta t}\right)^{n^*} & \le & \left(\frac{1-5\Delta t}{1+5\Delta t}\right)^{\frac{T}{\Delta t}} \\
\| & & & & \| \\
e^{(T-\Delta t)\frac{\ln(1-5\Delta t)-\ln(1+5\Delta t)}{\Delta t}} & & & & e^{T\frac{\ln(1-5\Delta t)-\ln(1+5\Delta t)}{\Delta t}} \\
\| & & & & \| \\
e^{(T-\Delta t)\frac{-5\ln(1-5\Delta t)-5\ln(1+5\Delta t)}{5\Delta t}} & & & & e^{T\frac{-5\ln(1-5\Delta t)-5\ln(1+5\Delta t)}{5\Delta t}} \\
\downarrow & & & & \downarrow \\
e^{-10T} & & & & e^{-10T}
\end{array}$$

on conclut que

$$\lim_{\Delta t\to 0} u_{n^*} = u_0 \lim_{\Delta t\to 0}\left(\frac{1-5\Delta t}{1+5\Delta t}\right)^{n^*} = u_0 e^{-10T}.$$

7. La suite définissant le schéma d'EULER explicite pour l'EDO assignée s'écrit

$$\frac{v_{n+1}-v_n}{\Delta t} = f(t_n, u_n) \qquad \Longrightarrow \qquad v_{n+1} = v_n - 10\Delta t\, v_n = (1-10\Delta t) v_n = (1-10\Delta t)^{n+1} v_0.$$

Il s'agit à nouveau d'une suite géométrique de raison

$$r_e = 1 - 10\Delta t$$

qui converge si et seulement si $|r_e| < 1$, *i.e.* si et seulement si $\Delta t < 0{,}2$ (le schéma d'EULER pour cette EDO est conditionnellement stable).

Soit $T > 0$ fixé et considérons $n^* = n^*(\Delta t)$ tel que $T - \Delta t < n^* \Delta t \le T$. Alors

$$\begin{array}{ccccc}
(1-10\Delta t)^{\frac{T}{\Delta t}-1} & \le & (1-10\Delta t)^{n^*} & \le & (1-10\Delta t)^{\frac{T}{\Delta t}} \\
\| & & & & \| \\
e^{(T-\Delta t)\frac{\ln(1-10\Delta t)}{\Delta t}} & & & & e^{T\frac{\ln(1-10\Delta t)}{\Delta t}} \\
\| & & & & \| \\
e^{-10(T-\Delta t)\frac{\ln(1-10\Delta t)}{-10\Delta t}} & & & & e^{-10T\frac{\ln(1-10\Delta t)}{-10\Delta t}} \\
\downarrow & & & & \downarrow \\
e^{-10T} & & & & e^{-10T}
\end{array}$$

d'où

$$\lim_{\Delta t\to 0} v_{n^*} = u_0 \lim_{\Delta t\to 0} (1-10\Delta t)^{\frac{T}{\Delta t}} = u_0 e^{-10T}.$$

De plus, on sait (cf. cours) que la suite $\{u_n\}_{n\in\mathbb{N}}$ converge à l'ordre 2 tandis que la suite $\{v_n\}_{n\in\mathbb{N}}$ converge à l'ordre 1.

Exercice 5.5

Soit le problème de CAUCHY :

$$\begin{cases} u'(t) + u^5(t) = 0, & \forall t \in \mathbb{R}^+, \\ u(0) = u_0 > 0. \end{cases} \tag{5.5}$$

1. Montrer qu'il existe une unique solution globale $u \in \mathscr{C}^\infty(\mathbb{R}^+, \mathbb{R}^+)$.
2. Soit le schéma numérique défini par la suite $(u_n)_{n\in\mathbb{N}}$ suivante

$$\frac{u_{n+1} - u_n}{\Delta t} + u_{n+1} u_n^4 = 0, \qquad \forall n \in \mathbb{N},$$

   pour $\Delta t > 0$ fixé. Expliciter l'expression de u_{n+1} en fonction de u_n.
3. Montrer que la suite $(u_n)_{n\in\mathbb{N}}$ est une suite décroissante et convergente vers 0. Ce schéma est-il inconditionnellement A-stable ?

SOLUTION. C'est un problème de CAUCHY du type

$$\begin{cases} u'(t) = f(t, u(t)), & \forall t \in \mathbb{R}, \\ u(0) = u_0 > 0, \end{cases} \tag{5.6}$$

avec $f(t, u(t)) = g(u(t)) = -u^5(t)$.

1. Comme $g \in \mathscr{C}^1(\mathbb{R}^+, \mathbb{R}^+)$, d'après CAUCHY-Lipschitz, il existe $T > 0$ et une unique solution $u \in \mathscr{C}^1([0, T], \mathbb{R}^+)$. Par récurrence, en exploitant l'EDO et la régularité de g, on grimpe en régularité sur u et $u \in \mathscr{C}^\infty([0, T], \mathbb{R}^+)$.
   La fonctionne nulle est solution de l'équation différentielle $g(0) = 0$. Comme $u_0 > 0$, par l'unicité de la solution du problème de CAUCHY on a $u(t) > 0$ pour tout $t \in [0, T]$ (car deux trajectoires ne peuvent pas se croiser). De plus, u est décroissante, ainsi la solution est bornée ($u(t) \in]0, u_0[$). On en déduit par le théorème des extrémités que la solution u admet un prolongement sur $\mathbb{R}^+$ solution de l'EDO.
2. Pour $\Delta t > 0$ fixé on obtient

$$u_{n+1} = \frac{u_n}{1 + u_n^4 \Delta t}, \qquad \forall n \in \mathbb{N}.$$

3. On étudie la suite

$$\begin{cases} u_0 > 0, \\ u_{n+1} = \frac{u_n}{1+u_n^4 \Delta t}, & \forall n \in \mathbb{N}, \end{cases}$$

   pour $\Delta t > 0$ fixé.
   Par récurrence on montre que si $u_0 > 0$ alors $u_n > 0$ pour tout $n \in \mathbb{N}$. De plus, $\frac{u_{n+1}}{u_n} = \frac{1}{1+u_n^4 \Delta t} < 1$ pour tout $n \in \mathbb{N}$: la suite est monotone décroissante. On a alors une suite décroissante et bornée par zéro, donc elle converge. Soit ℓ la limite de cette suite, alors $\ell \geq 0$ et $\ell = \frac{\ell}{1+\ell^4 \Delta t}$ donc $\ell = 0$. Ce schéma est donc inconditionnellement A-stable.

Exercice 5.6

Soit le problème de CAUCHY :

$$\begin{cases} u'(t) + \sin(u(t)) = 0, & \forall t \in \mathbb{R}, \\ u(0) = u_0 > 0. \end{cases} \tag{5.7}$$

1. Montrer qu'il existe une unique solution globale $u \in \mathscr{C}^\infty(\mathbb{R}, \mathbb{R})$.
2. Écrire le schéma le schéma d'EULER explicite pour ce problème de CAUCHY en explicitant vos notations.
3. Montrer que la suite $(u_n)_{n\in\mathbb{N}}$ construite par ce schéma vérifie

$$|u_{n+1}| \leq |u_n| + \Delta t, \qquad \forall n \in \mathbb{N},$$

   où $\Delta t > 0$ est le pas de la suite.
4. En déduire que

$$|u_n| \leq |u_0| + n\Delta t, \qquad \forall n \in \mathbb{N}.$$

SOLUTION. C'est un problème de CAUCHY du type

$$\begin{cases} u'(t) = f(t, u(t)), & \forall t \in \mathbb{R}, \\ u(0) = u_0 > 0, \end{cases} \tag{5.8}$$

avec $f(t, u(t)) = g(u(t)) = -\sin(u(t))$.

1. Comme $g \in \mathscr{C}^1(\mathbb{R},\mathbb{R})$, d'après CAUCHY-Lipschitz, il existe $T > 0$ et une unique solution $u \in \mathscr{C}^1([-T, T], \mathbb{R})$. Par récurrence, en exploitant l'EDO et la régularité de g, on grimpe en régularité sur u et $u \in \mathscr{C}^\infty([-T, T], \mathbb{R})$.
   Toutes les fonctions constante $u(t) = k\pi$ pour $k \in \mathbb{Z}$ sont solutions de l'équation différentielle car $g(k\pi) = 0$. Pour u_0 donné, soit $\tilde{k} \in \mathbb{Z}$ tel que $u_0 \in [\tilde{k}\pi; (\tilde{k}+1)\pi]$; par l'unicité de la solution du problème de CAUCHY on a $u(t) \in [\tilde{k}\pi; (\tilde{k}+1)\pi]$ pour tout $t \in [-T, T]$ (car deux trajectoires ne peuvent pas se croiser), *i.e.* la solution est bornée. On en déduit par le théorème des extrémités que la solution u admet un prolongement sur $\mathbb{R}$ solution de l'EDO.
2. Soit $\Delta t > 0$ fixé et $t_n = n\Delta t$ pour tout $n \in \mathbb{Z}$. Le schéma d'EULER explicite pour l'EDO donnée construit la suite
$$u_{n+1} = u_n - \Delta t \sin(u_n), \qquad \forall n \in \mathbb{N}.$$
3. Comme $|a + b| \le |a| + |b|$ et comme $|-\sin(x)| \le 1$ pour tout $x \in \mathbb{R}$, on conclut que
$$|u_{n+1}| = |u_n - \Delta t \sin(u_n)| \le |u_n| + |\Delta t \sin(u_n)| \le |u_n| + \Delta t$$
   pour $\Delta t > 0$ fixé.
4. Par récurrence : $|u_{n+1}| \le |u_n| + \Delta t \le |u_{n-1}| + 2\Delta t \le \cdots \le |u_0| + (n+1)\Delta t$.

Exercice 5.7 *Loi de* NEWTON

Considérons une tasse de café à la température de 75°C dans une salle à 25°C. On suppose que la température du café suit la loi de Newton, c'est-à-dire que la vitesse de refroidissement du café est proportionnelle à la différence des températures. En formule cela signifie qu'il existe une constante $K < 0$ telle que la température vérifie l'équation différentielle ordinaire (EDO) du premier ordre.

$$T'(t) = K(T(t) - 25).$$

La condition initiale (CI) est donc simplement

$$T(0) = 75.$$

Pour calculer la température à chaque instant on a besoin de connaître la constante K. Cette valeur peut être déduite en constatant qu'après 5 minutes le café est à 50°C, c'est-à-dire

$$T(5) = 50.$$

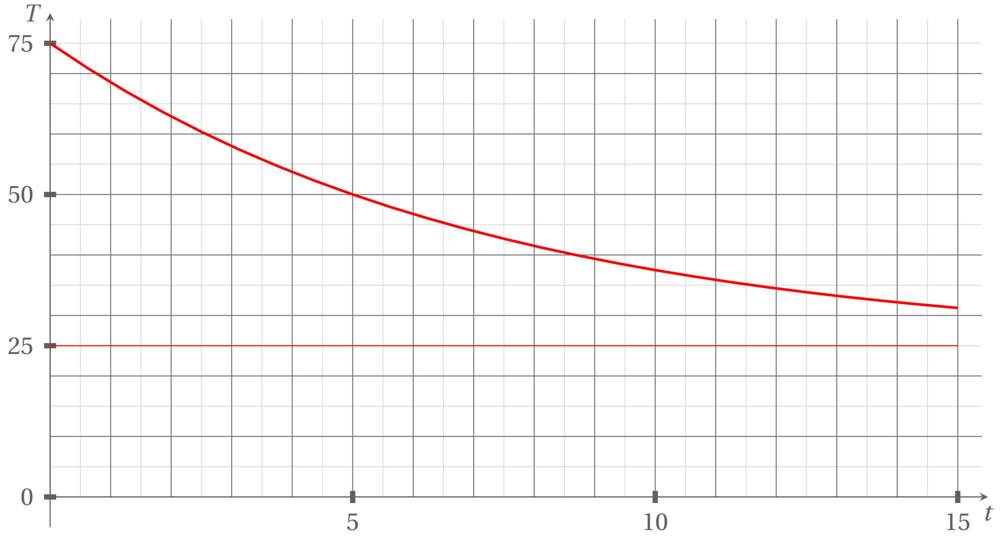
Calculer la solution exacte de ce problème de CAUCHY et la comparer avec la solution approchée obtenue par la méthode d'EULER explicite.

SOLUTION. ***Solution exacte***

1. On commence par calculer toutes les solutions de l'EDO. Étant une équation différentielle du premier ordre, la famille de solutions dépendra d'une constante qu'on fixera en utilisant la CI. Il s'agit d'une EDO à variables séparables donc formellement on a
$$T'(t) = K(T(t) - 25) \implies \frac{T'(t)}{(T(t) - 25)} = K \implies \frac{\mathrm{d}T}{(T - 25)} = K\mathrm{d}t \implies$$
$$\int \frac{1}{(T - 25)}\mathrm{d}T = K\int \mathrm{d}t \implies \ln(T - 25) = Kt + c \implies T - 25 = De^{Kt} \implies T(t) = 25 + De^{Kt}$$
2. La valeur numérique de la constante d'intégration D est obtenue grâce à la CI :
$$75 = T(0) = 25 + De^{K\cdot 0} \implies D = 50 \implies T(t) = 25 + 50e^{Kt}$$
3. Il ne reste qu'à établir la valeur numérique de la constante de refroidissement K grâce à l'«indice» :
$$50 = T(5) = 25 + 50e^{Kt} \implies K = -\frac{\ln(2)}{5} \implies T(t) = 25 + 50e^{-\frac{\ln(2)}{5}t}$$

On peut donc conclure que la température du café évolue selon la fonction

$$T(t) = 25 + 50e^{-\frac{\ln(2)}{5}t}.$$



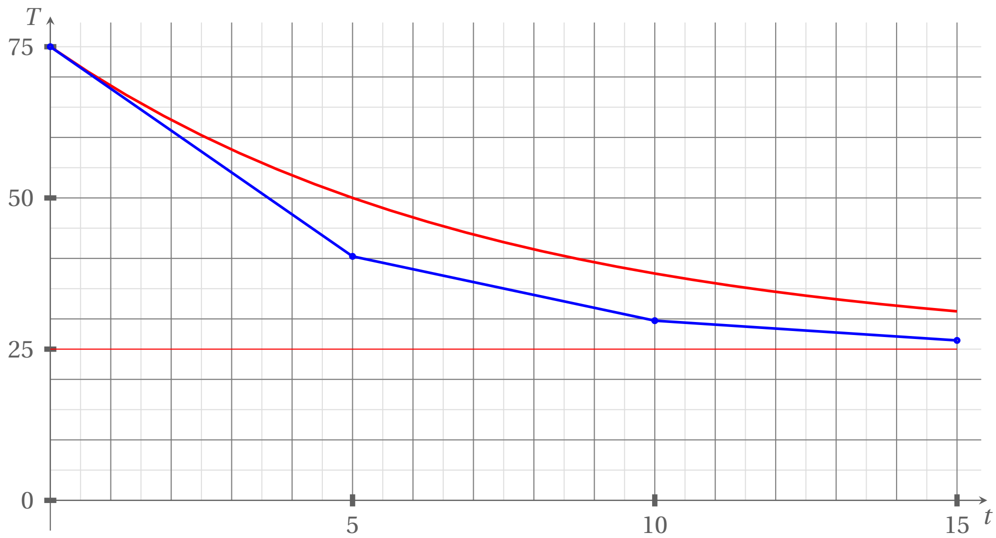
***Solution approchée par la méthode d'*Euler** Supposons de connaître K mais de ne pas vouloir/pouvoir calculer la fonction $T(t)$. Grâce à la méthode d'EULER on peut estimer la température à différentes instantes t_i en faisant une discrétisation temporelle du futur (i.e. on construit une suite de valeurs $\{t_i = 0 + i\Delta t\}_i$) et en construisant une suite de valeurs $\{T_i\}_i$ où chaque T_i est une approximation de $T(t_i)$. Si on utilise la méthode d'EULER, cette suite de température est ainsi construite :

$$\begin{cases} T_{i+1} = T_i - \frac{\ln(2)}{5}\Delta t\,(T_i - 25),\\ T_0 = 75,\end{cases}$$

qu'on peut réécrire comme

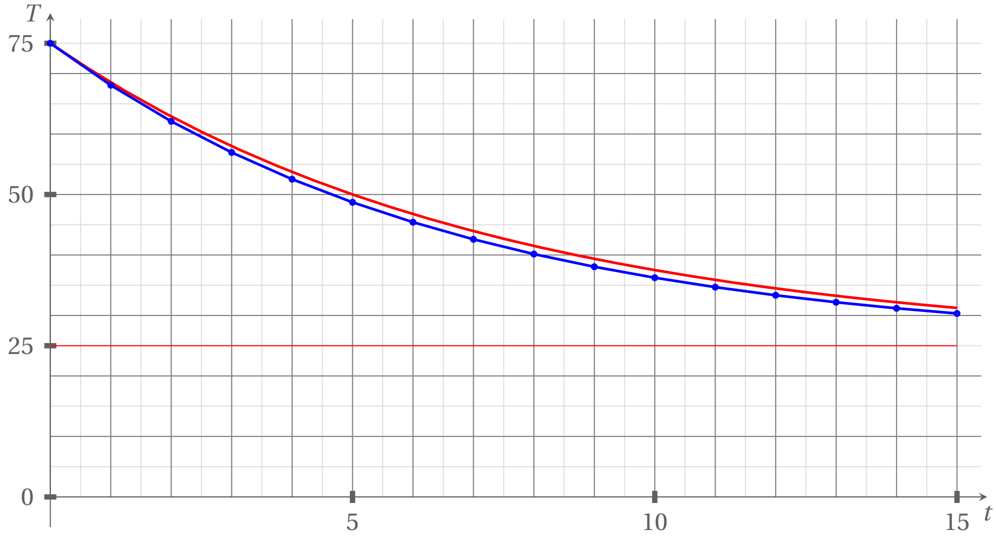
$$\begin{cases} T_{i+1} = (1 - \frac{\ln(2)}{5}\Delta t)T_i + 5\ln(2)\Delta t,\\ T_0 = 75.\end{cases}$$

1. Exemple avec $\Delta t = 5$:



t_i	$T(t_i)$	T_i	$T(t_i) - T_i$
0.000000	75.000000	75.000000	0.000000
5.000000	50.000000	40.342641	9.657359
10.000000	37.500000	29.707933	7.792067
15.000000	31.250000	26.444642	4.805358

2. Exemple avec $\Delta t = 1$:



t_i	$T(t_i)$	T_i	$T(t_i) - T_i$
0.000000	75.000000	75.000000	0.000000
1.000000	68.527528	68.068528	0.459000
2.000000	62.892914	62.097962	0.794952
3.000000	57.987698	56.955093	1.032605
4.000000	53.717459	52.525176	1.192283
5.000000	50.000000	48.709377	1.290623
6.000000	46.763764	45.422559	1.341205
7.000000	43.946457	42.591391	1.355066
8.000000	41.493849	40.152707	1.341142
9.000000	39.358729	38.052095	1.306634
10.000000	37.500000	36.242691	1.257309
11.000000	35.881882	34.684123	1.197759
12.000000	34.473229	33.341618	1.131610
13.000000	33.246924	32.185225	1.061700
14.000000	32.179365	31.189141	0.990224
15.000000	31.250000	30.331144	0.918856

Exercice 5.8 *«Les experts - Toulon»*

La loi de Newton affirme que la vitesse de refroidissement d'un corps est proportionnelle à la différence entre la température du corps et la température externe, autrement dit qu'il existe une constante $K < 0$ telle que la température du corps suit l'équation différentielle

$$\begin{cases} T'(t) = K(T(t) - T_{\text{ext}}), \\ T(0) = T_0. \end{cases}$$

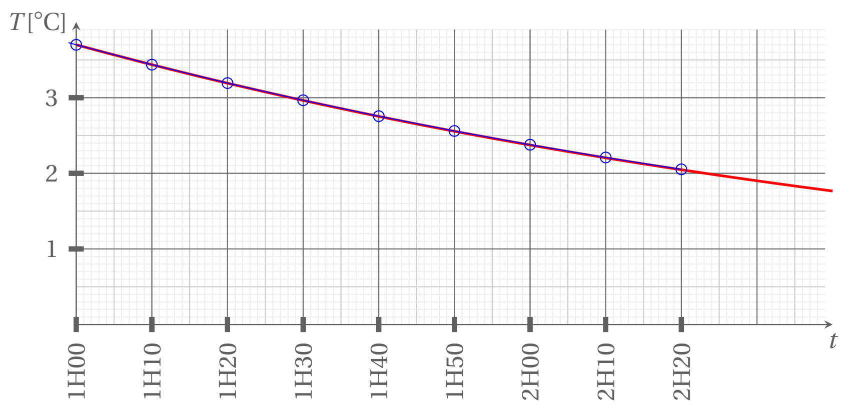
1. Soit Δt le pas temporel. Écrire le schéma d'EULER implicite pour approcher la solution de cette équation différentielle.
2. Soit $T_{\text{ext}} = 0$°C. En déduire une forme du type
$$T_{n+1} = g(\Delta t, n, T_0)$$
avec $g(\Delta t, n, T_0)$ à préciser (autrement dit, l'itéré en t_n ne dépend que de Δt, de n et de T_0). Que peut-on en déduire sur la convergence de la méthode ?
3. *Problème*. Un homicide a été commis. On veut établir l'heure du crime sachant que
   - ▷ pour un corps humaine on peut approcher $K \approx -0.007438118376$ (l'échelle du temps est en minutes et la température en Celsius),
   - ▷ le corps de la victime a été trouvé sur le lieu du crime à $2H20$ du matin,
   - ▷ à l'heure du décès la température du corps était de 37°C,
   - ▷ à l'heure de la découverte la température du corps est de 20°C,
   - ▷ la température externe est $T_{\text{ext}} = 0$°C.

   Approcher l'heure de l'homicide en utilisant le schéma d'EULER implicite avec $\Delta t = 10$ minutes.
4. Pour cette équation différentielle, il est possible de calculer analytiquement ses solutions. Comparer alors la solution exacte avec la solution approchée obtenue au point précédent.

Solution.

1. La méthode d'Euler implicite est une méthode d'intégration numérique d'EDO du premier ordre de la forme $T'(t) = F(t, T(t))$. En choisissant un pas de discrétisation Δt, nous obtenons une suite de valeurs (t_n, T_n) qui peuvent être une excellente approximation de la fonction $T(t)$ avec
$$\begin{cases} t_n = t_0 + n\Delta t, \\ T_{n+1} = T_n + F(t_{n+1}, T_{n+1})\Delta t. \end{cases}$$
La méthode d'Euler implicite pour cette EDO s'écrit donc
$$T_{n+1} = T_n + K\Delta t(T_{n+1} - T_{\text{ext}}).$$
2. Si $T_{\text{ext}} = 0°$C, en procédant par récurrence sur n on obtient
$$T_{n+1} = g(\Delta t, n) = \frac{1}{1-K\Delta t}T_n = \frac{1}{(1-K\Delta t)^{n+1}}T_0,$$
autrement dit, l'itérée en t_n ne dépend que de Δt et de n mais ne dépend pas de T_n. Comme $0 < \frac{1}{1-K\Delta t} < 1$ pour tout $\Delta t > 0$, la suite est positive décroissante ce qui assure que la solution numérique est stable et convergente.
3. On cherche combien de minutes se sont écoulés entre le crime et la découverte du corps, autrement dit on cherche n tel que
$$20 = \frac{1}{(1-K\Delta t)^{n+1}}37 \implies (1-K\Delta t)^{n+1} = \frac{37}{20} \implies n+1 = \log_{(1-K\Delta t)}\frac{37}{20} = \frac{\ln\frac{37}{20}}{\ln(1-K\Delta t)} \implies n \approx 8.$$
Comme $t_n = t_0 + n\Delta t$, si $t_n = 2H20$ alors $t_0 = t_n - n\Delta t = 2H20 - 1H20 = 01H00$.
4. Calcule analytique de toutes les solutions de l'équation différentielle :
   ▷ On cherche d'abord les solutions constantes, i.e. les solutions du type $T(t) \equiv c \in \mathbb{R}$ quelque soit t. On a
   $$0 = K(c - T_{\text{ext}})$$
   d'où l'unique solution constante $T(t) \equiv T_{\text{ext}}$.
   ▷ Soit $T(t) \neq T_{\text{ext}}$ quelque soit t. Puisqu'il s'agit d'une EDO à variables séparables on peut calculer la solution comme suit :
   $$T'(t) = K(T(t) - T_{\text{ext}}) \implies \frac{T'(t)}{T(t) - T_{\text{ext}}} = K \implies \frac{\mathrm{d}T}{T - T_{\text{ext}}} = K\mathrm{d}t \implies$$
   $$\int \frac{1}{T - T_{\text{ext}}}\mathrm{d}T = K\int \mathrm{d}t \implies \ln(T - T_{\text{ext}}) = Kt + c \implies T - T_{\text{ext}} = De^{Kt} \implies T(t) = T_{\text{ext}} + De^{Kt}.$$

   La valeur numérique de la constante d'intégration D est obtenue grâce à la CI :
   $$T_0 = T(0) = De^{K\cdot 0} \implies D = -T_0 \implies T(t) = T_0e^{Kt}$$
   Ici $T_0 = 37°$C donc la température du cadavre suit la loi
   $$T(t) = 37e^{Kt}.$$
   Pour déterminer l'heure du meurtre il faut alors résoudre l'équation
   $$20 = 37e^{Kt}$$
   d'où $t = \frac{1}{K}\ln\frac{20}{37} \approx 82{,}70715903$ minutes, c'est-à-dire 83 minutes avant 2H20 : le crime a été commit à 00H57.



Exercice 5.9

Un modèle pour la diffusion d'une épidémie se base sur l'hypothèse que sa vitesse de propagation est proportionnelle au nombre d'individus infectés et au nombre d'individus sains.
Si on note $I(t) \geq 0$ le nombre d'individus infectés à l'instant $t \geq 0$ et $A > 0$ le nombre d'individus total, il existe une constante $k \in \mathbb{R}^+$ telle que $I'(t) = kI(t)(A - I(t))$.

1. Montrer qu'il existe $T > 0$ et une unique solution $I \in \mathscr{C}^\infty([0,T])$ au problème de CAUCHY :
$$\begin{cases} I'(t) = kI(t)(A - I(t)), \\ I(0) = I_0 > 0. \end{cases}$$
2. Montrer que si $0 < I_0 < A$ alors $0 < I(t) < A$ pour tout $t > 0$.
3. Montrer que si $0 < I_0 < A$ alors $I(t)$ est croissante sur $\mathbb{R}^+$.
4. Soit $0 < I_0 < A$. On considère le schéma semi-implicite
$$\frac{I_{n+1} - I_n}{\Delta t} = kI_n(A - I_{n+1}).$$
Montrer que ce schéma est inconditionnellement A-stable.

SOLUTION. C'est un problème de CAUCHY du type

$$\begin{cases} I'(t) = f(t, I(t)), & \forall t \in \mathbb{R}^+, \\ I(0) = I_0 > 0, \end{cases} \tag{5.9}$$

avec $f(t, I(t)) = g(I(t)) = kI(t)(A - I(t))$.

1. Comme $g \in \mathscr{C}^1(\mathbb{R}, \mathbb{R})$, d'après CAUCHY-Lipschitz, il existe $T > 0$ et une unique $I \in \mathscr{C}^1([0,T], \mathbb{R})$ solution du problème de CAUCHY. Par récurrence, en exploitant l'EDO et la régularité de g, on grimpe en régularité sur I et $I \in \mathscr{C}^\infty([0,T], \mathbb{R})$.
2. Puisque la fonction nulle et la fonction constante $I(t) = A$ sont solutions de l'équation différentielle, si $0 < I_0 < A$ alors $0 < I(t) < A$ pour tout $t \in [0,T]$ (car, par l'unicité de la solution du problème de CAUCHY, deux trajectoires ne peuvent pas se croiser).
3. Puisque $I'(t) = kI(t)(A - I(t))$, si $0 < I_0 < A$ alors I est croissante pour tout $t \in [0,T]$. On en déduit par le théorème des extrémités que la solution I admet un prolongement sur $\mathbb{R}^+$ solution de l'EDO et que I est croissante pour tout $t \in \mathbb{R}^+$.
4. Soit $0 < I_0 < A$. On considère le schéma semi-implicite
$$\frac{I_{n+1} - I_n}{\Delta t} = kI_n(A - I_{n+1})$$
pour $\Delta t > 0$ fixé. On obtient une formule de récurrence rendue explicite par un calcul élémentaire :
$$I_{n+1} = \frac{1 + kA\Delta t}{1 + kI_n\Delta t} I_n.$$
Si $0 < I_0 < A$ alors
▷ $I_n > 0$ quelque soit n ;
▷ I_n est majorée par A car
$$I_{n+1} \leq A \quad \Longleftrightarrow \quad (1 + kA\Delta t)I_n \leq (1 + kI_n\Delta t)A \quad \Longleftrightarrow \quad I_n \leq A$$
donc par récurrence $I_{n+1} \leq A$ quelque soit n ;
▷ I_n est une suite monotone croissante (encore par récurrence on montre que $|I_{n+1}| \geq |I_n| \geq \cdots \geq |I_0|$) ;
donc ce schéma est inconditionnellement A-stable.

Calcul analytique de toutes les solutions :

On a déjà observé qu'il y a deux solutions constantes de l'EDO : la fonction $I(t) \equiv 0$ et la fonction $I(t) \equiv A$.

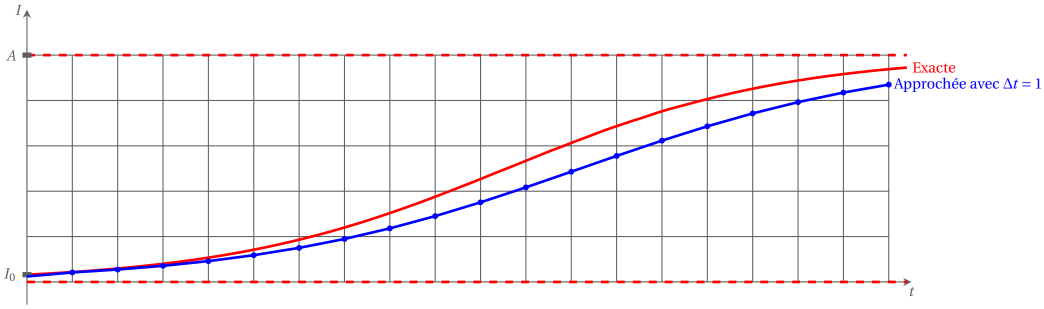
Pour chercher toutes les solutions non constantes on remarque qu'il s'agit d'une EDO à variables séparables donc on a

$$I(t) = \frac{A}{De^{-Akt} + 1}$$

La valeur numérique de la constante d'intégration D est obtenue grâce à la CI :

$$D = \frac{A - I_0}{I_0}$$

Exemple avec $A = 5000$, $I_0 = 160$, $k = \frac{\ln(363/38)}{35000}$ et $\Delta t = 1$:



Exercice 5.10

Considérons une population de bactéries. Soit $p(t)$ le nombre d'individus (≥ 0) à l'instant $t \geq 0$. Un modèle qui décrit l'évolution de cette population est l'«équation de la logistique» : soit k et h deux constantes positives, alors $p(t)$ vérifie l'équation différentielle ordinaire (EDO) du premier ordre

$$p'(t) = kp(t) - hp^2(t).$$

On veut calculer $p(t)$ à partir d'un nombre initiale d'individus donné

$$p(0) = p_0 \geq 0.$$

Solution.

Solution exacte

1. On commence par calculer toutes les solutions de l'EDO. Étant une équation différentielle du premier ordre, la famille de solutions dépendra d'une constante qu'on fixera en utilisant la CI. Il s'agit d'une EDO à variables séparables.
   On cherche d'abord les solutions constantes, c'est-à-dire les solutions du type $p(t) \equiv c$ pour tout $t \in \mathbb{R}^+$:

   $$0 = kc - hc^2.$$

   On a donc deux solutions constantes :

   $$p(t) \equiv 0 \qquad \text{et} \qquad p(t) \equiv \frac{k}{h}.$$

   Étant donné que deux solutions d'une EDO ne s'intersectent jamais, dorénavant on supposera $p(t) \neq 0$ et $p(t) \neq \frac{k}{h}$ pour tout $t \in \mathbb{R}^+$, ainsi

   $$\frac{p'(t)}{kp(t) - hp^2(t)} = 1.$$

   Formellement on a

   $$\begin{aligned}
   \frac{\mathrm{d}p}{kp - hp^2} &= 1\,\mathrm{d}t & \implies && \int \frac{1}{p(k-hp)}\,\mathrm{d}p &= \int 1\,\mathrm{d}t & \implies\\
   \frac{1}{k}\int \frac{1}{p}\,\mathrm{d}p - \frac{1}{k}\int \frac{-h}{k-hp}\,\mathrm{d}p &= \int 1\,\mathrm{d}t & \implies && \frac{1}{k}\ln(p) - \frac{1}{k}\ln(k-hp) &= t + c & \implies\\
   \ln\left(\frac{p}{k-hp}\right) &= kt + kc & \implies && \frac{p}{k-hp} &= De^{kt} & \implies\\
   p(t) &= \frac{k}{\frac{1}{De^{kt}} + h}.
   \end{aligned}$$

2. La valeur numérique de la constante d'intégration D est obtenue grâce à la CI :

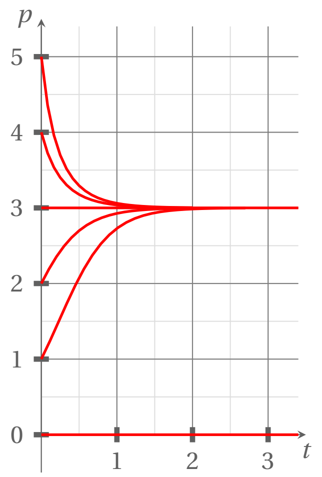
$$p_0 = p(0) = \frac{kD}{1+hDe^{0k}} \qquad \Longrightarrow \qquad D = \frac{p_0}{k-hp_0}.$$

On peut donc conclure que la population évolue selon la fonction

$$p(t) = \begin{cases} 0 & \text{si } p_0 = 0, \\ \frac{k}{h} & \text{si } p_0 = \frac{k}{h}, \\ \frac{k}{\frac{k-hp_0}{p_0 e^{kt}}+h} & \text{sinon.} \end{cases}$$

Une simple étude de la fonction p montre que

- ▷ si $p_0 \in]0;k/h[$ alors $p'(t) > 0$ et $\lim_{t\to+\infty} p(t) = k/h$,
- ▷ si $p_0 \in]k/h;+\infty[$ alors $p'(t) < 0$ et $\lim_{t\to+\infty} p(t) = k/h$.



Exemple avec $k = 3$, $h = 1$ et différentes valeurs de p_0.

Solution approchée Supposons de ne pas vouloir/pouvoir calculer la fonction $p(t)$. Grâce à la méthode d'EULER on peut estimer le nombre d'indivus à différentes instantes t_i en faisant une discrétisation temporelle du futur (i.e. on construit une suite de valeurs $\{t_i = 0 + i\Delta t\}_i$) et en construisant une suite de valeurs $\{p_i\}_i$ où chaque p_i est une approximation de $p(t_i)$. Si on utilise la méthode d'EULER, cette suite est ainsi construite :

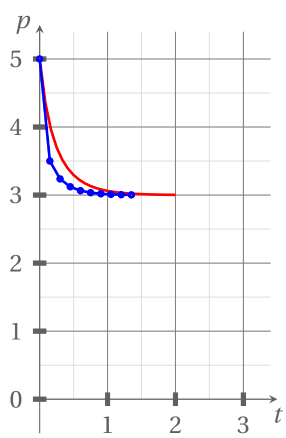
$$\begin{cases} p_{i+1} = p_i + \Delta t \; p_i(k - hp_i), \\ p_0 \text{ donné,} \end{cases}$$

qu'on peut réécrire comme

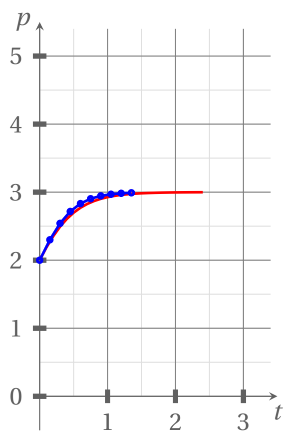
$$\begin{cases} p_{i+1} = (1 + k\Delta t - h\Delta t p_i)p_i, \\ p_0 \text{ donné.} \end{cases}$$

On veut appliquer cette méthode au cas de la figure précédente, i.e. avec $k = 3$, $h = 1$ et les valeurs initiales $p_0 = 5$ et $p_0 = 2$. Si on choisit comme pas temporelle $\Delta t = 0{,}15$, on obtient les figures suivantes :

t_i	$p(t_i)$	p_i	$p(t_i) - p_i$
0.00	5.000000	5.000000	0.000000
0.15	4.027123	3.500000	0.527123
0.30	3.582637	3.237500	0.345137
0.45	3.347079	3.122164	0.224915
0.60	3.212403	3.064952	0.147451
0.75	3.132046	3.035091	0.096956
0.90	3.082874	3.019115	0.063759
1.05	3.052319	3.010459	0.041861
1.20	3.033151	3.005736	0.027415
1.35	3.021054	3.003150	0.017904
1.50	3.013390	3.001731	0.011659
1.65	3.008524	3.000952	0.007573
1.80	3.005430	3.000523	0.004907



t_i	$p(t_i)$	p_i	$p(t_i)-p_i$
0.00	2.000000	2.000000	0.000000
0.15	2.274771	2.300000	−0.025229
0.30	2.493175	2.541500	−0.048325
0.45	2.655760	2.716292	−0.060532
0.60	2.770980	2.831887	−0.060907
0.75	2.849816	2.903298	−0.053483
0.90	2.902469	2.945411	−0.042942
1.05	2.937070	2.969529	−0.032459
1.20	2.959567	2.983102	−0.023535
1.35	2.974092	2.990663	−0.016571
1.50	2.983429	2.994852	−0.011423
1.65	2.989412	2.997164	−0.007752
1.80	2.993240	2.998439	−0.005199



Exercice 5.11 *Méthode de Taylor*

La méthode de Taylor est basé sur la relation

$$y(x+h) \simeq y(x) + y'(x)h + \frac{1}{2!}y''(x)h^2 + \frac{1}{3!}y'''(x)h^3 + \cdots + \frac{1}{m!}y^{(m)}(x)h^m$$

Cette relation prédit $y(x+h)$ à partir de $y(x)$, ainsi elle permet d'écrire une formule d'intégration numérique. Le dernier terme indique l'ordre de la méthode et l'erreur de troncature, due aux termes omis, est

$$E = \frac{1}{(m+1)!}y^{(m+1)}(\xi)h^{m+1} \qquad \text{pour } x<\xi<x+h,$$

que l'on peut approcher par

$$E \simeq \frac{h^m}{(m+1)!}\left(y^{(m)}(x+h) - y^{(m)}(x)\right).$$

Considérons le problème de Cauchy

$$\begin{cases} y'(x)+4y(x)=x^2,\\ y(0)=1. \end{cases}$$

Estimer $y(0.1)$ par la méthode de Taylor d'ordre 4 avec un seul pas d'intégration.

Solution. Le développement de Taylor en 0 jusqu'à l'ordre 4 est

$$y(h) \simeq y(0) + y'(0)h + \frac{1}{2!}y''(0)h^2 + \frac{1}{3!}y'''(0)h^3 + \frac{1}{4!}y^{IV}(0)h^4.$$

En dérivant l'EDO on trouve

$$\begin{aligned} & & y(0)&=1,\\ y'(x)&=-4y(x)+x^2, & y'(0)&=-4,\\ y''(x)&=-4y'(x)+2x, & y''(0)&=16,\\ y'''(x)&=-4y''(x)+2, & y'''(0)&=-62,\\ y^{IV}(x)&=-4y'''(x), & y^{IV}(0)&=248. \end{aligned}$$

donc, pour $x=0$ et $h=0.1$, on obtient

$$y(0.1) \simeq 1 + \frac{-4}{10} + \frac{16}{200} + \frac{-62}{6000} + \frac{248}{240000} = 0.6707$$

et comme

$$y^{IV}(x+h) = -4y'''(x) = -4\left(-4y''(x)+2\right) = \left(-4\left(-4y'(x)+2x\right)+2\right) = \left(-4\left(-4\left(-4y(x)+x^2\right)+2x\right)+2\right)$$

alors $y^{IV}(0.1) \simeq \left(-4\left(-4\left(-4\times 0.6707+(0.1)^2\right)+0.2\right)+2\right) = 166.259$ et on obtient l'estimation de l'erreur

$$E \simeq \frac{248}{960000}\left(y^{IV}(0.1)-y^{IV}(0)\right) = \frac{248}{960000}(166.259-248) = -0.000068.$$

A. Rappels d'analyse et d'algèbre linéaire

A.1. Suites numériques

Suites

Une suite numérique est une application de $\mathbb{N}$ dans $\mathbb{R}$. La notation traditionnelle est $(x_n)_{n\in\mathbb{N}}$.

Limite d'une suite.

On dit qu'une suite $(x_n)_{n\in\mathbb{N}}$ converge vers $x \in \mathbb{R}$ (on écrit $\lim_n x_n = x$) si

$$\forall \varepsilon > 0 \quad \exists N \in \mathbb{N} \quad \forall n \in \mathbb{N} \quad n > N \implies |x_n - x| < \varepsilon.$$

On dit qu'une suite $(x_n)_{n\in\mathbb{N}}$ diverge vers plus l'infini (on écrit $\lim_n x_n = +\infty$) si

$$\forall M \in \mathbb{R} \quad \exists N \in \mathbb{N} \quad \forall n \in \mathbb{N} \quad n > N \implies x_n > M.$$

On dit qu'une suite $(x_n)_{n\in\mathbb{N}}$ diverge vers moins l'infini (on écrit $\lim_n x_n = -\infty$)

$$\forall m \in \mathbb{R} \quad \exists N \in \mathbb{N} \quad \forall n \in \mathbb{N} \quad n > N \implies x_n < m.$$

Remarque

La suppression d'un nombre fini de termes ne modifie pas la nature de la suite, ni sa limite éventuelle.

Suite convergente

Soit $(x_n)_{n\in\mathbb{N}}$ une suite. On dit que $(x_n)_{n\in\mathbb{N}}$ est convergente s'il existe $x \in \mathbb{R}$ tel que $\lim_n x_n = x$.

Suite bornée, suite monotone

- ▷ Une suite $(x_n)_{n\in\mathbb{N}}$ est croissante lorsque pour tout n, $x_n \le x_{n+1}$.
- ▷ Une suite $(x_n)_{n\in\mathbb{N}}$ est décroissante lorsque pour tout n, $x_n \ge x_{n+1}$.
- ▷ Une suite $(x_n)_{n\in\mathbb{N}}$ est majorée s'il existe $a \in \mathbb{R}$ tel que, pour tout n, $x_n \le a$. On dit que a est un majorant de la suite.
- ▷ Une suite $(x_n)_{n\in\mathbb{N}}$ est minorée s'il existe $a \in \mathbb{R}$ tel que, pour tout n, $x_n \ge a$. On dit que a est un minorant de la suite.
- ▷ Une suite $(x_n)_{n\in\mathbb{N}}$ est dite bornée si elle est à la fois majorée et minorée, c'est-à-dire s'il existe $M \in \mathbb{R}$ tel que $|x_n| \le M$ pour tout n.

Théorème (condition nécessaire)

Tout suite convergente est bornée.

Théorème de la convergence monotone (conditions suffisantes)

- ▷ Une suite $(x_n)_{n\in\mathbb{N}}$ qui est croissante et majorée est convergente et $\lim_n x_n = \sup_n x_n$.
- ▷ Une suite $(x_n)_{n\in\mathbb{N}}$ qui est croissante et non majorée diverge vers $+\infty$.
- ▷ Une suite $(x_n)_{n\in\mathbb{N}}$ qui est décroissante et minorée est convergente et $\lim_n x_n = \inf_n x_n$.
- ▷ Une suite $(x_n)_{n\in\mathbb{N}}$ qui est décroissante et non minorée diverge vers $-\infty$.

Théorème d'encadrement (condition suffisante)

Soit $(x_n)_{n\in\mathbb{N}}$ une suite. S'il existe deux suites convergentes $(u_n)_{n\in\mathbb{N}}$ et $(v_n)_{n\in\mathbb{N}}$ ayant une même limite $x\in\mathbb{R}$ et satisfaisant

$$\exists N\in\mathbb{N}\quad \forall n\in\mathbb{N}\quad n>N \implies u_n\le x_n\le v_n,$$

alors $\lim_n x_n = x$.

Valeur d'adhérence

On dit que $\ell\in\mathbb{R}$ est une valeur d'adhérence de la suite $(x_n)_{n\in\mathbb{N}^*}$ si et seulement si il existe une sous-suite extraite qui converge vers ℓ.

Théorème

Soit $(x_n)_{n\in\mathbb{N}}$ une suite. Si (x_n) converge vers ℓ, toute sous-suite converge aussi vers ℓ.

Si une suite extraite de $(x_n)_{n\in\mathbb{N}}$ diverge, ou si deux suites extraites de $(x_n)_{n\in\mathbb{N}}$ ont des limites différentes, alors $(x_n)_{n\in\mathbb{N}}$ diverge.

Si deux suites extraites de $(x_n)_{n\in\mathbb{N}}$ convergent vers la même limite ℓ et si x_n est un terme d'une de ces suites extraites, alors $(x_n)_{n\in\mathbb{N}}$ converge aussi vers ℓ. Par exemple, si (x_{2n}) et (x_{2n+1}) convergent vers ℓ, alors (x_n) converge vers ℓ.

Limites fondamentales

$$\lim_n \frac{P(n)}{Q(n)} = \begin{cases} 0 & \text{si } a<b,\\ \frac{p_a}{q_b} & \text{si } a=b,\\ \infty & \text{si } a>b,\end{cases} \qquad \text{avec} \qquad \begin{aligned} P(n) &= p_0+p_1n+\dots+p_an^a,\\ Q(n) &= q_0+q_1n+\dots+q_bn^b.\end{aligned}$$

$$\begin{array}{llll}
\lim_n n\sin\left(\frac{1}{n}\right)=1 & (n \text{ en radiant}) & \lim_n n^2\left(1-\cos\left(\frac{1}{n}\right)\right)=\frac{1}{2} & (n \text{ en radiant})\\
\lim_n \left(1+\frac{\alpha}{n}\right)^n = e^\alpha & (\alpha\in\mathbb{R}) & \lim_n n\left(\ln\left(1+\frac{1}{n}\right)\right)=1 & \\
\lim_n n\left(a^{1/n}-1\right)=\ln a & (a>0) & \lim_n n\left(\left(1+\frac{1}{n}\right)^\alpha-1\right)=\alpha & (\alpha\in\mathbb{R})
\end{array}$$

Critère du rapport (ou de D'ALEMBERT)

Soit $(u_n)_{n\in N}$ une suite telle que $u_n>0$ pour tout $n\in\mathbb{N}$ et $\frac{u_{n+1}}{u_n}\to\ell$; alors

- ▷ si $0\le\ell<1$ alors $u_n\to 0$,
- ▷ si $1<\ell\le+\infty$ alors $u_n\to+\infty$
- ▷ si $\ell=1$ on ne peut pas conclure.

Critère de la racine (ou de CAUCHY)

Soit $(u_n)_{n\in N}$ une suite telle que $u_n>0$ pour tout $n\in\mathbb{N}$ et $\sqrt[n]{u_n}\to\ell$; alors

- ▷ si $0\le\ell<1$ alors $u_n\to 0$,
- ▷ si $1<\ell\le+\infty$ alors $u_n\to+\infty$
- ▷ si $\ell=1$ on ne peut pas conclure.

Suite arithmétique

Une suite $(v_n)_{n\in\mathbb{N}}$ est arithmétique de raison $r\in\mathbb{R}$ si

$$v_{n+1}=v_n+r \qquad \forall n\in\mathbb{N}$$

et donc $v_n=v_0+nr$ et l'on a

$$\sum_{k=0}^{n-1} v_k = n\frac{v_0+v_{n-1}}{2}.$$

Suite et série géométriques

Une suite $(v_n)_{n\in\mathbb{N}}$ est géométrique de raison $q \in \mathbb{R}$ si

$$v_{n+1} = q v_n \qquad \forall n \in \mathbb{N}$$

et donc $v_n = v_0 q^n$.

- ▷ Si $q < -1$, la suite (v_n) diverge et ne possède pas de limite.
- ▷ Si $q = -1$, la suite (v_n) diverge et possède deux valeurs d'adhérence 1 et -1.
- ▷ Si $|q| < 1$, la suite (v_n) converge vers 0.
- ▷ Si $q = 1$, la suite (v_n) est constante et converge vers 1.
- ▷ Si $q > 1$, la suite (v_n) est divergente mais possède une limite égale à $+\infty$.

Considérons la suite $s_{n-1} := \sum_{k=0}^{n-1} v_k = v_0 \sum_{k=0}^{n-1} q^k = \begin{cases} v_0 \frac{1-q^n}{1-q} & \text{si } q \neq 1 \\ n v_0 & \text{si } q = 1 \end{cases}$

- ▷ Si $q < -1$, la suite (s_n) est divergente et ne possède pas de limite.
- ▷ Si $q = -1$, la suite (s_n) est divergente et possède deux valeurs d'adhérence 0 et 1.
- ▷ Si $|q| < 1$, la suite (s_n) converge vers $\frac{1}{1-q}$.
- ▷ Si $q = 1$, la suite (s_n) est divergente mais possède une limite égale à $+\infty$.
- ▷ Si $q > 1$, la suite (s_n) est divergente mais possède une limite égale à $+\infty$.

A.2. Primitives et intégrales

Primitive

Soit I un intervalle de $\mathbb{R}$. Une fonction $f : I \to \mathbb{R}$ est *intégrable* s'il existe une fonction dérivable $F : I \to \mathbb{R}$ telle que pour tout $x \in I$, $F'(x) = f(x)$. Une telle fonction F est une *primitive* (ou *intégrale indéfinie*) de f.

Existence des primitives

Soit I un intervalle de $\mathbb{R}$ et $f : I \to \mathbb{R}$ une fonction continue. Alors f est intégrable.

Propriété

- ▷ Si F est une primitive de f alors, pour tout réel c, la fonction $F + c$ est aussi une primitive de f.
- ▷ Toute primitive de f est nécessairement de la forme $F + c$ pour une certaine constante c.

Notation

L'ensemble des primitives d'une fonction f est noté $\int f$ ou encore $\int f(x)\,\mathrm{d}x$.

Remarque

Si $a \in I$ alors $F(x) = \int_a^x f(t)\,\mathrm{d}t$ est l'unique primitive qui s'annule en a.

Linéarité

Soit I un intervalle de $\mathbb{R}$, f et $g : I \to \mathbb{R}$ deux fonctions intégrables et $k \in \mathbb{R}$. Alors

$$\int (f(x) + g(x))\,\mathrm{d}x = \int f(x)\,\mathrm{d}x + \int g(x)\,\mathrm{d}x \qquad \text{et} \qquad \int k f(x)\,\mathrm{d}x = k \int f(x)\,\mathrm{d}x.$$

Intégration directe

Dans le tableau qui suit on sous-entend que l'intégration est réalisée sur un intervalle contenu dans l'ensemble de définition de la fonction à intégrer.

$$\int x^n\,\mathrm{d}x = \frac{x^{n+1}}{n+1} + c \qquad \Longrightarrow \qquad \int [f(x)]^n f'(x)\,\mathrm{d}x = \frac{[f(x)]^{n+1}}{n+1} + c \qquad \text{pour } n \neq -1$$

$$
\begin{aligned}
&\int \tfrac{1}{x}\,\mathrm{d}x = \ln(x)+c &&\Longrightarrow \int \tfrac{f'(x)}{f(x)}\,\mathrm{d}x = \ln(|f(x)|)+c\\
&\int e^x\,\mathrm{d}x = e^x+c &&\Longrightarrow \int e^{f(x)}f'(x)\,\mathrm{d}x = e^{f(x)}+c\\
&\int a^x\,\mathrm{d}x = (\log_a e)a^x+c &&\Longrightarrow \int a^{f(x)}f'(x)\,\mathrm{d}x = (\log_a e)a^{f(x)}+c\\
&\int \sin(x)\,\mathrm{d}x = -\cos(x)+c &&\Longrightarrow \int \sin(f(x))f'(x)\,\mathrm{d}x = -\cos(f(x))+c\\
&\int \cos(x)\,\mathrm{d}x = \sin(x)+c &&\Longrightarrow \int \cos(f(x))f'(x)\,\mathrm{d}x = \sin(f(x))+c\\
&\int \tfrac{1}{\cos^2(x)}\,\mathrm{d}x = \tan(x)+c &&\Longrightarrow \int \tfrac{f'(x)}{\cos^2(f(x))}\,\mathrm{d}x = \tan(f(x))+c\\
&\int \tfrac{1}{\sqrt{1-x^2}}\,\mathrm{d}x = \arcsin(x)+c = -\arccos(x)+c &&\Longrightarrow \int \tfrac{f'(x)}{\sqrt{1-(f(x))^2}}\,\mathrm{d}x = \arcsin(f(x))+c = -\arccos(f(x)+c\\
&\int \tfrac{1}{1+x^2}\,\mathrm{d}x = \arctan(x)+c &&\Longrightarrow \int \tfrac{f'(x)}{1+(f(x))^2}\,\mathrm{d}x = \arctan(f(x))+c\\
&\int \cosh(x)\,\mathrm{d}x = \sinh(x)+c &&\Longrightarrow \int \cosh(f(x))f'(x)\,\mathrm{d}x = \sinh(f(x))+c\\
&\int \sinh(x)\,\mathrm{d}x = \cosh(x)+c &&\Longrightarrow \int \sinh(f(x))f'(x)\,\mathrm{d}x = \cosh(f(x))+c\\
&\int \tfrac{1}{\cosh^2(x)}\,\mathrm{d}x = \tanh(x)+c &&\Longrightarrow \int \tfrac{f'(x)}{\cosh^2(f(x))}\,\mathrm{d}x = \tanh(f(x))+c
\end{aligned}
$$

Intégration par changement de variable

Soit F une primitive de f et g une fonction dérivable. Alors la fonction $f(g(x))g'(x)$ est intégrable et l'on a

$$\int f(g(x))g'(x)\,\mathrm{d}x = F(g(x))+c.$$

Autrement dit, en posant $u=g(x)$ on obtient $\frac{\mathrm{d}u}{\mathrm{d}x}=g'(x)$, soit encore $\mathrm{d}u = g'(x)\,\mathrm{d}x$ et donc

$$\int f(g(x))g'(x)\,\mathrm{d}x = \int f(u)\,\mathrm{d}u = F(u)+c.$$

Intégration par parties

Soit f et g deux fonctions dérivables. Alors

$$\int f(x)g'(x)\,\mathrm{d}x = f(x)g(x) - \int f'(x)g(x)\,\mathrm{d}x.$$

Exemple

Calculer une primitive de $\frac{\ln x}{x}$ avec les trois méthodes décrites ci-dessus.

Intégration directe :

comme $\int \frac{\ln x}{x}\,\mathrm{d}x = \int f(x)f'(x)\,\mathrm{d}x$ avec $f(x)=\ln(x)$ et comme $\int f(x)f'(x)\,\mathrm{d}x = \frac{[f(x)]^2}{2}$ on conclut que $\int \frac{\ln x}{x}\,\mathrm{d}x = \frac{\ln^2(x)}{2}+c$.

Intégration par changement de variable :

on pose $u=\ln(x)$ donc $\mathrm{d}u = \frac{\mathrm{d}x}{x}$ et $\int \frac{\ln x}{x}\,\mathrm{d}x = \int u\,\mathrm{d}u = \frac{u^2}{2}+c = \frac{\ln^2(x)}{2}+c$.

Intégration par parties :

si on pose $g(x)=\ln(x)$ et $f'(x)=\frac{1}{x}$ alors $g'(x)=\frac{1}{x}$ et $f(x)=\ln(x)$ donc $\int \frac{\ln x}{x}\,\mathrm{d}x = \ln^2(x) - \int \frac{\ln x}{x}\,\mathrm{d}x$, i.e. $2\int \frac{\ln x}{x}\,\mathrm{d}x = \ln^2 x + c$ et finalement $\int \frac{\ln x}{x}\,\mathrm{d}x = \frac{\ln^2(x)}{2}+k$.

Fonction rationnelle

Soit $N(x)$ et $D(x)$ deux polynômes de degré respectivement ν et δ. Toute fonction du type $P(x) = \frac{N(x)}{D(x)}$ est dite *fonction rationnelle*.

▷ Si $\nu \geq \delta$ on dit que P est une *fonction rationnelle impropre*.

▷ Si $\nu < \delta$ on dit que P est une *fonction rationnelle propre.*

Propriété

Soit N et D deux polynômes de degré respectivement ν et δ et $P(x) = \frac{N(x)}{D(x)}$ une fonction rationnelle impropre (*i.e.* $\nu \ge \delta$). Alors, en effectuant la division euclidienne de N par D, on peut réécrire P comme

$$P(x) = Q(x) + \frac{R(x)}{D(x)}$$

où Q est un polynôme de degré $\nu - \delta$ et R un polynôme de degré au plus $\delta - 1$, ainsi $\frac{R(x)}{D(x)}$ est une fonction rationnelle propre.

On en déduit que

$$\int P(x)\,\mathrm{d}x = \int \frac{N(x)}{D(x)}\,\mathrm{d}x = \int Q(x)\,\mathrm{d}x + \int \frac{R(x)}{D(x)}\,\mathrm{d}x.$$

L'intégration de Q étant triviale, on conclut que la difficulté de l'intégration d'une fonction rationnelle se réduit à l'intégration d'une fonction rationnelle propre.

Propriété

Si $\frac{R(x)}{D(x)}$ est une fonction rationnelle propre et si D possède

▷ k racines réelles a_k chacune de multiplicité m_k et

▷ h couples de racines complexes conjuguées qui sont racines du polynôme $x^2 + b_h x + d_h$ chacune de multiplicité n_h (ainsi $\Delta = b_h^2 - 4d_h < 0$ pour tout h),

alors D s'écrit

$$D(x) = c(x-a_1)^{m_1}(x-a_2)^{m_2}\dots(x-a_k)^{m_k}(x^2+b_1x+d_1)^{n_1}(x^2+b_2x+d_2)^{n_2}\dots(x^2+b_hx+d_h)^{n_h}$$

et $\frac{R(x)}{D(x)}$ se décompose en *fractions simples* sous la forme

$$\begin{aligned}
\frac{R(x)}{D(x)} = &\frac{A_{1,1}}{x-a_1} + \frac{A_{1,2}}{(x-a_1)^2} + \dots + \frac{A_{1,m_1}}{(x-a_1)^{m_1}}\\
&+ \frac{A_{2,1}}{x-a_2} + \frac{A_{2,2}}{(x-a_2)^2} + \dots + \frac{A_{2,m_1}}{(x-a_2)^{m_2}}\\
&+ \dots +\\
&+ \frac{A_{k,1}}{x-a_k} + \frac{A_{k,2}}{(x-a_k)^2} + \dots + \frac{A_{k,m_k}}{(x-a_k)^{m_k}}\\
&+ \frac{B_{1,1}x+C_{1,1}}{x^2+b_1x+d_1} + \frac{B_{1,2}x+C_{1,2}}{(x^2+b_1x+d_1)^2} + \dots + \frac{B_{1,n_1}x+C_{1,n_1}}{(x^2+b_1x+d_1)^{n_1}}\\
&+ \frac{B_{2,1}x+C_{2,1}}{x^2+b_2x+d_2} + \frac{B_{2,2}x+C_{2,2}}{(x^2+b_2x+d_2)^2} + \dots + \frac{B_{2,n_2}x+C_{2,n_2}}{(x^2+b_2x+d_2)^{n_2}}\\
&+ \dots +\\
&+ \frac{B_{h,1}x+C_{h,1}}{x^2+b_hx+d_h} + \frac{B_{h,2}x+C_{h,2}}{(x^2+b_hx+d_h)^2} + \dots + \frac{B_{h,n_h}x+C_{h,n_h}}{(x^2+b_hx+d_h)^{n_k}}
\end{aligned}$$

où les $A_{i,j}$, $B_{i,j}$ et $C_{i,j}$ sont des constantes.

Pour intégrer une fonction rationnelle il suffit alors de connaître les primitives des quatre fractions simples suivantes :

$$f_1(x) = \frac{A}{x-a}, \qquad f_2(x) = \frac{A}{(x-a)^n}, \qquad f_3(x) = \frac{Bx+C}{x^2+bx+d}, \qquad f_4(x) = \frac{Bx+C}{(x^2+bx+d)^n}.$$

Intégration des fractions simples

Supposons

▷ $A, B, C, a, b, d \in \mathbb{R}$

▷ $n \in \mathbb{N}$, $n > 1$

▷ $\Delta = b^2 - 4d < 0$

alors

1. la primitive de $f_1(x) = \frac{A}{x-a}$ est

$$\int \frac{A}{x-a}\,\mathrm{d}x = A\ln|x-a| + \text{cnst};$$

2. la primitive de $f_2(x)=\frac{A}{(x-a)^n}$ est
$$\int \frac{A}{(x-a)^n}\,\mathrm{d}x=\frac{A}{(1-n)(x-a)^{n-1}}+\text{cnst};$$
3. la primitive de $f_3(x)=\frac{Bx+C}{x^2+bx+d}$ est
$$\begin{aligned}\int \frac{Bx+C}{x^2+bx+d}\,\mathrm{d}x&=\int \frac{Bx+C}{\left(x+\frac{b}{2}\right)^2+\left(d-\frac{b^2}{4}\right)}\,\mathrm{d}x\\&=\frac{1}{\sqrt{d-\frac{b^2}{2}}}\int \frac{B\left(\sqrt{d-\frac{b^2}{4}}\,t-\frac{b}{2}\right)+C}{t^2+1}\,\mathrm{d}t\\&=\frac{B}{2}\int\frac{2t}{1+t^2}\,\mathrm{d}t+\frac{C-B\frac{b}{2}}{\sqrt{d-\frac{b^2}{4}}}\int\frac{1}{1+t^2}\,\mathrm{d}t\\&=\frac{B}{2}\ln|1+t^2|+\frac{C-B\frac{b}{2}}{\sqrt{d-\frac{b^2}{4}}}\arctan(t)+\text{cnst}\\&=\frac{B}{2}\ln\left|1+\frac{\left(x+\frac{b}{2}\right)^2}{d-\frac{b^2}{4}}\right|+\frac{C-B\frac{b}{2}}{\sqrt{d-\frac{b^2}{4}}}\arctan\left(\frac{x+\frac{b}{2}}{\sqrt{d-\frac{b^2}{4}}}\right)+\text{cnst};\end{aligned}$$
$$x+\frac{b}{2}=\sqrt{d-\frac{b^2}{4}}\,t\qquad \mathrm{d}x=\sqrt{d-\frac{b^2}{4}}\,\mathrm{d}t$$
4. la primitive de $f_4(x)=\frac{Bx+C}{(x^2+bx+d)^n}$ est
$$\int \frac{Bx+C}{(x^2+bx+d)^n}\,\mathrm{d}x=\frac{B}{2}\underbrace{\int\frac{2x+b}{(x^2+bx+d)^n}\,\mathrm{d}x}_{I_1}+\left(C-\frac{Bb}{2}\right)\underbrace{\int\frac{1}{(x^2+bx+d)^n}\,\mathrm{d}x}_{I_2}$$
avec
$$\begin{aligned}I_1&=\frac{B}{2}\frac{1}{(1-n)(x^2+bx+d)^{n-1}}\\I_2&=\int\frac{1}{(x^2+bx+d)^n}\,\mathrm{d}x\\&=\left(d-\frac{b^2}{4}\right)^{\frac{1}{2}-n}\underbrace{\int\frac{1}{(1+t^2)^n}\,\mathrm{d}t}_{I^n}\end{aligned}$$
$$x+\frac{b}{2}=\sqrt{d-\frac{b^2}{4}}\,t\qquad \mathrm{d}x=\sqrt{d-\frac{b^2}{4}}\,\mathrm{d}t$$
et l'intégrale I^n se calcule par récurrence
$$I^n=\frac{t}{2(n-1)(1+t^2)^{n-1}}+\frac{2n-3}{2(n-1)}I^{n-1}.$$

Exemple

1. On veut intégrer la fonction rationnelle propre $f(x)=\frac{x-3}{x^2-4x+5}$. Comme $\Delta=4^2-4\times5<0$ il s'agit d'un intégrale du type $\int\frac{Bx+C}{x^2+bx+d}\,\mathrm{d}x$. Le dénominateur se décompose comme $x^2-4x+5=(x-2)^2+1$ et l'intégrale s'écrit
$$\begin{aligned}\int f(x)\,\mathrm{d}x&=\int\frac{x-3}{(x-2)^2+1}\,\mathrm{d}x=\int\frac{(t+2)-3}{t^2+1}\,\mathrm{d}t=\frac{1}{2}\int\frac{2t}{t^2+1}\,\mathrm{d}t-\int\frac{1}{t^2+1}\,\mathrm{d}t\\&=\frac{1}{2}\ln(1+t^2)-\arctan(t)+\text{cnst}=\frac{1}{2}\ln((x+2)^2+1)-\arctan(x-2)+\text{cnst}\end{aligned}$$
2. On veut intégrer la fonction rationnelle propre $f(x)=\frac{3x+1}{x^3-4x}$. On doit d'abord la décomposer en fractions simples ; comme $x^3-4x=x(x^2-4)=x(x-2)(x+2)$ la fonction admet la décomposition
$$f(x)=\frac{3x+1}{x(x-2)(x+2)}=\frac{A_1}{x}+\frac{A_2}{x-2}+\frac{A_3}{x+2}$$
Pour calculer les constantes A_i on peut utiliser le principe d'identité des polynômes :
$$\frac{3x+1}{x(x-2)(x+2)}=\frac{A_1(x-2)(x+2)+A_2x(x+2)+A_3x(x-2)}{x(x-2)(x+2)}=\frac{(A_1+A_2+A_3)x^2+(2A_2-2A_3)x-4A_1}{x(x-2)(x+2)}\iff\begin{cases}A_1+A_2+A_3=0\\2A_2-2A_3=3\\-4A_1=1\end{cases}$$

ainsi

$$\int f(x)\,\mathrm{d}x = -\frac{1}{4}\int \frac{1}{x}\,\mathrm{d}x - \frac{5}{8}\int \frac{A_2}{x-2}\,\mathrm{d}x + \frac{7}{8}\int \frac{A_3}{x+2}\,\mathrm{d}x = -\frac{1}{4}\ln|x| - \frac{5}{8}\ln|x-2| + \frac{7}{8}\ln|x+2| + c.$$

3. On veut intégrer la fonction rationnelle impropre $f(x) = \frac{3x^3+2x-5}{3x^2-5x-2}$. On effectue d'abord la division euclidienne

$$\begin{array}{cccc|l} 3x^3 & & +2x & -5 & 3x^2-5x-2 \\ \cline{5-5} -3x^3 & +5x^2 & +2x & & x+\frac{5}{3} \\ \cline{1-4} & 5x^2 & +4x & -5 & \\ & -5x^2 & +\frac{25}{3}x & +\frac{10}{3} & \\ \cline{2-4} & & \frac{37}{3}x & -\frac{5}{3} & \end{array}$$

ainsi $f(x) = x+\frac{5}{3}+\frac{\frac{37}{3}x-\frac{5}{3}}{3x^2-5x-2}$. Maintenant on décompose le terme $\frac{\frac{37}{3}x-\frac{5}{3}}{3x^2-5x-2}$ en fractions simples : on a $3x^2-5x-2 = (x+\frac{1}{3})(x-2)$ et on doit chercher les deux constantes A_1 et A_2 telles que

$$\frac{\frac{37}{3}x-\frac{5}{3}}{3x^2-5x-2} = \frac{A_1}{x+\frac{1}{3}} + \frac{A_2}{x-2}$$

En utilisant le principe d'identité des polynômes on a

$$\frac{\frac{37}{3}x-\frac{5}{3}}{3x^2-5x-2} = \frac{A_1(x-2)+A_2(x+\frac{1}{3})}{3x^2-5x-2} = \frac{(A_1+A_2)x-2A_1+\frac{A_2}{3}}{3x^2-5x-2} \iff \begin{cases} A_1+A_2 = \frac{37}{3} \\ -2A_1+\frac{A_2}{3} = -\frac{5}{3} \end{cases}$$

On conclut que

$$f(x) = x+\frac{5}{3}+\frac{52/63}{x+\frac{1}{3}}+\frac{207/63}{x-2}$$

et

$$\int f(x)\,\mathrm{d}x = \frac{x^2}{2}+\frac{5}{3}x+\frac{52}{63}\ln\left|x+\frac{1}{3}\right|+\frac{207}{63}\ln|x-2|+c.$$

4. On veut intégrer la fonction rationnelle propre $f(x) = \frac{x-4}{x^3-x^2-5x-3}$. On doit d'abord la décomposer en fraction simples ; comme $x^3-x^2-5x-3 = (x-3)(x+1)^2$ la fonction f admet la décomposition

$$f(x) = \frac{A_{1,1}}{x-3}+\frac{A_{2,1}}{x+1}+\frac{A_{2,2}}{(x+1)^2}.$$

On détermine les constantes en utilisant le principe d'identité des polynômes

$$\begin{aligned}\frac{x-4}{x^3-x^2-5x-3} &= \frac{A_{1,1}(x+1)^2+A_{2,1}(x-3)(x+1)+A_{2,2}(x-3)}{x^3-x^2-5x-3} \\ &= \frac{(A_{1,1}+A_{2,1})x^2+(2A_{1,1}-2A_{2,1}+A_{2,2})x+A_{1,1}-3A_{2,1}-3A_{2,2}}{x^3-x^2-5x-3} \iff \begin{cases} A_{1,1}+A_{2,1}=0 \\ 2A_{1,1}-2A_{2,1}+A_{2,2}=1 \\ A_{1,1}-3A_{2,1}-3A_{2,2}=-4 \end{cases}\end{aligned}$$

On conclut que

$$f(x) = \frac{-1/16}{x-3}+\frac{1/16}{x+1}+\frac{-5/4}{(x+1)^2}$$

et

$$\int f(x)\,\mathrm{d}x = -\frac{1}{16}\ln|x-3|+\frac{1}{16}\ln|x+1|-\frac{5}{4(x+1)}+c.$$

5. On veut intégrer la fonction rationnelle propre $f(x) = \frac{x^2+2}{(x^2-2x+5)^2}$. Comme $\Delta = 4-20 < 0$, la fonction f se décompose comme

$$f(x) = \frac{B_1x+C_1}{x^2-2x+5}+\frac{B_2x+C_2}{(x^2-2x+5)^2}.$$

On détermine les constantes en utilisant le principe d'identité des polynômes

$$\frac{x^2+2}{(x^2-2x+5)^2} = \frac{B_1x+C_1}{x^2-2x+5}+\frac{B_2x+C_2}{(x^2-2x+5)^2} = \frac{B_1x^3+(C_1-2B_1)x^2+(5B_1-2C_1+B_2)x+5C_1+C_2}{(x^2-2x+5)^2} \iff \begin{cases} B_1=0 \\ C_1-2B_1=1 \\ 5B_1-2C_1+B_2=0 \\ 5C_1+C_2=2 \end{cases}$$

On obtient alors que

$$\int f(x)\,\mathrm{d}x = \underbrace{\int \frac{1}{x^2-2x+5}\,\mathrm{d}x}_{I_1}+\underbrace{\int \frac{2x-3}{(x^2-2x+5)^2}\,\mathrm{d}x}_{I_2}.$$

▷ On calcule I_1 :

$$\begin{aligned}\int \frac{1}{x^2-2x+5}\,\mathrm{d}x &= \int \frac{1}{(x-1)^2+4}\,\mathrm{d}x \qquad \begin{aligned}x-1&=2t\\ \mathrm{d}x&=2\,\mathrm{d}t\end{aligned}\\ &= \frac{1}{2}\int \frac{1}{t^2+1}\,\mathrm{d}t\\ &= \frac{1}{2}\int \frac{1}{1+t^2}\,\mathrm{d}t = \frac{1}{2}\arctan(t)+c = \frac{1}{2}\arctan\left(\frac{x-1}{2}\right)+c;\end{aligned}$$

▷ On calcule I_2 :

$$\begin{aligned}\int \frac{2x-3}{(x^2-2x+5)^2}\,\mathrm{d}x &= \int \frac{2x-2}{(x^2-2x+5)^2}\,\mathrm{d}x - \int \frac{1}{(x^2-2x+5)^2}\,\mathrm{d}x = -\frac{1}{x^2-2x+5} - \int \frac{1}{(x^2-2x+5)^2}\,\mathrm{d}x\\ &= -\frac{1}{x^2-2x+5} - \int \frac{1}{((x-1)^2+4)^2}\,\mathrm{d}x = -\frac{1}{x^2-2x+5} - \frac{1}{8}\int \frac{1}{(t^2+1)^2}\,\mathrm{d}t\\ &= -\frac{1}{x^2-2x+5} - \frac{1}{8}\left(\frac{1}{2}\frac{t}{1+t^2} + \frac{1}{2}\int \frac{1}{1+t^2}\,\mathrm{d}t\right) = -\frac{1}{x^2-2x+5} - \frac{1}{16}\left(\frac{t}{1+t^2}+\arctan(t)\right)\\ &= -\frac{1}{x^2-2x+5} - \frac{1}{16}\left(\frac{2(x-1)}{x^2-2x+5}+\arctan\left(\frac{x-1}{2}\right)\right)+c = -\frac{x+7}{8(x^2-2x+5)} - \frac{1}{16}\arctan\left(\frac{x-1}{2}\right)+c.\end{aligned}$$

On conclut que

$$\int f(x)\,\mathrm{d}x = \frac{7}{16}\arctan\left(\frac{x-1}{2}\right) - \frac{x+7}{8(x^2-2x+5)}+c.$$

6. On veut intégrer la fonction rationnelle propre $f(x)=\frac{1}{x^3(x^2+1)^2}$ qui se décompose comme

$$f(x)=\frac{A_1}{x}+\frac{A_1}{x^2}+\frac{A_1}{x^3}+\frac{B_1x+C_1}{x^2+1}+\frac{B_2x+C_2}{(x^2+1)^2}.$$

On détermine les constantes en utilisant le principe d'identité des polynômes

$$\begin{aligned}\frac{1}{x^3(x^2+1)^2} &= \frac{A_1}{x}+\frac{A_2}{x^2}+\frac{A_3}{x^3}+\frac{B_1x+C_1}{x^2+1}+\frac{B_2x+C_2}{(x^2+1)^2}\\ &= \frac{(A_1+B_1)x^6+(A_2+C_1)x^5+(2A_1+A_3+B_1+B_2)x^4+(2A_2+C_1+C_2)x^3+(A_1+2A_3)x^2+(A_2)x+(A_3)}{x^3(x^2+1)^2}\\ &\iff \begin{cases}A_1+B_1=0\\ A_2+C_1=0\\ 2A_1+A_3+B_1+B_2=0\\ 2A_2+C_1+C_2=0\\ A_1+2A_3=0\\ A_2=0\\ A_3=1\end{cases}\end{aligned}$$

On obtient alors que

$$\begin{aligned}\int f(x)\,\mathrm{d}x &= -2\int \frac{1}{x}\,\mathrm{d}x + \int \frac{1}{x^3}\,\mathrm{d}x + \int \frac{2x}{x^2+1}\,\mathrm{d}x + \int \frac{x}{(x^2+1)^2}\,\mathrm{d}x\\ &= -2\ln|x| - \frac{1}{2x^2} + \ln|x^2+1| + \frac{1}{2}\int \frac{2x}{(x^2+1)^2}\,\mathrm{d}x\\ &= -2\ln|x| - \frac{1}{2x^2} + \ln|x^2+1| - \frac{1}{2}\frac{2x}{(x^2+1)}+c.\end{aligned}$$

Théorème fondamentale du calcul différentiel et intégral

Soit f une fonction continue sur $[a;b]$ et F une primitive de f sur $[a;b]$. Alors

1. la dérivée de $g(x)=\int_a^x f(x)\,\mathrm{d}x$ existe et est égale à $f(x)$;
2. $\int_a^b f(x)\,\mathrm{d}x = F(b)-F(a)$.

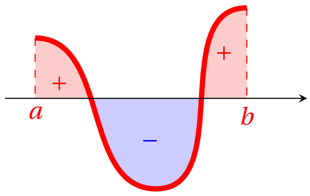
L'expression $F(b)-F(a)$ se note aussi $[F(x)]_a^b$ ou encore $F(x)|_a^b$.

Remarque

Si f est de classe $\mathscr{C}^1$ sur $[a;b]$ alors $g(x)=f(x)-f(a)=\int_a^x f'(x)\,\mathrm{d}x$.

Interprétation géométrique

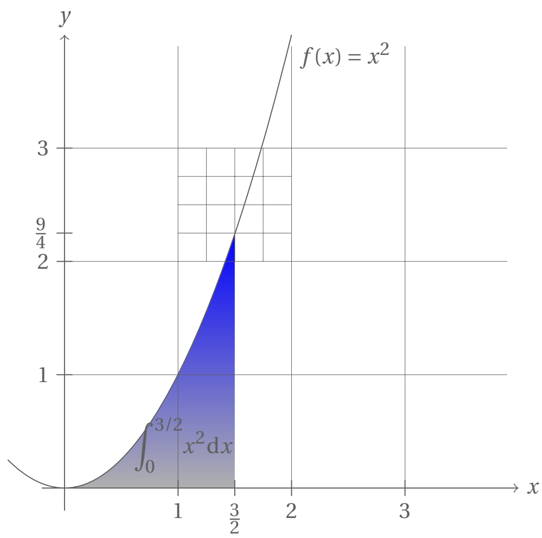
L'intégrale définie $\int_a^b f(x)\,\mathrm{d}x$ correspond à l'aire du domaine du plan situé entre le graphe de f et l'axe des abscisses, compté positivement pour la partie située au-dessus de l'axe des abscisses et négativement pour la partie située en dessous.

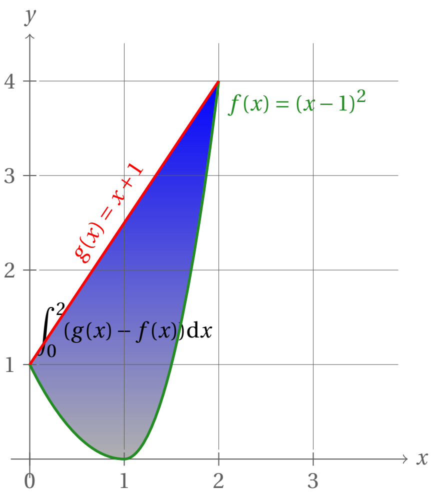


Calcul d'aires

Soit f et g deux fonctions intégrables sur $[a;b]$ telles que $f(x) \ge g(x)$ sur $[a;b]$. L'aire de la surface comprise entre les graphes de f et g entre a et b est définie par l'intégrale $\int_a^b (f(x) - g(x))\,\mathrm{d}x$.

Exemple





Propriété

Soit $[a;b]$ un intervalle de $\mathbb{R}$ avec $a < b$, f et $g : [a;b] \to \mathbb{R}$ deux fonctions intégrables.

Relation de Chasles : $\int_a^b f(x)\,\mathrm{d}x = \int_a^c f(x)\,\mathrm{d}x + \int_c^b f(x)\,\mathrm{d}x$ pour tout $c \in [a;b]$.

Relation d'ordre : si $f(x) \le g(x)$ pour toute $x \in [a;b]$ alors $\int_a^b f(x)\,\mathrm{d}x \le \int_a^b g(x)\,\mathrm{d}x$.

Nullité : si f est continue sur $[a;b]$ et $f(x) \ge 0$ pour toute $x \in [a;b]$ alors $\int_a^b f(x)\,\mathrm{d}x = 0$ si et seulement si $f(x) \equiv 0$.

Parité : si f est paire sur $[-a;a]$ avec $a \ge 0$ alors $\int_{-a}^a f(x)\,\mathrm{d}x = 2\int_0^a f(x)\,\mathrm{d}x$; si f est impaire sur $[-a;a]$ avec $a \ge 0$ alors $\int_{-a}^a f(x)\,\mathrm{d}x = 0$.

Valeur absolue : $|\int_a^b f(x)\,\mathrm{d}x| \le \int_a^b |f(x)|\,\mathrm{d}x$.

Moyenne : si $m \le f(x) \le M$ pour toute $x \in [a;b]$ alors $m \le \frac{\int_a^b f(x)\,\mathrm{d}x}{b-a} \le M$. Le nombre $\frac{\int_a^b f(x)\,\mathrm{d}x}{b-a}$ est la valeur moyenne de f sur $[a;b]$.

A.3. Matrices et calcul pratique d'un déterminant

Définition : matrice

On appelle MATRICE $m \times n$ (ou d'ordre $m \times n$) à coefficients dans $\mathbb{K}$ tout tableau de m lignes et n colonnes d'éléments de $\mathbb{K}$. L'ensemble des matrices $m \times n$ à coefficients dans $\mathbb{K}$ est noté $\mathscr{M}_{m,n}(\mathbb{K})$.
Si $m = n$ on dit qu'on a une MATRICE CARRÉE. L'ensemble des matrices carrées n à coefficients dans $\mathbb{K}$ est noté $\mathscr{M}_n(\mathbb{K})$.
Une matrice $m \times 1$ est appelée VECTEUR-COLONNE et une matrice $1 \times n$ est appelée VECTEUR-LIGNE.
On convient de noter a_{ij} l'élément de la matrice situé sur la i-ème ligne et j-ème colonne ($1 \le i \le m$ et $1 \le j \le n$).

Une matrice $\mathbb{A}$ est représentée entre deux parenthèses ou deux crochets :

$$\mathbb{A} = \begin{pmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1n} \\ \vdots & & \vdots & & \vdots \\ a_{i1} & \dots & a_{ij} & \dots & a_{in} \\ \vdots & & \vdots & & \vdots \\ a_{m1} & \dots & a_{mj} & \dots & a_{mn} \end{pmatrix} \quad \text{ou} \quad \mathbb{A} = \begin{bmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1n} \\ \vdots & & \vdots & & \vdots \\ a_{i1} & \dots & a_{ij} & \dots & a_{in} \\ \vdots & & \vdots & & \vdots \\ a_{m1} & \dots & a_{mj} & \dots & a_{mn} \end{bmatrix}$$

ou encore

$$\mathbb{A} = (a_{ij})_{\substack{1\le i\le m\\1\le j\le n}} \quad \text{ou} \quad \mathbb{A} = [a_{ij}]_{\substack{1\le i\le m\\1\le j\le n}}$$

Nous travaillerons avec $\mathbb{K} = \mathbb{R}$, $\mathbb{Q}$ ou $\mathbb{Z}$.

Définition : addition de matrices

Si $\mathbb{A} = (a_{ij})_{\substack{1\le i\le m\\1\le j\le n}}$ et $\mathbb{B} = (b_{ij})_{\substack{1\le i\le m\\1\le j\le n}}$ sont deux matrices $m\times n$, on définit l'ADDITION des matrices par

$$\mathbb{A}+\mathbb{B} = (a_{ij}+b_{ij})_{\substack{1\le i\le m\\1\le j\le n}}.$$

La MATRICE NULLE, notée $\mathbb{O}_{m,n}$, est la matrice dont tous les éléments sont nuls.
La MATRICE OPPOSÉE D'UNE MATRICE $\mathbb{A}$ est notée $-\mathbb{A}$. Si $\mathbb{A} = (a_{ij})_{\substack{1\le i\le m\\1\le j\le n}}$ alors $-\mathbb{A} = (-a_{ij})_{\substack{1\le i\le m\\1\le j\le n}}$.

La somme de deux matrices d'ordres différents n'est pas définie.

Propriété

Si $\mathbb{A}$, $\mathbb{B}$ et $\mathbb{C}$ sont des matrices de même ordre, alors nous avons
- ▷ $\mathbb{A}+\mathbb{B} = \mathbb{B}+\mathbb{A}$ (commutativité),
- ▷ $\mathbb{A}+(\mathbb{B}+\mathbb{C}) = (\mathbb{A}+\mathbb{B})+\mathbb{C}$ (associativité).

Définition

On appelle MATRICE DIAGONALE toute matrice carrée $\mathbb{D} = (d_{ij})_{1\le i,j\le n}$ telle que $i\ne j \implies d_{ij} = 0$. Si on note $d_i = d_{ii}$, une matrice diagonale est de la forme

$$\mathbb{D}_n = \begin{pmatrix} d_1 & 0 & \dots & 0 & 0 \\ 0 & d_2 & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & d_{n-1} & 0 \\ 0 & 0 & \dots & 0 & d_n \end{pmatrix}.$$

On la note $\text{Diag}(d_1, d_2, \dots, d_n)$.
La MATRICE IDENTITÉ d'ordre n, notée $\mathbb{I}_n$, est la matrice diagonale $\text{Diag}(1,1,\dots,1)$.

Définition

On dit qu'une matrice carrée $\mathbb{A} = (a_{ij})_{1\le i,j\le n}$ est TRIANGULAIRE SUPÉRIEURE (resp. INFÉRIEURE) si $i > j \implies a_{ij} = 0$ (resp. si $i < j \implies a_{ij} = 0$).

Une matrice triangulaire supérieure et inférieure est une matrice diagonale.

Définition : produit d'une matrice par un scalaire

Si $\mathbb{A} = (a_{ij})_{\substack{1\le i\le m\\1\le j\le n}}$ est une matrice $m\times n$ et si $\alpha\in\mathbb{K}$, on définit le PRODUIT D'UNE MATRICE PAR UN SCALAIRE par

$$\alpha\cdot\mathbb{A} = (\alpha\cdot a_{ij})_{\substack{1\le i\le m\\1\le j\le n}}$$

Propriété

Soient $\mathbb{A}$ et $\mathbb{B}$ deux matrices de même ordre et $\alpha \in \mathbb{K}$ un scalaire, on a

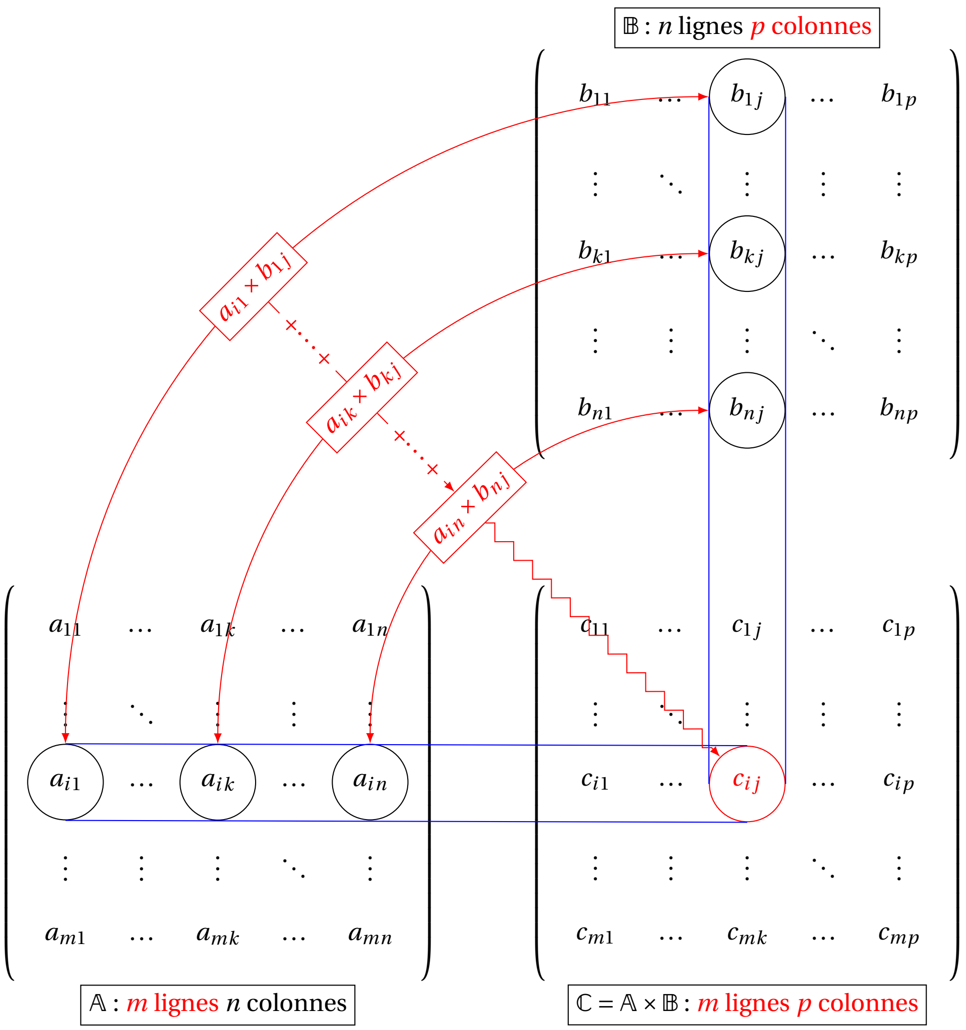
▷ $\alpha \cdot (\mathbb{A}+\mathbb{B}) = \alpha \cdot \mathbb{A} + \alpha \cdot \mathbb{B}$ (distributivité).

Définition : produit de matrices

Si $\mathbb{A} = (a_{ik})_{\substack{1\le i\le m\\ 1\le k\le n}}$ est une matrice $m\times n$ et si $\mathbb{B} = (b_{kj})_{\substack{1\le k\le n\\ 1\le j\le p}}$ est une matrice $n\times p$, on définit le PRODUIT DES MATRICES par

$$\mathbb{A}\times\mathbb{B} = \left(\sum_{k=1}^{n} a_{ik}b_{kj}\right)_{\substack{1\le i\le m\\ 1\le j\le p}}$$

C'est une matrice $m\times p$.



Exemple

Soient les deux matrices

$$\mathbb{A} = \begin{pmatrix} 1 & 3 & 0\\ -1 & 1 & 2\end{pmatrix} \qquad \text{et} \qquad \mathbb{B} = \begin{pmatrix} 1 & 2 & 0\\ 0 & 2 & 3\\ 0 & -1 & -2\end{pmatrix}$$

La matrice $\mathbb{A}$ est d'ordre 2×3, la matrice $\mathbb{B}$ est d'ordre 3×3, donc la matrice produit $\mathbb{A}\times\mathbb{B}$ est une matrice d'ordre 2×3 :

$$\mathbb{A}\times\mathbb{B} = \begin{pmatrix} 1\times1+3\times0+0\times0 & 1\times2+3\times2+0\times(-1) & 1\times0+3\times3+0\times(-2)\\ -1\times1+1\times0+2\times0 & -1\times2+1\times2+2\times(-1) & -1\times0+1\times3+2\times(-2)\end{pmatrix} = \begin{pmatrix} 1 & 7 & 9\\ -1 & -2 & -1\end{pmatrix}.$$

Propriété

Si $\mathbb{A} \in \mathscr{M}_{m,n}(\mathbb{K})$, $\mathbb{B} \in \mathscr{M}_{n,p}(\mathbb{K})$, $\mathbb{C} \in \mathscr{M}_{p,q}(\mathbb{K})$, alors
- ▷ $\mathbb{A} \times (\mathbb{B} \times \mathbb{C}) = (\mathbb{A} \times \mathbb{B}) \times \mathbb{C}$ (associativité) ;
- ▷ $\mathbb{A}(\times(\mathbb{B} + \mathbb{C}) = \mathbb{A} \times \mathbb{B} + \mathbb{A} \times \mathbb{C}$ (distributivité) ;

Si $\mathbb{A} \in \mathscr{M}_n(\mathbb{K})$ alors $\mathbb{A} \times \mathbb{I}_n = \mathbb{I}_n \times \mathbb{A} = \mathbb{A}$.

Attention

$\mathbb{A} \times \mathbb{B} \neq \mathbb{B} \times \mathbb{A}$ en général (non commutativité).

Prenons le cas général avec $\mathbb{A}$ d'ordre $m \times p$ et $\mathbb{B}$ d'ordre $p \times n$. Le produit $\mathbb{A} \times \mathbb{B}$ est défini, c'est une matrice d'ordre $m \times n$. Qu'en est-il du produit $\mathbb{B} \times \mathbb{A}$? Il faut distinguer trois cas :
- ▷ *si $m \neq n$ le produit $\mathbb{B} \times \mathbb{A}$ n'est pas défini ;*
- ▷ *si $m = n$ mais $p \neq n$, le produit $\mathbb{A} \times \mathbb{B}$ est défini et c'est une matrice d'ordre $m \times n$ tandis que le produit $\mathbb{B} \times \mathbb{A}$ est défini mais c'est une matrice d'ordre $p \times p$ donc $\mathbb{A} \times \mathbb{B} \neq \mathbb{B} \times \mathbb{A}$;*
- ▷ *si $m = n = p$, $\mathbb{A}$ et $\mathbb{B}$ sont deux matrices carrées d'ordre m. Les produits $\mathbb{A} \times \mathbb{B}$ et $\mathbb{B} \times \mathbb{A}$ sont aussi carrés et d'ordre m mais là encore, en général, $\mathbb{A} \times \mathbb{B} \neq \mathbb{B} \times \mathbb{A}$;*

Exemple

Soient les matrices

$$\mathbb{A} = \begin{pmatrix} 1 & -1 \\ 3 & 0 \end{pmatrix}, \qquad \mathbb{B} = \begin{pmatrix} 6 & -5 \\ 2 & 1 \end{pmatrix}.$$

On obtient

$$\mathbb{A} \times \mathbb{B} = \begin{pmatrix} 4 & -6 \\ 18 & -15 \end{pmatrix} \qquad \text{et} \qquad \mathbb{B} \times \mathbb{A} = \begin{pmatrix} -9 & -6 \\ 5 & -2 \end{pmatrix}.$$

Définition : matrice transposée

Si $\mathbb{A} = (a_{ij})_{\substack{1 \le i \le m \\ 1 \le j \le n}}$ est une matrice $m \times n$, on définit la matrice TRANSPOSÉE de $\mathbb{A}$, notée $\mathbb{A}^T$, par $\mathbb{A} = (a_{ji})_{\substack{1 \le j \le n \\ 1 \le i \le m}}$.

C'est donc une matrice $n \times m$ obtenue en échangeant lignes et colonnes de la matrice initiale.

Propriété

- ▷ $(\mathbb{A}^T)^T = \mathbb{A}$ si $\mathbb{A} \in \mathscr{M}_{m,n}(\mathbb{K})$,
- ▷ $(\alpha\mathbb{A})^T = \alpha\mathbb{A}^T$ si $\alpha \in \mathbb{K}$ et $\mathbb{A} \in \mathscr{M}_{m,n}(\mathbb{K})$,
- ▷ $(\mathbb{A} + \mathbb{B})^T = \mathbb{A}^T + \mathbb{B}^T$ si $\mathbb{A}, \mathbb{B} \in \mathscr{M}_{m,n}(\mathbb{K})$,
- ▷ $(\mathbb{A} \times \mathbb{B})^T = \mathbb{B}^T \times \mathbb{A}^T$ si $\mathbb{A} \in \mathscr{M}_{m,n}(\mathbb{K})$ et $\mathbb{B} \in \mathscr{M}_{n,p}(\mathbb{K})$.

Définition : matrice symétrique, matrice anti-symétrique

- ▷ Une matrice $\mathbb{A}$ est dite SYMÉTRIQUE si $\mathbb{A}^T = \mathbb{A}$.
- ▷ Une matrice $\mathbb{A}$ est dite ANTI-SYMÉTRIQUE si $\mathbb{A}^T = -\mathbb{A}$.

Exemple

$$\mathbb{A} = \begin{pmatrix} 1 & 5 & -9 \\ 5 & 4 & 0 \\ -9 & 0 & 7 \end{pmatrix} \qquad \text{est une matrice symétrique.}$$

$$\mathbb{B} = \begin{pmatrix} 1 & 5 & -9 \\ -5 & 4 & 0 \\ 9 & 0 & 7 \end{pmatrix} \qquad \text{est une matrice anti-symétrique.}$$

Définition : matrice inversible, matrice singulière

Une matrice carrée $\mathbb{A} \in \mathscr{M}_n(\mathbb{K})$ est dite INVERSIBLE ou régulière si elle est symétrisable pour le produit matriciel, autrement dit s'il existe une matrice $\mathbb{B} \in \mathscr{M}_n(\mathbb{K})$ telle que $\mathbb{A} \times \mathbb{B} = \mathbb{B} \times \mathbb{A} = \mathbb{I}_n$. Une matrice non régulière est dite SINGULIÈRE. L'inverse, s'il existe, d'une matrice $\mathbb{A}$ est noté $\mathbb{A}^{-1}$.

Proposition

Soit $\mathbb{A}$ et $\mathbb{B}$ deux matrices inversibles, alors
- ▷ $\mathbb{A}^{-1}$ l'est aussi et $\left(\mathbb{A}^{-1}\right)^{-1} = \mathbb{A}$,
- ▷ $\mathbb{A}^T$ l'est aussi et $\left(\mathbb{A}^T\right)^{-1} = \left(\mathbb{A}^{-1}\right)^T$,
- ▷ $\mathbb{A} \times \mathbb{B}$ l'est aussi et $(\mathbb{A} \times \mathbb{B})^{-1} = \mathbb{B}^{-1} \times \mathbb{A}^{-1}$.

Opérations élémentaires sur les matrices

Les opérations (ou manipulations) élémentaires sur les lignes d'une matrices sont
- ▷ la multiplication d'une ligne par un scalaire non nul :
$$L_i \leftarrow \alpha L_i;$$
- ▷ l'addition d'un multiple d'une ligne à une autre ligne :
$$L_i \leftarrow L_i + \alpha L_j;$$
- ▷ l'échange de deux lignes :
$$L_i \leftrightarrow L_j.$$

Ces transformations sont équivalentes à la multiplication à gauche (pré-multiplication) par la matrice inversible obtenue en appliquant à la matrice identité la transformation correspondante.

Les opérations analogues sur les colonnes sont
- ▷ *la multiplication d'une colonne par un scalaire non nul :*
$$C_i \leftarrow \alpha C_i;$$
- ▷ *l'addition d'un multiple d'une colonne à une autre colonne :*
$$C_i \leftarrow C_i + \alpha C_j;$$
- ▷ *l'échange de deux colonnes :*
$$C_i \leftrightarrow C_j.$$

Elles sont équivalentes à la multiplication à droite (post-multiplication) par la matrice inversible obtenue en appliquant à la matrice identité la transformation correspondante.

Théorème

En partant d'une matrice $\mathbb{A}$, l'utilisation d'un nombre fini d'opérations élémentaires conduit à une matrice équivalente.

Déterminant d'une matrice d'ordre n

Soit $\mathbb{A}$ une matrice carrée d'ordre n.
Étant donné un couple (i, j) d'entiers, $1 \le i, j \le n$, on note $\mathbb{A}_{ij}$ la matrice carrée d'ordre $n-1$ obtenue en supprimant la i-ème ligne et la j-ème colonne de $\mathbb{A}$.
On défini le DÉTERMINANT de $\mathbb{A}$, et on le note $\det(\mathbb{A})$, par récurrence sur l'ordre de la matrice $\mathbb{A}$:
- ▷ si $n = 1$: le déterminant de $\mathbb{A}$ est le nombre
$$\det(\mathbb{A}) = a_{11},$$
- ▷ si $n > 1$: le déterminant de $\mathbb{A}$ est le nombre
$$\det(\mathbb{A}) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(\mathbb{A}_{ij}) \qquad \text{quelque soit la ligne } i,\ 1 \le i \le n,$$
ou, de manière équivalente, le nombre
$$\det(\mathbb{A}) = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det(\mathbb{A}_{ij}) \qquad \text{quelque soit la colonne } j,\ 1 \le j \le n.$$

Astuce

Pour se souvenir des signes de ces deux formules, on peut remarquer que la distribution des signes + et − avec la formule

$(-1)^{i+j}$ est analogue à la distribution des cases noirs et blanches sur un damier :

$$\begin{vmatrix} + & - & + & - & \dots \\ - & + & - & + & \dots \\ + & - & + & - & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{vmatrix}$$

Exemple

Soit la matrice

$$\mathbb{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

alors

$$\det(\mathbb{A}_{11}) = a_{22}, \qquad \det(\mathbb{A}_{12}) = a_{21}, \qquad \det(\mathbb{A}_{21}) = a_{12}, \qquad \det(\mathbb{A}_{22}) = a_{11},$$

donc on peut calculer $\det(\mathbb{A})$ par l'une des formules suivantes :
- ▷ $a_{11}\det(\mathbb{A}_{11}) - a_{12}\det(\mathbb{A}_{12}) = a_{11}a_{22} - a_{12}a_{21}$ (développement suivant la ligne $i = 1$)
- ▷ $-a_{21}\det(\mathbb{A}_{21}) + a_{22}\det(\mathbb{A}_{22}) = -a_{21}a_{12} + a_{22}a_{11}$ (développement suivant la ligne $i = 2$)
- ▷ $a_{11}\det(\mathbb{A}_{11}) - a_{21}\det(\mathbb{A}_{21}) = a_{11}a_{22} - a_{21}a_{12}$ (développement suivant la colonne $j = 1$)
- ▷ $-a_{12}\det(\mathbb{A}_{12}) + a_{22}\det(\mathbb{A}_{22}) = -a_{12}a_{21} + a_{22}a_{11}$ (développement suivant la colonne $j = 2$)

Ces formules donnent bien le même résultat.

Exemple

Soit la matrice

$$\mathbb{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

alors

$$\det(\mathbb{A}_{11}) = \det\begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix} = a_{22}a_{33} - a_{23}a_{32}, \quad \det(\mathbb{A}_{12}) = \det\begin{pmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{pmatrix} = a_{21}a_{33} - a_{23}a_{31}, \quad \det(\mathbb{A}_{13}) = \det\begin{pmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} = a_{21}a_{32} - a_{22}a_{31},$$

$$\det(\mathbb{A}_{21}) = \det\begin{pmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{pmatrix} = a_{12}a_{33} - a_{13}a_{32}, \quad \det(\mathbb{A}_{22}) = \det\begin{pmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{pmatrix} = a_{11}a_{33} - a_{13}a_{31}, \quad \det(\mathbb{A}_{23}) = \det\begin{pmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{pmatrix} = a_{11}a_{32} - a_{12}a_{31},$$

$$\det(\mathbb{A}_{31}) = \det\begin{pmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{pmatrix} = a_{12}a_{23} - a_{13}a_{22}, \quad \det(\mathbb{A}_{32}) = \det\begin{pmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{pmatrix} = a_{11}a_{23} - a_{13}a_{21}, \quad \det(\mathbb{A}_{33}) = \det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{12}a_{21},$$

donc on peut calculer $\det(\mathbb{A})$ *par l'une des formules suivantes* :
- ▷ $a_{11}\det(\mathbb{A}_{11}) - a_{12}\det(\mathbb{A}_{12}) + a_{13}\det(\mathbb{A}_{13})$ (développement suivant la ligne $i = 1$)
- ▷ $-a_{21}\det(\mathbb{A}_{21}) + a_{22}\det(\mathbb{A}_{22}) - a_{23}\det(\mathbb{A}_{23})$ (développement suivant la ligne $i = 2$)
- ▷ $a_{31}\det(\mathbb{A}_{31}) - a_{32}\det(\mathbb{A}_{32}) + a_{33}\det(\mathbb{A}_{33})$ (développement suivant la ligne $i = 3$)
- ▷ $-a_{11}\det(\mathbb{A}_{11}) + a_{21}\det(\mathbb{A}_{21}) - a_{31}\det(\mathbb{A}_{31})$ (développement suivant la colonne $j = 1$)
- ▷ $a_{12}\det(\mathbb{A}_{12}) - a_{22}\det(\mathbb{A}_{22}) + a_{32}\det(\mathbb{A}_{32})$ (développement suivant la colonne $j = 2$)
- ▷ $-a_{13}\det(\mathbb{A}_{13}) + a_{23}\det(\mathbb{A}_{23}) - a_{33}\det(\mathbb{A}_{33})$ (développement suivant la colonne $j = 3$)

Quelques calculs montrent que ces formules donnent bien le même résultat.

Astuce

Il convient d'utiliser cette définition après avoir fait apparaître sur une même rangée le plus possible de zéro sachant que
- ▷ si deux colonnes (resp. deux lignes) sont identiques ou proportionnelles, alors $\det(\mathbb{A}) = 0$;
- ▷ si on échange deux colonnes (resp. deux lignes), alors le déterminant est changé en son opposé ;
- ▷ on ne change pas un déterminant si on ajoute à une colonne (resp. une ligne) une combinaison linéaire des autres colonnes (resp. lignes).

Exemple

Soit la matrice

$$\mathbb{A} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 3 & 5 \end{pmatrix}$$

alors

$$\det(\mathbb{A}_{11}) = \det\begin{pmatrix} 2 & 0 \\ 3 & 5 \end{pmatrix} = 10, \qquad \det(\mathbb{A}_{12}) = \det\begin{pmatrix} 0 & 0 \\ 0 & 5 \end{pmatrix} = 0, \qquad \det(\mathbb{A}_{13}) = \det\begin{pmatrix} 0 & 2 \\ 0 & 3 \end{pmatrix} = 0,$$

$$\det(\mathbb{A}_{21}) = \det\begin{pmatrix}0 & 1\\ 3 & 5\end{pmatrix} = -3, \qquad \det(\mathbb{A}_{22}) = \det\begin{pmatrix}1 & 1\\ 0 & 5\end{pmatrix} = 5, \qquad \det(\mathbb{A}_{23}) = \det\begin{pmatrix}1 & 0\\ 0 & 3\end{pmatrix} = 3,$$

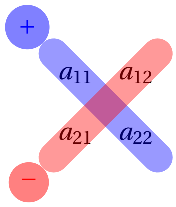
$$\det(\mathbb{A}_{31}) = \det\begin{pmatrix}0 & 1\\ 2 & 0\end{pmatrix} = -2, \qquad \det(\mathbb{A}_{32}) = \det\begin{pmatrix}1 & 1\\ 0 & 0\end{pmatrix} = 0, \qquad \det(\mathbb{A}_{33}) = \det\begin{pmatrix}1 & 0\\ 0 & 2\end{pmatrix} = 2,$$

donc on peut calculer $\det(\mathbb{A})$ par l'une des formules suivantes :
- ▷ $1\det(\mathbb{A}_{11}) + 0\det(\mathbb{A}_{12}) + 1\det(\mathbb{A}_{13}) = 10 + 0 + 0 = 10$
- ▷ $0\det(\mathbb{A}_{21}) + 2\det(\mathbb{A}_{22}) + 0\det(\mathbb{A}_{23}) = 0 + 2 \times 5 + 0 = 10$ ⇠ formule pratique car il n'y a qu'un déterminant à calculer
- ▷ $0\det(\mathbb{A}_{31}) + 3\det(\mathbb{A}_{32}) + 5\det(\mathbb{A}_{33}) = 0 + 0 + 5 \times 2 = 10$
- ▷ $1\det(\mathbb{A}_{11}) + 0\det(\mathbb{A}_{21}) + 0\det(\mathbb{A}_{31}) = 10 + 0 + 0 = 10$ ⇠ formule pratique car il n'y a qu'un déterminant à calculer
- ▷ $0\det(\mathbb{A}_{12}) + 2\det(\mathbb{A}_{22}) + 3\det(\mathbb{A}_{32}) = 0 + 2 \times 5 + 0 = 10$
- ▷ $1\det(\mathbb{A}_{13}) + 0\det(\mathbb{A}_{23}) + 5\det(\mathbb{A}_{33}) = 0 + 0 + 5 \times 2 = 10$

Déterminant d'une matrice d'ordre 2

Soit $\mathbb{A}$ une matrice carrée d'ordre $n = 2$.

$$\det(\mathbb{A}) = \det\begin{pmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{pmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$



Exemple

Soit la matrice

$$\mathbb{A} = \begin{pmatrix}5 & 7\\ 4 & 3\end{pmatrix}$$

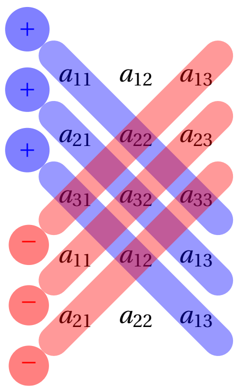
alors

$$\det(\mathbb{A}) = 5 \times 3 - 7 \times 4 = 15 - 28 = -13.$$

Règle de Sarrus

Soit $\mathbb{A}$ une matrice carrée d'ordre $n = 3$.

$$\det(\mathbb{A}) = \det\begin{pmatrix}a_{11} & a_{12} & a_{13}\\ a_{21} & a_{22} & a_{23}\\ a_{31} & a_{32} & a_{33}\end{pmatrix} = \big(a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{33}\big) - \big(a_{13}a_{22}a_{31} + a_{23}a_{32}a_{11} + a_{33}a_{12}a_{31}\big)$$



Exemple

Soit la matrice

$$\mathbb{A} = \begin{pmatrix}1 & 0 & 1\\ 0 & 2 & 0\\ 0 & 3 & 5\end{pmatrix}$$

alors

$$\det(\mathbb{A}) = (1 \times 2 \times 5 + 0 \times 0 \times 0 + 0 \times 0 \times 0) - (1 \times 2 \times 0 + 0 \times 3 \times 1 + 5 \times 0 \times 0) = 10.$$

Exemple

Soit la matrice

$$\mathbb{A}=\begin{pmatrix}5&7&1\\4&3&2\\2&1&6\end{pmatrix}$$

alors

$$\det(\mathbb{A})=(5\times3\times6+4\times1\times1+2\times7\times2)-(1\times3\times2+2\times1\times5+7\times4\times6)=-62.$$

Propriété

Le déterminant d'une matrice triangulaire est égal au produit des éléments diagonaux.

Théorème

$\mathbb{A}$ est inversible si et seulement si $\det(\mathbb{A})\neq 0$.

Propriété

- ▷ $\det(\mathbb{A}^T)=\det(\mathbb{A})$,
- ▷ $\det(\mathbb{A}^{-1})=\frac{1}{\det(\mathbb{A})}$,
- ▷ $\det(\mathbb{A}\times\mathbb{B})=\det(\mathbb{A})\cdot\det(\mathbb{B})$.

Définition : rang

Le RANG d'une matrice quelconque $\mathbb{A}$ est égal au plus grand entier s tel que l'on puisse extraire de $\mathbb{A}$ une matrice carrée d'ordre s inversible, c'est-à-dire de déterminant non nul. Les opérations élémentaires sur les lignes ou les colonnes d'une matrice ne modifient pas le rang.

Exemple

Soit $\mathbb{A}$ et $\mathbb{B}$ les deux matrices suivantes

$$\mathbb{A}=\begin{pmatrix}1&3&2\\1&3&1\end{pmatrix}\qquad\mathbb{B}=\begin{pmatrix}1&0&1\\0&5&-1\\-1&0&-1\end{pmatrix}.$$

Le rang de $\mathbb{A}$ est 2 car

- ▷ $\mathbb{A}$ est d'ordre 2×3 donc $s\leq\min\{2,3\}$ donc $s=0$, 1 ou 2 ;
- ▷ comme le déterminant de la sous-matrice composée de la première et de la deuxième colonne est nul, on ne peut pas conclure ;
- ▷ comme le déterminant de la sous-matrice composée de la première et de la troisième colonne est non nul, alors $s=2$.

Le rang de $\mathbb{B}$ est 2 car

- ▷ $\mathbb{B}$ est d'ordre 3×3 donc $s\leq3$
- ▷ le déterminant de $\mathbb{B}$ est 0 donc $s\neq3$
- ▷ le déterminant de la sous-matrice $\begin{pmatrix}1&0\\0&5\end{pmatrix}$ est 5, donc $s=2$.

A.4. Systèmes linéaires

Définition : système linéaire

Soit $n,p,\geq1$ des entiers. Un SYSTÈME LINÉAIRE $n\times p$ est un ensemble de n équations linéaires à p inconnues de la forme

$$(S)\quad\begin{cases}a_{11}x_1 & + & \dots & + & a_{1p}x_p & = & b_1,\\ \vdots & & & & \vdots & & \vdots\\ a_{n1}x_1 & + & \dots & + & a_{np}x_p & = & b_n.\end{cases}$$

▷ Les COEFFICIENTS a_{ij} et les SECONDES MEMBRES b_i sont des éléments donnés de $\mathbb{R}$. Les INCONNUES $x_1,x_2,\dots,x_p$ sont à chercher dans $\mathbb{R}$.

- ▷ Le SYSTÈME HOMOGÈNE associé à (S) est le système obtenu en remplaçant les b_i par 0.
- ▷ Une SOLUTION de (S) est un p-uplet $(x_1, x_2, \dots, x_p)$ qui vérifie simultanément les n équations de (S). Résoudre (S) signifie chercher toutes les solutions.
- ▷ Un système est IMPOSSIBLE, ou incompatible, s'il n'admet pas de solution. Un système est POSSIBLE, ou compatible, s'il admet une ou plusieurs solutions.
- ▷ Deux systèmes sont ÉQUIVALENTS s'ils admettent les mêmes solutions.

Écriture matricielle

Si on note

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \qquad \mathbb{A} = \begin{pmatrix} a_{11} & \dots & a_{1p} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{np} \end{pmatrix}$$

le système (S) est équivalent à l'écriture matricielle $\mathbb{A}\mathbf{x} = \mathbf{b}$.

Définition : système échelonné

Un système (S) est EN ESCALIER, ou ÉCHELONNÉ, si le nombre de premiers coefficients nuls successifs de chaque équation est strictement croissant.

Autrement dit, un système est échelonné si les coefficients non nuls des équations se présentent avec une sorte d'escalier à marches de longueurs variables marquant la séparation entre une zone composée uniquement de zéros et une zone où les lignes situées à droite de l'escalier commencent par des termes non nuls, comme dans l'exemple suivant de 5 équations à 6 inconnues :

$$\begin{cases} 5x_1 & -x_2 & -x_3 & +2x_4 & & +x_6 & = & b_1 \\ & & 3x_3 & -x_4 & +2x_5 & & = & b_2 \\ & & & & -x_5 & +x_6 & = & b_3 \\ & & & & & 5x_6 & = & b_4 \\ & & & & & 0 & = & b_5 \end{cases}$$

Réduction

Quand un système contient une équation du type

$$0x_1 + \cdots + 0x_p = b,$$

- ▷ si $b \neq 0$ le système est impossible,
- ▷ si $b = 0$, on peut supprimer cette équation, ce qui conduit à un système équivalent à (S) dit SYSTÈME RÉDUIT.

Définition : matrice augmentée

Si on ajoute le vecteur-colonne des seconds membres $\mathbf{b}$ à la matrice des coefficients $\mathbb{A}$, on obtient ce qu'on appelle la matrice augmentée que l'on note $[\mathbb{A}|\mathbf{b}]$.

Méthode du pivot de Gauss

Soit $\mathbb{A} = (a_{ij})_{\substack{1\le i\le n\\ 1\le j\le p}}$ la matrice des coefficients du système (S). En permutant éventuellement deux lignes du système, on peut supposer $a_{ii} \neq 0$ (appelé pivot de l'étape i).
Étape j : pour $i > j$, les transformations

$$L_i \leftarrow L_i - \frac{a_{ij}}{a_{jj}} L_j$$

éliminent l'inconnue x_j dans les lignes L_i.
En réitérant le procédé pour i de 1 à n, on aboutit à un système échelonné.

Exemple

Soit le système linéaire

$$\begin{cases} x_1 + 2x_2 + 3x_3 + 4x_4 = 1,\\ 2x_1 + 3x_2 + 4x_3 + x_4 = 2,\\ 3x_1 + 4x_2 + x_3 + 2x_4 = 3,\\ 4x_1 + x_2 + 2x_3 + 3x_4 = 4. \end{cases}$$

1. Résolution par la méthode du pivot de Gauss :

$$\begin{cases} x_1+2x_2+3x_3+4x_4=1\\ 2x_1+3x_2+4x_3+x_4=2\\ 3x_1+4x_2+x_3+2x_4=3\\ 4x_1+x_2+2x_3+3x_4=4 \end{cases} \xrightarrow[L_4\leftarrow L_4-4L_1]{\substack{L_2\leftarrow L_2-2L_1\\ L_3\leftarrow L_3-3L_1}} \begin{cases} x_1+2x_2+3x_3+4x_4=1\\ -x_2-2x_3-7x_4=0\\ -2x_2-8x_3-10x_4=0\\ -7x_2-10x_3-13x_4=0 \end{cases}$$

$$\xrightarrow{\substack{L_3\leftarrow L_3-2L_2\\ L_4\leftarrow L_4-7L_2}} \begin{cases} x_1+2x_2+3x_3+4x_4=1\\ -x_2-2x_3-7x_4=0\\ -4x_3+4x_4=0\\ 4x_3+36x_4=0 \end{cases} \xrightarrow{L_4\leftarrow L_4+L_3} \begin{cases} x_1+2x_2+3x_3+4x_4=1\\ -x_2-2x_3-7x_4=0\\ -4x_3+4x_4=0\\ 40x_4=0 \end{cases}$$

donc $x_4=0$, $x_3=0$, $x_2=0$ et $x_1=1$.

2. Résolution par la méthode du pivot de Gauss en écriture matricielle :

$$[\mathbb{A}|\mathbf{b}] = \left(\begin{array}{cccc|c} 1&2&3&4&1\\ 2&3&4&1&2\\ 3&4&1&2&3\\ 4&1&2&3&4 \end{array}\right) \xrightarrow{\substack{L_2\leftarrow L_2-2L_1\\ L_3\leftarrow L_3-3L_1\\ L_4\leftarrow L_4-4L_1}} \left(\begin{array}{cccc|c} 1&2&3&4&1\\ 0&-1&-2&-7&0\\ 0&-2&-8&-10&0\\ 0&-7&-10&-13&0 \end{array}\right)$$

$$\xrightarrow{\substack{L_3\leftarrow L_3-2L_2\\ L_4\leftarrow L_4-7L_2}} \left(\begin{array}{cccc|c} 1&2&3&4&1\\ 0&-1&-2&-7&0\\ 0&0&-4&4&0\\ 0&0&4&36&0 \end{array}\right) \xrightarrow{L_4\leftarrow L_4+L_3} \left(\begin{array}{cccc|c} 1&2&3&4&1\\ 0&-1&-2&-7&0\\ 0&0&-4&4&0\\ 0&0&0&40&0 \end{array}\right)$$

donc $x_4=0$, $x_3=0$, $x_2=0$ et $x_1=1$.

Exemple

On veut résoudre le système linéaire suivant par la méthode de Gauss.

$$\begin{cases} x_1+2x_2-x_3=2,\\ x_1-2x_2-3x_3=-6,\\ x_1+4x_2+4x_3=3. \end{cases} \xrightarrow{\substack{L_2\leftarrow L_2-L_1\\ L_3\leftarrow L_3-L_1}} \begin{cases} x_1+2x_2-x_3=2,\\ -4x_2-2x_3=-8,\\ 2x_2+5x_3=1. \end{cases} \xrightarrow{L_3\leftarrow L_3+L_2/2} \begin{cases} x_1+2x_2-x_3=2,\\ -4x_2-2x_3=-8,\\ 4x_3=-3. \end{cases}$$

donc $x_3=\frac{-3}{4}$, $x_2=\frac{19}{8}$ et $x_1=\frac{-7}{2}$.

Exemple

On veut résoudre le système linéaire suivant par la méthode de Gauss

$$\begin{cases} -x_1+x_2+3x_3=12\\ 2x_1-x_2+2x_3=-8\\ 4x_1+x_2-4x_3=15 \end{cases} \xrightarrow{\substack{L_2\leftarrow L_2+2L_1\\ L_3\leftarrow L_3+4L_1}} \begin{cases} -x_1+x_2+3x_3=12\\ x_2+8x_3=16\\ 5x_2+8x_3=63 \end{cases} \xrightarrow{L_3\leftarrow L_3-5L_2} \begin{cases} -x_1+x_2+3x_3=12\\ x_2+8x_3=16\\ -32x_3=-17 \end{cases}$$

donc $x_3=\frac{17}{32}$, $x_2=\frac{47}{4}$ et $x_1=\frac{43}{32}$.

Exemple

On utilise la méthode du pivot de Gauss pour résoudre le système linéaire suivant :

$$\begin{cases} -2u-4v+3w=-1\\ 2v-w=1\\ u+v-3w=-6 \end{cases} \xrightarrow{L_3\leftarrow L_3+L_1/2} \begin{cases} -2u-4v+3w=-1\\ 2v-w=1\\ -v-\frac{3}{2}w=-{}^{13}/_{2} \end{cases} \xrightarrow{L_3\leftarrow L_3+L_2/2} \begin{cases} -2u-4v+3w=-1\\ 2v-w=1\\ -2w=-6 \end{cases}$$

donc $w=3$, $v=2$ et $u=1$.

Exemple

On utilise la méthode du pivot de Gauss pour résoudre le système linéaire :

$$\begin{cases} -2x-y+4t=2\\ 2x+3y+3z+2t=14\\ x+2y+z+t=7\\ -x-z+t=-1 \end{cases} \xrightarrow[L_4\leftarrow L_4-\frac{1}{2}L_1]{\substack{L_2\leftarrow L_2+L_1\\ L_3\leftarrow L_3+\frac{1}{2}L_1}} \begin{cases} -2x-y+4t=2\\ 2y+3z+6t=16\\ \frac{3}{2}y+z+3t=8\\ \frac{1}{2}y-z-t=-2 \end{cases} \xrightarrow[L_4\leftarrow L_2-\frac{1}{4}L_2]{L_3\leftarrow L_3-\frac{3}{4}L_2} \begin{cases} -2x-y+4t=2\\ 2y+3z+6t=16\\ -\frac{5}{4}z-\frac{3}{2}t=-4\\ -\frac{7}{4}z-\frac{5}{2}t=-6 \end{cases} \xrightarrow{L_4\leftarrow L_4-\frac{7}{5}L_3} \begin{cases} -2x-y+4t=2\\ 2y+3z+6t=16\\ -\frac{5}{4}z-\frac{3}{2}t=-4\\ -\frac{2}{5}t=-\frac{2}{5} \end{cases}$$

donc $t=4$, $z=2$, $y=2$ et $x=0$.

Exemple

On veut résoudre le système linéaire

$$\begin{pmatrix}1&2&3&4\\2&3&4&1\\3&4&1&2\\4&1&2&3\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}=\begin{pmatrix}10\\10\\10\\10\end{pmatrix}.$$

par la méthode du pivot de Gauss.

$$[\mathbb{A}|\mathbf{b}]=\left(\begin{array}{cccc|c}1&2&3&4&10\\2&3&4&1&10\\3&4&1&2&10\\4&1&2&3&10\end{array}\right)\xrightarrow{\substack{L_2\leftarrow L_2-2L_1\\ L_3\leftarrow L_3-3L_1\\ L_4\leftarrow L_4-4L_1}}\left(\begin{array}{cccc|c}1&2&3&4&10\\0&-1&-2&-7&-10\\0&-2&-8&-10&-20\\0&-7&-10&-13&-30\end{array}\right)$$

$$\xrightarrow{\substack{L_3\leftarrow L_3-2L_2\\ L_4\leftarrow L_4-7L_2}}\left(\begin{array}{cccc|c}1&2&3&4&10\\0&-1&-2&-7&-10\\0&0&-4&4&0\\0&0&4&36&40\end{array}\right)\xrightarrow{L_4\leftarrow L_4+L_3}\left(\begin{array}{cccc|c}1&2&3&4&10\\0&-1&-2&-7&-10\\0&0&-4&4&0\\0&0&0&40&40\end{array}\right)$$

donc

$$\begin{cases} x_1+2x_2+3x_3+4x_4=10\\ -x_2-2x_3-7x_4=-10\\ -4x_3+4x_4=0\\ 40x_4=40 \end{cases} \implies x_4=1,\quad x_3=1,\quad x_2=1,\quad x_1=1.$$

Exemple

On veut résoudre le système linéaire suivant par la méthode du pivot de Gauss :

$$\begin{pmatrix}6&1&1\\2&4&0\\1&2&6\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}12\\0\\6\end{pmatrix}.$$

$$[\mathbb{A}|\mathbf{b}]=\left(\begin{array}{ccc|c}6&1&1&12\\2&4&0&0\\1&2&6&6\end{array}\right)\xrightarrow{\substack{L_2\leftarrow L_2-\frac{2}{6}L_1\\ L_3\leftarrow L_3-\frac{1}{6}L_1}}\left(\begin{array}{ccc|c}6&1&1&12\\0&\frac{11}{3}&-\frac{1}{3}&-4\\0&\frac{11}{6}&\frac{35}{6}&4\end{array}\right)\xrightarrow{L_3\leftarrow L_3-\frac{\frac{11}{6}}{\frac{11}{3}}L_2}\left(\begin{array}{ccc|c}6&1&1&12\\0&\frac{11}{3}&-\frac{1}{3}&-4\\0&0&6&6\end{array}\right)$$

donc

$$\begin{cases} 6x_1+x_2+x_3=12,\\ \frac{11}{3}x_2-\frac{1}{3}x_3=-4\\ 6x_3=6 \end{cases} \implies x_3=1,\quad x_2=-1,\quad x_1=2.$$

Exemple

On veut résoudre les systèmes linéaires suivants :

$$\begin{cases} x-5y-7z=3\\ 2x-13y-18z=3\\ 3x-27y-36z=3 \end{cases} \quad\text{et}\quad \begin{cases} x-5y-7z=6\\ 2x-13y-18z=0\\ 3x-27y-36z=-3 \end{cases} \quad\text{et}\quad \begin{cases} x-5y-7z=0\\ 2x-13y-18z=3\\ 3x-27y-36z=6. \end{cases}$$

Le trois systèmes s'écrivent sous forme matricielle

$$\begin{pmatrix}1&-5&-7\\2&-13&-18\\3&-27&-36\end{pmatrix}\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}3\\3\\3\end{pmatrix}\quad\text{et}\quad\begin{pmatrix}1&-5&-7\\2&-13&-18\\3&-27&-36\end{pmatrix}\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}6\\0\\-3\end{pmatrix}\quad\text{et}\quad\begin{pmatrix}1&-5&-7\\2&-13&-18\\3&-27&-36\end{pmatrix}\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}0\\3\\6\end{pmatrix}$$

On remarque que seul le seconde membre change. On calcule alors d'abord la décomposition $\mathbb{LU}$ de la matrice $\mathbb{A}$:

$$\begin{pmatrix}1&-5&-7\\2&-13&-18\\3&-27&-36\end{pmatrix}\xrightarrow[L_3\leftarrow L_3-3L_1]{L_2\leftarrow L_2-2L_1}\begin{pmatrix}1&-5&-7\\0&-3&-4\\0&-12&-15\end{pmatrix}\xrightarrow{L_3\leftarrow L_3-4L_2}\begin{pmatrix}1&-5&-7\\0&-3&-4\\0&0&1\end{pmatrix}$$

donc

$$\mathbb{L}=\begin{pmatrix}1&0&0\\2&1&0\\3&4&1\end{pmatrix}\qquad\qquad\mathbb{U}=\begin{pmatrix}1&-5&-7\\0&-3&-4\\0&0&1\end{pmatrix}$$

Pour résoudre chaque système linéaire on résout les systèmes triangulaires $\mathbb{L}\mathbf{y}=\mathbf{b}$ et $\mathbb{U}\mathbf{x}=\mathbf{y}$.

1. Pour le premier système on a

$$\begin{pmatrix}1&0&0\\2&1&0\\3&4&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix}=\begin{pmatrix}3\\3\\3\end{pmatrix}\qquad\Longrightarrow\qquad y_1=3,\quad y_2=-3,\quad y_3=6;$$

$$\begin{pmatrix}1&-5&-7\\0&-3&-4\\0&0&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}3\\-3\\6\end{pmatrix}\qquad\Longrightarrow\qquad x_3=6,\quad x_2=-7,\quad x_1=10.$$

2. Pour le seconde système on a

$$\begin{pmatrix}1&0&0\\2&1&0\\3&4&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix}=\begin{pmatrix}6\\0\\-3\end{pmatrix}\qquad\Longrightarrow\qquad y_1=6,\quad y_2=-12,\quad y_3=27;$$

$$\begin{pmatrix}1&-5&-7\\0&-3&-4\\0&0&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}6\\-12\\27\end{pmatrix}\qquad\Longrightarrow\qquad x_3=27,\quad x_2=-32,\quad x_1=35.$$

3. Pour le dernier système on a

$$\begin{pmatrix}1&0&0\\2&1&0\\3&4&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix}=\begin{pmatrix}0\\3\\6\end{pmatrix}\qquad\Longrightarrow\qquad y_1=0,\quad y_2=3,\quad y_3=-6;$$

$$\begin{pmatrix}1&-5&-7\\0&-3&-4\\0&0&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}6\\-12\\27\end{pmatrix}\qquad\Longrightarrow\qquad x_3=-6,\quad x_2=7,\quad x_1=-7.$$

Exemple

Soit le système linéaire

$$\begin{pmatrix}6&1&1\\2&4&0\\1&2&6\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}12\\0\\6\end{pmatrix}.$$

1. Résoudre les systèmes linéaires par la méthode du pivot de Gauss.
2. Factoriser la matrice $\mathbb{A}$ (sans utiliser la technique du pivot) et résoudre le système linéaire.

1. Méthode du pivot de Gauss :

$$[\mathbb{A}|\mathbf{b}]=\left(\begin{array}{ccc|c}6&1&1&12\\2&4&0&0\\1&2&6&6\end{array}\right)\xrightarrow[L_3\leftarrow L_3-\frac{1}{6}L_1]{L_2\leftarrow L_2-\frac{2}{6}L_1}\left(\begin{array}{ccc|c}6&1&1&12\\0&\frac{11}{3}&-\frac{1}{3}&-4\\0&\frac{11}{6}&\frac{35}{6}&4\end{array}\right)\xrightarrow{L_3\leftarrow L_3-\frac{\frac{11}{6}}{\frac{11}{3}}L_2}\left(\begin{array}{ccc|c}6&1&1&12\\0&\frac{11}{3}&-\frac{1}{3}&-4\\0&0&6&6\end{array}\right)$$

donc

$$\begin{cases}6x_1+x_2+x_3=12,\\\frac{11}{3}x_2-\frac{1}{3}x_3=-4\\6x_3=6\end{cases}\qquad\Longrightarrow\qquad x_3=1,\quad x_2=-1,\quad x_1=2.$$

2. Factorisation de la matrice $\mathbb{A}$:

$$\begin{pmatrix}6&1&1\\2&4&0\\1&2&6\end{pmatrix}\xrightarrow[L_3\leftarrow L_3-\frac{1}{6}L_1]{L_2\leftarrow L_2-\frac{2}{6}L_1}\begin{pmatrix}6&1&1\\\frac{2}{6}&\frac{11}{3}&-\frac{1}{3}\\\frac{1}{6}&\frac{11}{6}&\frac{35}{6}\end{pmatrix}\xrightarrow{L_3\leftarrow L_3-\frac{\frac{11}{6}}{\frac{11}{3}}L_2}\begin{pmatrix}6&1&1\\\frac{2}{6}&\frac{11}{3}&-\frac{1}{3}\\\frac{1}{6}&\frac{\frac{11}{6}}{\frac{11}{3}}&6\end{pmatrix}$$

donc

$$\mathbb{L}=\begin{pmatrix}1&0&0\\\frac{1}{3}&1&0\\\frac{1}{6}&\frac{1}{2}&1\end{pmatrix}\qquad\qquad\mathbb{U}=\begin{pmatrix}6&1&1\\0&\frac{11}{3}&-\frac{1}{3}\\0&0&6\end{pmatrix}$$

Pour résoudre le système linéaire on résout les systèmes triangulaires $\mathbb{L}\mathbf{y}=\mathbf{b}$

$$\begin{pmatrix}1&0&0\\\frac{1}{3}&1&0\\\frac{1}{6}&\frac{1}{2}&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix}=\begin{pmatrix}12\\0\\6\end{pmatrix}\qquad\Longrightarrow\qquad y_1=12,\quad y_2=-4,\quad y_3=6$$

et $\mathbb{U}\mathbf{x}=\mathbf{y}$

$$\begin{pmatrix}6&1&1\\0&\frac{11}{3}&-\frac{1}{3}\\0&0&6\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}1\\-4\\6\end{pmatrix}\qquad\Longrightarrow\qquad x_3=1,\quad x_2=-1,\quad x_1=2.$$

Méthode de Gauss-Jordan

Dans cette variante de la méthode du pivot de Gauss, à chaque étape on fait apparaître des zéros à la fois au-dessus et en-dessous du pivot.
Étape j : pour $i\neq j$, les transformations

$$L_i\leftarrow L_i-\frac{a_{ij}}{a_{jj}}L_j$$

éliminent l'inconnue x_j dans les lignes L_i.
En réitérant le procédé pour i de 1 à n, on aboutit à un système diagonal.

Exemple

Résoudre le système linéaire

$$\begin{pmatrix}1&2&3&4\\2&3&4&1\\3&4&1&2\\4&1&2&3\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}=\begin{pmatrix}1\\2\\3\\4\end{pmatrix}$$

par la méthode de Gauss-Jordan.

$$[\mathbb{A}|\mathbf{b}]=\left(\begin{array}{cccc|c}1&2&3&4&1\\2&3&4&1&2\\3&4&1&2&3\\4&1&2&3&4\end{array}\right)\xrightarrow[\substack{L_3\leftarrow L_3-3L_1\\L_4\leftarrow L_4-4L_1}]{L_2\leftarrow L_2-2L_1}\left(\begin{array}{cccc|c}1&2&3&4&1\\0&-1&-2&-7&0\\0&-2&-8&-10&0\\0&-7&-10&-13&0\end{array}\right)\xrightarrow[\substack{L_3\leftarrow L_3-2L_2\\L_4\leftarrow L_4-7L_2}]{L_1\leftarrow L_1+2L_2}\left(\begin{array}{cccc|c}1&0&-1&-10&1\\0&-1&-2&-7&0\\0&0&-4&4&0\\0&0&4&36&0\end{array}\right)$$

$$\xrightarrow[\substack{L_2\leftarrow L_2-L_3/2\\L_4\leftarrow L_4+L_3}]{L_1\leftarrow L_1-L_3/4}\left(\begin{array}{cccc|c}1&0&0&4&1\\0&-1&0&-7&0\\0&0&-4&4&0\\0&0&0&40&0\end{array}\right)\xrightarrow[\substack{L_2\leftarrow L_2+9L_4/40\\L_3\leftarrow L_3+4L_4/40}]{L_1\leftarrow L_1+11L_4/40}\left(\begin{array}{cccc|c}1&0&0&0&1\\0&-1&0&0&0\\0&0&-4&0&0\\0&0&0&40&0\end{array}\right)$$

donc

$$x_1=1,\quad x_2=0,\quad x_3=0,\quad x_4=0.$$

Définition : système de Cramer

Un SYSTÈME est dit DE CRAMER s'il a une solution, et une seule.

Propriété

Considérons un système carré d'ordre n à coefficients réels. Le système est de Cramer si une des conditions équivalentes suivantes est remplie :

1. $\mathbb{A}$ est inversible ;
2. $\text{rg}(\mathbb{A}) = n$;
3. le système homogène $\mathbb{A}\mathbf{x} = \mathbf{0}$ admet seulement la solution nulle.

Méthode de Cramer

La solution d'un système de Cramer d'écriture matricielle $\mathbb{A}\mathbf{x} = \mathbf{b}$ est donnée par

$$x_j = \frac{\det(\mathbb{A}_j)}{\det(\mathbb{A})}, \qquad 1 \le j \le n$$

où $\mathbb{A}_j$ est la matrice obtenue à partir de $\mathbb{A}$ en remplaçant la j-ème colonne par la colonne des seconds membres $\mathbf{b}$.

Cette formule est cependant d'une utilité pratique limitée à cause du calcul des déterminants qui est très couteux.

Exemple

On veut résoudre le système linéaire

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

par la méthode de Cramer. On a

$$\mathbb{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \qquad \det(\mathbb{A}) = a_{11}a_{22} - a_{12}a_{21},$$

$$\mathbb{A}_1 = \begin{pmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{pmatrix}, \qquad \det(\mathbb{A}_1) = b_1a_{22} - a_{12}b_2,$$

$$\mathbb{A}_2 = \begin{pmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{pmatrix}, \qquad \det(\mathbb{A}_2) = a_{11}b_2 - b_1a_{21},$$

donc

$$x_1 = \frac{b_1a_{22} - a_{12}b_2}{a_{11}a_{22} - a_{12}a_{21}}, \qquad x_2 = \frac{a_{11}b_2 - b_1a_{21}}{a_{11}a_{22} - a_{12}a_{21}}.$$

Exemple

On veut résoudre par la méthode de Cramer le système linéaire

$$\begin{pmatrix} 1 & -1 & 2 \\ 2 & 1 & 0 \\ 3 & 2 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}$$

On a

$$\mathbb{A} = \begin{pmatrix} 1 & -1 & 2 \\ 2 & 1 & 0 \\ 3 & 2 & 0 \end{pmatrix}, \qquad \det(\mathbb{A}) = 2,$$

$$\mathbb{A}_1 = \begin{pmatrix} 2 & -1 & 2 \\ -1 & 1 & 0 \\ 1 & 2 & 0 \end{pmatrix}, \quad \det(\mathbb{A}_1) = -6; \quad \mathbb{A}_2 = \begin{pmatrix} 1 & 2 & 2 \\ 2 & -1 & 0 \\ 3 & 1 & 0 \end{pmatrix}, \quad \det(\mathbb{A}_2) = 10; \quad \mathbb{A}_3 = \begin{pmatrix} 1 & -1 & 2 \\ 2 & 1 & -1 \\ 3 & 2 & 1 \end{pmatrix}, \quad \det(\mathbb{A}_3) = 10,$$

donc

$$x = \frac{-6}{2} = -3, \qquad y = \frac{10}{2} = 5, \qquad z = \frac{10}{2} = 5.$$

Définition : cofacteur

Soit $\mathbb{A}$ une matrice carrée d'ordre n. Étant donné un couple (i, j) d'entiers, $1 \le i, j \le n$, on note $\mathbb{A}_{ij}$ la matrice carrée d'ordre $n-1$ obtenue en supprimant la i-ème ligne et la j-ème colonne de $\mathbb{A}$. On appelle COFACTEUR de l'élément a_{ij} le nombre $(-1)^{i+j}\det(\mathbb{A}_{ij})$.

Exemple

Soit $\mathbb{A}=\begin{pmatrix} a & b\\ c & d\end{pmatrix}$. Alors la matrice des cofacteurs de $\mathbb{A}$ est la matrice $\begin{pmatrix} d & -c\\ -b & a\end{pmatrix}$.

Calcul de $\mathbb{A}^{-1}$

$\mathbb{A}$ étant inversible, pour obtenir $\mathbb{A}^{-1}$ il suffit de résoudre le système $\mathbb{A}\mathbf{x}=\mathbf{b}$ qui admet pour solution $\mathbf{x}=\mathbb{A}^{-1}\mathbf{b}$. On peut alors calculer $\mathbb{A}^{-1}$ en résolvant n systèmes linéaires de termes sources $(1,0,0,\dots,0)$, $(0,1,0,\dots,0)$, …, $(0,0,0,\dots,1)$. Les méthodes suivantes résolvent ces n systèmes linéaires simultanément.

Première méthode.

1. On calcul la matrice des cofacteurs des éléments de $\mathbb{A}$, appelée comatrice de $\mathbb{A}$;
2. on transpose la comatrice de $\mathbb{A}$;
3. on divise par $\det(\mathbb{A})$.

Cette méthode est quasi-impraticable dès que $n>3$.

Deuxième méthode.

La matrice $\mathbb{A}$ est inversible si et seulement si on obtient par opérations élémentaires sur les lignes de $\mathbb{A}$ une matrice triangulaire sans zéros sur la diagonale ; non inversible si et seulement si on obtient une matrice triangulaire avec un zéro sur la diagonale. Si $\mathbb{A}$ est inversible, on effectue les mêmes opérations sur la matrice $[\mathbb{A}|\mathbb{I}_n]$ jusqu'à obtenir $[\mathbb{I}_n|\mathbb{A}^{-1}]$:

$$[\mathbb{A}|\mathbb{I}_n]\xrightarrow{\text{Opérations élémentaires}}[\mathbb{I}_n|\mathbb{A}^{-1}].$$

Exemple

Soit $\mathbb{A}=\begin{pmatrix} a & b\\ c & d\end{pmatrix}$ avec $\det(\mathbb{A})=ad-bc\neq 0$. Alors $\mathbb{A}^{-1}=\frac{1}{ad-bc}\begin{pmatrix} d & -b\\ -c & a\end{pmatrix}$.

Exemple

On veut calculer l'inverse de la matrice $\mathbb{A}=\begin{pmatrix} 1 & 1 & -1\\ -1 & 1 & 1\\ 1 & -1 & 1\end{pmatrix}$.

Première méthode.

1. On calcule la matrice des cofacteurs des éléments de $\mathbb{A}$, appelée comatrice de $\mathbb{A}$:

$$\text{comatrice}=\begin{pmatrix} (-1)^{1+1}\begin{vmatrix} 1 & 1\\ -1 & 1\end{vmatrix} & (-1)^{1+2}\begin{vmatrix} -1 & 1\\ 1 & 1\end{vmatrix} & (-1)^{1+3}\begin{vmatrix} -1 & 1\\ 1 & -1\end{vmatrix}\\ (-1)^{2+1}\begin{vmatrix} 1 & -1\\ -1 & 1\end{vmatrix} & (-1)^{2+2}\begin{vmatrix} 1 & -1\\ 1 & 1\end{vmatrix} & (-1)^{2+3}\begin{vmatrix} 1 & 1\\ 1 & -1\end{vmatrix}\\ (-1)^{3+1}\begin{vmatrix} 1 & -1\\ 1 & 1\end{vmatrix} & (-1)^{3+2}\begin{vmatrix} 1 & -1\\ -1 & 1\end{vmatrix} & (-1)^{3+3}\begin{vmatrix} 1 & 1\\ -1 & 1\end{vmatrix}\end{pmatrix}=\begin{pmatrix} 2 & 2 & 0\\ 0 & 2 & 2\\ 2 & 0 & 2\end{pmatrix};$$

2. on transpose la comatrice de $\mathbb{A}$:

$$\text{comatrice}^T=\begin{pmatrix} 2 & 0 & 2\\ 2 & 2 & 0\\ 0 & 2 & 2\end{pmatrix};$$

3. on divise par $\det(\mathbb{A})$:

$$\mathbb{A}^{-1}=\frac{1}{4}\begin{pmatrix} 2 & 0 & 2\\ 2 & 2 & 0\\ 0 & 2 & 2\end{pmatrix}=\begin{pmatrix} 1/2 & 0 & 1/2\\ 1/2 & 1/2 & 0\\ 0 & 1/2 & 1/2\end{pmatrix}.$$

Deuxième méthode.

$$
[\mathbb{A}|\mathbb{I}_3]=\left(\begin{array}{ccc|ccc}1&1&-1&1&0&0\\-1&1&1&0&1&0\\1&-1&1&0&0&1\end{array}\right)\xrightarrow[L_3\leftarrow L_3-L_1]{L_2\leftarrow L_2+L_1}\left(\begin{array}{ccc|ccc}1&1&-1&1&0&0\\\boxed{0}&2&0&1&1&0\\\boxed{0}&-2&2&-1&0&1\end{array}\right)
$$

$$
\xrightarrow{L_2\leftarrow L_2/2}\left(\begin{array}{ccc|ccc}1&1&-1&1&0&0\\0&\boxed{1}&0&1/2&1/2&0\\0&-2&2&-1&0&1\end{array}\right)\xrightarrow[L_3\leftarrow L_3+2L_2]{L_1\leftarrow L_1-L_2}\left(\begin{array}{ccc|ccc}1&\boxed{0}&-1&1/2&-1/2&0\\0&1&0&1/2&1/2&0\\0&\boxed{0}&2&0&1&1\end{array}\right)
$$

$$
\xrightarrow{L_3\leftarrow L_3/2}\left(\begin{array}{ccc|ccc}1&0&-1&1/2&-1/2&0\\0&1&0&1/2&1/2&0\\0&0&\boxed{1}&0&1/2&1/2\end{array}\right)\xrightarrow{L_1\leftarrow L_1+L_3}\left(\begin{array}{ccc|ccc}1&0&\boxed{0}&1/2&0&1/2\\0&1&\boxed{0}&1/2&1/2&0\\0&0&1&0&1/2&1/2\end{array}\right)=[\mathbb{I}_3|\mathbb{A}^{-1}].
$$

Exemple

On veut calculer $\mathbb{A}^{-1}$ où $\mathbb{A}$ est la matrice $\begin{pmatrix}1&0&-1\\4&-1&-2\\-2&0&1\end{pmatrix}$.

$$
[\mathbb{A}|\mathbb{I}_3]=\left(\begin{array}{ccc|ccc}1&0&-1&1&0&0\\4&-1&-2&0&1&0\\-2&0&1&0&0&1\end{array}\right)\xrightarrow[L_3\leftarrow L_3+2L_1]{L_2\leftarrow L_2-4L_1}\left(\begin{array}{ccc|ccc}1&0&-1&1&0&0\\0&-1&2&-4&1&0\\0&0&-1&-2&0&1\end{array}\right)
$$

$$
\xrightarrow[L_3\leftarrow -L_3]{L_2\leftarrow -L_2}\left(\begin{array}{ccc|ccc}1&0&-1&1&0&0\\0&1&-2&4&-1&0\\0&0&1&2&0&-1\end{array}\right)\xrightarrow[L_2\leftarrow L_2+2L_3]{L_1\leftarrow L_1+L_3}\left(\begin{array}{ccc|ccc}1&0&0&-1&0&-1\\0&1&0&0&-1&-2\\0&0&1&-2&0&-1\end{array}\right)=[\mathbb{I}_3|\mathbb{A}^{-1}].
$$

Exemple

On veut calculer $\mathbb{A}^{-1}$ où $\mathbb{A}$ est la matrice $\begin{pmatrix}1&0&1\\0&1&2\\2&0&1\end{pmatrix}$.

$$
[\mathbb{A}|\mathbb{I}_3]=\left(\begin{array}{ccc|ccc}1&0&1&1&0&0\\0&1&2&0&1&0\\2&0&1&0&0&1\end{array}\right)\xrightarrow{L_3\leftarrow L_3-2L_1}\left(\begin{array}{ccc|ccc}1&0&1&1&0&0\\0&1&2&0&1&0\\0&0&-1&-2&0&1\end{array}\right)
$$

$$
\xrightarrow{L_3\leftarrow -L_3}\left(\begin{array}{ccc|ccc}1&0&1&1&0&0\\0&1&2&0&1&0\\0&0&1&2&0&-1\end{array}\right)\xrightarrow[L_2\leftarrow L_2-2L_3]{L_1\leftarrow L_1-L_3}\left(\begin{array}{ccc|ccc}1&0&0&-1&0&1\\0&1&0&-4&1&2\\0&0&1&2&0&-1\end{array}\right)=[\mathbb{I}_3|\mathbb{A}^{-1}].
$$

A.5. Équations différentielles d'ordre 1

Équations différentielles

Une équation différentielle (EDO) est une équation, dont l'inconnue est une fonction y, exprimée sous la forme d'une relation $F(y,y',y'',\dots,y^{(n)})=g(x)$ dans laquelle cohabitent à la fois $y=y(x)$ et ses dérivées y', y'', ... (n est appelé l'ordre de l'équation). Si la fonction g, appelée «second membre» de l'équation, est nulle, on dit que l'équation en question est homogène.

Problème de Cauchy et théorème de Cauchy-Lipschitz

Soient a et g deux fonctions continues d'un intervalle I dans $\mathbb{R}$ et $(x_0,y_0)\in I\times\mathbb{R}$. Considérons le problème de Cauchy

$$
\begin{cases}y'(x)+a(x)y(x)=g(x),\\y(x_0)=y_0.\end{cases}
$$

Alors il existe une unique solution y de l'équation différentielle telle que $y(x_0)=y_0$.

Remarque

Graphiquement, ce théorème signifie que par tout point du plan dont l'abscisse est dans I, il passe une courbe intégrale et une seule, autrement dit deux trajectoires ne peuvent pas se croiser. En particulier, si une équation différentielle admet comme solution la solution nulle, alors toute autre solution est soit toujours positive soit toujours négative.

Équations différentielles du premier ordre à variables séparables

Lorsque l'équation est de la forme

$$f(y(x))y'(x) = g(x)$$

où f et g sont des fonctions données dont on connaît des primitives F et G, on a

$$F(y(x)) = G(x) + C \qquad \text{où } C \in \mathbb{R},$$

et si F possède une fonction réciproque F^{-1}, on en déduit

$$y(x) = F^{-1}(G(x) + C),$$

relation qui donne toutes les solutions de l'équation. Cette solution générale dépend de la constante d'intégration C.

Astuce mnémotechnique

En pratique, on peut écrire l'équation sous la forme

$$f(y)\,\mathrm{d}y = g(x)\,\mathrm{d}x,$$

puis intégrer formellement les deux membres

$$\int f(y)\,\mathrm{d}y = \int g(x)\,\mathrm{d}x,$$

et exprimer y en fonction de x.

Exemple

On veut résoudre l'équation différentielle $y'(x) = xy(x)$ sur des intervalles à préciser. Il s'agit d'une EDO du premier ordre à variables séparables : la fonction $y(x) \equiv 0$ est solution, toute autre solution $y(x)$ sera donc non nulle. On peut alors réécrire l'EDO comme

$$\frac{y'(x)}{y(x)} = x \quad \Longrightarrow \quad \int \frac{1}{y}\,\mathrm{d}y = \int x\,\mathrm{d}x \quad \Longrightarrow \quad \ln|y| = \frac{x^2}{2} + C \text{ avec } C \in \mathbb{R}.$$

Ainsi, toute solution non nulle est

$$y(x) = De^{x^2/2} \qquad \text{avec } D \in \mathbb{R}.$$

Exemple

Le carbone 14 est un isotope présent dans tout organisme vivant. Le nombre d'atomes de carbone 14 est constant tant que l'organisme est en vie. À la mort de l'organisme, le nombre d'atomes décroît avec une vitesse proportionnelle au nombre d'atomes. On note $n(t) > 0$ le nombre d'atomes au temps t, exprimé en années, après la mort de l'organisme. Ce mécanisme se traduit par l'équation

$$n'(t) = -kn(t)$$

où k est une constante positive.

1. Il s'agit d'une «EDO du premier ordre à variables séparables». Si $n(t) \equiv c$ est solution alors $0 = -kc$ d'où $c = 0$: l'unique solution constante est la solution $n(t) = 0$ quelque soit $t \in \mathbb{R}^+$.
   Si $n(t) \neq 0$, on peut écrire
   $$\frac{n'(t)}{n(t)} = -k$$
   d'où la famille de solutions
   $$n(t) = De^{-kt}, \qquad D \in \mathbb{R}^+.$$
   On conclut que, quelque soit la condition initiale $n(0) = n_0 \geq 0$, l'unique solution est $n(t) = n_0 e^{-kt}$ pour tout $t \in \mathbb{R}^+$.
2. Sachant qu'il faut 5700 ans pour que la quantité de carbone 14 diminue de moitié dans un organisme mort, on peu estimer k. Puisque $n_0/2 = n(5700) = n_0 e^{-5700t}$, on obtient $k = \ln 2^{-5700} \approx 1.216\,10^{-4}$.

3. On suppose que des ossements anciens récemment exhumés contiennent 9 fois moins de carbone 14 que des ossements similaires d'aujourd'hui. On peut alors déterminer l'âge des ossements exhumés : puisque $n_0/9 = n(\hat{t}) = n_0 e^{-k\hat{t}}$, on obtient $\hat{t} = 5700\frac{\ln 9}{\ln 2} \approx 18000$ ans.

Équations différentielles linéaires du premier ordre

Elles sont de la forme

$$a(x)y'(x) + b(x)y(x) = g(x)$$

où a, b et g sont des fonctions données, continues sur un intervalle $I \subset \mathbb{R}$. Pour la résolution, on se place sur un intervalle $J \subset I$ tel que la fonction a ne s'annule pas sur J. Toute solution $y(x, C)$ de cette EDO est de la forme $y_H(x, C) + y_P(x)$ où y_P est une solution particulière de l'EDO et y_H est la solution générale de l'équation homogène associée (c'est-à-dire de l'EDO $a(x)y'(x) + b(x)y(x) = 0$). On est donc conduit à deux problèmes : rechercher d'abord la solution générale de l'équation homogène et ensuite une solution particulière de l'équation complète.

- ▷ *Résolution de l'équation homogène associée : $a(x)y'(x) + b(x)y(x) = 0$.*
  C'est une équation à variables séparables. Elles comportent donc la fonction nulle et des fonctions qui ne s'annulent jamais. On peut alors écrire formellement

$$\frac{y'(x)}{y(x)} = -\frac{b(x)}{a(x)} \qquad \Longrightarrow \qquad \ln|y| = -\int \frac{b(x)}{a(x)}\,\mathrm{d}x.$$

  Ses solutions sont du type

$$y_H(x, C) = Ce^{-A(x)} \qquad \text{où } A(x) = \int \frac{b(u)}{a(u)}\,\mathrm{d}u$$

  avec C constante arbitraire.
- ▷ *Recherche d'une solution particulière (méthode de* LAGRANGE*).*
  Si $y_1(x)$ est une solution non nulle de l'EDO homogène, on introduit une fonction auxiliaire inconnue $K(x)$ telle que $y(x) = K(x)y_1(x)$ soit solution de notre EDO. On calcule alors $y'(x)$ et on reporte $y'(x)$ et $y(x)$ dans notre EDO. On observe que $K(x)$ disparaît, ce qui fournit une auto-vérification. Il ne reste que $K'(x)$, ce qui permet de calculer $K(x)$ et donc $y_P(x)$.

L'intégrale générale est donc

$$y(x, C_1) = y_H(x, C_1) + y_P(x).$$

Exemple

On veut résoudre le problème de Cauchy

$$\begin{cases} y'(x) + (4x^3 + 5)y(x) = x^3 e^{-5x}, \\ y(0) = 1. \end{cases}$$

Il s'agit d'une EDO du premier ordre linéaire, sa solution est donc du type $y(x) = y_H(x, C) + y_P(x)$ où $y_H(x, C)$ est la famille de solutions de l'EDO homogène $y'(x) + (4x^3 + 5)y(x) = 0$ et $y_P(x)$ est une solution particulière de l'EDO complète $y'(x) + (4x^3 + 5)y(x) = x^3 e^{-5x}$ qu'on cherchera par exemple sous la forme $y_P(x) = K(x)y_H(x, 1)$.

- ▷ Pour trouver toutes les solutions de l'EDO homogène, on cherche d'abord les solutions constantes, i.e. des fonctions $y(x) \equiv A$ solution de $y'(x) + (4x^3 + 5)y(x) = 0$; pour cela il faut que

$$0 + (4x^3 + 5)A = 0;$$

  l'unique solution constante est la fonction $y(x) \equiv 0$ pour tout $x \in \mathbb{R}$. Comme deux trajectoires ne peuvent pas se croiser, toutes les autres solutions seront soit toujours positives soit toujours négatives. Soit donc $y(x) \neq 0$; comme c'est une EDO à variables séparables on peut alors écrire

$$\frac{y'(x)}{y(x)} = -(4x^3 + 5) \quad \Longrightarrow \quad \int \frac{1}{y}\,\mathrm{d}y = -\int 4x^3 + 5\,\mathrm{d}x \quad \Longrightarrow \quad \ln|y| = -4\frac{x^4}{4} - 5x + D \quad \Longrightarrow \quad y = Ce^{-x^4 - 5x} \text{ pour } C \in \mathbb{R}^*.$$

  On conclut que les solutions de l'EDO homogène sont les fonctions du type $y_H(x, C) = Ce^{-x^4 - 5x}$ pour $C \in \mathbb{R}$.
- ▷ Pour que $y_P(x) = K(x)y_H(x, 1) = K(x)e^{-x^4 - 5x}$ soit une solution particulière de l'EDO complète $y'(x) + (4x^3 + 5)y(x) = x^3 e^{-5x}$ il faut qu'elle vérifie l'EDO ; en injectant la dérivée première $y_P'(x) = K'(x)y_H(x, 1) + K(x)y_H'(x, 1) = K'(x)e^{-x^4 - 5x} + K(x)((-4x^3 - 5)e^{-x^4 - 5x})$ dans l'EDO complète on obtient

$$K'(x)e^{-x^4 - 5x} + \cancel{K(x)((-4x^3 - 5)e^{-x^4 - 5x})} + \cancel{(4x^3 + 5)K(x)e^{-x^4 - 5x}} = x^3 e^{-5x} \quad \Longrightarrow \quad K'(x) = x^3 e^{x^4} \quad \Longrightarrow \quad K(x) = \frac{e^{x^4}}{4}.$$

  On conclut qu'une solution de l'EDO complète est la fonction $y_P(x) = \frac{1}{4}e^{x^4} e^{-x^4 - 5x} = \frac{1}{4}e^{-5x}$.

Toutes les solutions de l'EDO sont donc les fonctions $y(x) = \left(C - \frac{e^{-x^4}}{4}\right)e^{-x^4 - 5x}$ pour $C \in \mathbb{R}$.

On cherche parmi ces solutions celle qui vérifie $y(0) = 1$; comme $y(0) = C + \frac{1}{4}$, l'unique solution du problème de Cauchy donné est la fonction $y(x) = \left(\frac{3}{4} + \frac{e^{x^4}}{4}\right)e^{-x^4 - 5x}$.

Équations différentielles de BERNOULLI

Elles sont du premier ordre et de la forme

$$a(x)y'(x)+b(x)y(x)=g(x)(y(x))^{\alpha}, \qquad \alpha\in\mathbb{R}\setminus\{0;1\}$$

où a, b et g sont des fonctions données, continues sur un intervalle $I\subset\mathbb{R}$. Pour la résolution, on se place sur un intervalle $J\subset I$ tel que la fonction a ne s'annule pas sur J et on fait le changement de fonction $z=y^{1-\alpha}$. On arrive à

$$a(x)z'(x)+(1-\alpha)b(x)z(x)=(1-\alpha)g(x)$$

qui est une équation linéaire du premier ordre.

Exemple

On veut résoudre l'équation différentielle

$$y'(x)+\frac{1}{2}y(x)=\frac{1}{2}(x-1)y^3(x).$$

Il s'agit d'une équation différentielle de BERNOULLI. Comme $a(x)=1$, on cherche sa solution générale sur $\mathbb{R}$.

▷ *Réduction à une équation différentielle linéaire du premier ordre :* si on pose $z(x)=1/y^2(x)$ elle se réécrit

$$z'(x)-z(x)=1-x.$$

▷ *Résolution de l'équation homogène associée :* $z_H'(x)-z_H(x)=0$.

$z_H(x)=0$ est solution et on sait que les autres solutions ne s'annulent pas sur $\mathbb{R}$. On peut donc écrire $\frac{z_H'(x)}{z_H(x)}=1$, ce qui conduit à la solution générale sur $\mathbb{R}$ de l'équation homogène

$$z_H(x,C)=Ce^x \qquad \text{avec } C\in\mathbb{R}^+.$$

▷ *Recherche d'une solution particulière (méthode de* LAGRANGE*).*

Considérons une nouvelle fonction inconnue $K(x)$ telle que $z_P(x)=K(x)z_H(x,C)=K(x)e^x$ soit solution de $z_P'(x)-z_P(x)=1-x$. On calcule $z_P'(x)=(K'(x)+K(x))e^x$ et on le reporte dans l'EDO ; on obtient

$$\left[K'(x)+K(x)\right]e^x-K(x)e^x=1-x \qquad\Longrightarrow\qquad K'(x)=\frac{1-x}{e^x}.$$

On en déduit

$$K(x)=\int(1-x)e^{-x}\,\mathrm{d}x=-(1-x)e^{-x}-\int e^{-x}\,\mathrm{d}x=xe^{-x}$$

et donc

$$z_P(x)=x.$$

La solution générale est donc

$$z(x,C)=z_H(x,A)+z_P(x)=Ae^x+x \qquad \text{avec } A\in\mathbb{R}$$

et on conclut

$$y(x,C)=\frac{1}{\sqrt{x+Ae^x}}.$$